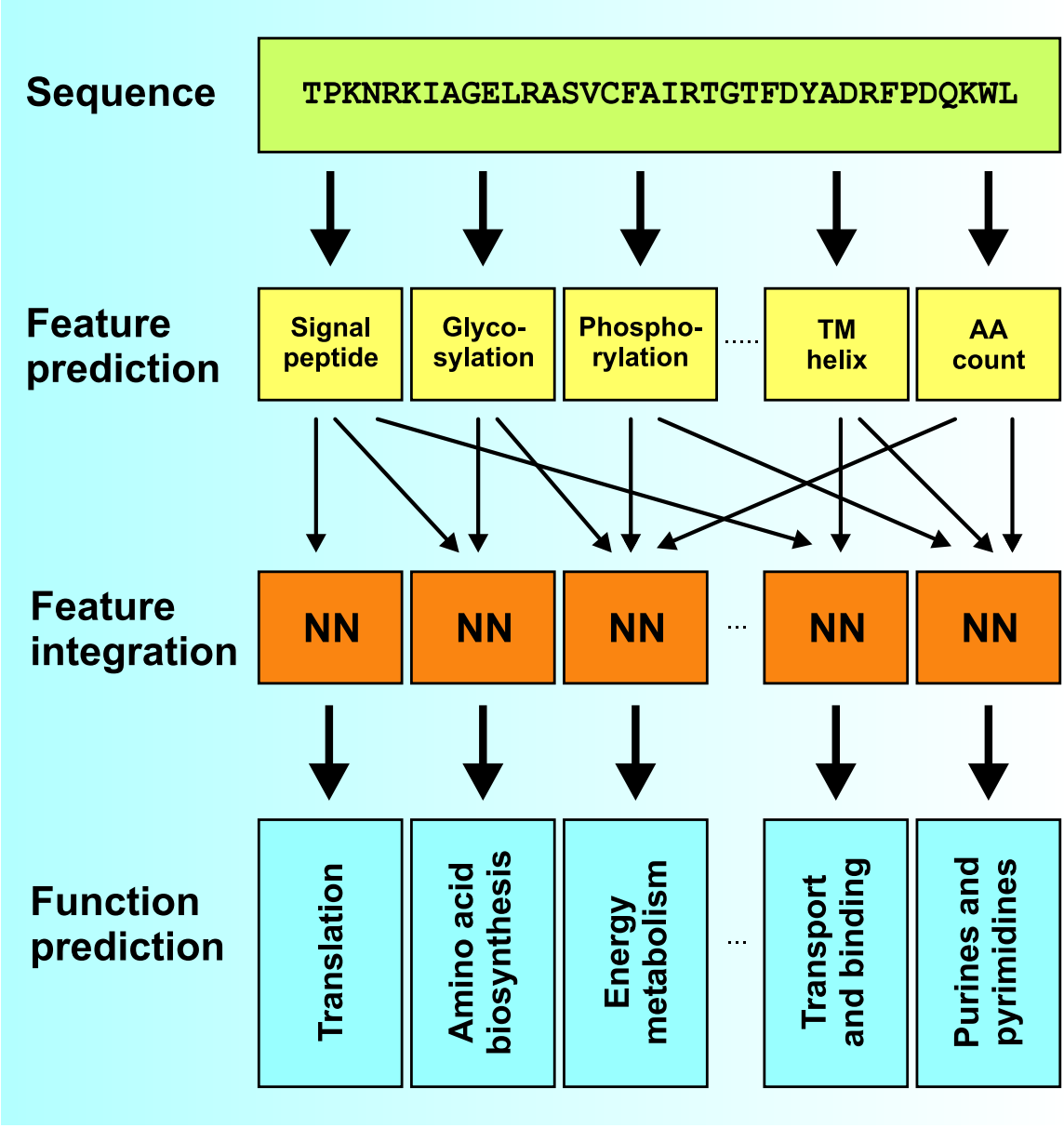
Prediction of Protein Function
from Sequence Derived Protein Features



Lars Juhl Jensen
2002

DTU

Prediction of Protein Function from Sequence Derived Protein Features

Ph.D. thesis
Lars Juhl Jensen

Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby
Denmark

Lyngby 2002

Abstract

Today, many genomes have been sequenced (including that of humans), and the pace at which new genomes are being sequenced is constantly increasing. When analyzing the protein coding genes, it is typically only possible to assign function to half of all protein sequences based on sequence similarity. There is thus a great need for sequence-based protein function prediction methods.

Even if a protein sequence is not similar to any other known protein, it will still have to perform its function in the same cellular context—making use of the same molecular machinery. As this will be reflected in the properties of the protein, it is reasonable to expect that proteins with related function will have similar properties, even if they are not evolutionarily related. By using neural networks to integrate a large number of predicted protein features, a protein function prediction method has been successfully developed. The features used for encoding a protein sequence include the predicted subcellular localization of the protein, post/co-translational modifications, and protein structure (in the form of secondary structure and transmembrane helices).

As protein function is multi-faceted, many definitions of "function" exist. To cope with this, predictors have been trained for both cellular role categories, enzyme classification, and a subset of Gene Ontology classes (several of which are interesting from a pharmaceutical point of view). Even though the method is capable of generalizing between very distantly related organisms, predictors have also been trained for several organisms.

Furthermore, the representation of proteins in feature space has been studied, leading to several interesting discoveries concerning the evolution as well as function of proteins.

Dansk resume

Forudsigelse af proteinfunktion fra sekvensafledte proteinegenskaber

I dag, hvor adskillige hele genomer er sekventeret (herunder det humane), kan et bud på sekvensen af alle proteiner i de tilsvarende proteomer fås ved genfindingsmetoder. På basis af sekvenssammenligning med proteiner fra andre organismer er det muligt at tilskrive en funktion til omkring halvdelen af disse—funktionen af den anden halvdel er fortsat ukendt. Det vil derfor være meget værdifuldt at have en metode til at forudsige funktionen af disse proteiner ud fra deres sekvens.

Selv om proteinernes sekvens ikke ligner noget andet vi kender, skal de dog fungere i det samme cellulære miljø og håndteres af de samme biologiske mekanismer som øvrige proteiner. Det kan derfor forventes, at proteiner med lignende funktion har visse lignende egenskaber, til trods for at deres sekvenser er helt forskellige. Ved at forudsige et stort antal egenskaber for proteinerne og benytte disse som input til kunstige neurale netværk, er det lykkedes at udvikle en metode til forudsigelse af proteiners funktion. Egenskaberne som benyttes repræsenterer hvor i en celle proteinet befinder sig, hvordan proteinet bliver modificeret (såkaldte post-translationelle modifikationer) samt forskellige aspekter af proteinets struktur (specielt sekundær struktur og position af eventuelle transmembrane segmenter).

Da der findes flere forskellige definitioner af "proteinfunktion" er adskillige forudsigere blevet trænet. Metoden er således nu i stand til at forudsige såvel proteiners cellulære rolle, deres eventuelle enzymklasse samt et antal farmaceutisk interessante proteinklasser (herunder receptorer, hormoner og vækstfaktorer).

For at forstå den biologiske baggrund for at forudsigerne fungerer, er de kunstige neurale netværk som er blevet trænet for hver proteinklasse, blevet analyseret. Herved er et antal interessante karakteristika for forskellige typer af proteiner blevet opdaget.

Preface

This Ph.D. thesis is written for BioCentrum-DTU, Technical University of Denmark under the Biotechnology programme.

The research has been done at the Center for Biological Sequence Analysis at BioCentrum-DTU, where it has been supervised by Professor Søren Brunak.

The project has been financed by a Ph.D. stipend from the Technical University of Denmark.

Acknowledgments

I wish to thank everybody who have helped make this project not only possible but also a very pleasurable experience indeed.

- My supervisor, Søren Brunak, for his inspiration and for trusting me and giving me an unusual amount of freedom to do what I felt was right.
- Damien Devos, Javier Tamames, and Alfonso Valencia for providing valuable information on how to do automated function assignment and its many flaws.
- David Ussery for proofreading this thesis and correcting countless grammatical errors.
- All the other co-authors on the papers included in this thesis: Marie Skovgaard, Ramneek Gupta, Anders Krogh, David Ussery, Steen Knudsen, Can Kesmir, Nikolaj Blom, Peder Worning, Ulrik de Lichtenberg, Thomas Skøt Jensen, Carsten Friis, Hanne Jarmer, Christopher Workman, Henrik Nielsen, Claus A. F. Andersen, Hans-Henrik Stærfeldt, and Kristoffer Rapacki.
- Thanks in particular to Ramneek Gupta and Can Kesmir for mapping out the mine field of function prediction before I started—often by personally stepping on the mines, I might add.
- Kristoffer Rapacki for always being helpful, for countless intellectual conversations, and not the least for a fantastic hike on the Mt. Olympus.
- David Ussery for lots of pleasant discussions—and liquorice.
- Lone Boesen, Ilsabe Lampe, and Johanne Keiding for patiently helping the rest of us with all kinds of problems we constantly get ourselves into.
- My parents for always supporting me and never questioning my decisions.
- Finally, everybody at CBS for making it such a fun and friendly place to work.

Publications

Publications included in this thesis

Paper I

Marie Skovgaard, Lars Juhl Jensen, Søren Brunak, David W. Ussery, and Anders Krogh.
On the total number of genes and their length distribution in complete microbial genomes.
Trends in Genetics, **17**, 425–428, 2001.

Paper II

Lars Juhl Jensen, Ramneek Gupta, Nikolaj Blom, Damien Devos, Javier Tamames, Can Kesmir, Henrik Nielsen, Hans-Henrik Stærfeldt, Kristoffer Rapacki, Christopher Workman, Claus A. F. Andersen, Steen Knudsen, Anders Krogh, Alfonso Valencia, and Søren Brunak.
Prediction of human protein function from post-translational modifications and localization features.
Journal of Molecular Biology, **319**, 1257–1265, 2002.

Paper III

Ramneek Gupta, Lars Juhl Jensen, and Søren Brunak.
Orphan function prediction and its relation to glycosylation.
Ernst Schering Research Foundation Proceedings, **38**, Chapter 13, 275–294, 2002.

Paper IV

Lars Juhl Jensen, David W. Ussery, and Søren Brunak.
Functionality of system components: Conservation of protein function in protein feature space.
Genome Research, **13**, 2444–2449, 2003.

Paper V

Ulrik de Lichtenberg, Thomas Skøt Jensen, Lars Juhl Jensen, and Søren Brunak.
In-silico Proteomics of the yeast cell cycle.
Journal of Molecular Biology, **329**, 663–674, 2003.

Paper VI

Lars Juhl Jensen, Marie Skovgaard, and Søren Brunak.
Prediction of novel enzymes in hyperthermophile archaea.
Protein Science, **11**, 2894–2898, 2002.

Paper VII

Lars Juhl Jensen, Ramneek Gupta, Hans-Henrik Stærfeldt, and Søren Brunak.
Prediction of human protein function according to Gene Ontology categories.
Bioinformatics, **19**, 335–342, 2003.

Paper VIII

Marie Skovgaard, Lars Juhl Jensen, Carsten Friis, Hans-Henrik Stærfeldt, Peder Worning, Søren Brunak, and David W. Ussery.
The atlas visualization of genome-wide information.
Methods in Microbiology, **33**, Chapter 3, 49–63, 2002.

Additional papers

Peder Worning, Lars Juhl Jensen, Karen E. Nelson, Søren Brunak, and David W. Ussery.
Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*.
Nucleic Acids Research, **28**, 706–709, 2000.

Christopher Workman, Lars Juhl Jensen, Hanne Jarmer, Randy Berka, Laurent Gautier, Hans-Henrik Saxild, Claus Nielsen, Søren Brunak, and Steen Knudsen.
A new non-linear normalization method to reduce variability in DNA microarray experiments.
Genome Biology, **3**, 0048.1-0048.16, 2002.

Vera van Noort and Lars Juhl Jensen.
A new method to reduce variability in oligonucleotide microarray data based on probe DNA sequences.
Submitted to *Bioinformatics*, 2002.

Contents

Part I

Introduction

Chapter 1

Systems biology and protein function

Currently one the hottest topics in molecular biology is *systems biology*. It covers the study of entire biological systems as one entity—in sharp contrast to the traditional reductive approach of molecular biology where one gene is studied at a time. However, the term is still so new that it is unclear exactly what is meant by "systems biology" (Bonetta, 2002).

The vast amounts data generated by modern high throughput methods (e.g. whole genome shotgun sequencing) forms the foundation of systems biology. One of the major challenges will be to turn the raw data into biological discoveries. Due to the sheer amounts of data, computational methods must be developed for analyzing them. For this reason, bioinformatics is quickly becoming a central part of systems biology.

1.1 The age of 'omes

The systems biology approach to molecular biology has brought with it not only vast amounts of data, but also a number of new words. While the word *genome* itself was coined by H. Winkler in 1920[1] (even though scientists did not then know what it was made of), the related word *genomics* as well as *transcriptomics* and *proteomics* are all fairly new terms describing previously non-existent areas of research.

1.1.1 Whole genome sequencing

The age of 'omes is still young, not really starting until 1995 when the 1,830,138 basepair genome of *Haemophilus influenzae* was published (Fleischmann et al., 1995). This, the sequencing of the of the first complete bacterial genome, is one of the milestones in molecular biology.

Only 3 month later the same research group published the second complete bacterial genome, that of *Mycoplasma genetalium* (Fraser et al., 1995). Consisting of a mere 580,074 basepairs, it is the smallest and most compact genome

[1] First recorded use of the word according to The Oxford English Dictionary

sequenced yet (with the exception of viral genomes). From thereon, the pace at which new genomes are being sequenced has been constantly increasing. In 1996 three more genomes were published: a second mycoplasma (Himmelreich et al., 1996), a cyanobacterium (Kaneko et al., 1996), and the first archaeal genome (Bult et al., 1996). Yet another 6 genomes were sequenced in 1997 alone. These included three of the most important model organisms, namely the Gram-negative *Escherichia coli* (Blattner et al., 1997), the Gram-positive *Bacillus subtilis* (Kunst et al., 1997), and the yeast *Saccharomyces cerevisiae* (Goffeau et al., 1997). The latter got much attention, as it was the first eukaryotic genome to be sequenced—and that only two years after sequencing of the first bacterial genome was completed.

Around this time people in the field started to realize, that while genomes might contain all the genetic information underlying an organism, extracting and interpreting this information is far from trivial. Since then it has become increasingly evident, that the work has only just begone when a genome has been sequenced. After all, the purpose of the recently finished sequencing of the human genome (Int. Human Genome Sequencing Consortium, 2001; Venter et al., 2001) was not to get strings of three billion G's, A's, T's, and C's—we want to gain insight of how human cells work.

1.1.2 The transcriptome

1997 also marked the entry of another 'ome: the transcriptome. By spotting cDNA of each individual gene in a genome onto glass slides, the mRNA expression levels of all genes can be measured simultaneously by hybridization (Schena et al., 1995). This technique is commonly referred to as *Stanford cDNA microarrays* or just microarrays for short.

The first transcriptome analysis using microarrays was an analysis of the diauxic shift in *S. cerevisiae* (DeRisi et al., 1997). Around the same time, the first transcriptome analysis of the *S. cerevisiae* mitotic cell cycle was also published (Velculescu et al., 1997). In this experiment the mRNA concentrations were measured at three timepoints using a very different method called *serial analysis of gene expression* (Velculescu et al., 1995). The year after, the cell cycle was analyzed at more timepoints using a cDNA microarray based analysis (Spellman et al., 1998).

Today two methods are being used extensively for transcriptome analysis: the *Stanford cDNA microarrays* that have already been described and the *Affymetrix GeneChips*. The latter also works by hybridization, although instead of spotting the probes, they are synthesized *in situ* by a photolithographic process (Lipshutz et al., 1999).

Transcriptome data contain much information which is not evident from the genome sequence. Clustering of genes that show similar expression patterns can be used for identifying genes with similar cellular functions (Eisen et al., 1998; Hughes et al., 2000a). Analysis of the promoter regions of co-expressed genes can further reveal novel promoter elements (Pesole et al., 1992; Brazma et al., 1997, 1998; van Helden et al., 1998; Jensen and Knudsen, 2000; Lawrence et al., 1993; Workman and Stormo, 2000). Systematic analysis of gene expression changes

in knockout (and/or over-expression) mutants can reveal regulatory coupling of genes. This is called regulatory network reconstruction, and is probably the most ambitious use of transcriptome data to date.

Although it is easy to get carried away by this, it is important to keep in mind that the ultimate product of most genes is not RNA but protein. Therefore, many aspects of a cell cannot be understood by studying the transcriptome. This is a large part of the reason why reconstruction of regulatory networks have proven a very elusive goal indeed.

1.1.3 Proteomics

The proteome covers the full complement of proteins in a cell and is immensely complex compared to the genome and the transcriptome. Proteomics take into account not only the sequence of each protein but also their location and all co- and post-translational modifications (Pandey and Mann, 2000). This includes reversible modifications responsible for regulation at the protein level, making the proteome a dynamic entity like the transcriptome. Because the proteome encompasses essentially all aspects of proteins, it is also intimately related to understanding protein function at the systems level.

This increase in complexity when going from the genome to the proteome is not equal for all organisms, as more complex organisms tend to have much more elaborate regulatory mechanisms at the protein level. Consequently, the number of genes in a genome has turned out to be a poor measure for the complexity of the organism (Claverie, 2001), like the genome size was realized to be many years ago (Callan, 1972). Another more unfortunate consequence is that knowing the genome sequence of a higher eukaryote (e.g. *Homo sapiens*) may turn out to be much less informative than would be anticipated from the analysis of lower eukaryotes (e.g. *Saccharomyces cerevisiae*).

Due to the inherent complexity of the proteome, no methods currently exist for "measuring the proteome". Instead different experiments focus on different aspects of proteomes: protein concentration, subcellular localization, phosphorylation or glycosylation.

1.1.4 The many other 'omes

The words genome, transcriptome, and proteome seems to have started a trend: whenever something new is studied for a complete cell, a new 'ome is coined (together with the related 'omics research field). Examples include the physiome, metabolome, interactome, glycome and secretome (Greenbaum et al., 2001).[2]

As more and more 'omes are being defined, they also start to overlap more and more. For instance it can be argued that both the glycome and the secretome are parts of the proteome. Although the proteome is very complex and difficult to understand, this tendency to break the proteome into smaller parts is unfortunate, as it would seem to defy the purpose of proteomics: large scale data integration through which proteins can be analyzed in their proper context.

[2]The many 'omes invented by now have caused Henrik Nielsen from CBS to suggest one more: the omome, defined as the complete set of 'omes describing an organism.

There are several databases which claim to be genome or proteome databases, e.g. the Yeast Proteome Database (YPD) (Hodges et al., 1998). Although data are collected for the whole genome, the focus is on providing as much information as possible for the individual gene—*not* on standardizing these data to allow simultaneous analysis of many genes (Greenbaum et al., 2001). It may sound like a paradox, but today's genome databases are simply not designed for doing genomics.

However many 'omes may be defined, the ultimate goal is to elucidate all functions of all gene products and thereby understand how a cell works—only then can we really claim to have "solved" a genome (Johnson, 2000; Greenbaum et al., 2001). This task will most likely require all the 'omes to be integrated.[3]

1.2 "Protein function"—an important but fuzzy concept

The majority of functions in cells are governed by proteins, be they enzymatic or not. Assigning functions to the proteins encoded by a genome it therefore one of the crucial steps in gaining understanding of the organism. Because the function of half of all proteins in newly sequenced genomes often is completely unknown, complete genome sequencing gives much less insight into the organism than initially hoped for (Walhout and Vidal, 2001).

Even though everybody talk about the "function" of various proteins, there is still no clearcut definition of what "function" actually means. Instead the is a multitude of different interpretations of what the word means.

One interpretation is the precise biochemical actions of the protein, which today is often termed the *molecular function*. This definition is particularly useful for enzymes, where the function then becomes a matter of which chemical reactions the enzyme catalyze. However attractive this definition may seem, knowing the chemistry a protein is involved in does not always provide much understanding of why the cell needs the protein.

1.2.1 Function and interactions

Part of the reason why it is difficult to relate the chemical function of a protein to its biological purpose is that proteins do not function alone. To understand the function of a protein, it must be considered in its proper cellular context, for example by appreciating how the cell would behave without it (Attwood and Miller, 2001). Many proteins are parts of larger complexes, which are the functional units that fulfill a role in the cell (Gavin et al., 2002). In this case it can be argued that all the proteins that form the complex should also have the same function.

Since a protein does not perform its function alone but in the context of many other proteins as well as other biomolecules, it is highly relevant to study

[3]Although the omome was originally intended as a joke, it may turn out to be a completely meaningful term.

the interaction partners of a protein in order to understand its function (Bork, 2000; Eisenberg et al., 2000; Yaspo, 2001; Ho et al., 2002). In fact the somewhat fuzzy concept of protein "function" is by some defined as the protein's interactions with other substances (Karp, 2000; Ho et al., 2002). To see that this is not an unreasonable definition, one only needs to try to come up with a possible function of a protein which does not interact with anything.

In line with this argumentation, a very different definition of "function", has been suggested: that of *cellular role*. Rather than giving a precise description of what a protein does at the molecular level, the cellular role tries to describe what use the cell has of a protein. As it is not clear exactly how these cellular roles are best defined, several closely related classification systems have been developed for use on different organisms (Riley, 1993; Andrade et al., 1999b).

1.2.2 Breadth and depth of function classification

Classifying proteins is difficult because many mutually overlapping definitions of function exist. In addition to molecular functions and cellular roles, one can also look at how a protein is involved in, for example, physiology, development or diseases (Liu and Rost, 2001). When biologists ask for the function of a protein, they are usually not referring to any one of the above definitions of protein function. Instead they want a sentence or two describing the biologically interesting aspects of the protein (Benner and Gaucher, 2001).

These different parallel systems for classifying different aspects of protein function, is what I call the breadth of protein function prediction. Many of the schemes also have different levels of detail at which the function can be described, which lead to the distinction between "narrow and deep" function descriptions and "broad and shallow" descriptions (Lewis et al., 2000). This is commonly known as the depth of function classification.

The accuracy with which function can be predicted by various methods should be expected to depend strongly on both the definition of function and the level of detail that is used (Benner and Gaucher, 2001). For instance co-expressed genes could be expected to be involved in the same cellular role, but do not necessarily have similar molecular function. Conversely, proteins with very similar structure would be likely to have similar molecular functions, but they may play entirely different roles in the cell.

1.2.3 Controlled vocabularies for describing function

Until recently, databases contained only free text descriptions of protein function. While such annotation is great for the biologist studying a few proteins, it is next to worthless for large scale studies.

The main problem with free text annotation is that it is only human readable—it cannot be parsed by a computer. A second problem is, that having different people write down protein function as free text almost encourages inconsistent annotation. Together these problems make it next to impossible to use free text annotation for large automated jobs, e.g. annotation of new genomes by comparative genomics methods (Lewis et al., 2000).

The solution is to create controlled vocabularies or ontologies for protein function. Several such function classification schemes exist, many of which have been developed in the past few years: the enzyme classification system (Enzyme Nomenclature, 1965, 1992), SWISS-PROT keywords (Bairoch and Apweiler, 2000), cellular role classes and and interaction based ontology for *E. coli* proteins (Riley, 1993; Karp, 2000), the MIPS classes (Mewes et al., 2002), the Gene Ontology system (Ashburner et al., 2000) as well as specialized classifications like the IMGT-ONTOLOGY (Giudicelli and Lefranc, 1999). Making these many alternative classification systems converge to one single standardized ontology for protein function is one of the key challenges in automated large scale function annotation and comparative genomics (Weir et al., 2001). Fortunately a *de facto* standard appears to be emerging, this standard being Gene Ontology.

Chapter 2

Assignment and prediction of protein function

One of the most fundamental tools in the field of bioinformatics is sequence alignment. By aligning sequences to one another, it is possible to evaluate how similar the sequences are and identify conserved regions in sets of related sequences. This is used extensively to assign function to genes in newly sequenced genomes.

Sequence alignment also forms the basis for many other types of bioinformatics analysis. All practical methods for reconstructing phylogenetic trees rely on sequence alignment, some explicitly require a multiple alignment as input while others require a matrix of all pairwise evolutionary distances. However, such matrices are estimated either from a multiple alignment or from a set of all pairwise alignments. Furthermore alignment is used for ensuring independence between test and training sets used as input for machine learning algorithms like artificial neural networks.

2.1 Pairwise alignment

The first and simplest type of alignment to be developed was pairwise alignment, which as the name implies allows only two sequences to be compared at a time. Today it is still the most commonly used form of sequence alignment, because it can be used to answer one of the most basic questions in sequence analysis: how similar is sequence A to sequence B?

2.1.1 Alignment scores

To be able to optimally align two sequences, one must first define a scoring function to evaluate the quality of an alignment. The scoring function used is sum over the similarity of aligned residues combined with a penalty function for gaps in the alignment. A substitution matrix is used to define the pairwise similarity scores between residues. Many different substitution matrices have been derived by various approaches (Dayhoff et al., 1978; Henikoff and Henikoff, 1992; Benner et al., 1994), the most commonly used one being the BLOSUM62 matrix. The gap penalties are usually defined as affine functions, i.e. the penalty for each gap is a linear function of the length of the gap.

It is important to realize that both the alignment score as well as the alignment itself will depend on the choice of substitution matrix and gap penalties. The "optimal alignment" found for two sequences is therefore not necessarily the best one from a biological perspective.

2.1.2 Global and local alignment algorithms

In theory the optimal alignment of two protein sequences could be found by scoring all possible alignments to find the best one. In practice this is not possible as the number of possible alignments grows exponentially with the sequence length. Fortunately, there is a smarter solution for the scoring function described above, as the optimal alignment can be found using dynamic programming algorithms. The runtime for aligning two sequences using these algorithms is proportional to the product of the two sequence lengths.

Two different variations of dynamic programming are frequently used in bioinformatics: the Needleman–Wunsch algorithm for global alignment (Needleman and Wunsch, 1970) and the Smith–Waterman algorithm for local alignments (Smith and Waterman, 1981). Global alignment algorithms align two sequences in their entirety—as a result, if two sequences only share a common protein domain, the alignment score will be very poor. While local alignment methods are also capable of aligning two entire sequence if they are similar over their entire length, they are also able to find regions of similarity between two otherwise unrelated sequences. This makes local alignment the more flexible of the two, for which reason it is also the most used of the two.

2.1.3 Fast heuristic alignment methods

While the dynamic programming algorithms are very fast for aligning two typical protein sequences, it is still prohibitively slow for comparing a complete proteome with all known protein sequences from other organisms. To address such massive pairwise alignment problems various heuristic alignment methods have been developed, the two best known programme packages being FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1997). Since the introduction of gapped BLAST in 1997 most people have been using BLAST.

The heuristic alignment methods can drastically speedup database searches compared to dynamic programming. The price paid is that the method is no longer guaranteed to find the optimal alignment, hence matches that would be found by a full dynamic programming approach can in theory be lost when using methods like BLAST. However, benchmarks have shown that in practice this happens very rarely (Park et al., 1998). The BLAST programme has therefore been used instead of the Smith–Waterman algorithm for the majority of sequence comparisons in this thesis.

2.1.4 Statistics of pairwise alignment

At this point it should not come as a surprise to anyone, that pairwise alignment algorithms align sequences. However, one should bear in mind that they always

do so—whether it makes sense or not. It is for this reason important to have a framework for evaluating the statistical significance of a pairwise alignment.

The significance of a match can be evaluated by comparing its alignment score to the distribution of maximal alignment scores for independent sequences. In the case of ungapped alignment, the scores for independent sequence alignments have been proven to follow a normal distribution, and the distribution for the highest score can be approximated then by the *extreme value distribution* (Karlin and Altschul, 1990). A similar result has never been proven for gapped alignments, but it has been convincingly argued to hold for reasonable choices of gap penalties (Mott, 2001; Waterman and Vingron, 1994). From this it follows that the expected number of matches with a score better than S between two sets of sequences (either of which could be a single sequence) can be written as:

$$E(S) = Knme^{-\lambda S},$$

where n and m are the number of residues in the two sets of sequences, while K and λ are constants that depend on the choice of substitution matrix and gap penalties.

It may seem odd at first that the same alignments become less significant (or even insignificant) when the database is made larger. It is obvious that the alignment scores arising from truly evolutionarily related proteins remain constant as the database grows. However, the noise arising from the database increases with the size of the database, giving a worse signal to noise ratio, thus decreasing the credibility of the alignment (Spang and Vingron, 2001).

2.2 The problems of inferring function by similarity

Pairwise alignment is likely the single most important tool in a bioinformatician's toolbox. However, this does not mean that pairwise alignment methods are flawless. An overview will be given of the most important issues when using pairwise local alignment algorithms in the context of database searches.

2.2.1 Similarity vs. homology

When using sequence alignment methods to assign function based on similarity, there are several different concepts that tend to get confused with each other. The most common case is to confuse the concepts "similarity" and "homology". When saying that two proteins are "similar", one simply states the observation that two protein sequences look alike. In sharp contrast to this, stating that two proteins are "homologs" implies that the two proteins have evolved from a common ancestor (or that one protein has evolved from the other) (Fitch, 1970).

Homology search methods are based on a strictly statistical rationale: if the similarity detected is very unlikely to occur by chance, there must be a different explanation (Spang and Vingron, 2001). It is normally assumed that the explanation is that the two proteins in question are related through evolution.

Although this is often the case, one should keep in mind that statistical tests never provide explanations—they only state that something is highly unlikely to occur by chance alone.

One alternative explanation to evolutionary relationship that can give rise to similar sequences is low complexity regions. For instance, two serine–threonine rich proteins are bound to have local high similarity, simply because they have the same highly biased amino acid composition. Several other such examples exist (e.g. leucine rich proteins). Similar simple sequences can easily evolve independently and thus tend to fool pairwise alignment search tools. To circumvent this problem, sequences are commonly masked by a so-called low complexity filter, which simply excludes all residues that, according to some criteria, belong to a low complexity region. However, this means that it is very difficult to find homologs for sequences containing large amounts of low complexity regions.

A related problem to low complexity regions are transmembrane helices. While they are not low complexity regions as such, they share certain features, such as a bias in amino acid composition. Transmembrane helices tend to cause the same types of problems in sequence searches as low complexity regions, namely that two transmembrane helices will look alike to alignment methods simply because they are both transmembrane helices. One way to approach this problem is to create special substitution matrices designed for transmembrane helices (Ng et al., 2000).

2.2.2 The detection limit of pairwise alignment methods

Not only can two proteins be similar without being homologs—remotely related homologs can also be quite dissimilar in sequence. A consequence of the latter is that it is very difficult to transfer functional information from remote homologs because the homology cannot be established based on very weak sequence similarity. There is thus an inherent limit to how remote homologs can be detected by pairwise alignment methods. This is the main reason why bioinformaticians are constantly on the lookout for new ways to improve sequence similarity search methods.

Although it is never possible to infer homology with absolute certainty based on sequence similarity, it is generally safe to assume that very similar sequences are close homologs. For this reason the function of a protein is rarely assigned incorrectly based on a match to a very similar protein.

2.2.3 Orthologs vs. paralogs

There are unfortunately exceptions to even this rule. While very similar sequences are almost guaranteed to be homologs, this guarantee does not extend to the proteins necessarily having the same function. Knowing that two proteins are related through evolution is not enough, it is also important how they are related. This is in particular true when dealing with remote homologs.

In this context it is crucial to distinguish between homologs, orthologs and paralogs. Two proteins are said to be orthologous if they have arisen through a speciation event, while they are said to be paralogous if they they stem from

a gene duplication (Fitch, 2000). In the case of paralogous proteins, one of the two copies that have arisen by the gene duplication will often mutate to fulfill a different functional role in the cell. It is thus more dangerous to transfer functional information between paralogs than to do so between orthologs (Jensen, 2001).

This has sparked an interest in computational methods for determining if two homologous sequences are in fact orthologs or paralogs. The right way to do this would be to reconstruct phylogenetic trees for related genes to establish their internal evolutionary relationship. Once this has been established, it is trivial to extract which pairs of proteins are orthologs and which are paralogs. Unfortunately, phylogenetic tree reconstruction is as much a craft as it is a computational method, and it has therefore proven very difficult to successfully automate this procedure. However, it has turned out to be possible to discriminate between orthologs and paralogs using a much simpler heuristic method (INPARANOID) which is closer related to clustering than to tree reconstruction (Remm et al., 2001).

2.3 Methods for assigning function based on pairwise sequence similarity

The most commonly employed method for assigning function to a protein is to infer it from a homologous protein found by sequence alignment. This is likely the main usage of sequence alignment methods. Recent improvements in search algorithms combined with the tremendous growth in sequence databases have resulted in this approach being ever more powerful.

2.3.1 Best match

The simplest possible way to utilize sequence alignment for assignment of protein function, is to simply search the sequence of unknown function against a database of proteins with known function. If one or more significant matches are found, the function of the best match is transfered to the query sequence. There is of course always a risk that the homologs have different functions, especially for remote homologs (Devos and Valencia, 2000).

One of the big problems with the best match method is to know how much information can be safely transfered. If, for instance, a protein is a remote homolog of a tryptophan transporter, it may be that it is an amino acid transporter but for a different amino acid than tryptophan (Casari et al., 1996). It is difficult to do anything about this problem, in particular if the functional annotation is not available in a computer readable form.

2.3.2 EUCLID

A more advanced method for assigning protein function from pairwise sequence alignment is the EUCLID method (Tamames et al., 1998; Andrade et al., 1999a). This method assigns proteins to the broad cellular role categories originally de-

fined for the *E. coli* genome (Riley, 1993). The EUCLID method has been used extensively for the work presented in this thesis.

For each protein sequence to be assigned, the method first identifies all BLAST matches in SWISS-PROT with an expectation value better then 10^{-5}. The keywords assigned to each of these SWISS-PROT entries are looked up, and a consensus subset of keywords is identified.

Based on large number of SWISS-PROT entries that were manually assigned to cellular role categories, a dictionary has been created. Every keyword in this dictionary is associated with a score for each cellular role category. To classify a protein of unknown function, a score is calculated for each cellular role by summing over the dictionary scores for all the consensus keywords. The protein is then assigned to the category that receives the highest score. A quality factor for the assignment is also calculated as the relative difference between the highest and the second highest score. However, this quality measure does not compare well across different categories (Damien Devos, personal communication). For our purpose the scores will be used differently, to allow a single protein to belong to more than one class. This modification is described in Chapter 4.

By looking for the consensus keyword of all matches rather than looking at only the best match, EUCLID avoids the danger of transferring too much functional information from a single homolog. But the most important advantage over the best match method is, that EUCLID assigns protein function according to a classification system whereas the best match algorithm transfers the annotated function from a database entry. The latter is problematic because most of the functional annotation in databases is available only as free text, as was discussed in Chapter 1.

2.4 Utilizing multiple sequences by iterative search methods

One inherent problem of pairwise alignment of sequences is, that there is only so much information in two sequences. Sometimes it is therefore simply not possible from two sequences to reliably tell if they share a common evolutionary background—and if they are thereby likely to have a common function. The way to address this problem is quite simply to use more than two sequences. EUCLID does this partially as the functional annotation of all database matches are considered when assigning function. However, EUCLID does not make use of the additional information available when having several related sequences in the alignment step.

2.4.1 Intermediate sequence search

The gradual change of one sequence into another through evolution can—although the mutations are small discrete events—be viewed as a trajectory in sequence space. This is the rationale behind a simple way to make use of more than two sequences when inferring evolutionary relationships: *intermediate sequence search* also known as ISS or *transitive homology search*. Such methods

have been shown to give 24% better sensitivity than pairwise alignment at same specificity (Bolten et al., 2001).

In its simplest form, an intermediate sequence search consists of two steps. First the query sequence is searched against a large database to identify all sequences with significant similarity to the query. These sequences, the intermediate sequences, are then all searched against the database to again identify all similar proteins. The sequences found in this second step are considered to be linked to the query sequence through an intermediary sequence. Sequences identified in this way may not themselves align sufficiently well to the query sequence to be picked by traditional pairwise alignment methods.

An obvious extension of this system presents itself: rather than stopping after the second round of database searches, why not continue and use the new sequences obtained to search once again? This method, which could be continued for any number of steps or until convergence, is known as multiple intermediary sequence search (MISS). While it may be an obvious extension, it has however not been a very successful one.

ISS methods in general and MISS in particular, seem to have been plagued by one problem especially. While good methods exist for estimating the significance of pairwise alignment scores, it is not obvious how to combine the E-values of the individual pairwise alignments into one ISS E-value. It is thus difficult to assess if two sequences found through ISS methods are in fact significantly related.

As a result of the lacking statistics, it has turned out to be very difficult to automate MISS methods (Li et al., 2000). Since it is not possible to reliably evaluate if a sequence is significantly related to the query, MISS methods tend to suffer from convergence problems. Rather than converge after finding a set of related sequences, MISS tends to at some stage pick up one or more false positives and then begin to diverge. This has made multiple intermediate sequence search methods next to impossible to automate.

Because the first stage of ISS (and MISS) is a simple pairwise similarity search, these methods shed no light on sequences where close homologs (possibly of unknown function) are not already known.

2.4.2 Sequence profile methods

A better known class of methods that make use of multiple sequences are the sequence profile methods. These rely on multiple alignment algorithms to align a set of related protein sequences to obtain a sequence profile, which describes the amino acid propensities of each individual position in the sequence.

Such a profile provides a better description of a protein than the single sequence itself. Not only do profiles reveal which amino acid is preferred at each position (the consensus sequence), they also contain information of the degree of conservation and preferred mutations along the amino acid chain. This additional information allows for much improved scoring schemes of similarity, where matches to highly conserved positions are emphasized over matches to highly variable ones.

Similar to the multiple intermediate sequence search method, many sequence profile methods iteratively pick up more and more sequences. As new sequences

are picked up, they are included in the multiple alignment and thus help improve the quality of the profile. Like the intermediary sequence search methods and so many other iterative approaches, iterative profile methods suffer from convergence problems.

Two such methods deserve to be mentioned, the first is SAM-T99 which represents the current state of the art of profile methods. It makes use of HMMs to model the sequence profiles, thereby having not only position specific amino acid scores but also position specific gap penalties. Its predecessor, SAM-T98, is sufficiently good at finding remote homologs that it was entered in the CASP3 assessment in the fold prediction class where it performed comparable to more advanced fold recognition methods (Sternberg et al., 1999). It was in other words capable of identifying remote homologs with known structure for some of the targets for which the CASP organizers thought that none existed. The main problem with this method is that it is computationally very intensive.

The other method which should be mentioned is PSI-BLAST, which is by far the fastest and most used of the profile methods. Like the BLAST programme, PSI-BLAST is a heuristic method which cuts corners to gain speed. In contrast to SAM-T99, PSI-BLAST uses a position specific scoring matrix rather than an HMM and therefore has fixed gap penalties. Furthermore, PSI-BLAST uses pairwise alignments rather than a multiple alignment to estimate the position specific amino acid distributions. It is for these reasons less sensitive than the SAM T99 method, although much more sensitive than simple pairwise alignment methods. However, this small loss in sensitivity is well compensated for by the much higher speed of PSI-BLAST compared to SAM-T99.

2.4.3 Manual intervention in iterative approaches

It has already been mentioned that the iterative methods for identifying remote homologs all suffer from the problem that they are hard to automate. In general to obtain the best possible performance, an expert should judge if the sequence found should be included in the sequence profiles (or used as intermediate sequence in the case of ISS). This is both a difficult and a very time consuming job and would thus be desirable to eliminate. A somewhat successful approach for doing this with PSI-BLAST is the SAWTED method (Liu and Rost, 2000). By comparing the textual similarity of matches found in SWISS-PROT, it is possible to imitate some of what an expert would do.

2.5 Databases of known protein families

Although much more powerful than simple pairwise alignment, the iterative schemes for using multiple sequences have two things in common: they are difficult to automate and they do not work for orphan sequences. Both of these problems can be addressed by making databases of multiple alignments rather than using multiple alignments to search sequence databases.

2.5.1 Databases of protein families

One of the first such databases is Pfam, which is simply an abbreviation for Protein Families. Pfam is split in two parts: a curated database of known protein domains (Pfam-A) and an automatically generated set of non-characterized protein families (Pfam-B). For all Pfam-A families, a curated multiple alignment of a representative set of sequences has been constructed, and from this alignment an HMM has been built. It is thus possible to search a query sequence against all known protein families to see if there are any matches.

Functional annotation describing the functions is also present for domains in Pfam-A. This is a very valuable source as it gives a good idea of the variation of functions within the family, and hence a good idea of how much functional annotation can safely be transferred.

Other databases like PROSITE (Falquet et al., 2002) and SMART (Schultz et al., 2000) also contain this type of information. Both of them—together with Pfam—have been integrated in the InterPro database (Apweiler et al., 2000), which also comes with a search tool called InterPro-Scan. Furthermore all InterPro families have been annotated according to the Gene Ontology classification scheme, which makes it an even more powerful resource for function annotation.

2.5.2 SUPERFAMILY

Unlike the other domain databases mentioned so far, SUPERFAMILY is a database of protein superfamilies rather than protein families (Gough and Chothia, 2002). A protein superfamily contains several families, which have very similar structure and are believed to be evolutionarily related. However, the sequence similarity is often so weak, that it is difficult to detect the relationship at the sequence level.

In SUPERFAMILY, HMMs have been constructed for each superfamily by aligning the sequence profiles of the individual protein families using structural information. Results show that using these HMMs, it is possible to assign some sequences to a superfamily even though they do not belong to any of the known families making up the superfamily. This allows assignment of structure to these sequences and can also give an idea of the function.

2.6 Alignment based methods for obtaining functional links

In many cases it is not possible to infer a protein's function based on sequence similarity to a sequence or sequence family of known function. Sometimes this is due to the protein being a so-called "orphan" protein, i.e. no homolog protein sequences can be detected at all. In such cases it is obviously impossible to get any further by using alignment methods.

Often homologous proteins are detected by similarity based methods, although the function of these related proteins is unknown. In recent years a number of methods have been developed to allow alignment to be used to assign a putative

function to such proteins. What all these methods have in common is that they try to establish functional links between proteins. These links will then hopefully link proteins of unknown function to proteins of known function.

2.6.1 The Rosetta stone method

The Rosetta stone method—named after the famous Rosetta stone—is also commonly referred to as the "protein fusion", a name which is somewhat more descriptive of how it works. The basic idea is that if two proteins A and B in one organism are found as one large fused protein sequence (C) in one or more other organisms, then the two proteins are likely to be somehow functionally related. It is important to note that although the proteins A and B are both similar to protein C, the are not themselves similar. It is thus possible to transfer function from B to A due to the existence of C (Enright et al., 1999; Marcotte et al., 1999a).

One can however argue that this is not that big a step forward. Since protein C was similar to B, it would be tempting to infer a putative function of C based on this similarity. When A is then subsequently similar to C, this function could again be transferred. However, this would normally be considered a dangerous thing to do. The protein fusion method refines the approach, but more importantly it gives a biologically meaningful reason for why we should believe such predictions.

Although the method is as such biologically sound, it should be approached with care. The two proteins linked together by a protein fusion will almost always correspond to different domains in the fused protein. There are however certain domain families that are found in combination with a wide variety of other domains. These are called "promiscuous domains" and cause huge problems for the protein fusion method since they will often link together functionally unrelated proteins. This problem can be dealt with by either masking all occurrences of promiscuous domains or by filtering the search results. When used correctly, the protein fusion method can be a powerful tool: it provides many putative links for orphan proteins—although these might not be of the highest quality.

2.6.2 Phylogenetic profiles

By aligning a protein sequence against the proteins from all completely sequenced genomes, it is possible find out which organisms have homologs of the protein and which do not. By constructing a binary vector from this information one gets an evolutionary fingerprint of the protein, known as a *phylogenetic profile* (Pellegrini et al., 1999; Marcotte et al., 2000). If proteins have identical (or at least very similar) phylogenetic profiles, it is an indication that they are involved in the same function. The reason is that if a set of proteins form a functional unit, an organism would often have no advantage of having only part of the proteins—it should either have all of them or none of them.

The quality of the functional links obtained by comparing phylogenetic profiles appears to be much higher than that of protein fusion links (Marcotte et al., 1999b), and the quality can be expected to improve as more genomes are se-

quenced. The main weakness of the method is that some types of phylogenetic profiles are non-informative: knowing that a protein exist in all sequenced organisms only indicates that it is essential—not why it is essential, that is its function. Similarly, phylogenetic profiles will not tell much about proteins present in all eukaryotes but not in prokaryotes, nor will they tell anything about proteins not seen in other genomes. This problem is inherent to the method and will not be solved by sequencing more genomes.

2.6.3 Genome proximity

Several genes involved in the same biochemical pathways are often found within a small region of the genome. This is in particular true in bacteria where such genes are often encoded by an operon (Jacob and Monod, 1961). The simple fact that two proteins are located close to each other within a genome obviously does not provide much evidence for the two proteins having related functions. But if two proteins appear close to each other in several different genomes, the evidence becomes stronger. This method for providing functional links between proteins is known as "conserved gene order" and "genome proximity"(Dandekar et al., 1998; Galperin and Koonin, 2000; Yanai et al., 2001).

The probability that two genes are functionally related increases with the number of genomes in which they are in chromosomal proximity (Yanai et al., 2001). Accuracy better than 90% can be obtained by using a stringent threshold. However, quite few links are observed for such stringent thresholds (Yanai et al., 2001). Links with an estimated accuracy of 80% (on KEGG pathways) were reported to the Predictome database, averaging only approximately 400 links per genome for medium sized prokaryotic genomes (Yanai et al., 2001).

This method for linking together proteins of common function is somewhat related to the *phylogenetic profile* method described above, as genes obviously have to be conserved in the same genomes in order to have conserved relative location . It still remains to be seen how well this method works for eukaryotes—although a poorer performance should be expected as eukaryotes lack operons (Marcotte, 2000).

2.6.4 Co-evolution of proteins

If two proteins perform a function together—often by forming a complex—it is not only their phylogenetic profiles that should be expected to be similar, but also the phylogenetic trees. This is because changes to one component of a complex will often require some adaptation of the other components as well. This is known as co-evolution.

Similar phylogenetic trees of proteins thus indicate that the proteins may interact and that they are likely to have the same function (Pazos and Valencia, 2001). The main strength of phylogenetic profiles is, that this method can be used for proteins where the phylogenetic profiles are very uninformative, e.g. proteins that are present in all eukaryotes but not in prokaryotes or even essential proteins which are present in all organisms.

There are two obstacles standing in the way of using this method on complete proteomes. One is that reconstruction of phylogenetic trees is computationally intensive. The other is that the comparison of phylogenetic trees is difficult to automate, as no good scoring function for the similarity of two phylogenetic trees exists (to this authors knowledge). Both of these obstacles can be circumvented by inferring co-evolution of proteins by comparing distance matrices directly, without actually having to reconstruct the phylogenetic trees (Pazos and Valencia, 2001).

2.6.5 Combining evidence

If used individually, the types of functional links described in this section only give hints that proteins have similar function. More confident predictions can be made by combining the methods to look for functional links supported by more than one method (Marcotte et al., 1999b; Eisenberg et al., 2000; Pellegrini, 2001).

Still, two points should be noted. First of all, the high confidence links obtained by requiring linkage by either phylogenetic profiles or two other methods only provide a coverage of around 15% of the proteome (Galperin and Koonin, 2000). Secondly, providing a functional link between two proteins does not clearly state how the two proteins are related in terms of function (Galperin and Koonin, 2000). Predictions made from functional links are thus interpretations that are open for discussion.

2.7 Prediction of protein function via structure

When function cannot be inferred based on sequences similarity, one must rely on true *ab initio* prediction methods. It is a generally accepted paradigm that the function of a protein is determined by its three-dimensional structure, and that the structure is determined by the sequence of the protein (Anfinsen, 1973).

Given this paradigm, it would be logical to think that *ab initio* function prediction could be done by first predicting the structure of the protein, and subsequently predict the function from the structure. However, both steps in this approach are likely to be very difficult to solve.

2.7.1 From sequence to structure

The challenge of predicting, *ab initio*, the structure of a protein from its sequence is known as the protein folding problem. This is a problem which has been worked on intensively for many years. There is still belief that structure prediction is feasible (Pillardy et al., 2000), *ab initio* methods have outperformed dedicated fold recognition methods for a few targets in fold recognition part of CASP (Sternberg et al., 1999). However, judged from the biannual CASP experiments *ab initio* prediction of protein structure is currently not a viable option (Lesk et al., 2001).

It therefore seems that the only realistic approach to obtaining a good model is homology modeling. While a model created this way can be very interesting for understanding the function of a protein, it is not very useful from a function

prediction point of view: as very few structures exist for proteins of unknown function, one might as well infer the function directly based on sequence similarity rather than go via structure. However, it is possible that this will change with the many new structures being determined by structural genomics efforts.

2.7.2 Predicting protein function from structure

Knowing the structure of a protein does not mean that we can necessarily figure out what the protein does, even though it is of course a big help (Norin and Sundström, 2002). This is in part because the function of a protein depends on its cellular context. Also, post-translational modifications can profoundly alter the function (and structure) of a protein. Predicting the function of a protein from its structure may therefore very well turn out to be as difficult as the protein folding problem. No methods known to the author have been developed which attempt to predict protein function from structure alone.

A somewhat less ambitious approach, is to simply infer the function of a protein from other proteins with similar structure. Conceptually this is very similar to inferring function from proteins with similar sequence, but could work for more remote homologs as structural similarities are detectable over longer evolutionary distances than sequence.

However, when stating that similar structure implies similar function, people often forget to define what they mean by "similar structure" and "similar function". Without these terms defined, the statement is empty. A reasonable definition of "similar structure" would be that the proteins belong to the same superfamily, as it is clear the proteins with the same fold do not necessarily have the same function (Todd et al., 2001). Defining "similar structure" to mean the same family also would not be meaningful, as the sequence would then always be similar as well.

It is more difficult to say how "similar function" is best defined in this context, although some people argue that the method is more likely to work well for predicting molecular function than cellular role (Weir et al., 2001). Still, in about 10% of the cases where structure is known for two remotely homologous enzymes, the substrate binding site has changed indicating a complete change of molecular function. It will therefore be difficult to predict the enzyme activity of an uncharacterized protein from knowing its superfamily (Todd et al., 2001).

2.7.3 Structure is important for understanding function

From the discussion above, one might think that the author disagrees with the paradigm that structure determines function. This is however not the case: the structure of a protein is responsible for its interactions with the environment and hence its function. However, it should be noted that the structure mentioned includes post-translational modifications like glycosylations, which are very important for protein function but are typically not included when protein structures are determined. Although studying structure is very important for understanding protein function, this author will argue that it is not very useful from a function prediction point of view.

Part II

Orphan protein function prediction

Chapter 3

To be or not to be—that's the *first* question

When it comes to analysis of whole genomes nothing is perfect. The sad fact is that many things are far from perfect. This includes gene finding: even the best of gene finders make many false positive predictions. That is in particular true for eukaryotes where gene finding is complicated by the presence of introns.

Even for prokaryotes gene finding turns out to be a nontrivial problem though. This is mainly due to the occurrence of random open reading frames (ORFs) that lead to false positive predictions. The expected length distribution of random ORFs is an exponential distribution, i.e. most of the ORFs will be very short while only a very small fraction of the ORFs will be long. However, the number of random ORFs as long or longer than many real protein coding genes is still large enough to pose a significant problem (Johnson, 2000). Also, a number of very short ORFs turn out to in fact be protein coding genes presenting a serious problem for gene finders (Basrai et al., 1997; Johnson, 2000). A largely arbitrary cutoff of 100 aa is often used when annotating ORFs as being coding genes. While it may give a reasonable compromise between false positives and false negatives, it unfortunately gives plenty of both.

Perhaps because it is much easier than gene finding in eukaryotes, gene finding in prokaryotes is by many considered to be a "solved" problem (Lewis et al., 2000). A possible explanation for this optimism is the often good agreement between genome annotations and predictions made by a gene finder. One must however not forget that most of the genes annotated in complete genomes have not been experimentally verified, but are themselves predictions made by a gene finder. Because the many different gene finders suffer under the same inherent difficulties described above, the errors they make can be expected to be heavily correlated. Evaluating gene finders on current complete genome annotations thus paints too pretty a picture of the current state of prokaryotic gene finding.

3.1 False positive predictions look like orphans

One property that the false positive predictions share is that the corresponding protein sequences are not similar to any other proteins in the databases (Fischer

and Eisenberg, 1999). Thus they all appear as orphan proteins, making it very difficult to assess the number of genes as well as the number of novel genes.

This problem only becomes more pronounced when studying eukaryotes where gene finding is further hampered by the presence of intervening sequences called introns. One illustration of these difficulties is that 30% of all introns predictions in *S. cerevisiae* have been shown to be wrong (Davis et al., 2000)—although yeast have unusually short introns which should cause relatively few problems. A perhaps even better illustration of the current state of Eukaryotic gene finding is the four fold spread of the guesses on the number of genes in the human genome. The guesses published in high profile journals range from 30,000 human genes (Claverie, 2001) to 120,000 human genes (Liang et al., 2000), where the lower figure appears to be closest to the current consensus.[1] However, the dust has not quite settled. There are still convincing arguments that the estimates of 30,000 to 40,000 genes may be quite a bit too low (Hogenesch et al., 2001).

The difficulties of making reliable gene finding in eukaryotes do not only result in prediction of spurious genes and real genes being missed. It also causes errors in genes that are otherwise correctly predicted: wrong splice sites may be predicted, exons can be missed, non-existing exons may be included, two genes may be predicted as one transcript or one gene may be split into two. All in all, these errors cause approximately half of the "correctly" predicted genes in the human genome to contain serious errors (Dunham et al., 1999; Hattori et al., 2000). These errors are of course reflected in the predicted protein sequence, which is likely to cause trouble for functional annotation efforts.

[1] Partially as a joke, a sweepstake on the number of genes in the human genome, Genesweep, was organized by the Cold Spring Harbor Laboratory. Here the guesses ranged from 25,000 to 312,000 (Attwood and Miller, 2001). Predicting more genes is not necessarily better.

Paper I

3.2 On the Total Number of Genes and Their Length Distribution in Complete Microbial Genomes

Marie Skovgaard, Lars Juhl Jensen, Søren Brunak, David W. Ussery, and Anders Krogh*

Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby, Denmark

* To whom correspondence should be addressed (email: krogh@cbs.dtu.dk)

In sequenced microbial genomes some of the annotated genes, usually marked in the public databases as hypothetical, are actually not protein coding genes, but rather open reading frames (ORFs) that occur by chance. Therefore, the number of annotated genes is higher than the actual number of genes for most of these microbes. Comparison of the length distribution of the annotated genes with the length distribution of those matching a known protein reveals that too many short genes are annotated in many genomes. Here we estimate the true number of protein coding genes for sequenced genomes by two different methods. The first estimate is based on the assumption that the fraction of long genes matching the SWISS-PROT database equals the fraction of all genes that matches SWISS-PROT. The second estimate, which is included as a database-independent check of the first, is the number of non-overlapping open reading frames longer than 100 triplets reduced by the number expected by chance. While it is often claimed that *E. coli* has about 4300 genes (Blattner et al., 1997), we show that it is likely to have only around 3800 genes, and that a similar discrepancy exists for almost all published genomes (see Table 3.1). In one extreme case we estimate that half of the annotated genes are wrong.

The most reliable method for identifying genes is by similarity to a protein in another organism. Genes with no match to known proteins can be predicted using statistical measures. The most important measure is the codon usage: the frequency of codons in true genes is different from what would be expected at random from a given base composition. However, the discriminatory power of a codon usage measure becomes less reliable for shorter ORFs, and this is true also for other measures of coding potential, such as dicodon or hexamer statistics. Together with the large number of short random ORFs, this tends to give an over-prediction of short genes. Since stop triplets (TAA, TGA, TAG) are A-T rich, their frequency is generally higher in A-T rich organisms than in G-C rich organisms, so the likelihood of long ORFs occurring by change is higher the higher the G-C content of the organism. Therefore the problem of discriminating between short proteins and random ORFs is generally less in A-T rich organisms than in G-C rich organisms, as shown in Figure 3.1.

The growing use of sequence databases in molecular biology makes it important to consider the correctness of the information stored in them. Careful annotators have clearly marked non-confirmed genes as hypothetical. However, many users of the databases assume that all annotated genes indeed correspond to true genes, and this can easily lead to wrong conclusions. An example is a recent study of protein length distributions for the three kingdoms of life (Zhang,

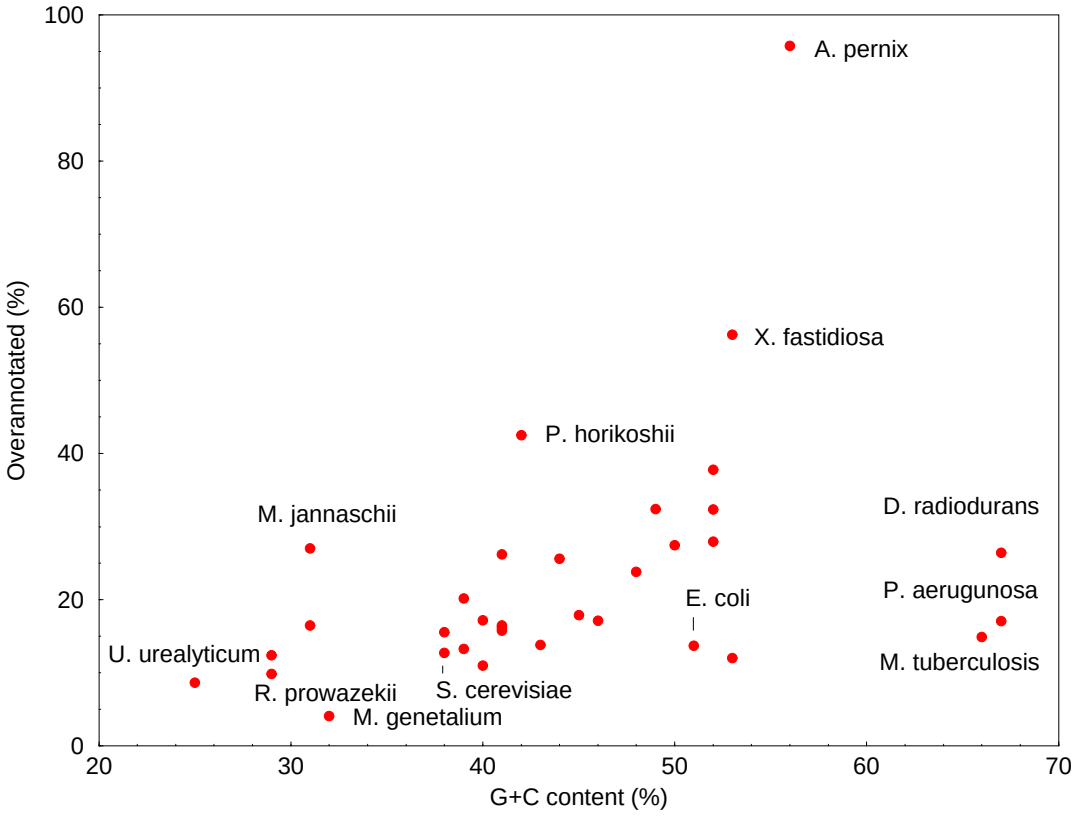


Figure 3.1: **Estimated over-annotation of genes in sequenced genomes.** For each organism the SWISS-PROT based estimate is calculated and the difference to the number of annotated genes shown in percent of the estimated number of genes.

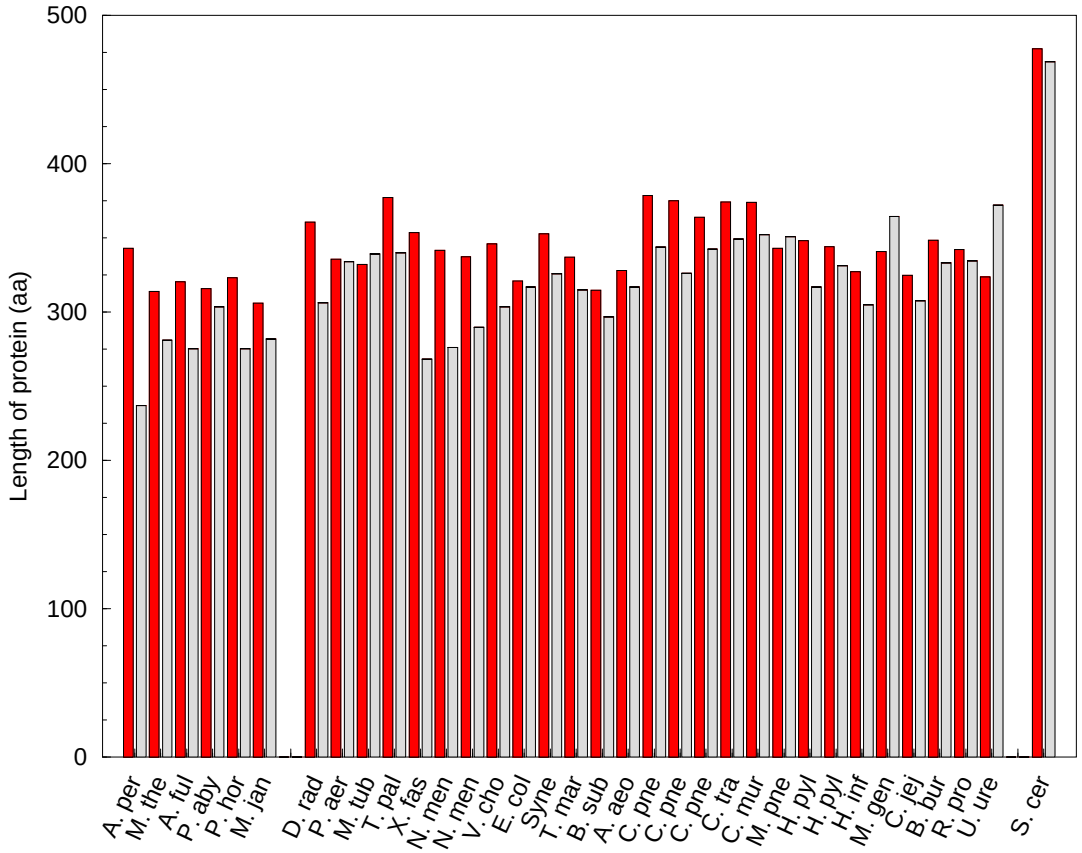


Figure 3.2: **Average length of annotated and confirmed proteins.** Red bars show average length of all proteins having matches to non-hypothetical proteins in SWISS-PROT (see Methods) and gray bars show the average length of all annotated proteins. The ordering of organisms is the same as in Table 3.1.

2000). Based on the annotation, it was concluded (among other things), that the average or median length of proteins is smaller in Archaea than Bacteria. This is due to the fact that a very large number of short ORFs are annotated as genes in some of the Archaeal organisms. When including only proteins confirmed by a match to a known protein, there seems to be no significant difference in the average (or median) lengths (Figure 3.2).

Length distributions

Shortly after the publication of the complete *S. cerevisiae* sequence it was shown that there was a systematic error in the CDS assignments. More than 400 sequences with lengths between 100 and 110 amino acids had no matches to previously assigned proteins (Das et al., 1997). This group stood out as a peak in the length distribution and seemed to be an artifact.

Similarly, we have plotted the distribution of lengths for each organism found in GenBank (rel. 119), see `http://www.cbs.dtu.dk/krogh/genomes/`. Figures 3.3 and 3.4 are examples showing the length distribution of the unique data set confirmed by a match to a non-hypothetical protein in SWISS-PROT, and the length distribution of those that are not. A very large protein family would result

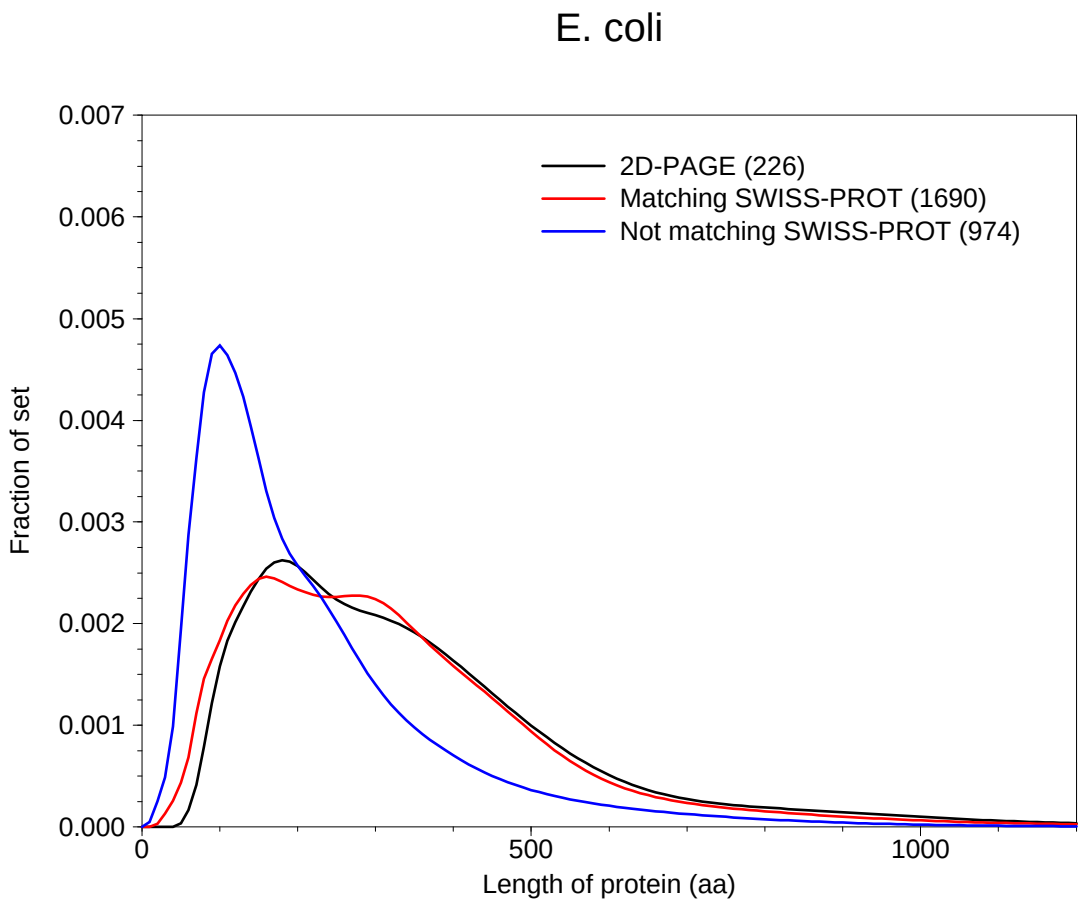


Figure 3.3: **Protein length distributions for *E. coli*.** The red and blue lines show the length distributions of unique sequences matching and not matching entries from SWISS-PROT, respectively. The black line shows the distribution of unique proteins from SWISS-2D-PAGE. All length distributions are normalized and have been smoothened by Gaussian kernel density estimation.

in a peak in the length distribution, which is avoided when using the unique data set, in which sequences with similarity to others are taken out (see Methods). At the same time, strong similarity reduction is expected to make the random sequences more dominant in the unique set, because it is not very likely that two random ORFs are similar.

E. coli is one of the most studied microbial organisms and the plot in Figure 3.3 reveals a significant difference between the length distributions of unique sequences matching and not matching SWISS-PROT. The sequences not matching SWISS-PROT are generally shorter than the ones matching. Actually, 81% of the 974 proteins that were excluded as 'not matching SWISS-PROT' did have matches to *hypothetical* proteins in SWISS-PROT, which were not counted. These were mostly from *E. coli* or closely related organisms. The definition for the keyword hypothetical in a SWISS-PROT entry is "predicted proteins for which there is no experimental evidence that they are expressed in vivo."

The length distribution of the annotated coding sequences in the Archaea *A. pernix* is shown in Figure 3.4. This is quite extreme, because rather than performing actual gene finding, all ORFs with a length of at least 100 triplets

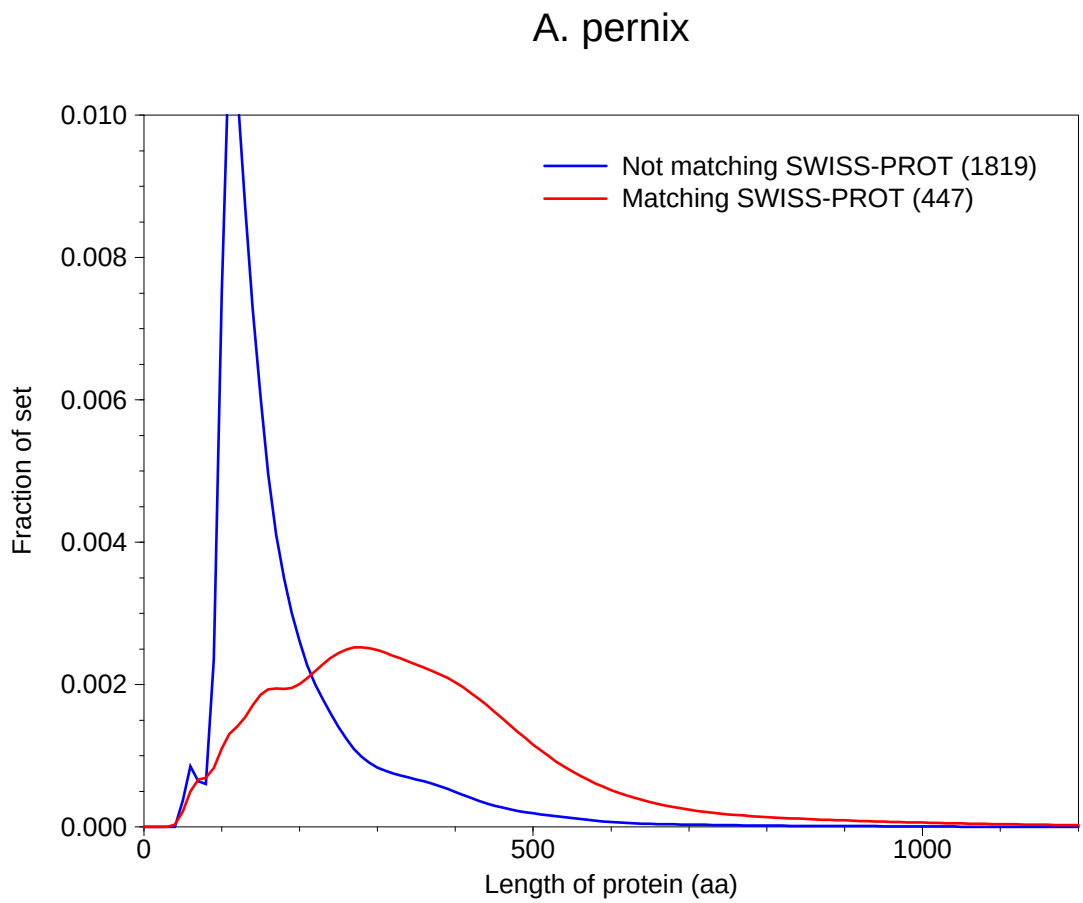


Figure 3.4: **Protein length distributions for *A. pernix*.** Similar curves as in Figure 3.3 except that there is no 2D-gel data.

were annotated as coding regions in the GenBank entry (Kawarabayasi et al., 1999). The subset of the unique data set with matches in SWISS-PROT has a distribution comparable to the distribution seen in other prokaryotic organisms, whereas the length distribution of non-matching annotated genes has a peak which is not too different from the geometric distribution expected in a random sequence of DNA.

Estimating the true number of protein coding genes

The length distributions indicate that too many protein coding genes are annotated. To obtain an estimate of the true number of proteins in each organism we have used the proteins in the SWISS-PROT database (Bairoch and Apweiler, 2000), which are not labeled hypothetical, as a reference. The estimate is based on the assumption that the fraction of proteins with a match in SWISS-PROT is independent of the length of the proteins. Since ORFs longer than 200 amino acids are unlikely to occur by chance in most organisms (apart from long repetitive sequences), the fraction of those matching SWISS-PROT was used as an estimate of the fraction of the total number of true proteins that match SWISS-PROT. Then the estimated number of genes is easily obtained by dividing the total number of matching proteins with this fraction. For instance, assume that 1400 of

Table 3.1: **Complete microbial genomes from GenBank release 119** The table contains the number of annotated proteins, the A+T content, the number of proteins estimated from matches to SWISS-PROT and from stop triplet frequency. The last column shows the number of clusters of orthologous genes (COGs) for the organism. The list is ordered by kingdoms (Archaea, Bacteria, Eukaryota) and A+T content.

	$A+T$ (%)	No. annotated genes	No. genes estimated from SWISS-PROT	No. genes estimated from stop triplet	Number of COGs
A. pernix	44	2694	1376	1423	1169
M. thermoautotro.	50	1869	1466	1535	1375
A. fulgidus	51	2407	1818	1927	1849
P. abyssi	55	1765	1497	1635	1443
P. horikoshii	58	2064	1448	1616	1365
M. jannaschii	69	1715	1350	1573	1320
D. radiodurans	33	2937	2323	1904	2176
P. aeruginosa	33	5565	4753	3508	4191
M. tuberculosis	34	3918	3410	2537	2668
T. pallidum	47	1031	920	820	707
X. fastidiosa	47	2766	1770	1792	1491
N. meningitidis A	48	2121	1539	1447	1455
N. meningitidis B	48	2025	1530	1500	-
V. cholerae	48	3828	2991	2931	2745
E. coli	49	4289	3771	3463	3327
Synechocystis	52	3169	2559	2550	2113
T. maritima	54	1846	1576	1564	1507
B. subtilis	56	4100	3263	3330	2803
A. aeolicus	57	1522	1337	1412	1317
C. pneumoniae	59	1052	903	909	647
C. pneumoniae AR39	59	997	790	913	-
C. pneumoniae J138	59	1070	921	910	-
C. trachomatis	59	894	772	754	631
C. muridarum	60	818	698	763	-
M. pneumoniae	60	677	610	617	423
H. pylori	61	1566	1303	1384	1081
H. pylori J99	61	1491	1316	1351	1062
H. influenzae Rd	62	1709	1479	1526	1504
M. genitalium	68	480	461	474	376
C. jejuni	69	1654	1420	1494	1289
B. burgdorferi	71	850	756	772	694
R. prowazekii	71	834	759	795	674
U. urealyticum	75	613	564	556	401
S. cerevisiae	62	6269	5560	5728	2175

2000 annotated genes longer than 200 amino acids have a match in SWISS-PROT (70%). If there is a total of 2100 annotated genes with a match in SWISS-PROT, we estimate that the total number of genes is $2100/0.7 = 3000$. These estimates are shown in Table 3.1 and the percent over-annotation according to the estimate is shown in Figure 3.1.

We have argued that some of the short annotated genes are not real. Due to the difficulty in discriminating a short ORF from a truly expressed gene, it is also quite likely that there are short genes which has not been annotated. The alignment based estimate implicitly assumes that there are quite few of these unannotated genes. This might not hold for all organisms, which would also mean that the estimate is too low. Thus, the alignment based estimate is really an estimate of how many of the annotated genes are correct.

It is possible that proteins matching SWISS-PROT are biased in length, because of the local search method (e.g. due to domain structure) and possibly an inherent length bias in SWISS-PROT, which would both invalidate our as-

sumption that the fraction of matching genes is approximately the same for long and short genes. If matching proteins are biased towards long proteins (which is perhaps most likely) our estimate is too low. For the genomes we have studied, approximately 15–20% of the proteins matching SWISS-PROT have lengths below 200 amino acids. To get an idea of the worst-case error, let us assume that proteins longer than 200 amino acids are twice as likely to match SWISS-PROT as compared to those shorter than 200. Then the true number of genes would be 15–20% larger than our estimate (equal to the fraction of matching proteins shorter than 200).

As a control of the above alignment estimate, we have calculated another estimate, which is completely independent of database matches. The maximal number of non-overlapping open reading frames longer than 100 triplets was found and the estimate of the true number of genes was obtained by reducing this number by those expected to occur at random. The reduction was calculated approximatively from the stop triplet frequency. These estimates are also shown in Table 3.1. ORFs shorter than 100 triplets were excluded since relatively few genes are expected, and the estimate becomes ill-behaved because of the huge number of short ORFs. For high A-T content the number of non-overlapping ORFs is a reasonably good estimate in itself, whereas the corrections become more and more important the lower the A-T content. The approximation of the corrections is quite crude, and overall the estimate is not very good for low A-T content. Indeed, the largest discrepancies between the two estimates is seen for the organisms with low A-T, whereas they are quite small for intermediate to high A-T content. The stop triplet based estimate is usually higher than the alignment estimate.

The number of genes that are members of clusters of orthologous genes (Tatusov et al., 2001) (COGs) are also shown in Table 3.1. A COG is defined if a gene is found in at least three lineages, so this should be an approximate lower bound on the number of protein coding genes in an organism. These numbers are lower than the alignment based estimates except for *A. fulgidus* and *H. influenzae* where they are very close.

It is notable that such a large fraction of the hypothetical *E. coli* ORFs had no significant matches to verified SWISS-PROT entries. The two-dimensional polyacrylamide gel electrophoresis data from SWISS-2DPAGE (Hoogland et al., 2000) was used as an independent control of the length distribution. 271 *E. coli* sequences were retrieved from the database, and the length distribution of the similarity reduced set of 226 is shown in Figure 3.3. Although it is a fairly small set of genes, it is experimentally confirmed and independent of biases in alignment. These data support our assumption that the length distribution of SWISS-PROT matches is reasonably unbiased. We investigated the possibility of using DNA expression array experiments, but they turned out to be unreliable as a control, because a probe for a wrongly annotated gene can be located in an untranslated region of an mRNA from an expressed gene, and will therefore appear to be expressed. Secondly, probes are often made for the annotated genes, making the results dependent on the annotation.

Conclusion

Our estimates of the number of real protein coding genes reduce the number of true proteins by 10–30% for the majority of microbial organisms. The two extremes are represented by *M. genetalium* where the estimates are 1–5% lower and *A. pernix* where they are close to 50% lower. The large over-annotation of *A. pernix* has previously been noted (Natale et al., 2000; Cambillau and Claverie, 2000). Natale et al. (2000) estimate the correct number of protein coding genes to be between 1,550 and 1,700 based on the assumption that the total fraction of confirmed genes should be about the same as for other organisms. However, since the other organisms are also over-annotated, this estimate is perhaps still too high, which explains our slightly lower estimate of about 1,400 genes in *A. pernix*. It is also possible that our estimates are a bit too low, as discussed above.

The problem with wrongly annotated protein coding genes is almost entirely due to the difficulty in distinguishing short non-coding ORFs from real genes. The problem cannot be solved at present, but there are several ways in which the situation can be improved until better gene identification methods are developed. Firstly, a measure of statistical significance for gene prediction is needed. Secondly, an ORF should only be annotated as a coding sequence (CDS) if it has either a trustworthy protein match or if it has very high significance. Thirdly, other possible genes should be annotated as ORFs, clearly showing that they are hypothetical.

Methods

We have analyzed 34 fully sequenced microbial genomes as found in GenBank release 119. For each organism all sequences annotated as 'CDS' in the feature table were extracted and translated to proteins. To generate the unique set, these sequences were aligned against themselves using gapped BLASTP (Altschul et al., 1997). With a threshold of 10^{-3} on the expectation scores, we subsequently generated maximal similarity reduced versions of the data sets using the algorithms by Hobohm et al. (1992). This procedure reduced the sets by 13–36%.

The full sets were searched against the SWISS-PROT database (Bairoch and Apweiler, 2000) (releases 38 to 39.7) using BLASTP. Matches to sequences with the keyword 'hypothetical' were disregarded. Sequences giving no hits in SWISS-PROT with an expectation score better than 10^{-3} were categorized as not matching SWISS-PROT, while sequences were considered to match SWISS-PROT only if at least one match with a score better than 10^{-6} was obtained. Sequences for which the best match had an expectation score between 10^{-3} and 10^{-6} were considered in the 'gray zone' and were not included in any of the categories (typically 3–5%).

Average lengths of all annotated genes and for all matching SWISS-PROT were calculated and used for the histogram in Figure 3.2.

Length distributions were calculated for all annotated CDSs, the unique set of annotated CDSs, the unique set having matches to SWISS-PROT, and the ones not matching SWISS-PROT. Rather than plotting raw histograms we made a Gaussian kernel density estimation of the logarithmic length distribution and

log-transformed the distribution back to an ordinary length distribution. The width of the kernel was estimated from the data (Silverman, 1986).

An estimate of the true number of genes was calculated by extrapolating from the proportion of annotated genes of length greater than 200 amino acids (ORF_{200}) that matches SWISS-PROT entries (SP_{200}). The total number of annotated genes matching SWISS-PROT (SP_{all}) was then divided by this ratio to get an estimate of the total number of genes ($SP_{all}\,ORF_{200}/SP_{200}$).

Another estimate of the total number of genes, independent of database matches, can be obtained from the ORF lengths and stop triplet frequencies:

$$G = \sum_{i=100}^{L} N_{max}(i) \frac{N_{orf}(i) - A p_{ran}(i)}{N_{orf}(i)},$$

where L is the length of the genome divided by 3, $N_{orf}(i)$ is the observed number of ORFs of triplet-length i, and $N_{max}(i)$ is the number of these in a set of *non-overlapping* ORFs longer than 100 triplets constructed by excluding the shortest ORFs first. $p_{ran}(i)$ is the probability of finding an ORF of triplet length i at a specific position in the genome, which can be approximated by $p_{ran}(i) = 2p_{stop}^2(1-p_{stop})^i$, where p_{stop} is the stop triplet frequency. A is the number of triplets in the genome not occupied by true genes, which can be found by solving the self-consistency equation,

$$L - A = \sum_{i=100}^{L} i\; N_{max}(i) \frac{N_{orf}(i) - A p_{ran}(i)}{N_{orf}(i)}.$$

The number of open reading frames grows exponentially as the length i goes to zero. Therefore the difference $N_{orf}(i) - Ap_{ran}(i)$ between the two very large numbers in these formulas is not well determined for short ORFs. This is why we estimate G only from ORFs longer than 100 triplets.

Acknowledgment

This work was supported by a grant from the Danish National Research Foundation.

3.3 The consequences of poor annotation

At a first one might ask why it is so important that the annotated genes are correct. After all, random protein sequences should not give significant matches when using alignment methods. One should bear in mind that the significance of a given alignment depends upon the size of the database against which the search was performed. This means that a heavy pollution of the databases with non-proteins will reduce the significance of all hits made by any search method against this database.

While this is serious enough, there are worse problems. Many people who are not themselves working on doing whole genome annotation, are not aware of the poor quality of the data sets. They therefore tend to trust the annotation almost blindly, which can lead to wrong conclusions since conceptual translations of large amounts of non-coding DNA regions become included in their analysis.

3.3.1 The *Aeropyrum pernix* sequel

There are several examples of researchers, who have been fooled by the *A. pernix* genome annotation, in which all open reading frames over 300 bp (corresponding to proteins of 100 aa) have been annotated as coding regions. Mistakes made on this genome further tend to become generalized to Archaea due to the comparatively poor annotation of most Archaeal genomes.

One such mistake has already been mentioned in Paper I: *A. pernix* proteins in particular and archaeal proteins in general appear to be shorter on average than bacterial and eukaryotic proteins (Zhang, 2000). Sadly, this is not a lone example. Essentially all types of analysis of protein sequences from complete genomes are bound to give odd results if the annotation is trusted as is.

In a review of hyperthermophilic proteins, the amino acid composition of all proteins from a range of different organisms was compared with the purpose of revealing how hyperthermophile proteins achieve their stability (Vieille and Zeikus, 2001). Unfortunately conceptual translations of all annotated coding regions were once again used as the basis of the analysis. The results were much as should be anticipated: archaeal proteins were found to have different amino acid compositions that those of bacteria. Also *A. pernix* proteins were noted to have a very different amino acid composition—both when compared to proteins from bacterial and archaea. While it may very well be true that archaeal and bacterial proteins differ, such conclusions should not be made based on all annotated coding regions. Instead one ought to limit the analysis to only include proteins with clear sequence matches to known proteins (for instance SWISS-PROT).

The differences between *A. pernix* protein coding genes and other of other organisms seems to be something many people can agree on. When again taking the annotation for granted, *A. pernix* was further noted to contain an unusually low percentage of coiled coil proteins (Liu and Rost, 2001). Also only 20% of all predicted protein sequences from *A. pernix* can be assigned a fold—this should be compared to 30–40% for most other prokaryotes. The difference is not as drastic when viewed at the residue level where 23% of all protein residues annotated in the *A. pernix* genome can be assigned to a fold, which is only slightly lower than

for other prokaryotes (Liu and Rost, 2001). The obvious explanation is that short ORFs cannot be assigned to folds, most likely because they do not correspond to real proteins. Although the length distributions of proteins from all the analyzed genomes were actually plotted in the article, the authors did not comment on this possible source of biases.

3.3.2 Similar problems in other organisms

More recently, the *Xylella fastidiosa* genome sequence has been published (Simpson et al., 2000). The genome annotation contains 1,083 ORFs with no significant matches and has an average ORF length of 799 bp—about 20% shorter than most other prokaryotes. Because of these many supposedly orphan proteins, only 47% of CDSs annotated in the *X. fastidiosa* genome could be assigned a function, which is suggested to be due to no phytopathogenic bacteria having been sequenced before (Simpson et al., 2000).

3.4 Synonymous vs. non-synonymous substitutions

While our approach can estimate how many ORFs are incorrectly annotated as protein coding genes in a genome, we make no attempt at predicting which of the genes without matches to known proteins are true genes. Recently a technique for discriminating between short protein coding genes and random ORFs has been suggested (Ochman, 2002). The method relies on calculating $^{K_a}/_{K_s}$ ratios for short ORFs that are conserved in a pair of closely related organisms, e.g. *E. coli* and *Salmonella typhimurium*.

In $^{K_a}/_{K_s}$ ratios, K_a is rate of non-synonymous substitutions while K_S is the rate of synonymous substitutions, i.e. substitutions at the DNA level which do not affect the protein sequence. A ratio of 1 corresponds to neutral evolution, in the sense that there is neither selection for or against changing the protein sequence. For most protein coding regions K_a is significantly lower than K_s, indicating a selection against changing the protein sequence. In very special cases K_a can be significantly larger than K_s. This is known as positive selection—this indicates a strong selection for changing the protein sequence.

Because the method only works for pairs of closely related organisms, it currently cannot be applied to all sequenced genomes and the method is also not applicable to all short ORFs. But for the ORFs for which the method can be used, the results are very interesting: more than 90% of the ORFs with a $^{K_a}/_{K_s}$ ratio greater than 1 are shorter than 300 bp. In summary the majority of the genes that do not exhibit a preference for synonymous substitutions are short, have an unusual codon usage, and are of unknown function—all signs pointing in the direction that they are not protein coding regions (Ochman, 2002). These results are in good agreement with our own conclusions.

3.5 Experimental verifications

Since Paper I was submitted, one of our predictions have been experimentally verified by others. Based on both comparative genomes of four yeasts and whole genome expression analysis, Lander has concluded that approximately 500 of the 6,100 protein coding genes currently annotated in the Saccharomyces Genome Database (SGD) are in fact not genes[2]. This corresponds to an estimated 5,600 genes which is only 40 genes more than our SWISS-PROT estimate.

Although this is obviously only *one* example, it does add to the credibility of our estimates. Many people would tend to reject our estimates, which indicate that approximately 15% of all annotated protein coding genes in the most studied organisms are false positives. The validation of our prediction for *S. cerevisiae* makes it harder to simply reject the rest of our predictions.

[2]Announced in BioMedNet News (`http://news.bmn.com/news/`) by Eric S. Lander, May 30, 2002.

Chapter 4

Predicting protein function from sequence derived features

In Chapter 1 it was discussed that the function of a protein can be defined in terms of the proteins interactions with its environment. Since proteins performing similar functions would thus have to interact with the same environment, they should be expected to share certain properties. This should be the case also for proteins which are not evolutionarily related through a common ancestor.

The cell and its different subcellular localizations make up the environment in which the majority of proteins function. Proteins use the same cellular machine to achieve the same things—it is thus natural to expect that prediction of post-translational modifications can give hints to function.

The fundamental idea behind ProtFun is to integrate all protein features described above in order to predict protein function. However, most of the properties described above are unknown to us. Since all we have is the protein sequence, we have to rely on predicted (or in some cases calculated) features to form the input for the method (see Figure 4.1). Fortunately many methods for prediction of various protein properties had already been developed both at the Center for Biological Sequence Analysis and by other research groups.

Considering the starting point—nothing but a sequence with no known homologs—one should not expect the unreasonable. It is in my opinion unlikely that protein function in general can be predicted with the accuracy of similarity based methods. The predictions made by the method I will now describe are therefore much better suited for computational screening purposes than for predicting the function of individual sequences.

4.1 Similar approaches

The idea of using different types of sequence derived features to predict function has been used by others as well. A protein function prediction system for *E. coli*, which can predict protein function both in the presence and absence of similar sequences was developed around the same time as ProtFun (King et al., 2001). When not making use of matches to similar sequence of known function, this predictor has several similarities to the ProtFun method. The system makes

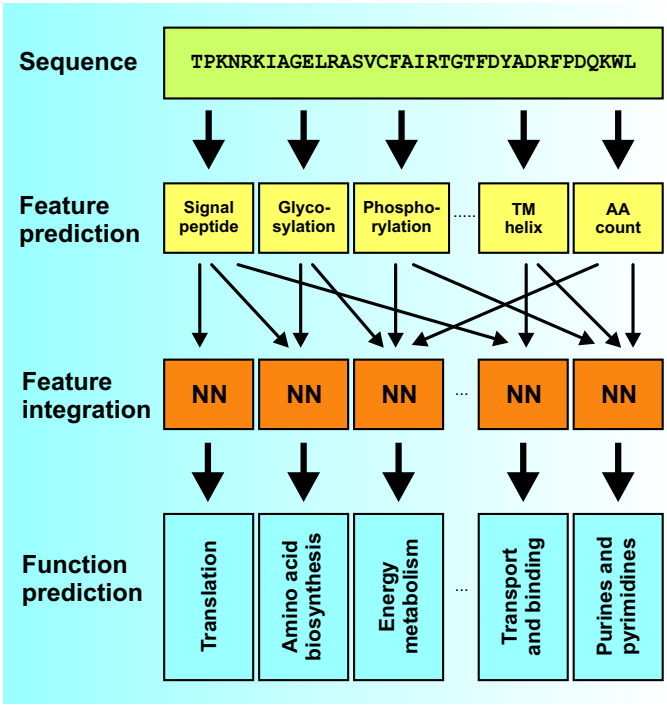


Figure 4.1: **The concept of ProtFun.** Based on sequence alone, a large number of protein properties are calculated or predicted. These features are integrated into functional category predictions by using neural networks.

use of sequence derived properties such as the amino acid composition, physical/chemical properties estimated from the composition, and secondary structure prediction. The crucial difference between their system and ours is that they do not make use of predicted subcellular localization signals nor co- and post-translational modifications. This is largely because they have focused on prokaryotic rather than eukaryotic proteins.

The company Virtual Genetics has developed a method for prediction of brain specific proteins which is very similar to ProtFun (Huss et al., 2001). The method makes use of many of the same prediction servers in ProtFun including the ones developed at CBS that are publicly available. The main difference is that Virtual Genetics use various types of decision trees (a different class of machine learning algorithms) instead of neural networks. This way they manage to predict 68% correct on a balanced data set of brain specific and non-brain specific proteins.

4.2 Generation of an "orphan" data set

In order to be able to train a method for predicting protein function of orphan proteins, proteins of known function are used. To emulate the "orphan situation" a test set must be constructed so that it shares no similarity with the training set. The most conventional way to achieve this would be to perform a homology reduction of the data and subsequently split the data set into a training and a test set. To ensure that no two sequences share significant similarity, a very strict threshold would have to be used, thereby reducing the data set enormously. To

avoid this undesirable reduction of the data set, a different scheme is used. The simple idea is to split the set of sequences into two sets of specified size, with the objective to minimize the total similarity between the sets. This is described in more detail in Paper II.

4.3 Function assignment of the sequences

4.3.1 Using the EUCLID dictionary to annotate cellular roles

A large data set of human protein sequences assigned to functional categories was unfortunately not available. For this reason the scoring system of the EUCLID method developed at Alfonso Valencias lab was used to assign human sequences from the SWISS-PROT database to cellular role categories (Tamames et al., 1998; Andrade et al., 1999a). These categories themselves (Andrade et al., 1999b) were derived from an earlier proposed scheme for *E. coli* (Riley, 1993) and comprise 13 functional classes which are subsets of three superclasses: Energy, Communication and Information. Proteins which cannot be assigned to one of the 13 categories are assigned to *other* or to *unknown*. There is a subtle difference between *other* and *unknown*: contrary to proteins assigned to the *unknown* category, the function of proteins assigned to *other* is known but it does not fit into any of the classes. Although neither of these two groups of proteins can be used for training a function prediction method, it would not be meaningful to attempt to use the predictor on the proteins assigned as *other*.

The EUCLID automatic assignment method was developed as part of the GeneQuiz annotation project, where it is used to assign proteins to cellular role categories based on BLAST matches. To assign a protein of unknown function, it is first compared to all protein sequences in the SWISS-PROT database, and a consensus set of keywords associated with the matches is extracted. These keywords are then used as input to an additive scoring system (Z-scores) in which each keyword has a score for each cellular role category. What is obtained is thus a Z-score for each category. Based on these scores, the protein is assigned to the category having the highest Z-score and a confidence class is assigned based on the difference between two highest Z-scores.

Since we use sequences from SWISS-PROT, they already have keywords. The BLAST step is thus not performed—instead the keywords of each entry simply enter the scoring system directly. Since a protein can belong to more than one functional class, we do not use the scores as done originally in EUCLID. Instead we studied the spread of the Z-scores and assigned a “high” and a “low” threshold of 3 and 0 respectively. We thus assign a protein as belonging to a class if the corresponding Z-score is above 3, while it is considered a negative example for a class if the Z-score is negative. Scores between the “high” and the “low” thresholds were considered unclear and were consequently not labeled. A protein can thus be assigned to multiple classes—or if no sufficiently high scores occur it may not be assigned to any of the classes.

4.3.2 Extracting enzyme classification from SWISS-PROT

The human SWISS-PROT sequences were also assigned to enzyme classes, both to enzyme vs. non-enzyme and to their major enzyme class, i.e. the first digit of the EC number (Enzyme Nomenclature, 1965, 1992). In the SWISS-PROT database most enzymes have the EC number annotated in the description line. These entries were assigned as enzymes and to the major enzyme class in question. Entries without an EC number but with a description word ending in "ase"' were manually assigned to the enzyme and non-enzyme categories or excluded from training if ambiguous. Sequences without EC numbers or the suffix "ase" were assumed to be non-enzymatic.

4.4 Correlations between different classification systems

While many different schemes for classifying protein function exist, these are correlated with one another. This can be exemplified by the two schemes described above, cellular role and enzyme classes (see Table 4.1).

As should be expected, the majority of proteins involved in various types metabolism are enzymatic whereas most proteins belonging to other functional categories, e.g. *transport and binding*, are non-enzymatic (see Table 4.1).

The enzyme classification does not only correlate with cellular roles at the enzyme/non-enzyme level. A large overrepresentation of oxidoreductases is observed for the *energy metabolism* category, which is consistent with the many oxidative steps taking place in aerobic respiration. Also, a large fraction of ligases turn out to be involved in translation. This too is consistent with the underlying biology.

A consequence of these correlations is that the performance of the predictors are bound to be correlated as well. If *energy metabolism* proteins can be predicted with high accuracy, it will be able to predict oxidoreductases with a certain

Table 4.1: **Correlations between cellular role classes and enzyme classes.** For the labeled data set the number of sequences belonging to each pair of a cellular role and an enzyme category is listed.

	nonenzyme	enzyme	EC1	EC2	EC3	EC4	EC5	EC6
Amino acid biosynthesis	14	70	20	22	1	30	3	3
Biosynthesis of cofactors	75	163	77	39	14	19	7	7
Cell envelope	107	65	2	7	54	1	1	0
Cellular processes	172	87	46	0	35	1	0	0
Central intermediary metabolism	5	208	0	96	103	7	1	1
Energy metabolism	38	289	204	22	38	14	10	1
Fatty acid metabolism	0	52	8	32	0	5	1	10
Purines and pyrimidines	175	345	76	168	88	8	3	10
Regulatory functions	551	30	1	9	4	0	1	0
Replication and transcription	621	115	0	56	24	1	7	3
Translation	135	38	1	0	12	1	0	21
Transport and binding	1201	237	15	69	131	5	2	0

accuracy and vice versa.

Because the cellular role predictors and the enzyme predictors are correlated, self consistency of a prediction is not as strong an indication of the accuracy as could be anticipated. However, inconsistency of a prediction is still an indication that something is likely to be wrong, e.g. a prediction of a non-enzymatic *central intermediary metabolism* protein is unlikely to be correct.

4.5 Protein length distributions revisited

In Paper I we made use of length distributions of proteins to estimate the number of genes in genomes. We saw that the shorter a claimed protein coding regions is, the more likely it is to only be a random ORF. What else can the length of a protein sequence tell us?

From the set of 5,494 SWISS-PROT sequences annotated with cellular roles as described above, a histogram of the protein length distribution was constructed for each cellular role category (see Figure 4.2). It is evident that the protein length distribution is not identical for the 12 different cellular roles: *cell envelope* and *translation* related proteins tend to be short while proteins from the *central intermediary metabolism* have a quite narrow distribution around 400 aa.

Although knowing the length of a protein is obviously not sufficient to predict its function, the length does give a weak hint to the function. The fundamental idea of ProtFun is to integrate many such hints, to obtain a much more qualified guess of what the function might be.

4.6 Sequence derived features

In ProtFun many different predicted protein features are used as input and even more have been tried but rejected. Two features are very closely correlated to the protein length discussed above: the *number of atoms* in the protein and the *molecular weight*. Both of the numbers are reported by the ExPASy ProtParam tool, a programme from the Swiss Institute of Bioinformatics which also provides a number of other simple protein properties. I will here give an overview of all the the features used, starting with the simple properties reflecting the amino acid composition, continuing with structural features and ending with localization features and post-translational modifications.

4.6.1 Number of positively/negatively charged residues

Charged residues play an important role in a proteins interactions with its environment, including other proteins, DNA and RNA. Two very simple features related to this have been used, both of which are reported by the ExPASy ProtParam tool: the total number of negatively charged residues (aspartic and glutamic acid) and the total number of positively charged residues (arginine and lysine). As large proteins will tend to have a larger number of charged residues simply due to their size, these two features also contain some information about the length of the protein.

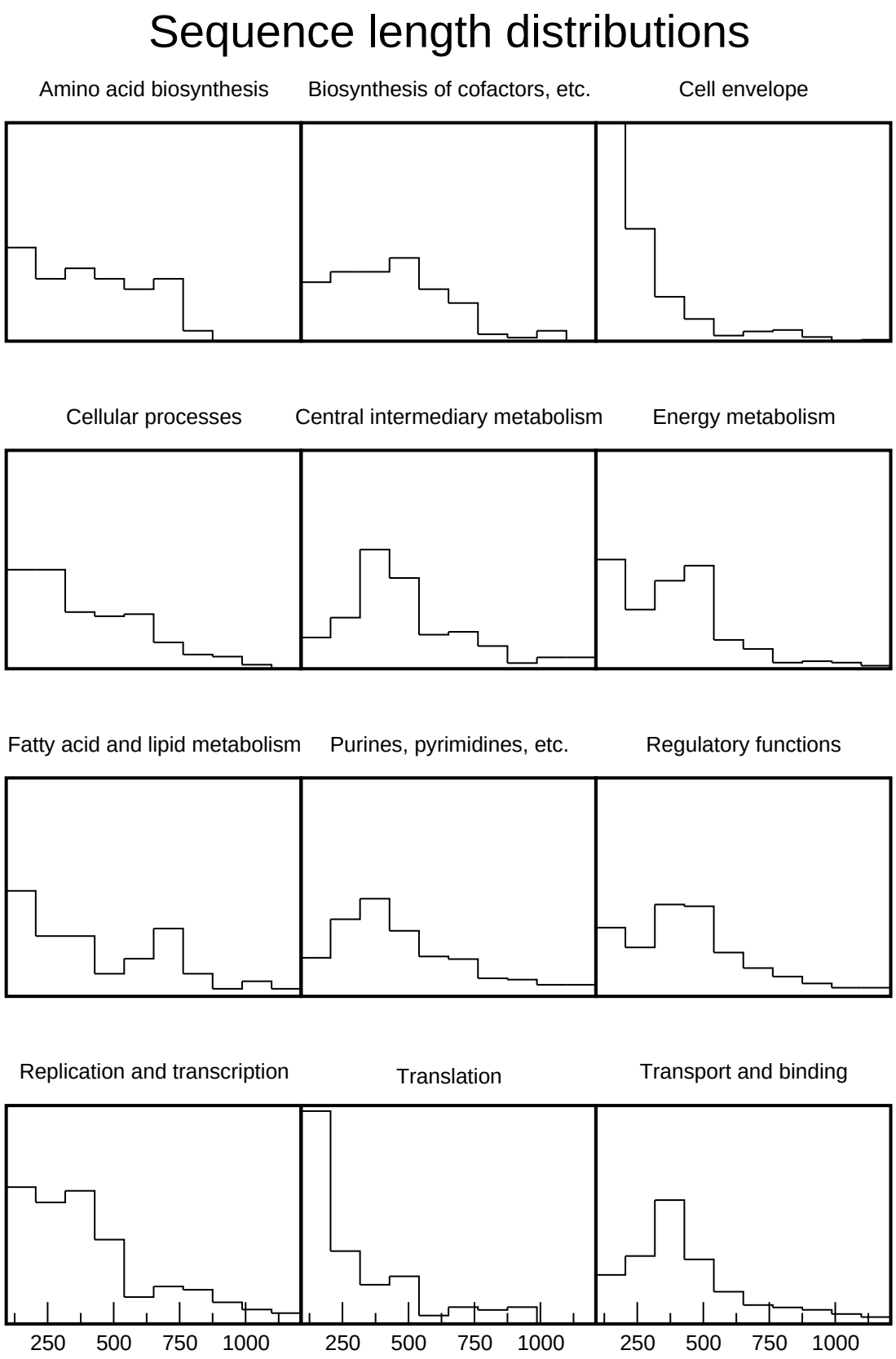


Figure 4.2: **Protein length distributions for cellular role categories.** Histograms with a bin size of 200 aa were created from the positive examples for each category in our labeled data set. The length distributions for different classes of proteins are clearly not identical.

4.6.2 Estimated isoelectric point

A alternative way to represent composition of charged residues is to estimate the *isoelectric point* (pI). The isoelectric point of a protein is defined as the pH at which the protein has no net charge. Proteins with a surplus of basic residues have high isoelectric points as high pH is required for these residues to titrate. Conversely, proteins with a surplus of acidic residues have low isoelectric points.

For a given pH value, the net charge of a protein can be estimated by using the formulas for mixtures of acids and bases in basic chemistry. In this equation all chemical groups in the protein that can titrate at biologically relevant pH values must be taken into account: aspartic acid (D), glutamic acid (E), cysteine (C), histidine (H), arginine (R), lysine (K) and tyrosine (Y) as well as the N- and C-terminal groups. By numerically solving the mixture equation to find the pH at which the net charge is zero, the ExPASy ProtParam tool finds an estimate of the *isoelectric point* of the protein.

The distribution of isoelectric points has been observed to be bimodal for other completely sequenced genomes (Klenk et al., 1997; Kawashima et al., 2000). Figure 4.3 shows that the distribution of estimated isoelectric points is also bimodal for our data set. It may seem puzzling that bimodal distributions are almost always obtained—however this can be fully explained by the fairly low frequencies of histidine and cysteine residues, the only residues to titrate around neutral pH. The pI thus tend to be determined by the titration points of either very basic or very acidic residues (Kawashima et al., 2000).

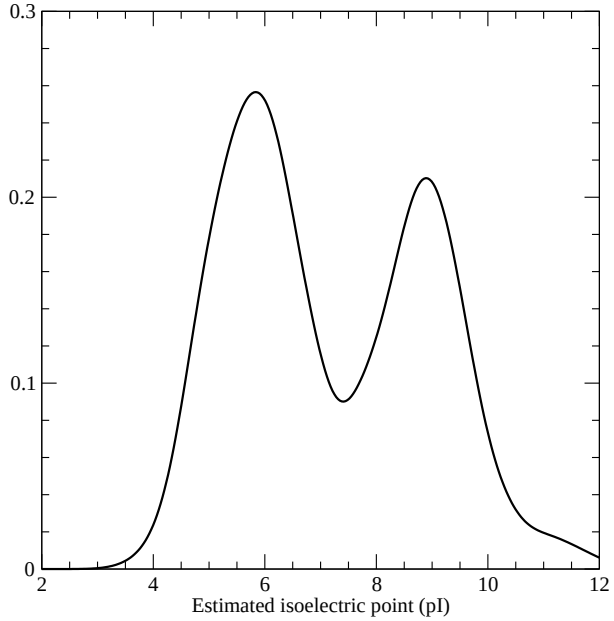


Figure 4.3: **Distribution of isoelectric points for the data set used to train ProtFun.** Based on the estimated isoelectric points of all proteins in our data set, a Gaussian kernel density estimate of the pI distribution was constructed. This distribution is bimodal with peaks around pH 6 and pH 9, which is also observed proteins from other organisms (Klenk et al., 1997; Kawashima et al., 2000).

4.6.3 Extinction coefficient

The *extinction coefficient* is a protein parameter that is commonly used in the laboratory for determining the protein concentration in a solution by spectrophotometry. It describes to what extend light is absorbed by the protein and depends upon the protein size and composition as well as the wavelength of the light.

For a wavelength of 280 nm, the *extinction coefficient* of a protein can be calculated in a simple additive fashion from the number of tryptophans (n_{Trp}), tyrosines (n_{Tyr}), and cystines (n_{Cys}) in the protein:

$$\epsilon_{protein} = n_{Trp}\epsilon_{Trp} + n_{Tyr}\epsilon_{Tyr} + n_{Cys}\epsilon_{Cys},$$

where ϵ_{Trp}, ϵ_{Tyr}, and ϵ_{Cys} are the extinction coefficients of the individual amino acids residues. This calculation is performed by the ExPASy ProtParam tool. It should be noted that ϵ_{Cys} is much smaller than ϵ_{Trp} and ϵ_{Tyr}, meaning that the extinction coefficient mainly reflects the number of tryptophans and tyrosines in the protein.

4.6.4 Grand average of hydropathicity

The hydropathicity of an amino acid residue is defined as the average hydrophobicity of the residues in a window centered at the residue, where the window size is typically 19 residues (Kyte and Doolittle, 1982). The ExPASy ProtParam tool calculates a measure called the *grand average of hydropathicity* ($GRAVY$) which is simply defined as the average hydropathicity over the entire sequence. Whereas the hydropathicity has a value for each position in the sequence, $GRAVY$ is only a single value for each protein.

This measure becomes less mystical when realizing that $GRAVY$ is simply the average of a running average of the hydrophobicity, which is the same as the average hydrophobicity (except for sequence end effects). Examples of the proteins with high $GRAVY$ values include membrane proteins.

4.6.5 Aliphatic index

The *aliphatic index* is defined as the relative volume of a protein occupied by aliphatic side chains (Alanine, Valine, Isoleucine, Leucine), and is again calculated by the ExPASy ProtParam tool. Although the mentioned residues are all hydrophobic, the *aliphatic index* is not a measure of hydrophobicity like the *grand average of hydropathicity* since aromatic residues are not included in the measure. As the aliphatic side chains are both hydrophobic and flexible, they play an important role in the hydrophobic packing of the protein core. Consistent with this, the *aliphatic index* has been proposed to be correlated with the thermostability in the case of globular proteins.

4.6.6 Protein instability index

The stability of a protein as defined by its half-life turns out to be correlated to the dipeptide composition. These correlations can be summarized in a measure

called the *protein instability index*, which relies on a weight value (DIWV) for each of the 380 possible dipeptides (Guruprasad et al., 1990). The *instability index* (*ii*) is defined as:

$$ii = \frac{10}{L} \sum_{i=1}^{i=L-1} DIWV(AA_i, AA_{i+1}),$$

where L is the sequence length and AA_i the amino acid in position i. The score for a protein is in other words simply 10 times the average instability contribution of all dipeptides in the sequence. Like the other features mentioned so far, it is calculated by the ExPASy ProtParam tool. Proteins with high instability scores (above 40) are typically unstable, whereas low scores are characteristic of stable proteins.

4.6.7 Predicted PEST regions

A different type of instability stems from the active degradation of certain proteins by the proteasome complex. The degradation of some proteins by the proteasome has been shown to be caused by sequences called *PEST regions* (Rechsteiner and Rogers, 1996). The regions are called this because they are rich in Proline (P), Glutamic acid (E), Serine (S), Threonine (T) and to a lesser extend Aspartic acid (D) residues. There does not seem to be a true consensus sequence, only strong sequence biases.

The requirements for a *PEST region* appear to be a hydrophilic stretch of at least 12 residues rich in P, E, S, and T residues. This stretch should be flanked by the positively charged residues Lysine (K), Arginine (R) and Histidine (H), which are not allowed with in the PEST sequence itself (Rechsteiner and Rogers, 1996). An algorithm developed elsewhere called PESTfind has been used to predict candidate PEST sequences in proteins.

4.6.8 Low complexity regions

PEST regions constitute only one type of region with a compositional bias, that is regions that are composed of only a few amino acid residue types. Other examples include homo-polymeric amino acid runs and short tandem repeats. Although such "simple" sequences have been largely neglected in protein sequence studies, it is becoming increasingly evident that various types of biased regions exist and play important structural and/or functional roles (Wootton, 1994b; Karlin, 1995).

Compositionally biased regions can be defined in terms of their low sequence complexity, and were identified using the *SEG* programme (Wootton, 1994a), which is the programme used for masking of protein sequences in the BLAST package (Altschul et al., 1997).

4.6.9 Protein secondary structure

Because the structure of a protein is very important for its function, predicted secondary structure was included as a feature. While it is not clear how much of

the structurally important features can be captured at the secondary structure level, it is all that we can currently predict with reasonable accuracy for globular proteins.

Prediction of secondary structure was done using *PSI-Pred*, which compared very favorably to other methods in the CASP experiment (Sternberg et al., 1999). However, it should be kept in mind that we will get a poorer performance as *PSI-Pred* must run in single sequence mode to allow prediction on orphan proteins. The output of *PSI-Pred* is a three state prediction of the secondary structure (helix, sheet, or coil) for each residue in the protein.

4.6.10 Transmembrane helix predictions

In addition to globular proteins, a large fraction of non-globular proteins exist, many of which reside in the various membranes of the cell. With the exception of β-barrel porins, such proteins all span the membranes by hydrophobic alpha-helices known as transmembrane helices. The perhaps most famous class of transmembrane proteins are the 7TM receptors.

From a function prediction point of view, it is interesting to predict not only whether a protein is transmembrane or not, but also the amount and sequence location of transmembrane segments as well as the membrane topology of the protein, i.e. the orientation of the protein in the membrane.

To make these predictions we made use of the TMHMM method also developed at CBS (Sonnhammer et al., 1998; Krogh et al., 2001). As the name indicates, the central part of the method is an HMM of transmembrane proteins. By Viterbi decoding of a protein sequence, each residue of the sequence is assigned as being either inside the membrane, outside, or part of a transmembrane helix.

4.6.11 Signal peptide prediction

Many proteins have N-terminal signal peptides which cause the protein to be translocated across the membrane to enter the endoplasmic reticulum (ER). This is initiated by interaction between the signal peptide and the signal recognition particle (SRP) during translation. This guides the ribosome, mRNA, and emerging polypeptide chain to the ER membrane where the protein chain is co-translationally translocated into the ER and the signal peptide is cleaved off. Once in the ER, the protein can either become secreted, be transferred to the Golgi apparatus or a lysosome, or be retained in the ER.

Signal peptides are characterized by having a short positively charged N-terminal region, a longer hydrophobic region and finally a cleavage site region required to contain small residues at positions -3 and -1 relative to the point of cleavage. Signal peptides are predicted using the *SignalP* prediction method which combines several neural networks trained to pick up these signals (Nielsen et al., 1997a,b, 1999). For each of the first 60 residues of the protein sequence several scores are calculated: an S-score which tells if the residue appears to be part of the signal peptide, a C-score which should ideally be high for the cleavage

site only, and finally a combined Y-score, being the geometric mean of the C-score and a smoothed derivative of the S-score.

Since the development of the original ProtFun method, SignalP has been combined with a separate mitochondrial transit peptide predictor and a chloroplast transit peptide predictor to make a method called TargetP (Emanuelsson et al., 2000). For non-plants, this method gives three scores predicting the type of protein: secreted/cell surface protein, mitochondrial protein, or protein without N-terminal signal peptides. The three TargetP scores have been included as a feature in later applications of the ProtFun approach.

4.6.12 Subcellular localization

The eukaryotic cell is divided into several subcellular compartments or localizations each containing a different complement of proteins. Two features related to subcellular localization have already been described, namely prediction of transmembrane helices and signal peptides which are in part controlling the sorting of proteins into their respective compartments. However, many other sequence signals exist as well.

To predict the subcellular localization of proteins, we used the *PSORT* prediction method (Nakai and Horton, 1999). PSORT discriminates between 11 different localizations: extracellular (including cell wall), endoplasmic reticulum, vacuolar, mitochondrial, cytoplasmic, nuclear, cytoskeletal, Golgi, vesicles of secretory system, peroxisomal, and plasma membrane.

PSORT relies on an array of different prediction methods, including predictors for signal peptides and transmembrane helices. These are combined by a rule based expert system to assign a probability to each of the subcellular localizations for each protein. A full list of the features used by *PSORT* can be found at `http://psort.nibb.ac.jp`.

4.6.13 N-linked GlcNAc glycosylation sites

Many types of sugar-amino acid linkage are known today, making glycosylation the most complex of the co- and post-translational modifications. Different types of glycoproteins are found to be involved in a wide variety of functions, in particular in eukaryotes (Spiro, 2002).

N-linked GlcNAc glycosylation of asparagine (N) is the most widely occurring type of glycosylation, mainly targeting secreted and membrane bound proteins. The oligosaccharide is co-translationally linked to the protein in the endoplasmic reticulum (ER) by the enzymatic protein complex oligosaccharyltransferase, which is conserved among eukaryotes and also has been found in archaea (Spiro, 2002).

The oligosaccharyltransferase recognizes the motif Asn-Xaa-Ser/Thr, where Xaa is any other amino acid than proline. As this motif is a necessary but not sufficient condition for glycosylation to take place, we used the neural network based predictor *NetNglyc* developed at CBS for predicting N-linked GlcNAc glycosylation sites.

4.6.14 O-linked GalNAc glycosylation sites

O-linked glycosylations constitute a varied class of sugar-amino acid linkages, as several types of sugars can be linked to various amino acid residues and in different configurations (Spiro, 2002). The most studied of such glycosylation is the O-α-GalNAc glycosylation of serines and threonines. Like the N-linked GlcNAc glycosylation described above, this type of glycosylation is mainly observed for secreted and membrane bound proteins. However, this type of glycosylation takes place post-translationally in the Golgi, where sugars are added successively to form branched oligosaccharide structure.

O-linked GalNAc glycosylation is mediated by several different GalNAc-transferase, which are likely to have different sequence specificities. As a result of this, a real consensus sequence is not known. However, O-linked GalNAc glycosylations often can be found clustered in repeat regions rich in serines, threonines, prolines, glycines and alanines, which make PEST regions likely targets for this type of glycosylation.

Although no good consensus sequence is known, a prediction method for O-linked GalNAc glycosylation, *NetOglyc*, has been successfully developed at CBS (Hansen et al., 1998). This method uses a set of neural networks to score possible glycosylation sites and subsequently subtracts a variable threshold based on the surface accessibility predicted by other networks. The *NetOglyc* scores for potential O-α-GalNAc were used as an input feature for function prediction.

4.6.15 O-linked β-GlcNAc glycosylation

A different type of O-linked glycosylation of serines and threonines, which has recently received much attention, is the O-β-GlcNAc glycosylation. This type of glycosylation targets nuclear and cytoskeletal proteins, and is the first type of glycosylation reported to occur outside the endoplasmic reticulum (ER) (Hart, 1997). It is also characteristic by being a monosaccharide modification. O-β-GlcNAc glycosylations are linked by the enzyme GlcNAc-transferase, which like the oligosaccharyltransferase complex is highly conserved among eukaryotes (Spiro, 2002).

Intriguingly, this type of glycosylation appear to often target serines and threonines that can also be phosphorylated. Such sites that can be both glycosylated and phosphorylated are known as Yin-Yang sites. Since these two post-translational modifications are mutually exclusive, O-β-GlcNAc glycosylation has been proposed to play a reciprocal regulatory role to phosphorylation.

At CBS a neural network based method for prediction of O-β-GlcNAc glycosylation has been developed as part of the *YinOYang* prediction server (`http://www.cbs.dtu.dk/services/YinOYang/`). Only the O-β-GlcNAc predictions were included in this feature as phosphorylations were handled separately.

4.6.16 Serine and threonine phosphorylation

Phosphorylation of serines and threonines, which was mentioned in parsing above, is one of the most important regulatory mechanisms. Because phosphate groups

can be reversibly added and removed by kinases and phosphorylates, phosphorylation often acts as a direct on/off switch of protein activity. This is a very common regulatory mechanism, which is used for regulation of essentially all processes that take place in a cell (Cohen, 2000).

Although phosphorylation of serines and threonines are performed by a large number of different kinases, each with its own motif specificity, the same kinases are often involved in phosphorylation of serines as well as threonines. It is thus not meaningful to discriminate between serine and threonine phosphorylation. Also, because ≈ 1000 kinases are believed to be encoded by the human genome (Cohen, 2000), no general consensus sequence for phosphorylation sites exist (Blom et al., 1999).

As it is rarely known which kinase is responsible for the phosphorylation of a particular site, kinase specific predictors cannot be constructed. Therefore, the generic phosphorylation predictor, *NetPhos*, was used for predicting serine and threonine phosphorylation sites (Blom et al., 1999).

4.6.17 Tyrosine phosphorylation

Like serines and threonines, tyrosines residues are also subject to reversible phosphorylation. Although tyrosine phosphorylation is also used as a regulatory mechanism, it is governed by an entirely different complement of kinases and phosphorylases. It might therefore be used to control different processes, for which reason it has been considered a separate input feature. Tyrosine phosphorylation sites were also predicted using NetPhos (Blom et al., 1999).

4.7 Feature representation

The features described are of two fundamentally different types: global features and local position specific features. The global features are characterized by consisting of one or more values for each protein—but always the same number of values. Examples of global features included are all the properties calculated by the ExPASy ProtParam program. After normalization, these features can be used as input to neural networks.

The reason for normalizing the data is, that the values representing different features can differ by orders of magnitude. Experience shows that such large differences often give rise to numerical difficulties when training neural networks by backpropagation. This is avoided by normalizing all values to the same scale. With the exception the PSORT probabilities which are already on the right scale, we converted all values to Z-scores, i.e. subtracted the mean value and divided by the standard deviation. For features related to the sequence length this was preceded by a log-transformation.

Position specific features, in contrast to global features, consist of one or more values per residue. Consequently, the number of values will vary with the sequence length and cannot be used directly as input to neural networks, as they require a fixed number of inputs. It is important to find a good way to encode this type of features as there are many of them. Examples include all the glycosylations,

phosphorylations and structure related features.

4.7.1 Encoding positional information

In order to preserve some positional information, the position specific features were encoded as average values within a number of bins representing different parts of the sequence. Several binning schemes were tested. The simplest was to simply divide the sequence into a number of evenly sized bins—we call this dynamic bins. The main drawback of dynamic is that a bin does not represent the same amount of sequence for all proteins. In an alternative approach called fixed bins, a number of bins of fixed size are located along the sequence with equal distance between their centers. The size of the bins is such that for all but the longest sequences the bins are overlapping, thus covering the complete sequence. Using this scheme a bin always represent the same amount of sequence—the price paid is that a residue can be part of more than one bin. Finally a composite scheme was tried where a fixed size bin was located at either end of the sequence and the remainder of the sequences was put into dynamic bins. This was motivated from the knowledge that the ends of sequences often convey special information, e.g. signal peptides and GPI anchors.

4.7.2 The choice of feature of representations

Encoding of the global features is as mentioned not a big problem, and the final representations were decided based by careful considerations rather than experimentation. Many features were simply subjected to a linear transformation to normalize them to a proper interval. Features with very skewed distributions, e.g. number of positively and negatively charged residues, were log-transformed first (see Table 4.2).

The representation of the SignalP feature deserved special attention. The SignalP predictor does not output one but four different scores for judging if a sequence starts with a signal peptide or not. The two most indicative of these (the meanS and maxY scores) were included in the representation leaving it to the ProtFun neural networks to join them into one prediction. Also the log transformed position of the maxY score was included as an indicator of the length of the possible signal peptide.

It is not at all obvious which of the binning schemes to use to encode each of the various positional features. Neural networks were thus trained for each positional feature individually to find the best representation. As it is also important to keep the dimensionality of each feature down to allow more features to be combined later on, the choice of feature encoding was often a tradeoff between high individual feature performance and low feature complexity. For this reason the final choices were largely subjective (see Table 4.2), but could instead have been made based on a minimal description length criterion.

Table 4.2: **Encoding of the individual features.** For each feature is listed the abbreviation of the feature, the encoding scheme used, and a brief description of the feature.

Abbreviation	Encoding	Description
EC	single value	Extinction coefficient predicted by ExPASy ProtParam
GRAVY	single value	Hydrophobicity predicted by ExPASy ProtParam
Nneg	single value	Number of negatively charged residues counted by ExPASy ProtParam
Npos	single value	Number of positively charged residues counted by ExPASy ProtParam
Nglyc	potential in 5 bins	N-glycosylation sites predicted by NetNGlyc
Oglyc	potential-threshold in 10 bins	GalNAc O-glycosylations predicted by NetOGlyc
PEST	fraction in 10 bins	PEST rich regions identified by PESTfind
PhosST	potential in 10 bins	Ser and Thr phosporylations predicted by NetPhos
PhosY	potential in 10 bins	Tyr phosporylations predicted by NetPhos
PSI-Pred	helix, sheet, and coil in 5 bins	Predicted secondary structure from PSI-Pred
PSORT	20 probabilities	Subcellular location predtions by PSORT
SEG	fraction in 5 bins	Low complexity regions masked by SEG filter
SignalP	meanS, maxY, and log(cleavage position)	Signal peptide predictions made by SignalP
TMHMM	inside, outside, and membrane in 5 bins	Transmembrane helix predictions made by TMHMM

4.8 Developing the prediction method

At this point, a labeled data set, split into test and training sets, has been created for each cellular role and enzyme category. A wide range of predicted or calculated proteins as well as their representations are also in place. All is ready for developing the actual prediction method.

4.8.1 Individual neural network training

In the development of ProtFun, all neural networks were trained using the back-propagation algorithm (Werbos, 1974; Parker, 1982; Rumelhart et al., 1986) to minimize the squared error function. For finding the best networks, the Pearson correlation coefficient on the test set has been used consistently as the measure for neural network performance. This may appear as strange, since most people would prefer to use Mathews correlation coefficient (Mathews, 1975) for such classification problems. The main reasoning for this is, that the Pearson correlation coefficient fluctuates much less during training than Mathews correlation coefficient, which makes the feature selection procedure described next more robust. Also, I will argue that the Pearson correlation is appropriate when the final predictions are to be probabilistic rather than discretized.

4.8.2 Optimization of feature combinations and network architecture

The first strategy attempted to find the best feature combination for each functional class was, to initially train a neural network for each class and subsequently prune it to remove features not contributing to the performance. However, this approach proved to not be possible due to the large number of sequence derived

features. The neural networks could simply not generalize to the test set at all, as the input space was of too high dimensionality compared to the number of training examples. There were thus no networks to prune.

It was also not possibly to simply try out all possible combinations of features, as this would give rise to a combinatorial explosion in the number of networks that should be trained. Instead we opted for a bootstrap approach, in which networks were initially trained for each individual feature, the best of which were subsequently combined to form larger and larger feature combinations.

The traditional way to do this would be to use a greedy search algorithm for selecting the features to include. Here, we instead choose to use the greedy algorithm for selecting features to exclude. Given a finite number of features, specifying which features *not* to be included is obviously just as good as specifying which ones to include. The motivation for doing it this way was that, due to cross correlations between features, there were not always features that clearly would be a part of the optimal feature combination. Yet, there were always some features that showed no correlation at all to the functional class in question.

Briefly, our feature selection algorithm works as follows: First neural networks are trained for each of the individual features. Judging from the maximal test set Pearson correlation coefficient, the best nine features are selected and networks are trained for all pairwise combinations of these. Each feature is assigned the correlation coefficient of the best combination it is involved in, and best eight features are retained. For these, all combinations of three features are used as input for neural networks and the worst feature is again discarded, leaving seven features. All combinations of 4 to 7 of these were then tested.

The five best feature combinations encountered during the training procedure were noted. For each of these, the number of units in the hidden layer was varied in steps of 10, to find an approximate optimum, again maximizing the test set Pearson correlation coefficient. The resulting feature usages and network architectures for cellular role and enzyme categories are listed in Table 4.3.

4.8.3 Estimating probabilities neural network outputs

The many very different neural networks used present a problem: the output scores of such neural networks follow different distributions, reflecting that the outputs from different networks are incomparable. This makes simple averaging of the network ensemble dangerous as well as making interpretation of the scores for different functional classes difficult.

To my knowledge, the solution that I came up with is novel although it should be applicable to a wide range of such problems. Based on the test set output scores of each neural network, Gaussian kernel density estimates were made of the score distributions for positive and negative examples. These probability densities are for a given neural network denoted $f_{pos}(x)$ and $f_{neg}(x)$. From these it is possible to estimate the probability that an example is positive given a network output x:

$$P(x) = \frac{N_{pos} f_{pos}(x)}{N_{pos} f_{pos}(x) + N_{neg} f_{neg}(x)},$$

where N_{pos} and N_{neg} are the number of positive and negative examples in our

labeled data set, specifying for the prior probability of a positive example.

Having estimated $P(x)$ for a neural network, this function can now be used to normalize the raw output scores from the network to P-values, which are comparable both within and across categories. In the final output of the method, the P-values of the networks in each ensemble are averaged to estimate the probability for each cellular role and enzyme class.

Table 4.3: **Architecture and feature usage of the individual neural networks used for prediction of cellular role and enzyme categories by the ProtFun server.** All five neural networks in each ensemble are listed with their number of hidden neurons (H) and input features. Refer to Table 4.2 on page 55 for feature abbreviations.

Category	H	Input features
Amino acid biosynthesis	30	EC, PSI-Pred, PSORT, TMHMM
	30	EC, PSI-Pred, TMHMM
	30	EC, netOglyc, PSI-Pred, PSORT
	30	GRAVY, PSI-Pred, PSORT
	30	Oglyc, PSI-Pred, PSORT
Biosynthesis of cofactors	50	GRAVY, PEST, PSORT, TMHMM
	50	EC, PSI-Pred, PSORT, SEG
	30	EC, PSI-Pred, PSORT, TMHMM
	30	GRAVY, PSI-Pred, PSORT, SEG
	40	EC, PSI-Pred, PSORT, SEG
Cell envelope	40	SignalP, TMHMM
	40	Nglyc, PSORT, SignalP, TMHMM
	30	Nglyc, PSI-Pred, PSORT, SignalP TMHMM
	30	PSORT, SignalP, TMHMM
	30	PSI-Pred, PSORT, SignalP, TMHMM
Cellular processes	30	GRAVY, Nglyc, PSORT, SEG
	30	GRAVY, PhosST, PSI-Pred
	30	PhosST, PEST, PSORT, SEG
	30	PEST, PSORT, SEG
	30	PSORT, SEG
Central intermediary metabolism	50	EC, Nneg, Npos, PSI-Pred, PSORT, TMHMM
	30	Nneg, Npos, PSI-Pred, PSORT, TMHMM
	30	Nneg, Npos, PSI-Pred, TMHMM
	30	EC, PSI-Pred, PSORT
	30	Nneg, Npos, PSI-Pred
Energy metabolism	10	EC, PhosST, PhosY, PEST, PSORT, SignalP
	40	EC, PhosST, PhosY, PSORT, SignalP
	50	EC, PEST, PSORT, SignalP
	30	PhosY, PEST, PSORT, SignalP
	30	PEST, PSORT, SignalP
Fatty acid metabolism	30	GRAVY, PhosST, PEST, SignalP
	30	GRAVY, SEG, SignalP
	30	GRAVY, PEST, SEG, SignalP
	30	GRAVY, SEG
	30	PEST, PSI-Pred, SEG, SignalP
Purines and pyrimidines	10	GRAVY, Nneg, Npos, PSORT
	40	GRAVY, Nneg, TMHMM
	20	GRAVY, Nneg, Npos, TMHMM
	30	GRAVY, Nneg, Npos, PSORT
	30	EC, GRAVY, Nneg, TMHMM
Regulatory functions	30	PhosST, PhosY, PEST, PSORT
	30	Oglyc, Nglyc, PEST, PSORT
	30	Oglyc, PhosST, PhosY, PEST, PSORT
	30	Npos, Nglyc, PEST, PSORT
	40	PEST, PSORT
Replication and transcription	30	Oglyc, Nglyc, PSORT, TMHMM
	30	Oglyc, PSORT,
	30	Oglyc, PSORT, TMHMM,
	30	Nglyc, PSORT, TMHMM,
	30	GRAVY, Nglyc, PSORT, TMHMM
Translation	30	PhosY, Nglyc, PEST, SignalP
	30	Oglyc, PEST, SignalP
	30	Oglyc, PhosY, Nglyc, SignalP
	10	Oglyc, PEST, SignalP, TMHMM
	10	PhosY, Nglyc, PEST, SignalP
Transport and binding	40	EC, Nglyc, PSORT, SignalP
	30	Npos, PSORT
	40	EC, GRAVY, Nglyc, PSORT, SignalP
	30	GRAVY, Nglyc, PSORT, SignalP
	30	EC, PSORT
Enzyme/ non-enzyme	40	EC, Nneg, Npos, PSORT, TMHMM
	40	EC, Npos, PhosY, PSI-PRED, PSORT
	10	EC, Nneg, PSORT
	10	EC, Nneg, Npos, PSORT
	40	EC, PSORT, TMHMM
Oxireductase	10	EC, Oglyc, PEST, PSORT, SignalP
	30	EC, Oglyc, PSORT, SignalP
	30	EC, GRAVY, PhosY, PEST, PSORT
	50	EC, PhosY, PEST, PSORT, SignalP
	40	EC, Oglyc, PSORT, SignalP
Transferase	50	Nneg, Nglyc, PSORT, TMHMM
	30	Nneg, PhosY, PSORT, SEG
	30	Nneg, Nglyc, PSORT, SEG
	20	Nneg, phosY, PSORT
	50	EC, Nneg, PhosY, PSORT, TMHMM
Hydrolase	50	EC, GRAVY, PSI-PRED, PSORT, SignalP, TMHMM
	30	EC, Nneg, PSI-PRED, TMHMM
	20	EC, GRAVY, PSI-PRED, SignalP
	30	EC, GRAVY, PSI-PRED, TMHMM
	50	EC, PSORT, SignalP, TMHMM
Lyase	50	GRAVY, Nneg, PhosST, Nglyc, TMHMM
	30	Nglyc, PSORT, TMHMM
	50	Nneg, Npos, PhosST, PSORT, TMHMM
	30	GRAVY, Nneg, PhosST, Nglyc, PSORT, TMHMM
	30	GRAVY, Nneg, Nglyc, PSORT, TMHMM
Isomerase	30	Nneg, Npos, PSI-PRED, SEG, TMHMM
	30	Nneg, PEST, PSI-PRED, TMHMM
	30	Nneg, PEST, PSI-PRED, SEG
	30	EC, Nneg, PSI-PRED
	30	EC, Npos, PEST, PSI-PRED, TMHMM
Ligase	30	Nneg PSI-PRED, PSORT
	30	PEST, PSI-PRED, PSORT, TMHMM
	10	Nneg, PSI-PRED, SignalP, TMHMM
	30	EC, GRAVY, Nneg, PSI-PRED
	30	GRAVY, Nneg, PSI-PRED

Paper II

4.9 Prediction of human protein function from post-translational modifications and localization features

Lars Juhl Jensen[1†], Ramneek Gupta[1†], Nikolaj Blom[1], Damien Devos[2], Javier Tamames[2], Can Kesmir[1], Henrik Nielsen[1], Hans-Henrik Stærfeldt[1], Kristoffer Rapacki[1], Christopher Workman[1], Claus A. F. Andersen[1], Steen Knudsen[1], Anders Krogh[1], Alfonso Valencia[2], and Søren Brunak[1*]

[1] Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby, Denmark

[2] Protein Design Group
National Center for Biotechnology
CNB-CSIC
Cantoblanco, Madrid E-28049, Spain

[†] These authors contributed equally
[*] To whom correspondence should be addressed (email: brunak@cbs.dtu.dk)

We have developed an entirely sequence-based method which identifies and integrates relevant features that can be used to assign proteins of unknown function to functional classes, and for enzymes enzyme categories. We show that strategies for the elucidation of protein function may benefit from a number of functional attributes which are more directly related to the linear sequence of amino acids—and hence easier to predict—than protein structure. These attributes include features associated with post-translational modifications and protein sorting, but also much simpler aspects such as the length, isoelectric point and composition of the polypeptide chain.

Introduction

Out of the 30,000 to 45,000 genes believed to be present in the human genome no more than 40-60% can be assigned a functional role based on homology to proteins with known function (Int. Human Genome Sequencing Consortium, 2001; Venter et al., 2001). Traditionally, protein function has been related directly to the three-dimensional structure of the poly-peptide chain, which currently, for an arbitrary sequence, is quite hard to compute (Lesk et al., 2001). The method presented here operates in the "feature" space of all sequences, and is therefore complementary to methods which are based on alignment and the inherent, position-by-position quantification of similarity between two sequences. The method does not require knowledge of gene expression (Eisen et al., 1998), gene fusion and/or phylogenetic profiles (Marcotte et al., 1999b,a; Hughes et al., 2000a; Pellegrini et al., 1999). Although the latter type of method does not rely on finding direct matches to proteins of known function, it does require sequence similarity to other candidates that can be phylogenetically linked to a protein of known function.

For any function assignment method, the ability to correctly predict the relationship depends strongly on the function classification scheme used. One would for example not expect that a method based on co-regulation will work well for a category like "enzyme", since enzymes and the genes coding for their substrates or substrate transporters often may display strong co-regulation. A similar argument holds true for the phylogenetic profile method.

Our approach to function prediction is based on the fact that a protein is not alone when performing its biological task. It will have to operate using the same cellular machinery for modification and sorting as all the other proteins do. Essential types of post-translational modifications (PTMs) include: N- and O-glycosylation, (S/T/Y) phosphorylation, and cleavage of N-terminal signal peptides controlling the entry to the secretory pathway, but hundreds of other types of modifications exist (Garavelli et al., 2001) (a subset of these will be present in any given organism). Many of the PTMs are enabled by local consensus sequence motifs, while others are characterized by more complex patterns of correlation between the amino acids (Blom et al., 1999).

This suggests an alternative approach for function prediction, as one may expect that proteins performing similar functions would share some attributes even though they are not at all related at the global level of primary structure. As several predictive methods for PTMs have been constructed (Blom et al., 1999; Hansen et al., 1998; Nielsen et al., 1999; Nakai and Horton, 1999), a function prediction method based on such attributes can be applied to all proteins where the sequence is known.

Results and discussion

The ProtFun method described here integrates (using a neural network approach) 14 individual attribute predictions and calculated sequence statistics (out of more than 25 tested for discriminative value). The integrated method predicts functional categories as defined originally by Riley for *E. coli*, that in modified form

has been used to describe many entire genomes in recent publications (Int. Human Genome Sequencing Consortium, 2001; Venter et al., 2001; Riley, 1993; Fleischmann et al., 1995). In addition, it predicts whether a sequence is likely to function as an enzyme, and if so, its category according to the classes defined by the Enzyme Commission (Enzyme Nomenclature, 1965, 1992). The same scheme can be used to predict any other set of functional classes including more narrowly defined class ones. We have applied the approach to for instance specifically identify hormones, receptors and ion channels in the human genome as defined by the Gene Ontology Consortium (Ashburner et al., 2000).

We have used combinations of attributes in a collection of neural network ensembles for predicting the functional category of a protein. Combinations of attributes were selected by evaluating their discriminative value for a specific functional category, say proteins involved in transcription. Attributes useful function prediction must not only correlate well with the functional classification scheme, but must also be predictable from sequence with reasonable accuracy.

Interestingly, the combinations of attributes selected for a given category also implicitly characterize a particular functional class in an entirely new way. The method identifies without any *a priori* ranking of their importance, the biological features relevant for a particular type of functionality, see Figure 4.4. It appears that the use of post-translational modifications (PTMs) is essential for the prediction of several functional classes. In addition to attributes related to subcellular location the most important features for predicting if a protein is for example, regulatory or not, are PTMs. Similarly PTMs are very important for correct assignment of proteins related to the cell envelope, replication and transcription.

The fact that (predicted) PTMs correlate strongly with the functional categories fits well with biological knowledge. For example, predicted N-glycosylation sites turn out to be important for prediction of cell envelope proteins. In fact, it has been shown that removal of carbohydrates linked to asparagines from a protein normally targeted for the cell envelope retains it in the endoplasmic reticulum (Chen and Colley, 2000).

For proteins with "regulatory function" two of the most important features were S/T phosphorylation and Y-phosphorylation, respectively (Figure 4.4). It is very satisfying that this correlation was found by the neural networks when considering that reversible phosphorylation is a well known and widely used regulatory mechanism (Cohen, 2000). Glycosylation was also found to be a strong indicator for regulatory proteins. This is true for both N-glycosylation and O-GalNAc (mucin type) glycosylation of serine and threonine residues. For these proteins, two additional features had significant predictive value: The predicted subcellular location, and PEST regions (rich in proline, glutamic acid, serine, and threonine residues), where the latter targets proteins for degradation. Again, it makes sense that proteins involved in fast regulatory mechanisms should be degraded quickly (Rechsteiner and Rogers, 1996).

In order to understand further how PTMs correlate with the functional categories, we investigated the effect of alternative representations of the PTM attributes. For example, when phosphorylation of serine and threonine residues were encoded as two separate features the result was a slightly reduced predictive performance. Joining all three types of phosphorylation (S/T/Y) into one single

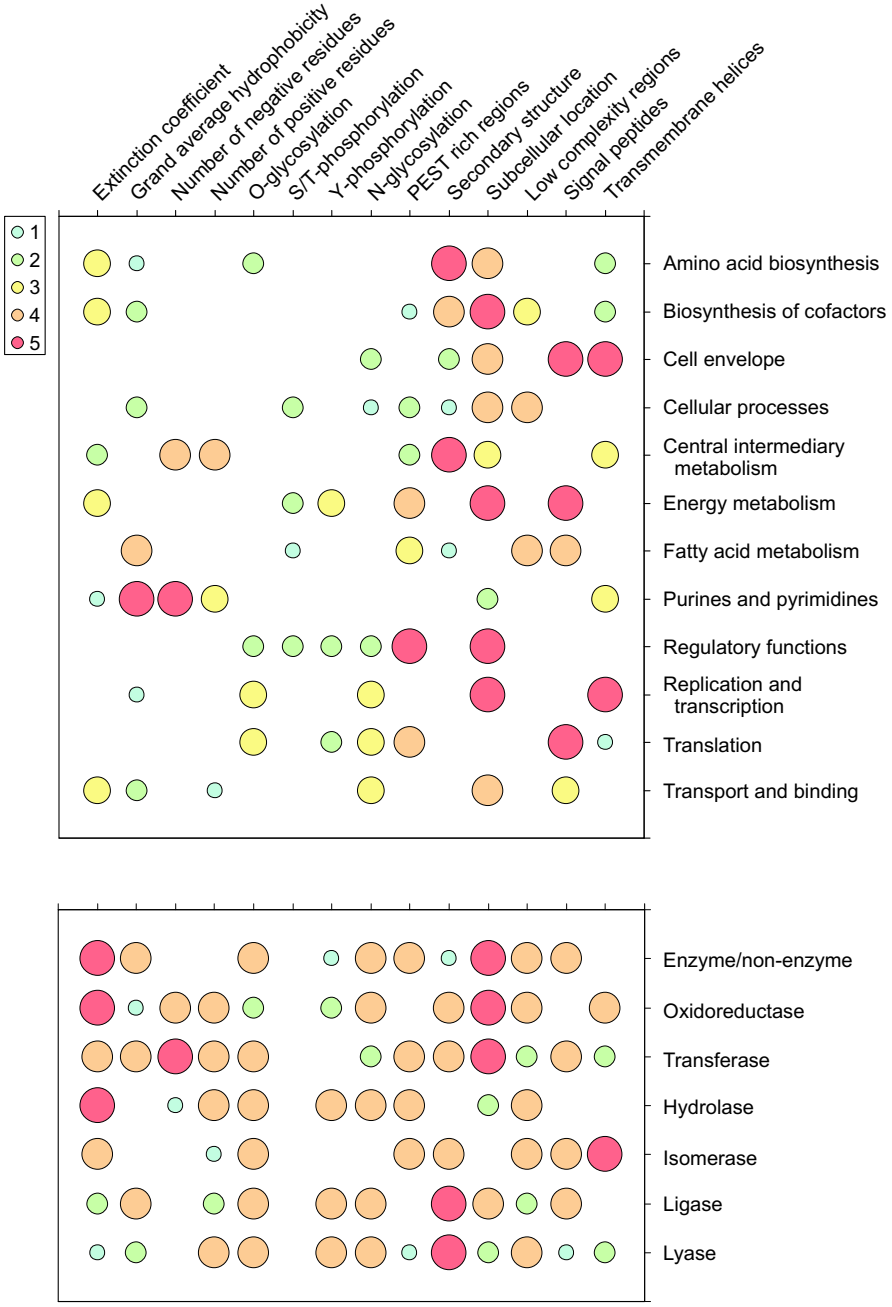


Figure 4.4: **The discriminative impact of features for the different functional categories and enzyme classes.** The figure shows how often a given feature (out of the 14 retained) was included in the five network ensemble performing the classification for a given category and thus it importance. The retained feature predicting methods were: NetNGlyc (Ramneek Gupta, unpublished results), NetOGlyc (Hansen et al., 1998), NetPhos (Blom et al., 1999), PEST regions (Rechsteiner and Rogers, 1996), PSIPRED (Jones, 1999), SEG filter (Wootton, 1994a), SignalP (Nielsen et al., 1999), PSORT (Nakai and Horton, 1999), TMHMM (Krogh et al., 2001). In addition, a number of calculated features were retained: extinction coefficient, grand average hydrophobicity, and the numbers of positively and negatively charged residues. During the feature selection process 11 features were *not* retained due to their low discriminatory value (or their correlation to other features retained): the amino acid composition, the composition of residues predicted to be buried or exposed, the aliphatic index, instability index, number of atoms, the net charge, the isoelectric point, predicted GlcNAc sites, the sequence length, and predicted coiled coil regions.

feature led to a much larger drop in performance. Again this makes biological sense as serine and threonine residues are known often to be phosphorylated by the same group of kinases, while tyrosine residues typically are phosphorylated by different kinases.

The most important single feature for distinguishing between enzymes and non-enzymes turned out to be protein secondary structure as predicted by PSIPRED (Jones, 1999). This also makes sense, as enzymes are known to be overrepresented among all-alpha proteins, and more rarely are found to be e.g. all-beta proteins.

We also trained networks for predicting the enzyme subclasses. Although these networks were trained specifically to discriminate between a given enzyme subclass and all other enzyme subclasses, they implicitly make an enzyme/non-enzyme prediction as well. The enzyme class predictions can thus be used as additional support for the predictions made by the enzyme/non-enzyme networks.

Quantitative description of the ProtFun predictive performance

The selection of category-relevant attributes is based on quantitative assessment of the ability to predict (assign) categories for new sequences non-similar to the sequences used to train the method (see below). Figure 4.5 shows how the ProtFun method fairs for the prediction of functional and enzyme categories in terms of sensitivity and the level of false positives. When the sensitivity is below 40%, the level of false positive predictions is very low. The confidence in the predictions can be used directly to sift out those predictions which almost certainly are correct. The way the probabilities are estimated gives rise to an almost linear

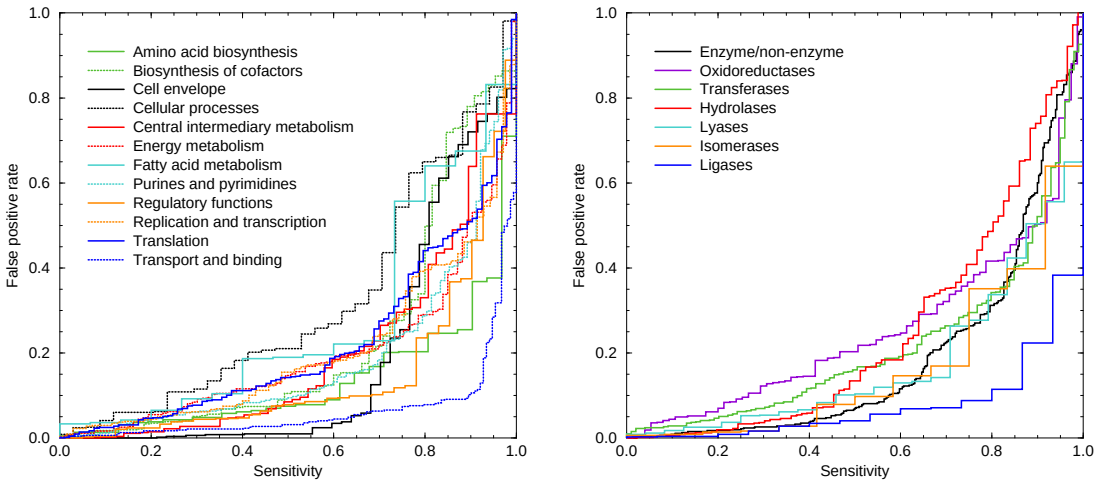


Figure 4.5: **The predictive performance shown as sensitivity vs. false positive rate for cellular role and enzyme categories.** The plot was constructed from results obtained for the independent test set, and corresponds to the expected performance for novel, uncharacterized proteins. For a given category, e.g. *transport and binding* a sensitivity of 90% can be archived with false positive rate of 10% corresponding to 90% correct prediction on both categories. Random performance would correspond to a line along the diagonal.

relationship between the probability threshold used and the false positive rate. For a given probability threshold, the rate of false positives is essentially 1 minus the threshold. The best performance values are comparable to the upper limits estimated from the consistency in the assignment of the SWISS-PROT keywords (Devos and Valencia, 2000).

Relation to the linguistic analysis of SWISS-PROT entries

The classification into functional categories is based on linguistic analysis and clustering of SWISS-PROT keywords by the EUCLID method (Blaschke et al., 1999; Tamames et al., 1998). The EUCLID method computes scores for placing a protein into each of the 14 categories (Riley, 1993; Andrade et al., 1999b). The classification is not mutually exclusive, i.e. a protein sequence may have scores high enough to be placed into two or more categories.

In general we found a strong relation between the prediction quality of EUCLID and ProtFun. This should not come as a surprise considering that the quality of prediction from EUCLID determines the quality of the data set on which ProtFun was trained. Two of the worst categories are "cellular processes" and "central intermediary metabolism" which are both very loosely defined—especially the former in which many different functions ranging from cell division to chaperones and detoxification are included.

One noticeable exception from this rule is the class of regulatory proteins. Because the class of regulatory proteins tends to overlap other categories a lot, these proteins are hard to categorize correctly by EUCLID. However, this has not been a problem for ProtFun which allows a protein to belong to more than one category. Indeed regulatory proteins are one of the best predicted categories.

Functional characterization of the complete human genome

Using ProtFun it is possible to estimate the breakdown on functional categories of the entire human genome. Ideally a data set with all proteins encoded by the human genome should be used. As no final and highly reliable set is yet available, we have used the database of confirmed sequences made available by the Ensembl initiative (Birney et al., 2001). This database consists of ∼27,000 protein sequences from the human genome, all of which are supported by EST matches. One should be aware that this database is likely to have a bias towards highly expressed proteins. Using the predicted probability for each category, the number of proteins in each category was subsequently estimated by summing over the probability of the category in question for every protein (Figure 4.6). The functional breakdowns in the human genome publications (Int. Human Genome Sequencing Consortium, 2001; Venter et al., 2001) are based on function assigned by sequence similarity and are therefore based on approximately 50-70% of the genes (depending on the gene number). Direct comparison to what we predict is also made difficult by the fact that different classification systems are used in the two articles. The most striking difference from the Venter et al. (2001) paper is that we predict a much larger fraction of the proteins to be enzymes, while the distributions over enzyme subcategories agree quite well. Part of the

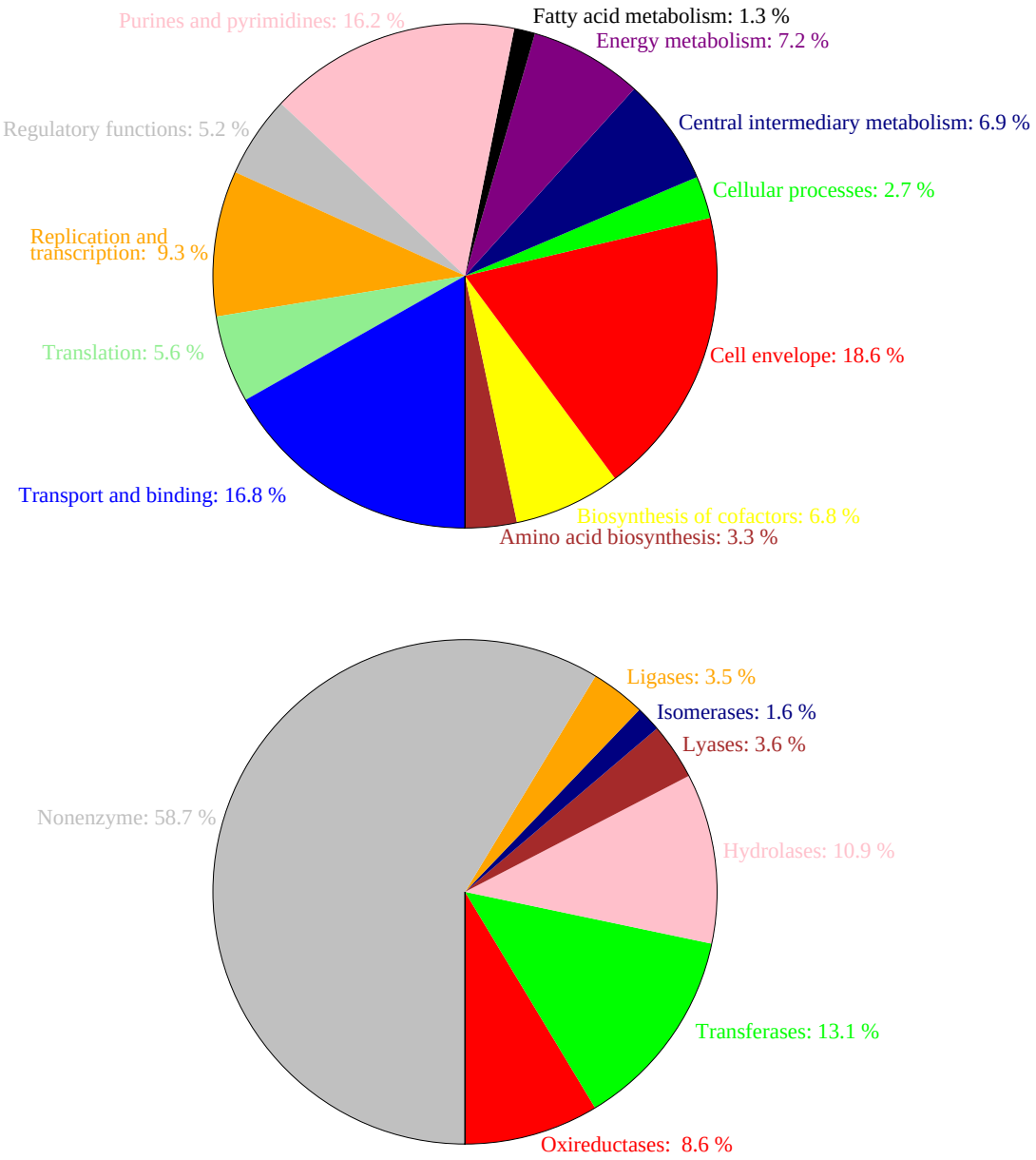


Figure 4.6: **Statistics for the human genome based on the Ensembl gene set (Birney et al., 2001).** From the probabilistic ProtFun output, the number of proteins belonging to a given category was estimated by summing over all 27,000 sequences. The most striking difference from the Venter et al. (2001) paper is that a larger fraction of the proteins is predicted to be enzymes, while the distributions over enzyme subcategories agree quite well.

explanation for the enzyme bias can be that the complete Ensembl data set (Birney et al., 2001) may have a bias towards highly expressed proteins. We also investigated the spread of functionally related proteins across the different chromosomes (data can be found at the ProtFun WWW site). Among several interesting observations (for example that endoplasmic and Golgi proteins are highly abundant at chromosome 6), was the fact that chromosome 11 seems to contain many uncharacterized proteins (belonging to *other categories*), which falls outside the classification used in this study.

Individual sequence prediction

The ProtFun method is perhaps best suited for obtaining functional hints for individual sequences for later use in assay selection and design. The first example shown here relates to the human prion sequence (ACC PRIO_HUMAN) which is being associated with the Creutzfelt-Jacobs syndrome. The functionality of this protein, which seems to produce no phenotype when knocked out in mice (Collinge et al., 1995), is still not fully understood. The ProtFun method predicts with high confidence that the human prion sequence belongs to the *transport and binding* category, and also that it is very unlikely to be an enzyme (Table 4.4). Indeed prions have been shown to be able to bind and transport copper while no catalytic activity has ever been observed (Brown et al., 1997; Brown, 2001). Interestingly, as the prion is a cell surface glyco-protein (expressed by neurons) it has a distinct pattern of post-translational modification, which most likely contain information which can be exploited by the system for functional inference. Incidentally, the cell envelope category is the third highest scoring category for the prion sequence. The purine/pyrimidine class is second, however no experimental evidence supports this functionality. Functional information was *not*

Table 4.4: ProtFun output for the human prion (PRIO_HUMAN P04156) and for an interacting pair of proteins, the amyloid A4 protein (A4_HUMAN P05067; P09000; Q16011) and transthyretin (TTHY_HUMAN P02766), which at the sequence level are entirely unrelated. Both of these proteins are with high confidence predicted to be cell envelope related as well as transport and binding proteins in agreement with the known functionality of these proteins.

	Prion	A4	TTHY
Amino acid biosynthesis	0.011	0.011	0.011
Biosynthesis of cofactors	0.041	0.161	0.034
Cell envelope	0.146	0.804	0.698
Cellular processes	0.027	0.027	0.051
Central intermediary metabolism	0.047	0.139	0.059
Energy metabolism	0.029	0.023	0.046
Fatty acid metabolism	0.017	0.017	0.023
Purines and pyrimidines	0.528	0.417	0.153
Regulatory functions	0.013	0.014	0.014
Replication and transcription	0.020	0.029	0.040
Translation	0.035	0.027	0.032
Transport and binding	0.831	0.827	0.812
Enzyme	0.233	0.367	0.227
Nonenzyme	0.767	0.633	0.773
Oxidoreductase (EC 1.-.-.-)	0.070	0.024	0.055
Transferase (EC 2.-.-.-)	0.031	0.208	0.037
Hydrolase (EC 3.-.-.-)	0.101	0.090	0.208
Lyase (EC 4.-.-.-)	0.020	0.020	0.020
Isomerase (EC 5.-.-.-)	0.010	0.010	0.010
Ligase (EC 6.-.-.-)	0.017	0.078	0.017

transferred by sequence similarity from the nearest neighbor, as the maximal similarity between the prion sequence and the data set (training and test) is 14.8% to proline-arginine-rich end leucine-rich repeat protein (PARG_HUMAN P02766). We believe that predictions like these are very useful when resolving protein function, because they can be used to generate specific hypotheses and direct laboratory experiments.

Using function prediction in conjunction with protein–protein interactions data

The method is also relevant for obtaining additional evidence on protein-protein interactions, where database information may contain many false negatives (as not all possible interactions have been screened). We did, as an example, predict the function of both sequences in all interaction partners found in the Database of Interacting Proteins (DIP) (Xenarios et al., 2001). If the functional categories of the interacting proteins are predicted to be the same for otherwise unrelated sequences, that should increase the likelihood of the prediction being correct (as well as the validity of the interaction). Others have successfully employed a similar approach based on subcellular localizations to lower the rate of false positives on yeast two-hybrid data (Schwikowski et al., 2000).

An interesting example of such an interacting protein-pair is the Alzheimer's disease amyloid A4 protein and transthyretin, which at the sequence level are entirely unrelated. Both of these proteins are with high confidence predicted to be *cell envelope* related as well as *transport and binding* proteins, see Table 4.4. Amyloid A4, a neural receptor, and transthyretin a thyroid hormone binding protein believed to transport thyroxine into the brain, have functionalities which are in full agreement with the prediction. The two sequences have a maximal similarity to the data set of 12.4% (to Citrate synthase (CISY_HUMAN O75390)).

When evaluating the functional category profiles for all interacting pairs in DIP (vs. non-interacting pairs) we found that interacting pairs indeed more often tend to have the same functional categorization (data not shown). However, while interacting non-enzymes in many cases will have the same functional role (belong to the same pathway), it may be more typical for enzyme-substrate pairs to belong to different categories.

Conclusion

The method presented here has the ability to transfer functional information between sequences which are far apart in sequence space. Not even the primary structures of the individual features (which are integrated by the method) need to be alike, or be related by evolution. The ProtFun method performs its non-linear classification in the feature space defined by 14 predicted and calculated attributes, which have been selected by the approach (out of more than 25 different attributes considered initially for discriminative value). The mapping between the space of all sequences and this feature space is also highly non-linear as very different sequences, by the individual feature predictors, may be converted to the

same patterns of, for example, posttranslational modification or secondary structure. This paper demonstrates that it is indeed possible to transfer functional information from the knowledgebase accumulated by experimental biology, even to proteins which are completely isolated in sequence space.

WWW server

Correspondence and requests for materials should be addressed to brunak@cbs.dtu.dk. The ProtFun method is made available online at the URL `http://www.cbs.dtu.dk/services/ProtFun/`.

Materials and methods

Data sets and functional class assignment

Classes of cellular function were defined after the 14 class classification originally proposed by M. Riley for the *E. coli* genome (Riley, 1993) and later extended by the TIGR group. The automatic class assignment to sequences was made by an extension of the EUCLID system performing linguistic analysis of SWISS-PROT keywords (Tamames et al., 1998). The system detects sequences similar to a query sequence by a BLAST search in the SWISS-PROT database and extracts common keywords from the entries. As we work with sequences from SWISS-PROT (with known function) we used the keywords directly and include no alignment step. For each functional class the informative weight (Z-score) of each keyword was extracted from a dictionary (Tamames et al., 1998). For each sequence a keyword sum leads to scores for the 14 classes.

The central point of the EUCLID system is the dictionary. The primary version of this dictionary was generated from an initial set of carefully, hand annotated proteins from different organisms spanning every kingdom of life. From this initial set, a first dictionary was defined which was used to assign all SWISS-PROT proteins and the process of dictionary definition and assignment was reiterated until convergence.

This final dictionary was used to assign functional classes to around 5,500 human proteins from SWISS-PROT. In the selection we omitted SWISS-PROT sequences representing fragments and hypothetical proteins and therefore the "high-quality" subset is smaller that the entire set of human proteins in this database.

The values obtained from the method were then compared to two thresholds: If the score for a category was above 3 it was considered a positive example while examples with a score below 0 were used as negative examples. Examples scoring between 0 and 3 were considered unclear and were thus not used. By labeling our data set this way we eliminate the most uncertain functional annotations thereby improving the quality of our data set. The composition of the data set obtained is shown in Table 4.5.

Table 4.5: **The number of sequences included in the data sets when training networks for the various categories.**

Category	Pos.	Neg.
Amino acid biosynthesis	85	3691
Biosynthesis of cofactors	240	2964
Cell envelope	173	2599
Cellular processes	259	3339
Central intermediary metabolism	216	3127
Energy metabolism.id	330	3310
Fatty acid metabolism	53	3846
Purines and pyrimidines	535	1649
Regulatory functions	586	3037
Replication and transcription	746	2591
Translation	174	3677
Transport and binding	1461	2126
Enzyme	1620	4038
Oxidoreductase	319	1213
Transferase	529	1003
Hydrolase	485	1047
Lyase	72	1460
Isomerase	49	1483
Ligase	78	1454

Enzyme class assignment

SWISS-PROT provides enzyme class information for most enzymes in the "DE" field. For those without an EC assignment, the suffix "ASE" and the presence (or absence) of the words "INHIBITOR" and "PRECURSOR" were additional considerations when assigning proteins into the categories "Enzyme", "Non-enzyme" or "Neither". The "Neither" category comprised ambiguous cases which were excluded from training. For the six enzyme classes only proteins with EC assignments were used. The negative set for each class contained enzymes assigned to other enzyme classes.

Similarity screening of test sets vs. training sets

To generate a training set A, and a test set B, in which the similarity between the two sets were minimal, the following heuristic algorithm was used: A similarity measure $D(a, b)$ between all pairs of sequences (a, b) in the original set was calculated using the Smith-Waterman score for the optimal local alignment between each sequence pair. A similarity measure $H(A, B)$ was defined as the sum of similarities between sequences in set A and sequences in set B. $H(A, B) = sum(D(a, b)|a \in A \wedge b \in B)$

The algorithm for generating the two sets A and B started by having all sequences in A. The algorithm selected a sequence $x \in A$ that maximized the

value $H(A,B) - H(A|x, B \cap x)$. New sequences were selected from A until set B had the desired size.

Feature prediction and encoding

A number of different prediction methods were used as input features for the method (see Figure 4.4). All the servers were run on all 5,500 sequences constituting the training and test sets. The output was parsed and the scores obtained from the different predication servers were normalized and/or converted into probabilities.

Encoding positional feature information (e.g. phosphorylation sites) for proteins of variable length is non-trivial. We tested a number of encoding schemes for the positional information in the input to the neural networks. One approach was to divide each sequence into a number of equally sized bins, where the bin size was dynamically calculated for each sequence. The most important disadvantage of this approach is that each bin does not represent the same number of residues for sequences of different length. An alternative method is to define a fixed number of bins of fixed size. Because the sequences have different lengths this can result in overlapping bins and redundant information in the encoding. For very long sequences the fixed size binning scheme gives rise to gaps between the bins in which case the method will fail to encode all the features fully. Finally a combined approach was tried with one fixed size bin in either of the sequence and dynamic bins to encode the rest of the sequence. Full details of the feature encoding can be found at the ProtFun WWW site (`http://www.cbs.dtu.dtk/services/ProtFun`).

Each of these encodings were tested with different number of bins on all features having positional information. The performance of each binning scheme was evaluated by training a neural network on each feature separately. For each feature the binning scheme that gave the highest test set correlation coefficients across the different functional categories on most categories was chosen.

Feature combinations and network ensembles

Optimal combination of parameters for each of the different categories were found using a boot-strap strategy. First, for every category a simple network with one fully connected hidden layer was trained on each separate feature. Details can be found elsewhere (Brunak et al., 1991), while information about specific network architectures can be found at the ProtFun www site (`http://www.cbs.dtu.dk/services/ProtFun`). Based on the test set performance of these networks we judged which features were potentially useful for prediction of at least one category. Networks were then trained for every pair of these features, to obtain information on the correlations between features. Many networks using increasing numbers of these features were then trained, and the best five were picked as an ensemble.

The output of these networks were subsequently transformed into probabilistic scores. Based on the predictions performed by each network on the test set, the network output distributions for positive ($f_{pos}(x)$) and negative examples

($f_{neg}(x)$) were estimated using Gaussian kernel density estimators on the output activities with no squashing function applied. From these density estimates and the number of the positive and negative examples (N_{pos} and N_{neg}) the probability that an example is positive ($P(x)$) can be calculated from the network output:

$$P(x) = \frac{N_{pos} f_{pos}(x)}{N_{pos} f_{pos}(x) + N_{neg} f_{neg}(x)}$$

To calculate the combined prediction of an ensemble of networks we simply take the average of their probabilistic predictions. It is these values that are reported by the method.

Chromosomal gene location

Chromosome locations for the human SWISS-PROT sequences were obtained by web-linking through SWISS-PROT references to the OMIM database (Online Mendelian Inheritance in Man) maintained at NCBI (`http://www3.ncbi.nlm.nih.gov/Omim/`). From OMIM, by further linking to LocusLink (`http://www.ncbi.nlm.nih.gov/LocusLink/`), one could obtain the chromosome number for the gene being considered. Not all proteins could be tracked down to their chromosome number in this fashion. For the remaining sequences, BLASTing them against the human genome database at NCBI revealed the chromosome number in most cases.

4.10 Caveats of ProtFun

4.10.1 Highly skewed distribution over cellular roles

The priors used for calculating the probability scores in the ProtFun and genome wide predictions both show a very uneven distribution over the different cellular role categories (see Paper II). In humans large numbers of proteins are believed and predicted to be involved in *transport and binding* and *regulatory functions* whereas a relatively small fraction of the proteins belong to the many categories related metabolism (Liu and Rost, 2001). This is a bit unfortunate since it means that *transport and binding* is a very broad category, thus predicting a protein to belong to this category is not very informative.

This is a consequence of the system originally having been designed for prokaryotes, more specifically for *E. coli* (Riley, 1993). For prokaryotes, in which a much larger fraction of the genes is devoted to metabolism, the distribution over cellular roles is much more even. The problems that EUCLID encounters when sorting human proteins into cellular roles may also in part be explained by the lack of a clear definitions of the categories for eukaryotes.

4.10.2 Labeling errors in the training data sets

Because of the sheer size of the data sets, the assignment of functional categories to human protein sequences had to be done in an automated fashion. As has already been described in Chapter 1, it is difficult to extract functional information from current databases where most of the information is present as free text.

The EUCLID method that we used for assignment of cellular roles thus relies only on the SWISS-PROT keywords, which are a controlled vocabulary. Unfortunately, these keywords often capture only part of the proteins function. It is thus unavoidable that EUCLID will make some misclassifications on this basis.

A clear example of this from our training set is the BCL2 protein, which is involved in the complex regulation of apoptosis. It should be assigned to the *regulatory functions* category, but is in fact labeled as a negative example for this category. Instead it is labeled as a positive example for both *biosynthesis of cofactors* and *purines and pyrimidines*. The prediction made by ProtFun are unfortunately consistent with the incorrect labeling.

It cannot be denied that these types of errors exist in the data set used for training the neural networks predicting cellular role. However, given that the categories are not clearly defined for eukaryotes, it is very difficult to assess how much higher the error rate is than what would be expected by manual curation. The accuracy of a manual curated data set would also depend on who did the curation.

4.10.3 Biologically meaningful prediction errors

In some cases neural networks simply cannot learn certain examples, in which case there is usually a good reason. One such example from our training set is Protein Z, which is involved in hemostasis by binding thrombin. Based on the

SWISS-PROT keywords, it has correctly been assigned as a *transport and binding* protein which has also been learned by ProtFun.

The interesting part is, that while Protein Z is also correctly labeled as a non-enzymatic, the neural networks insist on incorrectly predicting this training example to be an enzyme. In fact, Protein Z is homologous with vitamin K-dependent clotting factors which are enzymes. In the MEROPS database (Rawlings et al., 2002), Protein Z is classified as a non-enzymatic member of peptidase family S1A. It is thus for a good reason why Protein Z looks like an enzyme according to ProtFun: it was an enzyme but has lost its enzymatic activity.

4.11 Good but far from perfect

At this point it is clear that ProtFun works well for predicting both cellular roles and enzyme classes of human proteins. But although the predictions are good, even the predictions with the highest confidence are too uncertain to be used for annotation purposes. The predictions made by ProtFun should be considered qualified guesses—hints telling what it looks like. However, proteins are not always what they appear to be, which BCL2 is a good example of. The ProtFun predictions should therefore always be considered putative functions which can be used for guiding experimental work.

Chapter 5

A look into the black box

Having convinced ourselves that ProtFun indeed is capable of predicting protein function far better than should be expected from a random expectation leads to a new question: What rules have the neural networks learned, which allow them to predict the function of proteins they have never been shown before? Neural networks have often been accused of being "black boxes", indicating that it is not possible to understand how they work.

While it is certainly difficult to fully understand what a neural network has learned, one should keep in mind that a neural network is nothing but a mathematical function that is used to fit a given set of data—it is not black magic.

5.1 Known relations between protein properties and protein function

The first approach taken to interpret the neural networks is to look at which sequence derived features they make use of for predicting the different functional categories. It is comforting to realize that many of the correlations between protein features and protein function, discovered by our completely data driven approach, are in fact supported by current biological knowledge.

5.1.1 Protein structure and function

Differences between the secondary structure composition of enzymes and non-enzymes have previously been observed (Zhang and Zhang, 1999). Compared to other proteins, a strikingly different content of secondary structure elements have also been noted for proteases, which are subclass of hydrolases (Stawiski et al., 2000). Both of these studies rely on secondary structure derived from the three-dimensional protein structures in PDB.

There are two major differences between these two observations and our results. The first is that we work with predicted secondary structure, which allows us to analyze a much larger (and hopefully less biased) data set. The downside is that the predictions are not perfect. The second difference is that we take into account the positional information of the secondary structure—not only the overall content. This should to a limited extent allow us to find correlations between

protein fold classes and function.

We find that secondary structures play a quite important role for classification of enzymes, whereas it appears to be of much more limited use for predicting cellular roles. This observation is consistent, with the idea that structure is more closely related to molecular function than to cellular role.

In addition to secondary structure, we have also made use of transmembrane helix predictions, from which prediction of *replication and transcription* proteins benefits greatly. This is a clear illustration that ProtFun makes use of not only positive, but also negative evidence: very few if any transmembrane proteins are involved in *replication and transcription*. *Cell envelope* is the only other cellular role for which transmembrane helices play such an important role. Again this makes biological sense as a large fraction of cell envelope proteins span the cell membrane. One can in fact argue whether it is most reasonable to regard transmembrane helix prediction as a structural feature or a subcellular localization feature, as proteins with transmembrane helices are restricted to membranes.

5.1.2 Different compartments serve different functions

Features representing subcellular localization of proteins appear to constitute the overall most important single class of features. This is well supported by our knowledge of organization of the eukaryotic cell—the different compartments tend to serve different functions. Knowing (or predicting) where in the cell a protein is localized can therefore give valuable clues to its function (Chou and Elrod, 1999; Johnson, 2000; Bannai et al., 2002).

In the ProtFun method, we make use of several features for predicting subcellular localization. In addition to prediction of transmembrane helices, which has already been discussed, we make use of both predicted signal peptides, made by SignalP, and the more general subcellular localization predictions by PSORT.

PSORT predicts a large number of different compartments, but with low accuracy on particularly the minor localizations. Overall it gets 57% correct (on a non-balanced data set) which, although far from perfect, is still very useful (Nakai and Horton, 1999). The SignalP method predicts only two protein sorting categories: proteins that have a signal peptide and those that do not. While this gives a lot less information than PSORT, the advantage is that the quality the predictions made by SignalP is much higher than that of PSORT (Nielsen et al., 1999). The TargetP method included in later versions of ProtFun can be thought of as an intermediate between SignalP and PSORT, as it predicts more localizations than SignalP but fewer than PSORT (Emanuelsson et al., 2000).

While predictors that make use of sorting signals like N-terminal signal peptides are closest related to and give most insight into the biology underlying protein sorting, they are not without caveats. One objection has been raised by several groups predicting subcellular localization based on amino acid or dipeptide frequencies: given the accuracy of current gene finding algorithms, the N-terminal sequence can easily be incorrectly predicted, causing methods relying on signal peptide prediction to fail (Chou and Elrod, 1999; Hua and Sun, 2001). While this is a valid point, I still favor the more "biological" methods over composition based ones, in particular for protein function prediction purposes, where composition

derived features are likely better used directly.

Due to different chemical characteristics of the various cell compartments, differences are observed between the amino acid compositions of proteins from different compartments. Given that it is the surface of a protein that interacts with its environment, it is not surprising that the majority of such adaptations involve the exposed residues (Andrade et al., 1998). Because of the current methods for prediction exposed residues mainly rely on amino acid propensities, the predicted surface composition can just as well be encoded as the global composition. In our encoding, the protein surface composition would thus be represented by features such as the number of positive/negative residues.

Perhaps the best example of a compartment serving a particular function are the mitochondria. Although it does serve other purposes as well, the main purpose of mitochondria is oxidative phosphorylation. As a result of this, essentially all proteins imported into the mitochondria will be somehow involved in the energy metabolism. Although this is the case, the reverse is not true: proteins from the *energy metabolism* category also exist outside of the mitochondria. For which reason a perfect predictor of mitochondrial proteins would not suffice for making a perfect predictor of *energy metabolism* proteins.

There is thus firm support for using protein subcellular localization predictors for predicting the function of proteins. The usefulness for function prediction was actually a large part of the motivation behind the development of methods for predicting protein localization (Chou and Elrod, 1999; Bannai et al., 2002).

5.1.3 Protein lifetime and protein degradation

As was mentioned in Chapter 4, the degradation of many proteins is a highly regulated process that takes place through several different pathways. Degradation mediated by the ubiquitin–proteasome system is the best understood pathway for regulated degradation of proteins. Proteins are tagged for destruction by becoming polyubiquitinated and are then degraded by the huge proteasome complex. Proteins can also become monoubiquitinated, although this does not result in the protein being degraded. Instead it appears to be involved in membrane trafficking at least in yeast (Pole et al., 2002).

One likely mechanism for initiating ubiquitin mediated degradation of proteins is phosphorylation of PEST regions (Nakai, 2001). PEST regions are as described earlier regions that are rich in proline, glutamic acid, serine and threonine residues. Such regions contain many possible phosphorylation sites, and there is evidence that once phosphorylated, these regions act as recognition signals for ubiquitin ligases that target the protein for subsequent degradation.

Consistent with this theory, PEST regions are often found in regulatory proteins and other proteins with short life spans (Nakai, 2001). This trend was picked up by ProtFun, which uses PEST regions as one of the most important features for the prediction of regulatory proteins.

5.1.4 Phosphorylation

As mentioned earlier, reversible phosphorylation of serines, threonines and tyrosines constitute possibly the most important regulatory mechanism, one which is used for regulation of essentially all cellular processes (Cohen, 2000). This provides a well supported biological explanation for the predictor of *regulatory functions* to make use both predicted serine/threonine and tyrosine phosphorylation sites.

The biological relation between PEST regions an serine/threonine phosphorylation also seems to be reflected in the ProtFun feature usage, as all predictors that include serine/threonine phosphorylation sites also include PEST regions.

5.1.5 Glycosylation

The by far most varied type of covalent modification of proteins is glycosylation which happens either co- or post-translationally. Post-translational glycosylations can play much the same kind of regulatory purposes as phosphorylation. In fact one type of glycosylation, O-β-GlcNAc, is known in some cases to compete with serine/threonine-phosphorylation for the same residues (so-called Yin-Yang sites), allowing the two modifications to regulate protein function in a reciprocal fashion.

Unlike O-β-GlcNAc glycosylation which takes place in the cytoplasm, most other forms of glycosylation take place in either the endoplasmic reticulum or the Golgi. The two forms of glycosylation that we make use of in ProtFun are both this type of "permanent" glycosylation and mainly target secreted and membrane associated proteins. From a function prediction point of view they can thus be thought of as complementing the subcellular localization predictions described above. For a longer discussion of the relations between glycosylation and protein function, refer to Paper III.

Paper III

5.2 Orphan protein function and its relation to glycosylation

Ramneek Gupta, Lars Juhl Jensen, and Søren Brunak*

Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby, Denmark.

* To whom correspondence should be addressed (email: brunak@cbs.dtu.dk)

Since the first bacterial genomes were completely sequenced, the surge in genome sequence data has overwhelmed the scientific community's efforts of elucidating protein function. Computational methods have made it possible to work with sequences from complete genomes and proteomes, and inference of protein function by exploiting direct sequence similarity indeed goes a long way in describing a proteome's functional capacity. However, at least 40% of the gene products in newly sequenced genomes typically remain uncharacterized. Proteins without an annotated function are also known as *orphan* proteins since they do not belong to a functionally characterized protein family. Many sequences must therefore be compared using their features rather than by direct comparison in the conventional sequence space. Here we focus on one such feature—glycosylation—which is common in eukaryotic proteomes.

Conventional assignment of protein function

Proteins can be characterized in different ways such as their cellular role (the biological process they are involved in, e.g. transcription), their molecular function (e.g. ion transporter) or the cell cycle phase they are involved in. *Protein function* can thus be interpreted in various ways, but the "cellular role" descriptor has traditionally been popular in a number of genome sequencing projects.

Since the majority of housekeeping proteins are similar amongst different organisms, it is convenient to use accumulated experimental knowledge to accelerate the identification of new protein sequences. Structural and functional annotations can be transferred from a sequence sufficiently similar to the query sequence. This

process, also known as "transfer by homology", has assigned function to most gene products in newly sequenced genomes so far (Bork et al., 1998; Attwood, 2000; Iliopoulos et al., 2000; Eisenberg et al., 2000). Sequence similarity across complete genomes has also been used to construct protein families (Tatusov et al., 1997). The traditional paradigm that sequence determines structure which in turn determines function, is still the most widespread technique for assigning function. However:

- Proteins similar in sequence are not always analogous in function (Devos and Valencia, 2000). Indeed, similar sequence need not even imply similar protein structure.
- Transferring function from structural homologues is hampered by the slow growth in the amount of new folds in structural protein databases. As of early 2001, while the SWISS-PROT sequence database contains over 92,000 sequences, the highly redundant structural database PDB contains a little over 14,000 structures in all. SWISS-PROT itself contains only a fifth of the estimated proteins coded by the human genome using the widespread estimate of around 40,000 genes.
- It still leaves a large fraction of unidentified proteins in a genome (Iliopoulos et al., 2000; Rubin et al., 2000). More than 40-50% of proteins in the eukaryotic genomes sequenced so far (*Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*) bear no direct sequence similarity to proteins in any other proteome (Rubin et al., 2000; Initiative, 2000). Using GeneQuiz (Casari et al., 1996) to derive automated functional annotation by homology, in a study over 31 complete genome sequences, functions could be predicted for 62% of the proteins on average (Iliopoulos et al., 2000).

Non-homology based methods

Attempts have been made recently, to minimize the use of sequence similarity in deducing function (Marcotte, 2000). *Phylogenetic profiles* may be used to assign protein function by comparative genome analysis (Pellegrini et al., 1999). Another method, the *domain fusion method* (Marcotte et al., 1999a; Enright et al., 1999), also locates functionally related proteins, by analyzing patterns of domain fusion. The fused protein may then reveal functional aspects of its components. A similar method is one in which related proteins can be identified by being neighbors on a chromosome (Tamames et al., 1997; Dandekar et al., 1998; Overbeek et al., 1999). This is analogous to operon models in prokaryotic systems in which neighboring genes are involved in collective metabolic regulation.

Yet another way of linking functionally related proteins across genomes, is to associate a collection of genes to a phenotype (Huynen et al., 1998). The collection is enriched from organisms that share the same phenotype, and filtered for genes which occur in organisms which do not exhibit the phenotype. This selective enrichment, or *differential genome analysis*, can also be used for attributing genes to a functional cluster.

All these methods are however, not strictly non-homologous or of the *ab initio* type, since they rely to some degree on the presence of similar sequences, or on a protein having a functional partner with known attributes. Performance of these methods will increase with availability of more characterized protein sequences or, as is the case for phylogenetic profiles, with more complete proteomes.

One promising approach is the use of gene expression data obtained from DNA chip (or *microarray*) technology. Many coregulated genes have been found to be functionally related (Eisen et al., 1998), which forms an important assumption to a lot of microarray research. However, the degree and type of functional relation is still an open question. As remarked in a recent article (Heyer et al., 1999), functionally related genes need not be coexpressed, such as DNA repair genes responding to different types of damage. Conversely, coregulated genes need not be functionally related either since coregulation could just happen by chance if, e.g. gene products are needed at the same phase of the cell cycle. This also leads to the question of functional classification and clustering different gene products into a single biologically meaningful category.

A recent application of support vector machines to cluster gene expression data (Brown et al., 2000) could classify five functional classes with reasonable success. There was pre-evidence that the gene products clustered into these classes (Eisen et al., 1998). However, considering that only five classes could be predicted from over 200 functional classes (the MIPS categorization[1] was used), the problem of protein classification and predicting classes from microarray data remains a challenge.

Sequence-based identification in feature space

Our attempt at protein function prediction has been to use features inherent to the protein sequence. The general idea is to predict the cellular role using calculated global features such as molecular weight, sequence length, isoelectric point, etc. as well as more indirect features such as the predicted presence of potential glycosylation sites and of phosphorylation sites. This approach relies on the fact that the sequence of a protein contains many signals and properties relevant to processing by the subcellular machinery. Since all proteins in a cell are subject to the same subcellular environment, proteins with similar properties are likely to be processed and modified in a similar fashion (Jensen et al., 2002).

Apart from using these features ourselves, our main focus has been on sequence signals governing post-translational modifications of proteins. Proteins, once synthesized in a cell, are subject to many types of post-translational modifications which influence protein function. Among several modifications (e.g. phosphorylation, glycosylation, methylation, acetylation), some may be more complex than others and attribute a range of functional and structural properties to the protein's role in the cell.

Most post-translational modifications occur on well-defined residues in a protein, but usually without a consensus sequence. Such sequence signals (around the acceptor sites) can be predicted with reasonable accuracy using methods such

[1] `http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/index.html`

as artificial neural networks (Nielsen et al., 1997b; Hansen et al., 1998; Blom et al., 1999; Gupta et al., 1999b) and hidden Markov models (Sonnhammer et al., 1998; Nielsen and Krogh, 1998).

In our work on the ProtFun method (Jensen et al., 2002), the use of post-translational modifications (PTMs) appeared essential for the prediction of several functional classes. For instance, in the prediction of proteins with "regulatory function", the most important features were phosphorylation as well as glycosylation. Reversible phosphorylation is a well known and widely used regulatory mechanism (Cohen, 2000) and there is growing evidence that the O-β-GlcNAc glycosylation plays a reciprocal role with reversible phosphorylation (Comer and Hart, 2000). Other features useful in predicting regulatory proteins were predicted subcellular location and PEST regions (rich in proline, glutamic acid, serine, and threonine residues), where the latter is known to target proteins for degradation (Rechsteiner and Rogers, 1996).

PTMs are also very important for correct assignment of proteins related to the cell envelope, replication and transcription. The fact that (predicted) PTMs correlate strongly with the functional categories fits well with biological knowledge. For example, predicted N-glycosylation sites turn out to be important for prediction of cell envelope proteins. It has been shown experimentally that removal of carbohydrates linked to asparagines, from a protein normally targeted for the cell envelope, retains it in the endoplasmic reticulum (Chen and Colley, 2000).

Contribution of glycosylation to protein function

Prediction of glycosylation sites

The addition of a carbohydrate moiety to the side-chain of a residue in a protein chain influences the physicochemical properties of the protein. Glycosylation is known to affect proteolytic resistance, intracellular targeting, cell-cell interactions, protein regulation, solubility, stability, local structure, lifetime in circulation and immunogenicity (Lis and Sharon, 1993; Varki, 1993; Hounsell et al., 1996; van den Steen et al., 1998; Comer and Hart, 1999).

Of the various forms of protein glycosylation found in eukaryotic systems, the most important types are N-linked, O-linked GalNAc (mucin-type) and O-β-linked GlcNAc (intracellular/nuclear) glycosylation. N-linked glycosylation is a co-translational process involving the transfer of the precursor oligosaccharide, $GlcNAc_2Man_9Glc_3$, to asparagine residues in the protein chain. The asparagine usually occurs in a sequon Asn-Xaa-Ser/Thr, where Xaa is not Proline. This is however, not a specific consensus since not all such sequons are modified in the cell. O-linked glycosylation involves the post-translational transfer of an oligosaccharide to a serine or threonine residue. In this case, there is no well-defined motif for the acceptor site other than the near vicinity of proline and valine residues.

The biological roles of oligosaccharides on proteins are rather diverse (Varki, 1993; Kukuruzinska and Lennon, 1998; van den Steen et al., 1998). N-linked and

O-linked GalNAc glycosylation occur in the endoplasmic reticulum and Golgi apparatus, and thus modify proteins that go through the secretory pathway (secreted and membrane proteins). Glycosylation in these cases, lends structural stability and contributes to binding and immunogenic properties. In contrast, O-β-GlcNAc is a dynamic modification that occurs on cytoplasmic and nuclear proteins, and is known to play a regulatory or signaling role (Comer and Hart, 1999, 2000; Hanover, 2001).

Experimental determination of glycosylation sites is difficult to achieve as large amounts of purified protein are needed for the analysis of glycosylation sites. In addition, glycosylation can be an organism- and tissue specific event. Therefore only a few glycoproteins have been characterized so far as reflected in the low percentage of glycoprotein entries in SWISS-PROT (approx. 10% of human proteins, *see also* (Apweiler et al., 1999)). This motivates the need for developing theoretical means of predicting the glycosylation potential of sequons.

Methods for predicting glycosylation sites for the above three types have been developed[2] using artificial neural networks that examine correlations in the local sequence context and surface accessibility. These predictions were used as features for protein function prediction in the ProtFun method outlined above. In the following section, predicted glycosylation site information on human proteins is used to illustrate the contribution of glycosylation to protein function and assess how widespread this modification is across the human proteome.

N-Glycosylation

N-linked glycosylation modifies membrane and secreted proteins. This co-translational process occurs in the endoplasmic reticulum and is known to influence protein folding. The modification attributes various functional properties to a protein. To examine if certain categories of proteins were more prone to glycosylation than others, we studied the spread of known glycosylation sites across different categories.

In our data set of approximately 5,500 human proteins, only 189 proteins (at 453 *confirmed* sites) were annotated in SWISS-PROT as N-glycosylated (not considering proteins with only POTENTIAL or PROBABLE sites). Figure 5.1 illustrates the spread of human glycosylation sites along the protein chain and across predicted subcellular locations and keyword based assignment of cellular role categories (Jensen et al., 2002). Relative positions of sites on proteins were calculated with respect to normalized sequence lengths. To construct this plot, sequence lengths were normalized, and relative position expressed on a percent (0-100) scale. Glycosylation sites were binned (10 bins across each sequence), and their frequency plotted across different categories.

N-glycosylated proteins appeared to almost exclusively belong to the functional category, "Transport and binding". This may not be too surprising considering that this category consists largely of membrane and secreted proteins. The few proteins not belonging here were mostly involved in central intermediary metabolism. Subcellularly, extracellular proteins were the most favored and

[2]Glycosylation site prediction methods are available online—`http://www.cbs.dtu.dk/services/`

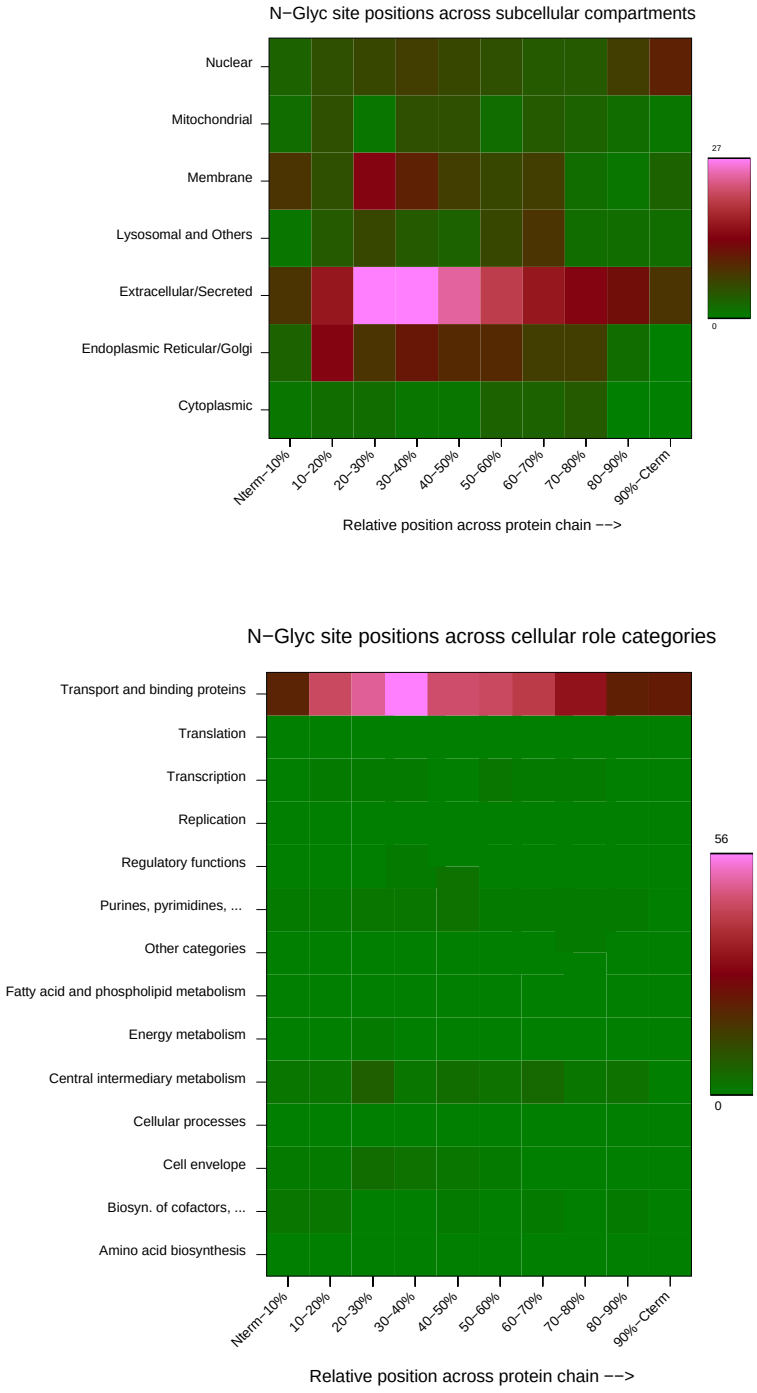


Figure 5.1: **Categorical distribution of known N-glycosylation sites across the protein chain.** Grey-scale indicates frequency of sites (light to dark in increasing order). Protein chains, normalized in length, are represented across the x-axis from N-terminal to C-terminal (divided into tenths). Subcellular locations (top) were predicted using PSORT, and cellular role classification (bottom) by lexical analysis of SWISS-PROT keywords (Jensen et al. 2001). Most N-glycosylation sites were clustered in the first half of all protein chains, and mainly occurred in extracellular transport and binding proteins.

others occurred in membrane proteins and in the endoplasmic reticulum or Golgi.

A clear positional preference for glycosylation sites on protein chains was apparent. The terminal ends of proteins seemed unfavorable and most sites seemed to occur N-terminal to the center of the protein chain (20 to 40% along the length from the N-terminal start). The frequency of sites smoothly tapered off on both ends from this peak with a longer C-terminal tail. This statistical observation agrees with experimental indications of glycosylation sites being at least 12-14 residues distant from the N-terminal end and 60 residues away from the C-terminal end of a protein chain (Nilsson and von Heijne, 1993, 2000). One peculiar observation from Figure 5.1 was the appearance of glycosylated sites in the C-terminal region of nuclear proteins. On examination, these turned out to be in around 10 proteins which were indeed annotated to be N-glycosylated in the C-terminal. The sub-cellular location, however, appeared to be mis-annotated by PSORT. For instance, some secreted proteins among these were the Vasopressin-Neurophysin 2-Copeptin precursor, Von Willebrand Factor Precursor and Immunoglobulin Delta Chain C.

O-linked GalNAc Glycosylation

The addition of GalNAc linked to serine or threonine residues of secreted and cell surface proteins, and further addition of Gal/GalNAc/GlcNAc residues (Hounsell et al., 1996), is also known as mucin type glycosylation and is catalyzed by a family of GalNAc-transferases (UDP-N-acetylgalactosamine: polypeptide N-acetylgalactosaminyltransferases). The modification, a post-folding event, takes place in the cis-Golgi compartment (Roth et al., 1994) after N-glycosylation and folding of the protein, and affects secreted and membrane bound proteins.

There is no acceptor motif defined for O-linked glycosylation. The only common characteristic among most O-glycosylation sites is that they occur on serine and threonine residues in close vicinity to proline residues, and that the acceptor site is usually in a beta-conformation. A prediction method (Hansen et al., 1995, 1998) for this type of glycosylation on mammalian proteins has been built earlier and made available as a web server[3]. A database of O-glycosylated sequences is also available[4] and was used in constructing the O-glycosylation site prediction methods (Gupta et al., 1999a).

Figure 5.2 shows the spread of predicted glycosylation sites (O-GalNAc, mucin-type) across different categories and across the protein chain (a similar binning was used as in the N-glycosylation case). Sites tend to cluster towards the C- and N-termini of proteins for some categories. This figure also shows that O-glycosylation acceptor sites occur in a wide range of proteins, though glycosylation patterns (frequency, positions across chain) may differ for different types of proteins.

[3] `http://www.cbs.dtu.dk/services/NetOGlyc/`

[4] `http://www.cbs.dtu.dk/databases/OGLYCBASE/`

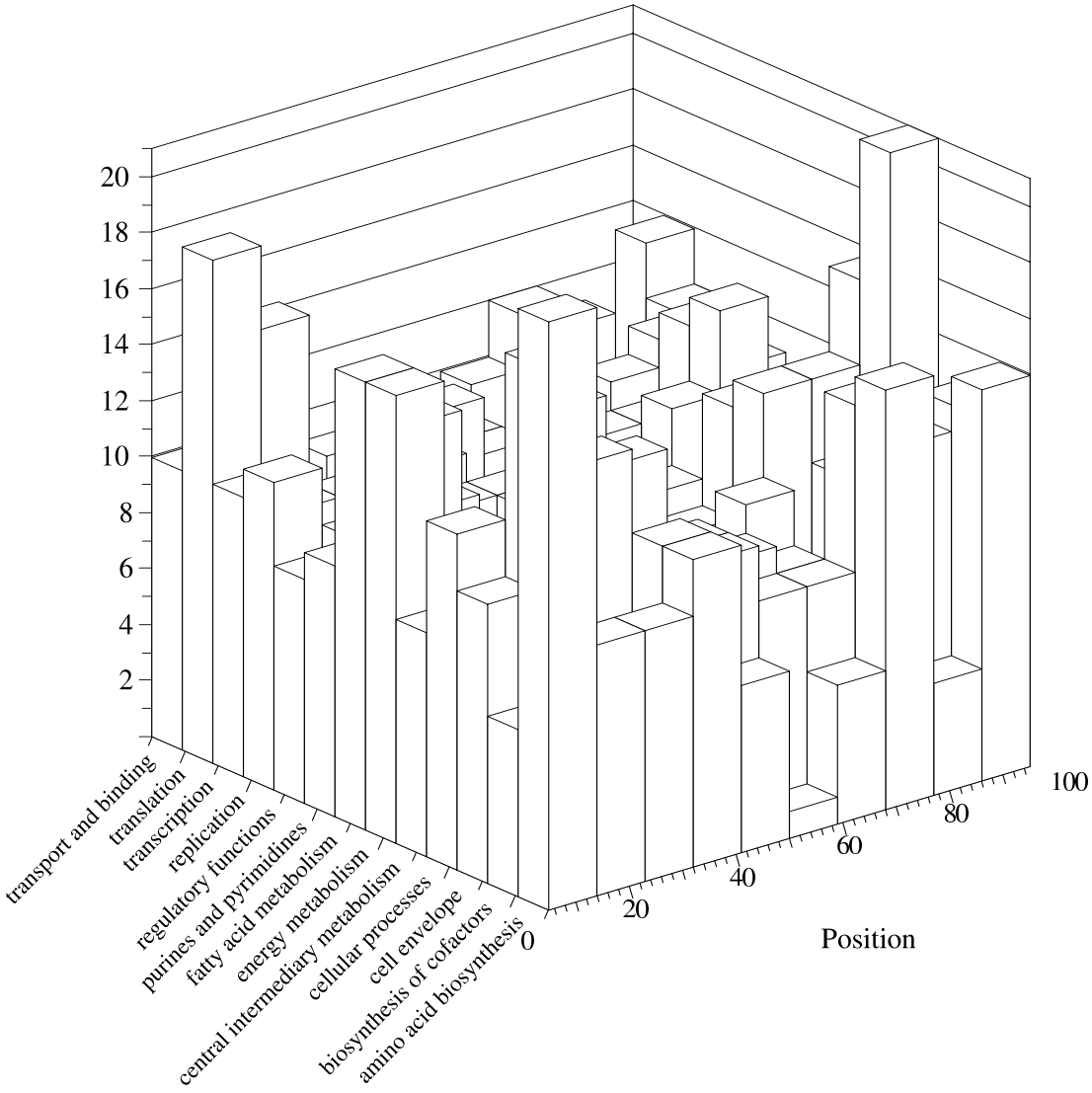


Figure 5.2: **Positional O-GalNAc glycosylation.** O-GalNAc (mucin type) glycosylation displays preference for position across a protein chain which could be significant across different categories. The *Position* axis reflects normalized protein chain length from N-terminal (0 on the axis) to C-terminal (100). The height of the bars indicates the number of predicted O-GalNAc sites (in ∼5,500 human proteins) for a particular category in a particular *position bin*.

O-β-linked GlcNAc Glycosylation

Glycosylation of cytosolic and nuclear proteins by single *N*-acetylglucosamine (GlcNAc) monosaccharides is known to be highly dynamic and occurs on proteins with wide-ranging functions and cellular roles (Hart et al., 1995; Snow and Hart, 1998). *N*-acetylglucosamine, donated by the nucleotide precursor UDP-*N*-acetylglucosamine, is attached in a beta-anomeric linkage to the hydroxyl group of serine or threonine residues. The attachment is, in short, known as O-β-GlcNAc[5] and the modification process as O-β-GlcNAcylation.

So far, all proteins with O-β-GlcNAc linked residues, are also known to be

[5]as opposed to the O-α-GlcNAc modification that has been found on certain membrane and secreted proteins of e.g. *Dictyostelium discoideum*

phosphorylated. Evidence suggests that at least in some cases, these two post-translational modification events may share a reciprocal relationship (Hart et al., 1995; Comer and Hart, 2000). This peculiar behavior strongly suggests a regulatory role for this modification. Sites which can be both glycosylated and alternatively phosphorylated are also known as "yin-yang" sites (Hart et al., 1995).

The acceptor site for O-β-GlcNAc glycosylation does not display a definite consensus sequence, nor are there many annotated sites in public databases. However, the fuzzy motif is marked by the close proximity of proline and valine residues, a downstream tract of Serines and an absence of Leucine and Glutamine residues in the near vicinity (data not shown). Out of approximately 5,500 human sequences from SWISS-PROT (rel. 38), over 4,600 had at least one predicted O-β-GlcNAc site. 1,535 of these proteins had at least one high scoring O-β-GlcNAc site prediction (with 3,154 high scoring Ser/Thr sites). A number of these were DNA-binding proteins and involved in transcriptional regulation. When ranked according to scores, a large fraction at the top of this list were found to be nuclear proteins (as annotated in SWISS-PROT). The O-β-GlcNAc transferase itself (P100 subunit) was found to have predicted O-β-GlcNAc sites.

While the O-β-GlcNAc modification seems to potentially affect almost all types of proteins, most O-β-GlcNAcylated proteins were either regulatory proteins or "transport and binding" proteins. A large fraction of unclassified proteins ("unknown" in role categories) were also predicted to contain this modification. Over half of all nuclear proteins contained a high ranking O-β-GlcNAc modified site.

The number of potential O-β-GlcNAc sites in proteins was studied with respect to function and cellular location. Figure 5.3 illustrates the number of predicted (high-scoring) sites per 100 Ser/Thr residues (per protein). Proteins with 1-2 predicted GlcNAc sites (per 100 Ser/Thr) were predominantly nuclear, cytoplasmic or membrane proteins. Nuclear and cytoplasmic proteins carried the highest densities of sites, a few cytoplasmic proteins having as many as 50 high-scoring O-β-GlcNAc sites among 100 Ser/Thr residues. With respect to cellular roles, proteins belonging to the category "Purines, pyrimidines, nucleosides and nucleotides" contained well spaced out sites (only a few sites among 100 Ser/Thr residues). Proteins with a wider distribution of sites included regulatory, transcription, replication, transport and binding, cell envelope and the "unknown" category proteins. The highest density of sites (30-40 per 100 Ser/Thr) was found in transcription and regulatory proteins, though some "unknown" proteins had over 40 sites (per 100 Ser/Thr). In general, the intracellular O-β-GlcNAc modification does not seem to cluster among close residues or display any characteristic spacing as was observed in another study of O-α-GlcNAc modifications affecting surface and membrane proteins of *Dictyostelium discoideum* (Gupta et al., 1999b).

Human proteome-wide scans revealed that the O-β-GlcNAc acceptor pattern occurs across a wide range of functional categories and subcellular compartments. For humans, the most populated functional categories were regulatory proteins and transport and binding proteins. Nuclear and cytoplasmic proteins were prominent, though membrane and secreted proteins were surprisingly also in high numbers. It is interesting to observe that acceptor patterns exist on

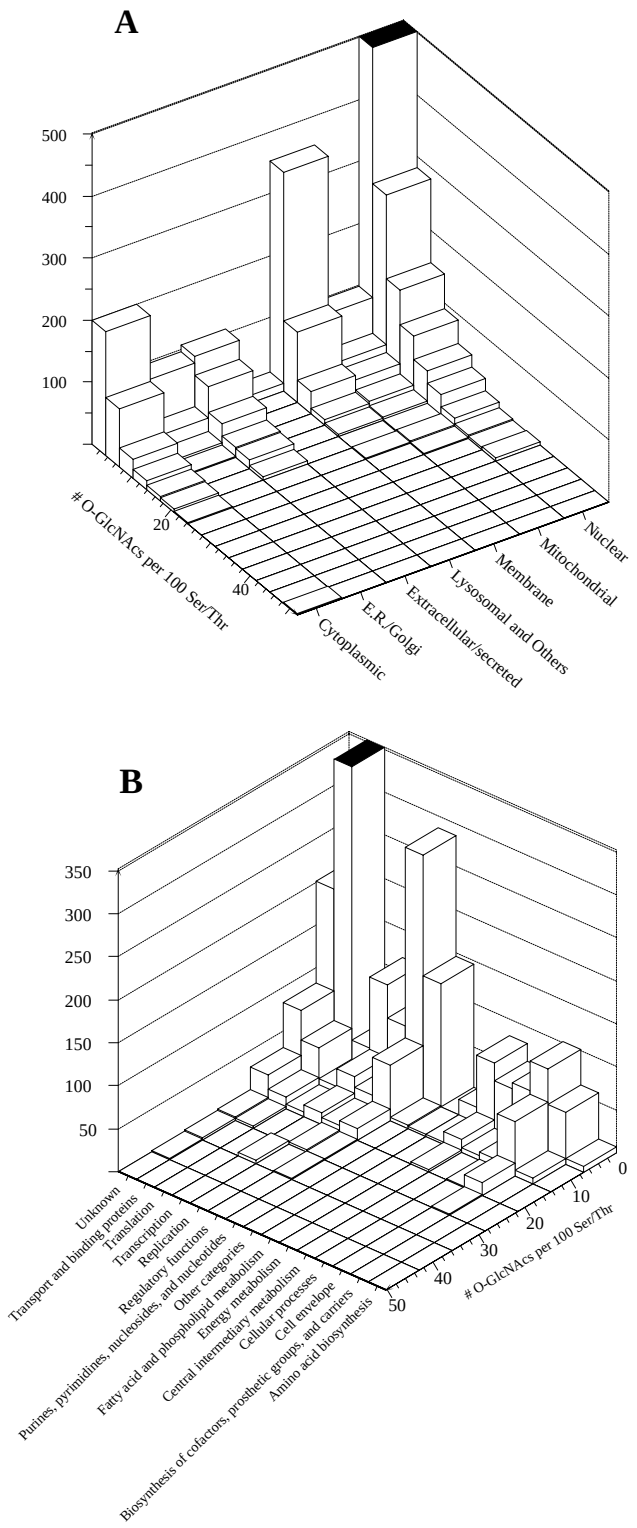


Figure 5.3: **Number of predicted O-β-GlcNAc sites per 100 Ser/Thr, in different categories of human proteins.** (A) shows proteins in different subcellular locations and (B) indicates cellular role categories. The z-scale (0-500 in A or 0-350 in B) is a frequency count for a particular bin; e.g. 0-2 O-β-GlcNAcs (per 100 Ser/Thr) occur most frequently for nuclear proteins in (A). These modifications usually do not occur in clusters. Although potential acceptor sites are largely found in nuclear/cytoplasmic proteins (usually regulatory), they also surprisingly occur in membrane proteins (mostly transport and binding proteins).

these proteins too, but the cellular machinery defines protein targeting and consequently influences their modifications. The prediction server guards against this possibility by generating a warning when a potential signal peptide is detected by SignalP[6].

PEST regions, rich in the amino acids Proline (P), Glutamic acid (E), Serine (S) and Threonine (T), are hypothesized to be degradative signals for constitutive of conditional protein degradation (Rechsteiner and Rogers, 1996). Phosphorylation, a common mechanism to activate the PEST-mediated degradation pathway, may be signaled by deglycosylation in the same region. Our scans revealed that a small fraction of O-β-GlcNAc sites appeared in PEST regions. Such sites were mostly found in proteins involved in regulatory functions.

Perspective

Glycosylation is clearly a modification affecting a wide range of proteins, and is now known to affect both intracellular and secreted proteins. Different types of glycosylation have varying site preferences on proteins, and occur in different patterns across the protein chain. Information for determining protein function lies not only in the presence or absence of glycosylation, but also in the glycosylation type and occurrence across the protein chain. This type of feature has a high degree of *discriminatory* value and combinations of other features cannot replace this class of modifications.

Predicting protein function remains a challenging task in bioinformatics. The approach outlined here uses a collection of sequence derivable features to predict the cellular role of a protein. The information rich features used in this work included a limited set of predicted post-translational modifications. Performance is likely to improve with the use of other correlated features. Some immediate features which can be tried out include GPI-anchor prediction (Eisenhaber et al., 1999) and motifs for other post-translational modifications such as the W-X-X-W C-Mannosylation motif (Krieg et al., 1998). Another feature worth considering is the "N-end rule" which relates the metabolic stability of a protein to the identity of its N-terminal residue (Varshavsky, 1996). In other words, the N-terminal residue of a protein can determine the *in vivo* half-life of a protein. Other possible features include dipeptide composition, nucleotide composition (of the genes coding for the proteins concerned), codon bias, and globularity of the protein.

The action of a protein in an organism can be described at different levels. A limitation in working with human proteins has been the absence of a well characterized functional classification scheme. The functional classification used in the approach explained above was reflective of the cellular role of the protein. This was a 14-category classification adopted from earlier work on *E. coli* proteins (Riley, 1993). Because of the diverse properties of a protein, matching multiple protein features to a functional classification is not trivial. However, with the availability of the human genome sequence and the development of comparative genomics, a biological vocabulary for protein function is called for.

[6] `http://www.cbs.dtu.dk/services/SignalP/`

One controlled vocabulary being developed is to be found in the gene ontology project (Ashburner et al., 2000) which defines a protein in context of the *biological process* it is involved in, the *molecular function* it carries out and what *cellular component* it constitutes. The attributes in the vocabulary are represented as directed acyclic graphs (DAGs) or networks. This is similar to a hierarchy with multiple sublevels for a higher level parent, except that a DAG allows a child to have more than one parent. Such ontologies seem promising for further work and narrowing down the specific roles of a protein.

Most traditional approaches for *in silico* characterization of proteins rely on protein function being determined by its structure or homology to already functionally annotated proteins. As shown here, protein sequences, on their own, contain a wealth of information concerning protein properties and can give vital clues to their molecular function and cellular role. This helps predicting the function of even those proteins for which no sequence homologues are to be found in databases.

It is worthwhile at this juncture to understand the importance of post-translational modifications and develop tools for predicting modified sites. This is essential information for deciphering protein function and characterizing complete proteomes.

Acknowledgments

We thank Alfonso Valencia's group for protein category data and Vera van Noort for useful comments with the text. The Danish National Research Foundation is acknowledged for financial support.

5.3 Evolutionary conservation of protein properties

If the protein features that we use for function prediction determine protein function, they should be subject to evolutionary pressure. One would thus expect that mutations changing these features would be selected against in proteins where the function is to be retained. As function is more often conserved among orthologous proteins than paralogous proteins, the features should be expected to be most conserved for orthologs.

5.3.1 Finding orthologs and paralogs between *H. sapiens* and *D. melanogaster*

Given only two homologous sequences from two different organisms, it is impossible to tell if they are orthologs or paralogs. If, however, one has access to the complete genomes of two organisms, it is possible to discern pairs of orthologous proteins from pairs of paralogous proteins. If only one copy of the gene exists in each genome, it is usually safe to assume that the two are orthologs (although theoretically they could be paralogs). In the case where more copies exist, it is possible to make a good guess at the phylogenetic relationships between the genes by analyzing the pairwise similarities, both within and between the genomes.

13,562 pairs of orthologous proteins between a set of 23,740 human protein sequences from Ensembl (Hubbard et al., 2002) and a set of 14,334 *Drosophila melanogaster* protein sequences from FlyBase (The FlyBase Consortium, 2002) were identified using the INPARANOID tool (Remm et al., 2001). It utilizes BLAST (Altschul et al., 1997) to find homologous pairs of sequences, further requirering that sequences match over more than half of their length to avoid matches to individual domains. Based on all such matches, the INPARANOID method identifies which pairs of proteins are most likely to be orthologs, and the remainder were assumed to be paralogs

5.3.2 Distances in feature space

Sequence derived features used by ProtFun were calculated for all sequences in the data set. The features were encoded and normalized as described in section 4.7, and the Euclidian distance between the members of each orthologous or paralogous pair was calculated for each feature.

The Euclidian distances were plotted as functions of the sequence identity within the pairs, as feature similarity can be expected to correlate with sequence similarity. Figure 5.4 reveals that most of the features used by ProtFun are more conserved (i.e. give smaller Euclidian distances) within pairs of orthologs compared to pairs of paralogs. As this is the case independent of the sequence identity, it indicates an evolutionary pressure for conserving the features.

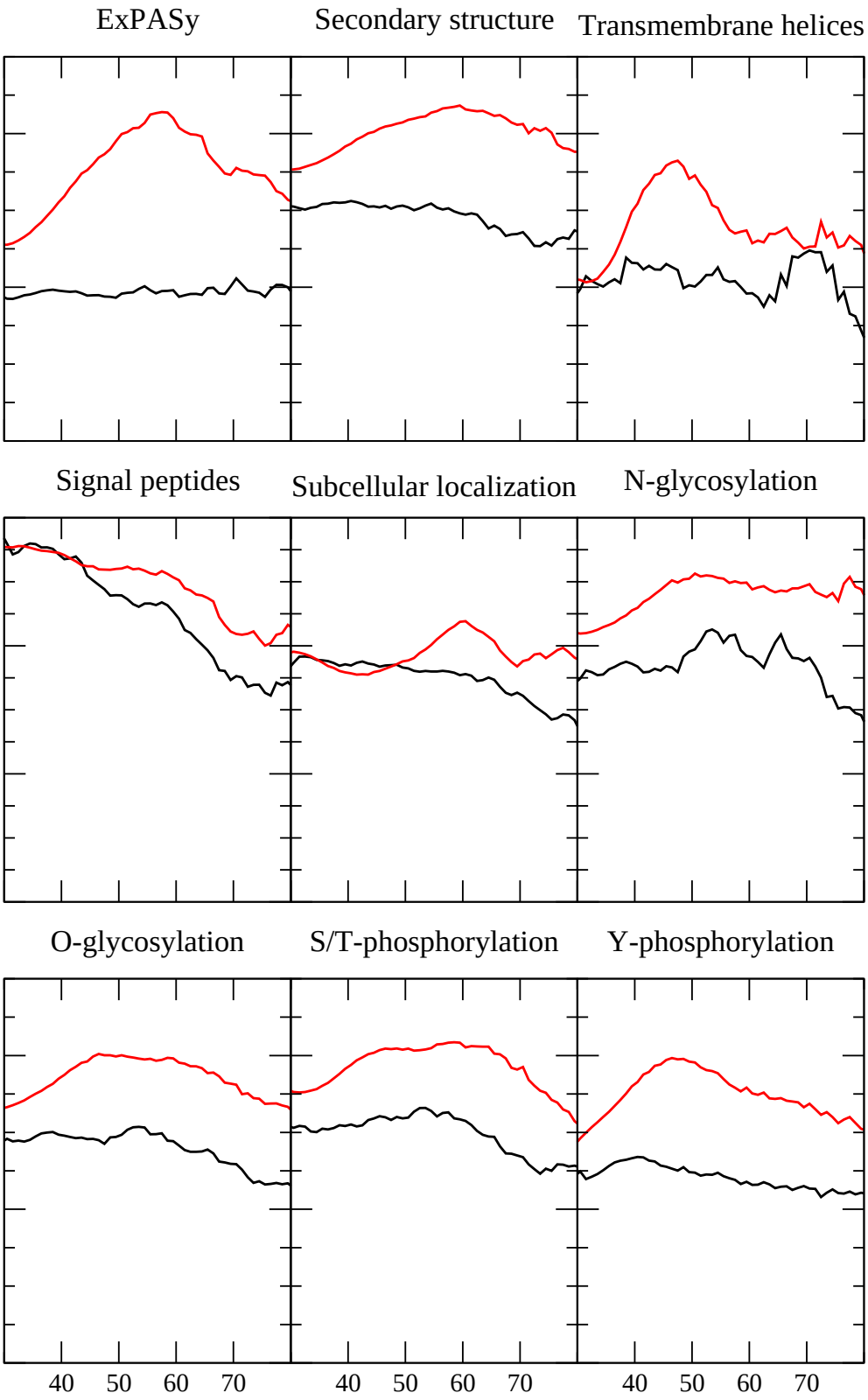


Figure 5.4: **Distances in feature space between orthologs and paralogs.** The Euclidian distance between normalized features is plotted as a function of sequence identity for pairs of orthologous (black) and paralogous (red) proteins. It is noteworthy that most features are more conserved over the entire range of sequence identity for homologs compared to paralogs.

5.3.3 Predicted functional similarity

Because the mapping between sequence derived and the ProtFun predictions is quite complex, the implications of the feature conservation on predicted function are not clear. The analysis of Euclidian distances suggest, that homologs are more likely to be predicted by ProtFun to have the same function than paralogs are.

To examine this, the function of all the *H. sapiens* and *D. melanogaster* was predicted with ProtFun. Based on the probabilistic assignment of cellular role categories, the probability that two proteins belong to the same cellular role can be estimated as the scalar product of the probability vectors. This way it was possible to plot as a function of sequence identity, the probability that two orthologous/paralogous proteins belong to the same cellular role category (see Figure 5.5 on page 99). It clearly shows that ProtFun predicts orthologs to be more likely to have the same function than paralogs, which is in perfect agreement with current belief (Remm et al., 2001).

5.4 Ability to generalize to other organisms

The ProtFun method was developed using human sequences only. However, all the biological support mentioned above holds for eukaryotes in general, indicating that ProtFun can be expected to work for other species than humans. The big questions are: how well do the individual predictors generalize to other species, and does it make biological sense when predictors do not work.

5.4.1 Creating data sets for cross-species comparison

The main obstacle standing in the way of making such a cross-species analysis is the lack of a common standard for function annotation (Lewis et al., 2000). In the annotation of several genomes, no classification system or controlled vocabulary was employed, rendering the current annotation next to useless for any automated purposes. The genome projects which have used well defined function descriptors have not all used the same—and the ones that have used the same function descriptors often used different criteria for assigning them to individual sequences. All in all, this makes functional comparison across genomes difficult at best. The way I have chosen to deal with this problem may seem a bit drastic: first all functional annotations present in the genomes are discarded and then new annotations are made using the same fully automated system to reannotate all the genomes.

To make the analysis as comprehensive and unbiased as possible, the largest possible set of predicted protein sequences was downloaded for a number of whole genome sequencing projects. These were selected to give a very broad coverage of eukaryotes: an insect (*D. melanogaster*), a round worm (*C. elegans*) a monocot plant (*A. thaliana*) and two yeasts (*S. cerevisiae* and *S. pombe*). In addition to these, sets consisting of all proteins annotated in 14 archaeal and 29 bacterial genomes were included in the analysis. A set of all predicted protein sequences in the human genome was also included as a reference.

Using the EUCLID method (Tamames et al., 1998; Andrade et al., 1999a), a score was calculated for each cellular role category from common keywords occurring in the SWISS-PROT entries found by a BLAST search. A labeling of the sequences was made, based on these scores by applying the same rules that were used for creating the ProtFun training set, i.e. a score above 3 signifies a positive example whereas a score below zero means a negative example.

Each protein sequence in all the data sets was also assigned to enzyme categories, based on the SWISS-PROT entries identified by the BLAST search. It was assigned as an enzyme only if at least two thirds of the database matches had an EC number in their description line. Equivalent to this, a protein was assigned as a non-enzyme if EC numbers were absent in more than two thirds of the database matches. Similarly, assignment of a major enzyme class to an enzyme required two thirds majority for the enzyme class (among the matches having EC numbers in their description line).

While the quality of this annotation will in many cases be worse than the original annotation, the automatically generated annotation has one crucial advantage: function has been assigned to proteins from all organisms using the same functional classification system and based on the same criteria. This allows for comparison across species.

5.4.2 Choosing the right performance measure

It is always difficult to boil the performance of a prediction method down to a single number. In the heart of this problem lies the tradeoff between sensitivity and specificity—which of two predictors is the better may very well depend on the application. There is thus no *right* way to do it.

So far in this thesis the Pearson correlation coefficient has been used when performance was to be reduced to a single number. It has several virtues, the most important being that no matter the data set, random performance will give a correlation coefficient of zero. However, there is one issue that makes it unsuitable for the problem at hand—correlation coefficients are not comparable across data sets which have different ratios of positive examples to negative examples (Baldi et al., 2000). This means that correlation coefficients cannot be compared across genomes because the protein coding genes have a different distribution over functional categories.

Instead the ROC area measure is used, which is defined as the area under the receiver output characteristic (ROC) curve—a graph showing the negative category sensitivity as function of positive category sensitivity. While not as convenient to calculate as the correlation coefficient, this measure has the advantage of being completely independent of the balancing data set. Random performance gives a ROC area of 0.5 while perfect performance corresponds to a ROC area of 1. A ROC area below 0.5 is equivalent to a negative correlation coefficient, i.e. a performance which is worse than random.

5.4.3 Cross-species evaluation of category and feature performance

Armed with comparably labeled data sets for a wide range of organisms, it is easy to calculate the ROC area performance measure of ProtFun for each functional category and each organism. A visualization of the resulting performance estimated is shown in Figure 5.6 on page 101, with the exception of genomes that encode too few proteins to give a good estimate of the performance (*M. genetalium*, *M. pneumoniae*, and *U. urealyticum*). The results verify that ProtFun does indeed appear to work for all eukaryotes, and surprisingly for some categories also on prokaryotes (although with poorer performance).

To understand why predictors of certain categories work on prokaryotes while others do not, the cellular role performance estimates were mapped onto features. For each feature and organism, a performance contribution was calculated as a weighted average of the ROC areas for cellular roles. Each cellular role enters a weight corresponding to the number of neural networks using the feature in question (see Figure 4.4 on page 62). The resulting feature performance contributions can also be seen in Figure 5.6, which reveal that it is mainly trends in physical/chemical properties and structural features that carry over from eukaryotes to prokaryotes.

Paper IV

5.5 Functionality of system components: Conservation of protein function in protein feature space

Lars Juhl Jensen, David W. Ussery, and Søren Brunak*
Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby, Denmark

* To whom correspondence should be addressed (email: brunak@cbs.dtu.dk)

A large number of protein features useful for prediction of protein function can be predicted from sequence, including post-translational modifications, subcellular localization, and physical/chemical properties. We show here that such protein features are more conserved among orthologs than paralogs, indicating they are crucial for protein function and thus subject to selective pressure. This means that a function prediction method based on sequence derived features may be able to discriminate between proteins with different function even when they have highly similar structure. Also, such a method is likely to perform well on other organisms than the one it was trained on. We evaluate the performance of such a method, ProtFun, which relies on such features as its sole input, and show that the method gives similar performance for most eukaryotes and performs much better than anticipated on archaea and bacteria. From this analysis we conclude that for the post-translational modifications studied, both the cellular use and the sequence motifs are conserved within Eukarya.

Introduction

Biological systems modeling at the molecular level normally requires knowledge about the functionality of the interacting components. The determination of protein function is an essential basis for many type of systems biology. It is a fundamental axiom that the structure of a protein determines its function. However, whether this is true or not depends very strongly on at what level one defines

"function". A close relationship between structure and function is observed if the detailed biochemical function is studied, such as which reaction is catalyzed by an enzyme. This type of functionality is often termed the "molecular function" and it is highly conserved within superfamilies, members of which according to the SCOP definitions are required to be related in both sequence, structure and function (Todd et al., 2001).

When studying the much broader "cellular role" categories, the relationship between structure and function becomes much less clear. For example, predicted protein secondary structure is much more useful for predicting enzyme class membership than cellular roles (Jensen et al., 2002). Several examples exist where proteins have different cellular roles although they belong to the same superfamily. The reverse is also true: proteins from many different superfamilies are involved in each of the particular cellular role categories.

Even for the chemically related EC classification the relationship to structure is unclear, for example the α/β-hydrolase superfamily contains—contrary to what the name indicates—not only hydrolases but also oxidoreductases, transferases and lyases (Todd et al., 2001). Another example is the zinc peptidase superfamily which includes a non-enzymatic receptor. Still, conservation of the enzyme class is seen for the majority of the enzyme superfamilies.

Typically, in any genome the function of only half the proteins can be assigned by sequence similarity search methods, while the rest remain unassigned. Some of these sequences of unknown function do not resemble any other known protein sequence; others have homologs but the function of these are also unknown. In any case, it is very difficult to suggest a function for these proteins.

For a long time, the paradigm behind solving this daunting task has been based on protein structure determination and prediction. The rationale has been that the structure of a protein is what determines its function, for which reason the function could be predicted via the structure, for example by homology building.

To be able to do this, several structural genomics initiatives have been started. These initiatives will be very useful for gaining new insight into the detailed chemical function of proteins that are today poorly understood. But given the relatively weak correlation between protein structure and cellular role combined with the vast number of unrelated proteins of unknown function, we believe that a different approach to predicting the cellular role of these proteins should be taken.

Instead, we have attempted to predict protein function based on predicted properties of proteins, such as physico-chemical properties, predicted post-translational modifications and subcellular localization signals (Jensen et al., 2002; Gupta et al., 2002). Although predicted from sequence, they are more conserved among orthologs than paralogs, given the same degree of sequence conservation. This is in contrast to three-dimensional structure, which is conserved for paralogs as well as orthologs.

We furthermore demonstrate that the sequence derived protein properties, characterize proteins of different cellular roles in ways that are conserved not only within Eukarya, but in several cases within all three domains of life: Eukarya, Archaea and Bacteria. These discoveries have been made through a cross-species analysis of the performance of the ProtFun prediction method (Jensen et al.,

2002) for a wide variety of organisms covering mammals, invertebrates, plants and fungi as well as Crenarchaeota, Euryarchaeota and Eubacteria.

Results and discussion

Features are more conserved among orthologs than paralogs

It is known and often utilized in function assignment that orthologs more often have identical function than paralogs (Jensen, 2001). If the sequence derived protein features we use are indeed indicative of protein function, they should then be expected to be more conserved within pairs of orthologous proteins than within pairs of paralogous proteins. As a consequence of this, the ProtFun method should more often predict the same function for orthologous proteins than paralogous proteins.

We have verified this on a data set consisting of all orthologs and paralogs between the complete genomes of *H. sapiens* and *D. melanogaster*. Because orthologs typically are more similar than paralogs at the sequence level, we have examined the feature similarities as function of the sequence identity (see Fig-

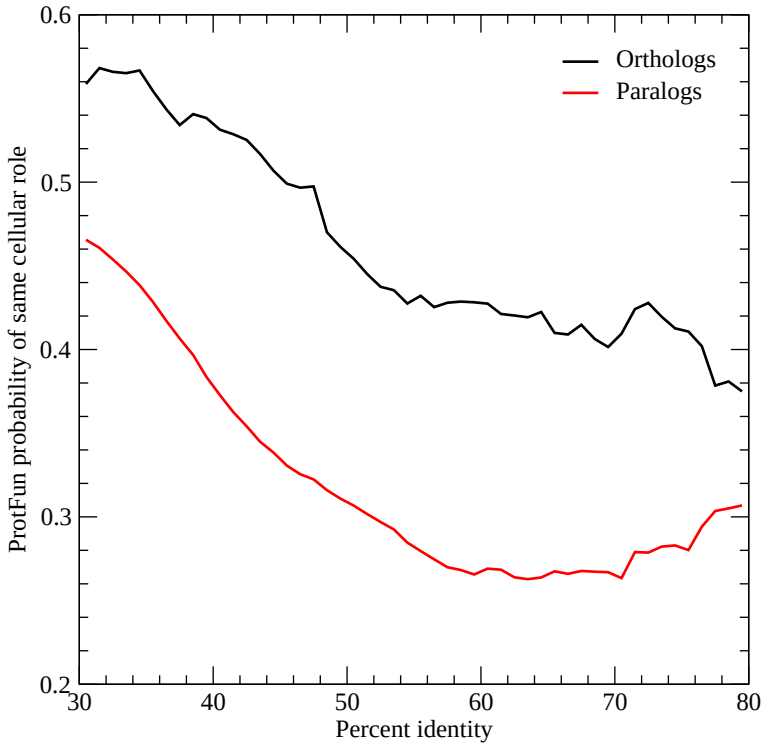


Figure 5.5: **Estimated probability for same cellular role as function of similarity for orthologs and paralogs.** These probabilities were estimated as the overlap integral of the ProtFun predictions for *H. sapiens* and *D. melanogaster* proteins involved in each pair. The probabilities could not be reliably estimated outside the range 30–80% identity as orthology vs. paralogy cannot be reliably predicted for distant homologs and because very closely related paralogs are likely predicted to be orthologs.

ure 5.5). It is clear that the functional similarity as predicted from sequence derived features is most conserved for orthologs over then entire range of sequence similarity studied.

The ProtFun predictions rely exclusively on sequence derived features as input. The ability of the method to discern between orthologs and paralogs therefore have the implication that the protein features are selectively conserved for orthologs.

Cross-species comparison

The ProtFun function prediction method has been trained on human protein sequences of known function only (Jensen et al., 2002), but given that it relies on protein features that occur in all eukaryotes it should be expected to be able to generalize to other organisms as well. Considering the results above, it would appear that the method is able to generalize at least to metazoans.

To investigate this further, we evaluated the performance of ProtFun on the complete genomes of 48 organisms. When doing this kind of comparative analysis of genomes/proteomes from many different species there are several potential sources of artefacts. Because the genes have been annotated using different gene finding methods or different similarity cutoffs, there can be vast differences in the quality of the annotations (Skovgaard et al., 2001). Also, because genomes have been annotated by different groups inconsistencies in the functional annotations are likely to occur. In order to compare protein function across multiple genomes, one has to make sure that the annotation is consistent.

We address these problems by reannotating the function of all proteins based on sequence similarity using the EUCLID method (Tamames et al., 1998; Andrade et al., 1999a) restricting ourselves to use proteins where a function could be assigned reliably. Since questionable ORFs that might have been annotated as genes are very unlikely to display significant sequence similarity to proteins in SWISS-PROT, these will automatically be rejected. The fully automated assignments into functional classes ensure comparability across organisms, but are likely to be less accurate than the original annotations. The fact that not only our own predictions will contain errors, but also the labeling to which we compare, means that we will obtain a conservative estimate of the ProtFun performance.

Good performance on all eukaryotes

To our surprise the ProtFun method performs almost equally well on all other eukaryotes tested including yeasts (see Figure 5.6). This ability to generalize across very different phyla shows that the trends found by the artificial neural networks do not only hold for human proteins but have in fact been conserved throughout the eukaryotic domain of life.

Sequence derived input features

Our approach to function prediction relies on sequence derived input features. These represent physical/chemical and functional biological properties of the pro-

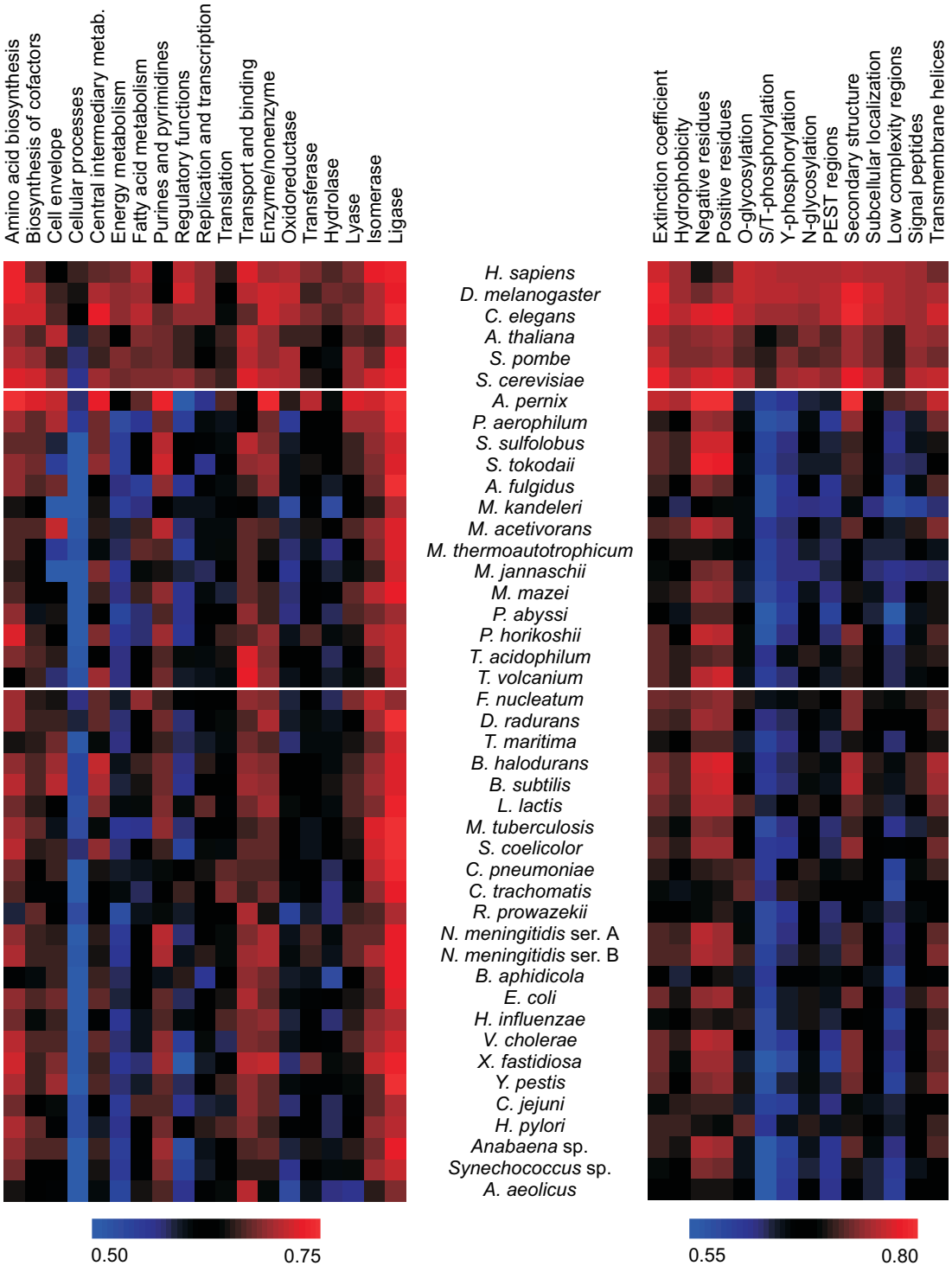


Figure 5.6: **ProtFun performance for functional classes and performance contributions from input features.** For 44 organisms the area under the receiver output characteristic (ROC) curve has been plotted for all cellular role categories and enzyme classes (left panel). These performances were mapped onto input features based on the feature usage matrix (see Figure 1 in (Jensen et al., 2002)).

tein which can be either calculated or predicted from the amino acid sequence alone. These features include predicted protein secondary structure, transmembrane helices, subcellular localization and post-translational modifications.

While all of these features make biological sense for eukaryotes in general, many of the feature predictors had been trained on mammalian or vertebrate data sets. Their performance on other eukaryotes was therefore unknown.

In the case of prokaryotes some of the features make no sense at all. For instance the lack of compartmentation in prokaryotes means that prediction of most subcellular localization makes little sense. How well other features like post-translational modifications (PTMs) will work for prokaryotes is even less clear: the functional role of a modification may be a different from that in eukaryotes, the motif may be different or the modification may not take place at all.

We have analyzed this in a systematic and quantitative fashion. The performances obtained for cellular roles were mapped according to feature importance. The resulting values represent the performance contributed by each sequence derived feature. Figure 5.6 shows these values visualized in the same way as the functional category performances. The general trends are as follows: Features representing structural properties like predicted secondary structure and membrane spanning helices as well as more general physico-chemical properties of the proteins generalize well to prokaryotes. On the other hand, most of the features representing predicted PTMs and protein sorting signals are of limited value in archaeal and bacterial genomes.

There are certain organisms that deviate from the patterns described above. One example is *Buchnera aphidicola* which belongs to the γ subdivision of proteobacteria. In contrast to most other organisms, even the correlations between simple physical/chemical properties (extinction coefficient, hydrophobicity and number of negative/positive residues) appear to break down (see Figure 5.6). All of these features reflect different aspects of the amino acid composition. The lack of correlation is thus likely to result from the unusual amino acid composition of *B. aphidicola* proteins, which is reflected in the predicted isoelectric points of *B. aphidicola* proteins (Shigenobu et al., 2000, 2001).

Many features fail on prokaryotes

It was anticipated, due to the very different organization of the eukaryotic and prokaryotic cells, that the predicted protein subcellular localization (according to PSORT) would be of little use in prokaryotes. Still, one could have expected N-terminal signal peptide prediction to work, as the signal peptides not only exist in prokaryotes but can be accurately predicted by the SignalP method which we use in ProtFun (Nielsen et al., 1997a).

The problem is that signal peptides do not play the exact same role in eukaryotes and in prokaryotes. Also, eukaryotes have several types of similar N-terminal targeting sequences, which can all be detected from the SignalP scores. For example, eukaryotic proteins targeted for the mitochondria will have mitochondrial targeting peptides, while their prokaryotic counterparts would be expected to be cytoplasmic and thus not have signal peptides. This difference in the meaning of similar biological motifs in prokaryotes and eukaryotes, explains the very poor

performance of the *energy metabolism* predictor for prokaryotes.

Two types of predicted glycosylation sites, both targeting secreted and membrane associated proteins, are being used by ProtFun. One is N-linked β-GlcNAc glycosylation of asparagines, which takes place in the endoplasmic reticulum. The other is O-linked α-GalNAc glycosylation of serines and threonines, which takes place in the Golgi. Possibly because glycosylation has not been studied nearly as much in prokaryotes as in eukaryotes, only one of the two types (N-linked β-GlcNAc glycosylation) has been observed in prokaryotes (Spiro, 2002). As with signal peptides, the consensus sequence for this modification appears to be the same in prokaryotes and eukaryotes. It is thus reasonable to expect the NetNGlyc predictor to work on prokaryotes even though it was trained on eukaryotic sequences.

Glycosylation also seems to play much the same role in prokaryotes and eukaryotes, but much fewer proteins appear to be glycosylated in prokaryotes (Spiro, 2002). The small number of glycoproteins may explain why glycosylation predictions appear to be of limited value for predicting functional classes in prokaryotes, despite both the consensus sequence and function being conserved.

Similar to glycosylation, phosphorylation is known to play an important role in prokaryotes, where it is involved in regulation like in eukaryotes. This makes phosphorylation sites a biologically relevant feature, which could be used for function prediction in prokaryotes. However, it was questionable if predicted phosphorylation sites could be used since the NetPhos predictor was trained solely on eukaryotic data. This depends entirely on whether the specificities of some of the prokaryotic kinases are sufficiently close to those of eukaryotic kinases. In our cross-species analysis we find that predicted phosphorylation sites contribute little to the performance on prokaryotic proteins, which indicates that the specificities of prokaryotic kinases are quite different from those of eukaryotic kinases.

Considering that so many of the input features used by ProtFun make little or no sense for prokaryotic, it is somewhat surprising that the method works at all for them. Figure 5.6 shows that the features mainly responsible for this are the physical/chemical properties, in particular the size and charge of the protein represented by the number of negative/positive residues. The only other features that contribute significantly are those related to structure, i.e. secondary structure and transmembrane helix prediction.

Universal feature usage and consensus in Eukarya

An interesting implication of the ability to generalize across species is that the different post-translational modifications apparently serve the same purposes for most if not all eukaryotes. Not only do eukaryotes have the gene repertoire for making the same modifications, they also employ them in a consistent manner.

The fact that all feature–function correlations hold within Eukarya has one further implication. It indirectly indicates that most (if not all) of the predictors that are used by ProtFun can be expected to work with reasonable accuracy for all eukaryotes. As mentioned above, this could not be taken for granted as some of them have been trained on data sets consisting exclusively of human proteins.

Same structure—different functions

In the introduction several examples of SCOP superfamilies containing enzymes from entirely different enzyme classes and superfamilies containing both enzymes and non-enzymes. These cases show, that conservation of structure at the superfamily level is not sufficient to guarantee that function is also conserved.

With respect to enzyme classification the Cupredoxin superfamily is one of the most diverse, containing an almost equal proportion of enzymes and non-enzymes. Table 5.1 shows the enzyme probabilities predicted by ProtFun along with the experimental assignment (Todd et al., 2001). Although all the proteins have the same conserved three-dimensional structure, our approach is able to correctly discriminate between the enzymatic and non-enzymatic members of the Cupredoxin superfamily. It should be pointed out though, that all the enzymatic members of the superfamily belong to the same protein family, even though some of them are less than 30% identical at the amino acid level.

Evolution of members within the same superfamily of proteins both with and without enzymatic activity is likely to have happened through gene duplication events and subsequent adaption of one of the copies for a new function. An enzymatic and a non-enzymatic member of the same superfamily are thus likely to be paralogs. It is therefore plausible that the stronger conservation of protein

Table 5.1: **Predictions for members of the Cupredoxin superfamily.** For each member of the superfamily the enzyme probability score from ProtFun is listed along with the experimental enzyme/non-enzyme assignment (Todd et al., 2001). The non-enzymes marked with an asterisk are part of enzymatic complexes, but do not contain active sites.

PDB identifier	Chain	Enzyme prob.	Experimental assignment
1NWP	A	0.257	Non-enzyme
1NWP	B	0.257	Non-enzyme
2CBP		0.289	Non-enzyme
1AAC		0.301	Non-enzyme
1PLC		0.310	Non-enzyme
1RCY		0.325	Non-enzyme
2CUA	B	0.354	Non-enzyme *
2CUA	A	0.368	Non-enzyme *
1JER		0.404	Non-enzyme
1PAZ		0.416	Non-enzyme
1CYW		0.483	Non-enzyme *
1A65	A	0.652	Enzyme
1NIF		0.688	Enzyme
1AOZ	A	0.773	Enzyme
1AOZ	B	0.773	Enzyme
1KCW		0.792	Enzyme

features observed for orthologs compared to paralogs is related to the ability to discriminate between structurally similar but functionally dissimilar proteins.

Conclusions

For a long time, there has been a very strong focus on the importance of protein structure for understanding protein function. However, based on our analysis we conjecture that many other protein properties, e.g. post-translational modifications, may in fact be more, or at least, equally important for determining and maintaining the function of a protein. These properties appear to be conserved among proteins of similar function, both in cases where the evolutionary relationship can be detected by sequence similarity and in more distantly related proteins of similar structure.

Materials and methods

Generation of the data set

A set of 23,740 protein sequences corresponding to predicted human genes was downloaded from the Ensembl database (Hubbard et al., 2002). Similarly a set of 14,334 *Drosophila melanogaster* protein sequences was obtained from FlyBase (The FlyBase Consortium, 2002), 20,263 *Caenorhabditis elegans* sequences from the protein database WormBase and 25,617 *Arabidopsis thaliana* protein sequences from The Arabidopsis Information Resource (TAIR) (Huala et al., 2001).

In addition to these eukaryotic data sets, the complete genome sequences of the two yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* were downloaded from GenBank and all translations of annotated protein coding regions were extracted (Benson et al., 2002). Protein data sets for 14 archaea and 28 bacteria were extracted in the same manner from the complete genome sequences (see Table 5.2).

To ensure comparable function annotation among these many genomes, all existing (if any) information on protein function was discarded and the proteins were automatically reassigned to cellular role categories by using the EUCLID method (Tamames et al., 1998; Andrade et al., 1999a) and to enzyme categories based on the following criteria.

Based on BLAST matches to known proteins in SWISS-PROT, EUCLID collect keywords which are used in an additive scoring system to calculate a Z-score for each cellular role category. We annotated each category separately based on the EUCLID Z-scores using the same rules used to label the training examples used for development of the ProtFun method (Jensen et al., 2002). Sequences were labeled as *positive examples* if their Z-score was above 3 while sequences with a Z-score less than 0 were labeled as *negative* examples. Any sequences having a Z-score from 0 to 3 were left out of the analysis for the category in question.

Sequences were assigned to enzyme classes based on the same BLAST matches used for selecting keywords above. As SWISS-PROT provides enzyme class infor-

Table 5.2: **Data sets used for cross-species evaluation.** The column "protein sequences" lists the number of protein coding regions annotated in the genomes, with the exception of the organisms *H. sapiens*, *D. melanogaster*, *C. elegans*, and *A. thaliana* (see text for details on these data sets). The protein sequences that could be assigned to a cellular role by the EUCLID method (last column), shows the amount of data available for validation of the ProtFun method for each organism.

Organism	Protein sequences	Assigned by EUCLID
H. sapiens	23,740	13,419
D. melanogaster	14,334	7,235
C. elegans	20,263	7,840
A. thaliana	25,617	11,771
S. pombe	4,952	2,786
S. cerevisiae	6,329	3,302
A. pernix	2,694	684
P. aerophilum	2,605	867
S. solfataricus	2,977	1,186
S. tokodaii	2,826	1,045
A. fulgidus	2,407	1,074
M. thermoautotrophicum	1,869	867
M. jannaschii	1,715	781
M. mazei	3,371	1,420
M. kandleri	1,691	653
M. acetivorans	4,540	1,850
P. abyssi	1,765	855
P. horikoshii	2,064	786
T. acidophilum	1,031	783
T. volcanium	1,499	792
Anabaena sp.	5,366	2,444
A. aeolicus	1,522	926
B. burgdorferi	850	461
B. halodurans	4,066	2,223
B. subtilis	4,100	2,240
Buchnera sp.	564	469
C. jejuni	1,654	975
C. pneumoniae	1,052	530
C. trachomatis	894	498
D. radiodurans	2,937	1,332
E. coli	4,289	2,883
F. nucleatum	2,068	1,083
H. influenzae	1,709	1,183
H. pylori	1,566	815
L. lactis	2,266	1,229
M. genitalium	480	332
M. pneumoniae	677	441
M. tuberculosis	3,918	1,973
N. meningitidis ser. A	2,121	1,132
N. meningitidis ser. B	2,025	1,088
R. prowazekii	834	548
S. coelicolor	7,848	3,625
Synechocystis sp.	3,169	1,598
T. maritima	1,846	1,064
T. pallidum	1,031	507
V. cholerae	3,828	2,054
X. fastidiosa	2,766	1,184
Y. pestis	4,008	2,566

mation for most enzymes in the description field, this information was extracted for all BLAST matches identified by running EUCLID. Labeling of enzyme vs. non-enzyme as well as the major enzyme class was decided by voting among the matches. At least two thirds majority for either "yes" or "no" was required for a sequence to be labeled.

Performance evaluation

One of the best commonly used performance evaluation criteria is the correlation coefficient, which we have used as the main criteria during development of the ProtFun method. However, the correlation coefficient cannot be used for the problem at hand because of its dependence upon the relative frequency of positive and negative examples in the data set (Baldi et al., 2000). Correlation coefficients can thus not be used for comparing the performance of our prediction method across genomes with different break down on functional categories.

Instead we opt for using the area under the receiver output characteristic (ROC) curve, a plot of true negative rate vs. true positive rate. The area under this curve will be 1 for a perfect predictor and 0.5 for a predictor performing no better than random. Like the correlation coefficient, this performance measure is balanced, taking into account the tradeoff between high sensitivity and low rate of false positives. In addition to this, it is also independent of the data set composition in terms of positive and negative examples.

Feature mapping

The ROC area performances functional classes were mapped onto the sequence derived features. These sequence derived feature were the prediction methods NetNGlyc (manuscript in preparation), NetOGlyc (Hansen et al., 1998), NetPhos (Blom et al., 1999), PEST regions (Rechsteiner and Rogers, 1996), PSIPRED (Jones, 1999), PSORT (Nakai and Horton, 1999), SEG filter (Wootton, 1994a), SignalP (Nielsen et al., 1999) and TMHMM (Krogh et al., 2001) as well as the number of calculated features: extinction coefficient, grand average hydrophobicity, and the numbers of positively and negatively charged residues.

For each organism, the performance of each of these 14 input features was calculated as a weighted average of the ROC areas of the 12 cellular role categories. Each cellular role category entered with a weight corresponding to the number of neural networks in its ensemble of predictors that make use of the feature in question (see Figure 1 in (Jensen et al., 2002)). We decided to not include the enzyme classifiers in this mapping procedure because all of the neural network ensembles make use of a large number of sequence derived features. This makes it very difficult to correctly attribute the predictive performance to the right features for these classifiers.

Obtaining sets of orthologs and paralogs

Assignment of orthologs vs. paralogs is far from being a trivial problem. To obtain a large data set of orthologs/in-paralogs and out-paralogs, we have made use of

the INPARANOID tool to classify the pairs of homologous proteins between the *H. sapiens* and *D. melanogaster* described above (Remm et al., 2001). Paralogs were assigned based on BLAST matches covering at least 50% of the sequence length, which were not listed as orthologs by INPARANOID. By this approach we predicted 13,562 pairs orthologous proteins and 151,923 pairs of paralogous proteins. In order ensure comparability of the two data sets only pairs of paralogs consisting of one *H. sapiens* and one *D. melanogaster* protein were included.

Acknowledgments

The authors wish to thank Ulrik de Lichtenberg and Thomas Skøt Jensen for valuable discussions and ideas. Thanks also to Marie Skovgaard for comments on the manuscript. This work was supported by the Danish National Research Foundation.

5.6 Case stories for non-human proteins

5.6.1 Essential proteins of unknown function in *M. genetalium*

A minimal set of approximately 300 essential genes have been identified in the tiny genome of *Mycoplasma genetalium* by systematic single gene knockout. Surprisingly, the function of 112 of was still unknown in 1999 (Hutchison III et al., 1999).

In an attempt to predict the function of these, human homologs of them were looked for using gapped BLAST. Homologs were identified for only 15 of 112 genes, none of which resulted in strong predictions by ProtFun. After having realized ProtFuns ability to generalize to prokaryotes for certain functional categories, it was attempted to use the *M. genetalium* sequences directly.

To strengthen the predictions by ProtFun, its predictions were combined with predicted functional links. Based on the criteria by Marcotte and coworkers, high confidence links could be identified for only eight of the 112 essential genes of unknown function. These genes were: MG120, MG121, MG134, MG200, MG259, MG349, MG387 and MG448. Both of the proteins MG120p and MG121p have high confidence links to MG040p which belongs to the very common protein family "Basic membrane lipoprotein" (IPR003760)—a class of prokaryotic lipid binding proteins.

MG120p as well as MG121p are predicted by ProtFun to belong to the class *transport and binding* with high probability scores (0.816 and 0.773 respectively). These predictions are in agreement with the functional links since MG040p EUCLID gives MG040p a score of 4.19 for the category *transport and binding*, i.e. above the threshold used for identifying positive examples when creating the training set for ProtFun. There are thus several reasons for believing that MG120p and MG121p are indeed *transport and binding* proteins.

5.6.2 Circular proteins

It seems that whenever bioinformaticians believe to have addressed a biological issue, nature turns out to be more inventive. An example of this is trans-splicing, which breaks every rule built into eukaryotic gene finders.

In the ProtFun method very few assumptions are made about proteins—one is that protein sequences are linear. Given that, the existence of circular proteins hardly comes as a surprise.

All circular proteins known are very small (less than 100 aa) and have presumably evolved from ancestral linear proteins (Trabi and Craik, 2002). Little is known about the biosynthesis of circular proteins, but they all appear to be made from longer linear precursors through cleavage and cyclization.

In order to test the performance of ProtFun on circular proteins, one must first find a way to represent them as input vectors. It is not clear what is the best representation—in this test the mature circular protein sequence was simply cut at the site where cyclization took place, this way getting a short linear sequence corresponding to the mature circular protein.

Table 5.3: **ProtFun predictions for cyclic proteins.** The ProtFun prediction method was run on a number of cyclic proteins. To turn the sequences into linear sequences that can be used as input, all the cyclic proteins were cut at the site of cyclization.

	varv				wAgaIV		
	A	B	C	D	A	B	BC
Amino acid biosynthesis	0.500	0.500	0.500	0.500	0.500	0.500	0.818
Biosynthesis of cofactors	0.472	1.431	0.472	0.653	0.417	0.458	0.889
Cell envelope	5.262	3.393	5.000	1.787	5.082	4.689	1.180
Cellular processes	0.548	0.740	0.562	0.507	0.658	0.370	0.534
Central intermediary metabolism	0.651	0.651	0.651	0.651	0.651	0.651	0.857
Energy metabolism	1.878	0.900	2.000	1.811	1.089	0.478	1.722
Fatty acid metabolism	4.846	5.308	4.538	5.000	2.154	1.308	1.308
Purines and pyrimidines	0.193	0.206	0.193	0.222	0.193	0.193	0.214
Regulatory functions	0.329	0.596	0.323	0.236	0.925	0.882	0.081
Replication and transcription	0.373	0.873	0.366	0.325	1.142	1.056	0.075
Translation	2.614	4.205	2.545	2.432	6.750	4.250	1.023
Transport and binding	0.463	0.134	0.485	0.451	0.041	0.068	1.498
Enzyme	1.187	1.470	1.187	1.655	0.744	0.733	0.747
Nonenzyme	0.925	0.811	0.925	0.737	1.103	1.107	1.102
Oxidoreductase	0.255	0.144	0.255	0.216	0.139	0.139	0.596
Transferase	0.096	0.241	0.096	0.130	0.099	0.090	0.104
Hydrolase	0.180	0.180	0.180	0.205	0.192	0.196	0.180
Lyase	0.426	0.426	0.426	0.426	0.426	0.426	0.447
Isomerase	0.313	0.313	0.313	0.313	0.313	0.313	0.313
Ligase	0.334	0.334	0.334	0.334	0.334	0.334	0.334

	CMT-III	CPI	cyclo-psycho	Kalata BI	Tx31	Tx32	viola-peptid
Amino acid biosynthesis	0.500	0.500	0.773	0.500	1.273	0.727	0.545
Biosynthesis of cofactors	0.417	0.417	0.417	0.472	0.639	1.139	0.458
Cell envelope	0.525	1.377	3.016	5.262	0.525	1.066	4.574
Cellular processes	0.397	0.438	0.370	0.630	1.014	0.370	0.466
Central intermediary metabolism	0.683	0.651	0.651	0.651	0.730	0.714	0.651
Energy metabolism	0.456	0.811	2.778	1.833	1.878	0.344	0.733
Fatty acid metabolism	4.385	1.385	2.923	4.615	1.846	1.308	4.231
Purines and pyrimidines	0.193	0.193	0.321	0.193	0.313	0.222	0.193
Regulatory functions	0.752	0.534	0.398	0.404	0.093	0.286	0.398
Replication and transcription	1.075	1.060	0.623	0.358	0.075	0.097	0.347
Translation	6.841	5.864	2.432	2.500	0.750	0.682	2.568
Transport and binding	0.073	0.161	0.098	0.715	1.283	1.512	0.661
Enzyme	0.754	0.747	1.655	1.187	0.827	0.894	0.621
Nonenzyme	1.099	1.102	0.737	0.925	1.069	1.043	1.152
Oxidoreductase	0.130	0.120	0.144	0.269	0.298	0.125	0.115
Transferase	0.136	0.093	0.342	0.096	0.180	0.116	0.093
Hydrolase	0.186	0.180	0.180	0.180	0.180	0.180	0.180
Lyase	0.426	0.426	0.426	0.426	0.958	0.426	0.426
Isomerase	0.313	0.313	0.313	0.313	0.313	0.313	0.313
Ligase	0.334	0.334	0.334	0.334	0.334	0.334	0.334

The results obtained for the circular protein sequences known (Trabi and Craik, 2002) are not impressive. In fact the majority of the proteins give high odds scores for both of the categories *cell envelope* and *translation* (see for examples Table 5.3). This is easily explained comparing the small size of all the cyclic proteins with the length distributions shown in Figure 4.2 on page 46. It is possible that better results could be obtained by using the full precursor sequences as input instead of the mature sequences, but this was not tested.

5.6.3 ATP binding proteins from a random library

As most of the neural networks in ProtFun rely on a combination of protein properties that are required for the chemical function and properties that are important only in a cellular context, it is hard to know if ProtFun will produce meaningful output for protein sequences constructed through *in vitro* selection methods.

By employing an ingenious *in vitro* selection procedure in which proteins and their corresponding cDNA sequences were covalently cross-linked, four families of ATP binding proteins have been developed starting from a pool of random sequences (Keefe and Szostak, 2001). Table 5.4 shows the ProtFun predictions

Table 5.4: **ProtFun predictions for ATP binding proteins developed by *in vitro* selection.** As the ProtFun method depends on the cellular context to predict function, it would be surprising if it worked for these artificially created proteins. As expected it does not seem to work.

	Family A	Family B	Family C	Family D
Amino acid biosynthesis	0.818	0.500	0.727	0.682
Biosynthesis of cofactors	1.167	0.708	1.472	0.458
Cell envelope	0.475	0.492	0.508	0.541
Cellular processes	1.027	0.863	0.507	0.411
Central intermediary metabolism	0.714	0.667	1.429	0.651
Energy metabolism	2.722	2.911	3.278	0.511
Fatty acid metabolism	1.308	1.308	1.462	1.308
Purines and pyrimidines	0.193	0.593	0.214	0.193
Regulatory functions	0.348	0.106	0.205	0.534
Replication and transcription	0.713	0.437	0.530	1.216
Translation	3.523	5.227	6.705	5.045
Transport and binding	0.059	0.051	0.080	0.107
Enzyme	0.859	0.852	1.215	0.744
Nonenzyme	1.057	1.059	0.914	1.103
Oxidoreductase	0.586	0.793	0.668	0.216
Transferase	0.180	0.119	0.145	0.099
Hydrolase	0.183	0.180	0.375	0.192
Lyase	0.426	0.426	0.426	0.426
Isomerase	0.313	0.313	0.313	0.313
Ligase	0.334	0.334	0.334	0.334

for a representative of each family.

The predictions all give high odds scores to the categories *translation* like the circular proteins just discussed. This is again explained by the fairly small size of the proteins (the longest of which is 109 aa). Quite high scores are also seen for *energy metabolism*, which could perhaps be argued to make sense for ATP binding proteins. However, this is not very convincing since, given the number of processes in which ATP is involved, similar arguments could be made for any of the other categories.

5.7 A lighter shade of dark

To begin with, the ProtFun method may appear as black magic—how can neural networks predict the function of protein they have never seen before from its sequence alone? I hope that the analysis presented in this chapter has shed some light on (or into) the black box.

Several lines of evidence suggest that the ProtFun method works by recognizing the same protein features that a eukaryotic cell does: proteins are sorted into different parts of the cell where they perform their function and are modified in various ways to make them better suited for their purpose. Furthermore, the chemical properties of the proteins have become adapted to the environment in which they perform their function.

It thus makes sense that the ProtFun method works almost equally well for proteins from all eukaryotes whereas the performance on prokaryotic proteins is much poorer. Also the method should not be expected to work well for proteins designed or selected for having a particular function *in vitro*.

Chapter 6

Comparison of ProtFun with other existing methods

The first kind of large volume data sets to be analyzed by bioinformaticians was sequence data. Since then, protein–protein interaction data and several other types of data have been analyzed, most if not all of which can be used to assist in resolving the function of proteins. A general trend is that these methods help predict function by linking together proteins of similar function. Perhaps mainly due to the difficulties of defining what is meant by "similar function", very few quantitative statements have been made about the prediction quality of such links.

Because these methods provide links between proteins rather than assignment of proteins to classes, it is difficult to compare the methods. The way chosen here to address this problem is by only evaluating the quality of links between proteins that can both be reliably assigned cellular roles by EUCLID. For any given category, a true positive can then be defined as a link between positive examples and a true negative as a link between two negative examples. Links between a positive and a negative example are counted as $1/2$ false positive and $1/2$ false negative. From these numbers the Matthews correlation coefficient for functions of linked proteins can be calculated.

The rationale behind this scheme is, that using a true link to assign function of one protein based on the other will result in a true assignment regardless of the direction in which the link is used. Using a false link on the other hand will result in either a false positive or a false negative assignment depending on the direction in which the link is used. The false links should thus be evenly distributed over the two possibilities.

It should be noted that this method for evaluating the predictions of functional links actually favors the link methods over ProtFun. This is because such methods do not link all proteins of unknown function to a protein of known function—the coverage is thus not 100%. ProtFun on the other hand is forced to assign every single protein to a class. If allowed to discard the most uncertain predictions, higher correlation coefficients could no doubt be reached.

Most of these alternative methods for function assignment have mainly been used for assigning function to *S. cerevisiae* proteins, for which the best data sets are thus available. This restricts the comparison of the performance of the

different methods (including ProtFun) to yeast. While it has already been shown that ProtFun works on yeast, one should bear in mind that it was developed for human proteins and that the performance on yeast is the poorest we have observed so far for a eukaryote. This further puts ProtFun in an unfavorable position for the coming comparison.

6.1 Clustering of expression profiles

The emergence of DNA array technologies marked the beginning of a new era of bioinformatics. Using either Stanford type cDNA arrays or Affymetrix GeneChips with short oligo probes, it is possible to simultaneously monitor the expression levels of all genes in an organism. This technology was first used on *S. cerevisiae* and today this is still the organism for which most array data is publicly available. Expression data is probably the only source of biological information growing faster in volume than the DNA sequence databases.

From the very beginning this type of data has been used to predict protein function. By clustering proteins showing similar expression profiles it was discovered that proteins involved in the same cellular processes clustered together, an observation that was instantly exploited to infer function (Eisen et al., 1998; Spellman et al., 1998).

In addition to such unsupervised machine learning approaches, SVMs have also been employed to predict from gene expression profiles a set of five functional classes (Brown et al., 2000). While this prediction method achieves high accuracy, it should be noted that the classes were selected based on the knowledge that genes from these particular classes clustered together. In fact there are only five more classes out of the approximately 200 classes in the MIPS classification system that can be learned from the expression profiles used (Gustavo Stolovitsky, personal communication). Also it is worth mentioning, that although the authors claim the superiority of SVMs over other machine learning approaches, their performance was not compared to that of neural networks, with which very similar results are obtained (Gustavo Stolovitsky, personal communication).

In order to investigate the correlation between protein function and expression profile clustering in a way that allows for comparison with ProtFun, the Rosetta compendium by Hughes et al. (2000a) was downloaded. It consists of whole genome expression data measured for *S. cerevisiae* under 300 different conditions, including both environmental changes and knockout mutants. It is the most comprehensive gene expression data set to date, likely making it the best suited for prediction of cellular roles.

The Rosetta compendium covers 6230 yeast transcripts each measured for 300 conditions. All pairwise Pearson correlation coefficients were calculated, with the exception of pairs where more than 50 dimensions were lacking due to missing values in the data set. For each gene the most correlated profile with a Pearson correlation of at least 0.6 was identified. Based on these links the function correlation coefficient was calculated as described above (see Table 6.1 on page 117). It is thus a very optimistic measure of the performance of array clustering methods, namely the performance that can be expected if function is only inferred when

the function of the closest neighbor is known. As a gene is almost guaranteed to cluster together with its closest neighbor, any clustering method is unlikely to perform better.

The differences in performance of the different functional classes are very much as should be expected (see Table 6.1). The expression profile approach excels in the *translation* category, a category known to contain sets of highly coregulated genes, e.g. ribosomal proteins. Similarly it works well on several of the metabolism categories. As a consequence of this and the distribution of enzymes over cellular roles (see Table 4.1), it also gives a decent performance on predicting enzymes vs. non-enzymes, although enzymes as such should not be expected to cluster together. Conversely, the method gives a quite poor performance for *regulatory functions*—again as should be anticipated. The only slight surprise is the equally poor performance on the *purines and pyrimidines* category.

6.2 Protein–protein interaction screening

Another source of information on the function of proteins is the interactions between different proteins. This information could come either from experimental determination by yeast two-hybrid methods or other methods, or it may be from predicted interactions. When working with low resolution function prediction, it would be expected that interacting proteins usually belong to the same category. This can obviously be used if the function of one or more of the interaction partners is known, but even when this is not the case it could be used as a constraint on the prediction method by tying the function prediction of interacting proteins.

Protein–protein interactions often imply that two or more proteins together form a larger functional component. It is thus likely that interacting proteins share a common function and this has been used by several groups to infer function of proteins which share no sequence similarity to sequences of known function. However, not much has been done to quantify the strength of this relationship. The reasons for this are likely the same that are listed above for expression profile clustering.

Several methods exist for identifying protein–protein interactions: co-immuno precipitation, mass spectrometry (Ho et al., 2002) and the yeast two-hybrid screening approach. The latter has attracted a lot of attention as it allows for fast systematic screening of protein–protein interactions in whole genomes (Ito et al., 2001; Uetz et al., 2000).

Two research groups have performed systematic screens of the *S. cerevisiae* genome using the yeast two-hybrid approach (Ito et al., 2001; Uetz et al., 2000). Several databases of protein–protein interaction data currently exist, the most complete of which appears to be Database of Interacting Proteins (DIP) (Xenarios et al., 2002), which covers all organisms. In the case of yeast interaction data, it should be noted though that the MIPS database contains additional such data. Schwikowski et al. (2000) have published a combined data set consisting of interactions from several of these sources . This data set forms the basis for the present evaluation of the correlation between functions of physically interacting

proteins.

The correlation between function of linked proteins, calculated as described above, can be seen in Table 6.1. In light of interactions between translation related proteins being underrepresented in yeast two-hybrid screens compared to other methods (von Mering et al., 2002), it may seem strange that *translation* turns out to be the cellular role on which protein interactions predict best. There are however two different properties that influence how well the predictor will work. The first is how well the interactions are detected by the assay used, in this case yeast two-hybrid. But it is equally important how much the members of the cellular role category in reality interact with each other. Given the huge ribosome complex, the category *translation* is probably the most favorable in this respect.

6.3 *In silico* methods for obtaining functional links

A number of methods for predicting protein–protein interactions or functional interactions have already been presented earlier in this thesis (Eisenberg et al., 2000; Marcotte et al., 1999b,a; Pellegrini et al., 1999). Since these methods can be thought of as the main protein function prediction methods competing with ProtFun, it is interesting to compare the performance of them.

To perform the analysis, the Predictome database was downloaded and all predicted functional links in *S. cerevisiae* were extracted and divided among classes (Mellor et al., 2002). In order to not mix up methods of varying performance, an individual set of predicted interactions was made for each computational method. As was anticipated from the original articles, it turned out that the vast majority of predicted interactions were predicted by either the phylogenetic profile method or the Rosetta stone approach. The analysis was thus focused on these two methods.

Proteins fused together are likely to attain an increased mutual affinity. It has even been speculated that fusion proteins might be an evolutionary path to interacting proteins (Marcotte et al., 1999a). Thus it is likely that the Rosetta stone methods for function prediction should more correctly be called a methods for protein–protein interaction prediction—which in turn implies a functional relationship between the proteins.

Although the predictions made by these *in silico* methods are far from perfect, an evaluation on a set of experimentally verified protein–protein interaction revealed their accuracy to be comparable to that of yeast two–hybrid screens but with better coverage (von Mering et al., 2002).

The degree of functional relationship between proteins linked by these methods was of course already investigated by the authors of the original papers. The method they used was in fact very much similar to the analysis presented here. However, rather than assigning proteins to categories and calculating the correlation coefficient within each, the authors have chosen to look at the overlap of SWISS-PROT keywords between interaction partners. Since we assign function using the EUCLID method (which in turn look at SWISS-PROT keywords), the

Table 6.1: **Comparison of the performance of several function prediction methods.** The performance of ProtFun was compared to that of several other function prediction methods. The performance attainable by clustering of microarray expression data was evaluated on the data by Hughes et al. (2000a). The data set by Schwikowski et al. (2000) was used for evaluating the use of yeast two-hybrid interactions.

Functional category	Array	Y2H	Fusion	Phyl.	ProtFun
Amino acid biosynthesis	0.473	0.239	0.531	0.910	0.344
Biosynthesis of cofactors	0.242	-0.012	0.820	0.657	0.308
Cell envelope	0.293	0.394	0.017	-0.027	0.271
Cellular processes	0.427	0.502	0.118	0.950	0.324
Central intermediary metabolism	0.253	0.529	0.210	-0.020	0.341
Energy metabolism	0.449	0.351	0.697	0.922	0.385
Fatty acid metabolism	0.412	0.095	0.009	1.000	0.188
Purines and pyrimidines	0.160	0.315	-0.198	0.390	0.448
Regulatory functions	0.175	0.553	-0.039	-0.001	0.481
Replication and transcription	0.311	0.541	0.339	0.688	0.659
Translation	0.758	0.598	0.198	0.531	0.362
Transport and binding	0.353	0.343	-0.126	-0.001	0.686
Enzyme/non-enzyme	0.303	0.259	-0.079	0.559	0.579
Oxidoreductase	0.352	0.440	0.594	-0.162	0.386
Transferase	0.199	0.346	0.014	0.147	0.489
Hydrolase	0.451	0.327	-0.141	0.673	0.519
Lyase	0.245	-0.016	-0.007	-0.027	0.236
Isomerase	0.179	0.175	-0.016	-0.031	0.202
Ligase	0.148	0.588	0.115	0.337	0.286

two approaches are quite similar. The primary reason for redoing the analysis is that the keyword assay simply cannot be used to evaluate the performance of the ProtFun, which does not link proteins. Also, their approach did not reveal for which type of proteins the methods work and for which it fails. This is quite valuable information to have when using the methods to predict function.

Marcotte and coworkers have also compared the performance of their method to both clustering of array data and protein–protein interaction data. They arrive at somewhat different results than us, as protein–protein interactions and phylogenetic profiles were found to be roughly equally good, while clustering of array data was clearly worst. The comparable ability of *in silico* predictions and yeast two–hybrid to predict function is consistent with their similar accuracy at identifying protein–protein interactions (von Mering et al., 2002).

The much worse performance of expression clustering observed in that study compared to the one presented here, is most likely explained by the choice of data sets. Where the present comparison uses the very comprehensive Rosetta compendium (which was not available in 1999), they had to resort to data sets for the diauxic shift, mitotic cell cycle and sporulation (DeRisi et al., 1997; Spellman et al., 1998; Chu et al., 1998). The conditions under which expression has been measured is obviously crucial to how well clustering proteins will have related functions.

6.4 Comparison of the methods

Based on the results presented in Table 6.1, it can be concluded that ProtFun compares favorably to all of the other methods tested. If looking at the average performance (ROC area) over the 12 cellular role categories, ProtFun is the second best method, only surpassed by phylogenetic profiles. However, the coverage obtained with phylogenetic profiles is less than 10% (data not shown), more than an order of magnitude lower than ProtFun which makes a prediction for every sequence. If focusing on enzyme categories rather than cellular roles, ProtFun has by far the highest average performance of all methods—and of course still gives full coverage in contrast to the other methods.

Compared to high-throughput expression or interaction data analysis, the predictions made by ProtFun give better quality and coverage without the need for doing the experiments. Needless to say, these high throughput methods have other virtues than being usable for predicting function.

It is also interesting to note, that the performance obtained for various functional categories varies much less for ProtFun than for the two types of *in silico* functional links (see Table 6.1). The ProtFun thus seems to be the most generally applicable of the methods, while the others are more likely to excel on some classes of proteins and fail completely on others.

6.5 Comparison with a method for *E. coli*

Although our predictor has been trained on human sequences and is not very well suited for predicting on bacterial protein sequences (see Paper IV), nonetheless an attempt will be made to compare the performance of ProtFun to that of King et al. (2001). At the cellular role level they obtained an specificity (called accuracy in the paper) of 69% and a coverage of 20% when not making use of sequence similarity. This is the performance that should be compared to that of ProtFun.

It is not obvious how this should be done as ProtFun does not assign each protein to one cellular role, but instead assigns a probability for each category. To make matters worse, these probabilities cannot be interpreted as true probabilities when working with organisms where the breakdown on cellular roles is very different from that of *H. sapiens*.

To address these problems, the specificity was plotted as a function of sensitivity for each cellular role category using the labeled *E. coli* data set described in Paper IV. From visual inspection of these plots, a "probability" threshold was selected for each categories for which ProtFun works acceptably on *E. coli* proteins. For categories where the performance was not acceptable, a threshold of 1 was used, i.e. no proteins were predicted to belong to those categories.

Using the selected thresholds, the overall specificity and coverage for the *E. coli* genome was estimated to be 44% and 38% respectively. Compared to the predictor by King et al., ProtFun makes predictions of the function of almost twice as many proteins, but with a much lower specificity. It is worth noting that the specificity of 44% percent is still far better than the 11% obtained by simply guessing on the most common category (*transport and binding*).

Chapter 7

Applying the scheme to other functional classifications

Having convinced ourselves that the ProtFun approach to function prediction is indeed a competitive and a generally applicable one, it is time to collect the fruits by applying the method to other classification schemes and data sets.

7.1 A baseline for protein–protein interaction prediction

The first application of the ProtFun approach which will be briefly mentioned is a quite unusual one, as the method is not applied to classification of individual proteins but pairs of proteins.

In Chapter 6, it was shown that interacting proteins often have similar function and that this can be utilized to predict protein function. As correlations always go both ways, the reverse is of course also true: proteins with similar function are more likely to interact than randomly selected proteins. We can thus expect to predict pairs of interacting proteins from sequence derived features with a performance which is better than a random guess.

Still, the method will mainly rely on the trivial fact that proteins have to be the same place in the cell in order to interact. It does not take into account the three-dimensional structure of the proteins that interact nor the energetics of their interaction. Hence the performance of this method should be regarded as a baseline performance for protein–protein interaction prediction. A method relying on structure should at least perform better than this baseline in order to be interesting.

7.1.1 Creating positive and negative sets for protein interactions

From the Database of Interacting Proteins (DIP) (Xenarios et al., 2001), 2,669 pairwise interactions between *S. cerevisiae* proteins were extracted. The majority of these were identified from high throughput yeast two–hybrid screens. Duplicate

interactions were removed so that interactions were only counted in one direction. This resulted in a set of 1,867 protein pairs which make up the positive set.

As is so often the case in the field of bioinformatics, it is much more difficult to obtain a good quality set of negative examples. Databases of protein–protein interactions simply do not report if two proteins have been shown *not* to interact. A negative set was constructed by a 32-fold downsampling of the pairs of proteins involved in the interactions making up the positive set. Discarding interactions included in the positive set, this resulted in a negative set of 13,879 protein pairs.

7.1.2 Training on pairs of proteins

The protein–protein interaction predictor was trained very much as original ProtFun method, except for two adaptations made to encompass pairs of proteins. Each protein pair was simply represented in feature space by concatenating the feature vectors for the two proteins. In order to ensure symmetry when training the method, all examples were duplicated with the two proteins swapped.

The best feature combination was searched for, using the previously described heuristic (see Paper II), with the modification that the performance of each feature combination was evaluated by five-fold cross validation rather than a single test set.

7.1.3 Performance mainly due to subcellular localization features

Table 7.1 shows the performance of the top five feature combinations as well as the best performing individual features. The correlation coefficient of 0.188 is not good enough for this method to be used in practice for interaction prediction, but that was never the intention either. The purpose is to define a lower limit for the performance of structure based protein–protein interaction predictors.

Table 7.1: **Baseline performance for protein–protein interaction prediction.** Using a ProtFun-inspired approach, protein–protein interaction predictors were trained on yeast two–hybrid data to establish a baseline for other prediction methods.

Feature(s)	Correlation
AI, Natom, Nneg, PSORT	0.188
GRAVY, Natom, Nneg, PSORT	0.183
AI, GRAVY, Nneg, TargetP	0.183
Natom, Nneg, Npos, PSORT	0.182
AI, GRAVY, SignalP	0.178
TargetP	0.146
PSORT	0.143
SignalP	0.136
GRAVY	0.123
SEG	0.116

That the method does indeed represent the baseline described above, can be seen from analyzing the feature usage that underlies the performance. By comparing the performances of the best feature combination and the best single feature (PSORT) it is evident that most of the performance can be attributed to predicted subcellular localization. It is not only the PSORT feature that can provide this information, it can largely be replaced by predictions by either the SignalP or the TargetP method (see Table 7.1).

The baseline performance presented here is likely to be a conservative estimate of the actual correlation between subcellular localization of protein–protein interactions. One reason for this is that the subcellular localization predictors are not perfect. Another more important reason is, that as the positive set is mainly based on yeast two–hybrid experiments it can be expected to contain in the order of 50% false positives (Ito et al., 2001; Mrowka et al., 2001; von Mering et al., 2002). The error rate of the negative set is more difficult to estimate, but could easily be equally large. These large error rates are likely to lower the correlation coefficients obtained.

7.1.4 Too good to be true

Other researches have also tried training methods for predicting protein–protein interactions from sequence alone. Bock and Gough have trained a Support Vector Machine (SVM) using DIP as their positive set in a similar manner to the method described above (Bock and Gough, 2001). In their sequence representation all sequences are normalized to the same length, and each amino acid residue is converted into residue properties such as charge and hydrophobicity. This is similar to our binning schemes but is more fine grained and retains more of the original sequence information, thus possibly allowing their method to recognize conserved domains.

Using this approach, they manage to predict 80% correct on a balanced data set. Given that a specificity in the order of only 50% has been estimated for the type of interaction data they used to create the test and training sets, the reported accuracy is literally too good to be true. The method would have to not only correctly predict which proteins interact, but also which ones are erroneously listed as interaction partners in DIP.

I strongly suspect that the problem lies in the choice of the negative set: lacking a database of non-interacting protein pairs, the authors chose to use shuffled sequences instead. This can give rise to serious problems as the machine learning method may simply learn to discriminate between real and shuffled sequences. Being aware of this problem, the shuffling was performed conserving mono-, di-, as well as tri-residue frequencies. Although the performances archived on these three sets were very similar, it is still possible that the shuffling procedure causes the problem and the the SVMs are merely capable of discriminating real protein sequences from shuffled ones. The proper way to test this is to check if the method predicts the majority of randomly sampled pairs of proteins to be interaction partners.

7.2 The yeast mitotic cell cycle

As *S. cerevisiae* is the best experimentally characterized eukaryote, the second application of ProtFun will also be on yeast proteins. In Chapter 6 it was discussed how well cellular role categories can be predicted from microarray expression studies.

Gene expression patterns are much better correlated to certain other functional classes of proteins. The example that will be discussed here is cell cycle related proteins, many of which will display periodic expression profiles through the mitotic cell cycle.

7.2.1 The need for a sequence based predictor

While many of the genes exhibit periodic expression, this is not sufficient for finding all proteins involved in the mitotic cell cycle. The concentration of mature proteins can also be regulated at the protein level by controlled conversion of precursors into mature proteins as well as by controlled degradation. Also, the activity of the proteins present can be regulated—for instance by reversible phosphorylation. This means that the active concentration of a protein can vary during the cell cycle even if the transcript concentration is constant.

A more technical problem has to do with the detection limit and background noise level of microarray data. Genes expressed at very low copy numbers in the cell cannot be measured reliably using current microarray technologies. Because there is a certain level of background noise on the measurements, the signal-to-noise ratio becomes worse for genes expressed at low levels. A periodicity in the signal can therefore be lost in the noise, making it difficult to identify the cell cycle related proteins that are expressed in low copy numbers. To make matters worse, the primary regulators of the cell cycle are among those proteins.

Making a sequence based predictor of cell cycle proteins is one way to circumvent these problems. Using the periodically expressed genes identified from microarray experiments for training, a ProtFun-like predictor of cell cycle proteins has been developed and subsequently applied to a data set of all *S. cerevisiae* protein sequences.

Paper V

7.3 *In-silico* Proteomics of the Yeast Cell Cycle

Ulrik de Lichtenberg†, Thomas Skøt Jensen†, Lars Juhl Jensen, and Søren Brunak*
Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby, Denmark

† These authors contributed equally
* To whom correspondence should be addressed (email: brunak@cbs.dtu.dk)

Summary

DNA microarrays have been extensively used to identify cell cycle regulated genes in yeast, but there is surprisingly little overlap in the genes identified. We present a characterization of cell cycle proteins and show that certain protein features can be used to distinguish cell cycle regulated genes from other genes (features include protein phosphorylation, glycosylation, subcellular location and instability/degradation). We demonstrate that co-expressed genes encode proteins which share combinations of features, and provide a first *in-silico* view of the proteome dynamics during the cycle. An entirely sequence-based machine-learning method was trained to identify cell cycle proteins in the yeast proteome. The method identified many novel putative cell cycle proteins, most of which presently have unknown function.

Introduction

The eukaryotic cell cycle is a large and complex system regulated at many levels by diverse mechanisms (Mendenhall and Hodge, 1998; Breeden, 2000; Tyers and Jorgensen, 2000). One goal of cell cycle research is to uncover the size and complexity of the underlying molecular system, by identifying all cell cycle genes and proteins (Nurse, 2000). Two DNA microarray studies have recently been performed in *S. cerevisiae* in which the expression level of each gene in the yeast genome was measured during the cell cycle (Cho et al., 1998; Spellman et al.,

1998). These data have been analyzed by visual inspection (Cho et al., 1998), Fourier analysis and correlation to profiles of known cell cycle genes (Spellman et al., 1998), as well as by a single-pulse statistical model (Zhao et al., 2001). Each study proposed a list of periodicly expressed genes based on their analysis of the data. However, pronounced discrepancies exist between these lists of cell cycle regulated genes, as shown in Figure 7.1.

In these studies, 940 genes were proposed to be periodic, yet less than half of them (397) were identified in at least two studies and only 144 genes were found in all three studies. Zhao et al. (2001) analyzed the three cell cycle experiments (synchronized with α-factor, Cdc28 and Cdc15) individually and concluded that 1,088 genes showed significant periodicity in one of the experiments, 260 were periodic in at least two of three and only 71 genes were significant in all three experiments (Zhao et al., 2001). A recent analysis of the data (Shedden and Cooper, 2002) also concludes that reproducibility is poor. Together, these observations demonstrate discrepancies both between conclusions drawn by different research groups (Figure 7.1), and between the three different synchronization experiments analyzed with the same method. Furthermore, our analysis indicates that weakly expressed genes are underrepresented among the proposed periodic genes, probably due to the low signal-to-noise ratio for such genes in DNA microarray experiments.

Our results demonstrate that many cell cycle proteins display correlations between features, which are different from the average yeast protein. These features include phosphorylation, glycosylation, stability and/or disposition for targeted degradation and localization in the cell. Further analysis reveals systematic temporal variations in protein *feature space* during the cell cycle, demonstrating that many co-expressed cell cycle genes encode proteins that share the same features. Each of the four cell cycle phases, G_1, S, G_2 and M, displays different characteristics that can be related to events specific to these phases. Based on the discriminative features as input, neural networks were trained to identify cell cycle proteins in the yeast proteome. A similar feature based classification approach has previously been employed for orphan function prediction of human proteins (Jensen et al., 2002). Our method identifies a large number of new putative cell

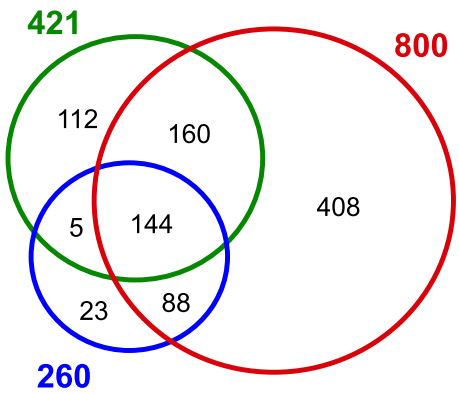


Figure 7.1: **Extent of Agreement between the Published Lists of Periodicly Expressed Transcripts.** Data shown for (Cho et al., 1998) (green, 421 genes), (Spellman et al., 1998) (red, 800 genes) and (Zhao et al., 2001) (blue, 260 genes).

cycle proteins undetected in DNA microarray studies.

Results and discussion

Based on a periodicity analysis of the publicly available microarray data (Spellman et al., 1998), a training set was selected consisting of 97 proteins displaying very significant periodicity in expression during the cell cycle, along with 556 proteins encoded by non-periodic genes. Although *S. cerevisiae* was the first eukaryotic organism to be sequenced (Goffeau et al., 1997), annotations of protein features such as post-translational modifications, subcellular localization, transmembrane helices, etc. are only available for a subset of the proteins encoded in the genome. Our method is therefore based on protein features calculated or predicted from the amino acid sequence by a set of well-documented bioinformatics tools and predictors (Jensen et al., 2002), see Figure 7.2.

Neural networks were then trained to distinguish the cell cycle proteins from the non-cell cycle proteins solely based on their features. Hundreds of protein feature combinations were tested in an iterative fashion, selecting for those combinations that led to the best classification performance (see Experimental Procedures). This heuristic approach resulted in an ensemble of five neural networks integrating different protein features, as depicted in Figure 7.2. The output from individual networks was combined into one final probability score, with high values corresponding to cell cycle proteins. The ensemble obtained a Matthews correlation coefficient in the range of 0.4–0.5, with a very low false positive rate of less than 1% and a sensitivity of 20–30%. This means that high scores are to be taken as strong supporting evidence for a cell cycle role, whereas low scores are less conclusive. In other words, the method will not identify all cell cycle proteins, rather it is suited for finding new putative candidates that may be missed by other techniques. The neural networks detect complex correlations in protein *feature space* characterizing diverse cell cycle proteins that do not necessarily display any similarity in amino acid sequence, three-dimensional structure or gene expression.

Characteristic features of cell cycle proteins

The discriminative features selected by the neural networks (Figure 7.2) provide an interesting characterization of cell cycle proteins as a class. Serine/threonine protein phosphorylation proved very useful for the classification. Our findings indicate that high potential for serine/threonine protein phosphorylation are overrepresented in cell cycle proteins—which is well known and consistent with the involvement of multiple serine/threonine kinases, e.g. the yeast Cdk, Cdc28p, in cell cycle regulation (Mendenhall and Hodge, 1998). The predicted subcellular localization also proved very valuable for the discrimination. Cell cycle proteins appear to be overrepresented in the nuclear and cell wall categories, most likely explained by their involvement in processes such as transcription, DNA replication, repair, chromatin functions, budding and cell wall formation. Other correlations picked up by the neural networks indicate that many cell cycle proteins are unsta-

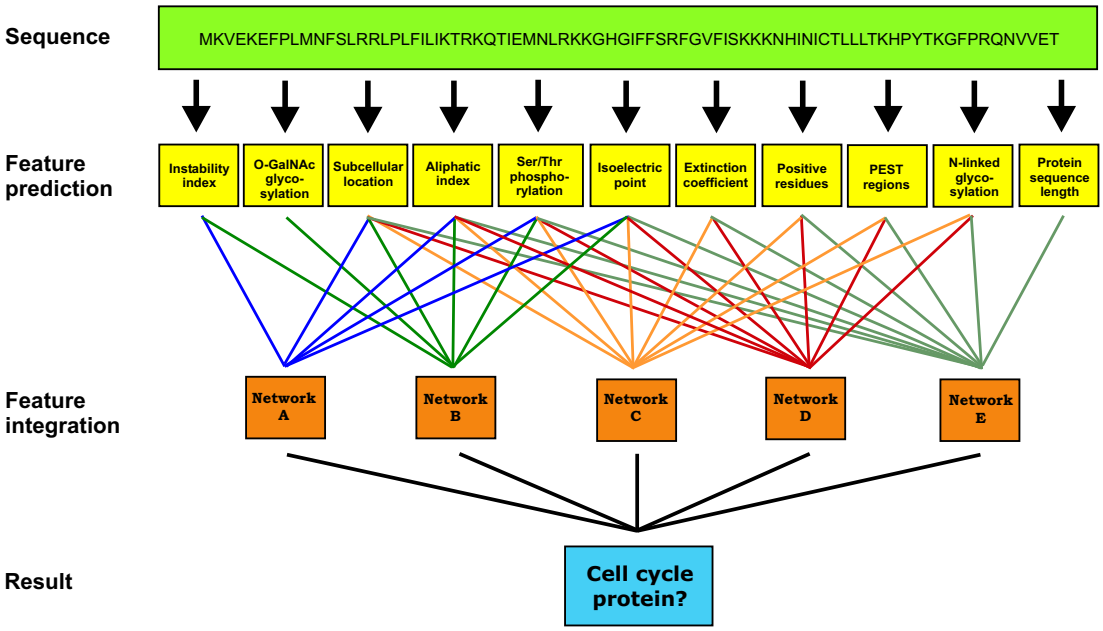


Figure 7.2: **Schematic Illustration of the Neural Network Approach.** Protein features were derived from the amino acid sequence and integrated in different combinations by neural networks. The features selected for best discriminative performance were: Instability index (Guruprasad et al., 1990), O-GalNAc glycosylation (Hansen et al., 1998), subcellular localization (Nakai and Horton, 1999), aliphatic index, serine/threonine phosphorylation (Blom et al., 1999), isoelectric point, extinction coefficient, number of positively charged residues, PEST regions (Rechsteiner and Rogers, 1996), N-linked glycosylation (Ramneek Gupta, unpublished results) and sequence length. The edges in the figure illustrate which features were integrated in each of the five networks that make up the ensemble. The following features were tested and discarded in the process due to their relatively poor discriminative value in input combinations: Tyrosine phosphorylation (Blom et al., 1999), signal peptides (Nielsen et al., 1997a), O-GlcNAc glycosylation (Ramneek Gupta, unpublished results), transmembrane helices (Krogh et al., 2001), hydrophaticity (GRAVY) (Kyte and Doolittle, 1982), amino acid composition and number of negatively charged residues.

ble (have a high instability index) and/or contain so-called PEST regions in their amino acid sequence—regions known to be recognized by ubiquitin ligating complexes as such the *Anaphase Promoting Complex/cyclosome* (APC/C) and the Skp1p-Cdc53p/Cullin-F-Box protein complexes (SCF) that target numerous cell cycle proteins for degradation by the proteasome (Tyers and Jorgensen, 2000). Many cell cycle regulated proteins appear to have high potentials for N-linked glycosylation—a post-translational modification found almost exclusively in secreted or extracellular proteins—suggesting that many of these proteins could be related to budding and cell wall formation. Although of less discriminative value, cell cycle proteins seem to contain, on average, more positively charged residues. All in all, these findings are consistent with existing knowledge that phosphorylation, localization and degradation are key regulatory mechanisms of the cell cycle, see e.g. Mendenhall and Hodge (1998); Tyers and Jorgensen (2000); Breeden (2000).

Proteome-wide prediction of cell cycle proteins

The trained neural network prediction method was applied to the entire *S. cerevisiae* proteome, to identify new putative cell cycle regulated proteins. Among the highest scoring 250 proteins (not used when training the method) were 75 previously suggested to be periodic in at least one of the three DNA microarray studies (Cho et al., 1998; Spellman et al., 1998; Zhao et al., 2001), along with 175 new putative cell cycle proteins. There is large diversity in both annotated function and subcellular localization of these proteins, which include nuclear, cell wall, membrane, cytoplasmic and cytoskeletal proteins. However, most of these potential cell cycle proteins have no known function, role or subcellular location, suggesting a high potential for new biological discoveries. A list of the top scoring 500 proteins is available from the website: `http://www.cbs.dtu.dk/cellcycle`.

The highest scoring of all proteins in the *S. cerevisiae* proteome is encoded by the gene *YIL169C*. It has no known function and was only identified as periodic in one of three microarray studies (Cho et al., 1998). The protein was earlier reported to interact with the two known cell cycle proteins, Mob1p and Fus3p (Ito et al., 2001). Mob1p is required for cytokinesis and mitotic exit (Luca et al., 2001), whereas the *FUS3* gene encodes a MAP Serine/Threonine kinase. Recent genome-wide location data suggest the promoter region of *YIL169C* to be associated with at least one known cell cycle transcriptional activator, Fkh2p, possibly also Ndd1p and Fkh1p (Simon et al., 2001). Taken together, these data support a cell cycle role for *YIL169C*, as suggested by the neural network ensemble. Prediction of phosphorylation sites (Blom et al., 1999) indicates Yil169p to be heavily phosphorylated on serine and threonine residues, making it a putative substrate of Fus3p. Also, the sequence is predicted (Rechsteiner and Rogers, 1996) to contain several PEST regions, indicative of a high potential for targeted degradation. The neural network ensemble suggests many other candidates for which we have been able to find supporting evidence, demonstrating the strength of the sequence-based approach in identifying new potential cell cycle proteins.

Some of the high scoring proteins are encoded by genes with low intensities in the microarray studies, indicating poor hybridization or weak expression. Figure 7.3 shows an underrepresentation of weakly expressed genes in all three sets of microarray identified periodic genes. Among the weakest 15% of the genes (with respect to signal intensity on the microarray chips) there also appears to be a correlation between the number of periodic genes proposed and the fraction of weakly expressed genes, e.g. among the 260 genes found significantly periodic by Zhao et al. (2001) only 7.4% fall within the 15% intensity fractile. In comparison, 18.2% of the genes identified by our method fall within the lowest 15% of the intensity distribution. Our method may therefore identify cell cycle proteins previously undetected on microarrays, due to the poor signal-to-noise ratio. The majority of these genes have no known function or cellular role.

Table 7.2 shows a selection of new putative cell cycle proteins suggested by the neural network ensemble, all of which are unidentified as periodic in microarray studies and encoded by weakly expressed genes (among the 15% signal intensity fractile). The last column in Table 7.2 lists additional supporting evidence, either interaction with other proteins or binding of known cell cycle transcription

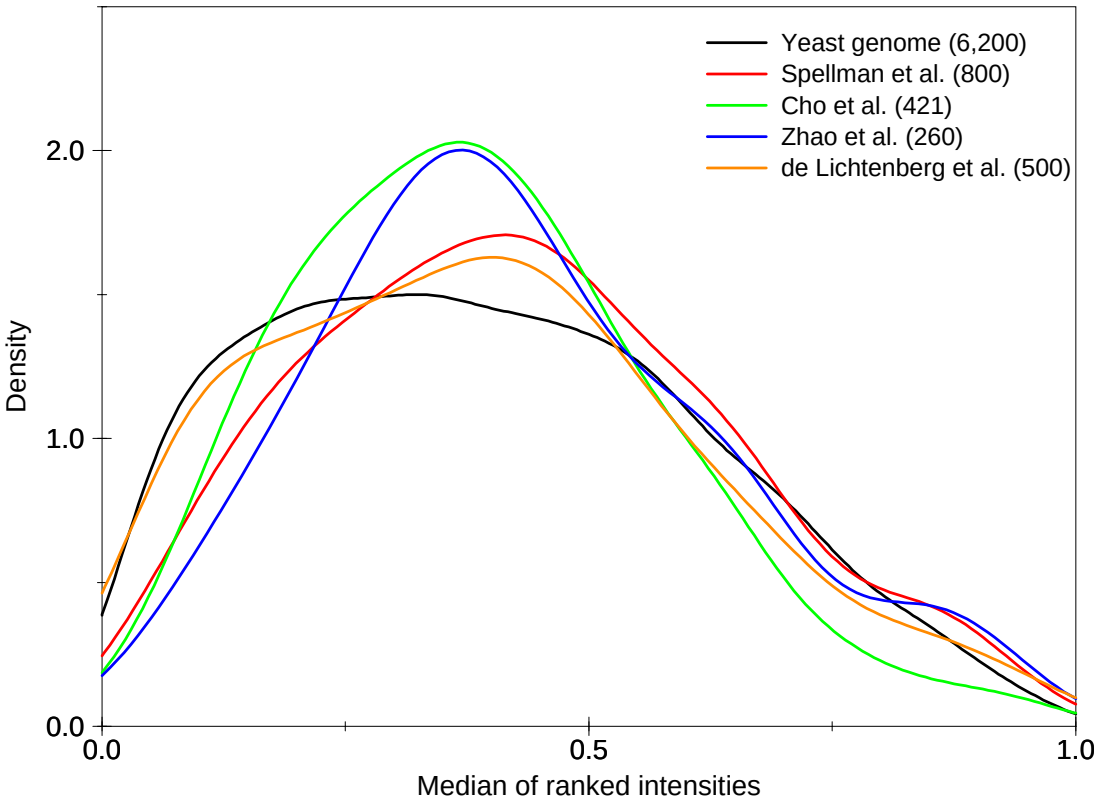


Figure 7.3: **Average Intensity Distributions of Selected Gene Sets.** The figure shows smoothed distributions of *average intensity* for the entire *S. cerevisiae* genome, the cell cycle genes proposed by Spellman et al. (1998), by Cho et al. (1998), by Zhao et al. (2001) and the 500 top-scoring genes from the neural network approach. The average fluorescence intensity was computed for each gene in each of the three cell cycle time series experiments (Cho et al., 1998; Spellman et al., 1998)). Within each experiment, the average intensities were converted into fractiles. The median fractile over all three experiments was then used as measure of the *average intensity* of each gene. For more details, see our website.

factors to the promoter region of the gene. For example, the protein encoded by *YPL070W*, which has no known function, but reported interactions with the cyclin-dependent kinase, Cdc28p, and with two proteins Nip29p (Spc29p) and Tem1p of the yeast spindle pole body (SPB). The gene product is predicted to be heavily phosphorylated, especially on serine residues. Ypl070p may thus be a potential substrate of Cdc28p and may play a role in the cell cycle in relation to the SPB.

Temporal variation of the cell cycle proteome subset

To assess temporal variations in *protein feature space* during the cell cycle, we mapped the 500 highest scoring proteins from the prediction to time points in the cell cycle, based on the time of maximal expression of their encoding genes in the three microarray experiments (Spellman et al., 1998). Of these genes, 309 could be confidently assigned a time of maximal expression. A circular plot was constructed (Figure 7.4) where each circle corresponds to a particular feature, with one color for time points where the proteins expressed have higher values

Table 7.2: **Weakly Expressed Putative Cell Cycle Proteins.** *S. cerevisiae* proteins suggested by the neural network ensemble, all of which appear non-periodic and weakly expressed (among the 15% intensity fractile, see Figure 7.3) in DNA microarray experiments (Cho et al., 1998; Spellman et al., 1998; Zhao et al., 2001). Data for promoter association with known cell cycle transcriptional activators were taken from the publicly available data by Simon et al. (2001). Protein–protein interaction data refers to 1) Uetz et al. (2000), 2) Ito et al. (2001), 3) Tong et al. (2002), 4) Bertram et al. (2002), 5) Bender et al. (1996), 6) Ho et al. (2002) and 7) Deane et al. (2002). Interactions with proteins previously identified as periodicly expressed are marked in bold.

Protein	Function	Selected physical interactions
Tri1p	Unknown	**Rad51p**[1], Top1p[1,2], Pet11p[1]
Yfr022p	Unknown	**Yor264p**[2]
Ylr312p	Unknown	**Mpd2p**[1], Nup116p[2]
Ypl070p	Unknown	Cdc28p[1], **Tem1p**[2], **Nip29p**[2]
Yjl215p	Unknown	**Yor264p**[2], Gcn3p[2]
Yil092p	Unknown	**Cpr8p**[2]
Gln3p	Transcription factor	**Sho1p**[3], Gts1p[1], Tor1/2p[4]
Npr2p	Nitrogen permease regulator	**Bbp1p**[2], Yor138p[1]
Mus81p	DNA repair	**Clb2p**[1], Rad53p[6].
Boi1p	Bem1p-binding protein	**Bem1p**[5], **Hof1p**[3], **Zds2p**[1,3]

Protein	Function	Promoter associated with
Ygl007p	Unknown	Fkh2p, Ndd1p, Mcm1p, Ace2p, Swi4p
Ynl269p	Unknown	Swi5p, Ace2p, Swi6p
Ydr287p	Unknown	Mcm1p
Ypr115p	Unknown	Ace2p, Mbp1p
Ydl187p	Unknown	Swi4p
Yjl160p	Unknown	Swi5p, Mcm1p, Ace2p
Tos3p	Ser/Thr protein kinase	Swi4p, Swi6p
Pcl10p	Cyclin	Swi6p

of this feature than the cell cycle average, and another color (usually dark) for values lower than the cell cycle average. The strength of a particular feature (e.g. isoelectric point) was computed at all time points during the cell cycle by averaging over proteins whose encoding genes peak in the neighborhood of the particular time point. This yielded a division of the cell cycle into a circular "clock", with each time point corresponding to one percent of the cell cycle. Zero time was set to be the presumed position of G_1 phase entry, right after cell division.

Figure 7.4 thus offers a novel *in-silico* view of the cell cycle proteome dynamics and reveals intriguing temporal variations in characteristic features of these proteins during the cell cycle. The time resolution of this "clock" is much higher than the conventional division of the cycle into four phases (G_1, S, G_2 and M) depicted inside the feature circles of Figure 7.4. Our results demonstrate that genes maximally expressed at the same stage in the cell cycle appear to share features at the proteome level. The patterns observed in Figure 7.4 were largely conserved in similar plots representing the sets of periodic cell cycle proteins previously identified in the microarray studies (Cho et al., 1998; Spellman et al., 1998; Zhao et al., 2001) (data not shown), suggesting that the feature patterns of Figure 7.4 may be considered representative of the entire yeast cell cycle proteome. Also depicted in Figure 7.4 are known cell cycle transcriptional activators (marked in blue) positioned at the time where they are reported to function (Simon et al.,

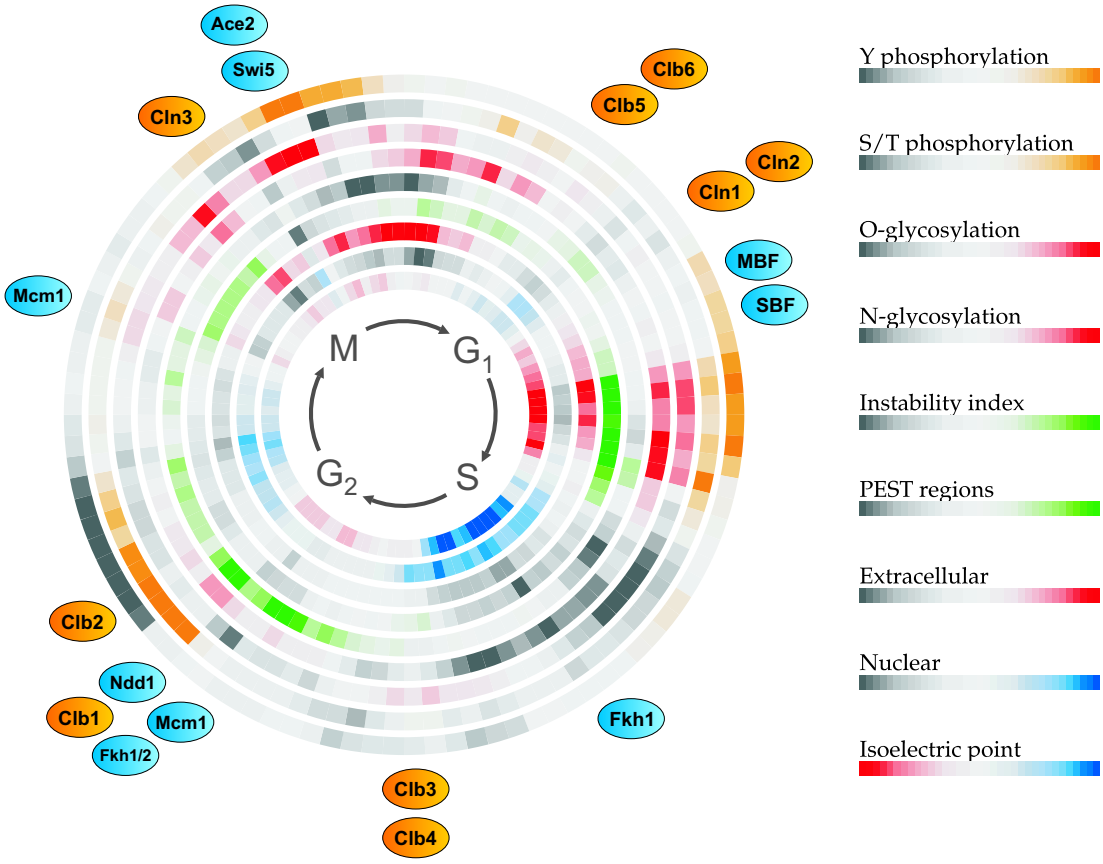


Figure 7.4: ***In-silico* Proteome Dynamics during the Cell Cycle.** Shows the temporal variation in 9 selected protein features during the cell cycle, with zero time (at the top of the plot) corresponding to the presumed time of cell devision (M/G_1 transition). The color scales correspond to ± two standard deviations from the cell cycle average. From inner to outer the circles correspond to: isoelectric point, nuclear and extracellular localization predictions (Nakai and Horton, 1999), PEST regions (Rechsteiner and Rogers, 1996), instability index (Guruprasad et al., 1990), N-linked glycosylation potential, O-GalNAc glycosylation potential (Hansen et al., 1998), serine/threonine- and tyrosine phosphorylation potential (Blom et al., 1999). Inside the circles are marked the presumed positions of the four cell cycle phases: G_1, S, G_2 and M. Also depicted are known cell cycle transcriptional activators (marked in blue), positioned at the time where they are reported to function (Simon et al., 2001), along with nine cyclins (marked in orange), placed at the time where their genes are maximally expressed. Most of the cyclins are believed to activate Cdc28 kinase activity when expressed, but it should be noted that Clb5p and Clb6p are kept inactive in G_1 phase by the inhibitor protein Sic1p (Mendenhall and Hodge, 1998; Breeden, 2000).

2001), along with nine cyclins (marked in orange) placed at the time where their genes are maximally expressed.

Figure 7.4 displays a complex pattern of up and down regulated features right at the top of the plot, at the suspected time of late M phase, cytokinesis and cell devision, where the cell completes the cycle. This pattern is followed by a large uniformly colored area in G_1, indicating that no features are over- or under-represented among the proteins expressed here. This is taken over by a very distinct feature pattern late in G_1, where proteins have a large number of features in common, particularly post-translational modifications and PEST regions. Most of these features display the opposite pattern upon entry into what

we believe to be the S phase, indicating the S phase proteins to be preferentially nuclear, characteristic of high isoelectric points, fewer PEST regions and lower potentials for glycosylation. Only few features stand out in the presumed G_2 phase, where proteins appear to have high instability index, indicating short life time. Almost as a burst, we observed the potential for serine and threonine phosphorylation to be significantly higher in a small window in G_2, where the tyrosine phosphorylation potential is at the same time very low. This pattern coincides with the reported activity of several transcriptional activators and the maximal expression the two cyclins genes *CLB1* and *CLB2*. The last differential feature pattern is seen towards the end of the cell cycle in M phase and at the transition to the next cycle.

Protein phosphorylation during the cell cycle

The changing patterns of phosphorylation are particularly interesting. Although only serine/threonine phosphorylation was found of discriminative value for identifying cell cycle proteins, we observe significant temporal variations in both tyrosine and serine/threonine phosphorylation at three stages in the cell cycle—all with a different correlation between the two types (Figure 7.4). Proteins mapped to time points 20–30% into the cell cycle have high potentials for both kinds of phosphorylation with the tyrosine potential rising first. The next differential phosphorylation pattern is seen 60–70% into the cell cycle, where proteins have high potentials for serine/threonine phosphorylation, but very low potentials for tyrosine phosphorylation. Towards the end of the cell cycle, before cell division, tyrosine phosphorylation stand out again, whereas the serine/threonine phosphorylation reaches its lowest level. These observations are at least in part consistent with previous biological observations that the activity of the cyclin dependent kinases rises during the cell cycle from the G_1/S transition (START) until the end of Mitosis, where it drops due to activity of inhibitors and targeted degradation of the cyclins, see e.g. Mendenhall and Hodge (1998); Lew et al. (1997). However, our results surprisingly suggest tyrosine phosphorylation to be significantly more abundant among proteins expressed both at the suspected G_1/S transition and towards the end of the cell cycle in mitosis. It may be that serine/threonine phosphorylation serves as the primary "engine" of the cell cycle, and that tyrosine phosphorylation is only involved in the initial and terminal stages of the cycle.

The G_1/S transition (START)

The rise in tyrosine phosphorylation potential 17% into the cell cycle, that reaches a local maximum at 20–30% together with the serine/threonine phosphorylation, correlates well with the maximal expression of several cyclins (*CLN1, CLN2, CLB5, CLB6*), with the expression of genes involved in budding (*BUD9, BNI4, GIN4, CSI2, CRH1, AXL2, SVS1, QRI1, MCD4, RSR1, MSB2, MNN1*) and DNA replication/repair (*PRI2, DPB2, POL12, POL30, DBP2, CDC9, CDC21, RAD53, MSH6, RFA1, RAD27, RNR1, SLD2, CTF18, TOF1, RFA2, OGG1, CDC45, HYS2, MSH2, POL2*). Together, these data suggest the *feature space* pattern 20–30% into the cell cycle to be a fingerprint of the G_1/S transition

Table 7.3: **Neural Network Identified G_1/S Proteins.** Protein identified by the neural network ensemble which are maximally expressed 20–30% into the cell cycle, at the presumed G_1/S transition. See Table 7.2 for references on protein–protein interactions and promoter association with known cell cycle transcriptional activators. MBF (a complex of Mbp1p and Swi6p) is believed to predominantly activate genes involved in DNA replication and repair, whereas SBF (a complex of Swi4p and Swi6p) activate genes related to budding and cell wall formation (Simon et al., 2001).

Protein	Selected physical interactions	Promoter	Function
Csi2p			Chitin synthesis
Msb2p	Bni4p[7], Cla4p[1]	SBF	Bud site selection
Rsr1p	Bem1/4p[1], Cdc24p[1], Sec15p[1]		Bud site selection
Axl2p			Bud site selection
Crh1p	Arr4p[2], Lys14p[2], Ynl092p[2]	SBF, Swi5p	Cell wall protein
Tos6p	Nup116p[2]	SBF	Cell wall protein, probable
Slg1p	Rom2p[7]		Cell wall integrity
Mpd1p	Ypr085p[2]		Disulfide isomerase
Scw10p	Sua7p[2]	SBF	Putative Glucanase
Yps1p			Aspartyl Protease
Spa2p	Bni1p[7], Ste7/11p[7], Msb3/4p[7]		Cell polarity, budding
Svs1p		SBF	Ser/Thr rich protein
Rad53p	Asf1p[7], Rad9p[7], Dbf4p[7]		Cell cycle checkpoint kinase
Asf1p	Rad53p[7], Hht1p[7]		Component of RCAF
Mcd1p	Cdc5p[7], Irr1p[7], Scc2p[7], Smc1/3p[7]	MBF	Cohesin
Sld2p	Dpb11p[7]		DNA replication
Zds2p	Bni1p[7], Sir1/2/4p[7], Cdc11p[7], Rho1p[7]		Transcriptional silencing
Opy2p		Swi6p	Probably cell cycle regulation
Yer028p	Ybr061p[2]		Contains zinc finger domains
Gpi16p	Arf1/2p[6], Pho89p[2], Yjr015p[1]		GPI-anchor transamidase
Pry2p		SBF	Function unknown
Tos2p	Cdc24p[7], Pkc1p[7]	SBF	Function unknown
Tos4p	Ylr037p[2]		Function unknown
Prm5p	Bud2p[1], Msn1p[2], Atp14p[2]		Function unknown
Erp3p		Mbp1p	Function unknown
Srl3p	Mec3p[2], Stb1p[2]		Function unknown
Ybr089p	Rpb9p[2]		Function unknown
Ydl211p	Ara1p[2]		Function unknown
Yil025p	Rpc10p[2], Soh1p[2]		Function unknown
Yil141p	Rrn10p[2], Mcm21p[2], Srb4p[2]		Function unknown
Yjl028p			Function unknown

(termed *START* in yeast) evident at the proteome level. Figure 7.4 also indicates that this group of G_1/S proteins contains many members with PEST regions, low isoelectric point, high potential for phosphorylation and glycosylation, and the group is also predicted to be rich in extracellular proteins or proteins related to the cell wall. The latter correlate well with proteins involved in budding and cell wall formation. Table 7.3 shows the neural network identified G_1/S proteins, their interactions, activators and annotated function.

The G_2 phase

Another interesting phosphorylation pattern is the burst in serine/threonine phosphorylation seen 60–70% into the cell cycle (Figure 7.4). It coincides with the maximal expression of the two cyclins Clb1p and Clb2p, and with the reported function of the transcriptional activators Mcm1p, Fkh1p, Fkh2p and Ndd1p that activate the transcription of G_2/M genes (Breeden, 2000; Simon et al., 2001). Interestingly, our neural network method identifies a number of serine/threonine

kinases whose genes are maximally expressed before this burst in phosphorylation potential, namely Ymr291p, Elm1p, Mps1p, Kin4p and Skm1p. Of these, only Elm1p is identified as a periodic cell cycle gene in microarray studies (Cho et al., 1998; Spellman et al., 1998; Zhao et al., 2001). Also expressed right before are the two cyclins Clb3p and Clb4p that activate Cdc28p, along with two other cyclins, Pcl7p and Pcl10p, that associate with another yeast Cdk, Pho85p. The proteins with increased phosphorylation potential 60–70% into the cycle may thus be potential substrates of these kinases and Cdks.

The two forkhead transcription factor genes FKH1 and FKH2 are maximally expressed around 50% into the cell cycle, in late S or early G_2 phase. They are known to promote transcription of a large number of cell cycle genes in $G2/M$ (Simon et al., 2001; Breeden, 2000). Both contain a protein domain, the Forkhead-association domain (FHA), demonstrated to specifically recognize and bind phosphothreonine epitopes on proteins (Durocher and Jackson, 2002). Such domains are also present in the DNA damage checkpoint protein Rad53p and in the protein Tos4p, both of which are recognized by the neural network ensemble and found to peak at the presumed G_1/S transition (see Table 7.3). Interestingly, the neural network ensemble also identifies a protein, encoded by the gene YDR200C, which in one of the cell cycle experiments (the $Cdc28$ arrest) displays a cyclic pattern of expression similar to that of FKH1/2, and which is also reported to contain an FHA domain (Letunic et al., 2002; Durocher and Jackson, 2002). It has no known function, but has reported interactions with another FHA containing protein of unknown function, YLR238W, with Far3p, which plays a role in pheromone-mediated cell cycle arrest and with a protein of unknown function, YNL127W, that shows weak similarity to Fus2p—a protein involved in cell fusion during mating. YDR200C only appears periodic in the synchronization experiment performed by Cho et al. (1998) and was consequently not included in any of the published lists of periodicly expressed genes (see Figure 7.1 and (Cho et al., 1998; Spellman et al., 1998; Zhao et al., 2001)). However, taken together, the data reviewed above may indicate a cell cycle role for the protein, as suggested by our neural network ensemble.

The S phase

As seen in Figure 7.4, the late G_1 pattern changes completely 35–45% into the cycle caused by the expression of a new group of proteins characteristic of being mostly nuclear, having very high isoelectric points, being very stable (low instability index and few PEST regions) and displaying lower potential for glycosylation and phosphorylation than the average yeast cell cycle protein. Among the proteins expressed here are eight histones (Hht1p, Htb1p, Htb2p, Hhf1p, Hhf2p, Hht2p, Hta1p and Hho1p) supporting the notion that this pattern in feature space corresponds to the S-phase of the cell cycle. Histones are stable, nuclear proteins with high isoelectric point and no potential for glycosylation. It might be suspected that these histones dominate the picture seen in the S phase part of Figure 7.4, masking a greater diversity in features among the other proteins expressed here. Since the neural network identification of cell cycle proteins is completely independent of the DNA microarray data used to map the proteins,

it would not be expected by chance that genes peaking at a specific phase encode proteins being similar in *feature space*. It turns out, however, that the majority of the non-histone proteins expressed 35–45% into the cell cycle share the same features that stand out in Figure 7.4, e.g. genes such as *IRS4, SHE1, TOF2, ENT4, YPL150W* and *YNR014W* all encode proteins with high isoelectric point (all above 8.0) and predicted to be nuclear. Most of these are previously unidentified as cell cycle proteins.

Our work has identified a set of protein features characteristic of cell cycle proteins and provides a first *in-silico* view into the temporal dynamics of these key features of the cell cycle proteome. Also, our neural network based method identifies a large number of new potential cell cycle proteins. Our results are largely supported by existing knowledge of the cell cycle and by other sources of experimental data. We hope that our research may inspire future experimental work to establish the validity of the hypothesizes that arise from our analysis.

Experimental procedures

Training set

A periodicity analysis was performed on the three publicly available synchronization experiments (α-factor, $Cdc28$ and $Cdc15$) compiled by Spellman et al. (1998), to identify periodicly as well as non-periodicly expressed genes in *S. cerevisiae*. A Fourier like analysis was applied to the data, such that each gene i was assigned a score D_i based on its temporal expression profile during the experiment, with cell cycle frequency $\omega = \frac{2\pi}{T}$:

$$D_i = \sqrt{(\sum_{\mathrm{t}} sin(\omega t) x_i(t))^2 + (\sum_{\mathrm{t}} cos(\omega t) x_i(t))^2}$$

The cell cycle periods, T, estimated by Zhao et al. (2001) were used (58 min for α experiment, 115 min for $cdc15$ experiment and 85 min for the $cdc28$ experiment), and a combined Fourier score, F_i, was computed as:

$$F_i = \frac{(D_{i,\alpha} + 0.8 \cdot D_{i,cdc15} + D_{i,cdc28})}{3}$$

The contribution from the $cdc15$ experiment was scaled in the combined score, because this experiment covers 2.5 cell cycles, whereas the α and $cdc28$ experiments cover only two (using the Zhao et al. (2001) estimates). The lowest scoring 556 genes (thresholding at 0.75) were used as examples of "non cell cycle proteins", which display no periodic regulation during the cell cycle. By an estimation method described in detail at our website (`http://www.cbs.dtu.dk/cellcycle`), we established a conservative threshold at 6.0 to include 115 genes in a set of very significantly periodic genes. To ensure consistent behavior over multiple cycle, we required the Pearson correlation between the expression profiles of the first and the second cycle to be above 0.4, thereby excluding 18 genes. The procedures outlined above resulted in a training set consisting of 97 "cell cycle proteins" and 556 "non cell cycle proteins". This set was used to train neural networks and is available at our website.

Neural network training

Three-fold cross validation was used (division of the data set in three different ways), each with 430 protein sequences for training and 215 for independent evaluation of the classification performance, which was measured as the Matthews test correlation coefficient over all three test sets. As input to neural networks we used protein features derived directly from the amino acid sequence of the proteins to give a *feature space* representation of each protein. An iterative heuristic was applied to select for the most discriminative features and the best performing combinations of these. Features that proved most discriminative in combinations of two were used to construct new combinations of three features, from which the best was selected to form combinations of four, etc. Figure 7.2 shows the five best input combinations, which were combined into an ensemble of neural networks. To bring the individual network output into comparable format, the distribution of test set scores was ranked and used as conversion table for output from that network, making it possible to simply average all 15 neural network output scores into one final score (between 0 and 1). Information on the detailed classification performance of the neural network ensemble can be found at our website. A prediction was obtained from the ensemble for the entire *S. cerevisiae* proteome (set of all translated ORFs from SGD, `http://genome-www.stanford.edu/Saccharomyces/`).

In-silico feature proteomics

The three publicly available cell cycle experiments (α-factor, *cdc*28 and *cdc*15) were used to determine the time of maximal expression of the identified cell cycle genes. The time series data was normalized within each experiment with the cycling times estimated by Zhao et al. (2001) to bring the data on a comparable time scale. Within each experiment, the time of maximal expression was compared between two consecutive cycles, averaging the two time points if they deviate less than 20% of the cell cycle. Experiments that met this self-consistency criteria were compared and subjected to the same criteria. This way, an average *peak time* was computed for the self-consistent genes (based on one, two or three experiments). The three experiments were aligned by comparing the distribution of peak times for genes known to peak in the G_1-phase (Spellman et al., 1998), and furthermore shifted to set zero time to the suspected time of cell devision (early G_1). The average *peak time* thus indicates how many percent into the cell cycle a given cell cycle protein is maximally expressed. Out of the neural network ensembles 500 top-scoring proteins, 309 met the self-consistency criteria and were assigned an average *peak time*.

The cell cycle was divided into 100 time points and the average value of a given feature was calculated in each of time point by averaging over the proteins expressed in a window of $\pm$ 5 time points. The average feature values were visualized with respect to their deviation from the average value of all 309 cell cycle proteins, using one color for values higher than the average and another color for lower values. The extremes of the color scale was set at $\pm$ two standard deviations. The nine most interesting features were combined into the *in-silico*

proteome "clock" shown in Figure 7.4.

Acknowledgments

The authors would like to thank Jiri Bartek for valuable comments on this work, as well as Hans Henrik Stærfeldt for developing the *Genewiz* software used to construct Figure 7.4. This work was supported by grants from Novo Nordisk and the Danish National Research Foundation.

7.3.1 DNA binding proteins

The histones discussed in Paper V are a good example of how the ProtFun approach works. One of the characteristics of the histone proteins are their high isoelectric points, which cause the histones to be positively charged in the cell. This is indeed a key property for proteins that bind constitutively to DNA, as it is negatively charged. This property can thus be expected to hold for bacterial proteins involved in DNA compaction as well, although they are not believed to be evolutionarily related to histones. This is at least true for the FIS protein from *E. coli*.

7.3.2 Predicting human cell cycle proteins

One of the most intriguing aspects of this predictor is, that it might be able to predict cell cycle related proteins in the human genome. Although it has not been verified yet, it is not unreasonable to expect the method to generalize to all eukaryotes given the cross-species comparison of ProtFun presented in Paper IV. Considering the close biological relation between cancer and regulation of the cell cycle, this would be a very interesting application of the method described in Paper V.

7.4 An archaeal enzyme predictor

From the cross species testing of ProtFun (see Paper IV), it is clear that although ProtFun works surprisingly well on prokaryotes, the performance is worse than for eukaryotes. A very important factor for the breakdown of protein subcellular localization prediction, which obviously makes no sense in a prokaryotic cell. In addition to these features there might be others, which do not have the same functional implications in prokaryotes as in eukaryotes. There is thus reason to believe, that better predictors for prokaryotic proteins can be obtained by retraining the method.

Because of the more limited selection of meaningful predicted protein features when working with prokaryotes, it can be expected to be difficult to predict cellular roles, as these predictors tend to rely heavily of PTMs and localization information. Enzyme/non-enzyme and enzyme classes predictors on the other hand tend to rely on a much broader set of features, hereunder protein structure features. As these have been shown to generalize well to prokaryotes (see Paper IV), making new predictors for these classes seems more feasible.

We have developed such a predictor for archaea (see Paper VI). The motivation for this has been the past few years intense research in using enzymes from hyperthermophilic archaea for industrial purposes. As many industrial processes run at fairly high temperatures, there has been a constant quest for finding or developing more thermostable enzymes. Looking for usable enzymes in hyperthermophilic archaea has been one of the most successful approaches—it has however been hampered by a lack of knowledge of archaea, which have not been studied nearly as much as bacteria and eukaryotes.

7.4.1 More proteins of unknown function in Archaea

Compared to bacteria very little experimental work has been done on archaea relative to the number of genomes that have been sequenced. For this reason much less is known about archaea and the quality of annotation of their genomes is correspondingly lower.

The fraction of genes of unknown function in the genomes reflect this fact, although one should be very careful not simply looking at the fraction of annotated genes that have been assigned a function in the GenBank entry. Again the problem is inconsistent annotation. In Paper I, it was shown that archeal genomes in general tend to be more overannotated than their bacterial counterparts. Since random ORFs cannot be assigned to a functional category this will automatically lead to the higher fraction of unknown proteins claimed in the whole genome papers of archaea (Kawarabayasi et al., 1999; Fitz-Gibbon et al., 2002; She et al., 2001; Kawarabayasi et al., 2001; Klenk et al., 1997; Smith et al., 1997; Bult et al., 1996; Deppenmeier et al., 2002; Slesarev et al., 2002; Galagan et al., 2002; Kawarabayasi et al., 1998; Ruepp et al., 2000; Kawashima et al., 2000).

An estimate for the fraction of genes in each genome that are of known function can be seen in Table 7.4. Here the number of annotated protein coding sequences that can be assigned to a cellular role by the EUCLID method (see Paper IV) is compared to the total number of protein coding genes estimated for the genome by the SWISS-PROT method (see Paper I). While there are several possible sources of error on both these numbers, resulting in a possibly inaccurate estimate that could be biased in either direction, these estimates are at least systematic and can be compared across species.

Looking at Table 7.4 it is realized that for most archeal genomes, 55–60% of the proteins estimated to be encoded can be assigned a cellular role by EUCLID. This figure should be compared to 60–75% that can typically be assigned for bacterial genomes. The reliability of these estimates is supported by the fact that the experimentally best characterized classes of bacteria fall in the high end of the interval, whereas the bacteria with the lowest fractions of known proteins are *D. radiodurans* and *M. tuberculosis*.

Among archaea, *M. kandleri* stand out by having only 44% of its genes assigned to cellular roles. The genome turns out to contain two big regions, constituting about a quarter of the genome, which contain essentially no genes of known function.

7.4.2 Creating the data set

As very few Archaeal proteins have been experimentally verified and characterized, it was clear that we could not obtain a sufficiently large data set of sequence to be able to train a ProtFun-style predictor. Instead we opted for a very large automatically generated data set, accepting that we would pay the price of having a higher error rate. In connection with Paper IV such data sets had already been generated for all available archaeal genomes where proteins were assigned as enzymes/non-enzymes and the enzymes were labeled with the major enzyme

Table 7.4: **Comparing the number of estimated protein coding genes to the number of genes that can be assigned to a cellular role.** The actual number of genes in each genome was estimated using the SWISS-PROT method described in Paper I. Proteins were automatically assigned to functional classes by the method described in Paper IV. As should be expected, there are more proteins of unknown function in extremophiles than in well studied bacteria.

Organism	No. genes estimated	Cellular roles		Enzyme/non-enzyme	
		No. genes assigned	% of estimate	No. genes assigned	% of estimate
A. pernix (Kawarabayasi et al., 1999)	1,376	684	50	688	50
P. aerophilum (Fitz-Gibbon et al., 2002)	1,706	867	51	863	51
S. solfataricus (She et al., 2001)	2,288	1,186	52	1,179	51
S. tokodaii (Kawarabayasi et al., 2001)	2,035	1,045	51	1,041	51
A. fulgidus (Klenk et al., 1997)	1,818	1,074	58	1,053	59
M. thermoautotrophicum (Smith et al., 1997)	1,466	867	60	878	59
M. jannaschii (Bult et al., 1996)	1,350	781	58	780	58
M. mazei (Deppenmeier et al., 2002)	2,686	1,420	53	1,440	54
M. kandleri (Slesarev et al., 2002)	1,477	653	44	658	45
M. acetivorans (Galagan et al., 2002)	3,456	1,850	54	1,827	53
P. abyssi (unpublished)	1,497	855	58	861	57
P. horikoshii (Kawarabayasi et al., 1998)	1,448	786	54	781	54
T. acidophilum (Ruepp et al., 2000)	1,250	783	63	791	63
T. volcanium (Kawashima et al., 2000)	1,243	792	64	778	62
Anabaena sp. (Kaneko et al., 2001)	4,020	2,444	61	2,425	60
A. aeolicus (Deckert et al., 1998)	1,337	926	73	983	69
B. burgdorferi (Fraser et al., 1997)	756	461	67	504	61
B. halodurans (Takami et al., 2000)	3,220	2,223	69	2,264	70
B. subtilis (Kunst et al., 1997)	3,263	2,240	72	2,351	69
Buchnera sp. (Shigenobu et al., 2000)	556	469	84	513	92
C. jejuni (Parkhill et al., 2000b)	1,420	975	71	1,015	69
C. pneumonicae (Kalman et al., 1999)	903	530	61	553	59
C. trachomatis (Stephens et al., 1998)	772	498	69	532	65
D. radiodurans (White et al., 1999)	2,323	1,332	59	1,375	57
E. coli (Blattner et al., 1997)	3,771	2,883	79	2,987	76
F. nucleatum (Kapatral et al., 2002)	1,660	1,083	65	1,115	67
H. influenzae (Fleischmann et al., 1995)	1,479	1,183	86	1,277	80
H. pylori (Tomb et al., 1997)	1,303	815	67	869	62
L. lactis (Bolotin et al., 2001)	1,737	1,229	71	1,266	70
M. genitalium (Fraser et al., 1995)	461	332	70	324	72
M. pneumoniae (Himmelreich et al., 1996)	610	441	62	379	72
M. tuberculosis (Cole et al., 1998)	3,410	1,973	55	1,868	58
N. meningitidis ser. A (Parkhill et al., 2000a)	1,539	1,132	78	1,205	74
N. meningitidis ser. B (Tettelin et al., 2000)	1,530	1,088	76	1,161	71
R. prowazekii (Andersson et al., 1998)	759	548	75	572	72
S. coelicolor (Bentley et al., 2002)	5,986	3,625	60	3,489	58
Synechocystis sp. (Kaneko et al., 1996)	2,559	1,598	63	1,633	62
T. maritima (Nelson et al., 1999)	1,576	1,064	72	1,127	68
T. pallidum (Fraser et al., 1998)	920	507	60	552	55
V. cholerae (Heidelberg et al., 2000)	2,991	2,054	74	2,202	69
X. fastidiosa (Simpson et al., 2000)	1,770	1,184	70	1,233	67
Y. pestis (Parkhill et al., 2001)	3,350	2,566	77	2,726	81
S. cerevisiae (Goffeau et al., 1997)	5,560	3,302	67	3,726	59

class.

From the analysis presented in Paper IV, it was concluded that there does not appear to be radical differences in the usage of protein features within Archaea. Combining this with the fairly low number of proteins which could be assigned to a functional category, we decided to pool all archaeal proteins rather than train individual predictors for Crenarchaea and Euarchaea or for individual organisms.

Out of a total of 33,143 annotated protein coding genes in the 14 sequenced

archaeal genomes, 13,816 could be assigned as either enzyme or non-enzyme. The vast majority of the proteins that were classified as enzymes could also be assigned to an enzyme class, resulting in a set of 8,872 enzymes with "known" enzyme class.

Because these data sets consist of pooled proteins from 14 species, we can expect a large number of homologous proteins within the sets. This was dealt with by first partitioning each of the two sets in five cross validation sets, using a heuristic that minimizes that similarity between sequences in different sets. Subsequently all sequences with more than five connections to sequences in other sets were removed, thereby eliminating essentially all similarity between the cross validation sets.

The end results were two sets of archaeal proteins which were used for five fold cross validation training: 6,905 enzymes/non-enzymes and 6,837 enzymes assigned to enzyme classes.

7.4.3 Training the networks

Because of the high expected error rate in these fully automatically generated data sets, quite few weights were used in neural networks compared to the number of training examples. This gives us reason to believe that the networks have not overfitted on incorrect examples, but instead simply failed to learn most of them. Feature selection was done like described in Paper II, with the exception of the use of cross validation rather than a single test set to evaluate the performances.

Paper VI

7.5 Prediction of novel archaeal enzymes from sequence derived features

Lars Juhl Jensen, Marie Skovgaard, and Søren Brunak*
Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby, Denmark

* To whom correspondence should be addressed (email: brunak@cbs.dtu.dk)

The completely sequenced archaeal genomes potentially encode, among their many functionally uncharacterized genes, novel enzymes of biotechnological interest. We have developed a prediction method for detection and classification of enzymes from sequence alone, which is available at `http://www.cbs.dtu.dk/services/ArchaeaFun`. The method does not make use of sequence similarity rather it relies on predicted protein features like co- and post-translational modifications, secondary structure and simple physical/chemical properties.

Introduction

The conservation among enzymatic pathways is very low in Archaea, and even lower between Bacteria and Archaea. Some of the main characterized pathway operons are not found in all archaea, showing the complete loss of metabolic pathways. This is seen in the case of the histidine pathway that is found in *P. furiosus* but not in *P. horikoshii* and *P. abyssi* (Lecompte et al., 2001). For most pathways one or two reactions are predicted to be catalyzed by Archaea specific enzymes (Makarova et al., 1999). As the most studied archaea are all extremophiles, their proteins are of interest to basic science and for commercial exploitation.

In general, the gene repertoire of an archaeal organism is specifically related to other archaea, but are not significantly different from that of bacteria (Bansal and Meyer, 2002). The basic components of transcription, translation and replication system is well conserved in Archaea, the same goes for genes involved in repair and recombination.

The archaeal domain of life has a prokaryotic cell organization, but is more similar to Eukarya in relation to transcription, translation and replication. The metabolic proteins in Archaea is often more similar to homologous genes in Bacteria than in Eukarya (Koonin et al., 1997). With the sequencing of the first Crenarchaeota it was seen that the gene repertoire overlapped more with Euryarchaeota than with Bacteria or Eukarya (Natale et al., 2000). These archaeal features makes the archaeal domain of life an interesting area for research in uncharacterized proteins.

The variations in metabolism and the extreme conditions lead to unique archaeal enzymes that need to be characterized. Determination of three-dimensional structure is the traditional approach to functional classification of genes that cannot be assigned a role based on homology to known proteins. This is a very time consuming process and need for a faster method of classification is obvious. ProtFun is such a method for functional prediction based on sequence derived features (Jensen et al., 2002), however it has been developed for eukaryotes.

ProtFun was developed for predicting function of human proteins and makes use of the fact the function of a protein is affected by its surroundings and compartment. Functional categories are predicted from correlations between functional features which can be derived from sequence. Some of the features used for human sequences are not biologically meaningful in the case of prokaryotes and thus cannot be expected to correlate to function. However, in a cell without compartments, the function of the protein will still be affected by interacting proteins and cellular components. Interactions between proteins and cellular components can also be derived from the sequence in archaea. Therefore we have used the ProtFun approach to make an archaeal enzyme prediction method.

Results and discussion

Predictability of enzymes and enzyme classes

From a practical point of view the most important aspect of a prediction method is its ability to make correct predictions. As prediction methods are never perfect, one is always faced with the dilemma of choosing between making few false positive predictions and having a high sensitivity, i.e. correctly identify as many positive examples as possible. This tradeoff can be visualized as what is known as the receiver output characteristic (ROC) curve, where the rate of false positives is plotted as a function of the sensitivity by varying the score threshold used for making positive prediction. Figure 7.5 shows the ROC curves for all seven predictors included in our method.

The ROC curve for enzyme/non-enzyme prediction breaks at a sensitivity around 75% and a false positive rate of 30%. Assuming that half of the proteins encoded by archaeal genomes are enzymes—which corresponds to the composition of our training set—this corresponds to a specificity just over 70%.

Out of proteins that are enzymes, hydrolases and in particular ligases can be predicted with high certainty. For sensitivities below 20% the rate of false

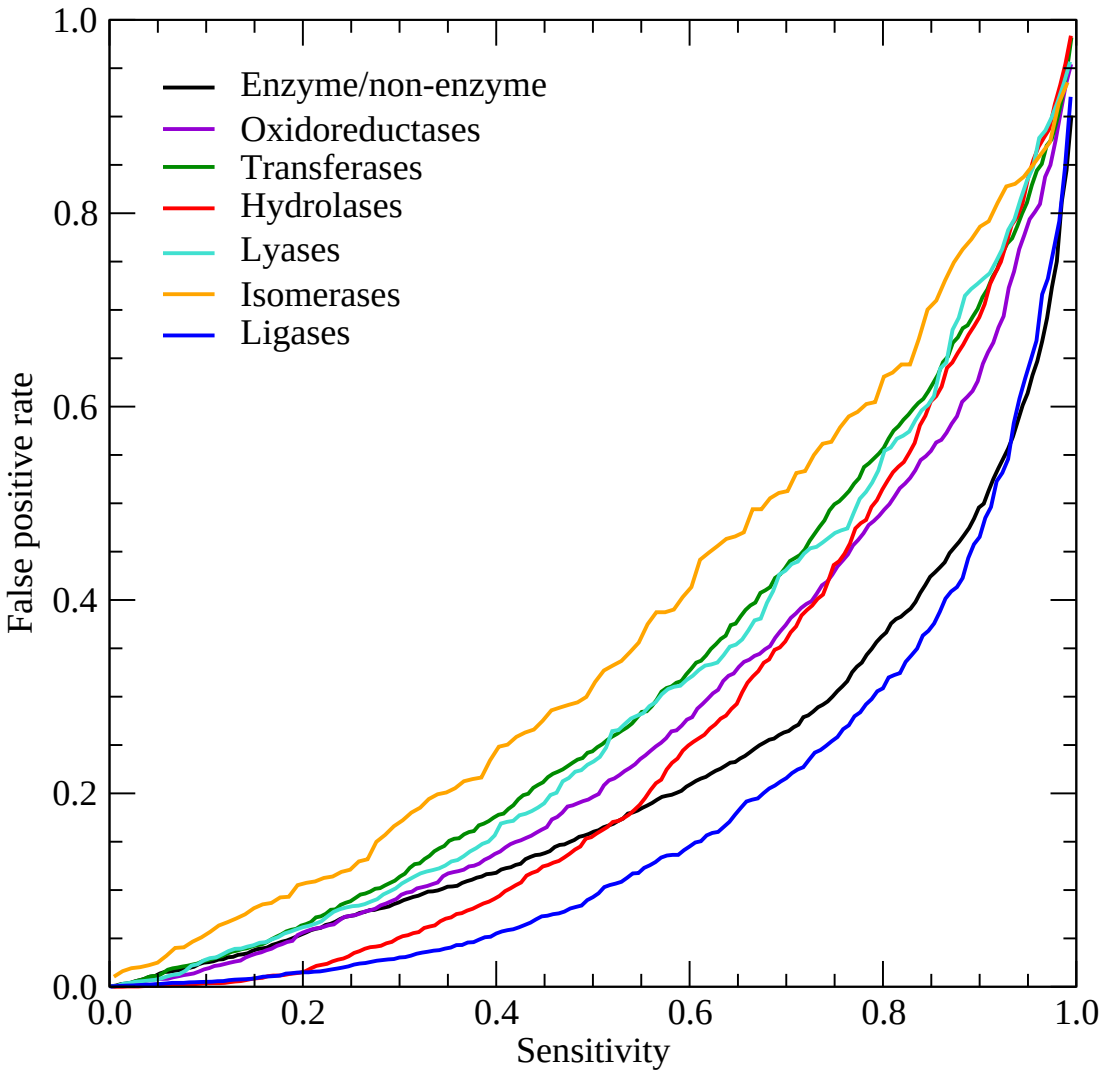


Figure 7.5: **Sensitivity and rate of false positives for the different predictors.** The plot was constructed based on the results obtained on the cross validation test set for enzyme/non-enzyme and each of the six major enzyme classes. Due to the careful partitioning of the cross validation data set, the performances shown here are what can be expected for real uncharacterized proteins. Random performance corresponds to a line along the diagonal.

positives is below 1% for both categories. In the case of hydrolases, which are of particular interest in the detergent industry, a sensitivity of 55% can be attained by sacrificing a bit on the specificity. By using the enzyme/non-enzyme and hydrolase predictors in combination, it should thus be possible to predict approximately 40% of all novel hydrolases in archaea.

Isomerases constitute the only class of enzymes where we have an unacceptably poor performance. This is somewhat surprising as this enzyme class is indeed predictable for human proteins (Jensen et al., 2002). However, the correlations found for human proteins clearly do not carry over to archaea, as the neural networks trained on human isomerases are not capable of predicting archaeal isomerases either (work in progress). Up to a sensitivity of around 30%, the remaining enzyme classes are predictable with false positive rates comparable to

those of the enzyme/non-enzyme prediction.

Biologically meaningful feature usage

While the performance is the most important aspect of the method from an engineering point of view, it is from a scientific point of view at least as interesting to understand how this performance is attained. We will now turn to analyzing the biological meaning of the features used for assigning enzymatic function.

Predicted structural properties

The structure of a protein is an important determinant for the detailed molecular function of proteins, and would consequently also be useful for prediction of enzymes and enzyme classes. As we only have the sequence available, the predicted secondary structure is one of the sequence derived features that we make use of. Although no clear trends are seen in the secondary structure of different types of enzymes, this feature is one of the most important features for our predictor overall (see Figure 7.6).

Based on the analysis of secondary structure derived from experimentally determined protein structures, differences in the secondary structure content of enzymes and non-enzymes have previously been shown to exist (Zhang and Zhang,

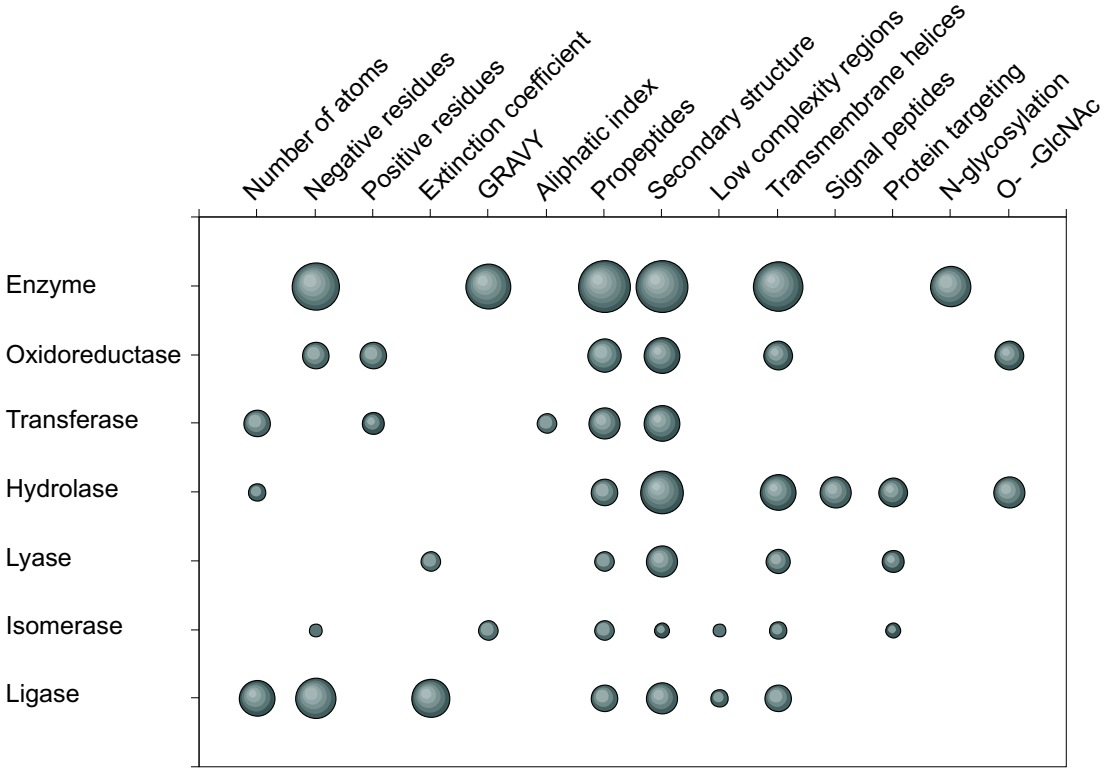


Figure 7.6: **Feature importance for the different classifiers.** The performance of each sequence derived feature for each category is visualized as a spot with area proportional to the correlation coefficient.

1999). This agrees well with our result that protein secondary structure can be used for predicting enzymes.

Different secondary structure contents have also been observed for particular classes of enzymes, e.g. proteases tend to have a lower helix content than other enzymes (Stawiski et al., 2000). Secondary structure is by far the most important feature for prediction of hydrolases, one of the two classes of enzymes that we are best at predicting. By comparing the positional secondary structure content for the six enzyme classes, we found that archaeal hydrolases have an unusually low content of α-helix and high content of β-sheet in their N-terminal region compared to other archaeal enzymes.

One other structural feature that can be predicted from sequence is transmembrane helices. This feature is in particular valuable for prediction of enzymes vs. non-enzymes as transmembrane proteins are underrepresented among enzymes. This observation is in sharp contrast to what we have observed for human proteins, where transmembrane helix prediction was found to be useless for enzyme prediction (Jensen et al., 2002). As prokaryotes do not have intracellular membranes, their transmembrane proteins will tend to be distributed over fewer functional categories.

Glycosylation of archaeal proteins

It may come as a surprise that predicted glycosylation sites can be used for the prediction of archaeal enzymes, given that glycosylation is mostly associated with eukaryotes. However, six different types of glycosylation have by now been confirmed to take place in archaea (Spiro, 2002).

We make use of two types of predicted glycosylation sites for this predictor: N-linked β-GlcNAc glycosylation sites are used for the enzyme/non-enzyme prediction, while O-β-GlcNAc sites are used for predicting oxidoreductases and hydrolases. The type of N-linked glycosylation is among the six types of glycosylation known to occur in archaea and is believed to mainly target cell surface proteins. This may explain why predicted N-glycosylation sites serve as an indicator that a protein is non-enzymatic. The other type of glycosylation that we make use of, O-β-GlcNAc, has not been observed for archaeal proteins so far.

It could be argued, that even if the two types of glycosylation sites used in our method exist in archaea, the predictors cannot be expected to work as they were trained exclusively on eukaryotic data. However, the proteins involved in both types of glycosylation appear to have developed early in evolution as they are highly conserved among all eukaryotes (Spiro, 2002).

In the case of N-linked β-GlcNAc glycosylation, homologs of the highly conserved STT3 subunit of oligosaccharyltransferase have been found, and it has been suggested that the Asn-Xaa-Ser/Thr consensus sequence known in eukaryotes does also hold for archaea (Spiro, 2002). Based on this we find it plausible that the glycosylation predictors, at least NetNglyc, do in fact work for archaea too.

Other important protein properties

In addition to the predicted structural features and glycosylation sites, all predictors make use of one or more of the simple protein properties that can be estimated from the sequence by the ExPASy ProtParam server. This is consistent with observations made in a cross species validation of the original ProtFun method (work in progress). In particular it is worth noting that the number of negatively charged residues is usable both for predicting enzymes/non-enzymes and ligases. This is because enzymes in general and ligases in particular contain a high number of negatively charged residues.

The most puzzling feature used in our prediction method is predicted Furin-type propeptide cleavage sites. No such cleavage sites are known in prokaryotic proteins, and the prediction method does not appear to correctly identify archaeal propeptides. In spite of this, the propeptide predictions are correlated to enzyme classes. As Furin-type cleavage sites are mainly characterized by being rich in positively charged residues (arginines and lysines), it is possible that it captures a different simple archaeal signal.

Combining enzyme predictions with phylogenetic patterns

The ProtFun prediction method can be used to obtain functional hints for some of the many archaeal genes where no functional assignments exist. Additional evidence can be obtained by using information of linked genes based on phylogenetic patterns (Pellegrini et al., 1999). Genes that are linked through a phylogenetic pattern are expected to be within the same cellular role category.

These cellular roles are related to enzyme classes as certain types of enzymes are overrepresented within particular cellular roles. In particular, many proteins involved in energy metabolism are oxidoreductases while central intermediary metabolism is biased towards transferases and hydrolases. Novel proteins from these two cellular role categories were predicted by extracting proteins of unknown function, which in the Predictome database (Mellor et al., 2002) were linked to proteins assigned to cellular roles by EUCLID (Tamames et al., 1998).

Five genes that had only a vague functional description if any, were found to be linked to proteins involved in central intermediary metabolism and were furthermore predicted by our method to be hydrolases. These proteins had a band 7 domain, that is also found in the major integral membrane protein stomatin and in the bacterial plasma membrane proteins *HflCK* (Tavernarakis et al., 1999). The function of the band 7 domain is unclear. The *HflCK* proteins have been suggested to be either proteases themselves or modulators of a protease (Cheng et al., 1988; Noble et al., 1993; Kihara et al., 1997). Our prediction agrees with the proposed protease activity.

Many archaeal protein coding genes have the annotation "conserved hypothetical protein". One such protein from *M. jannaschii*, MJ1681p, was linked by a phylogenetic pattern to a number of proteins that were classified as being involved in energy metabolism. Our method predicts MJ1681p to be an oxidoreductase. This prediction is in agreement with Pfam (Bateman et al., 2002) revealing a putative dicluster-type iron-sulphur center. The presence of an iron-

sulphur center strongly suggests that this is a ferredoxin, a function that has also been annotated for homologs of MJ1681p in more recently sequenced archaeal genomes.

Materials and Methods

Creating labeled data sets

The complete genome sequences of 4 crenarchaea and 10 euryarchaea were downloaded downloaded from GenBank (Benson et al., 2002) and the conceptual translations of all 33,143 annotated protein coding regions were extracted (see Table 7.5).

These sequences were searched against the SWISS-PROT database using BLAST (Altschul et al., 1997), recording all matches with an E-value better than $10^{-}3$. Using regular expressions, the description lines of all matches were searched for EC numbers, which are annotated for most enzymes in SWISS-PROT. The extracted EC numbers were then used in a majority voting scheme to assign the archaeal query sequences to enzyme/non-enzyme and possibly major enzyme class.

In order to avoid questionable labeling of the data sets, proteins were only labeled as *enzyme* if at least two-thirds of their database matches had an EC number in their description line. Equivalently, proteins annotated as *non-enzyme* were required to have EC numbers for less than one third of the database matches. If between one third and two-thirds of the matches had EC numbers or if no matches were found, the classification of the protein was considered unclear and it was removed from the enzyme/non-enzyme data set.

Despite these precautions, the data set will still contain some incorrectly labeled sequences. However, as the function of sufficiently many archaeal proteins

Table 7.5: **The data set size and breakdown on organisms.** Enzyme classes have been assigned based on sequence similarity to SWISS-PROT entries.

Organism	Annotated protein sequences	Assigned as enzyme/ nonenzyme	Assigned to enzyme class
Aeropyrum pernix	2,694	688	454
Pyrobaculum aerophilum	2,605	863	570
Sulfolobus solfataricus	2,977	1,179	782
Sulfolobus tokodaii	2,826	1,041	716
Archaeoglobus fulgidus	2,407	1,053	696
Methanobacterium thermoautotrophicum	1,869	878	584
Methanococcus jannaschii	1,715	780	516
Methanococcus mazei	3,371	1,440	908
Methanopyrus kandleri	1,691	658	455
Methanosarcina acetivorans	4,540	1,827	1,123
Pyrococcus abyssi	1,765	861	541
Pyrococcus horikoshii	2,064	781	489
Thermoplasma acidophilum	1,031	791	515
Thermoplasma volcanium	1,499	778	523
Total	33,143	13,816	8,872

has not yet been determined experimentally, one has to rely on function assigned based on sequence similarity.

From proteins annotated as *enzyme*, a second data set labeled with major enzyme class was assigned based on a similar scheme: if at least two-thirds of the database matches with EC numbers agree on the first digit of the EC number, the query sequence is assigned to the corresponding major enzyme class. Like above, sequences are removed if the two-thirds majority rule was not fulfilled.

These procedures resulted in two data sets for each genome: one data set proteins assigned as either *enzyme* or *non-enzyme* and a smaller set of enzymes annotated with major enzyme class. The sizes of these data sets are shown in Table 7.5.

Construction of pooled cross validation sets

Based on an analysis to be presented elsewhere, it was decided to pool the data sets for different organisms to make two large archaeal data sets. As these data sets consist of proteins from different organisms many orthologous proteins should be expected, making reduction of the similarity between training and test sets particularly important.

The two data sets were each partitioned into five equally sized subsets for cross validation. These sets were constructed with the objective to minimize the total number of significant BLAST hits between sequences in different sets. Each cluster of orthologous proteins will therefore reside in the same subset, allowing for reliable estimation of the performance by cross validation.

Neural network training

Individual cross validation ensembles of neural networks were trained for predicting the enzyme/non-enzyme classification as well as for predicting each of the six enzyme classes. To avoid problems with over-training on erroneously labeled examples, networks with very few weights compared to the data set size were used (Brunak et al., 1990). For each of these seven predictors, the optimal combination was found by a boot-strap strategy very similar to used for the development of the original ProtFun predictor (Jensen et al., 2002).

First a cross validation ensemble of neural networks were trained having only a single sequence derived feature as input. This was done for all features and all categories. The encoding used for each feature can be found at our web site (`http://www.cbs.dtu.dk/services/ProtFun/protfun\_add.html`). Based on the cross validated test set performance of these predictors, the worst performing features were rejected for each enzyme category and neural networks were trained for all pairs of the remaining features. From these the best features were again selected, progressively building up combinations of many features. In the end the feature combination with the best cross validation performance was selected for each category.

Prediction on new sequences is done by first running the many prediction methods to obtain the sequence derived features, which are subsequently used as input for the ensembles of five neural networks for each of the seven categories.

The neural network outputs were converted to probability scores as described in the ProtFun publication (Jensen et al., 2002). The probability for each category is estimated as the ensemble average of the probability scores from the individual networks.

Acknowledgments

The authors would like to acknowledge Ramneek Gupta and Peter Duckert for useful discussions on glycosylation and propeptides. This work was supported by grants from the Danish National Research Foundation and the Danish Natural Science Research Council. Marie Skovgaard is funded by EU Cell Factory Project, Screen, QLK3-CT-2000-00649.

7.5.1 Function prediction on bacterial proteins

Given the importance of secondary structure prediction for predicting enzymes and enzyme classes in archaeal proteomes (Paper VI), it is not unlikely that a similar predictor could be trained for bacterial proteins. This expectation is further supported by the cross-species comparison (Paper IV), which revealed that the enzyme predictors trained on human sequences gave similar results for archaea and eubacteria.

7.6 Prediction of Gene Ontology classes

It has already been discussed in this thesis, that while the cellular role categorization system is quite well standardized, it is not ideal for classification of eukaryotic proteins as it was designed with prokaryotes in mind.

Also one can ask if the cellular role categories are in fact a bit too broad. While broad categories have the advantage of giving many evolutionarily unrelated training examples for each category, it also comes at a price. The categories become very diverse and the proteins belonging to the same category might therefore not have much in common. Within more narrowly defined functional categories, the proteins can be expected to be more similar. However, one may not have enough examples to be able to identify these similarities.

To address both of these issues, a new version of ProtFun was trained for the Gene Ontology classification system. In contrast to the cellular role and enzyme classification systems looked at so far, Gene Ontology has a hierarchy of classes, which allows function prediction to be attempted at many different levels of detail. Equally important, it appears that Gene Ontology is quickly becoming the new standard for functional annotation, in particular for eukaryotic genomes.

As we were not yet certain of the cross-species capabilities of the ProtFun method, the data set was reduced to include only SWISS-PROT and TrEMBL entries containing human proteins. This also helps avoid some of the likely biases in the databases. All proteins which did give matches to any InterPro families were also removed, giving a set of 21,401 human protein sequences, each associated with one or more GO numbers.

Paper VII

7.7 Prediction of human protein function according to Gene Ontology categories

Lars Juhl Jensen*, **Ramneek Gupta, Hans-Henrik Stærfeldt, and Søren Brunak**
Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby, Denmark

* To whom correspondence should be addressed (email: ljj@cbs.dtu.dk)

Motivation: The human genome project has led to the discovery of many human protein coding genes which were previously unknown. As a large fraction of these are functionally uncharacterized, it is of interest to develop methods for predicting their molecular function from sequence.
Results: We have developed a method for prediction of protein function for a subset of classes from the Gene Ontology classification scheme. This subset includes several pharmaceutically interesting categories—transcription factors, receptors, ion channels, stress and immune response proteins, hormones and growth factors can all be predicted. Although the method relies on protein sequences as the sole input, it does not rely on sequence similarity, but instead on sequence derived protein features such as predicted post translational modifications (PTMs), protein sorting signals and physical/chemical properties calculated from the amino acid composition. This allows for prediction of the function for orphan proteins where no homologs can be found. Using this method we propose two novel receptors in the human genome, and further demonstrate chromosomal clustering of related proteins.
Availability: Sequences can be submitted to the prediction server via a web interface at `http://www.cbs.dtu.dk/services/ProtFun/`

Introduction

For most of the whole genome sequencing projects, the function of a large fraction of proteins remain unknown. Predicting the putative function of these so-called orphan proteins is an important but difficult task for bioinformaticians. We have

previously presented the ProtFun method for predicting the cellular role categories originally proposed by Riley (1993) as well as enzymatic function according the EC classification system (Jensen et al., 2002).

Using a similar approach we have now expanded the ProtFun prediction method to also cover a number of biologically as well as pharmaceutically interesting categories in the Gene Ontology (GO) classifications system (Ashburner, 1998; Ashburner et al., 2000). Unlike the cellular role categories, the GO categories chosen do not give complete coverage in the sense that some proteins will not belong to any of the categories. But for the ones that do, the GO predictor will provide a more specific description of the function than the rather broad cellular roles categories do.

System and methods

Generation of a labeled data set

The most crucial step in developing a good prediction method is always to obtain a good data set. Unfortunately it was not possible to directly obtain a large set of sequences annotated according to the Gene Ontology classification scheme (Ashburner et al., 2000). This is mainly due to the large amount of manual work involved in reannotating individual sequences.

Instead we made use of the InterPro database (Apweiler et al., 2000) in which protein families have been assigned with Gene Ontology numbers. By linking this with a list InterPro domain matches to SWISS-PROT and TrEMBL, a set of 21,401 human sequences with annotated Gene Ontology numbers was obtained.

An alternative source of Gene Ontology numbers would be the SWISS-PROT keywords. However, to make use of these we would have to discard all TrEMBL entries, leaving us with approximately 9,000 human protein sequences. For this reason we decided against using this source of Gene Ontology annotations and used only the InterPro derived set described above.

Typically, a homology reduction would be performed for this set to obtain a smaller set of sequences, none of which display significant sequence similarity. However, homology reducing this data set using Hobohm algorithm 2 to remove matches with BLAST expectation values below 10^{-6}, reduced the data set to less than 3,000 sequences (Hobohm et al., 1992; Altschul et al., 1997).

Data set partitioning

To avoid throwing away the majority of the available data, we instead divided the data set into five cross validation sets of equal size with minimal sequence similarity overlap between the sets. As finding the optimal solution to this problem is of combinatorial complexity, a heuristic was developed.

First all sequences in a set S were aligned against each other, finding the maximal scoring subsequence similarity using the BLAST program (Altschul et al., 1997). Given two sequences a and b one minus their P-value were used as a scoring weight $W(a, b)$. The algorithm divides S into k equally sized partitions

$P_1 \ldots P_k$, minimizing the total inter-partition similarity. Let

$$E_{i,j} = \sum_{a \in P_i} \sum_{b \in P_j} W(a, b)$$

define the weight of the sequence similarity between two sets, and let

$$E_{\text{ext}} = \sum_{i=1}^{k} \sum_{j=1}^{k} E_{i,j}, i \neq j$$

be the weight to be optimized. Let

$$I_j = \sum_{a \in P_j} \sum_{b \in P_j} W(a, b)$$

be the internal weight among the sequences assigned a partition P_j. The algorithm first leave $P_1..P_k$ empty, then chooses the sequence a from S and a set P_j that leads to the least increase in E_{ext} given a is assigned to P_j. In case of ambiguity, the partition P_j is chosen that has maximal I_j. Assignment is only considered among the sets that do not yet have the desired size.

Unfortunately it turned out to either be impossible to split the set of 21,401 proteins into five unrelated subsets or the heuristic at least failed to find a sufficiently good solution. By looking closer into the similar sequences between the subsets in the best solution, it was realized that almost all connections were caused by a fraction of the data set. Each of the five subsets were thus reduced to 2,500 sequences by removing the nodes with the highest connectivity first. This resulted in a five fold cross validation set of 12,500 sequences with no significant similarity between sequences in the different subsets.

This ensures that when training cross validation ensembles of neural networks, any two similar sequences will either both be used for training or both used for testing. This allows us to use a set much larger than the one obtained by homology reduction while still getting a correct measure of the performance on an independent test set.

Choosing the classes to predict

We have based our work on the Gene Ontology as of June 10th 2001, which defines a total of 7,949 different classes, of which 1,532 were represented in the data set described above. However, it would be neither feasible nor necessary to train neural networks for prediction of each of these categories.

Because we require sequence similarity to only occur within each of the five subsets used for cross validation, sequences belonging to the same protein family will be present in the same subset. This means that in order for a Gene Ontology category to be represented in all sets, multiple InterPro families must belong to the category. The requirement for multiple families is also necessary in order to be able to generalize further than simple sequence similarity, i.e. make an *ab initio* function prediction method.

To have sufficient diversity within each of the five cross validation data sets, we decided to only train neural networks for Gene Ontology categories which were annotated to at least 20 different InterPro families. This reduced the number of categories further, leaving 347 categories for which neural networks were trained.

Sequence derived protein features

In addition to the 14 features used for input in the original ProtFun prediction method (Jensen et al., 2002), features representing two new servers were added to the list. These were the newly developed method for prediction of propeptide cleavage sites (Blom et al., unpublished results) and the subcellular compartment predictor TargetP (Emanuelsson et al., 2000). Even though TargetP makes use of SignalP to predict extracellular proteins, the original encoding of the SignalP output was still retained as a separate feature, see Emanuelsson et al. (2000) for details.

Neural network training and feature selection

For each Gene Ontology class, standard feed-forward neural networks with a single layer of hidden neurons were used for predicting which examples belong to a given class. Different learning rates were used for positive and negative examples to avoid biased learning due to our data sets being heavily skewed towards negative examples. For each feature combination (including single features) the input vector for the neural networks consists of a concatenation of the respective feature vectors while the target output is a single value which is 1 for positive examples and 0 for negative examples for the Gene Ontology class in question. As neural networks were trained with different combinations of sequence derived features as input, the number of hidden units was varied to keep the size of the network as close to 100 weights as possible.

First cross validation ensembles of five neural networks each were trained using each protein feature as single input. For this, a neural network was trained using each of the five sets for testing and other four sets for training. A robust estimate of the performance of a feature for a particular functional class was calculated as the median test set Pearson correlation coefficient of the five neural networks in the ensemble.

The majority of the 347 selected Gene Ontology classes turned out to not be strongly correlated to any of the predicted features. Judging this from the cross validation performance of single feature neural networks, many categories were discarded because they did not appear to be predictable with any reasonable accuracy. Also, a number of trivial categories related to subcellular locations were removed, e.g. *cell* (and its alternative *extracellular*) and *membrane* which should be trivially predictable based on the signal peptide (SignalP) and transmembrane helix (TMHMM) features, respectively. For the majority of the categories for which acceptable performance was not attained, the training sets were quite small (data not shown).

Only 26 Gene Ontology classes remained after this reduction. For each, the optimal feature combination was searched for using a greedy search heuristic also described in the original ProtFun publication (Jensen et al., 2002). For the best performing single features, network ensembles were trained for all feature pairs. Judging the performance of each feature from the performance of the best pair in which it is involved, the worst features were once again removed. All combinations of three were then tested for the remaining feature, and so on so

forth. When the best feature combinations had been identified, the number of weights were optimized for the three best feature combinations to find the best performing combination of features and network architecture.

To avoid unnecessary redundancy in the prediction method, a number of categories were discarded because they were too closely related to better alternatives (see Figure 7.7). For instance, the molecular function *transcription factor* was discarded in favor of the cellular role category *transcription regulation*. Also, *defence response* was removed as it was found to be too closely related to both its superclass *stress response* and its subclass *immune response*. The final set of predictors consists of cross validation ensembles of five neural networks for each of 14 Gene Ontology categories.

Making predictions with the neural networks

To use the neural networks to predict the function of novel sequences, all sequence derived features are first calculated and encoded the same way as the training examples, then presented to each of the five neural networks corresponding to each Gene Ontology class. For each class, the average output is calculated and converted to a probability using a calibration curve. These calibration curves were estimated from score distributions as described in the original ProtFun publication (Jensen et al., 2002).

Table 7.6: **Overlap matrix for selected Gene Ontology categories.** The degree of overlap between functional classes (identified by their GO-number) is shown as the number of shared positive examples. Along diagonal the number of positive examples for each class is shown. The categories that were removed due to extensive overlap are shown in italics. The textual descriptions for the GO-numbers are: transcription factor (3700), transcription (6350), transcription regulation (6355), signal transducer (4871), receptor (4872), transmembrane receptor (4888) stress response (6950), defence response (6952), immune response (6955) response to biotic stimulus (9607) response to wounding (9611).

GO #	*3700*	6350	6355	4871	4872	*4888*	6950	*6952*	6955	*9607*	*9611*
3700	117	114	106								
6350		193	153								
6355			153								
4871				462	324	264					
4872					324	264					
4888						264					
6950							65	51	48	57	51
6952								51	48	51	51
6955									48	48	48
9607										57	51
9611											51

Support Vector Machines

Support Vector Machines (SVMs) were also tested on the classes that were found to be predictable by neural networks. We used the SVM-Torch software (Collobert and Bengio, 2000) to train SVMs with radial basis function kernels with varying standard deviation on the same data sets used for neural network training. The radial basis function kernel has been one of the best performing kernels in previous bioinformatics applications of SVMs (Brown et al., 2000; Hua and Sun, 2001). However, the performances we obtained with SVMs were not better than those of neural networks, for which reason the use of SVMs was not pursued further.

Discussion

While many categories turned out to not be predictable, fortunately many pharmaceutically interesting categories are among the predictable classes. Transcription factors, receptors, ion channels, stress and immune response proteins, hormones and growth factors are all among the predictable categories.

There are several reasons for why so relatively few Gene Ontology classes can be predicted using our method. One is lack of data: for 90% of the Gene Ontology classes we cannot assign a single positive example among human SWISS-PROT and TrEMBL entries—sufficient examples to even attempt training were only found for 2.4% of all Gene Ontology classes. Also, many of the categories are represented by the sequences which further reduces the set of categories to predict (Table 7.6). Predictors were successfully trained for the majority of the classes where many training examples were available. This could be seen as an indication that our approach is best suited for predicting broadly defined categories, but may simply reflect that insufficient data are available for more specific categories. It is also worth noting that that our method appears to be better at predicting *biological process* than *molecular function* which is consistent with previous observations (Jensen et al., 2002).

The performance obtained for the different Gene Ontology categories is shown in Figure 7.7. For all categories a sensitivity of at least 50% can be attained with a rate of false positives below 10%. For the best categories, hormones and receptors, a sensitivity of 70% can be obtained with a false positive rate of only 5%.

The method is well suited for gene discovery and assay selection purposes. For instance, it should be possible to predict a set of approximately 1,000 sequences containing 70% of all novel receptors or peptide hormones (assuming 40,000 genes in the human genome). Our prediction method can also be useful for getting an idea of a possible function for proteins known to be involved in a particular disease but otherwise of unknown function, thereby helping to select appropriate assays. However, even though the performance is much better than what could be expected considering that only sequence is used, the performance is still not good enough for annotation purposes.

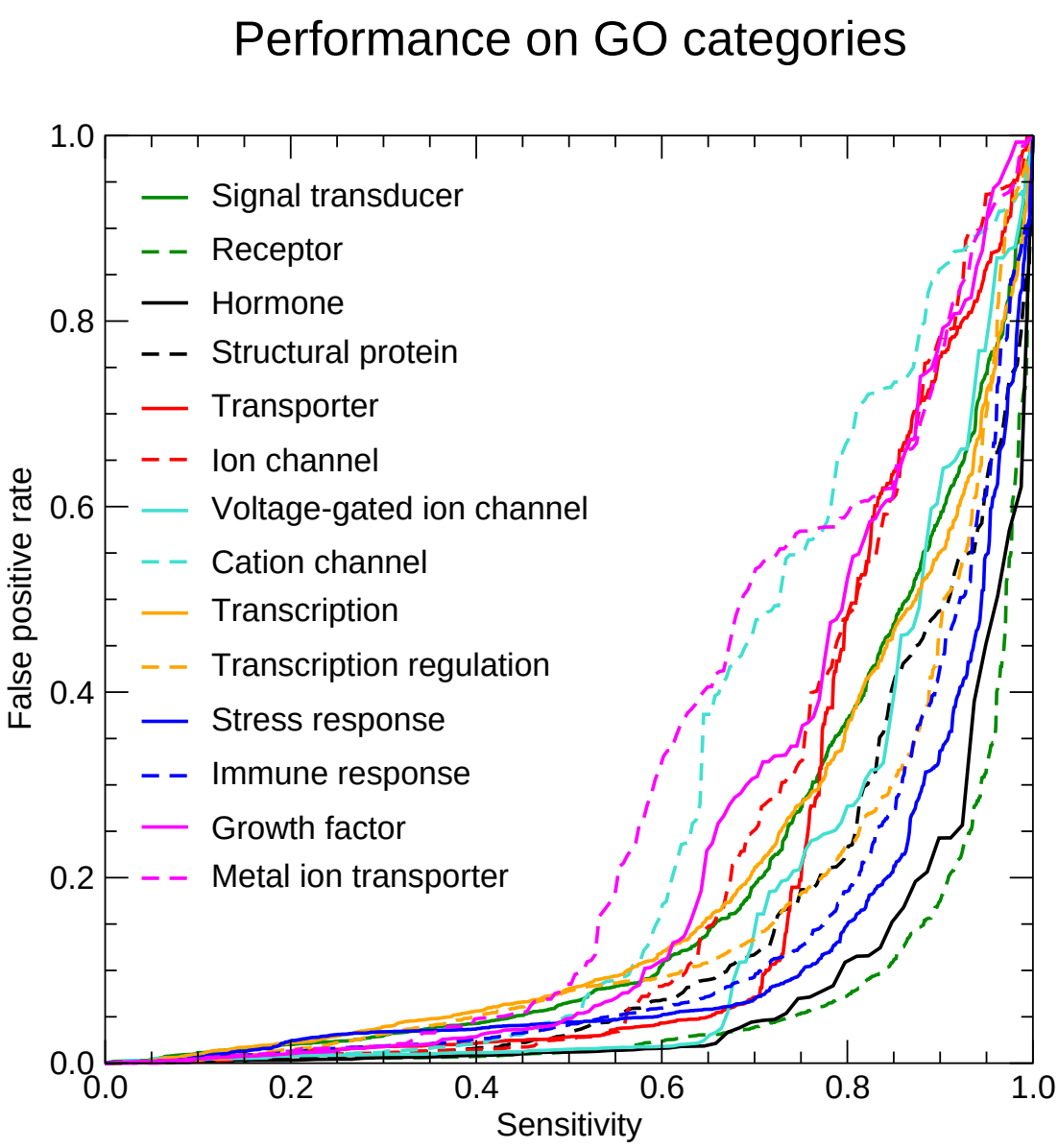


Figure 7.7: **Sensitivity and rate of false positives for the predictors.** The receiver output characteristic (ROC) curve for each class was constructed based on the cross validation test set performances. Due to the homology partitioning of the cross validation data set, the performances shown correspond to what can be expected for novel proteins.

Interpreting the features

In addition to evaluating how well the prediction method works, it is also interesting to interpret the method to understand how it works. Since the method works, the features used as input for a given category must somehow characterize the proteins belonging to that category.

Simply looking at which features are used can provide a first simple idea of what the different predictors look for. In addition to this simple binary description, we have also used two different approaches for evaluating the importance of the individual features.

The simplest of these is to look at the performance obtained by neural net-

works using each of the features individually. No extra networks have to be trained to do this, as these correlation coefficients were already found during the feature selection procedure. Figure 7.8 shows these values visualized with blue circles. The main problem of this approach is that correlations between features are lost, meaning that a feature which is mainly useful in combination with one or more other features will be judged as unimportant. This approach has also been used for analyzing the NetDrug prediction method (Frimurer et al., 2000).

Because of this problem we also used a second measure for feature importance: the loss in correlation coefficient by removing a particular feature. For a given feature and Gene Ontology category, the correlation coefficient was obtained for neural networks trained on the optimal feature combination with the feature in question removed. This value was subtracted from the correlation coefficient of optimal feature combination to find the loss in performance. These values are shown in Figure 7.8 with yellow circles.

The two measures for the feature importance both have advantages and disadvantages. By measuring feature importance as the loss in correlation coefficient, the problem with correlated features is solved, as a feature which adds a lot to the performance of other features will be judged as being important. However, this method has problems with redundant feature combinations where the same information is encoded by different features. For such encodings the performance is hardly affected by removing any one of the features, causing all features to be judged as unimportant. Since the networks perform well, this can clearly not be the case. We use both feature importance measures as they complement each other well, while none of them are perfect. Because both of them can miss important features but never give false positives, we consider a feature to be important if any of the two measures give a high score.

Transmembrane helix prediction is the most valuable feature for predicting the set of Gene Ontology categories that we work with (see Figure 7.8). It is very important for prediction of *signal transduction* proteins (especially *receptors*), *transporters*, and *ion channels* (in particular *cation channels*). It is interesting that we are able to predict both receptors and ion channels with high accuracy, considering that both of them are characterized by being transmembrane—in the case of ion channels, transmembrane helix prediction is essentially the only feature that matters. This can only be explained by the neural networks having learned a particular transmembrane structure which is characteristic of ion channels, rather than simply predicting all transmembrane proteins to be ion channels.

Figure 7.8 further reveals that secondary structure predictions are very useful for predicting *stress response* and its subclass *immune response*. From studying the feature distributions for these classes, it is clear that these proteins have a strong bias for β-sheets over α-helices, especially in the C-terminal part of the sequence. *Stress response* and *immune response* proteins are further characterized by having signal peptides.

Proteins related to *transcription* and more specifically to *transcription regulation* are recognized from feature combinations where all features are of approximately equal importance. From examining the loss in correlation coefficient (Figure 7.8, yellow circles) it is realized that this encoding is also highly robust as any one of the features can be removed without affecting the performance

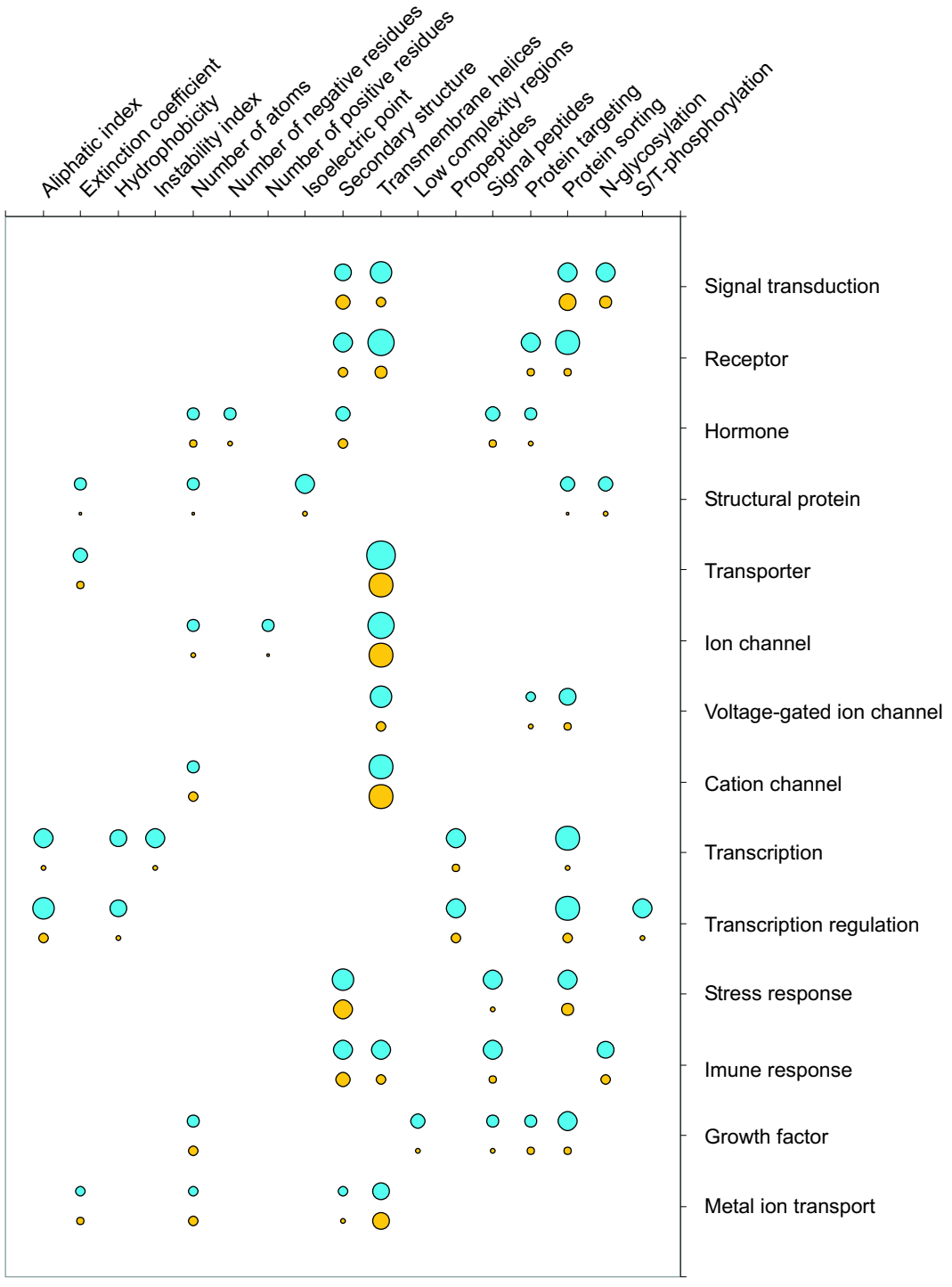


Figure 7.8: **Feature importance estimates.** The importance of each sequence derived was estimated for each of the categories in two ways: based on the correlation coefficient obtained by neural networks having the feature as their sole input (blue) and as the loss in correlation when the feature was removed from the optimal combination (yellow).

much. This makes it is difficult to decode the characteristics of transcription proteins from the neural networks, although it is clear that prediction of nuclear compartment at least is involved.

Novel putative receptors

We have used our method to screen the human proteome for possible novel receptors. For this purpose we used the Ensembl database as of November 12th 2001. Even though the rate of false positives is very low, the relatively low abundance of receptors results in a fairly high absolute number of false positives. While our set is believed to be highly enriched in receptors, further support for the prediction is needed—preferably in the form of experimental verification. We will now briefly discuss protein sequences from Ensembl that we predict to be novel receptors.

The protein sequence `ENSP00000257015` from Ensembl which is located on chromosome X, is predicted to be a putative receptor with a probability score of 73%. A BLAST search gave no clue as to the function of this protein, but a search against the Pfam database (Bateman et al., 2002) revealed a weak similarity to the TGF-beta type III receptor domain family zona_pellucida:

```
Alignments of top-scoring domains:
zona_pellucida: domain 1 of 1, from 29 to 111: score 4.6, E = 0.28
            *->qCtedgmvvsvvkdlltkpglnlpsllllgpndsaCqpvpvdstqaf
               +C+ d+++v+v+   l     +   +l lg    +C+p  v   ++
19577    29    LCSIDWFMVTVHPFMLNNDVCVHFHELHLG---LGCPPNHV--QPHA 70

            vifevplngCGtrlqv.tgdhlvYeNeLvaapsplvgpggsIt<-*
            ++f+  +++CG r +    d ++Y++e++++ + +  p+ +++
19577    71 YQFTYRVTECGIRAKAvSQDMVIYSTEIHYSSKGT--PSKFVI    111
```

Although the expectation value is 0.28 and the match is thus not evidence when viewed alone, it does add evidence to our prediction.

`ENSP00000252184` is another protein sequence which could not be assigned a GO number based on matches to InterPro, nor did BLAST or Pfam searches give any matches to proteins of known function. ProtFun predicts this protein to be a receptor with 72% probability. In fact, others have previously suggested that this is a G protein-coupled receptor based on a careful manual study of the predicted transmembrane helix structure (see GenBank entry `AF376725`).

Chromosomal clustering of proteins with similar function

It has been observed by several research groups, that genes with related function are often located close to each other on the chromosomes (Dandekar et al., 1998; Frishman et al., 1998; Galperin and Koonin, 2000; Yanai et al., 2001). This effect is strongest in prokaryotes due to the existence of operons, but clustering of related genes is also observed in eukaryotes (Wambutt et al., 2000). In prokaryotes, "chromosomal proximity" has been used for function prediction by claiming two genes to be functionally linked if their homologs are located close to each other in multiple genomes (Dandekar et al., 1998).

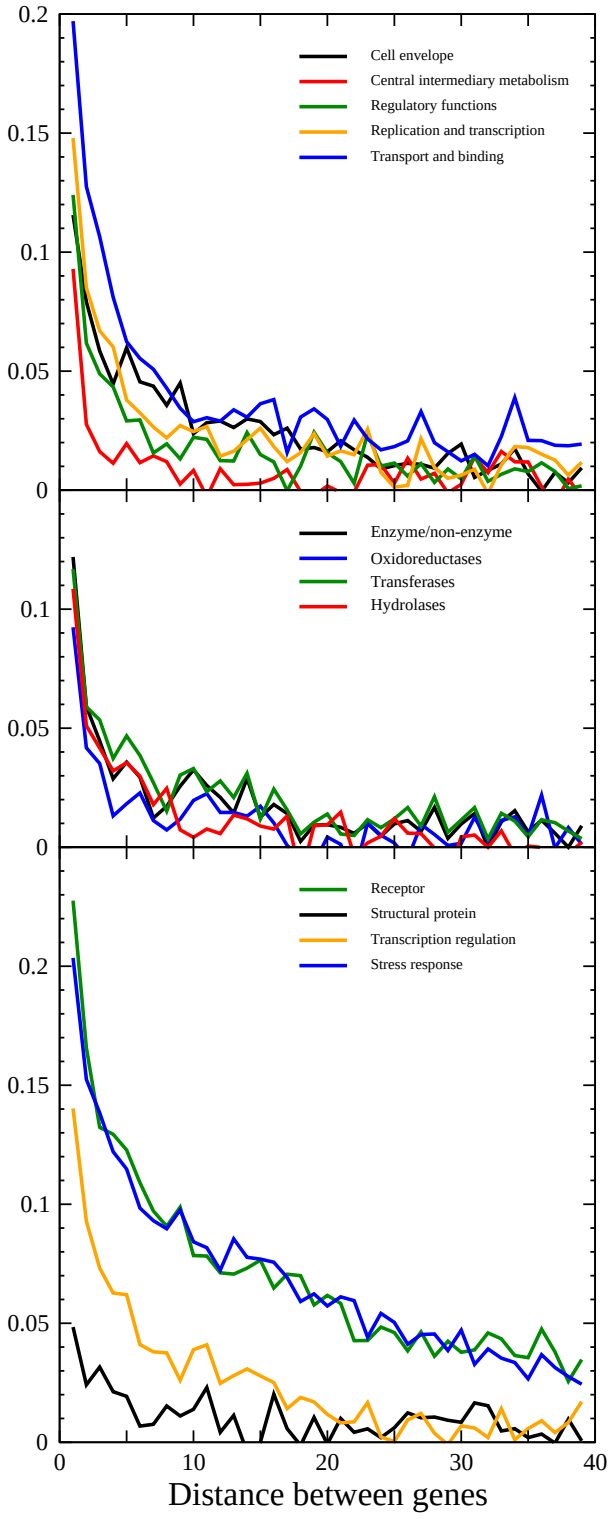


Figure 7.9: **Autocorrelation functions for predictions of selected categories from different classification schemes.** For each category in the three classification schemes predicted by ProtFun, the autocorrelation function of the probabilistic predictions was calculated with respect to the chromosomal localization. The gene distances are measured in "number of genes", i.e. neighboring genes have a distance of one. Chromosomal clustering is in particular observed for stress response genes and genes encoding receptors.

We have investigated how much the functionally related proteins cluster in the human genome, by studying the autocorrelation function of the probabilistic scores for each individual category in the cellular role, enzyme class, and Gene Ontology classification schemes. Figure 7.9 shows a representative subset of these autocorrelation functions for each classification scheme.

The autocorrelation functions reveal, that the extend to which genes with similar function cluster depends strongly on how "function" is defined. If defined in terms of the chemical function (e.g. enzyme classification), the clustering is very weak. Cellular role categories cluster more strongly, but not nearly as strongly as several of the Gene Ontology categories.

Even within the same classification scheme, there is a great deal of variance. This is in particular true for the Gene Ontology system, where many different types of functional categories occur. The two categories showing the strongest autocorrelations are *receptors* and *immune response* proteins, for which the autocorrelation functions also decay quite slowly. Strong correlations are of course also observed for their superclasses, *signal transduction* and *stress response*. It should be noted that although the autocorrelation functions for the *receptor* and *stress response* categories are very similar, the overlap in our training set is small between these two classes.

Conclusions

We have succeeded in making a sequence based function prediction method for a subset of the Gene Ontology. The method is well suited for computational screening of the human genome (and possibly other eukaryotic genomes) for novel drug targets. Based on complete genome predictions, we suggest two novel receptors and furthermore find strong indications that functionally related proteins are clustered in the human genome, although this varies between functional categories.

Acknowledgments

The authors wish to thank David Ussery for suggestions to the manuscript. This work was supported by grants from the Danish National Research Foundation and the Danish Natural Science Research Council.

7.7.1 An important addition to ProtFun

Although only a fairly small number of Gene Ontology classes could be predicted well by the ProtFun approach, the 14 classes that can be predicted constitute an important addition to ProtFun. They do so because these categories are much more specific and represent pharmaceutically interesting classes. The Gene Ontology part of the ProtFun output could therefore prove useful for identifying new drug targets—something which is unlikely to be true for the cellular role predictions.

7.8 Functional clustering at the global scale

In Paper VII, autocorrelation functions for the many different functional classes were analyzed on the human genome. It was seen that genes both from some of the cellular role categories and in particular from certain Gene Ontology classes tend to occur as clusters in the chromosomes. However, the autocorrelation functions say nothing about how the genes are clustered, except from giving an idea of the size of the clusters.

7.8.1 Clustering at the chromosome level

The distribution of functionally related proteins over chromosomes was mentioned briefly in Paper II. This can be considered the coarsest level at which clustering of functionally related proteins can be studied, yet it can provide information not available from autocorrelation functions.

Figure 7.10 shows the distribution of cellular role categories over the human chromosomes. The cellular roles were taken from our labeled data set created using EUCLID, and chromosome numbers were assigned to most of the proteins by linking to the LocusLink database via the OMIM database.

The presence of unusually many proteins from the class *other categories* on chromosome 20 has already been mentioned in Paper II. A few other noteworthy observations can be made from Figure 7.10, one being that judged from the number of proteins of unknown function, chromosomes 20 and 22 appear to be the best characterized of the human chromosomes.

In general the larger the chromosomes display fewer deviations from the expectation. A simple explanation for this is that the large chromosomes encode so many proteins that even a large cluster of functionally related proteins would not skew the distribution much. The only possible exception to this rule is an over-representation of *amino acid biosynthesis* genes on chromosome 2. Conversely, the chromosomes encoding few proteins (in particular the Y-chromosome) tend to deviate much more from the expectation. However, many of those deviations may not be statistically significant.

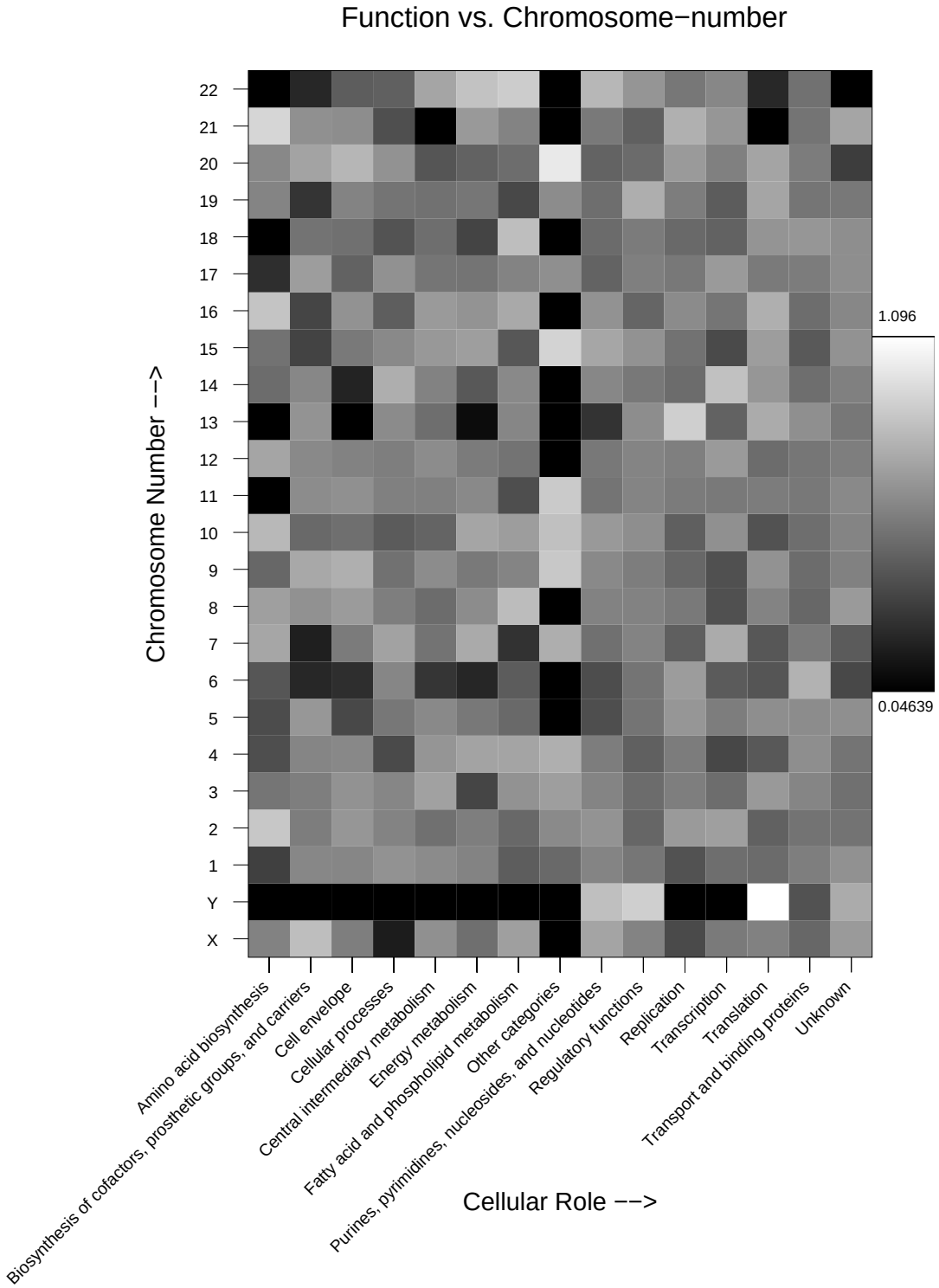


Figure 7.10: **Distribution of cellular roles over chromosomes.** Each spot represents the normalized frequency of the number of proteins (belonging to a certain category) transcribed from a particular chromosome. This number is obtained by dividing the absolute number of proteins (on a spot) by the number of proteins transcribed by the chromosome and by the number of proteins occurring in the category.

7.9 Whole genome visualization of protein function

The autocorrelation functions and the chromosome function distributions only allow the clustering of protein functions to be observed at a global level. Even though there might not be strong correlations at the global scale, there can still be local regions in the genome that are strongly biased towards genes of particular functional classes. To investigate such local properties, it is useful to visualize it. A quite successful genome visualization method known as "genome atlases" has previously been developed at CBS, and it has proven useful for visualizing a host of different properties at the DNA level, in particular DNA structure and repeat patterns (Jensen et al., 1999; Pedersen et al., 2000; Friis et al., 2000; Ussery et al., 2001, 2002). By mapping the probability scores given by ProtFun for each protein onto the corresponding region of the chromosomes, it is possible to visualize the spacial distributions of protein functions across entire genomes.

Paper VIII

7.10 The Atlas Visualization of Genome-wide Information

Marie Skovgaard, Lars Juhl Jensen, Carsten Friis, Hans-Henrik Stærfeldt, Peder Worning, Søren Brunak, and David W. Ussery*

Center for Biological Sequence Analysis
BioCentrum-DTU
Technical University of Denmark
DK-2800 Lyngby, Denmark

* To whom correspondence should be addressed (email: dave@cbs.dtu.dk)

Introduction

The wealth of information contained in a microbial genome is not easy to comprehend at all scales. Even after the genome of an organism has been sequenced, the problem of gaining an overview of the newly acquired data still remains. One way to get an overview is to visualize positional features at the chromosome level; we have developed a method, Atlases, for showing correlations between position dependent information in sequenced chromosomes.

The DNA sequence is not only hard to comprehend because of its size but also because the genomic sequence is not linked in a simple way to the biology of the organism. For example, examining the AT content of genomes, a property often reported in genome sequencing papers, considerable variation is observed. Figure 7.11 shows the percent AT of 25 different proteobacterial genomes, ranging from 33% to 74% AT. The AT content does not appear to correlate to the proteobacterial subdivisions.

The percent AT of a chromosome reflects only an *average* property of the chromosome. However, the AT content is not homogeneously distributed throughout the DNA. Often there are clusters of AT-rich and AT-poor regions; for example most promoter regions are more AT-rich than the average coding sequences (Ozoline et al., 1999; Pedersen et al., 2000). In many cases the variations between regions will tell more than the average value, as exemplified in Figure 7.12 where an AT-rich region is found to contain genes involved in pathogenesis.

The AT content within a region of a chromosome is a very simple property

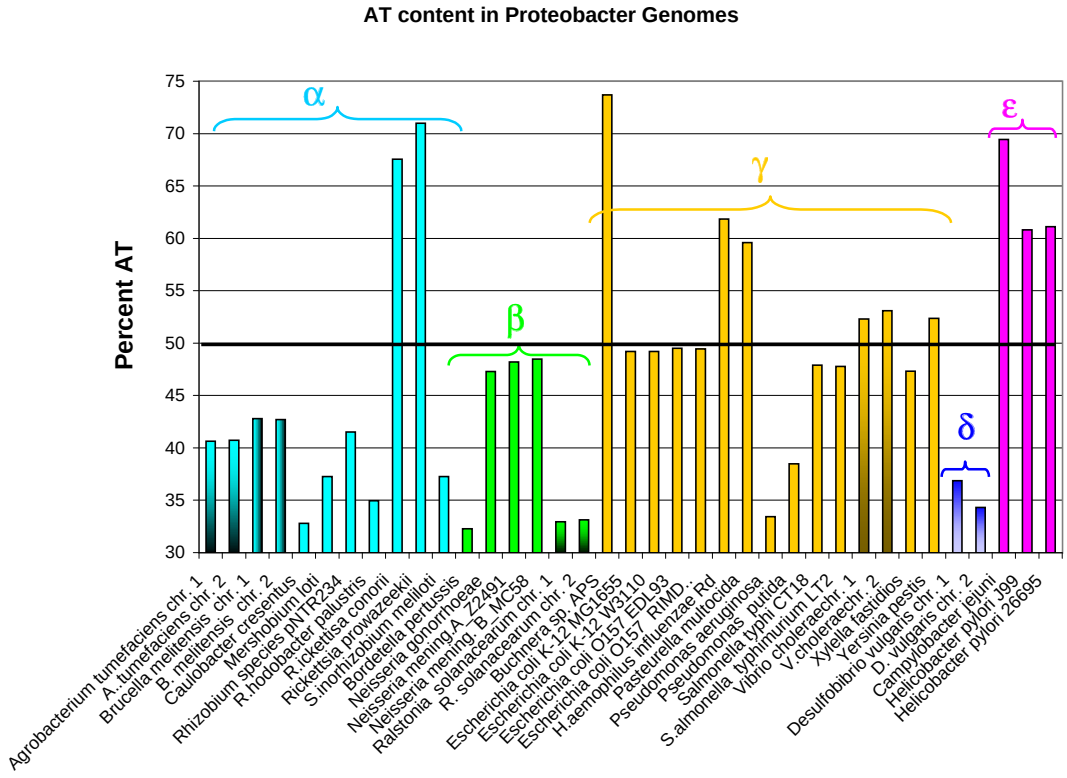


Figure 7.11: **Percent AT in 25 Proteobacter genomes.** The genomes are grouped into subdivisions, and the last bar is the average of all 25 genomes.

to calculate from the nucleotide sequence. More complex features like DNA curvature or major groove compressibility, that reflect structural properties of a given region, can be estimated directly from the sequence and give biological insight. Additional information can be accessed by looking at the genes encoded by the chromosome. Once the location of the genes is known, it is possible to visualize both experimentally determined expression levels and RNA sequence features. By translating the RNA sequence to protein sequence, it is also possible to visualize properties of the proteome, such as protein function.

Construction of the visualization software

In order to be able to visualize such diverse data, a flexible software tool is needed. We have developed a computer programme, GeneWiz, which enables us to visualize a complete chromosome compactly. The programme creates an either circular or linear graphical representation of the entire chromosome or of a specified subsection. Sufficiently large regions that display significant variation from the rest of the chromosome can be readily found—in order to be able to see deviations of smaller regions a zoomed Atlas must be made.

Each feature, such as AT content or gene expression level, is represented as a separate lane in the atlas. The value is at each position color coded according to a user specified color scale. We generally use color scales where regions of extreme values are highlighted (this can be one-ended or two-ended scales) whereas typical

values are grey. If wanted the plot can be smoothed by a running average.

The properties to be visualized must be present in the form of one value per basepair in the chromosome. For simple sequence features like the AT content this is the natural format, whereas for data such as gene expression levels, the value for each gene must be mapped onto the corresponding range of basepairs. In addition to the data series, the annotations from a GenBank file can be displayed using a series of icons with user-defined colors. This allows for the identification of short or long annotated regions of interest.

GeneWiz is solely a visualization program, and is not capable of calculating the data used in the different atlases. All data must be calculated and properly formatted. While this obviously adds to the work of creating an atlas, it is gives great flexibility as these data can come from any source. In this paper we use simple measures generated from lookup tables, publicly available programs like BLAST (Altschul et al., 1997), methods developed in-house like ProtFun (Jensen et al., 2002) as well as experimentally determined expression data.

Use of Atlases to Visualize DNA Information

Genome Atlases

The GenomeAtlas is a general atlas made for all the fully sequenced microbial chromosomes found in public databases (Pedersen et al., 2000; Jensen et al., 1999). The GenomeAtlas is a combination of some generally informative parameters and can be used as an offset for identifying unique regions or special features for the given chromosome. The GenomeAtlas for all public available sequenced chromosomes can be found at `http://www.cbs.dtu.dk/services/GenomeAtlas/`.

Introducing the parameters

To generate GenomeAtlas plots, a number of parameters are calculated for the DNA double helix based on the nucleotide sequence. These parameters belong to three categories: repeats, structural parameters, and parameters directly related to the base composition. These three categories are combined into a common atlas where the parameters are visualized, giving the values of the parameters as the intensity of the color (Jensen et al., 1999).

Structural parameters

A number of measures for the local structure of DNA have been devised, most of which are based dinucleotide or trinucleotide models that have been obtained by fitting either experimental results or theoretical estimates (Pedersen et al., 1998, 2000).

Intrinsic curvature is a property of DNA that is closely related to anomalous gel mobility, as DNA fragments with high intrinsic curvature will migrate slower on polyacrylamide gels than markers with the same length. In this work we have used the CURVATURE programme (Shpigelman et al., 1993), which is based on

a wedge model (Trifonov and Sussman, 1980; Ulanovsky et al., 1986), for prediction of intrinsic curvature. From a set of dinucleotide values for the twist, wedge, and direction angles the three-dimensional path of a 21 bp fragment is calculated. Curvature profiles for longer sequences can thus be calculated using a 21 bp running window. Curves are often encountered upstream of highly expressed genes (Bracco et al., 1989).

Stacking energy relates to the interaction energy between adjacent basepairs in the DNA double helix. The total stacking energy of a DNA segment can be estimated from the set of dinucleotide values determined by quantum mechanical calculations on crystal structures (Ornstein et al., 1978). All stacking energies are negative since base stacking is an energetically favorable interaction that serves to stabilize the double helix. This means that regions with large stacking energies are strongly stabilized and therefore less likely to destack or melt than regions with less negative stacking energies.

The position preference is a measure of helix flexibility based on a set of 32 trinucleotide values giving the log-odds of the minor groove facing outwards when wrapped around a histone octamer (Satchwell et al., 1986). On this scale a value of zero represents no preference of the trinucleotide for specific positions in the nucleosomes, while large absolute values means that the trinucleotide has strong preference. Because large absolute values thereby implies that the sequence is inflexible, a measure of flexibility is obtained by removing the sign from the original trinucleotide values (Pedersen et al., 1998). On that scale low values correspond to high bendability.

Base composition

The trivial way to parameterize the base composition is to simply use the G-, A-, T-, and C-contents. A drawback of this representation is that the four parameters are mutually correlated as they sum to 1. An alternative parameterization for the base composition is A+T and G−C. In addition to being mutually independent measures, they also have the advantage of being easier to interpret in a biological context.

The A+T content is strongly correlated to the structural parameters described above—especially the stacking energy. A+T rich regions usually destack more readily, have a higher intrinsic curvature, and are less flexible. The parameter G−C, known as the GC skew (McLean et al., 1998) reflects a general bias of purines towards the leading strand of DNA replication (Tillier and Collins, 2000). Since the GC skew has almost no correlation to the structural properties of DNA, the A+T content contains nearly all the structural information arising from the mononucleotide composition.

Repeat elements

Repeats are multiple copies of the same sequence at different locations on a piece of DNA. The repeats can be found either by a very accurate method using a basic algorithm which finds the highest degree of homology for an R bp long repeat within a window of length W (Jensen et al., 1999), or by cutting the sequence up

in fragments and using the heuristic alignment algorithm BLAST (Altschul et al., 1997) to find the homologous regions with the length R. The basic algorithm is more accurate than BLAST but it is also computationally demanding, therefore BLAST is used on large sequences. There are two kinds of repeats, a direct repeat is a sequence that is present in at least two copies on the same strand, whilst two copies located on opposite strands will give rise to an inverted repeat.

GenomeAtlases of Pathogenicity Plasmids

A GenomeAtlas can give at quick overview of a given chromosome and thereby be the reason for further analysis of a given organism or a more specific search for a given feature can be made by looking through a collection of atlases. The latter was the case when a study of pathogenicity islands in bacterial plasmids was based on the knowledge of the correlation between pathogenicity islands and variation in AT content, such as the toxin genes in plasmid pO157 from pathogenic *E. coli* strains (Friis et al., 2000). Another example of the correlation between pathogenicity islands and changes in AT content can also be found in the large virulence plasmid of *Shigella flexneri* (GenBank accession number AF348706) (Venkatesan et al., 2001).

The atlas of the *Shigella flexneri* 5a virulence plasmid pWR501 (Figure 7.12) reveals an A+T rich area, which is strongly curved, will destack or melt more

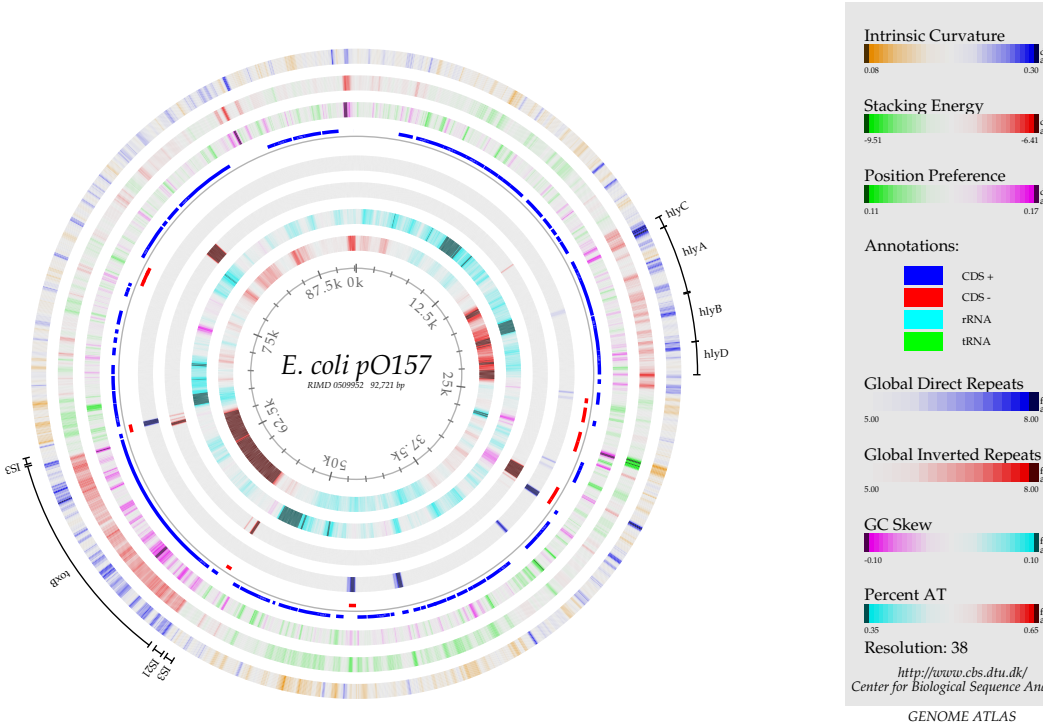


Figure 7.12: **GenomeAtlas for *Shigella flexneri* 5a virulence plasmid pWR501.** The marked regions contain three loci which codes for a total of 34 virulence-related genes.

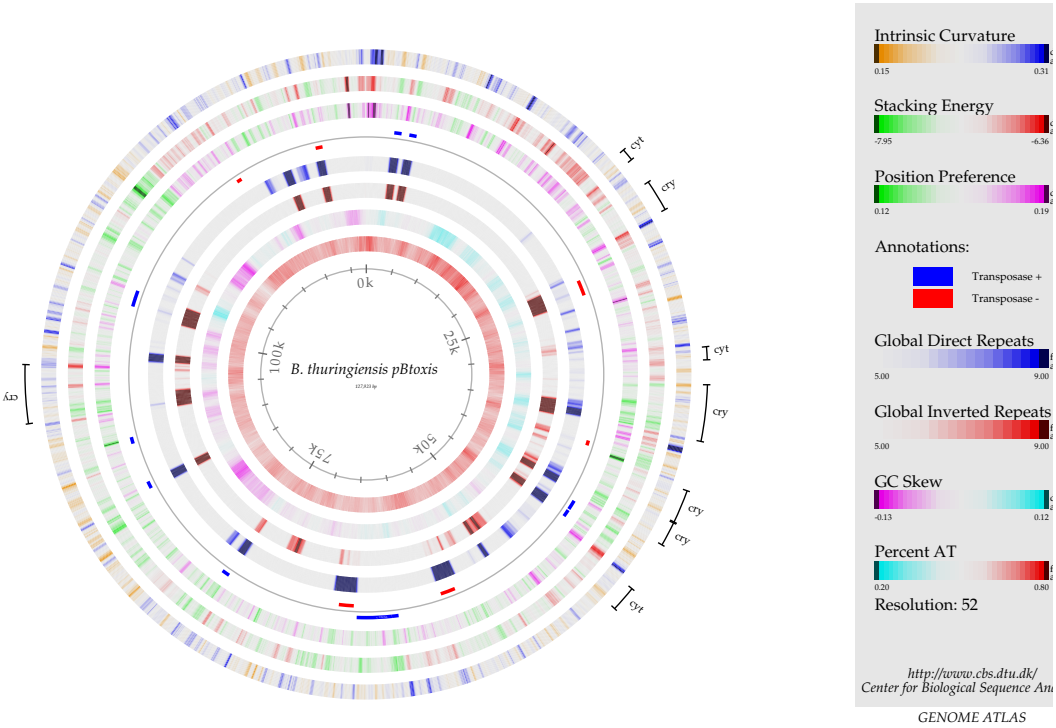


Figure 7.13: **GenomeAtlas for *Bacillus thuringenisis* pBtoxis.** The insecticide activity comes from the marked *cry* and *cyt* genes.

readily than the rest of the plasmid, and is more rigid. This region encodes a locus of genes (*ipa-mxi-spa*) involved in the pathogenic invasion of mammalian cells, and includes a type III secretion pathway (Schuch et al., 1999; Page et al., 2001).

Variations in AT content are obviously not always correlated with the presence of toxic genes; another indication of potential pathogenic regions can be the localization of multiple repeats (especially Insertion Sequence (IS) elements) (Hacker et al., 1997; Hacker and Kaper, 2000). Large numbers of direct and inverted repeats can be seen in Figure 7.12. Typically, global direct repeats account for around three percent or less of most bacterial chromosomes (data not shown, but "GenomeAtlases" for all sequenced genomes can be found on our web page). Many of the repeats (especially the global inverted repeats) are reflective of IS elements. Note that the A+T rich *ipa-mxi-spa* region is the largest region free of repeats in the plasmid.

Another example of a plasmid with many repeats is the plasmid pBtoxis[1] from the spore-forming bacteria *Bacillus thuringiensis* subsp. *israelensis*. Like pWR501, the repeats in pBtoxis are scattered all over the plasmid (Figure 7.13); a search was made for genes from transposable elements like transposases and integrases and by doing a simple BLAST search against SWISS-PROT these

[1]This sequence data were produced by the Microbial Genomes Sequencing Group at the Sanger Institute and can be obtained from ftp://ftp.sanger.ac.uk/pub/pathogens/bti/

genes were located and they were indeed found to be associated with the repeats. In the case of this plasmid the presence of transposable elements was known long before the plasmid was sequenced (Mahillon et al., 1994), but the GenomeAtlas can be used as an easy method for localization of transposons and IS elements.

B. thuringiensis is used in agriculture as an alternative to synthetic chemical pesticides. It produces parasporal crystals that have insecticidal activity, and the genes that are believed to be responsible for this activity are marked in Figure 7.13 (Schnepf et al., 1998). The transposable elements in pBtoxis are at least partly responsible for the high degree of genetic plasticity that makes *B. thuringiensis* adaptable to a variety of environments. However, it should also lead to caution in the use, since *B. thuringiensis*, based on genetic evidence, is from the same species as *Bacillus anthracis* and *Bacillus cereus*, both human pathogens (Helgason et al., 2000).

Some of the repeats that are not associated with genes from transposable elements seem to be copies of *cry*, the gene for the pesticide crystal protein. The three *cyt* genes produce cytolytic delta-endotoxins; the absence of repeats in this area indicates that they are not similar to each other at the nucleotide level.

In the case of the two plasmids presented here the atlas was used as a method to screen large plasmids for signs that indicated the presence of toxic genes; many pathogenic regions within plasmids might not be found in this way, but the atlas serves as a very strong method for initial examination of the sequences (Friis et al., 2000).

A custom made DNA Atlas

The GenomeAtlas is our "standard" atlas, which can capture interesting features of a chromosome. As an example, consider chromosome 1 from the protozoan *Leishmania major*, an intracellular pathogen of the immune system. This chromosome has an unusual organization of its genes, with the 79 protein coding genes being in two large clusters. The first 29 genes are coded on one strand whilst the last 50 genes are on the other strand (Myler et al., 1999). From the GenomeAtlas[2] a correlation between intergenic regions and global repeats can be observed. In order to further investigate the possible relationship between other structural parameters and intergenic regions, we constructed a custom atlas (see Figure 7.14).

Several properties of the chromosome are revealed by the base composition parameters (AT content and GC-skew). The telomeres and a region around 80 kbp have a much higher AT content than the rest of the chromosome. Also a shift in the GC-skew is observed around 80 kbp, correlating with the unusual gene organization. This is in agreement with the region being proposed as the origin of replication (McDonagh et al., 2000).

More direct than inverted repeats are found in this chromosome. Some of these arise from gene duplications—the most obvious example is the two genes around position 240 kbp. Even though gene duplications are observed, the direct repeats still exhibit a slight preference for non-coding regions. This preference is

[2]The atlas can seen at `http://www.cbs.dtu.dk/services/GenomeAtlas/Eukaryotes/Leishmania/major/`

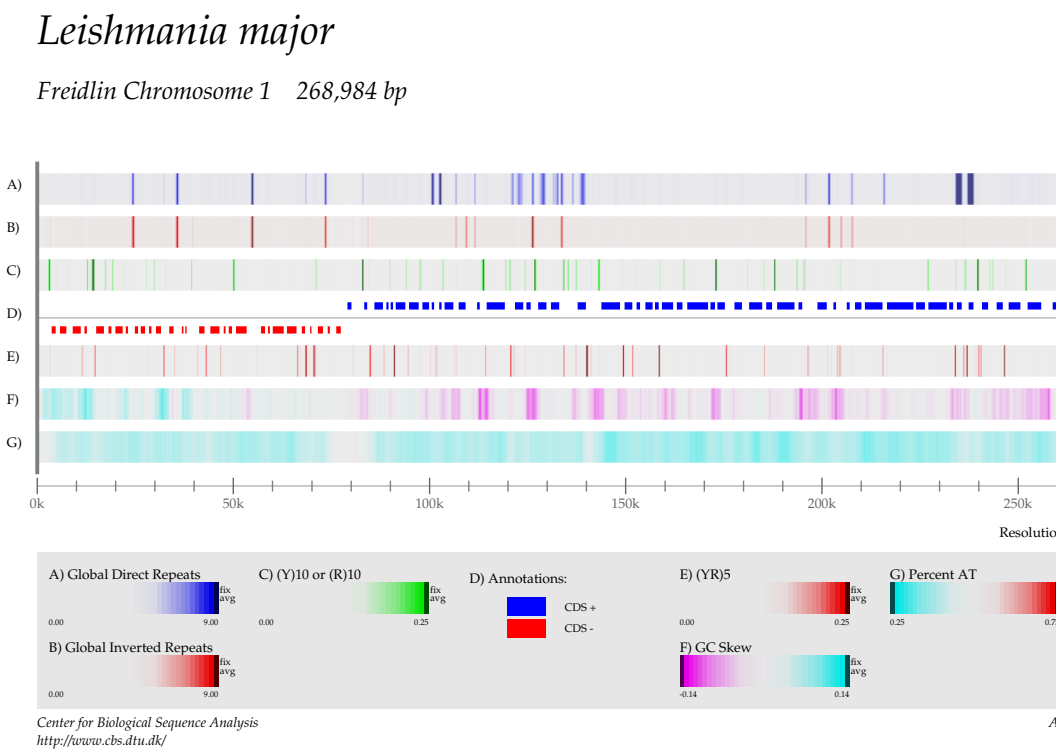


Figure 7.14: **Specialized Atlas for *Leishmania major* chromosome 1.**

much stronger for inverted repeats, which occur exclusively in intergenic regions, as shown in Table 7.7.

The exclusive localization of inverted repeats in intergenic regions prompted an interest in whether other DNA structural elements might also be preferentially positioned within non-coding regions. Runs of purines (or pyrimidines) as well as alternating pyrimidine/purine stretches occur more often than would be expected from the base-composition of *Leishmania major* (Ussery et al., 2002). The location of these regions was visualized by plotting the location of all such stretches of at least 10 bp. Many purine stretches can adopt an A-DNA conformation, whereas pyrimidine/purine stretches that are GC-rich can adopt a Z-DNA conformation. There is a strong preference (about 10-fold, see (Y)10 column in Figure 7.11) for purine stretches in the intergenic regions, while the pyr/pur regions are less strongly correlated with the non-coding DNA.

Table 7.7: **Characteristics of coding and non-coding sequences in *Leishmania major* chromosome 1.**

DNA Property	Length (bp)	% Direct	% Inverted	% (Y)10	% (YR)5	% AT
Coding	140,229	4.5	0.0	1.2	2.6	34.6
Non coding	128,755	6.0	4.6	11.2	6.9	39.4
Whole chromosome	268,984	5.2	2.2	6.0	4.7	36.9

Atlases for Visualizing genome-wide RNA expression

It has often been said that even non-coding DNA is far from a random string of bases. Since helix structure is a function of that same string of bases, this statement must apply to structural features as well. These structural features are suspected of affecting not just the termination of transcription as mentioned earlier, but also the rate of transcription itself.

Almost unheard of five years ago, genome-wide mRNA expression analysis has become mainstream in most major microbiological laboratories. With this technology it is feasible to examine the transcription levels for an entire microbial genome under a broad range of different circumstances, and to some degree reverse engineer regulatory pathways (Spellman et al., 1998).

By visualizing measured levels of transcription in an atlas it becomes possible to examine if correlations exist between the mRNA expression levels and DNA structural properties or base composition. Such correlations could be expected due to chromatin packing (Ussery et al., 2001). Also the relationship between the level of transcription and chromosomal location may reveal interesting aspects (Hughes et al., 2000b).

ExpressionAtlas

The strength of genome-wide RNA expression analysis lies in the ability to simultaneously measure the expression levels of an entire genome. When analyzing several arrays with thousands of genes one is faced with much the same problem as when analyzing whole genome sequences: the sheer amount of data makes it hard to get an overview. The ExpressionAtlas is a way of visualizing expression experiments taking into account chromosomal position and other factors suspected of being involved in transcription, such as DNA structure and repeats.

In the example shown, the average intensities from cDNA arrays (Cho et al., 1998) were used as an estimate of the constitutive expression levels of genes in *Saccharomyces cerevisiae*. Alternatively log-fold changes could be plotted to highlight regulated genes. We chose average intensities to ensure comparability to the predicted expression levels also displayed on the atlas.

Neural networks trained on average expression values from *E. coli* microarray experiments predicted the expression level of each gene. The predicted levels of expression were normalized to a range from 0 to 1. As input to the neural networks the trinucleotide frequencies of the coding regions were used. These 64 frequencies were calculated without taking the reading frame into account. This representation was chosen because the majority of DNA structural properties can be captured at the trinucleotide level. In this way we can capture possible correlation between the structural properties of the coding DNA and the expression levels.

Both the experimentally measured and the predicted expression levels are displayed in the atlas in Figure 7.15, together with position preference, global repeats and AT content. The AT content and the global repeats were included to give a general view of the composition of the chromosome, whereas the position

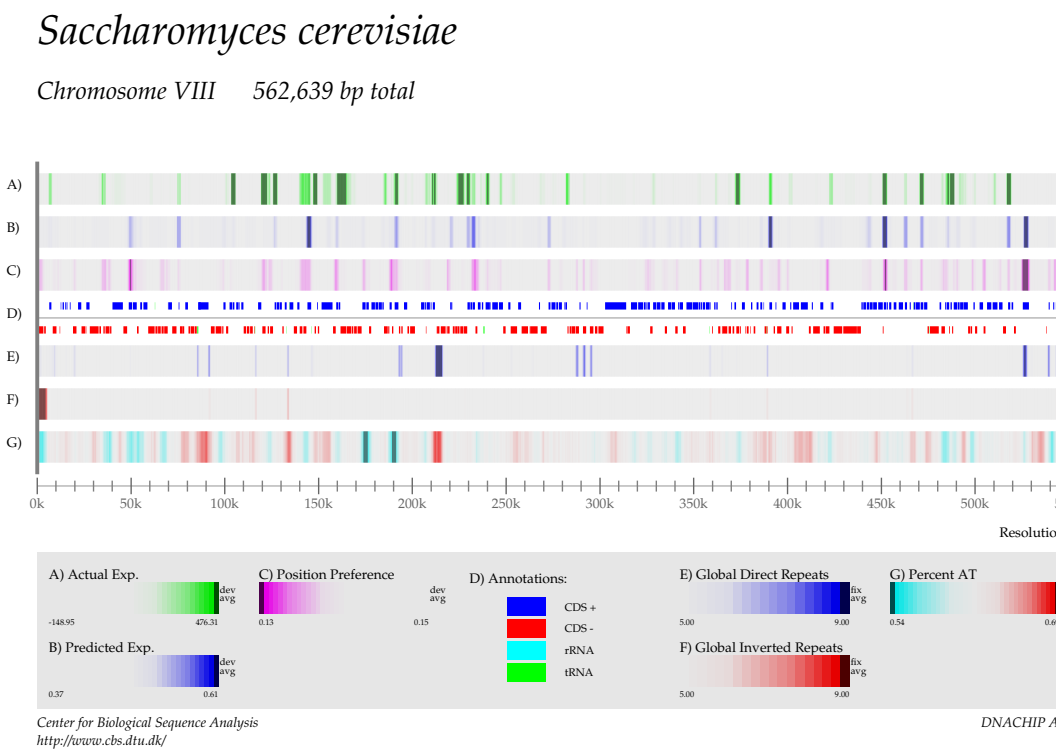


Figure 7.15: **The ExpressionAtlas of *S. cerevisiae* chromosome sko02mim.** Lane A shows the average intensities from cDNA arrays, indicating the constitutively expressed genes, whilst lane B is the predicted expression.

preference, being a measure of the flexibility of the double helix, is expected to be correlated with the expression levels (Pedersen et al., 2000). The inverted repeats clearly mark the telomeric regions.

Comparing the actual expression levels with the levels predicted by neural networks reveals a strong correlation between the two. A similar, albeit weaker, correlation is observed between expression levels and the position preference measure. The fact that neural networks trained on the prokaryote *E. coli* data can predict highly expressed genes in a eukaryote implies the existence of universal DNA properties that influence transcription. The correlation with the position preference measure suggests that helix flexibility plays a part in this. Speculations on such generic features of expressed genes have been proposed before (Sharp and Li, 1987).

Atlases for Visualizing Global Prediction of Protein Function

By looking at the ExpressionAtlas it is possible to identify regions with genes that are highly expressed (and possibly regulated) under one or more experimental conditions. It is at this point obvious to ask what the function of these genes might be.

Unfortunately the function of a large fraction genes remains unknown in most fully sequenced chromosomes. Of the 30,000 to 50,000 genes believed to be present in the human genome no more than 40-60% can be assigned a functional role based on homology to known proteins. Even though the situation is a bit more favorable when looking at simpler model organisms like *S. cerevisiae* and *C. elegans*, the function of more than 30% of the predicted protein sequences still remains unknown.

In newly sequenced chromosomes most of the functional annotation of genes is based on homology inference. Using methods such as BLAST (Altschul et al., 1997) homologous proteins are identified by sequence similarity and the function is inferred from the knowledge about the homologs. However it is usually the case that somewhere from 30–50% of the proteins give no matches to proteins of known function. These are known as "orphan" proteins.

Traditionally, protein function has been viewed as something directly related to the conformation of the polypeptide chain. However, as the three-dimensional structure currently is quite hard to calculate from the sequence (Lesk et al., 2001), a computational strategy for the elucidation of orphan protein function may benefit also from the prediction of functional attributes which are more directly related to the linear sequence of amino acids.

Our approach to function prediction is based on the fact that a protein is not alone when performing its biological task. As it will have to operate using the same cellular machinery for modification and sorting as all the other proteins do, one can expect some conservation of essential types of post-translational modifications (PTMs). Because reasonably precise methods for prediction of PTMs from sequence exist today, our prediction method which integrates such relevant features to assign orphan protein to functional class, can be applied to all proteins where the sequence in known (Jensen et al., 2002; Gupta et al., 2002). This is in contrast to methods that rely on clustering of co-expressed genes (Eisen et al., 1998), prediction of gene fusions and/or phylogenetic profiles (Marcotte et al., 1999a,b; Hughes et al., 2000a; Pellegrini et al., 1999).

For any function prediction method, the ability to correctly assign the relationship depends strongly on the function classification scheme used. We predict a scheme of twelve cellular functions that is closely related to the fourteen class classification originally proposed by Riley for the *E. coli* genome (Riley, 1993). The system consists of an ensemble of neural networks for each functional category, each neural network having a different combination predicted protein features as its input. The networks were trained exclusively on human protein sequences, but perform well on a wide selection of eukaryotes (including *S. cerevisiae*). For each protein sequence the outputs of these neural networks are subsequently combined into a probability for each category.

We have applied this software to all predicted protein sequences from *S. cerevisiae* chromosome sko02mim. Based on our performance estimates of the method on *S. cerevisiae* sequences, we have selected a subset of eight categories out of the original twelve category system. The probabilistic scores of each protein sequence were mapped onto the position in the chromosome of the corresponding gene. Figure 7.16 shows the resulting FunctionAtlas along with the actual expression levels also shown in the ExpressionAtlas.

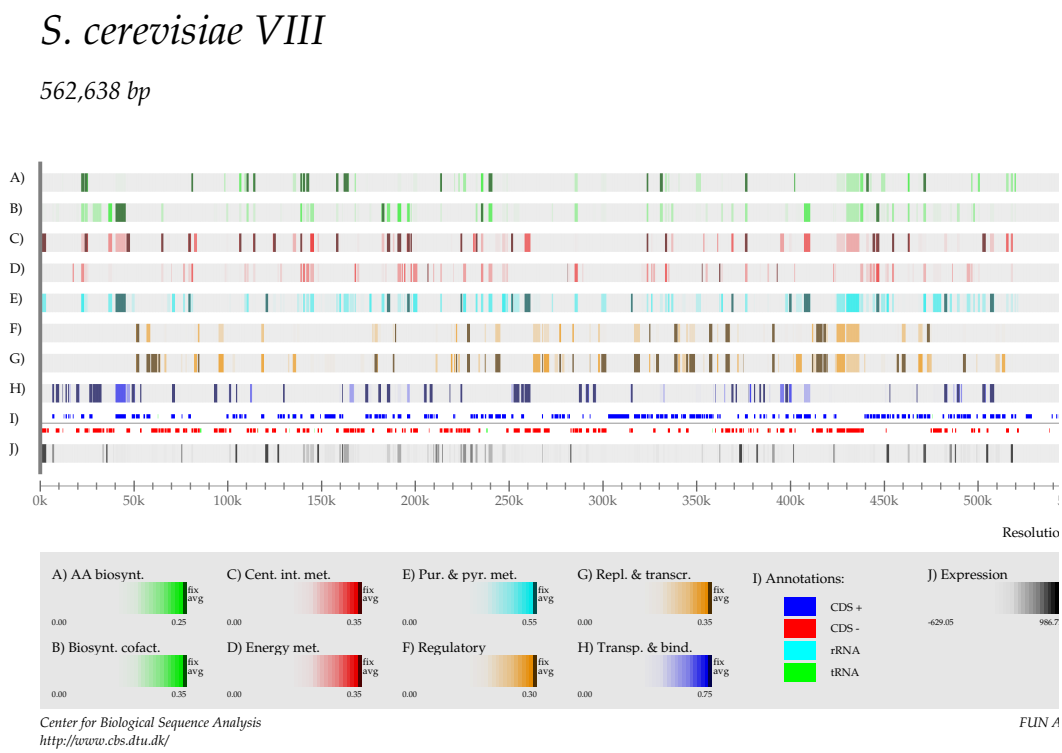


Figure 7.16: **The FunctionAtlas of *S. cerevisiae* chromosome sko02mim.** Lane A: Amino acid biosynthesis, Lane B: Biosynthesis of co-factors, Lane C: Central intermediary metabolism, Lane D: Energy metabolism, Lane E: Purine and pyrimidine metabolism, Lane F: Regulatory function, Lane G: Replication and transcription, Lane H: Transport and binding, Lane J: Average intensity from cDNA experiments, (the same as in Figure 7.15).

One feature that is visible from a FunctionAtlas is clusters of genes with related functions. Examples of this include the regions 10 kbp–50 kbp and 250 kbp–260 kbp that contain very large numbers of predicted transport and binding proteins. The regions 50 kbp–60 kbp, 335 kbp–350 kbp and 410 kbp–420 kbp that are predicted to contain a large number of genes involved in replication or transcription, several of which are likely to serve a regulatory role according to our predictions. Since genes of related function are known to often cluster (although the extent varies from organism to organism) predicted functional clusters can be trusted more than individual predictions. If the function of some of the genes within a cluster is known and in agreement with the prediction—as is the case for several of the mentioned clusters—this obviously adds to the evidence.

Another possibility is to correlate the predicted protein function to expression data. Close inspection of Figure 7.16 reveals that many of the constitutively highly expressed genes are predicted to be involved in energy metabolism although this is overall a quite rare category. A hypergeometric test of the underlying data verifies that this correlation is indeed significant at a 95% confidence level. A large number of highly expressed transcripts for proteins involved in replication and transcription can also be identified although no correlation between function and expression level is found in this case.

Concluding remarks

In summary, we have shown several different applications of Atlases for visualizing different type of information, within the context of the whole plasmid or chromosome. Essentially, any type of information concerning the DNA, RNA or protein can be plotted along the chromosome, allowing for rapid analysis of global properties in a serendipitous manner. The atlas gives the researcher the option to view the calculated or experimentally measured data in a position dependent way and thereby see correlations between a feature and its position or variation in a feature within the chromosome.

The atlas can be used to spot variation in different features within a region but not all the information can be viewed at the same time. For a bacterial chromosome of the same size as *E. coli* only features with approximately the same size as a gene can be observed, this means that some of the features showed for *L. major* with repeats in intergenic regions would not be visible in the *E. coli* genome. If variation in a smaller scale is to be seen for a large chromosome a shorter region should be visualized.

Acknowledgments

The authors would like to acknowledge and thank the people at CBS for their help, in particular Ramneek Gupta, Steen Knudsen, Anders Krogh, and Anders Gorm Pedersen. This work was supported by a grant from the Danish National Research Foundation.

7.10.1 Visualization enzyme and Gene Ontology categories

In addition to the atlas of cellular roles (see Figure 7.16 on page 180), two extra types of atlases covering the enzyme classes and the Gene Ontology categories have also been developed. These atlases are merely displaying different probabilities from the ProtFun output.

From studying a large number of atlases (data not shown) it appears that clusters of genes related in the enzyme classification are hardly ever seen. This is consistent with the very weak autocorrelation observed for enzyme classes (see Figure 7.9 on page 163). In contrast, clusters of *receptors* and *stress response* genes would be expected at least in human chromosomes. Unfortunately, the human genome—although "completely sequenced"—is still only available as a large number of contigs and maps, which is not well suited for making atlases.

7.11 Many possible uses for the ProtFun approach

In this chapter, the wide range of possible applications for our method has been illustrated. There is good reason to believe, that ProtFun-like predictors can be developed for most broad protein classes in eukaryotes. We have also enjoyed some success with predicting enzymatic function for archaeal proteins. Combined with the cross-species evaluation presented in Chapter 5, this gives reason to believe that the approach may also work for eubacteria. Furthermore, it has been shown how the function prediction method can be used for visualizing the chromosome wide distribution of gene functions.

Chapter 8

Room for improvement

8.1 Making use of DNA data

The ProtFun method has been developed to only require the protein sequence as input. This decision was based on two reasons. First of all, most of the information about the function of a protein should be expected to be in the protein sequence itself. Secondly, it allows the method to be applied to all proteins of unknown function.

There might be more information available in the DNA sequence than what is reflected in the protein sequence. This could possibly be utilized to improve the ProtFun prediction method in cases where this additional information is indeed available—which is the case for the many proteins of unknown function identified in complete genome sequencing projects.

8.1.1 Codon usage

It is well known that the genetic code is degenerate in the sense that several codons exist that encode the same amino acid. The so-called synonymous codons are not all used at the same frequencies as some codons are preferred over others. These codon preferences vary from organism to organism.

Not surprisingly the preferences in codon usage turn out to be correlated to the tRNA pool of the organism so that the commonly used codons correspond to the largest tRNA concentrations. As a result of this, the most highly expressed genes will tend to use only the preferred codons in order to allow for rapid translation of the messenger RNA into protein. Based on resulting differences in codon usage between highly expressed genes and all other genes, a measure called codon adaption index (CAI) can be defined (Sharp and Li, 1987). The CAI value for a gene is therefore an estimate of the constitutive expression of the gene.

As proteins from certain functional classes (e.g. *translation*) are likely to be more highly expressed than proteins from other classes (e.g. *regulatory functions*), one should expect that the codon adaptation index will differ between different classes for proteins. This has indeed previously been shown to be the case in *E. coli* (Karlin et al., 1998).

Figure 8.1 shows the distributions of log-CAI values for the positive and negative *S. cerevisiae* data sets for cellular roles (Paper IV). It is observed that

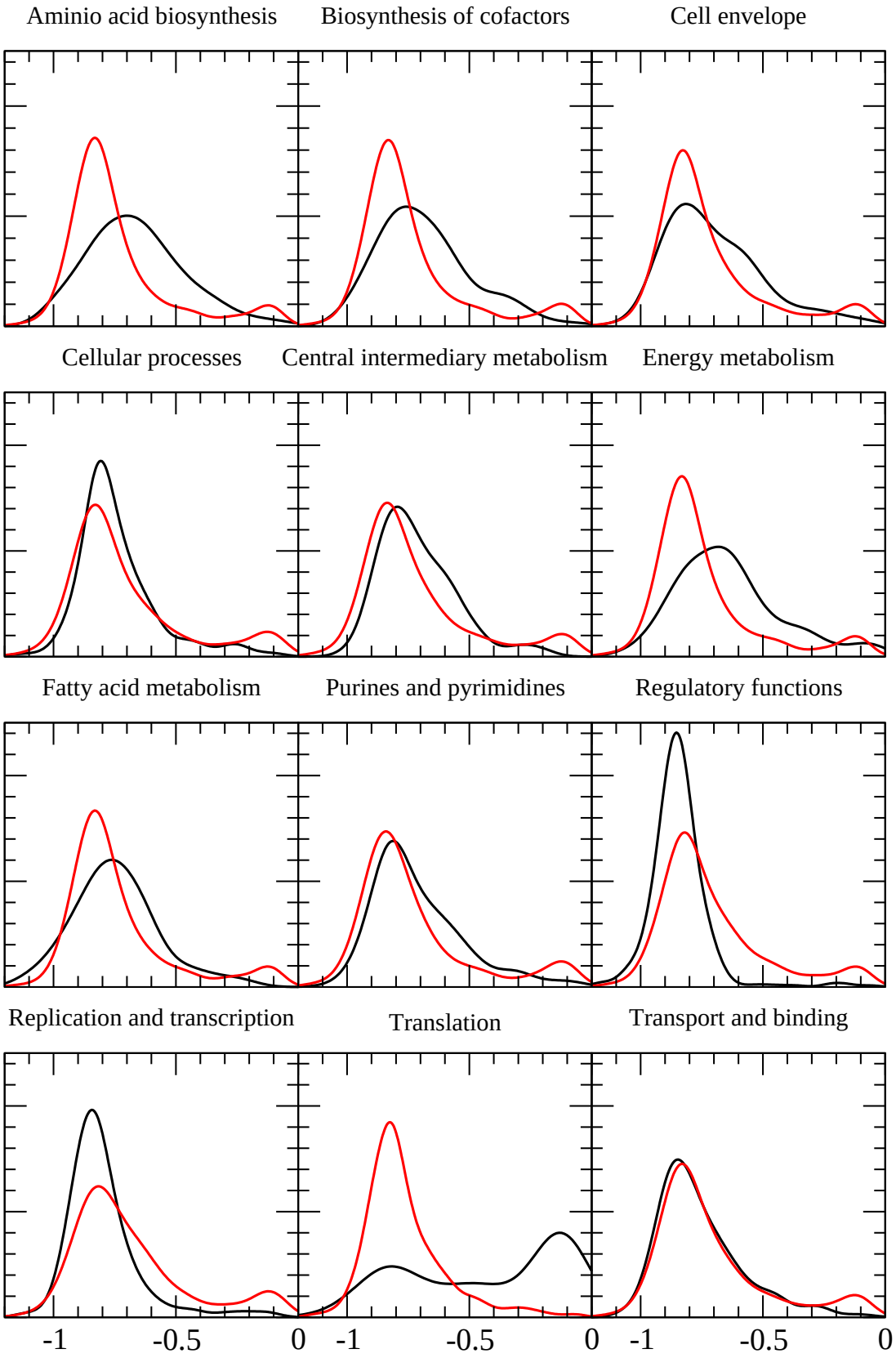


Figure 8.1: **Differences in codon adaptation index for functional classes of genes.** The codon adaptation index is shown for ORFs encoding the positive (black) and negative (red) *S. cerevisiae* sets for each cellular role category.

proteins involved in translation and several types of metabolism have particularly high CAI values whereas regulatory proteins have low CAI values. It is therefore likely that the CAI value of a protein could be used as an additional input feature to ProtFun and thereby improve performance.

8.1.2 DNA structure

In addition to calculating CAI values, it is also possible to estimate DNA structural properties of each gene given entire cDNA sequences. Such properties have previously been used for revealing regions containing genes involved in pathogenicity (Friis et al., 2000), and could be correlated with other classes of genes as well.

Figure 8.2 shows one estimated structural feature, the *position preference* (Satchwell et al., 1986), for *S. cerevisiae* genes encoding proteins involved in translation compared to other genes. The difference between the two distributions is highly significant according to a Kolmogorov–Smirnov test. The position preference score for a gene gives a measure of the anisotropic flexibility of the DNA, and has previously been suggested to correlate with high expression (Pedersen et al., 2000). This agrees with our observation since genes related to translation are known to be highly expressed.

8.1.3 Promoter elements

I have previously co-authored a paper in which a method for finding correlations between upstream patterns and protein function was presented (Jensen and Knudsen, 2000). While this paper focused on identifying the upstream patterns

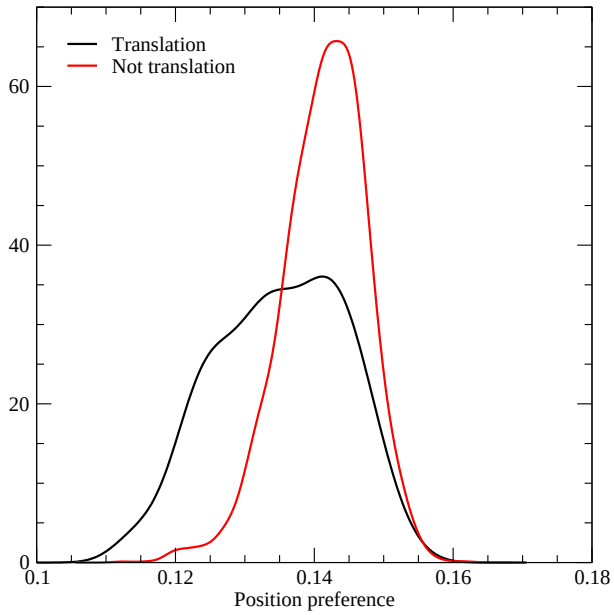


Figure 8.2: **Higher DNA flexibility of ORFs encoding proteins involved in translation.** The *position preference* measure of anisotropic flexibility (Satchwell et al., 1986) is compared for genes involved in translation and other genes.

Table 8.1: **Predicted *S. cerevisiae* promoter elements correlated to particular cellular roles.** Patterns were discovered using the hypergeometric method described by Jensen and Knudsen (2000). The statistical significance is reported as $-log(\alpha)$ in the $p\alpha$ column. Also reported is the correlation coefficient (Corr.) and the number of true positives (TP) and false positives (FP).

	Pattern	$p\alpha$	Corr.	TP	FP	Function
Cbf1p-Met2p-Met28p	...CACGTG.....	> 10	0.096	26	147	Amino acid biosynthesis
	...CACGTGA....	> 10	0.090	17	81	
Methionine element	..AATGACT.....	> 10	0.093	17	78	Amino acid biosynthesis
	...ATGACT.....	6.8	0.127	44	251	
	...ATGACTA....	> 10	0.096	15	61	
	...ATGACTC....	> 10	0.169	22	49	
	...TTGACTC....	> 10	0.100	13	45	
	...ATGACTCA...	> 10	0.163	12	14	
	....TGACTCA...	> 10	0.222	28	45	
	....TGACTCT...	> 10	0.118	17	56	
	.....GACTCA...	> 10	0.158	37	145	
	.....GACTCT...	> 10	0.106	29	157	
	.....GACTCTT..	> 10	0.123	19	64	
	.....GACTCAT..	> 10	0.095	12	42	
	...TATCGTTT...	> 10	0.113	13	37	Amino acid biosynthesis
	...CGTATAA....	> 10	0.110	16	70	Biosynthesis of cofactors
	...CACGTGA....	> 10	0.100	15	72	Biosynthesis of cofactors
TATA box	...TATATAAG...	3.5	0.112	40	113	Energy Metabolism
	....ATATAA....	5.2	0.114	194	1066	
HOMOL1	TACATCC.......	> 10	0.127	27	54	Translation
	.ACATCCG......	> 10	0.155	23	27	
	..CATCCGT.....	> 10	0.120	22	41	
	..AATCCGT.....	> 10	0.115	18	30	
	..AATCCGTA....	> 10	0.119	12	12	
	...ATCCGTA....	> 10	0.171	32	44	
	....TCCGTAC...	> 10	0.189	34	40	
	.....CCGTACAC.	6.3	0.122	10	7	
	......CGTACAT.	> 10	0.125	31	71	
	.......GTACAT.	4.7	0.110	78	337	
	.......GTACATT	3.9	0.110	38	116	
RPG	.AACACCCA.....	> 10	0.119	12	12	Translation
	..ACACCCA.....	> 10	0.120	26	55	
	...CACCCATA...	> 10	0.140	18	20	
	....ACCCATA...	> 10	0.145	34	67	
	.....CCCAT....	5.0	0.107	118	609	
	.....CCCATAC..	> 10	0.164	31	45	
	......CCATAC..	5.5	0.117	54	188	
	......CCATACA.	> 10	0.151	37	73	
Curved element	..AAAAATTT....	3.3	0.109	68	280	Translation
	...AAAATTTT...	3.4	0.110	69	284	

rather than the protein function, the correlations found between promoter elements and protein function of course hold both ways. One could thus use the occurrences of oligomers in the upstream regions of genes as hints to the function of proteins encoded.

A set of 500 bp 5'UTR sequence of yeast was combined with EUCLID assignments of cellular roles to construct both a positive and a negative set of promoter regions for each cellular role category. For each category, the saco_patterns soft-

ware (Jensen and Knudsen, 2000) was used to identify conserved DNA patterns that are significantly overrepresented in the positive set compared to the negative set according to a hypergeometric test. The results can be seen in Table 8.1.

Significant patterns were only discovered for five of the categories. The majority of these patterns are well described. Two of the three patterns found in the upstream regions of amino acid biosynthesis genes have previously been associated with methionine biosynthesis (Jensen and Knudsen, 2000). One of them is the consensus recognition sequence for the Cbf1p-Met2p-Met28p complex (O'Connell and Baker, 1992).

The only pattern found correlated to the *energy metabolism* category was the yeast consensus sequence for TATA-box motifs, which apparently occur more frequently in the upstream regions of these genes. This may be explained by the high expression levels observed for many energy metabolism genes.

Three patterns were found to be significantly correlated to genes from the translation category: the HOMOL1 (Larkin et al., 1987) and RPG (Vignais et al., 1987, 1990) consensus sequences that occur in the upstream regions of most ribosomal protein coding genes and the putative curved promoter element previously associated with protein synthesis (Jensen and Knudsen, 2000).

The approach described is likely to be most valuable for prokaryotes and simple eukaryotes where the regulatory regions are fairly short and localized close to the corresponding genes. In higher eukaryotes like humans where identification of promoter elements is much harder, this approach is likely to be of limited value.

8.2 Predicting protein function for complete genomes

If the complete genomic sequence of an organism is available, it should be possible to make use of this when predicting protein function. Currently all predictions are being made by the neural networks on a "gene by gene" basis.

8.2.1 Override predictions by database searches

When ProtFun is run on a complete genome, it will usually be an attempt to annotate a putative function to as many proteins as possible. It would thus make sense to automatically take into account matches to known protein families and include them in the prediction output. When present, these matches could be used to override the predictions made by neural networks to attain the best possible prediction for each protein. Alternatively the matches could simply be shown as additional information. This could be elegantly implemented using InterPro-scan, which could also provide additional features for neural networks as already described.

8.2.2 Making use of *in silico* functional links

Throughout this thesis a number of computational methods for linking together proteins of similar function have been mentioned: The Rosetta stone method

(gene fusions), phylogenetic profiles, and chromosomal localization. While many might view these methods as competitors to ProtFun, it would make much more sense to use these methods in conjunction with ProtFun, as they could improve the prediction quality in several ways. Used together with the search for conserved protein families described above, they would obviously provide an essentially independent prediction of protein function to a large number of proteins.

Even when no links to proteins assigned based on sequence similarity exist, the links can improve the quality of prediction. This can be done because it allows ProtFun predictions of several individual proteins to be combined into a more reliable prediction—if a group of proteins are predicted to be functionally linked and ProtFun assigns them to the same category, it adds confidence to these predictions.

8.2.3 Allow integration of additional experimental data

One possible step further would be to also allow the inclusion of additional data such as microarray expression studies and/or protein–protein interaction screens. These types of data are, as have been illustrated, very similar to the *in silico* functional links and can thus be included in much the same way. Integrating different data sources is in my opinion one of the most important challenges in systems biology.

8.3 The future of function prediction

Given the difficulties encountered when assigning protein function from homologous proteins, the *ab initio* function prediction problem is unlikely to be solved in a foreseeable future. Still, significant progress has been made in the past few years where several methods have been developed which do not rely on direct sequence similarity to proteins of known function.

All the methods have one thing in common: they make use of the fact that proteins interact with its environment and in particular with other proteins. Several of the methods developed rely on the prediction of protein–protein interactions to infer function of uncharacterized proteins. The method presented in this thesis (ProtFun) makes use of the cellular context in a different way. Central to the method is the idea that proteins performing some function will have to perform this function in the same context. They can therefore be expected to share certain characteristics even if they are not evolutionarily related.

An important aspect of the ProtFun prediction method is that it makes use of biologically relevant features as input. This allows the method to be used for more than merely just predicting protein function—the neural networks and the features they use can be analyzed to make biological discoveries. While it might be possible to predict function equally well using for instance k-mer frequencies as input, such a prediction method would be unlikely to give any insight into how proteins work. In my opinion it is more important that future function prediction methods capture the biology of the problem better rather than simply outperforming todays methods.

Bibliography

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402.

Andersson, S. G. E., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T. S., Alsmark, U. C. M., Podowski, R. M., Naslund, A. K., Eriksson, A., Winkler, H. H., and Kurland, C. G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, 396:133–140.

Andrade, M. A., Brown, N. P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C., and Sander, C. (1999a). Automated genome sequence analysis and annotation. *Bioinformatics*, 15:391–412.

Andrade, M. A., O'Donoghue, S. I., and Rost, B. (1998). Adaption of protein surfaces to subcellular location. *J. Mol. Biol.*, 276:517–525.

Andrade, M. A., Ouzounis, C., Sander, C., Tamames, J., and Valencia, A. (1999b). Functional classes in the three domains of life. *J. Mol. Evol.*, 49:551–557.

Anfinsen, C. B. (1973). Principles than govern the folding of protein chains. *Science*, 181:223–230.

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J., and Zdobnov, E. M. (2000). InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, 16:1145–1150.

Apweiler, R., Hermjakob, H., and Sharon, N. (1999). On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta.*, 1473:4–8.

Ashburner, M. (1998). On the representation of gene function in genetic databases. In *Proc., Intelligent Systems for Molecular Biology*, volume 6, Menlo Park, CA. AAAI Press.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Attwood, T. K. (2000). The quest to deduce protein function from sequence: the role of pattern databases. *Int. J. Biochem. Cell Biol.*, 32:139–155.

Attwood, T. K. and Miller, C. J. (2001). Which craft is best in bioinformatics? *Computers and Chemistry*, 25:327–337.

Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, 28:45–48.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18:298–305.

Bansal, A. K. and Meyer, T. E. (2002). Evolutionary analysis by whole-genome comparisons. *J. Bacteriol.*, 184:2260–2272.

Basrai, M. A., Hieter, P., and Boeke, J. D. (1997). Small open reading frames: beautiful needles in the haystack. *Genome Res.*, 7:768–771.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. L. (2002). The Pfam protein families database. *Nucleic Acids Res.*, 30:276–280.

Bender, L., Lo, H. S., Lee, H., Kokojan, V., and Peterson, V. (1996). Associations among PH and SH3 domain-containing proteins and rho-type GTPases in yeast. *J. Cell Biol.*, 133:879–894.

Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Prot. Eng.*, 7:1323–1332.

Benner, S. A. and Gaucher, E. A. (2001). Evolution, language and analogy in functional genomics. *Trends in Genetics*, 17:414–418.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. (2002). GenBank. *Nucleic Acids Res.*, 30:17–20.

Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quali, A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.-H., Kieser, T.,

Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M.-A., Rutherford, K., Rutter, S., Seeger, K., Saunder, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, K., Barrell, B. G., Parkhill, K., and Hopwood, D. A. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, 417:141–147.

Bertram, P. G., Choi, J. H., Carvalho, J., Ai, W., Zeng, C., Chan, T. F., and Zheng, X. F. (2002). Tripartite regulation of gln3p by TOR, ure2p, and phosphatases. *J. Biol. Chem.*, 275:35727–35733.

Birney, E., Bateman, A., Clamp, M. E., and Hubbard, T. J. (2001). Mining the draft human genome. *Nature*, 409:827–328.

Blaschke, C., Andrade, M. A., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein–protein interactions. In *Proc. of Intelligent Systems for Molecular Biology*, volume 7, pages 60–67, Menlo Park, CA. AAAI Press.

Blattner, F. R., Plunkett 3rd, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, 277:1453–1474.

Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, 294:1351–1362.

Bock, J. R. and Gough, D. A. (2001). Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17:45–460.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S. D., and Sorokin, A. (2001). The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. lactis IL1403. *Genome Res.*, 11:731–753.

Bolten, E., Schliep, A., Schneckener, S., Schomburg, D., and Schrader, R. (2001). Clustering protein sequences—structure prediction by transitive homology. *Bioinformatics*, 17:935–941.

Bonetta, L. (2002). Systems biology—the new R&D buzzword? *Nature Medicine*, 8:315–316.

Bork, P. (2000). Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res.*, 10:398–400.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.*, 283:707–725.

Bracco, L., Kotlarz, D., Kolb, A., Diekmann, S., and Buc, H. (1989). Synthetic curved DNA sequences can act as transcriptional activators in *Escherichia coli*. *EMBO J.*, 8:4289–4296.

Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. (1998). Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.*, 8:1202–1215.

Brazma, A., Vilo, J., Ukkonen, E., and Valtonen, K. (1997). Data mining for regulatory elements in yeast genome. *ISMB*, 5:65–74.

Breeden, L. (2000). Cyclin transcription: Timing is everything. *Curr. Biol.*, 10:R586–R588.

Brown, D. R. (2001). Copper and prion disease. *Brain Research Bulletin*, 55:165–173.

Brown, D. R., Qin, K., Herms, J. W., Madlung, A., Manson, J., Strome, R., Fraser, P. E., Kruck, T., von Bohlen, A., Schulz-Schaeffer, W., Giese, A., Westaway, D., and Kretzschmar, H. (1997). The cellular prion protein binds copper in vivo. *Nature*, 390:684–687.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, Jr., M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U.S.A.*, 97:262–267.

Brunak, S., Engelbrecht, J., and Knudsen, S. (1990). Cleaning up gene databases. *Nature*, 343:123.

Brunak, S., Engelbrecht, J., and Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, 220:49–65.

Bult, J. C., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., and Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, 273:1058–1073.

Callan, H. G. (1972). Replication of DNA in the chromosomes of eukaryotes. *Proc. R. Soc. Lond.*, 181:19–41.

Cambillau, C. and Claverie, J. M. (2000). Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.*, 275:32383–32386.

Casari, G., Ouzounis, C., Valencia, A., and Sander, C. (1996). GeneQuiz-II: Automatic function assignment for genome sequence analysis. In *Proceedings of the First Annual Pacific Symposium on Biocomputing*, pages 707–709, Hawaii. World Scientific.

Chen, C. and Colley, K. J. (2000). Minimal structural and glycosylation requirements for ST6Gal I activity and trafficking. *Glycobiology*, 10:531–583.

Cheng, H. H., Muhlrad, P. J., Hoyt, M. A., and Echols, H. (1988). Cleavage of the cII protein of phage lambda by purified HflA protease: control of the switch between lysis and lysogeny. *Proc. Natl. Acad. Sci. U.S.A.*, 85:7882–7886.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, 2:65–73.

Chou, K.-C. and Elrod, D. W. (1999). Protein subcellular location prediction. *Protein Eng.*, 12:107–118.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, L. (1998). The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705.

Claverie, J.-M. (2001). What if there are only 30,000 human genes. *Science*, 291:1255–1257.

Cohen, P. (2000). The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem. Sci.*, 25:596–601.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry 3rd, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osbourne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Ruter, S., Seeger, K., Skelton, J., Squares, R., Sulston, J. E., Taylor, K., Whitehead, S., and Barrell, B. G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393:537–544.

Collinge, J., Palmer, M. S., Sidle, K. C., Hill, A. F., Gowland, I., Meads, J., Asante, E., Bradley, R., Doey, L. J., and Lantos, P. L. (1995). Unaltered susceptibility to bse in transgenic mice expressing human prion protein. *Nature*, 378:779–783.

Collobert, R. and Bengio, S. (2000). Support vector machines for large-scale regression problems. Technical Report IDIAP-RR-00-17, IDIAP, Switzerland.

Comer, F. I. and Hart, G. W. (1999). O-GlcNAc and the control of gene expression. *Biochim. Biophys. Acta.*, 1473:161–171.

Comer, F. I. and Hart, G. W. (2000). O-glycosylation of nuclear and cytosolic proteins: Dynamic interplay between O-glcnac and O-phosphate. *J. Biol. Chem.*, 275:29179–29182.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, 23:324–328.

Das, S., Yu, L., Gaitatzes, C., Rogers, R., Freeman, J., Bienkowska, J., Adams, R. M., and Smith, T. F. (1997). Biology's new rosetta stone. *Nature*, 385:29–30.

Davis, C. A., Grate, L., Spingola, M., and Ares, Jr., M. (2000). Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res.*, 28:1700–1706.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In Dayhoff, M., editor, *Atlas of Protein Sequence and Structure*, volume 5, pages 345–352. National Biochemical Research Foundation, Washington D.C.

Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: Two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, 1:349–356.

Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., Huber, R., Feldman, R. A., Short, J. M., Olsen, G. J., and Swanson, R. V. (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature*, 392:353–358.

Deppenmeier, U., Johann, A., Hartsch, T., Merkl, R., Schmitz, R. A., Martinez-Arias, R., Henne, A., Wiezer, A., Bäumer, S., Jacobi, S., Brüggemann, H., Leinard, T., Christmann, A., Bömeke, M., Steckel, S., Bhattacharyya, A., Lykidis, A., Overbeek, R., Klenk, H.-P., Gunsalus, R. P., Fritz, H.-J., and Gottschalk, G. (2002). The genome of *Methanosarcina mazei*: Evidence for lateral gene transfer between bacteria and archaea. *J. Mol. Microbiol. Biotechnol.*, 4:453–461.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686.

Devos, D. and Valencia, A. (2000). Practical limits of function prediction. *Proteins*, 41:98–107.

Dunham, I., Shimizu, N., Roe, B. A., and Chissoe, S. *et al.*. (1999). The DNA sequence of human chromosome 22. *Nature*, 402:489–495.

Durocher, D. and Jackson, S. P. (2002). The FHA domain. *FEBS Lett.*, 513:58–66.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–14868.

Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, 405:823–826.

Eisenhaber, B., Bork, P., and Eisenhaber, F. (1999). Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.*, 292:741–758.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Prediction of subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, 300:1005–1016.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90.

Enzyme Nomenclature (1965). *Recommendations (1964) of the International Union of Biochemistry on the Nomenclature and Classification of Enzymes, Together with their Units and the Symbols of Enzyme Kinetics*. Elsevier, Amsterdam.

Enzyme Nomenclature (1992). *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, San Diego, CA.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and A, B. (2002). The PROSITE database, its status in 2002. *Nucleic Acids Res.*, 30:235–238.

Fischer, D. and Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics*, 15:759–762.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *J. Biol. Chem.*, 19:99–113.

Fitch, W. M. (2000). Homology a personal view on some of the problems. *Trends in Genetics*, 16:227–231.

Fitz-Gibbon, S. T., Ladner, H., Kim, U. J., Stetter, K. O., Simon, M. I., and Miller, J. H. (2002). Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. *Proc. Natl. Acad. Sci. U.S.A.*, 99:984–989.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Georghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., and Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269:496–512.

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., Peterson, J., Kerlavage, A. R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M. D., Gocayne, J., Weidman, J., Utterback, T., McDonald, L., Artiach, P., Bowman, C., Garland, S., Cotton, M. D., Horst, K., Roberts, K., Hatch, B., Smith, H. O., and Venter, J. C. (1997). Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, 390:580–586.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleishmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Petersen, S. N., Smith, H. O., Hutchison, C. A., and Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270:397–403.

Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R., Gwinn, M., Hickey, E. K., Clayton, R., Ketchum, K. A., Sodergren, E., Hardham, J. M., McLeod, M. P., Salzberg, S., Peterson, J., Khalak, H., Richardson, D., Howell, J. K., Chidambaram, M., Utterback, T., McDonald, L., Artiach, P., Bowman, C., Cotton, M. D., Fujii, C., Hatch, B., Roberts, K., Sandusky, M., Weidman, J., Smith, H. O., and Venter, J. C. (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*, 281:375–388.

Friis, C., Jensen, L. J., and Ussery, D. W. (2000). Visualization of pathogenicity regions in bacteria. *Genetica*, 108:47–51.

Frimurer, T. M., Bywater, R., Nærum, L., Lauritsen, L. N., and Brunak, S. (2000). Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J. Chem. Inf. Comp. Sci.*, 40:1315–1324.

Frishman, D., Mironov, A., Mewes, H.-W., and Gelfand, M. (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.*, 26:2941–2947.

Galagan, J. E., Nusbaum, C., Roy, A., Endrizzi, M. G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., Brown, A., Allen, N., Naylor, J., Stange-Thomann, N., DeArellano, K., Johnson, R., Linton, L., McEwan, P., McKernan, K., Talamas, J., Tirrell, A., Ye, W., Zimmer, A., Barber, R. D., Cann, I., Graham, D. E., Grahame, D. A., Guss, A. M., Hedderich, R., Ingram-Smith, C., Kuettner, H. C., Krzycki, J. A., Leigh, J. A., Li, W., Liu, J., Mukhopadhyay, B., Reeve, J. N., Smith, K., Springer, T. A., Umayam, L. A., White, O., White, R. H., Conway de Macario, E., Ferry, J. G., Jarrell, K. F., Jing, H., Macario, A. J., Paulsen, I., Pritchett, M., Sowers, K. R., Swanson, R. V., Zinder, S. H., Lander, E., Metcalf, W. W., and Birren, B. (2002).

The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res.*, 12:532–542.

Galperin, M. Y. and Koonin, E. V. (2000). Who's your neighbor? new computational approaches for functional genomics. *Nature Biotechnology*, 18:609–613.

Garavelli, J. S., Hou, Z., Pattabiraman, N., and Stephens, R. M. (2001). The RESID database of protein structure modifications and the NRL-3D sequence-structure database. *Nucleic Acids Res.*, 29:199–201.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Hfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147.

Giudicelli, V. and Lefranc, M.-P. (1999). Ontology for immunogenetics: IMGT-ONTOLOGY. *Bioinformatics*, 12:1047–1054.

Goffeau, A. et al. (1997). The yeast genome directory. *Nature*, 387 suppl.:5–105.

Gough, J. and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, 30:268–272.

Greenbaum, D., Luscombe, N. M., Jansen, R., Qian, J., and Gerstein, M. (2001). Interrelating different types of genomic data from proteome to secretome: 'oming in on function. *Genome Res.*, 11:1463–1468.

Gupta, R., Birch, H., Rapacki, K., Brunak, S., and Hansen, J. E. (1999a). O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res.*, 27:370–372.

Gupta, R., Jensen, L. J., and Brunak, S. (2002). Orphan protein function and its relation to glycosylation. In Mewes, H.-M., Seidel, H., and Weiss, B., editors, *Ernst Schering Research Foundation Proceedings*, volume 38, pages 275–294. Springer-Verlag, Berlin.

Gupta, R., Jung, E., Gooley, A. A., Williams, K. L., Brunak, S., and Hansen, J. (1999b). Scanning the available *Dictyostelium discoideum* proteome for O-linked GlcNAc glycosylation sites using neural networks. *Glycobiology*, 9:1009–1022.

Guruprasad, K., Reddy, B. V. B., and Pandit, M. W. (1990). Correlation between stability of a protein and its di-peptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.*, 4:155–161.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape, H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol*, 23:1089–97.

Hacker, J. and Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.*, 54:641–679.

Hanover, J. A. (2001). Glycan-dependent signaling: O-linked N-acetylglucosamine. *FASEB J.*, 15:1865–1876.

Hansen, J. E., Lund, O., Engelbrecht, J., Bohr, H., and Nielsen, J. O. (1995). Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase. *Biochem. J.*, 308:801–813.

Hansen, J. E., Lund, O., Tolstrup, N., Gooley, A. A., Williams, K. L., and Brunak, S. (1998). NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj. J.*, 15:115–130.

Hart, G. W. (1997). Dynamic O-linked glycosylation of nuclear and cytoskeletal proteins. *Ann. Rev. Biochem.*, 66.

Hart, G. W., Greis, K. D., Dong, L. Y., Blomberg, M. A., Chou, T. Y., Jiang, M. S., Roquemore, E. P., Snow, D. M., Kreppel, L. K., and Cole, R. N. (1995). O-linked N-acetylglucosamine: the "yin-yang" of Ser/Thr phosphorylation? nuclear and cytoplasmic glycosylation. *Adv. Exp. Med. Biol.*, 376.

Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H.-S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.-K., Soeda, E., Ohki, M., Takagi, T., Sakaki, Y., Taudien, S., Blechschmidt, K., Polley, A., Menzel, U., Delabar, J., Kumpf, K., Lehmann, R., Patterson, D., Reichwald, K., Rump, A., Schillhabel, M., Schudy, A., Zimmermann, W., Rosenthal, A., Kudoh, J., Shibuya, K., Kawasaki, K., Asakawa, S., Shintani, A., Sasaki, T., Nagamine, K., Mitsuyama, S., Antonarakis, S. E., Minoshima, S., Shimizu, N., Nordsiek, G., Hornischer, K., Brandt, P., Scharfe, M., Schön, O., Desario, A., Reichelt, J., Kauer, G., Blöcker, H., Ramser, J., Beck, A., Klages, S., Hennig, S., Riesselmann, L., Dagand, E., Haaf, T., Wehrmeyer, S., Borzym, K., Gardiner, K., Nizetic, D., Francis, F., Lehrach, H., Reinhardt, R., and Yaspo, M.-L. (2000). The DNA sequence of human chromosome 21. *Nature*, 405:311–319.

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L., Gill, S. R., Nelson, K. E., Read, T. D., Tettelin, H., Richardson, D., Ermolaeva, M. D., Vamathevan, J., Bass, S., Qin, H., Dragoi, I., Sellers, P., McDonald, L., Utterback, T., Fleishmann, R. D., Nierman, W. C., and White, O. (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, 406:477–483.

Helgason, E., Økstad, O. A., Caugant, D. A., Johansen, H. A., Fouet, A., Mock, M., Hegna, I., and Kolstø, A. (2000). *Bacillus anthracis*, *Bacillus cereus*, and

Bacillus thuringiensis—one species on the basis of genetic evidence. *Appl. Environ. Microbiol.*, 66:2627–30.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89:10915–10919.

Heyer, L. J., Kruglyak, S., and Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.*, 9:1106–1115.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C., and Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, 24:4420–4449.

Ho, Y., Gruhler, A., Hellbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennet, K., Boutiller, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sørensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hougue, C. W. V., Figeys, D., and Tyers, M. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415:180–183.

Hobohm, U., Scharf, M., Schneider, R., and Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.*, 1:409–17.

Hodges, P. E., Payne, W. E., and Garrels, J. I. (1998). Yeast protein database (YPD): A database for the complete proteome of *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, 26:68–72.

Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G., and Cooke, M. P. (2001). A comparison of the Celera and Ensembl gene sets reveals little overlap in novel genes. *Cell*, 106:412–415.

Hoogland, C., Sanches, J. C., Tonella, L., Binz, P. A., Bairoch, A., Hochstrasser, D. F., and Appel, R. D. (2000). The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.*, 28:286–288.

Hounsell, E. F., Davies, M. J., and Renouf, D. V. (1996). O-linked protein glycosylation structure and function. *Glycoconjugate J.*, 13:19–26.

Hua, S. and Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17:721–728.

Huala, E., Dickerman, A., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, J., Huang, W., Mueller, L., Bhattacharyya, D., Bhaya, D., Sobral, B., Beavis, B., Somerville, C., and Rhee, S. Y. (2001). The Arabidopsis Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, 29:102–105.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. (2002). The Ensembl genome database. *Nucleic Acids Res.*, 30:38–41.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, K. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000a). Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126.

Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., Burchard, J., Dow, S., Ward, T. R., Kidd, M. J., Friend, S. H., and Marton, M. J. (2000b). Widespread aneuploidy revealed by dna microarray expression profiling. *Nature Genetics*, 25:333–337.

Huss, M., Boström, H., Asker, L., and Cöster, J. (2001). Learning to recognize brain specific proteins based on low-level features from on-line prediction servers. In *Proceedings for the BIOKDD01 Workshop on Datamining in Bioinformatics*, volume 1, pages 45–49.

Hutchison III, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., and Venter, J. C. (1999). Global transposon mutagenesis and a minimal mycoplasma genome. *Science*, 286:2165–2169.

Huynen, M., Dandekar, T., and Bork, P. (1998). Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett.*, 426:1–5.

Iliopoulos, I., Tsoka, S., Andrade, M. A., Janssen, P., Audit, B., Tramontano, A., Valencia, A., Leroy, C., Sander, C., and Ouzounis, C. A. (2000). Genome sequences and great expectations. *Genome Biology*, 2:interactions0001.1–0001.3.

Initiative, T. A. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. the arabidopsis genome initiative. *Nature*, 408:796–815.

Int. Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.*, 98:4569–4574.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356.

Jensen, L. J., Friis, C., and Ussery, D. W. (1999). Three views of microbial genomes. *Res. Microbiol.*, 150:773–777.

Jensen, L. J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H. H., Rapacki, K., Workman, C., Andersen, C. A. F., Knudsen, S., Krogh, A., Valencia, A., and Brunak, S. (2002). Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, 319:1257–1265.

Jensen, L. J. and Knudsen, S. (2000). Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, 16:326–333.

Jensen, R. A. (2001). Orthologs and paralogs—we need to get it right. *Genome Biology*, 2:interactions1002.1–1002.3.

Johnson, M. (2000). The yeast genome: on the road to the golden age. *Curr. Opin. Genet. Dev.*, 10:617–623.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202.

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R. W., Olinger, L., Grimwood, J., Davis, R. W., and Stephens, R. S. (1999). Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genetics*, 21:385–389.

Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takazawa, M., Yamada, M., Yasuda, M., and Tabata, S. (2001). Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res.*, 8:205–213.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., and Tabata, S. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.*, 3:109–136.

Kapatral, V., Anderson, I., Ivanova, N., Reznik, G., Los, T., Lykidis, A., Bhattacharyya, A., Bartman, A., Gardner, W., Grechkin, G., Zhu, L., Vasieva, O., Chu, L., Kogan, Y., Chaga, O., Goltsman, E., Bernal, A., Larsen, N., D'Souza, M., Walunas, T., Pusch, G., Haselkorn, R., Fonstein, M., Kyrpides, N., and Overbeek, R. (2002). Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586. *J. Bacteriol.*, 184:2005–2018.

Karlin, S. (1995). Statistical significance of sequence patterns in proteins. *Curr. Opin. Struct. Biol.*, 5:360–371.

Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U.S.A.*, 87:2264–2268.

Karlin, S., Mrazek, J., and Campbell, A. M. (1998). Codon usages in different gene classes of the *Escherichia coli* genome. *Molecular Microbiology*, 29:1341–1355.

Karp, P. D. (2000). An ontology for biological function based on molecular interactions. *Bioinformatics*, 16:269–285.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Otsuka, R., Nakazawa, H., Takamiya, M., Kato, Y., Yoshizawa, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Masuda, S., Yanagii, M., Nishimura, M., Yamagishi, A., Oshima, T., and Kikuchi, H. (2001). Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. *DNA Res.*, 8:123–140.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Nakazawa, H., Takamiya, M., Masuda, S., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., and Kikuchi, H. . (1999). Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.*, 30:83–101.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., and H., K. (1998). Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.*, 5 suppl.:147–155.

Kawashima, T., Amano, N., Koike, H., Makno, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., Nunoshiba, T., Yamamoto, Y., Aramaki, H., Makino, K., and Suzuki, M. (2000). Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. *Proc. Natl. Acad. Sci. U.S.A.*, 97:14257–14262.

Keefe, A. D. and Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, 410:715–718.

Kihara, A., Akiyama, Y., and Ito, K. (1997). Host regulation of lysogenic decision in bacteriophage lambda: transmembrane modulation of FtsH (HflB), the cii degrading protease, by HflKC (HflA). *Proc. Natl. Acad. Sci. U.S.A.*, 94:5544–5549.

King, R. D., Karwath, A., Clare, A., and Deshaspe, L. (2001). The utility of different representations of protein sequence for predicting functional class. *Bioinformatics*, 17:445–454.

Klenk, H. P., Clayton, R. A., Tomb, J. F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., Richardson, D. L., Kerlavage, Graham, D., Kyrpides, N., Fleischmann, R., Quackenbush, J., Lee, N., Sutton, G., Gill, S., Kirkness, E., Dougherty, B., McKenney, K., M.D. Adams, M. D., Loftus, B., Peterson, S., McNeil, L. K., Badger, J. H., Zhou, L. X., Overbeek, R., Gocayne, J. D., Weidman, J. F., McDonald, L., Utterback, T., Cotton, M. D., Spriggs, T., Artiach, P., Kaine, B. P., Sykes, S. M., Sadow, P. W., Andrea, K. P., Bowman, C., Fujii, C., Garland, S. A., Mason, T. M., Olsen, G. J., Fraser, C. M., Smith, H. O., Woese, C. R., and Venter, J. C. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*, 390:364–370.

Koonin, E. V., Mushegian, A. R., Galperin, M. Y., and Walker, D. R. (1997). Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.*, 25:619–637.

Krieg, J., Hartmann, S., Vicentini, A., Glasner, W., Hess, D., and Hofsteenge, J. (1998). Recognition signal for C-mannosylation of Trp-7 in RNase 2 consists of sequence Trp-X-X-Trp. *Mol. Biol. Cell*, 9:301–309.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.*, 305:567–580.

Kukuruzinska, M. A. and Lennon, K. (1998). Protein N-glycosylation: molecular genetics and functional significance. *Crit. Rev. Oral Biol. Med.*, 9:415–48.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S. K., Codani, J. J., Connerton, I. F., and Danchin, A. *et al.*. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, 390:249–256.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157:105–132.

Larkin, J. C., Thompson, J. R., and Woolford, Jr., J. L. (1987). Structure and expression of the *Saccharomyces cerevisiae* CRY1 gene: a highly conserved ribosomal protein gene. *Mol. Cell. Biol.*, 7:1764–1775.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214.

Lecompte, O., Ripp, R., Puzos-Barbe, V., Duprat, S., Heilig, R., Dietrich, J., Thierry, J. C., and Poch, O. (2001). Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. *Genome Res.*, 11:981–993.

Lesk, A. M., , Conte, L. L., , and Hubbard, T. J. P. (2001). Assessment of novel fold targets in CASP4: Predictions of three-dimensional structures, secondary structures and interresidue contacts. *Proteins*, 45 suppl. 5:98–118.

Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P., and Bork, P. (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, 30:242–244.

Lew, D. J., Weinert, T., and Pringle, J. R. (1997). Cell cycle control in *Saccharomyces cerevisiae*. In Pringle, J., Broach, J., and Jones, E., editors, *The Molecular Biology of the Yeast Saccharomyces—3: Cell cycle and Cell Biology*, pages 697–. Cold Spring Harbor Laboratory Press, New York.

Lewis, S., Ashburner, M., and Reese, M. G. (2000). Annotating eukaryote genomes. *Curr. Opin. Struct. Biol.*, 10:349–354.

Li, W., Pio, F., Pawlowski, K., and Godzik, A. (2000). Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology. *Bioinformatics*, 16:1105–1110.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genetics*, 25:239–240.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics*, 21:20–24.

Lis, H. and Sharon, N. (1993). Protein glycosylation: Structural and functional aspects. *Cur. J. Biochem.*, 218:1–27.

Liu, J. and Rost, B. (2000). SAWTED: Structure Assignment With Text Description—enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics*, 16:125–129.

Liu, J. and Rost, B. (2001). Comparing function and structure between entire proteomes. *Protein Sci.*, 10:1970–1979.

Luca, F. C., Mody, M., Kurischko, C., Roof, D. M., Giddings, T. H., and Winey, M. (2001). *Saccharomyces cerevisiae* mob1p is required for cytokinesis and mitotic exit. *Mol. Cell. Biol.*, 21:6972–6983.

Mahillon, J., Rezsohazy, R., Hallet, B., and Delcour, J. (1994). 231 and other *Bacillus thuringiensis* transposable elements: a review. *Genetica*, 93:13–26.

Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I., and Koonin, E. V. (1999). Comparative genomics of archaea (euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.*, 9:608–628.

Marcotte, E. M. (2000). Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, 10:359–365.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999a). Detecting protein function and protein–protein interactions from genome sequences. *Science*, 285:751–753.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–86.

Marcotte, E. M., Xenarios, I., van Der Bliek, A. M., and Eisenberg, D. (2000). Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 97:12115–12120.

Mathews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451.

McDonagh, P. D., Myler, P. J., and Stuart, K. (2000). The unusual gene organization of *Leishmania major* friedlin chromosome 1 may reflect novel transcription processes. *Nucleic Acids Res.*, 28:2800–2803.

McLean, M. J., Wolfe, K. H., and Devine, K. M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.*, 47:691–696.

Mellor, J. C., Yanai, I., Clodfelter, K. H., Mintseis, J., and DeLisi, C. (2002). Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.*, 30:306–309.

Mendenhall, M. D. and Hodge, A. E. (1998). Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, 62:1191–1243.

Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkoetter, M., Rudd, S., and Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, 30:31–34.

Mott, R. (2001). Maximum likelihood estimation of the statistical distribution of smith–waterman local sequence similarity scores. *Bull. Math. Biol.*, 54:59–75.

Mrowka, R., Patzak, A., and H, H. (2001). Is there a bias in proteome research. *Genome Res.*, 11:1971–1973.

Myler, P. J., Audleman, L., de Vos, T., Hixson, G., Kiser, P., Magness, C., Rickel, E., Sisk, E., Sunkin, S., Swartzell, S., Westlake, T., Bastein, P., Fu, G., Ivens, A., and Stuart, K. (1999). *Leishmania major* friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc. Natl. Acad. Sci. U.S.A.*, 96:2902–2906.

Nakai, K. (2001). Prediction of *in Vivo* fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.*, 134:103–116.

Nakai, K. and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, 24:34–36.

Natale, D. A., Shankavaram, U. T., Galperin, M. Y., Wolf, Y. I., Aravind, L., and Koonin, E. V. (2000). Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biology*, 1:research0009.1–0009.19.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickeyand, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., and Fraser, C. M. (1999). Evidence of lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, 399:323–329.

Ng, P. C., Henikoff, J. G., and Henikoff, S. (2000). PHAT: a transmembrane–specific substitution matrix. *Bioinformatics*, 16:760–766.

Nielsen, H., Brunak, S., Engelbrecht, J., and von Heijne, G. (1997a). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, 10:1–6.

Nielsen, H., Brunak, S., and von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, 12:3–9.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997b). A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, 8:581–599.

Nielsen, H. and Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden markov model. In *Proc., Intelligent Systems for Molecular Biology*, volume 6, pages 122–130, Menlo Park, CA. AAAI Press.

Nilsson, I. and von Heijne, G. (2000). Glycosylation efficiency of Asn-Xaa-Thr sequons depends both on the distance from the C terminus and on the presence of a downstream transmembrane segment. *J. Biol. Chem.*, 275:17338–17343.

Nilsson, I. M. and von Heijne, G. (1993). Determination of the distance between the oligosaccharyltransferase active site and the endoplasmic reticulum membrane. *J. Biol. Chem.*, 268:5798–5801.

Noble, J. A., Innis, M. A., Koonin, E. V., Rudd, K. E., Banuet, F., and Herskowitz, I. (1993). The *Escherichia coli* hflA locus encodes a putative GTP-binding protein and two membrane proteins, one of which contains a protease-like domain. *Proc. Natl. Acad. Sci. U.S.A.*, 90:10866–10870.

Norin, M. and Sundström, M. (2002). Structural proteomics: development in structure-to-function predictions. *Trends in Biotechnology*, 20:79–84.

Nurse, P. (2000). A long twentieth century of the cell cycle and beyond. *Cell*, 100:71–78.

Ochman, H. (2002). Distinguishing the orfs from the elfs: short bacterial genes and the annotation of genomes. *Trends in Genetics*, 18:335–337.

O'Connell, K. F. and Baker, R. E. (1992). Possible cross-regulation of phosphate and sulfate metabolism in *Saccharomyces cerevisiae*. *Genetics*, 132:63–73.

Ornstein, R. L., Rein, R., Breen, D. L., and MacElroy, R. D. (1978). An optimized potential function for the calculation of nucleic acid interaction energies. I. Base stacking. *Biopolymers*, 17:2341–2360.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.*, 96:2896–2901.

Ozoline, O. N., Deev, A. A., Arkhipova, M. V., Chasov, V. V., and Travers, A. (1999). Proximal transcribed regions of bacterial promoters have a non-random distribution of A/T tracts. *Nucleic Acids Res.*, 27:4768–4774.

Page, A. L., Fromont-Racine, M., Sansonetti, P., Legrain, P., and Parsot, C. (2001). Characterization of the interaction partners of secreted proteins and chaperones of *Shigella flexneri*. *Mol. Microbiol.*, 42:1133–1145.

Pandey, A. and Mann, M. (2000). Proteomics to study genes and genomes. *Nature*, 405:837–846.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, 284:1201–1210.

Parker, D. B. (1982). Learning-logic. Technical Report 581-64, Office of Technology Licensing, Stanford University.

Parkhill, J., Achtman, M., James, K. D., Bentley, S. D., Churcher, C., Klee, S. R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., Davies, R. M., Davis, P., Devlin, K., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Leather, S., Moule, S., Mungall, K., Quail, M. A., Rajandream, M. A., Rutherford, K. M., Simmonds, M., Skelton, J., Whitehead, S., Spratt, B. G., and Barrell, B. G. (2000a). Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature*, 404:502–506.

Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcer, C., Basham, D., Chillingworth, T., Davies, R. M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A. V., Moule, S., Pallen, M. J., Penn, C. W., Quall, M. A., Rajandream, M.-A., Rutherford, K. M., van Vllet, A. H. M., Whitehead, S., and Barrell, B. G. (2000b). The genome sequence of the food-born pathogen *Campylobacter jejuni* reveals hypervariale sequences. *Nature*, 403:665–668.

Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B., Sebaihia, M., James, K. D., Churcher, C., Mungall, K. L., Baker, S., Basham, D., Bentley, S. D., Brooks, K., Cerdeno-Tarraga, A. M., Chillingworth, T., Cronin, A., Davies, R. M., Davis, P., Dougan, G., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Karlyshev, A. V., Leather, S., Moule, S., Oyston, P. C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S., and Barrell, B. G. (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, 413:523–527.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.*, 14:609–614.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl. Acad. U.S.A.*, 85:2444–2448.

Pedersen, A. G., Baldi, P., Chauvin, Y., and Brunak, S. (1998). DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, 281:663–673.

Pedersen, A. G., Jensen, L. J., Stærfeldt, H. H., Brunak, S., and Ussery, D. W. (2000). A DNA structural atlas of *E. coli*. *J. Mol. Biol.*, 299:907–930.

Pellegrini, M. (2001). Computational methods for protein function analysis. *Curr. Opin. Chem. Biol.*, 5:46–50.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 38:667–677.

Pesole, G., Prunella, N., Liuni, S., Attimonelli, M., and Saccone, C. (1992). WORDUP: An efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucleic Acids Res.*, 20:2871–2875.

Pillardy, J., Czaplewski, C., Liwo, A., Lee, J., Ripoli, D. R., Kaźmierkiewicz, R., Ołdziejl, S., Wedemeyer, W. J., Gibson, K. D., Arnautova, Y. A., Saunders, J., Ye, Y.-J., and Scheraga, H. A. (2000). Recent improvements in prediction of

protein structure by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. U.S.A.*, 98:2329–2333.

Pole, S., Sigismund, S., Faretta, M., Guldi, M., Capua, M. R., Bossi, G., Chen, H., De Camilli, P., and Di Fiore, P. (2002). A single motif responsible for ubiquitin recognition and monoubiquitination in endocytic proteins. *Nature*, 416:451–455.

Rawlings, N. D., O'Brien, E., and Barrett, A. J. (2002). MEROPS: the protease database. *Nucleic Acids Res.*, 30:343–346.

Rechsteiner, M. and Rogers, S. W. (1996). PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.*, 21:267–271.

Remm, M., Storm, C. E. V., and Sonnhammer, E. L. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparison. *J. Mol. Biol.*, 314:1041–1052.

Riley, M. (1993). Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.*, 57:862–952.

Roth, J., Wang, Y., Eckhardt, A. E., and Hill, R. L. (1994). Subcellular localization of the UDP-N-acetyl-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase-mediated O-glycosylation reaction in the submaxillary gland. *Proc. Natl. Acad. Sci. U.S.A.*, 91:8935–8939.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. H., Zhong, F., Zhong, W., Gibbs, R., Venter, J. C., Adams, M. D., and Lewis, S. (2000). Comparative genomics of the eukaryotes. *Science*, 287:2204–2215.

Ruepp, A., Graml, W., Santos-Martinez, M. L., Koretke, K. K., Volker, C., Mewes, H. W., Frishman, D., Stocker, S., Lupas, A. N., and Baumeister, W. (2000). The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature*, 407:508–513.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Satchwell, S. C., Drew, H. R., and Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, 191:659–675.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470.

Schnepf, E., Crickmore, N., van Rie, J., Lereclus, D., Baum, J., Feitelson, J., Zeigler, D. R., and Dean, D. H. (1998). *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiol. Mol. Biol. Rev.*, 62:775–806.

Schuch, R., Sandlin, R. C., and Maurelli, A. T. (1999). A system for identifying post-invasion functions of invasion genes: requirements for the *Mxi-Spa* type III secretion pathway of *Shigella flexneri* in intercellular dissemination. *Mol. Micro.*, 34:675–689.

Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P., and Bork, P. (2000). SMART: A web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, 28:231–234.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnology*, 18:1257–1261.

Sharp, P. M. and Li, W. H. (1987). The codon adaptation index — a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, 15:1281–1295.

She, Q., Singh, R. K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M., Chan-Weiher, C.-Y., Clausen, I. G., Curtis, B. A., de Moors, A., Erauso, G., Fletcher, C., Gordon, P. M. K., Jong, I. H., Jeffries, A. C., Kozera, C. J., Medina, N., Peng, X., Thi-Ngoc, H. P., Redder, P., Schenk, M., Theriault, C., Tolstrup, N., Charlebois, R. L., Doolittle, W. F., Duguet, M., Gaasterland, T., Garrett, R. A., and Ragan, M. A. (2001). The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl. Acad. Sci. U.S.A.*, 98:7835–7840.

Shedden, K. and Cooper, S. (2002). Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. *Nucleic Acids Res.*, 30:2920–2929.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, 407:81–86.

Shigenobu, S., Watanabe, H., Sakaki, Y., and Ishikawa, H. (2001). Accumulation of species-specific amino acid replacements that cause loss of particular protein functions in *Buchnera*, an endocellular bacterial symbiont. *J. Mol. Evol.*, 53:377–386.

Shpigelman, E. S., Trifonov, E. N., and Bolshoy, A. (1993). CURVATURE: Software for the analysis of curved DNA. *CABIOS*, 9:435–444.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London. Chap. 3.

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Ranaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708.

Simpson, A. J., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M., Alvarenga, R., Alves, L. M., Araya, J. E., Baia, G. S., Baptista, C. S., Barros, M. H., Bonaccorsi, E. D., Bordin, S., Bove, J. M., Briones, M. R., Bueno, M. R., Camargo, A. A., Camargo, L. E., Carraro, D. M., Carrer, H., Colauto, N. B., Colombo, C., Costa, F. F., Costa, M. C., Costa-Neto, C. M., Coutinho, L. L., Cristofani, M., Dias-Neto, E., Docena, C., El-Dorry, H., Facincani, A. P., Ferreira, A. J., Ferreira, V. C., Ferro, J. A., Fraga, J. S., Franca, S. C., Franco, M. C., Frohme, M., Furlan, L. R., Garnier, M., Goldman, G. H., Goldman, M. H., Gomes, S. L., Gruber, A., Ho, P. L., Hoheisel, J. D., Junqueira, M. L., Kemper, E. L., Kitajima, J. P., Krieger, J. E., Kuramae, E. E., Laigret, F., Lambais, M. R., Leite, L. C., Lemos, E. G., Lemos, M. V., Lopes, S. A., Lopes, C. R., Machado, J. A., Machado, M. A., Madeira, A. M., Madeira, H. M., Marino, C. L., Marques, M. V., Martins, E. A., Martins, E. M., Matsukuma, A. Y., Menck, C. F., Miracca, E. C., Miyaki, C. Y., Monteriro-Vitorello, C. B., Moon, D. H., Nagai, M. A., Nascimento, A. L., Netto, L. E., Nhani, Jr., A., Nobrega, F. G., Nunes, L. R., Oliveira, M. A., de Oliveira, M. C., de Oliveira, R. C., Palmieri, D. A., Paris, A., Peixoto, B. R., Pereira, G. A., Pereira, Jr., H. A., Pesquero, J. B., Quaggio, R. B., Roberto, P. G., Rodrigues, V., de M Rosa, A. J., de Rosa, Jr., V. E., de Sa, R. G., Santelli, R. V., Sawasaki, H. E., da Silva, A. C., da Silva, A. M., da Silva, F. R., da Silva, Jr, W. A., da Silveira, J. F., Silvestri, M. L., Siqueira, W. J., de Souza, A. A., de Souza, A. P., Terenzi, M. F., Truffi, D., Tsai, S. M., Tsuhako, M. H., Vallada, H., Van Sluys, M. A., Verjovski-Almeida, S., Vettore, A. L., Zago, M. A., Zatz, M., Meidanis, J., and Setubal, J. C. (2000). The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, 406:151–157.

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., and Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends in Genetics*, 17:425–428.

Slesarev, A. I., Mezhevaya, K. V., Makarova, K. S., Polushin, N. N., Shcherbinina, O. V., Shakhova, V. V., Belova, G. I., Aravind, L., Natale, D. A., Rogozin, I. B., Tatusov, R. L., Wolf, Y. I., Stetter, K. O., Malykh, A. G., Koonin, E. V., and Kozyavkin, S. A. (2002). The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proc. Natl. Acad. Sci. U.S.A.*, 99:4644–4649.

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani N, . Caruso, A., Bush, D., Safer, H., Patwell, D., Prabhakar, S., McDougall, S., Shimer, G., Goyal, A., Pietrokovski, S., Church, G. M., Daniels, C. J., Mao, J. I., Rice, P., Nolling,

J., and Reeve, J. N. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional analysis and comparative genomics. *J. Bacteriol.*, 179:7135–7155.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197.

Snow, D. M. and Hart, G. W. (1998). Nuclear and cytoplasmic glycosylation. *Int. Rev. Cytol.*, 181.

Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *ISBM*, 6:175–182.

Spang, R. and Vingron, M. (2001). Limits of homology detection by pairwise sequence comparison. *Bioinformatics*, 17:338–342.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *S. cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297.

Spiro, R. G. (2002). Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology*, 12:43R–56R.

Stawiski, E. W., Baucom, A. E., Lohr, S. C., and Gregoret, L. M. (2000). Predicting protein function from structure: Unique structural features of proteases. *Proc. Natl. Acad. Sci. U.S.A.*, 97:3954–3958.

Stephens, R. S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R. L., Zhao, Q., Koonin, E. V., and Davis, R. W. (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science*, 282:754–759.

Sternberg, M. J. E., Bates, P. A., Kelley, L. A., and MacCallum, R. M. (1999). Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.*, 9:368–373.

Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Hirama, C., Nakamura, Y., Ogasawara, N., Kuhara, S., and Horikoshi, K. (2000). Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Res.*, 28:4317–4331.

Tamames, J., Casari, G., Ouzounis, C., and Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, 44:66–73.

Tamames, J., Ouzounis, C., Casari, G., Sander, C., and Valencia, A. (1998). EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, 14:542–543.

Tatusov, R. L. and, K., EV and, L., and DJ (1997). A genomic perspective on protein families. *Science*, 278:631–637.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galpein, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, 29:22–28.

Tavernarakis, N., Driscoll, M., and Kyrpides, N. C. (1999). The SPFH domain: implicated in regulating targeted protein turnover in stomatins and other membrane-associated proteins. *Trends Biochem. Sci.*, 24:425–427.

Tettelin, H., Saunders, N. J., Heidelberg, J., Jeffries, A. C., Nelson, K. E., Eisen, J. A., Ketchum, K. A., Hood, D. W., Peden, J. F., Dodson, R. J., Nelson, W. C., Gwinn, M. L., DeBoy, R., Peterson, J. D., Hickey, E. K., Haft, D. H., Salzberg, S. L., White, O., Fleischmann, R. D., Dougherty, B. A., Mason, T., Ciecko, A., Parksey, D. S., Blair, E., Cittone, H., Clark, E. B., Cotton, M. D., Utterback, T. R., Khouri, H., Qin, H., Vamathevan, J., Gill, J., Scarlato, V., Masignani, V., Pizza, M., Grandi, G., Sun, L., Smith, H. O., Fraser, C. M., Moxon, E. R., Rappuoli, R., and Venter, J. C. (2000). Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, 287:1809–1815.

The FlyBase Consortium (2002). The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.*, 30:106–108.

Tillier, E. R. and Collins, R. A. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.*, 50:249–257.

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, 307:1113–1143.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzgerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M., and Venter, J. C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, 388:539–547.

Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W., Fields, S., Boone, C., and Cesareni, G. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295:321–324.

Trabi, M. and Craik, D. J. (2002). Circular proteins—no end in sight. *Trends Biochem. Sci.*, 27:132–138.

Trifonov, E. N. and Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. U.S.A.*, 77:3816–3820.

Tyers, M. and Jorgensen, P. (2000). Proteolysis and the cell cycle: With this RING I do thee destroy. *Curr. Opin. Gen. Dev.*, 10:54–64.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleish, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627.

Ulanovsky, L., Bodner, M., and Trifonov, E. N. (1986). Curved DNA: Design, synthesis, and circularization. *Proc. Natl. Acad. Sci. U.S.A.*, 83:862–866.

Ussery, D. W., Larsen, T. S., Wilkes, K. T., Friis, C., Worning, P., Krogh, A., and Brunak, S. (2001). Genome organisation and chromatin structure in *Escherichia coli*. *Biochimie*, 83:201–212.

Ussery, D. W., Soumpasis, D. M., Brunak, S., Stærfeldt, H. H., Worning, P., and Krogh, A. (2002). Bias of purine stretches in sequenced genomes. *Computers in Chemistry*, 26:in press.

van den Steen, P., Rudd, P. M., Dwek, R. A., and Opdenakker, G. (1998). Concepts and principles of o-linked glycosylation. *Crit. Rev. Biochem. Mol. Biol.*, 33:151–208.

van Helden, J., Andre, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842.

Varki, A. (1993). Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology*, 3:97–130.

Varshavsky, A. (1996). The N-end rule: functions, mysteries, uses. *Proc. Natl. Acad. Sci. U.S.A.*, 93:12142–12149.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270:484–487.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, Jr., D. E., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell*, 88:243–251.

Venkatesan, M. M., Goldberg, M. B., Rose, D. J., Grotbeck, E. J., Burland, V., and Blattner, F. R. (2001). Complete DNA sequence and analysis of the large virulence plasmid of *Shigella flexneri*. *Infect. Immun.*, 69:3271–3285.

Venter, J. et al. (2001). The sequence of the human genome. *Science*, 291:1304–1351.

Vieille, C. and Zeikus, G. J. (2001). Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.*, 65:1–43.

Vignais, M. L., Huet, J., Buhler, J. M., and Sentenac, A. (1990). Contacts between the factor TUF and RPG sequences. *J. Biol. Chem*, 265:14669–146674.

Vignais, M. L., Woundt, L. P., Wassenaar, G. M., Mager, W. H., Sentenac, A., and Planta, R. J. (1987). Specific binding of TUF factor to upstream activation sites of yeast ribosomal protein genes. *EMBO J.*, 6:1451–1457.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417:399–304.

Walhout, A. J. M. and Vidal, M. (2001). Protein interaction maps for model organisms. *Nature Rev. Mol. Cell Biol.*, 2:56–62.

Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K. D., Terryn, N., Harris, B., Ansroge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Muller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., Schmidtheini, T., Reichert, B., Portatelle, D., Perez-Alonso, M., Bountry, M., Bancroft, I., Vos, P., Hoheisel, J., Zimmermann, W., Wedler, H., Ridley, P., Langham, S. A., McCullagh, B., Bilham, L., Robben, J., van der Schueren, J., Grymonprez, B., Chuang, Y. J., Vandenbussche, F., Braeken, M., Weltjens, I., Voet, M., Bastiens, I., Aert, R., Defoor, E., Weitzenegger, T., Bothe, G., and Rose, M. (2000). Progress in arabidopsis genome sequencing and functional genomics. *J. Biotechnol.*, 31:281–292.

Waterman, M. S. and Vingron, M. (1994). Sequence comparison significance and poisson approximation. *Bioinformatics*, 9:367.

Weir, M., Swindells, M., and Overington, J. (2001). Insights into protein function through large-scale computational analysis of sequence and structure. *Trends in Biotechnology*, 19:S61–S66.

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioural sciences*. PhD thesis, Harvard University.

White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J., Haft, D. H., Gwinn, M. L., Nelson, W. C., Richardson, D. L., Moffat, K. S., Qin, H., Jiang, L., Pamphile, W., Crosby, M., Shen, M., Vamathevan, J. J., Lam, P., McDonald, L., Utterback, T., Zalewski, C., Makarova, K. S., Aravind, L., Daly, M. J., and Fraser, C. *et al.* (1999). Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science*, 286:1571–1577.

Wootton, J. C. (1994a). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, 18:269–285.

Wootton, J. C. (1994b). Sequences with "unusual" amino-acid compositions. *Curr. Opin. Struct. Biol.*, 4:413–421.

Workman, C. T. and Stormo, G. D. (2000). ANN-SPEC: A method for discovering transcription factor binding sites with improved specificity. In *Proceedings for the Pacific Symposium on Biocomputing*, volume 5, pages 681–688.

Xenarios, I., Fernandez, E., Salwínski, Ł., Duan, X. J., Thompson, M. J., Marcotte, E. M., and Eisenberg, D. (2001). DIP: The database of interacting proteins: 2001 update. *Nucleic Acids Res.*, 29:239–241.

Xenarios, I., Salwínski, Ł., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, 30:303–305.

Yanai, I., Mellor, J. C., and DeLisi, C. (2001). Identifying functional links between genes using conserved chromosomal proximity. *Trends in Biotechnology*, 19:S61–S66.

Yaspo, M.-L. (2001). Taking a functional genomics approach in molecular medicine. *Trends in Molecular Medicine*, 7:494–502.

Zhang, C.-T. and Zhang, R. (1999). Skewed distribution of protein secondary structure contents over the confomational triangle. *Prot. Eng.*, 12:807–809.

Zhang, J. (2000). Protein-length distributions for the three domains of life. *Trends in Genetics*, 16:107–109.

Zhao, L. P., Prentice, R., and Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 98:5631–5636.

List of Figures

List of Tables