Computational Analysis of Modular Protein Architectures

Rune Linding, Ivica Letunic, Toby J. Gibson, and Peer Bork
EMBL-Heidelberg, Heidelberg, Germany

Originally published in: Modular Protein Domains. Edited by Giovanni Cesareni, Mario Gimona, Marius Sudol and Michael Yaffe. Print ISBN: 3-527-30813-2

The sections in this article are

1
Introduction

In this chapter we discuss some of the bioinformatics and computational tools that exist for finding modular domains and their cognate ligands. We begin with a general discussion of protein architecture and relate this to the modular model of protein function. We pay particular attention to the SMART, ELM, GlobPlot, and DisEMBL resources, because these are the ones we are most familiar with. We apologize to those involved with all the great resources out there that we have not covered in this chapter.

2
Protein Architecture: Sequence, Structure, and Function

Nature seems to present protein functions in two states: structured and unstructured. Proteins are heteropolymers of amino acids; the sequence of amino acids determines not only the structure and folding of a protein but also the lack of structure. Molecular functions in proteins are associated with structural units, e.g., modular globular domains. However, an emerging large group of functional sites are found primarily in unstructured parts of proteins.

In this chapter we explore how these sites can be identified in proteins and describe some of the computational tools that can be used to analyze protein sequences.

Protein Science Encyclopedia - Online

DOI: 10.1002/9783527610754

2.1 The Modular Model of Protein Function

Multidomain proteins predominate in eukaryotic proteomes. The basic hypothesis in what one might call the modular model for protein function is that individual functions assigned to different sequence segments (often domains) combine to create a complex function for the whole protein.

The term 'modular' refers directly to the autonomous nature of the individual folding units determining these functions, whereas the term 'globular' describes the structural state of a domain whether or not it is modular. A dogma in structural biology is that atomic structure determines function. The modular model has grown out of this paradigm; however, we know that at the fold level this is not always true, one can find structures belonging to the same fold (e.g. in SCOP) that have completely different functions: an example of this is glutathione-S-transferase and S-crystallin. They share 75% sequence similarity and the same fold, but the former is an enzyme and the latter a structural protein [4]. This is mainly a problem when one is trying to infer function from sequence; having the atomic structure solved frequently helps to define the function.

Because single-domain globular proteins are often, although not always, less difficult to crystallize, for a long time they dominated our perceptions of typical protein structure (although fibrous proteins like collagen were of course well known). Gradually, as protein sequences have accumulated, the monodomain view of protein structure has been replaced by the realization that most proteins are multidomain, at least in higher eukaryotes. Multidomain architectures are usual for transmembrane receptors, signaling proteins, cytoskeletal proteins, chromatin proteins, transcription factors, and so forth. Multidomain proteins can be described as consisting of a series of modules or globular domains and a set of short linear functional sites. An archetypal protein of the modular model is Src (Figure 1). Src consists of three globular domains: SH3, SH2, and a tyrosine kinase domain (itself consisting of two structural domains), and eight known functional sites including four phosphorylation sites and three ligands of modular domains (SH3, SH2, and cyclin).

2.2 Partitioning of Protein Space

It is becoming increasingly clear that many functionally important protein segments occur outside globular domains [27, 99]. The set (or space) of all observations of protein structure and function is partitioned into two subspaces (Figure 2). The first consists of globular units having binding pockets, active sites, and interaction surfaces. The second subspace consists of nonglobular segments such as sorting signals, post-translational modification sites, and protein ligands (e.g., SH3 or WW ligands). Globular units are built of regular secondary structural elements and contribute the majority of the structural data deposited in PDB. The globular function space is described very well by domain databases such as SMART [49] and Pfam [11].

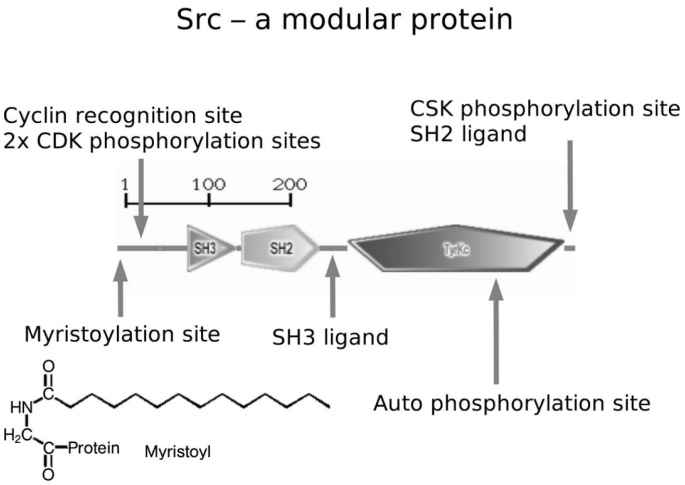

Fig. 1 Domain and functional site architecture of the well known proto-oncogenic protein kinase Src. About 60%–80% of proteins from higher eukaryotes have analogous modular architectures [6]. Even though 'only' ~30 000 ORFs are predicted to be in the human genome, most of these have several splice variants and, in addition, several functional sites, e.g., post-translational modification sites (PTMs). These sites exist in various states and thereby increase the number of different functional isoforms several fold. The arrows show only the approximate locations of the functional sites.

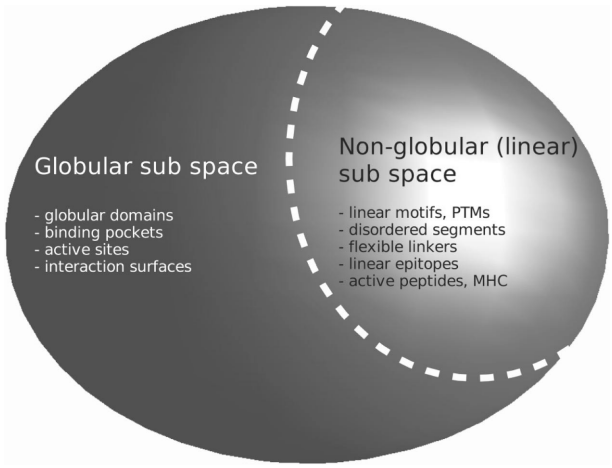

Fig. 2 A conceptual model of protein structure and function observation space. Many functions and their structures can be assigned to two different subspaces. A linear/nonglobular one and a globular one. It is important to notice that these two subspaces are not detached from each other – a good example is disordered loops that can protrude from a globular domain. Along the borderline (white) we find coiled coils, repeat proteins (often forming rods), and single transmembrane helices (TM1).

In contrast, the nonglobular subspace encompasses disordered, unstructured, and flexible regions lacking regular secondary structure. Functional sites within the nonglobular space are known as linear motifs catalogued by ELM [71], PROSITE [80], and Scansite [101]. This group of sites includes protein interaction sites, cell compartment targeting signals, post-translational modification sites, and cleavage sites.

Traditionally, protein function and interaction have been studied from a domain-centric view, and in fact, most large datasets that deal with protein interaction have also focused on this type of interaction [2]. This is because methods such as affinity purification tend to isolate 'sticky' and 'nontransient' protein complexes. The methods for isolating transient interactions, such as the binding of a cognate ligand to its modular domain, are different, and much smaller in-vivo datasets exist.

However, the nontransient network is only a part, and perhaps even a subpart, of the interaction networks within the cell. We hope that this book will encourage the scientific community to focus on collecting large amounts of data on domain-ligand interactions, since only by having these can we try to obtain a full picture of the cellular protein networks.

Below, we discuss how to annotate and analyze protein sequences for modular domains and linear motifs. Methods for finding domains and predicting their function are described first, followed by a description of how to identify unstructured regions and their potential functions.

3
Analyzing Globular Domains

The past three decades have seen relatively steady levels of domain discovery [18]. It seems likely that most of the more common mobile protein domains have already been described in the literature (see, Figure 3). However, because there is a large number of domains that are detectable only in a relatively small number of proteins, we predict that various domains are still hiding in many proteins and that each genome might harbor its own repertoire of species-specific or at least lineage-specific domains [26].

Since the late 1970s, homology search has been an extremely powerful computational technique for assigning novel functions to proteins. In addition to database search methods that were introduced in the early 1980s, an awareness of conserved entities such as local motifs was incorporated into software that was able to scan dedicated collections of such motifs in the mid 1980s (see, e.g., PROSITE, 1985, for an early resource). Yet, the statistics of more sophisticated search methods such as BLAST are still struggling with compositional biases due to nonglobularity or multiple occurrences of such structural entities within a search sequence. However, without discrimination of the rationale behind functionality (e.g., short functional motifs that can change quickly in time, such as glycosylation patterns vs. essential catalytic residues that stay conserved over billions of years), homology searches with the aim of function prediction remain limited. Thus, in this chapter

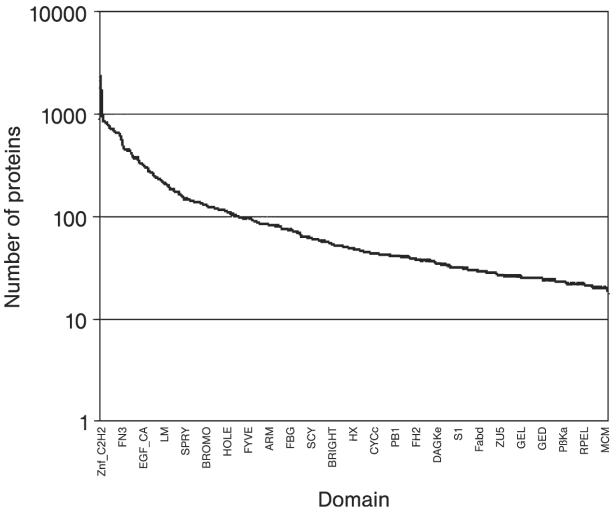

Fig. 3 Presence of SMART domains in the *Homo sapiens* proteome. Domains present in more than 20 proteins are shown. Domain names are displayed for only a few of the 314 domains.

we first define globular domains and then describe resources that help to annotate known domains. Finally, we explain the analysis options of one of these resources, SMART.

3.1 Globularity of Domains

Both 'domain' and 'globular' are terms defined in structural protein biology. They have since been used in other contexts such as function description and sequence analysis. Domains were first defined in structural terms, since early X-ray structures showed separate entities with defined subfunctions connected by flexible regions. The term globular refers to the globular structural state: a protein globule can be depicted as a soluble sphere having a hydrophobic core. An operational definition of globular proteins can be found in [19]:

> Most natural proteins in solution are much smaller in their dimensions than comparable polypeptides with random or repetitive conformations and have roughly spherical shapes; hence they are generally referred to as globular. Their physical properties do not change gradually as the environment is altered (e.g., by changes in temperature, pH or pressure) as do the properties of random polypeptides. Instead, globular proteins usually exhibit little or no change, until a point is reached at which there is a sudden drastic change and, invariably, a loss of biological function. This phenomenon is known as denaturation.

Structural biologists relate the term globular to domains that are compact and fold independently of the remainder of the host protein. Theoretical definitions of

domains exist [85, 86, 88], including several algorithms for classifying domains and folds [43, 87].

Resources for finding globular domains with determined tertiary structure include SCOP (fold level) [54], ASTRAL (domains from SCOP) [16], SUPERFAMILY (hidden Markov models of SCOP domains) [35], and CATH (architecture level) [37, 65]. However, these resources are confined to the structural knowledge base PDB [96]. Therefore, only a subset of the total fold and structure space is described; the remaining domains are to a certain extent described in Pfam and SMART.

3.2 Resources for Analysis of Globular Domains

There are numerous domain databases available that can be useful for detection and analysis of globular domains in your favorite protein sequence. They can be separated into several categories:

- Databases such as SMART (http://smart.embl.de/ [49]) and PFAM (http://www.sanger.ac.uk/software/pfam/ [11]) primarily make use of hand-edited sequence alignments representing single protein domains with well defined borders at the sequence level. PROSITE (http://www.expasy.org/prosite/ [80]) is also a handmade resource, but it contains a much more heterogeneous set of domains and motifs although, for linear motifs, it has been superseded by ELM.
- Other databases rely on various automatic methods to generate their domain signatures. This is so for ProDom (http://www.toulouse.inra.fr/prodom.html [79]) and BLOCKS (http://www.blocks.fhcrc.org/ [38]). Such resources predict domains that do not always correspond to known structural and globular domains and for this purpose may not be as sensitive. However, these resources are of substantial discovery value, since they collect conserved sequence segments that might specify novel functions.
- In addition to the search functionality of the databases themselves, several metaservers allow users to search multiple domain databases. The Interpro database, at the EBI (http://www.ebi.ac.uk/interpro/ [61]) allows searching of the PROSITE, PFAM, PRINTS, ProDom, and SMART model collections, and the Conserved Domain Database (CDD) at the NCBI (http://www.ncbi.nlm.nih.gov/structure/cdd/cdd.shtml [57]) allows searching of profiles derived from SMART and PFAM using a modified version of the BLAST algorithm.

We describe the SMART resource in greater detail, because it focuses on modular signaling domains.

3.3 SMART: Simple Modular Architecture Research Tool

The explosion of sequence data increases the need for computational sequence-analysis tools that annotate novel genes with predicted functions. Function prediction, however, is fraught with potential pitfalls, such as variable sequence

divergence, nonequivalent functions of homologs, and nonidentical multidomain architectures [25]. Detecting nonenzymatic regulatory domains is essential to predicting a protein's cellular role, binding partners, and subcellular localization.

Such domains can be divergent in sequence and occur in contrasting multidomain contexts. This leads to difficulties in unraveling the evolution and function of multidomain proteins. To help in solving these problems, SMART has been developed to identify and annotate protein domains, particularly those in eukaryotes that are genetically mobile and difficult to detect.

3.3.1 The SMART Alignment Set

Domain detection in SMART relies on multiple sequence alignments of representative family members.

Alignment Construction Protocol The starting point for constructing a multiple sequence alignment that optimally represents a domain family is an alignment of divergent family members based on known tertiary structures, where possible, or from homologs identified in a PSI-BLAST [3] analysis. These alignments are optimized manually and, after construction of a hidden Markov model (HMM), used to search current sequence databases (Figure 4). Each sequence of the alignment is also used as a query in a PSI-BLAST search. All sequences that are significantly similar [as detected by HMM ($E < 0.01$) or PSI-BLAST ($E < 0.001$) searches] are added to the alignment using the sequence-versus-HMM alignment method of HMMer. Alignments are checked manually for potential false positives or misassembled protein sequences derived from genomic sources. From this alignment, one of each sequence pair sharing $> 67\%$ identity is deleted to reduce redundancy. The resulting alignment is used as a starting point for a subsequent round of searches. This iterative procedure is pursued until no new homologs are detected.

Searching Method To maximize the sensitivity of domain and repeat detection, SMART uses hidden HMMer models as implemented in the HMMer software package (http://hmmer.wustl.edu/). HMMER provides statistically sound E values, thus giving a robust estimate of the significance of a domain hit. [The E value represents the number of sequences having a score $\geq X$ that would be expected absolutely by chance. The E value connects the score (X) of an alignment between a user-supplied sequence and a database sequence, generated by any algorithm, with how many alignments having similar or greater scores would be expected from a search of a random sequence database of equivalent size.] From a database search with an HMM derived from the SMART alignment, the highest per-protein E value of identified true positives (E_p) and the lowest per-protein E value of predicted true negatives (E_n) are stored within the SMART database. Similarly, for two or more repeats in a protein, the lowest E value of a false positive repeat (E_r) is stored. To ensure that the E value thresholds are independent of database size, the size of the protein database used when deriving the thresholds is also recorded. SMART predicts a domain homolog within any sequence that either has an E value $< E_p$ or else when $E_p <$ E value $< E_n$ and E value < 1.0. If no repeat threshold is defined, all hits in a protein are reported; otherwise only those with E values $< E_r$ are shown.

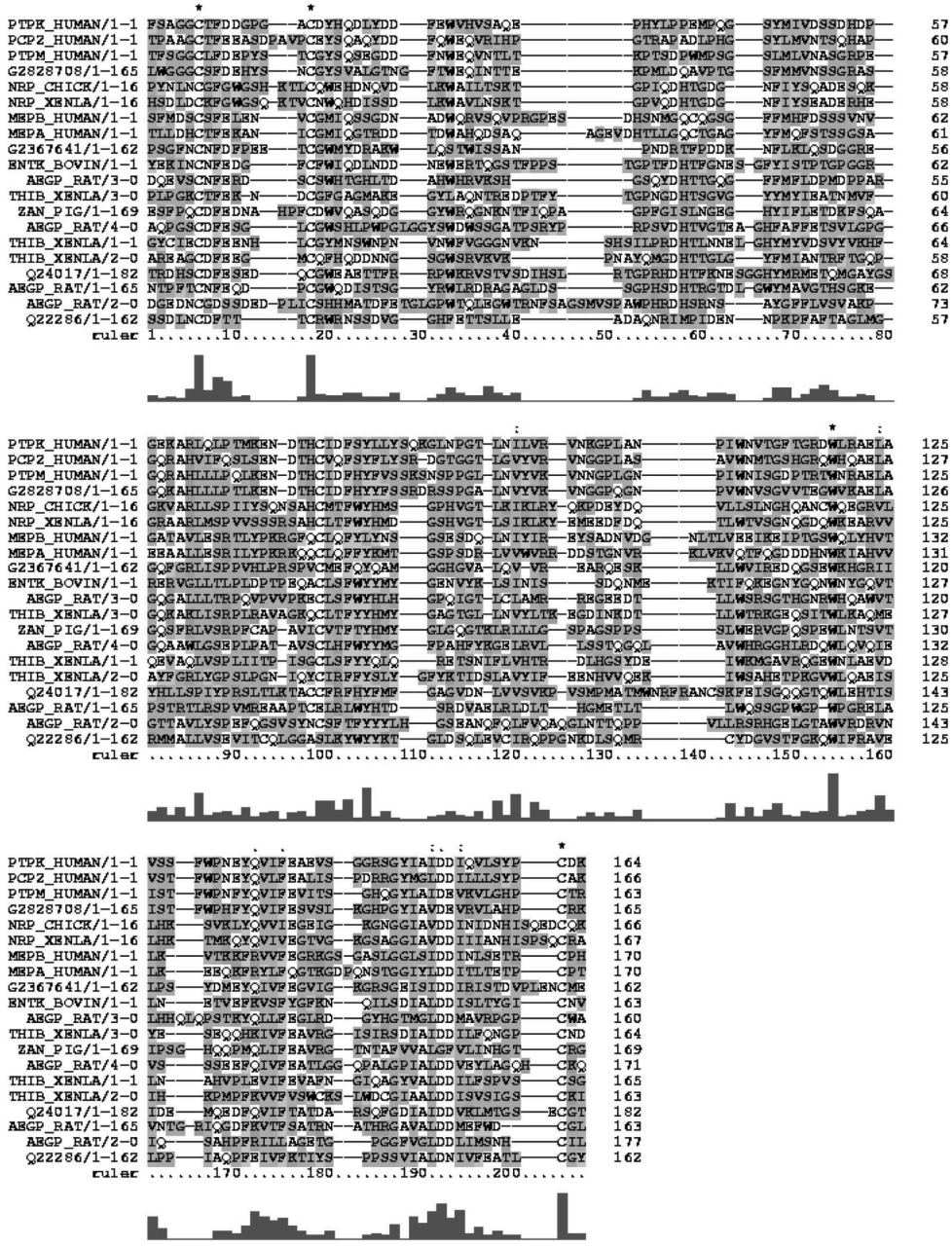

Fig. 4 SMART alignment representing the MAM domain.

Domain Coverage Originally, SMART was intended as a tool for the analysis of domains involved in eukaryotic signal transduction [76]; it was later expanded to detect domains of extracellular proteins and bacterial two-component regulatory systems. Gradually, various domains associated with DNA, RNA, chromatin, and cytoskeletal functions have been added. Over the past few years, to augment the SMART domain set, several semi-automatic search methods to identify new and biologically interesting domains were developed. The current release of SMART (version 4.0) includes nearly 700 protein domains.

3.3.2 SMART Relational Database System

The core of SMART is a relational database management system (RDBMS) powered by PostgreSQL (http://www.postgresql.org/), which stores information on SMART domains and the underlying nonredundant protein database.

Protein Database Basic components of SMART's source sequence database are the Swiss-Prot and SP-TrEMBL [10] protein sets, which have been used by SMART since its inception. This set was recently expanded by inclusion of all proteomes available in the Ensembl collection [17]. Sequences from all sources are compared, and a nonredundant set of proteins with multiple identifiers per sequence is generated. Sequences are retrievable and linkable via any of the original identifiers.

Domain Database The SMART domain database stores information on each domain's presence in all proteins in the relational database. Each domain's hit borders, raw bit score, and Expect (E) value are recorded, together with the protein accession code, description, and species name. In addition to domain information, other intrinsic features of each protein, such as transmembrane regions, coiled coils, signal peptides, and internal repeats are included.

3.3.3 Web Interface

SMART provides a web-based interface to its underlying relational database and HMMer-based search engine. There are two principal ways of using SMART: individual sequence analysis and domain architecture analysis. Here we describe major features of the current SMART (version 4.0).

Sequence Analysis SMART uses the CRC64 algorithm to calculate checksums for all user-supplied sequences. If a matching checksum is found in the SMART database, precalculated results are displayed. If there is no match, HMMer software is used to scan the sequence with all SMART profiles. It is also possible to include Pfam profiles in the search.

Resulting schematic protein representations (Figure 5) are easy to interpret: a gray line shows the protein backbone, and different colored shapes represent domains and features that are confidently predicted. If a user-supplied sequence has a matching checksum identified, several important features become available in the main results page.

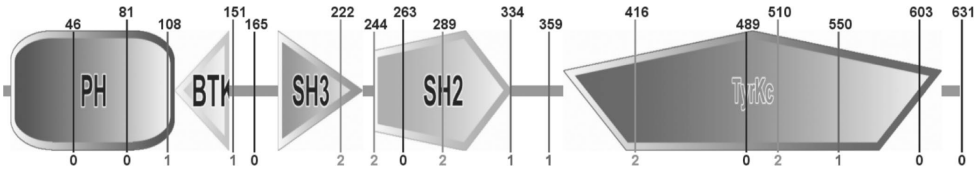

Fig. 5 SMART representation of mouse tyrosine protein kinase TEC (ENSMUSP00000006349). Gray lines show the protein backbone, with domains represented by different colored shapes. Intron positions are indicated by vertical lines showing the amino acid location and the intron phase. Intron positions are taken from Ensembl gene predictions.

Where available, intron positions are shown in schematic protein figures. For proteins that match any of the Ensembl predictions, SMART shows intron positions as vertical colored lines (Figure 5). This information is retrieved from a precalculated mapping of Ensembl gene structures to protein sequences.

Extra information may be associated with the sequence. If multiple IDs are associated with the same sequence, users receive a list of all IDs with links to corresponding source databases. Since SMART incorporates Ensembl genomes, users also receive a list of alternative splices of the gene encoding the analyzed protein (if there are any). It is possible to either display SMART protein annotation for any of the alternative splices or obtain a graphical multiple sequence alignment of all of them (Figure 6).

Orthology information: SMART provides orthology information for all Ensembl-predicted proteins. These relationships are distinct from those provided by Ensembl. There are two separate sets of orthologs for each protein: 1 : 1 reciprocal best matches in other genomes and orthologous groups with reciprocal best hits from all genomes analyzed (i.e., each of these proteins has exactly one

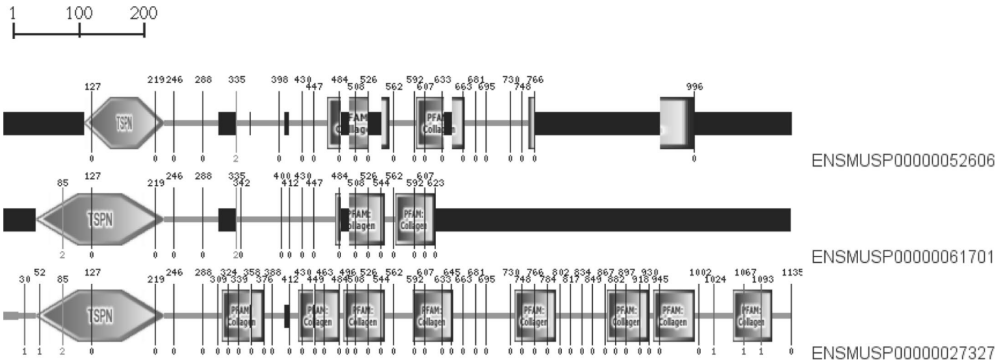

Fig. 6 SMART graphical alignment of alternative splice variants of *Mus musculus* procollagen gene *ENSMUSG00000026141*. Domain and intron positions are adjusted according to gaps in the alignment (black boxes).

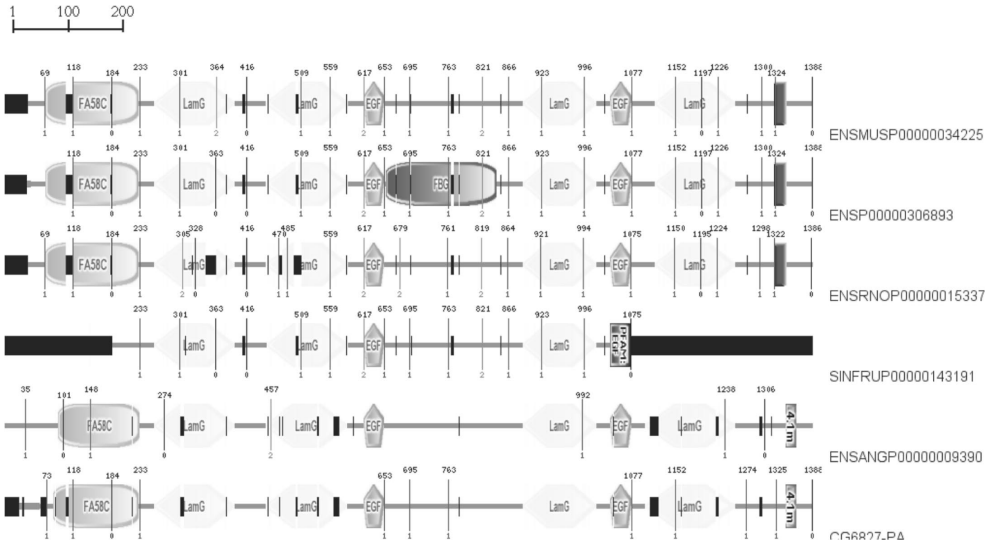

Fig. 7 SMART representation of an orthologous group alignment. Orthologous proteins from different species are aligned using Clustal W. Domains, intrinsic features, and introns are mapped onto the alignment with their positions adjusted according to gaps (black boxes). This tool allows easy visual comparison of intron positions and their relations to protein features. Proteins displayed: *H. sapiens* ENSP00000306893, *M. musculus* ENSMUSP00000034225, *Rattus norvegicus* ENSRNOP00000015337, *Fugu rubripes* SINFRUP00000143191, *Drosophila melanogaster* CG6827-PA, and *Anopheles gambiae*, ENSANGP00000009390.

ortholog in all six genomes). Orthologous groups are displayed as graphical multiple sequence alignments (Figure 7). All orthology information is extracted from all-against-all Smith–Waterman similarities for combined proteomes, using a previously described method [103].

Domain Architecture Analysis (Architecture SMART and Alert SMART) Architecture SMART allows users to search for specific domain architectures using AND/NOT logic. Since the SMART database includes intrinsic protein features as well as precalculated results for Pfam [11] domains, these can be used together with SMART domains. For example, it is possible to identify receptor tyrosine kinases by searching for proteins that contain both a tyrosine kinase domain and a predicted transmembrane region (query "TyrKc AND TRANS", Figure 8).

In addition to standard domain querying, SMART can be used to find proteins based on gene ontology (GO [8]) terms associated with domains. Associations of domains with GO are taken from Interpro [61]. Querying with GO terms is a two-step process. In the first step, the user obtains a list of domains matching the GO terms entered. After selecting the domains of interest from the list, proteins

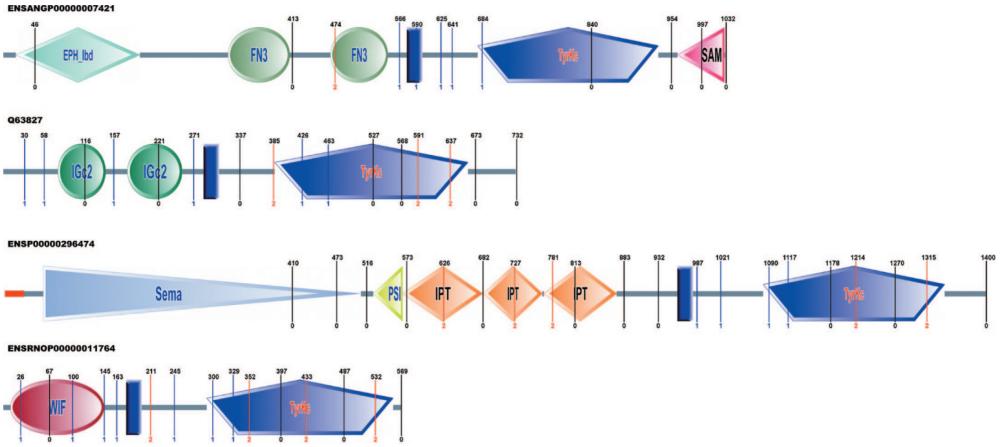

Fig. 8 Using intrinsic features in domain architecture queries. The SMART database was queried for all proteins containing a tyrosine kinase domain and a transmembrane region ("TyrKc AND TRANS"), and 660 proteins were found,including the four displayed here. The color of domain names correlates with both subcellular localization (blue = extracellular, black = intracellular)and catalytic activity (red = catalytically active).

containing those domains are displayed. As with standard domain querying, results can be limited to specific taxonomic ranges.

Finding Proteins with Similar Domain Architecture SMART can search for all proteins that have the same domain architecture as the query (having all the domains of the query protein in the same co-linear order) or that have an identical set of domains (at least one of all domain types of the query protein, irrespective of order). Identification of proteins having identical or near-identical domain architectures as the query sequence may improve predictions of protein functions. This feature also reveals, by using a taxonomic breakdown, the phyletic distribution of a given architecture.

3.3.4 Application of SMART

Apart from its use as a web tool, SMART has been applied to large-scale annotation projects, such as annotation of the human, mouse, and mosquito draft genome sequences [48, 95, 103]. It was also used in the investigation of single domain families in model organisms [40] and for the study of sequence conservation in multiple alignments [67]. In conjunction with genomic data, SMART was used for the study of conservation of gene (i.e., intron/exon) structure [13].

SMART has also been incorporated into other domain and protein family resources that are used for the primary annotation of sequence databases. It is a component database of Interpro [61], which contributes to the annotation of Swiss-Prot

sequences [10], and of the Conserved Domain Database (CDD), which contributes to the annotation of RefSeq sequences [70].

3.4 Other Features and Resources

3.4.1 Globular Repeats

Repeats can be hard to detect in protein sequences: most of them are short, and their sequences are often highly divergent. The numbers of repeats in different proteins are extremely variable. Finally, defining the first and last residues of a repeat is more contentious than for a domain, since repeats are more prone to circular permutation than are domains, particularly within closed structures [74], and to partial truncation, resulting in non-integer repeat numbers.

Repeat detection methods are often incorporated into domain prediction servers (for example, SMART uses the Prospero [60] program from the Ariadne package), but dedicated protein repeat prediction servers also exist, e.g., REPeats (http://www.embl-heidelberg.de/~andrade/papers/rep/search.html [5]). The GlobPlot method described in Section 4.1.4 is also fairly capable in detecting repeats.

3.4.2 Domain Interaction Prediction

Although homology-based methods are a primary source of globular domain discovery, protein-protein interactions are also being used and becoming more popular for this purpose. One of the servers for exploration of such interactions is STRING. The STRING database (http://string.embl.de/) is dedicated to proteomewide prediction of protein-protein associations [94]. It is an integrated resource relying on a wide range of experimental and computational datasets to make reliable interaction predictions. It contains genomic context associations (derived from genome comparisons), interactions derived from coexpression analysis, and various types of high-throughput experimental data, all of which are stringently bench-marked by using a common reference.

3.4.3 No Domains?

If no domains are found by, e.g., SMART or Pfam, this does not mean that your favorite protein does not contain any higher fold or globular domains. Most often, it simply indicates the presence of nonannotated domains, which of course has potentially higher discovery value. However, several resources exist to help identify potential domain boundaries and give hints as to the structure of what might be hidden in the sequence.

Secondary Structure Prediction: Good Prediction of secondary structure is the most mature of any structure prediction strategy, and accuracies of up to ~80% can be achieved [20, 21, 68]. An initial BLAST search to find homologous proteins is important to get a better idea of the function and to build a sequence set for a multiple alignment that can be used on secondary structure prediction servers

such as PredictProtein (PROFsec, http://www.predictprotein.org/) and JPRED (http://www.compbio.dundee.ac.uk/~www-jpred/).

Tertiary Structure Prediction: Difficult Prediction of tertiary structure and folds is still error-prone and difficult; however, having good secondary structure predictions at hand can assist this analysis. Perhaps the best approach is to submit the sequence to one of the homology-based prediction servers, such as the 3D-JURY meta-server (http://bioinfo.pl/Meta/ [34]) or SWISS-MODEL (http://www.expasy.org/swissmod/SWISS-MODEL.html [77]). Other resources can be found on the websites for the evaluation competitions CASP (http://predictioncenter.llnl.gov/casp5/Casp5.html) and CAFASP(http://bioinfo.pl/cafasp/).

Other Sequence Features: Narrowing Down Domain Boundaries Single transmembrane segments (TM1), coiled coils, and low-complexity regions are all incorporated in the SMART server. Sometimes low-complexity regions are disordered (see Section 4.1.3). Coiled coils are also disordered sequences; however, they behave like globular units after the coiled-coil structure is formed, which is a very clear example of disorder–order transition.

At EMBL in Heidelberg we have two additional methods that are useful in the definition of potential domain boundaries: DomCut [84] and GlobPlot [52], see Section 4.1.4. Another resource for potential domain boundary prediction is DomPred (http://bioinf.cs.ucl.ac.uk/dompred/ [58]).

Many proteins are entirely and natively unstructured and without globular domains, and the rest of this chapter is dedicated to the analysis of this part of protein space.

4
Analyzing Nonglobular Protein Segments

Since most attention in assigning function to proteins has been on globular domains, there are relatively few tools for analyzing the nonglobular protein space. Structural biology has tended to avoid unstructured proteins and regions (e.g., by removing them in recombinants), which has led to a skew toward globular proteins in structural datasets.

However, this neglect is not confined to structural biology – bioinformatics has also tended to keep nonglobular function prediction under the academic carpet. Although resources are readily available for revealing globular domains in sequences, until recently there has not been any comprehensive collection of short functional sites/motifs comparable to the globular domain resources. Yet these are just as important for the function of multidomain proteins. Indeed, it is impossible for a researcher to find a list of currently known motifs – going through the literature to retrieve them is impractical without foreknowledge in more areas than any one person has. This neglect is primarily due to the fact that short sequence motifs are

statistically insignificant and difficult to handle compared to domains for which accurate sequence models can be produced.

4.1 Unstructured Regions: Protein Disorder

The approach to finding functional sites is fundamentally different from the one described above for globular domains. Since linear motifs are often shorter than 10 amino acids, they overpredict massively even if they are described by using artificial neural networks or other sensitive probabilistic methods. However, linear motifs are context-dependent in the sense that they are functional only if they are exposed for interaction with a modular domain or in the right cell compartment. Structurally they prefer to be in nonglobular or disordered regions of the protein, both of which can be detected fairly accurately. A typical functional site is shown in Figure 9; notice the linear unstructured and flexible protein backbone, a requirement for the CSK kinase to be able to modify the tyrosine.

In the following we discuss how to find potentially nonglobular areas, including those that appear structurally disordered, and how to predict functional sites in them.

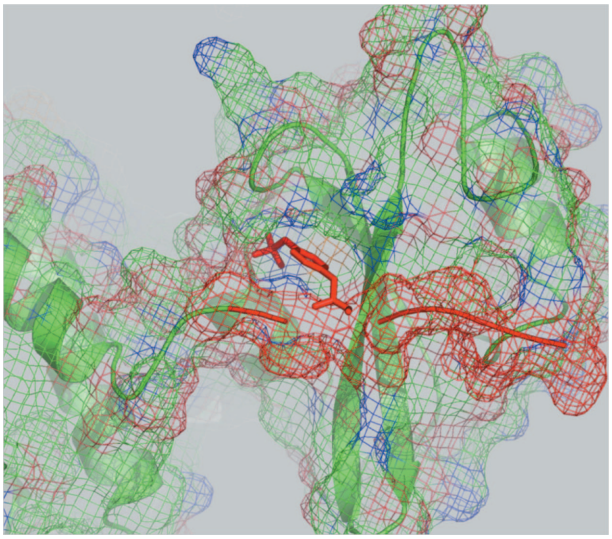

Fig. 9 The C-terminal CSK Tyr-phosphorylation site in Src (PDB: 1fmk) in the closed conformation bound to the SH2 domain. This linear motif (red) shows the general features of an ELM: it is linear in sequence and structure space. The sequence of the instance of this functional site is TEPQYQPGE. In the ELM resource this is called the MOD_TYR_CSK functional site and is described by the pattern [TAD][EA].Q(Y)[QE].[GQA][PEDLS]. The image was created with PyMOL (Table 2).

Recently, it has become possible to analyze natively unstructured proteins by methods such as NMR. Besides their high content of functional sites, disordered and nonglobular regions are exciting for many other reasons.

4.1.1 What Role Does Protein Disorder Play in Biology?

Target Selection In the post-genomic era, discovery of novel domains and functional sites in proteins is of growing importance. One focus of structural genomics initiatives is to solve structures for novel domains and thereby increase the coverage of fold and structure space [14]. During the target selection process in structural genomics/biology, it is important to consider intrinsic protein disorder, because disordered regions (at the N and C termini or even within domains) often lead to difficulties in protein expression, purification, and crystallization. It is therefore essential to be able to predict which regions of a target protein are potentially disordered/unstructured.

IDPs (Intrinsically Disordered Proteins) Although IDPs (also known as intrinsically unstructured proteins) are under-researched, an increasing number are being found. These are proteins or domains that, in their native state, are either entirely disordered or contain large disordered regions. More than 100 such proteins are known, including Tau, prions, Bcl-2, p53, 4E-BP1, and HMG proteins (see Figure 14) [7, 47, 56, 90, 91].

Protein disorder is important for understanding protein function as well as protein folding pathways [67, 92]. Although little is understood about the cellular and structural meaning of IDPs, they are thought to become ordered only when bound to another molecule (e.g., CREB–CBP complex [72]) upon changes in the biochemical environment [27, 29].

Function of Disorder and IDPs The current view on protein disorder is that it allows for more interaction partners and modification sites [53, 90, 99]. However, we have not been able to confirm this hypothesis by analyzing a large interaction dataset (unpublished results). This might be because such datasets are enriched in nontransient interactions, but interactions carried out by disordered proteins are transient.

Perhaps disordered proteins have evolved to provide a simple solution to having large intermolecular interfaces while keeping smaller protein, genome, and cell sizes [36]. It has been proposed that having several relatively low-affinity linear interaction sites allows for a flexible, subtle regulation as well as accounting for specificity and cooperative binding effects [31]. In light of the modular model described in Section 2.1, we can see how these sites can be used in a combinatorial manner to generate a very large set of potential interaction environments.

Protein Disorder and Disease Structural disorder in proteins is now known to play a central role in diseases mediated by protein misfolding and aggregation [12, 45, 78]. Amylogenic diseases such as Alzheimer's, Type II diabetes, and BSE are thought to be related to the occurrence of short linear motifs in unfolded regions. These motifs

are important for initiation of the formation of the amyloid fibers that cause great harm to the cellular environment, particularly in brain tissue. There are several proposed peptide models for these motifs, and the structural context in which they occur are under investigation [24, 55, 63, 83].

Other diseases such as Parkinson's, Huntington's, and serpinopathies are related to misfolding of proteins. The understanding of protein misfolding is related to analysis of the unstructured ensemble or the unfolded state of a polypeptide. This state can be analyzed in natively disordered proteins [30].

How does one characterize protein disorder and nonglobular regions? The field of protein disorder studies has, so far, failed to reach any agreement on this.

4.1.2 What is Protein Disorder?

No commonly agreed definition of protein disorder exists. The thermodynamic definition of disorder in a polypeptide chain is the random-coil structural state. The random-coil state can best be understood as the structural ensemble spanned by a given polypeptide in which all degrees of freedom are used within the conformational space. However, even under extremely denaturing solvent conditions, such as 8 M urea, this theoretical state is not observed in solvated proteins [46, 66, 89]. Proteins in solution thus seem to always retain a certain amount of residual structure.

Protein disorder is observed by a variety of experimental methods, such as X-ray crystallography; NMR, Raman, and CD spectroscopy; and hydrodynamic measurements [29, 82]. In vivo studies of disorder are possible with NMR spectroscopy on living cells (e.g., anti-sigma factor FlgM [22]). Each of these methods detects different aspects of disorder, resulting in several operational definitions of protein disorder (see [90] for a review).

Regions without regular secondary structure can be predicted by the NORSp (nonregular structure) server [53]; however, as the authors point out, such regions are not necessarily disordered. Structures such as the Kringle domain (PDB: 1krn) are almost entirely without regular secondary structure in their native state, but they still have tertiary structure in which the basic building block is coils. These loopy proteins are not necessarily IDPs, since they can still form a well defined globular tertiary structure.

In our work we have used four definitions of protein disorder:

- Loops/coils as defined by DSSP [44]. Residues are assigned to one of several secondary structure types. For this definition we consider residues in an α-helix (H), 3_{10} helix (G), or (β-strand (E) to be ordered and all other states (T, S, B, I) to be in loops (also known as coils). Loops/coils are not necessarily disordered (e.g., turns); however, protein disorder is found only within loops. It follows that one can use loop assignments as a necessary but not sufficient requirement for disorder; a disorder predictor based entirely on this definition is thus promiscuous.
- Hot loops constitute a refined subset of the above: namely, those loops having a high degree of mobility as determined from Cα temperature factors (B factors).

It follows that highly dynamic loops should be considered disordered. Several attempts have been made to try to use B factors for disorder prediction [15, 28, 32, 93, 104], but there are many pitfalls in doing so, because B factors can vary greatly within a single structure due to the effects of local packing and structural environment. Recent progress in deriving propensity scales for residue mobility based on B factors [81] has encouraged us to use B factors for defining protein disorder. The details for hot loops can be found in the methods part of [51].

- Missing coordinates/remark465 in X-ray structure, as defined by remark465 entries in the PDB. Nonassigned electron densities most often reflect intrinsic disorder and were used early, for disorder prediction [50].
- Russell–Linding propensities are parameters based on the hypothesis that the tendency for disorder can be expressed as P = RC – SS where RC and SS are the propensities for a given amino acid to be in random coil and regular secondary structure, respectively. This scale was defined during the development of the GlobPlot predictor described in Section 4.1.4.

Figure 10 shows the disorder propensities for each amino acid by our four definitions of disorder. A more detailed discussion of these values can be found in [51], but in general, hydrophobic residues promote order according to all definitions of

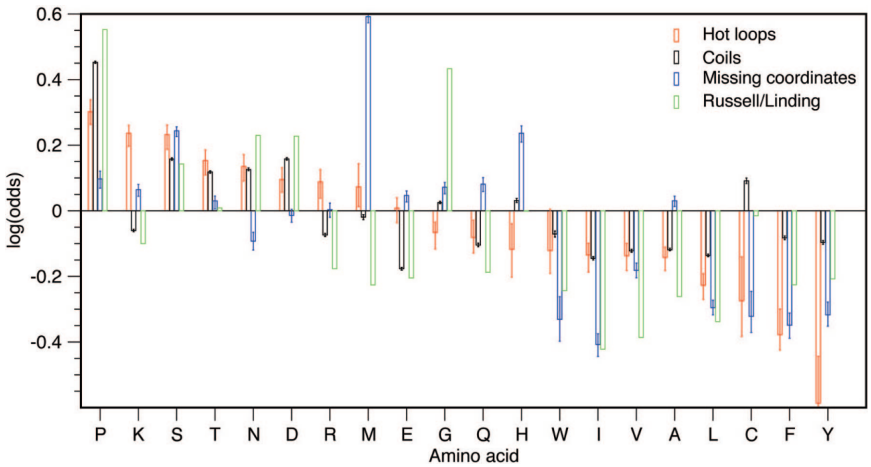

Fig. 10 Propensities of the amino acids to be disordered, according to the definitions used in DisEMBL and GlobPlot (sorted by hot loop preference). This scale directly reflects the datasets used for training; however, it is only a rough approximation of what the DisEMBL neural networks use in predicting disorder. Error bars correspond to the 25th and 75th percentiles as estimated by stochastic simulation. The Russell–Linding scale is an absolute scale. Methionine suffers a bias in the remark465 dataset for at least two reasons: (1) often the N-terminal methionine is missing; and (2) some structures are solved using selenomethionine derivatives for phasing, which can lead to deletion of the residue in the PDB entry. The same bias is seen in ([29], Figure 10).

disorder. Disorder-promoting residues include proline, lysine, serine, threonine, and methionine.

4.1.3 Methods for Finding Protein Disorder

Several other attempts have been made to predict disorder. Perhaps the earliest were methods of finding regions of low complexity. Although many such regions are structurally disordered, the correlation is far from perfect, because regions of low sequence complexity are not always disordered (and vice versa) [27]. Likely the strongest evidence for this correlation comes from the fact that low-complexity regions are rarely seen in protein 3D structures [75]. Methods to predict low complexity, like SEG [98] and CAST [69], are thus often used for this purpose. Methods using hydrophobicity can also give hints about disordered regions, because low-complexity regions are typically exposed and rarely hydrophobic.

The first tool designed specifically for prediction of protein disorder was PONDR (predictor of naturally disordered regions, http://www.pondr.com [32, 33, 73]). It is based on artificial neural networks. PONDR is, however, not freely accessible to academics. Refer to [59] for a recent evaluation of disorder prediction (DisEMBL was published after CASP5).

Prediction of protein tertiary structure may be an alternative route to disorder prediction, although such methods are computationally intensive and error-prone. Moreover, such methods are usually designed to predict the structure of globular domains, and their behavior with other sequences can be unpredictable.

At the EMBL in Heidelberg we have developed methods for finding unstructured regions from sequence data alone. These tools were primarily developed for use in the ELM project to help find regions potentially containing functional sites. However, these tools are now being used by several structural genomics initiatives and laboratories around the world who are either studying IDPs or trying to optimize their recombinant protein expression vectors by cutting out disordered segments.

4.1.4 GlobPlotting

GlobPlot was invented specifically to aid the ELM project; however, it proved to be of much wider interest [52]. From the beginning we wanted a graphical tool that could generate easy-to-interpret plots of the tendency within a sequence for structured or lack of structure. The basis for GlobPlot was the Russell–Linding scale mentioned earlier in Section 4.1.2. The combination of random coil and secondary structure in the Russell–Linding scale enhanced the discrimination of the graphs and was the key factor in the success of this scale at detecting both disorder and globular packing.

GlobPlot is not intended to be a competitor in secondary structure prediction, because it cannot give the same level of detail as can be obtained from secondary structure prediction based on multiple alignment. GlobPlot is an ab initio method, i.e., it requires only one sequence and can therefore be applied to novel sequences having no homologs, i.e., it does not use multiple alignment. The basic algorithm behind GlobPlot is beautifully simple and very fast: each amino acid a_i has a defined

propensity $P(a_i)R$ (see Russell–Linding in Figure 10). Given a protein sequence of length L, we define a sum function $Dis(a_i)$ as follows:

$$Dis(a_i) = \sum_{j=1}^{L} P(a_i)$$

where $P(a_i)$ is the propensity for the ith amino acid. The GlobPlot webserver plots the function, and the graphs are referred to as globplots. Before plotting, the digital-smoothing Savitzky-Golay algorithm is used to reduce noise on the curve.

Analyzing a GlobPlot Reading globplots is fairly easy, but different from, e.g., hydropathy plots, in that globplots are cumulative-sum curves rather than derivative curves. Because GlobPlot plots this running sum, the graph is analyzed by looking at the slope. The numbers on the ordinate do not matter, they equal the running sum, and we are interested only in whether or not a given segment of the graph is disordered. The latter is seen by the decrease or increase in the slope, because that is how the Russell–Linding scale works: negative values correspond to ordered residues, and positive values indicate disorder-promoting amino acids.

We designed GlobPlot like this because we think it results in profile-like, intuitive plots. In particular, we wanted to avoid a high-variation curve such as the derivative curve. The globplot in Figure 11 is a good example of one of these profile-like curves: the GlobPlot plot for mucin predicts that the central part of the protein is almost completely disordered (using the Russell–Linding disorder definition) – this is probably why this protein is so slimy.

Domain detection with GlobPlot is as easy as finding protein disorder, since both features are shown in the plot. To help you to navigate and understand the plots, the webserver overlays the graph with any predicted SMART domains. In domain hunting situations, you would look for downhill regions in the graph. As seen in Figure 12, GlobPlot can detect potential domains: notice the downhill slope whenever a domain is found by SMART/Pfam. GlobPlot often detects additional sequence to be ordered, this is because SMART and Pfam use only the most conserved sequence part of a domain to generate their hidden Markov models for the domain. This indicates that GlobPlotting is useful for domain boundary definition.

4.1.5 Prediction of Multiple Types of Disorder with DisEMBL

The performance of GlobPlot encouraged us to refine our approach and predict disorder in a more traditional biocomputational manner by training artificial neural network predictors for the various definitions of disorder mentioned above. This work led to the DisEMBL disorder predictor ensemble.

DisEMBL is a computational tool for prediction of disordered/unstructured regions within a protein sequence [51]. DisEMBL currently provides three alternative disorder definitions: hot loops, coils, and missing coordinates as defined in Section 4.1.2. The coils predictor is used primarily as a filter to require disorder to be within coil-predicted regions (see Section 4.1.2). DisEMBL is a highly accurate predictor, predicting more than 60% of hot loops with fewer than 2% false positives [51].

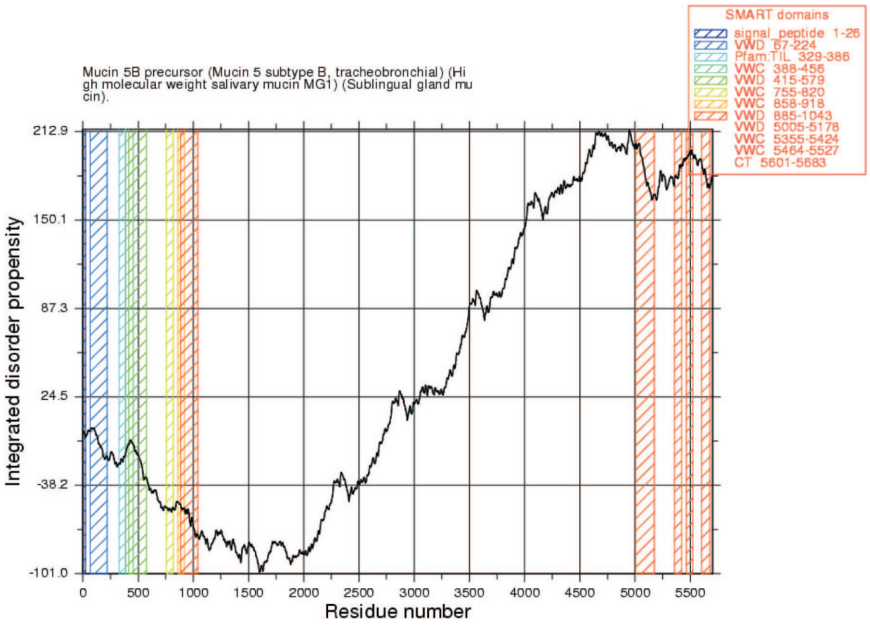

Fig. 11 Globplot of human mucin 5 protein (Swiss-Prot: MU5B_HUMAN). Most of this slimy protein is highly disordered. Since GlobPlot plots a running sum of the propensity to disorder, the graph is analyzed by looking at its slope. The numbers on the ordinate axis do not really matter, it is the uphill or downhill tendency that should be read. Referring to Figure 10 indicates that disorder-promoting propensities are positive, so 'uphill' on the graph is equivalent to disorder.

Hot Loops 'Hot loops' is a novel definition of disorder based on X-ray data. We think that it will prove difficult to pull out a much more precise definition of disorder based on crystallographic data. An example of hot loop results is shown in Figure 14, where we mapped the probabilities shown in Figure 13 onto the structure of nonhistone chromosomal protein 6A from yeast. It is remarkable that a definition based on X-ray data can predict so well for NMR structures, arguing that this novel definition of disorder is relevant. We also showed this correlation earlier, as well as a comparison of the correlations between our alternative definitions of protein disorder [51].

Using DisEMBL DisEMBL is freely available via a web interface (http://dis.embl.de/) and can be downloaded for use in large-scale studies. The web interface is fairly straightforward to use, you can submit a sequence or enter the Swiss-Prot/SWALL accession (e.g., P08630) or entry code (e.g., HMG1_HUMAN). The server fetches the sequence and description of the polypeptide from an ExPASy server using Biopython.org software. The probability of disorder is shown graphically, as illustrated at Figure 13. The random expectation levels for the different predictors are shown on the graph as horizontal lines, but should merely be considered absolute minima. The default parameters are set for optimal prediction and

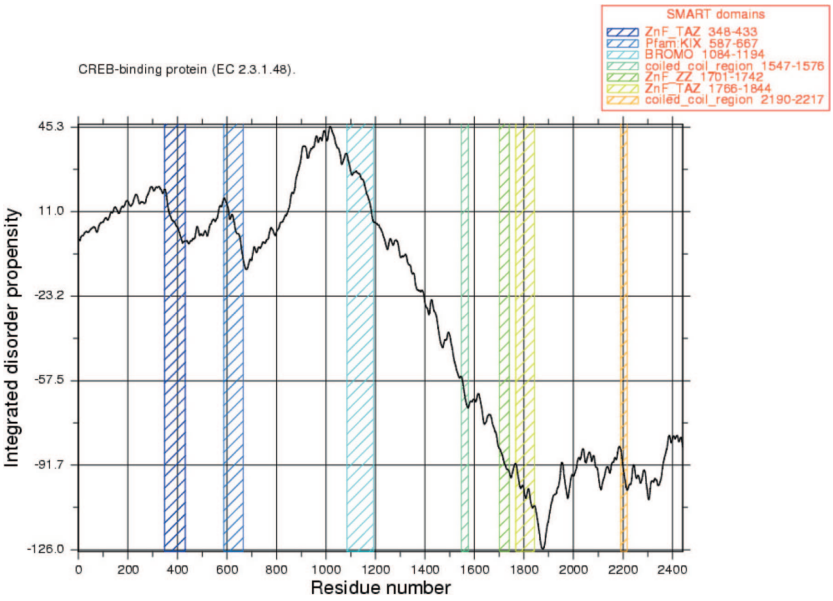

Fig. 12 Globplot of human CREB-binding protein (CBP_HUMAN). About half of the sequence appears to be disordered, with long flexible regions observed at the N and C termini. The flexible region just after the KIX domain might be important for induced binding of the pKID domain of CREB to CBP [23, 72]. For further discussion of disorder in CBP/CREB see Wright et al. [99].

should be changed only in rare situations. On-line documentation of the various settings is provided at http://dis.embl.de/help.html. If the query protein sequence is very long, > 1000 residues, you can download the predictions and use a local graphing/plotting tool such as Grace or OpenOffice.org to plot and zoom the data. A future version of DisEMBL may include a web applet for interactive plotting and zooming of the graphs.

GlobPlot and DisEMBL The GlobPlot algorithm is very simple and intuitive, which is appealing. Although it was originally designed for prediction of protein disorder, the Russell–Linding propensity scale functions just as well for detection of domain boundaries, repeats, and other globular features. The Russell–Linding scale and the SMART domain overlay feature are unique to GlobPlot. DisEMBL is more accurate than GlobPlot in coil prediction, which is related to the Russell–Linding scale. It furthermore provides the novel hot-loop definition of disorder. The two methods complement each other, since they approach disorder prediction differently. In general, we urge you to submit your sequences to both tools.

4.1.6 Design of Protein Expression Vectors

As mentioned earlier, protein disorder is related to problems during protein expression, purification, and crystallization. Other tools such as TANGO

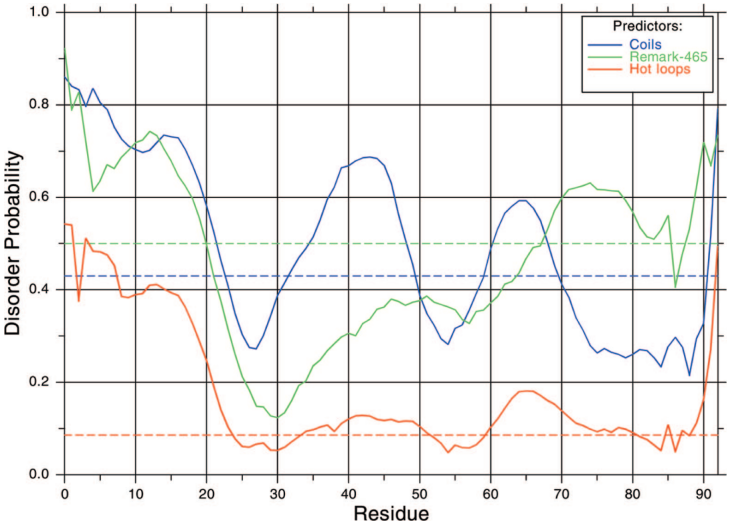

Fig. 13 Sample output from the DisEMBL web server, showing predictions for yeast nonhistone chromosomal protein 6A (high mobility group protein, Swiss-Prot: NHPA_YEAST). The green curve shows the predictions obtained for missing coordinates, red for the hot loop network, and blue for coils. The horizontal lines correspond to the random expectation level for each predictor: for coils and hot loops the prior probabilities were used, and a neural network score of 0.5 was used for remark465. From this plot it is seen that the N-terminal tail of the protein is especially predicted to be disordered. See Figure 14 for a mapping of the hot loop predictions onto the structure of this protein.

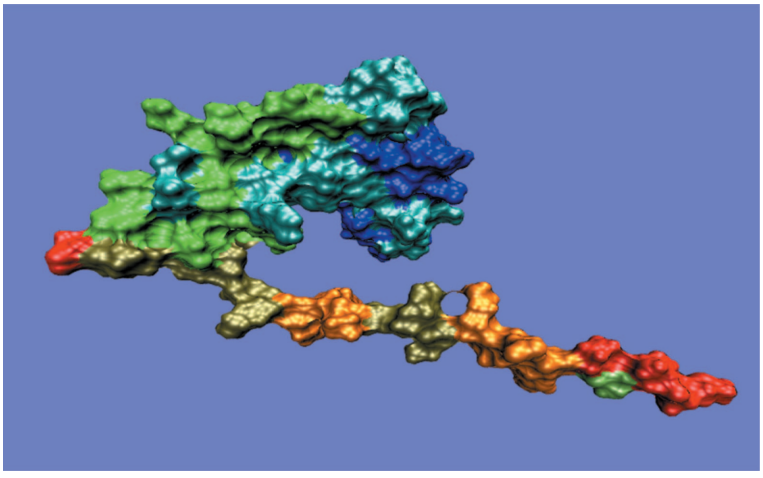

Fig. 14 DisEMBL hot loop predictions mapped on the NMR structure of nonhistone chromosomal protein 6A (high mobility group protein, PDB: 1cg7; model 1, Swiss-Prot: NHPA_YEAST). The predicted probabilities are indicated with a color scale going from blue to red, where red corresponds to the most likely disordered regions and blue to ordered regions. The unstructured tail clearly shows the highest disorder scores (see Figure 13). Surface plot generated with VMD (see Table 2).

(http://tango.embl.de/) deal with protein cross-beta aggregation which is different from the disorder in solvated proteins that our tools predict.

We believe that identification of potential disordered regions should provide a good basis for setting up expression vectors and/or comparing the data with obtained structural data. However, currently we cannot assess which of the definitions of disorder is most appropriate for design of protein expression vectors. We thus strongly encourage feedback on successes and failures in using DisEMBL for expression and structural analysis of proteins.

4.2 Function Prediction for Nonglobular Protein Segments

Having identified candidate unstructured regions, one can start searching for function in them. Most functions correlate with short linear peptide motifs that are used for cell compartment targeting, protein-protein interaction, regulation by phosphorylation, acetylation, glycosylation, and a host of other post-translational modifications. See Figure 15 for an overview of the many functions these sites perform. The number of known categories of functional sites has increased dramatically in the past few years, and it is obvious that there are more to be discovered. These sites are usually short and often reveal themselves in multiple sequence alignments as short patches of conservation, leading to their definition as linear motifs. In addition to occurring outside globular domains, some sites, e.g., phosphorylation sites, are often found in exposed flexible loops protruding from globular domains.

Functional Group	Functional Sites (ELMs)
Targeting signals	KDEL, NLS, NES Mitochondrial Import signals PTSI/II, Signal peptides
Linear protein interaction sites	EH, PDZ, SH2, SH3, WW, PCNA, LXXLE Calmodulin/Cyclin binding helix Protease inhibitors Active peptides and hormones
Covalent modification sites	Phosphorylation, methylation, acetylation glycosylation, fucosylation, prenylation myristoylation, ubiquitinylation
RNA interaction motifs	RX, RGY, RGG motifs
Processing sites	Diverse protease cleavage targets in secretion, apoptosis etc.

Fig. 15 Main classes of functional sites. Functional sites are as varied and numerous as domains are. On a proteome level we expect at least five sites per protein, resulting in about 150 000 instances in the human proteome. This indicates the presence of a gigantic and complex interaction and regulatory system.

Considering the abundance of targeting signals and post-translational modification sites, it is reasonable to assume that there are more functional sites than globular domains in a higher eukaryotic proteome.

4.2.1 Available Resources

ELM is the largest collection of linear motifs, followed by Scansite and PROSITE [64, 80, 101]. Scansite is a very capable resource focusing on cell signaling. It complements ELM in using position-specific scoring matrices (PSSMs) for prediction, which are more sensitive than the regular expressions ELM uses. However, Scansite does not provide an annotated database similar to ELM.

A series of individual predictors of functional sites can be found at http://www.cbs.dtu.dk/services/ which is hosted by the Center for Biological Sequence Analysis in Denmark. The CBS focuses on providing high-performance neural network predictors but without any annotated knowledgebase interface, taking a complementary approach to the other resources.

The PROSITE database has collected a number of linear protein motifs, representing them as regular expressions. PROSITE patterns have been very useful but suffered from severe overprediction; more recently the database has emphasized globular domain annotation at the expense of linear motifs.

Also of interest are protein interaction databases such as BIND and DIP [9, 100]. More informative protein interaction databases that store known instances of linear motifs include MINT [102] and Phospho.ELM at http://phospho.elm.eu.org/. Databases of instances are not directly useful for prediction but provide valuable data-mining resources.

It was recently demonstrated that short functional sites or protein features are crucial for the classification of protein function [41]. The Protfun method is an ab initio method for prediction of higher functional classes based on sequence features alone [42].

4.3 The Eukaryotic Linear Motif Resource: ELM

In this section we describe the ELM resource in detail, since it is the largest resource for linear motifs.

The Eukaryotic Linear Motif server (http://elm.eu.org/ [71]) is a new bioinformatics resource for investigating candidate short functional motifs in eukaryotic proteins. Some of the concepts used within ELM are defined in Table 1. An example of the concepts used in practice can be seen in Figures 16 and 17.

Linear motifs are short (usually < 10 amino acids) and therefore difficult to evaluate, since the usual significance assessments are inappropriate. Therefore, the ELM resource deploys logical context filters to eliminate false positives. The prediction strategy ELM uses is what we call knowledge-based decision support (KBDS). The basic idea is that, since we cannot discriminate ELMs based on sequence matching, we can use a knowledge base of contextual information regarding functional

Table 1 Definitions of concepts used in the ELM resource. Functional sites are, as opposed to, e.g., active sites, short and linear in sequence and structure space. In the ELM resource we describe functional sites as linear motifs. Here, the linear motif is shown as a regular expression or pattern, but it could as well have been another type of sequence model, e.g., a hidden Markov model

Concept	Definition	Example
A functional site	A set of short linear (sub)sequences that can be related to a molecular function	LIG_RBBD: Rb pocket interacting sequence
An ELM	The common pattern of a set of linear (sub)sequences that can be related to a molecular function	[LI].C.[DE]
An ELM instance	An instance of an ELM in a particular polypeptide	RBB1_HUMAN: LVCHE

sites and ELMs to filter out false positives. This knowledge base is created/curated manually from the scientific literature. Currently, KBDS filters are in place for cell compartment, globular domain clash, and taxonomic range. In favorable instances, the filters can reduce the number of retained matches by an order of magnitude or more.

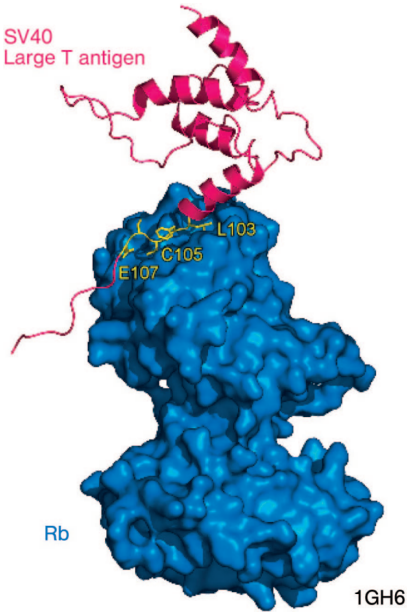

Fig. 16 LIG_RBBD is a functional site responsible for interaction with retinoblastoma (Rb) family proteins. Rb proteins are known for repression of E2F proteins, which are required for transcription of proteins important in the cell cycle. The figure shows the SV40 large T antigen interacting with the Rb pocket (PDB: 1gh6).

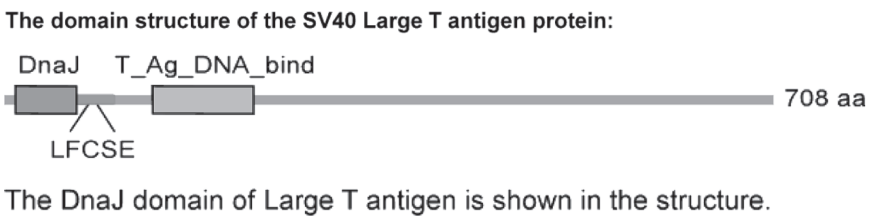

Fig. 17 The location of the instance (LFCSE) of LIG_RBBD in the SV40 large T antigen protein is C-terminal to the DnaJ globular domain.

4.3.1 **ELM Annotation – 'Site seeing'**

All data input is by hand curation, performed by trained molecular biologists. Annotating each ELM is called 'Site seeing' and includes the processes shown in Figure 18. To promote interoperability with other bioinformatics resources, ELM uses three public annotation standards. Gene ontology (GO) identifiers are used for cell compartment, molecular function, and biological process [8, 39], and the NCBI taxonomy database identifiers [97] are used for taxonomic nodes at the apex of phylogenetic groupings in which an ELM occurs. Annotations of ELM instances are assigned ontology terms from the Proteomics Standards Initiative Molecular Interaction ontologies for evidence methods (HUPO.org). In the future the ELM resource will be able to report known instances of ELMs with details about what kind of experiments were performed to show the instance, with links to the relevant literature.

The motif patterns are currently represented as POSIX regular expressions (usable in the Python and PERL languages), analogous to PROSITE patterns, but with a different syntax. For example, the FxDxF motif, which is responsible for the binding of accessory endocytic proteins to the alpha subunit of adaptor protein complex AP-2, has a consensus sequence of F-x-D-x-F and is written F. D. F. Linear motifs in ELM will in the future include motif descriptions according to the Seefeld convention nomenclature for linear motifs (see [1]).

In the future, ELM might incorporate HMMs or other sensitive search methods; nevertheless, linear motifs will continue to overpredict and require alternative approaches for reducing the levels of false positives.

4.3.2 **ELM Resource Architecture**

The core of the ELM resource is a relational database, powered by PostgreSQL, storing data about linear motifs. Figure 19 outlines how the ELM server is implemented. The user submits a protein sequence to the server and receives a list of matching ELMs that have been filtered to remove false positives (it may naturally include false negatives and residual false positives). Matched motifs are usually not statistically significant, and overprediction occurs despite filtering; hence matches should be considered to represent potential true instances of functional sites and should be used as guides to experimental determination.

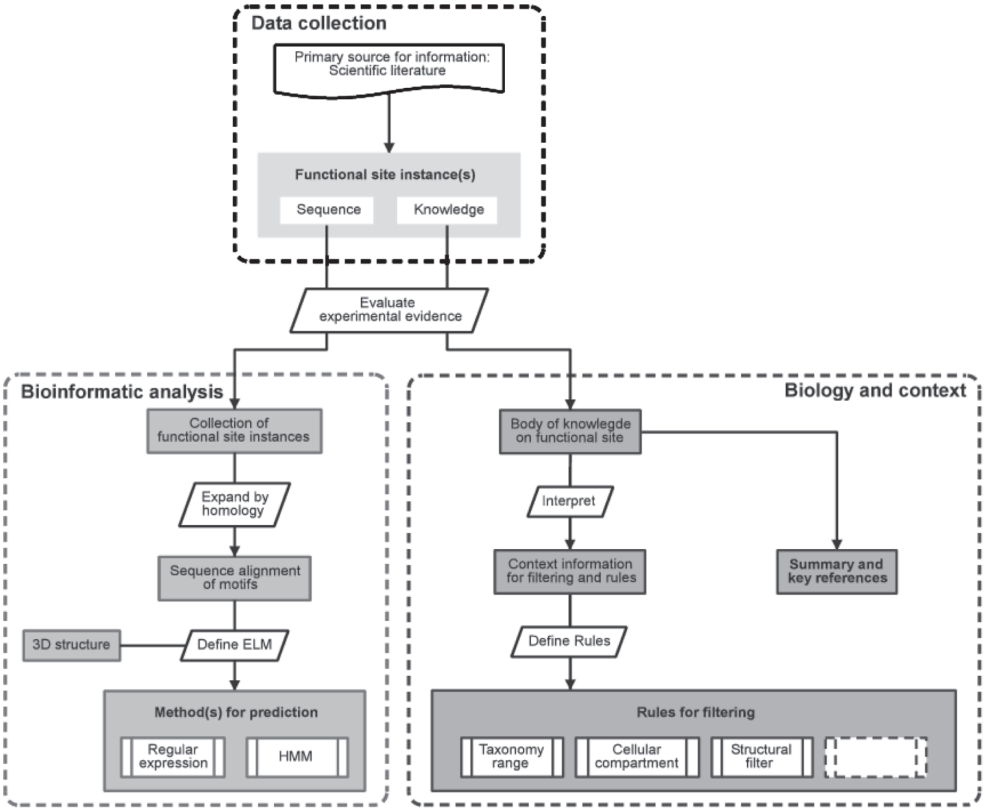

Fig. 18 The flow of the 'siteseeing' process typically involves extensive literature searches, BLAST runs, multiple alignment of relevant protein families, perusal of Swiss-Prot and other online databases, and, where practical, discussion with experimentalists from the field. The empty box symbolizes additional future strategies.

4.3.3 **Knowledge-based Decision Support (KBDS): ELM Filtering**

Sequence-matching methods find many false – but apparently plausible – instances of ELMs that somehow are not recognized by their cognate binding/modification domains. There are two explanations for this:

- One obvious reason why a sequence that matches a motif is not a true functional site is that the motif does not fully and accurately represent the functional site. This can partly be solved by deploying more sophisticated sequence models such as PSSMs or artificial neural networks, an approach used by Scansite and CBS.
- Another reason is that the sequence matches (potential ELM instances) occur in an irrelevant context. They may match a sequence from a wrong cellular compartment or from a species that does not use this functional site. As we have seen, the structural context is also of great importance for linear motifs to be reachable so as to be functional.

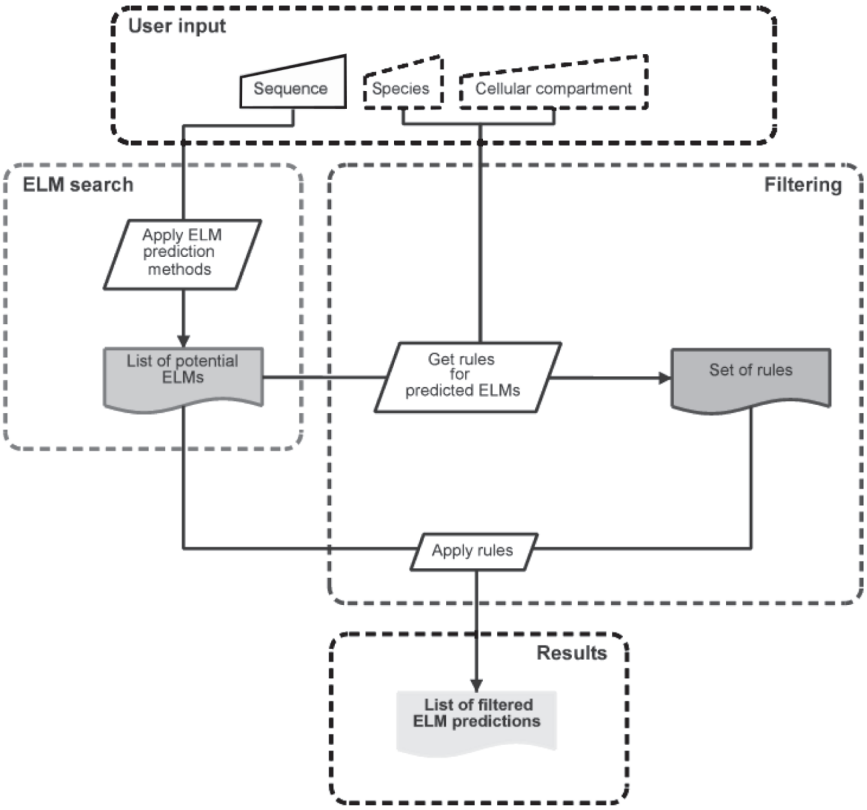

Fig. 19 Flowchart of the ELM server. Dashed boxes indicate the four stages from input to results. As the server is further developed, more filters will be added to allow more query-dependent data to be retrievable.

It is possible to develop context filters that remove such false positives. In ELM we do most of this inside the knowledge database that the 'siteseeing' process is building. The ELM database is designed to accommodate these types of filters or KBDS modules.

Currently, three filters are installed on the ELM server. These filters are not completely accurate and introduce false negatives occasionally, although we try to avoid this as much as possible. In general, the approach in ELM is to predict as few false positives as possible, but it is even more important to avoid false negatives.

Cell Compartment Filter In ELM every linear motif is annotated with GO terms for the set of cell compartments in which it is known to function. For example, KDEL is a signal for retention of the host protein by the endoplasmic reticulum, whereas the SUMO site applies to proteins in the nucleus and the PML body. The user specifies the compartments in which the query protein functions, and all matches for ELMs not found in these compartments are filtered out. In the future ELM may support prediction of compartments using LOC3D [62].

Globular Domain Filter: A Two-track Filtering Strategy Globular domains identified with the SMART and Pfam (domain subset) resources are used for filtering out ELMs. This filter has two tracks:

- a domain filter,
- an ELM rescue or reinstate module.

The domain filter works simply by removing all ELMs within the boundaries of the SMART/Pfam domains matching the same sequence, since they are false positives. The primitive assumption here is that sites within globular units are not accessible and therefore not functional, clearly an oversimplification.

ELMs can occur inside certain domains, e.g., the internal tyrosine phosphorylation sites in the active loops of tyrosine kinase domains, as is described in Section 2.1. This later group of ELMs are to a certain extent being 'rescued' by the ELM rescue module, i.e., for some ELMs certain SMART/Pfam domains are simply not used for filtering in the domain filter.

Given the limited accuracy of the domain filter, the unfiltered results are provided on the results front page. In many situations, users can investigate surface accessibility by examining an available 3D structure, by using a good-quality 2D structure prediction [20, 21, 68], or perhaps by using a homology modeling server such as SWISS-MODEL or the 3D-JURY metaserver [34, 77]. We are currently developing better domain filters, e.g., using surface accessibility from known structures to discriminate false from true positives. A good example of how this might work is shown in Figure 20.

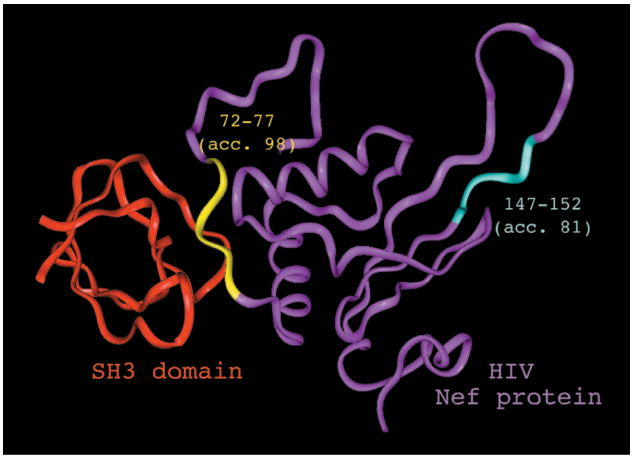

Fig. 20 The V-1 Nef protein (magenta) in complex with wild-type Fyn SH3 (red) domain (PDB: 1avz) contains two potential SH3 ELMs. Residue numbers are given as well as the accessibility (acc) of the highlighted fragments. The yellow sequence is the only one that binds to the Fyn partner. The cyan putative ELM is covered by a loop and has a accessibility of 81% (compared to the 98% of the true binding domain).

Taxonomic Filtering Some types of functional sites are found in all eukaryotes, e.g., the ER retention signal KDEL is universal. But others are restricted to specific eukaryotic taxa. Perhaps most strikingly, the large receptor tyrosine kinase multigene family is found only in metazoa. Each ELM is annotated with one or more NCBI taxonomy nodes to indicate its known phylogenetic distribution. The user provides the query species, and all ELMs that are not assigned to its lineage are filtered out.

4.3.4 Using ELM

The public ELM webserver allows you to retrieve filtered as well as unfiltered raw results. This approach should encourage you to think critically about ELM server results. Figure 21 shows the ELM server output using the human Src sequence as a query. This example indicates the potential of the KBDS approach for improving motif searches. A pipeline interface to ELM prediction for use in proteome analysis is currently being developed and implemented; this pipeline and the results will be made available as soon as possible.

The predictive power of the ELM resource can be enhanced by harnessing it to other data, including experimental results. For example, many protein kinase recognition sites are among those which severely overpredict. If a protein is known

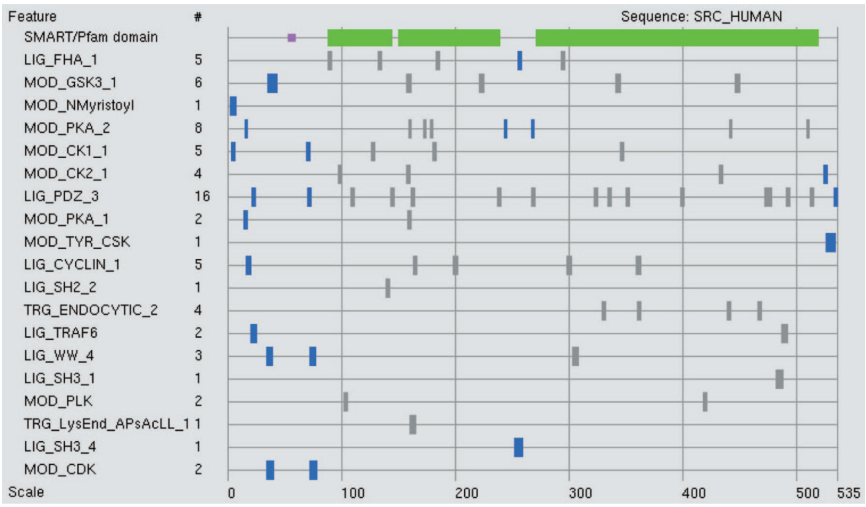

Fig. 21 Sample output from the ELM server. The query sequence was Src (SwissProt: SRC_HUMAN). The surviving ELMs are shown in blue, and the motifs that have been filtered out are shown in grey. This figure illustrates only how the globular domain (green) filter works: of the 103 ELMs in the resource at the time of writing, 27 match the sequence, but two are removed by the species filter, seven by the compartment filter, and five by the SMART/Pfam domain filter. Of the remaining 14 ELMs that survive the filtering, six are known to be true, and two are false negatives, i.e., not predicted by ELM (the C-terminal SH2 ligand and the autophosphorylation site within the tyrosine kinase domain; compare with Figure 1). The functionality of the ELM rescue module is not shown in this figure.

Table 2 Some resources referred to in this chapter. For more information please see the individual websites

Resource	Function classes	htpp://
SMART	globular modular domains	smart.embl.de
Pfam	globular modular domains	www.sanger.ac.uk/Software/Pfam/
Interpro	Globular domain meta server	www.ebi.ac.uk/interpro/
CDD	globular modular domains	web.ncbi.nlm.nih.gov/Structure/odd/odd.shtml
PROSITE	Domain signatures and a few linear motifs	www.expasy.ch/sprot/prosite.html
PredictProtein	Secondary structure prediction	www.predictprotein.org
ELM	Functional sites, ELMs, linear motifs	elm.eu.org
PyMOL	Very nice and easy to use molecule viewer and renderer	pymol.sourceforge.net
VMD	Feature rich molecule viewer	www.ks.uiuc.edu/Research/vmd/
Scansite	Phosphorylation and signaling motifs	scansite.mit.edu
Protfun	Enzyme categories and higher functional classes	www.cbs.dtu.dk/services/ProtFun
NetNglyc	N-glycosylation motifs	www.cbs.dtu.dk/services/NetNGlyc
PredictNLS	Nuclear localization signals	cubic.bioc.columbia.edu/predictNLS
SignalP	Cleavage sites & signal/non-signal peptide prediction	www.cbs.dtu.dk/services/SignalP
PSORT	Protein sorting signals	psort.nibb.ac.jp
Sulfinator	Tyrosine sulfation motifs	us.expasy.org/tools/sulfinator
GlobPlot	Protein disorder and globularity	globplot.embl.de
DisEMBL	Protein disorder	dis.embl.de
GO	biological function, component and process	www.geneontology.org
Ensemble	Genome browsing	www.ensembl.org
Phospho.ELM	Instances of Ser/Thr/Tyr phosphorylation	phospho.elm.eu.org
Perl	Script oriented language widely used in bioinformatics	www.perl.com
Python	Highly object oriented language designed for large projects	www.python.org
HMMER	Hidden Markov Model software suite	hmmer.wustl.edu
Biopython	Bioinformatics modules for perl	www.biopython.org
Bioperl	Bioinformatics modules for Python	www.bioperl.org

not to be phosphorylated, kinase sites can all be ignored; but if it is known to be phosphorylated, then the kinase-site matches can be targeted for experimental testing. Mass spectrometry can be a useful tool for revealing post-translational modifications. ELM can provide synergism with appropriate experiments and can help in mapping out a research program. In this way, the ELM resource should become increasingly useful to the research community

5 URLs

In Table 2 we have listed some URLs we thought might be useful for you to explore. Many more links can be found in the annual database (http://nar.oupjournals.org/content/vol31/issue1/) and webserver (http://nar.oupjournals.org/content/vol31/issue13/) open access issues of *Nucleic Acids Research*.

6 Conclusions

We hope that you now have a pretty clear idea of how to approach the analysis of your favorite modular protein. In this chapter we have not discussed all available resources for analysis of proteins, we apologize to the authors of these resources.

The mapping of globular domains should be considered mature – methods such as Pfam and SMART are highly reliable for determining potential domains in a sequence. The prediction of functional sites is a much younger field, although advancements have been made with nonsequence approaches such as the KBDS system in ELM.

The paradigm behind this chapter and the modular model of protein function is that sequence determines structure, which again determines function. This is clearly true in many instances; however, like any dogma, it is ultimately wrong and misleading. Our view of protein function is still very primitive. We expect the modular model to be enveloped by a more holistic model.

It does indeed seem as if nature is presenting molecular functions in two modes: structured domains that are folded and in which the fold/structure determines the function of the domain/protein, and an unstructured mode like the one we see for ELMs. These are modular units which seem to behave like autonomous bit/information strings carried within the host protein to accommodate certain functions or tuning of the host structure – they are themselves unstructured and only their sequence determines their function.

Acknowledgements

This work was partly supported by EU grant QLRI-CT-2000-00127. Thanks to Kresten Lindorff-Larsen, Sophie Chabanis-Davidson, Sara Quirk, and Francesca

Diella for commenting on this chapter. Thanks to Pål Puntervoll and Manuela Helmer Citterich for figures. Finally, we are deeply grateful to FreeBSD.org, (bio)Python.org, PostgreSQL.org, Debian.org, Gentoo.org, and Apache.org for fantastic open free software.

References

1 Aasland, R., et al., Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Lett.* 2002, **513**, 141–144.

2 Aloy, P., Russell, R., The third dimension for protein interactions and complexes. *Trends Biochem. Sci.* 2002, **27**, 633–638.

3 Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997, **25**, 3389–3402.

4 Andrade, M., *Bioinformatics and Genomes Current Perspectives*. Horizon Scientific Press 2003.

5 Andrade, M., Ponting, C., Gibson, T., Bork, P., Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.* 2000, **298**, 521–537.

6 Apic, G., Gough, J., Teichmann, S., An insight into domain combinations. *Bioinformatics* 2001, **17** Suppl 1, S83–89.

7 Aritomi, M., Kunishima, N., Inohara, N., Ishibashi, Y., Ohta, S., Morikawa, K., Crystal structure of rat Bcl-xL: implications for the function of the Bcl-2 protein family. *J. Biol. Chem.* 1997, **272**, 27886–27892.

8 Ashburner, M., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 2000, **25**, 25–29.

9 Bader, G., Betel, D., Hogue, C., BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 2003, **31**, 248–250.

10 Bairoch, A., Apweiler, R., The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 2000, **28**, 45–48.

11 Bateman, A., et al., The Pfam protein families database. *Nucleic Acids Res.* 2002, **30**, 276–280.

12 Bates, G., Huntingtin aggregation and toxicity in Huntington's disease. *Lancet* 2003, **361**, 1642–1644.

13 Betts, M., Guigo, R., Agarwal, P., Russell, R., Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? *EMBO J.* 2001, **20**, 5354–5360.

14 Brenner, S., Target selection for structural genomics. *Nat. Struct. Biol.* 2000, **7** Suppl, 967–969.

15 Brooks, B., Karplus, M., Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. USA* 1985, **82**, 4995–4999.

16 Chandonia, J., Walker, N., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S., ASTRAL compendium enhancements. *Nucleic Acids Res.* 2002, **30**, 260–263.

17 Clamp, M., et al., Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* 2003, **31**, 38–42.

18 Copley, R., Doerks, T., Letunic, I., Bork, P., Protein domain analysis in the era of complete genomes. *FEBS Lett.* 2002, **513**, 129–134.

19 Creighton, T., *Proteins Structures and Molecular Properties*, 2nd edit. Freeman, New York 1993.

20 Cuff, J., Barton, G., Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000, **40**, 502–511.

21 Cuff, J., Clamp, M., Siddiqui, A., Finlay, M., Barton, G., JPred: a consensus secondary structure prediction server. *Bioinformatics* 1998, **14**, 892–893.

22 Dedmon, M., Patel, C., Young, G., Pielak, G., FlgM gains structure in living

cells. *Proc. Natl. Acad. Sci. USA* 2002, **99**, 12681–12684.

23 Demarest, S., Martinez-Yamout, M., Chung, J., Chen, H., Xu, W., Dyson, H., Evans, R., Wright, P., Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* 2002, **415**, 549–553.

24 Dobson, C., Protein misfolding and human disease. *Scientific World Journal* 2002, **2** (1 Suppl 2), 132.

25 Doerks, T., Bairoch, A., Bork, P., Protein annotation: detective work for function prediction. *Trends Genet* 1998, **14**, 248–250.

26 Doerks, T., Copley, R., Schultz, J., Ponting, C., Bork, P., Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res.* 2002, **12**, 47–56.

27 Dunker, A., Brown, C., Lawson, J., Iakoucheva, L., Obradovic, Z., Intrinsic disorder and protein function. *Biochemistry* 2002, **41**, 6573–6582.

28 Dunker, A., et al., Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.* 1998, 473–484.

29 Dunker, A., et al., Intrinsically disordered protein. *J. Mol. Graph. Model* 2001, **19**, 26–59.

30 Dyson, H., Wright, P., Equilibrium NMR studies of unfolded and partially folded proteins. *Nat. Struct. Biol.* 1998, **5** Suppl, 499–503.

31 Evans, P., Owen, D., Endocytosis and vesicle trafficking. *Curr. Opin. Struct. Biol.* 2002, **12**, 814–821.

32 Garner, E., Cannon, P., Romero, P., Obradovic, Z., Dunker, A., Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. *Genome Inform Ser Workshop Genome Inform* 1998, **9**, 201–213.

33 Garner, E., Romero, P., Dunker, A., Brown, C., Obradovic, Z., Predicting binding regions within disordered proteins. *Genome Inform. Ser. Workshop Genome Inform.* 1999, **10**, 41–50.

34 Ginalski, K., Elofsson, A., Fischer, D., Rychlewski, L., 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 2003, **19**, 1015–1018.

35 Gough, J., The SUPERFAMILY database in structural genomics. *Acta Crystallogr. D Biol. Crystallogr.* 2002, **58**, 1897–1900.

36 Gunasekaran, K., Tsai, C., Kumar, S., Zanuy, D., Nussinov, R., Extended disordered proteins: targeting function with less scaffold. *Trends Biochem. Sci.* 2003, **28**, 81–85.

37 Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J., Orengo, C., Recognizing the fold of a protein structure. *Bioinformatics* 2003, **19**, 1748–1759.

38 Henikoff, J., Greene, E., Pietrokovski, S., Henikoff, S., Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* 2000, **28**, 228–230.

39 Hill, D., Blake, J., Richardson, J., Ringwald, M., Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res.* 2002, **12**, 1982–1991.

40 Hill, E., Broadbent, I., Chothia, C., Pettitt, J., Cadherin superfamily proteins in *Caenorhabditis elegans* and *Drosophila melanogaster*. *J. Mol. Biol.* 2001, **305**, 1011–1024.

41 Jensen, L., Ussery, D., Brunak, S., Functionality of system components: conservation of protein function in protein feature space. *Genome Res.* 2003, **13**, 2444–2449.

42 Jensen, L. J., et al., Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.* 2002, **319**, 1257–1265.

43 Jonassen, I., Eidhammer, I., Taylor, W., Discovery of local packing motifs in protein structures. *Proteins* 1999, **34**, 206–219.

44 Kabsch, W., Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, **22**, 2577–2637.

45 Kaplan, B., Ratner, V., Haas, E., Alpha-synuclein: its biological function

and role in neurodegenerative diseases. *J. Mol. Neurosci.* 2003, **20**, 83–92.

46 Klein-Seetharaman, J., et al., Long-range interactions within a nonnative protein. *Science* 2002, **295**, 1719–1722.

47 Kussie, P., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A., Pavletich, N., Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* 1996, **274**, 948–953.

48 Lander, E., et al., Initial sequencing and analysis of the human genome. *Nature* 2001, **409**, 860–921.

49 Letunic, I., et al., Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 2002, **30**, 242–244.

50 Li, X., Obradovic, Z., Brown, C., Garner, E., Dunker, A., Comparing predictors of disordered protein. *Genome Inform Ser Workshop Genome Inform* 2000, **11**, 172–184.

51 Linding, R., Jensen, L., Diella, F., Bork, P., Gibson, T., Russell, R., Protein disorder prediction: implications for structural proteomics. *Structure* 2003, **11**, 1453–1459.

52 Linding, R., Russell, R., Neduva, V., Gibson, T., GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 2003, **31**, 3701–3708.

53 Liu, J., Tan, H., Rost, B., Loopy proteins appear conserved in evolution. *J. Mol. Biol.* 2002, **322**, 53–64.

54 Lo Conte, L., Brenner, S., Hubbard, T., Chothia, C., Murzin, A., SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* 2002, **30**, 264–267.

55 Lopez De La Paz, M., Goldie, K., Zurdo, J., Lacroix, E., Dobson, C., Hoenger, A., Serrano, L., De novo designed peptide-based amyloid fibrils. *Proc. Natl. Acad. Sci. USA* 2002, **99**, 16052–16057.

56 Lopez Garcia, F., Zahn, R., Riek, R., Wuthrich, K., NMR structure of the bovine prion protein. *Proc. Natl. Acad. Sci. USA* 2000, **97**, 8334–8339.

57 Marchler-Bauer, A., et al., CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* 2003, **31**, 383–387.

58 Marsden, R., McGuffin, L., Jones, D., Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.* 2002, **11**, 2814–2824.

59 Melamud, E., Moult, J., Evaluation of disorder predictions in CASP5. *Proteins* 2003, **53** Suppl 6, 561–565.

60 Mott, R., Accurate formula for P-values of gapped local sequence and profile alignments. *J. Mol. Biol.* 2000, **300**, 649–659.

61 Mulder, N., et al., The InterPro Database, 2003, brings increased coverage and new features. *Nucleic Acids Res.* 2003, **31**, 315–318.

62 Nair, R., Rost, B., LOC3D: annotate sub-cellular localization for protein structures. *Nucleic Acids Res.* 2003, **31**, 3337–3340.

63 Nilsson, M., Dobson, C., In vitro characterization of lactoferrin aggregation and amyloid formation. *Biochemistry* 2003, **42**, 375–382.

64 Obenauer, J., Cantley, L., Yaffe, M., Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* 2003, **31**, 3635–3641.

65 Orengo, C., Pearl, F., Thornton, J., The CATH domain structure database. *Methods Biochem Anal* 2003, **44**, 249–271.

66 Pappu, R., Srinivasan, R., Rose, G., The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl. Acad. Sci. USA* 2000, **97**, 12565–12570.

67 Pei, J., Grishin, N., AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 2001, **17**, 700–712.

68 Pollastri, G., Przybylski, D., Rost, B., Baldi, P., Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* 2002, **47**, 228–235.

69 Promponas, V., Enright, A., Tsoka, S., Kreil, D., Leroy, C., Hamodrakas, S., Sander, C., Ouzounis, C., CAST: an iterative algorithm for the complexity analysis of sequence tracts: complexity

analysis of sequence tracts. *Bioinformatics* 2000, **16**, 915–922.

70 Pruitt, K., Maglott, D., RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* 2001, **29**, 137–140.

71 Puntervoll, P., et al., ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* 2003, **31**, 3625–3630.

72 Radhakrishnan, I., Perez-Alvarado, G., Parker, D., Dyson, H., Montminy, M., Wright, P., Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. *Cell* 1997, **91**, 741–752.

73 Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J., Dunker, A., Identifying disordered proteins from amino acid sequences. *Proc. IEEE Int. Conf. Neural Networks* 1997, **1**, 90–95.

74 Russell, R., Ponting, C., Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* 1998, **8**, 364–371.

75 Saqi, M., Sternberg, M., Identification of sequence motifs from a set of proteins with related function. *Protein Eng.* 1994, **7**, 165–171.

76 Schultz, J., Milpetz, F., Bork, P., Ponting, C., SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. USA* 1998, **95**, 5857–5864.

77 Schwede, T., Kopp, J., Guex, N., Peitsch, M., SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* 2003, **31**, 3381–3385.

78 Schweers, O., Schonbrunn-Hanebeck, E., Marx, A., Mandelkow, E., Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta structure. *J. Biol. Chem.* 1994, **269**, 24290–24297.

79 Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., Kahn, D., ProDom: automated clustering of homologous domains. *Brief Bioinform* 2002, **3**, 246–251.

80 Sigrist, C., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P., PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform.* 2002, **3**, 265–274.

81 Smith, D., Radivojac, P., Obradovic, Z., Dunker, A., Zhu, G., Improved amino acid flexibility parameters. *Protein Sci.* 2003, **12**, 1060–1072.

82 Smyth, E., Syme, C., Blanch, E., Hecht, L., Vasak, M., Barron, L., Solution structure of native proteins with irregular folds from Raman optical activity. *Biopolymers* 2001, **58**, 138–151.

83 Stefani, M., Dobson, C., Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.* 2003, **81**, 678–699.

84 Suyama, M., Ohara, O., DomCut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics* 2003, **19**, 673–674.

85 Taylor, W., Protein structural domain identification. *Protein Eng.* 1999, **12**, 203–216.

86 Taylor, W., A deeply knotted protein structure and how it might fold. *Nature* 2000, **406**, 916–919.

87 Taylor, W., Defining linear segments in protein structure. *J. Mol. Biol.* 2001, **310**, 1135–1150.

88 Taylor, W., Lin, K., Protein knots: a tangled problem. *Nature* 2003, **421**, 25.

89 Teilum, K., Kragelund, B., Poulsen, F., Transient structure formation in unfolded acylcoenzyme A–binding protein observed by site-directed spin labeling. *J. Mol. Biol.* 2002, **324**, 349–357.

90 Tompa, P., Intrinsically unstructured proteins. *Trends Biochem. Sci.* 2002, **27**, 527–533.

91 Uversky, V., Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 2002, **11**, 739–756.

92 Verkhivker, G., Bouzida, D., Gehlhaar, D., Rejto, P., Freer, S., Rose, P., Simulating disorder–order transitions in molecular recognition of unstructured proteins: where folding meets binding. *Proc. Natl. Acad. Sci. USA* 2003, **100**, 5148–5153.

93 Vihinen, M., Torkkila, E., Riikonen, P., Accuracy of protein flexibility predictions. *Proteins* 1994, **19**, 141–149.

94 von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B., STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 2003, **31**, 258–261.

95 Waterston, R., et al., Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002, **420**, 520–562.

96 Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H., The Protein Data Bank and structural genomics. *Nucleic Acids Res.* 2003, **31**, 489–491.

97 Wheeler, D., et al., Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.* 2002, **30**, 13–16.

98 Wootton, J., Nonglobular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 1994, **18**, 269–285.

99 Wright, P., Dyson, H., Intrinsically unstructured proteins: reassessing the protein structure–function paradigm. *J. Mol. Biol.* 1999, **293**, 321–331.

100 Xenarios, I., Salwinski, L., Duan, X., Higney, P., Kim, S., Eisenberg, D., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 2002, **30**, 303–305.

101 Yaffe, M., Leparc, G., Lai, J., Obata, T., Volinia, S., Cantley, L., A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.* 2001, **19**, 348–353.

102 Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G., MINT: a molecular interaction database. *FEBS Lett.* 2002, **513**, 135–140.

103 Zdobnov, E., et al., Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* 2002, **298**, 149–159.

104 Zoete, V., Michielin, O., Karplus, M., Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J. Mol. Biol.* 2002, **315**, 21–52.