A Multi-Volume Comprehensive Treatise

Biotechnology

Second, Completely Revised Edition

Edited by
H.-J. Rehm and G. Reed
in cooperation with
A. Pühler and P. Stadler

Volume 5a

Recombinant Proteins, Monoclonal Antibodies and Therapeutic Genes

Edited by
A. Mountain, U. Ney and D. Schomburg

WILEY-VCH

Weinheim · New York · Chichester · Brisbane · Singapore · Toronto

2 Sequence and Structure of Proteins

FRANK EISENHABER
PEER BORK

Heidelberg and Berlin, Germany

List of Abbreviations

ASC	computer program for the analysis of surface properties of proteins
BLAST	tool for protein or DNA sequence homology search leading to optimal local alignment
BLIP	inhibitor of TEM-1β-lactamase
BLITZ	tool for finding homologous sequences in databases
BLOCKS	motif database
BRCA1	breast cancer gene
CAP	BLAST output parser
CATH	classification of protein structures
CD	circular dichroism
COMBI	protein secondary structure prediction method
DEFINE	computer program for objectivation of secondary structure assignment
DSSP	computer program for objectivation of secondary structure assignment
ε-helix	extended polypeptide conformation outside β-sheet
EST	expressed sequence tag
FASTA	tool for finding homologous sequences in databases
FLyBase	molecular database for *Drosophila*
FSSP	classification of protein structures
FTP	file transfer protocol
GENEQUIZ	tool for functional assignment of proteins by sequence homology
GOR/GORIII	secondary structure prediction methods
HMG-box	high mobility group (HMG) domain in several families of nuclear proteins
HMM	Hidden Markov Model
MoST	motif search tool
NP-complete	problem which cannot be solved with a polynominal algorithm (*non-p*olynominal complete problem)
NSC	computer program for the analysis of surface properties of proteins
ORF	open reading frame
PATSCAN	WWW server for motif and profile searchers
PBE	Poisson-Boltzmann differential equation
P-CURVES	computer program for objectivation of secondary structure assignment
PDB	Brookhaven Protein Data Bank
PHD	engine for secondary structure prediction
PIMA	motif database
PPI-helix	extended conformation of *cis*-polyproline
PPII-helix	extended conformation of *trans*-polyproline
PREDATOR	engine for secondary structure prediction
PRINTS	motif database
PRODOM	motif database
PROFILE	WWW server for motif and profile searchers
PROSITE	motif database
r.m.s.d.	root mean square deviation
SCOP	classification of protein structures
SCR	structurally conserved regions
SOM	tool for pattern search
SOPM	program for secondary structure prediction
SRS	sequence retrieval system, query tool for biomolecular databases
SSCP	WWW program for secondary structure content prediction
SSPRED	engine for secondary structure prediction
STRIDE	modification of DSSP
SVR	structurally variable regions

SWISS-PROT	major annotated protein sequence database
TMAP	program for prediction of transmembrane regions
TREMBL	protein sequences obtained from translation of nucleotide sequences in the EMBL database
TTOP	program for prediction of topology of transmembrane proteins
URL	unique resource location (in the WWW)
UV	ultraviolet
WWW	world wide web
Yeast YPD	molecular database for yeast

1 Introduction

For biotechnological applications, it is very important to determine which type of specific structural information is necessary to understand or to modify the protein function in the desired manner. In this review, we will give an overview on protein structural features classified with increasing complexity and consider appropriate methodological approaches and theoretical concepts for their investigation. Special emphasis will be given to techniques applicable to predicting protein structural properties by relying only on the protein sequence (EISENHABER et al., 1995b). It is expected that the reader is familiar with basic knowledge on protein biochemistry and biophysics as given in standard university textbooks.

2 Hierarchical Description of Protein Structure

Whereas nucleic acids fulfill mainly the tasks of storage and transfer of genetic information in living organisms, the proteins form a complicated cellular machinery for realization of this genetic program dependent on and in response to changing environmental conditions. Some proteins are catalysts and enable chemical reactions which would otherwise not occur at temperatures (e.g., 37°C) and pH (e.g., pH 7) values typical for living organisms. Others are involved in storage and transport of particles ranging from electrons to macromolecules. Proteins mediate signal transmission between cells, tissues, and organisms; they control the passage of molecules through membranes surrounding cellular compartments. Yet other proteins are involved in the mechanochemistry of motion as in muscles or serve a structural purpose in the filamentous architecture of cells and tissues.

The functionality of different protein molecules is tightly connected with their structural and dynamic properties. Amino acid sequences for linear polypeptides forming protein molecules are directly encoded in genes. At the same time, the three-dimensional protein architecture represents the ultimate in molecular information, and from it springs a variety of significant scientific results: the understanding of protein folding and structural stability, interactions between subunits, receptors, ligands, substrates, and the like, enzymatic catalysis, the understanding of molecular evolution, the ability to engineer and design proteins through synthesis and mutation, the creation of drugs and the utilization of protein-based processes to confront human disease and suffering. Inspite of many years of intensive research, the complete description of structure and dynamics of a protein molecule based on fundamental physical and chemical principles is still an unsolved scientific task. Therefore, theoretical protein science must settle for lesser goals as well as for the quest, which is described in this article.

The main vehicle to organize the current knowledge of protein structure is the so-called hierarchical description (Fig. 1). A protein molecule in solution is a very complex system with a huge number of degrees of freedom, albeit many of them are of little importance for biological function. Traditionally, primary, secondary, tertiary, and quarternary structural levels are considered which correspond to biochemical events of coding, synthesis, and function of proteins as well as to physico-chemical properties of isolated polypeptides in solution. Supersecondary structure, protein fold, topology, and structural domain are historically younger terms. Short definitions are given below:

1. The lowest level of structural organization, the *primary structure,* is identical with the amino acid sequence. The order of amino acids as genetically encoded may be changed posttranscriptionally through splicing, during translation as a result of recoding mechanisms, and posttranslationally due to chemical modifications (backbone scission, side chain modifications). Generally, the chemical identity of a residue is of critical importance only for a few sequence positions. At most sequence positions, many amino acid exchanges have

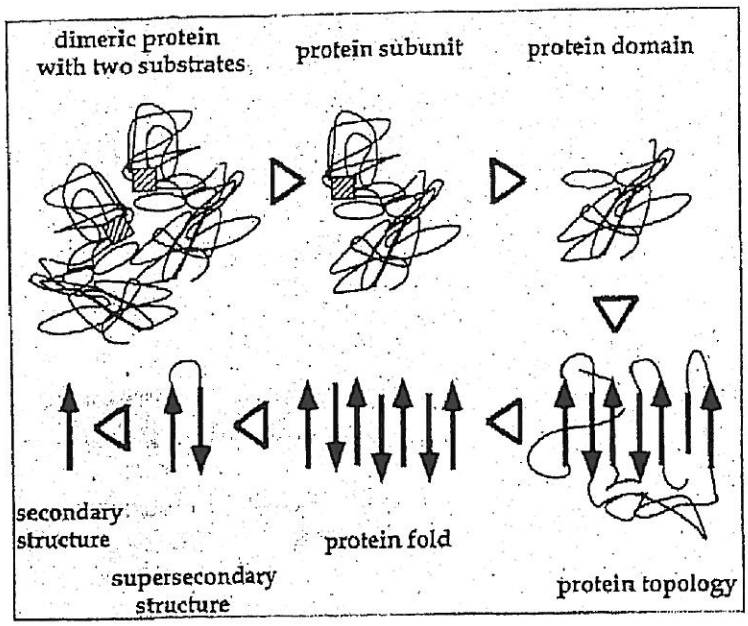

Fig. 1. Hierarchy of structural description of proteins. The flowchart presents the structural analysis of a dimeric β-sheet protein (at the upper left) with 2 ligands (squares) up to the level of secondary structural elements (β-strands represented as arrows, at the lower left). Successively, subunit, domain, topology, fold, and supersecondary structure are illustrated.

only a slight effect on three-dimensional structure and function.

2. The terms of *secondary structure* are used to describe preferred relative backbone locations of sequentially near residues mainly due to local interactions.
3. *Supersecondary structural motifs* are typical ways of packing between secondary structural units. A *protein fold* is commonly defined as the scaffold of secondary structural elements with repetitive backbone structure; i.e., α-helices and β-sheets. The term *"protein topology"* is more general and includes the spatial arrangement of all secondary structural elements including loop segments. The *tertiary structure* of a protein is described with the relative position of atoms in all residues of a polypeptide chain, both in regular secondary structural elements as well as in loops connecting them. The protein tertiary structure may also be represented in terms of pairwise distances between atoms of various residues. A particular feature of globular proteins is the existence of contacts between residues with large sequence separation (non-local interactions). As a rule, a tertiary structure has a typical densely packed hydrophobic core shielded from interaction with a solvent. A tertiary structure can comprise several *domains* which are distinguished structurally due to autonomous hydrophobic cores and, possibly thermodynamically as melting independently. A domain is often considered an autonomous folding unit. At the same time, a structural domain does not need to be contiguous in amino acid sequence.
4. The *quarternary structure* is composed of several subunits, each being a polypeptide with its own tertiary structure. The situation is similar as with domains but the chemical (peptide) link between subunits is missing.

Protein structural features are greatly influenced not only by the type of the cellular compartment or the solution conditions (e.g., temperature, water content, ionic strength, inclusion into membranes of organelles or cells, extracellular space), but also by binding of cofactors and other ligands.

3 Primary Structure of Proteins

The primary structural level, the lowest in the hierarchical description of protein structure, corresponds essentially to the chemical structure. After consideration of the known chemical possibilities of protein diversity, we

discuss modern aspects of protein sequence analysis.

3.1 Chemical Structure of Proteins

Proteins are biomacromolecules. Their main constituents are linear polypeptide chains. The monomeric units are α-L-amino acids or -imino acids, interconnected by peptide bonds. The succession of monomer types in the linear polypeptide is called the protein sequence. How do proteins achieve the impressing functional diversity?

1. No other biomacromolecule has such a variety of monomer types. In addition to the 20 traditional amino acid types with different side chains (plus selenocysteine, see Sect. 3.2), monomers already included into the polypeptide chain may also be chemically modified after translation (posttranslational modification). Both amino acid composition and the order of amino acid types in the sequence characterize an individual protein.
2. Proteins can contain small organic (prosthetic groups) or inorganic compounds as an integral part of their structure. For example, the type of cofactor can influence the reaction specificity of an enzyme.
3. The main source of functional diversity of proteins is the variety of three-dimensional domain structures each of which is preferred by a certain class of amino acid sequences. A crude classification distinguishes between globular and fibrillar domains, the latter being mainly in structural molecules.
4. Proteins vary widely in size. The range in the number of monomers is from a few dozen amino acids for small proteins like crambin up to the enormous value of 27,000 residues in the muscle protein titin. An eukaryotic domain consists typically of ~125 amino acid residues (~150 for prokaryota) as seen from sequence length distributions and the occurrence of leading methionines (BERMAN et al., 1994; KOLKER and TRIFONOV, 1995). Larger proteins are normally composed of many domains (BORK et al., 1996).

The experimental techniques for determining the sizes and the amino acid sequences of proteins have reached a very high level of maturity. As a result of standardization and automation, the productivity in genomic and protein sequencing has steadily increased. The size of sequence databases exploded during the last years. The number of entries in TREMBL (protein sequences obtained from translation of nucleotide sequences in the EMBL database, ftp://embl -ebi.ac.uk/pub/databases/trembl/) is larger than 200,000 and SWISS-PROT (http://expasy.hcuge.ch/) contains more than 60,000 annotated sequences. Even whole genomes become available (FLEISCHMANN et al., 1995).

3.2 Protein Primary Structure as a Result of Transcription and Translation of Genetic Information

The relation of genetic information and protein primary structure is very complex. Primarily, the succession of residue types in the polypeptide is genetically encoded in the triplet sequence of genes. The transcriptional events which may include synthesis of a transcript from one or several DNA segments and splicing (removal of intron segments and ligation of extrons) result in a messenger RNA (mRNA). Depending on cellular conditions, RNA splicing may even follow alternative pathways ("alternative splicing"). In the ribosomal machinery, this triplet code is translated into a polypeptide sequence. In dependence on the organism and the organelle type, different standard genetic translation tables apply. Additionally, probably in all organisms, a minority of genes relies on recoding of the canonical genetic table (RNA editing) for translation of their mRNAs (BÖRNER and PÄÄBO, 1996; GESTELAND and ATKINS, 1996; STADTMAN, 1996):

- Frameshifting at a particular site allows the expression of a protein from overlapping reading frames (possibly with several translation products in the case of non-100% frameshift efficiency).

- The meaning of code triplets may be altered. Stop codons can be redirected as tryptophane, glutamine, or even selenocysteine, the 21st translationally incorporated type of amino acid. The existence of UGA translation into a cysteine-type amino acid has already been predicted from symmetry considerations of the genetic code (SHCHERBAK, 1988, 1989). Also the editing of glycine codons into asparagine has been observed.
- Ribosomes may translate over coding gaps in mRNAs if the stop codon is hidden in some mRNA secondary structure (translational bypassing). If a stop codon is missing, the ribosome may use a piece of ribosomal RNA for completing the polypeptide (transtranslation).

RNA editing is one aspect of a more general phenomenon, namely the continuing evolution of the genetic code (OSAWA et al., 1992).

Finally, the polypeptide is subject to posttranslational modifications, necessary for a wide variety of reasons, e.g., protection against proteolysis, direction of transport, genetic regulation, membrane anchoring, and regulation of enzymatic activation or of degradation (CREIGHTON, 1992; HAN and MARTINAGE, 1992; RESH, 1994). In the simplest case, single amino acids are chemically altered. The modifications can be N-terminal (e.g., acetylation, myristoylation and pyroglutaminylation), C-terminal (e.g., amidation, isoprenylation and farnesylation), or affect the side chains (e.g., glycosylation, phosphorylation, hydroxylation, and disulphide bond formation). Another type of posttranslational chemical modification consists in cuts of peptide bonds with or without dissociation of the two resulting chains. The latter is the activation mechanism of trypsinogen and chymotrypsinogen. In the case of α-lytic protease, a segment of the precursor polypeptide chain acts as catalyst for achieving the native fold (BAKER et al., 1992a). Polypeptide chain scission may occur during functioning of a protein. For example, serpins are efficient inhibitors of plasma proteases. In the unbound state, serpins are in a metastable kinetically trapped state with a 5-stranded β-sheet and a largely unstructured reactive loop (MOTTONEN et al., 1992). Upon tight binding with serine proteases, the conformation rearranges to the 6-stranded latent form. During dissociation of the serpin molecules from the complex, they are slowly cleaved and inactivated as inhibitors (GOLDSMITH and MOTTONEN, 1994; HUANG et al., 1994; CARREL et al., 1994).

Multiple polypeptide chain scissions occur during the so-called protein splicing, the most recently discovered variant of posttranslational modification (COOPER and STEVENS, 1995). It involves the precise and autocatalytic excision of one or several intervening protein sequences from a precursor protein, coupled to the ligation (peptide bond formation) of the remaining sequence domains, which results in a spliced protein product. In full analogy to RNA splicing, the two types of segments of the precursor protein are named exteins (expressed part) and inteins (intervening part). In several cases, it was difficult to isolate the full-length precursor since inteins can actually splice while the C-terminal extein is still being translated (COOPER and STEVENS, 1995). Intein assignment was an important step in the genome analysis of *Methanococcus jannaschii* (BULT et al., 1996).

The consideration of RNA recoding and posttranslational chemical modification in an automatic manner is currently impossible (or, at least, very unreliable), and this adds another moment of uncertainty to the correctness of amino acid sequences derived from nucleotide sequence data canonically translated by computer programs.

3.3 Investigation of Primary Structural Features – Protein Sequence Analysis

Whereas in the early times of molecular biology, protein research concentrated mainly on the physico-chemical and functional characterization of selected proteins and only a few protein sequences were available, the situation has now reversed. In a typical case, the researcher has the protein sequence (mainly derived from corresponding cDNA se-

quences), but does know only little about its structure and function. The 28th of July, 1995, marks the beginning of a new era. The complete genome of *Haemophilus influenzae*, a bacterium, has been obtained as a result of world wide cooperation in large-scale sequencing projects (FLEISCHMANN et al., 1995). Other genomes followed (several eubacteria and archebacteria as well as bakers' yeast) or will be available in a foreseeable future (for nematode *Caenorhabditis elegans*, it will be available in 1998 and, for the human genome, it is expected in 2003, see Fig. 2). The genetic make-up (the complete DNA of an organism) should contain all information necessary for cells to mature, to reproduce and to interact with the environment as open, homeostatic system until their preprogramed death. Thus, knowledge about virtually all proteins in a living organism is available, albeit, at the moment, the experimental facts are primarily limited to the nucleotide se-

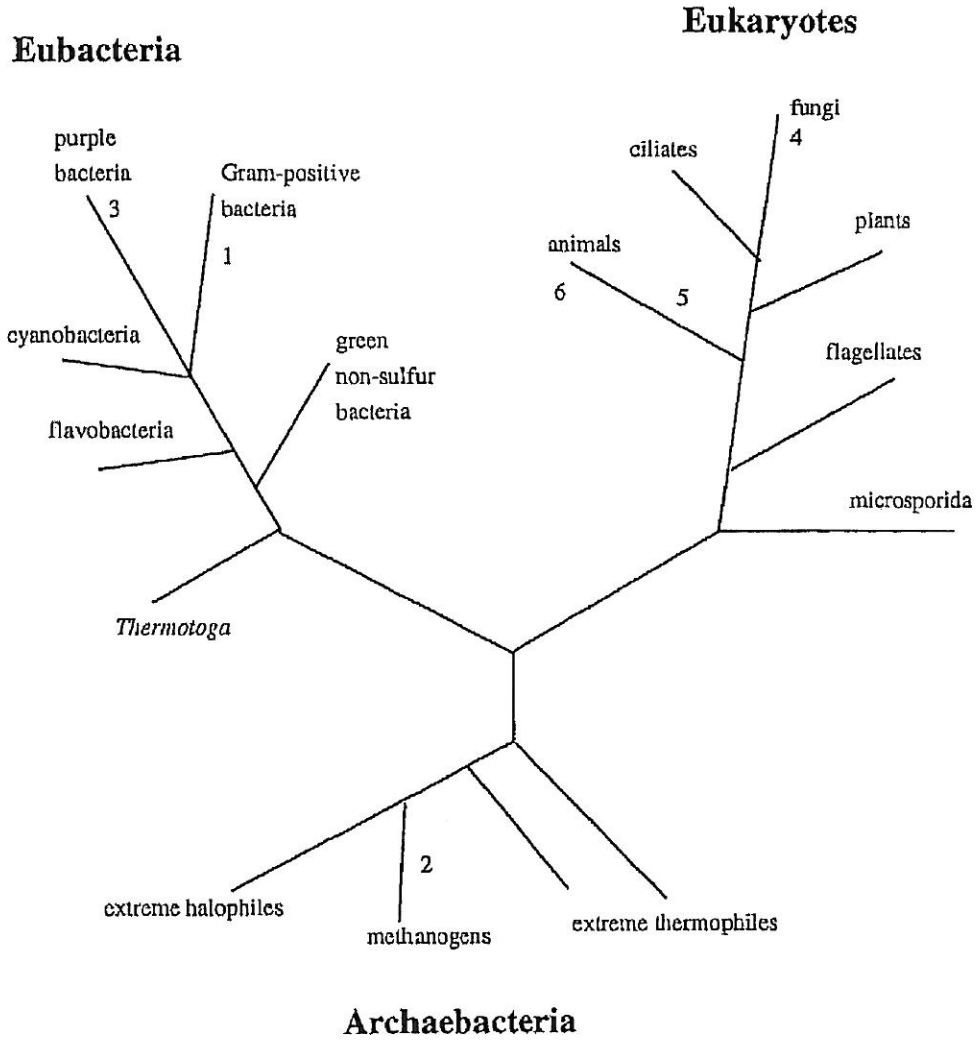

Fig. 2. Phylogeny of species and the size of genomes. A phylogenetic tree based on RNA data containing groups of organisms from all major phyla is shown.

Species	Genome Size [Mb]	Number of Genes (Proteins)
1 *M. genitalium*	0.6	470
2 *E. coli*	4.7	4000
3 *M. jannaschii*	1.7	1760
4 Yeast	13.5	6000
5 *C. elegans*	100	13000
6 Human	3000	70000

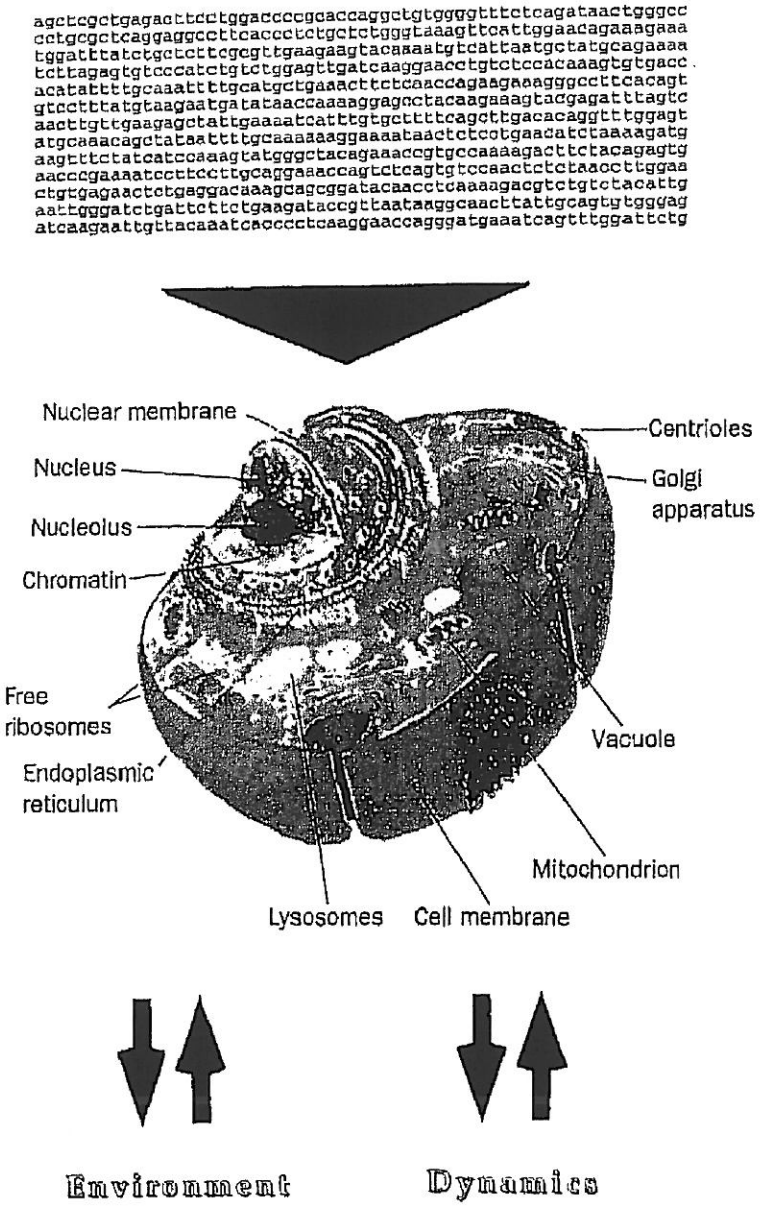

Fig. 3. Impact of the genetic make-up for a cell. In the upper part, a piece of the gene sequence of the breast cancer gene *BRCA*1 is presented (a few hundred bases). For comparison, the human genome comprises about $3 \cdot 10^9$ bases. The genetic information is sufficient to code for all cellular functions (lower part). It should be emphasized that living organisms composed of cells are not static devices but have different levels of ontogenetic development, exchange information, energy, and metabolic products with their environment, and finally react on impulses from outside.

quence of their genes (Fig. 3). The decoding of this enormous amount of data is an important goal and biological science is only starting to approach it, having in mind revolutionary applications in biotechnology, drug development, and in the treatment of diseases.

The main source of hypothetical information about such unknown proteins is amino acid sequence analysis and comparison (Fig. 4). This approach is knowledge-based and inductive. The query sequence or significant parts of it are compared with sequences or sequence motifs in databases. The annotated information (source organism, structural and functional data) of proteins or protein domains with similar sequences and the biochemical and molecular-biological context of the query protein are used to extrapolate possible structural and functional features (see, e.g., TATUSOV et al., 1996). To obtain finally scientific data, experimental verification of these conclusions need to follow.

Protein function (Tab. 1) requires also a multilevel description similar to protein structure (see Sect. 2). Primarily, a protein has a molecular function. It may catalyze a specific reaction or transmit a signal. A set of many co-operating proteins can fulfill a physiological function. In the simplest case, this is a metabolic pathway. At the next level, protein function determines phenotypic properties (phenotypic function or disfunction). Protein activity might be limited to certain cellular compartments, the extracellular space, and

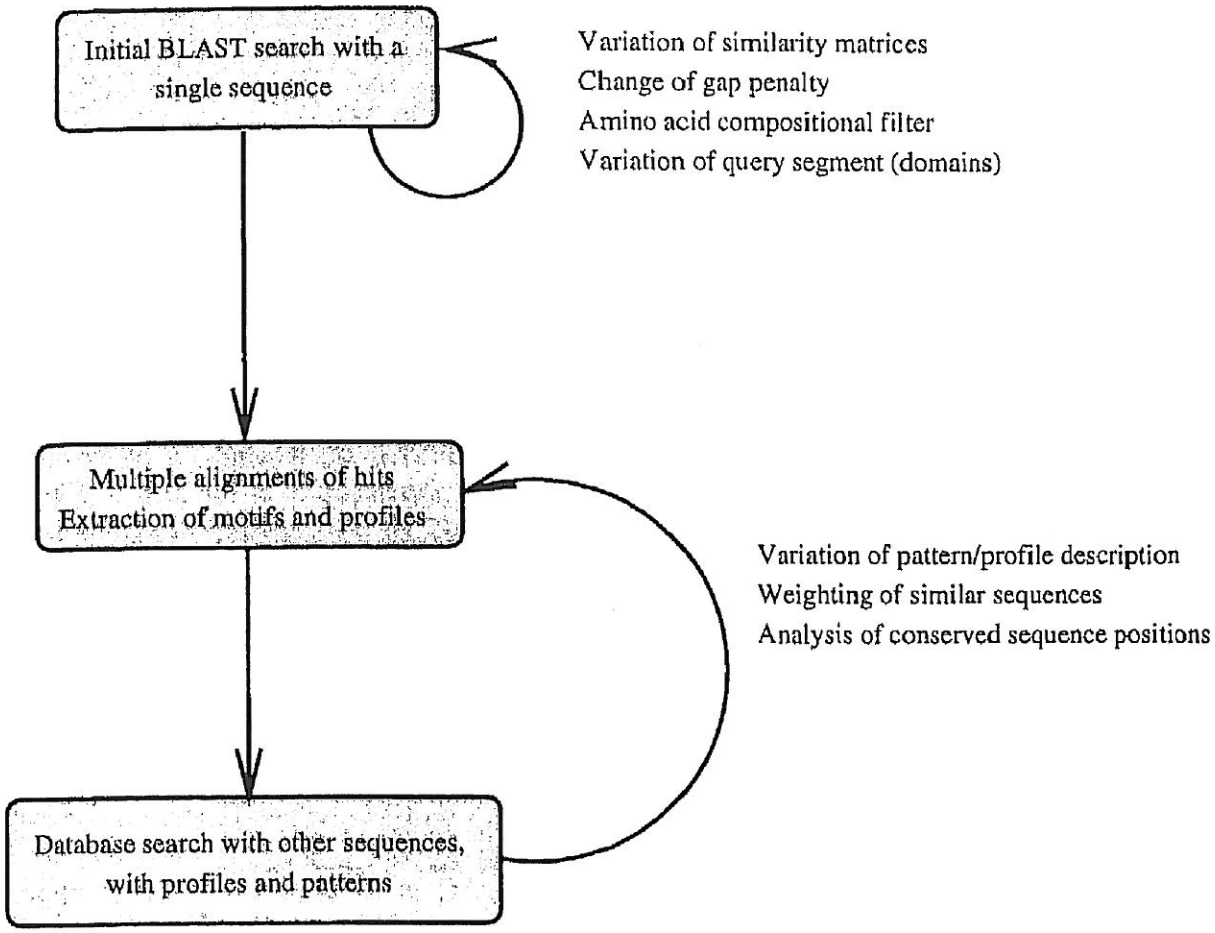

Fig. 4. Flowchart for iterative database searches. The sequence hits found at the initial searches are used for obtaining multiple alignments. They are analyzed and sequence motifs and/or profiles are extracted. With their help, a more sensitive database search is restarted.

types of cells (localization), to periods of the development (time restriction) or may depend on posttranslational modifications. Often, the corresponding genes are only expressed in a few phases of the ontogenesis and in some tissues (expression pattern).

In Sects. 3.3.1 and 3.3.2, we will consider both the issue of sequence databases and methods for knowledge-based sequence analysis. Protein property predictions relying on structural databases will be considered in Sect. 5.

3.3.1 Quality of Sequence Databases

The large amount of data on protein sequences calls for automation; also in sequence handling and analysis. Accurate storage and updating mechanisms as well as user-friendly retrieval software are required. Although database teams are aware of such demands and continuously improve the quality of data entries and computer software, databases have grown historically and are far from perfect. Thus, working with sequence databases requires knowledge about their pitfalls which can have a considerable influence on the interpretation of the data. We have already discussed in Sect. 3.2 possible modifications of translational decoding of mRNA sequences (organelle specific translation tables, RNA editing) and pre- (RNA splicing) and posttranslational changes, both resulting in a chemically altered polypeptide compared with the original DNA sequence. But there are also other, much more profane circumstances that result in uncertainty of sequence data. Errors can occur at each step of the multilevel procedure the result of which are protein sequences.

Tab. 1. What is Protein Function?

Molecular function	glycerol kinase contains an ATP-binding site
Physiological function	responsible for wing development nuclear transport
Phenotypic function	suppressor of SPT3 mutations involved in nucleotide metabolism
Dysfunction	deletion causes diabetis knockout lethal
Cellular localization	N-terminal myristilation transmembrane protein
Expression pattern	only in brain during embryogenesis activated by Gal4
Posttranslational modification	phosphotyrosine glycosylation

Many levels of protein function. Protein function is determined in the context of activity of other constituents of an organism. The primary molecular function of a protein may consist in catalyzing a specific reaction or in transmitting a signal. A set of many co-operating proteins accomplish a physiological function, e.g., a metabolic pathway. At the next level, proteins determine phenotype properties (phenotypic function or dysfunction). As a rule, protein activity is limited in space and time. The protein can be confined in certain cellular compartments, types of cells, or be located in the extracellular space. Posttranslational modifications often regulate protein activity. The genes coding proteins may be expressed in a few phases of the ontogenese and in some tissues only (expression pattern).

At the beginning, pieces of genomes are sequenced. Experimentally, the DNA sequences are obtained by reading-out electrophorese gels. Errors influence apparently 0.1% of all nucleotides (FLEISCHMANN et al., 1995) if DNA segments between 300 and 600 bases long are sequenced with 3–10 fold redundancy. This might affect 5–6% of all protein sequences (BIRNEY et al., 1996; TATUSOV et al., 1996). The rate is several orders of magnitude higher for so-called ESTs (expressed sequence tags), single gel reads of random cDNA. The comparison of 3,000 human proteins in SWISS-PROT that have been published more than once revealed differences in 0.3% of all amino acids (BORK and BAIROCH, 1996). This error rate seems to be a lower limit since sequences in different publications are often not independent, and many corrections have already been made. Anyway, mostly only erroneous stop codons and frameshift errors can be detected unambigiously. Single residue exchanges are hard to verify as strain differences or other types of natural polymorphism cannot be excluded. The rate appears small but the error may accumulate in the sequence considered. This can lead to functional misinterpretations (BORK, 1996). Another serious problem is the contamination of cDNA libraries with material from hosts (usually fungal, bacterial, or viral DNA). A prominent example is the "human" EST library of Genethon having a surprisingly high rate of matches with the yeast genome. Gene sequences of annexin I and insulin from a sponge (their "existence" in lower eukaryotes is very unlikely) were closely related to mammalian homologs (BORK and BAIROCH, 1996). It turned out that the biological sample had been contaminated by DNA of a rodent species.

As a next step, the processing of raw DNA data includes identification of genes, open reading frames (ORFs), and the exon–intron structure. In addition to sequence signals for exon–intron junctions, recent exon recognition algorithms integrate similarity searches in existing sequence databases since homology hits are a strong indicator for a coding sequence. Other properties used for exon–in-

tron classification are: (1) codon usage, (2) hexanucleotide frequencies, (3) local complexity (information content), (4) poly(A) ranges, regulatory sequences such as ribosome-binding segments (Shine-Delgarno segments) or promoters. The challenge is to improve the corresponding identification algorithms further since widely used programs for gene (exon) prediction in eukaryotes have an accuracy below 50% (BURSET and GUIGO, 1996). This rate drops further if the DNA sequences considered contain wrong nucleotide positions. The gene identification problem is avoided if mature mRNA is sequenced.

Finally, sequences are annotated by human beings or by their computer programs (BORK and BAIROCH, 1996). Hence, errors ranging from simple spelling ambiguities to semantic mistakes are common. For example, *SCD25*, the supressor gene of CDC25 in yeast, was so often misquoted as *SDC25* that it has become an accepted synonym. Database queries are hindered by differences between spelling variants [e.g., hemoglobin (US) and haemoglobin (UK), upper and lower case in *H* (Hairless) and *h* (hairy) in the *Drosophila* genetic nomenclature], representation of non-English characters (e.g., 'ü' in *Krüppel, Krueppel,* or *Kruppel*). The same gene may have several synonymes (e.g., *TUP*1, *AER*2, *SFL*2, *CYC*9, *UMR*7, *AAR*1, *AAM*1, and *FLK*1 are the same gene in yeast and *hns, bns*A, *drd*X, *osm*Z, *bgl*Y, *msy*A, *cur*, *pil*G, and *top*S are the same gene in *Escherichia coli*). Similar multiplicity exists on the level of protein names (e.g., annexin V was also called lipocortin V, endonexin II, calphobindin I, placental anticoagulant protein I, thromboplastin inhibitor, vascular anticoagulant α, and anchorin CII). The opposite is also frequent – different genes or proteins having the same name (e.g., cyclin is the name for a variety of cell cycle components or *MRF*1 is the gene name in mitochondria of yeast both for the peptide chain releasing factor 1 and for the respiratory function protein 1).

A major problem is the functional description of genes and proteins (BORK and BAIROCH, 1996). For example, the ORGanelle division in the EMBL, Genebank nomenclature should be used only for sequences in the genome of mitochondria and plastids. Often entries with nuclear-encoded genes for proteins targeted at organelles are wrongly entered in the same division. Since the number of sequences is large, functionalities are assigned automatically by the computer software based on sequence similarities. In this way, a single erroneous entry will lead to a whole sequence family with artificial functions [e.g., whether nifr3 has a role in nitrogen fixation, remains unclear, but it has already been assigned to several previously unannotated proteins (CASARI et al., 1994)].

Yet another layer of uncertainty consists of scope-related problems. Often, the retrieval system does not access all known sequence data but only a subset (due to license reasons or software limitations); sometimes even just a single sequence representative for a family. Many communities study particular classes of proteins and the information in such specialized databases is not pointed to from the general genetic databases but there is hope that links will appear in a near future (such as pointers between SWISS-PROT and the two databases FlyBase and yeast YPD within the SRS retrieval system, see Tab. 2, for WWW-links).

3.3.2 Knowledge-Based Prediction of Protein Structure and Function Using Protein Sequence Analysis

Generally, any property of the sequence representing an unknown protein can be compared with other sequences or families of sequences in databases. In cases of coincidence, the query protein is considered similar and the information on proteins which gave the hit is considered relevant also for the unknown protein. From the logical point of view, the discriminative power of this approach should be limited since it is not *a priori* clear whether a given sequence property is characteristic also for the sequence under consideration. But in practice, this way of thinking is unexpectedly often successful. The structure and function of an unknown protein can be tested experimentally if it is supposed to be related to a protein family with well-

Tab. 2. WWW Pointers for Important Programs and Databases for Similarity Searches in Protein Sequences are Presented
The list it not complete but the links here have been tested by the authors. A more detailed and updated list is available from http://www.embl-heidelberg.de/~bork/pattern.html

FTP Sites for Software Resources

Barton's flexible patterns	ftp://geoff.biop.ox.ac.uk
Propat (property pattern)	ftp://ftp.mdc.berlin.de/pub/makpat
SOM (neutral network)	ftp://ftp.mdc-berlin.de/pub/neural
SearchWise	http://www.ocms.ox.ac.uk/~birney/wise/topwise.html
PROFILE	ftp://ftp.ebi.ac.uk/pub/software/unix
MoST (motif search tool)	ftp://ncbi.nlm.nih.gov/pub/koonin/most
CAP (BLAST output parser)	ftp://ncbi.nlm.nih.gov/pub/koonin/cap

Searchable Motif Databases

PROSITE	http:/expasy.hcuge.ch/sprot/prosite.html
Motif search (ICR)	http://genome.ad.jp/SIT/MOTIF.html
Profile Scan (ISREC)	http://ulrec3.unil.ch/software/PFSCAN form.html
BLOCKS	http://www.blocks.fhcrc.org
PRINTS	http://www.biochem.ucl.ac.uk/~attwood/PRINTS/PRINTS.html
PIMA	http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html
PRODOM	http://protein.toulouse.inra.fr/prodom.html

WWW Servers for Motif and Profile Searches

Regular expressions	http://ibc.wustl.edu/fpat/
PROFILE	http://sgbcd.weizmann.ac.il/Bic/ExecAppl.html
PATSCAN	hppt://www.mcs.anl.gov/home/papka/ROSS/patscam.html
PatternFind (ISREC)	http://ulrec3.unil.ch/software/PATFND_mailform.html
Pmotif	http://alces.med.umn.edu/pmotif.html
HMM	http://genome.wustl.edu/eddy/hmm.html
Discover	http://hertz.njit.edu/~jason/help.html

studied members. At present, sequence analysis alone cannot give final knowledge.

1st, the meaning of terms denoting sequence properties needs to be clarified. Often, similarities are more obvious if only a relevant part of the sequence information is used for comparison.

- *Sequence composition* is traditionally the proportion of amino acid residues in the sequence. About 40% of all sequences in the SWISS-PROT database have a pronounced compositional bias (WOOTTON, 1994). The *dipeptide composition* is a much stronger criterion for sequence comparison (VAN HEEL, 1992). Already the frequent occurrence or absence of certain amino acid types in a protein is sometimes indicative for structure or function. Many transmembrane regions contain almost exclusively hydrophobic residues. A similar reduction of the amino acid alphabet is accompanied in coiled-coil regions with a position-dependent frequency of leucine and similar residue types (LUPAS et al., 1991; LUPAS, 1996). A glycine content of one third and a frequent occurrence of (hydroxy-)prolines with glycines mostly at every third sequence position are characteristic for a tropocollagen structure.
- A *motif* is a small conserved sequence region within larger entities. Sometimes, motifs are characteristic for structural and functional features (such as posttranslational glycosylation sites or SH3-binding sites) that develop independently from the

surrounding sequence. For these (relatively rare) motifs, the concept of sequence homology is irrelevant.

- The terms "alignment block" and "pattern" are more technical compared with "motif" and are used to deal with difficulties associated with gaps (insertions or deletions) in sequence comparisons. The *alignment block* refers to conserved parts of multiple alignments containing no gaps. A *pattern* consists of one or several alignment blocks and can also contain gaps.
- A *profile* implies a description of a sequence or an alignment in other terms than just the letters denoting amino acids. Usually, conserved physical properties among different residues are used to characterize an alignment position and to derive position-dependent weights and penalties for all amino acid types or a gap.

There is no contradiction between "profile" and "motif" since profiles may be restricted to smaller regions and patterns can be also described in amino acid property terms.

Thus, given a single query sequence, database-aided protein sequence analysis can be based on

(1) comparison of general sequence properties such as amino acid composition,
(2) motif searches (sequence pieces), or
(3) full sequence comparison.

All three variants have been attempted both as restricted to analysis of strings composed of letters denoting amino acids ("linguistic" analysis) or as profile-based. Approaches (1) and (2) are in some contrast to techniques of type (3). Whereas the latter make an effort to utilize the complete sequence information to maximize the overall signal, the former generalize from the noise in variable regions and concentrate on key features. The applicability of any of the approaches depends solely on the specific sequence studied and the resources. Generally, compositional and motif analysis are more suitable for large database searches since they are harnessed to fast word look-up algorithms which require only little computational resources, whereas full sequence comparisons, especially in profile representations, rely on exhaustive but slow dynamic programming algorithms and are applicable preferably to smaller subsets of sequences (BORK et al., 1995). Examples for motif and profile databases and searching software are listed in Tab. 2. BORK and GIBSON (1996) have evaluated different search algorithms in great detail.

3.3.2.1 Standard Procedure of Database-Aided Homology Search for a Query Sequence

The standard method in sequence analysis is to find distantly related proteins (grey zone homologies). Usually the only attempts undertaken are fast homology searches with one of the BLAST programs via World Wide Web (WWW) servers (ALTSCHUL et al., 1994), albeit FASTA or BLITZ are also in use. However, these techniques do not reveal many weakly homologous sequences. They omit proteins which are similar only in parts of their sequences (due to their multidomain structure) and, on the other hand, may find numerous similar proteins with very different functions if the sequence families are large. Sometimes, the application of special amino acid substitution matrices as offered in standard programs is an alternative if no significant hits are found (BORK and GIBSON, 1996). The probability to get a BLAST hit for a query sequence depends on the species: It is 70–90% for bacterial or yeast sequences but only ~50% for human sequences since the corresponding subset of the sequence databases is much less annotated. In the case of the genome analysis of *Haemophilus influenzae* and its comparison with the genome of *Escherichia coli* (TATUSOV et al., 1996), database homologies were found for about 90% of the genes. For about 80%, functional characterization of BLAST hits were available, but structural information was found only for 15%.

Given the concerns raised in Sect. 3.3.1, results of similarity searches in databases should be always considered with caution. Given the pressure on sequencing groups not

to overlook interesting homologies, the methodology is stretched and, with small thresholds, spurious hits are taken as real; consequently, misinterpretations cannot be avoided. Sometimes, sequence homology is just a result of structural constraints (in coiled-coil regions of muscle and other structural proteins) or the similarity is limited to a single domain. Therefore, regions with compositional bias should be identified in advance and removed from the query sequence before the BLAST search. Biological diversity appears unlimited and simple schemata do not work in all cases.

The BLAST results provide usually only a starting point for motif and profile searches. The next step is the extraction of patterns or profiles characterizing the alignment of similar sequences. A weighting scheme is useful if some type of sequences is overrepresented in the multiple alignment (HIGGINS et al., 1996). Local, conserved motifs are often the only observable markers of structural or functional regions. Iterative searches with such motifs and profiles in the sequence databases are often successful and increase the original BLAST success rate by an additional 10–20%. Sensitive searches including many cycles of analysis take several weeks and are possible only for a few sequences (Fig. 4). The analysis of thousands of sequences from large-scale sequencing projects within a few days requires automated evaluation of database searches such as with the GENEQUIZ program for gene function prediction (CASARI et al., 1994).

3.3.2.2 Significance of Weak Homologies

In assessing the significance of weak homologies, purely statistical methods are currently only of little help (BORK and GIBSON, 1996). At the beginning, formal properties of the alignment should be checked: possibility of frameshift errors, sequence weighting in multiple alignments for motif extraction, appropriate handling of gaps and amino acid substitutions given the protein family and sequence length, as well as completeness of database searches including novel sequences. The second level of checks includes structural constraints which must apply between sequences as a consequence of homology between them. Such conclusions are often possible even without knowing the three-dimensional structure of a single protein in the multiple alignment. For example, Cys patterns are expected to match in Cys-rich proteins (conservation of disulphide bonds). Gly and Pro are unlikely at positions where all other sequences have different amino acids. Insertions/deletions in highly conserved regions are suspicious. Similar logics proved successful in retrieving GAL4 and SH_2 domains (BORK and GIBSON, 1996).

The knowledge of complete genomes opens new indirect ways for functional predictions. If the enzymes belonging to a metabolic pathway are well known and a few members are found in the given organism, there is a chance to find also the others or to draw conclusions on a pathway modification (KOONIN et al., 1996). The order of genes in the genome also gives information about common regulatory blocks such as operons which might help in functional assignments (TATUSOV et al., 1996).

3.3.2.3 Search for Internal Repeats

In addition to this standard way of sequence analysis, special techniques have been developed for specific applications. A major concern are intrinsic sequence repeats that indicate duplication of structural elements and, possibly, a preceding gene duplication. Fourier spectral analysis and autocorrelation techniques have proven successful only for special classes of sequences (MCLACHLAN, 1983; MAKEEV and TUMANYAN, 1996) since insertions between repeats can be of greatly different length. Direct alignments of sequences with themselves combined with graph-theoretical methods are a more general and very efficient approach (HERINGA and ARGOS, 1993).

3.3.2.4 Complex Regression of Protein Sequence–Structure Relationships

The utilization of neural network techniques (FRISHMAN and ARGOS, 1992; HANKE et al., 1996) or hidden Markov models (KROGH et al., 1994; BAIROCH and BUCHER, 1994) was attempted for the derivation of a complex regression function between pairs of protein sequence–structure relationships since the underlying physics between sequence and three-dimensional structural properties are not well known. Due to the heuristic nature of these approaches, the real impact is difficult to assess and, probably, currently overestimated. Not only the whole sequence but also typical sequence properties have been taken as input information. For example, DUBCHAK et al. (1993) use the amino acid composition as input for neural networks trained to recognize 4 helix bundles, parallel $(\alpha\beta)_8$ barrels, nucleotide binding folds, and immunoglobulin folds. The matrix of size 20·20 containing dipeptide frequencies in a query sequence was used as input for neural networks for checking the relatedness to 45 folding classes and 4 folding types (RECZKO et al., 1994; RECZKO and BOHR, 1994).

4 Secondary Structural Features of Proteins

After a detailed consideration of different secondary structural elements, we consider the problem of an objective definition which could be the basis for a computer program to assign secondary structural states in protein tertiary structures resolved with X-ray crystallography or NMR techniques. The concept of secondary structural class, based mainly on secondary structural content, was the first attempt of a structural classification of proteins. Finally, we analyze methods for prediction of secondary structure from protein sequence alone.

4.1 Types of Secondary Structures

The primary structure, the sequence of amino acids in the polypeptide, is the basis for the formation of complex spatial structures of proteins. Historically, regularities in the spatial location of sequentially near residues are named secondary structure. More traditionally, the term secondary structure is even confined to repetitive local conformations along the polypeptide chain and includes mainly α-helices and β-sheets. In contrast, tertiary structural description puts emphasis on residue–residue contacts which are distant in sequence. Before entering the typology of secondary structural features, the physical framework of the description of protein structure needs clarification.

In the molecular-mechanical approximation, protein atoms are considered mass points and centers of spherical potential functions. A *conformation* is defined as a set of all relative positions of atomic centers. Whereas bond lengths and valence angles allow only moderate changes, rotations around single bonds have relatively small energetic thresholds and constitute the main source of conformational variability of polypeptides (Fig. 5). A torsional angle is defined as the angle between the 2 planes spanned by the first 3 atoms and the last 3 atoms respectively of a quadruple of chemically connected atoms. The *cis* position (all 4 atoms in one plane and both marginal atoms in the same halfplane with respect to the bond between the 2nd and the 3rd atom) is defined as zero degrees. Among the 3 backbone torsional angles φ (dihedral angle $C'_{i-1}N_iC\alpha_iC'_i$), ψ (dihedral angel $N_iC\alpha_iC'_i\ N_{i+1}$), and ω (dihedral angle $C\alpha_{i-1}C'_{i-1}N_iC\alpha_i$), the last one is constrained to values near 180° or 0° due to the resonance effects in an almost planar peptide group for *trans* and *cis* peptide bonds respectively. The results of studying possible atomic clashes in a monomer surrounded by peptide-groups can be summarized in a Ramachandran φ vs. ψ-plot (see, e.g., CREIGHTON, 1992, p. 183). There are generally 2 accessible regions: a large area with $-180° < \varphi < 0°$ and $100° < \psi < 190°$ (region β) as well as a smaller region α (or α_R) with $-100° < \varphi < -50°$ and $-70° < \psi < -30°$. For residues with small side

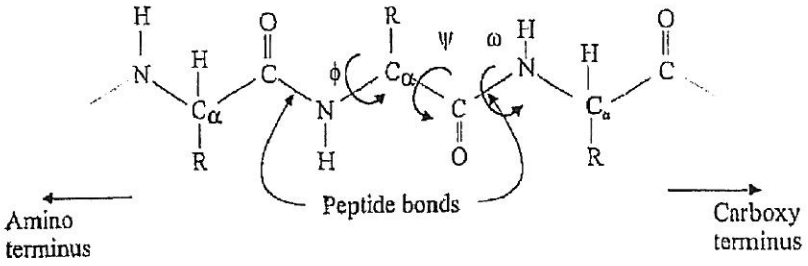

Fig. 5. The polypeptide backbone. The division of the polypeptide backbone into peptide units is a suitable methodological approach for describing the conformational properties of polypeptides.

chains, φ/ψ-values near 60°/40° are also possible (region α_L). In the case of proline, the φ is fixed to discrete values as a result of the ring structure including the backbone N-Cα bond. The conformational limitations described above restrict the possible forms of repetitive backbone structures. Repeated φ/ψ-combinations from region β result in extended chains, whereas many monomers with φ/ψ-values in the α-region form helical structures.

The analysis of local backbone conformations in crystallographic protein structure is in agreement with the molecular-mechanical treatment. Most of the residues belong to the allowed regions of the Ramachandran plot, although a few outsiders exist even in protein structures resolved with high resolution (KARPLUS, 1996). These residues are a sign of localized strain in the global protein structure which is compensated by other interactions and may indicate the non-optimality of the amino acid sequence for the given structure. In some cases, this strain has a functional role. Strain has also been observed in side chain conformations (SCHRAUBER et al., 1993).

Secondary structures observed in experimentally determined protein structures are characterized by two properties: (1) whether they are composed of conformationally (nearly) identical monomer units and (2) whether direct interactions between sequentially near residues exist. The first property distinguishes between different types of helices (including extended structures, see Tab. 3) and segments changing the direction of the polypeptide chain (loops and turns). With the second classification, α-helices and 3_{10}-helices are united with most tight turns as they include backbone hydrogen bonds between residues with small sequence separation. The extended structures [β-strands (both in parallel and antiparallel sheets or the poly-Gly-I helix), PPI-helix, PPII-helix (ε-helix)] and larger loops form the other group.

Generally, the secondary structural preferences of oligopeptide fragments describe only the ease or difficulty with which a specific se-

Tab. 3. Types of Repetitive Secondary Structures
Hydrogen bonds are denoted relative to a residue with sequence position i if applicable. Whereas most helices consist of repetitive *trans*-peptide bonds, poly-Pro-I is a polypeptide conformation with *cis*-peptide bonds

Type of Helix	Torsion Angle [°]			Hydrogen Bond	Twist [°]	Pitch [Å]
	φ	ψ	ω			
α-Helix	− 57	−47	180	i...i+4	100	1.50
3_{10}-Helix	− 49	−29	180	i...i+3	120	2.00
β-Strand (parallel sheet)	−119	113	180	n.a.	180	3.2
β-Strand (antiparallel)	−139	135	182	n.a.	180	3.4
Poly-Gly-II	− 80	150	180	n.a.	120	3.1
Poly-Pro-I	− 83	158	0	n.a.	108	1.9
Poly-ProII	− 78	149	180	n.a.	120	3.12

quence adopts the conformation in a given tertiary structure. Possible strain at this level can also be sacrificed for a tertiary topology to be achieved. The lability of preferences for secondary structural types by some amino acid sequence is obviously demonstrated by the ease of large-scale conformational changes of serpins and prion proteins. In the case of serpins, a metastable kinetically trapped 5-stranded β-sheet conformation and a largely unstructured reactive loop were observed which only slowly rearrange to the native 6-stranded sheet (MOTTONEN et al., 1992; GOLDSMITH and MOTTONEN, 1994; HUANG et al., 1994; CARREL et al., 1994). Prion proteins may trigger into a completely different α-structure followed by massive aggregation of such conformationally changed proteins (RIEK et al., 1994; CERPA et al., 1996).

4.1.1 The α-Helix and the 3_{10}-Helix

The right-handed α-helix is the most widely studied form of secondary structures. The detailed geometry may deviate from that given in Tab. 2 with respect to amino acid types constituting the helix and the helix environment in the tertiary structure. Helices have a specific sequence pattern of hydrophobic and hydrophilic residues depending on their environment (BLUNDELL and ZHU, 1995); for example, helices at the surface to solvent are amphiphatic with hydrophilic residues clustered at the side contacting solvent and hydrophobic residues directed to the center of the protein. Often, the carbonyl groups tend to point outwards for interaction with solvent or other donors. Only a fraction of the helices is truely linear, most exhibit some type of curvature or are even kinked (BARLOW and THORNTON, 1988).

The ends of α-helices are special regions since the corresponding clusters of NH and CO groups are not saturated by helical hydrogen bonds. In context also with the helix dipole originating from the dipoles of peptide bonds and accumulated with helix length, α-helix ends are a perfect place for specific substrate and ligand binding [see, for example, DORAN and CAREY (1996)]. At the termini, an α-helix may change into a 3_{10}-helix. Isolated 3_{10}-helices are rare in proteins but may be observed for small peptides (MILLHAUSER, 1995). Probably, 3_{10}-helices have a role in α-helix folding. The amino acid types have different propensities whether they are located in an α-helix or not and also depending on the specific locations, in the central region or at the N- or C-termini, the so-called capping boxes (PRESTA and ROSE, 1988; SERRANO et al., 1992; DOIG and BALDWIN, 1995; GONG et al., 1995; QIAN and CHAN, 1996). Generally, propensities vary among peptide host systems and are also different from those calculated from protein tertiary structures (BRYSON et al., 1995). Many helical segments observed in crystal structures of proteins are much less helical in the form of isolated peptides in solution.

4.1.2 Extended Structures in Protein Structures

The β-sheet is the second well characterized type of secondary structure. It consists of β-strands, a repetitive extended helical polypeptide conformation (Tab. 3). Poly-glycine I is essentially the same conformation. The polar NH and CO groups of the backbone are saturated with hydrogen bonds formed with peptide groups of neighboring strands. The helical parameters and the φ/ψ-values depend on the relative directionality of the strands inside the sheet, the amino acid composition, and the tertiary context in the protein structure. Peptides forming stable β-strands outside the protein context are difficult to design (MAYO et al., 1996).

The original models of β-sheets were planar and flat, but the structures observed in real proteins have a right-handed twist. Large sheets may even form a barrel. This is a consequence of tight packing of side chains on the surface of a sheet (MURZIN et al., 1994a; 1994b; VTYURIN, 1993; VTYURIN and PANOV, 1995). Few residues (generally 2) may not fit into the general β-sheet pattern, pucker out from an extended substructure between consecutive β-type hydrogen bonds

joining adjacent strands and form bulges (CHAN et al., 1993). Bending of β-strands is another structural distortion (DAFFNER et al., 1994). As a rule, parallel β-sheets are buried inside the protein and hydrophobic residues dominate. Antiparallel sheets have often one side exposed to solvent, resulting in alternation of hydrophobic and hydrophilic residues. This strict periodicity may be broken by a β-bulge where an extra residue is accomodated in edge β-strands. Sheets may contain also both parallel and antiparallel strands.

The left-handed poly-proline-II-helix (PPII-helix, form of poly-proline in water, acetic acid or benzyl alcohol) or ε-helix is essentially also an extended structure as in β-strands but without grouping in sheets and the formation of hydrogen bond networks. It was demonstrated that this type of secondary structure is common in globular proteins (ADZHUBEI and STERNBERG, 1993) and conserved in homologous structures (ADZHUBEI and STERNBERG, 1994). It is an important feature in structural motifs such as the HMG boxes for DNA-binding (ADZHUBEI et al., 1995).

4.1.3 Loop Segments

Polypeptide segments without a repetitive backbone structure are called loops or turns (short loops) and have been long considered together with PPII-helical fragments as "random coil" structures. They connect helical and extended segments and make changes in backbone directionality possible. Long loop regions are the most flexible parts in protein structures for the accommodation of insertions and deletions (PASCARELLA and ARGOS, 1992).

A sharp reversal in chain direction of about 180° within only 4 residues is possible with a β-turn. Typically, they occur in β-hairpins and, generally, antiparallel sheets. β-turns as observed in crystallographic protein structures cluster in discrete regions with respect to the backbone torsion angles of the central residues i+1 and i+2 (Tab. 4). Most of the tight turns are characterized by a main-chain hydrogen bond CO_i-NH_{i+3}. Sometimes, main-chain side-chain hydrogen bonds are observed in the case of serine or aspartate residues. These interactions give rise to amino acid type preferences, so that hydrogen bond acceptors such as aspartate, serine, or asparagine are usually the first residues in type I β-turns, where they can hydrogen bond to the NH group of the central peptide group (WILMOT and THORNTON, 1990; SIBANDA and THORNTON, 1993).

The complete classification of loops is an unsolved scientific task since the role of loops in protein folding and the energetic contribu-

Tab. 4. Types of Tight 4 Residue Turns (β-Turns)
The conformational characteristics of residues i+1 and i+2 are listed. The classification is given in accordance with WILMOT and THORNTON (1990). "Y" denotes the existence of the hydrogen bond CO_i-NH_{i+3}. The regions α_R and α_L are the right- and left-handed α-region respectively, β stands for β-regions. The regions γ_L and ε are located on the Ramachandran plot for glycine-like residues for positive φ angles, γ_L is similar to α_L and ε is the part for highly negative ψ-angles.

Type of β-Turn	Positon i+1			Positon i+1			Hydrogen Bond
	φ [°]	ψ [°]	Region	φ [°]	ψ [°]	Region	
I	− 60	− 30	α_R	− 90	0	α_R	Y
I′	60	30	α_L	90	0	γ_L	Y
II	− 60	120	β	80	0	γ_L	Y
II′	60	−120	ε	− 80	0	α_R	Y
VIa	− 60	120	β	− 90	0	α_R	Y
VIb	−120	120	β	− 60	0	α_R	Y
VIII	− 60	− 30	α_R	−120	120	β	N

tion to their stabilization are not known and purely geometric principles work the less, the longer the loops are. There are also examples that the identity of loops is not important for folding of a small protein at all (BRUNET et al., 1993). 5-Membered π-turns (RAJASHANKAR and RAMAKUMAR, 1996) and loops consisting of 3–8 residues (KWASIGROCH et al., 1996) have been exhaustively studied. Long loops ($\geq$10 residues) have been found to connect also mainly locally adjacent secondary structural elements ("long-closed loops"). Only 5% of the long loops (MARTIN et al., 1995) are between distant secondary structural units ("long-open loops"). Since they contain a larger percentage of proline residues the long-open loops probably contain some part of the PPII-helix.

4.2 Automatic Assignment of Secondary Structural Types in Three-Dimensional Structures

The definitions of secondary structural elements as described above are visual as derived from geometric models and are not quantitative. The secondary structure assignments given in the Brookhaven Protein Databank entries (ABOLA et al., 1987) by crystallographers and NMR spectroscopists are often subjective; therefore, a computer algorithm is necessary. The most widely used program is DSSP (KABSCH and SANDER, 1983), probably, since the corresponding software is widely available. DEFINE (RICHARDS and KUNDROT, 1988) or P-CURVES (SKLENAR et al., 1989) can also be utilized for objectification.

All 3 methods have been critically reviewed (COLLOCH et al., 1993). It was found that the assignments coincide only in 63% of the residues. This can be explained by the particularities of each method. The DSSP approach considers hydrogen bond patterns, while the P-CURVE algorithm finds regularities along the helicoidal axis and the DEFINE technique measures distances between Cα atoms. Therefore, when evaluating predictions, the "standard-of-truth" might vary depending on which property was used for the secondary structure assignment. In addition, the most widely used DSSP algorithm produces many α-helices comprising 4 or fewer residues and β-strands consisting of 2 or even only one monomer. STRIDE (FRISHMAN and ARGOS, 1995) is a recent modification of DSSP reported to give secondary structure assignments somewhat more coinciding with subjective assignments of crystallographers than DSSP, but not much.

An interesting alternative for finding regularities in backbone structures is the algorithm of ADZHUBEI and STERNBERG (1993) relying only on Cα coordinates.

4.3 The Concept of Secondary Structural Class

Early in 1976, when only about 40 crystallographic structures of proteins were known, LEVITT and CHOTHIA (1976) studied the succession of secondary structural elements along the amino acid sequence. Intuitively, they grouped the proteins into 4 structural classes (or folding types):

- all-α proteins having *only* α-helix secondary structural elements (more than 60% of the residues adopt helical conformation, no residues in β-strands);
- all-β proteins consisting *mainly* of (often antiparallel) β-strands;
- $\alpha+\beta$ proteins having *independent clusters* of α-helices and (often antiparallel) β-strands in the sequence; and
- α/β proteins with mixed (often *alternating*) segments of α-helix and (mostly parallel) β-strands.

Many more protein structures are known today, and, for an increasing number, it is not easy to classify them in accordance with the definitions of LEVITT and CHOTHIA (1976). A variety of class definitions in terms of secondary structural content has been presented (Fig. 6). At the same time, there are no clear clusters any more in the α-content vs. β-content plot for large sets of protein structures as available in the PDB. Also the notion of $\alpha+\beta$ and α/β proteins is not longer applicable. For

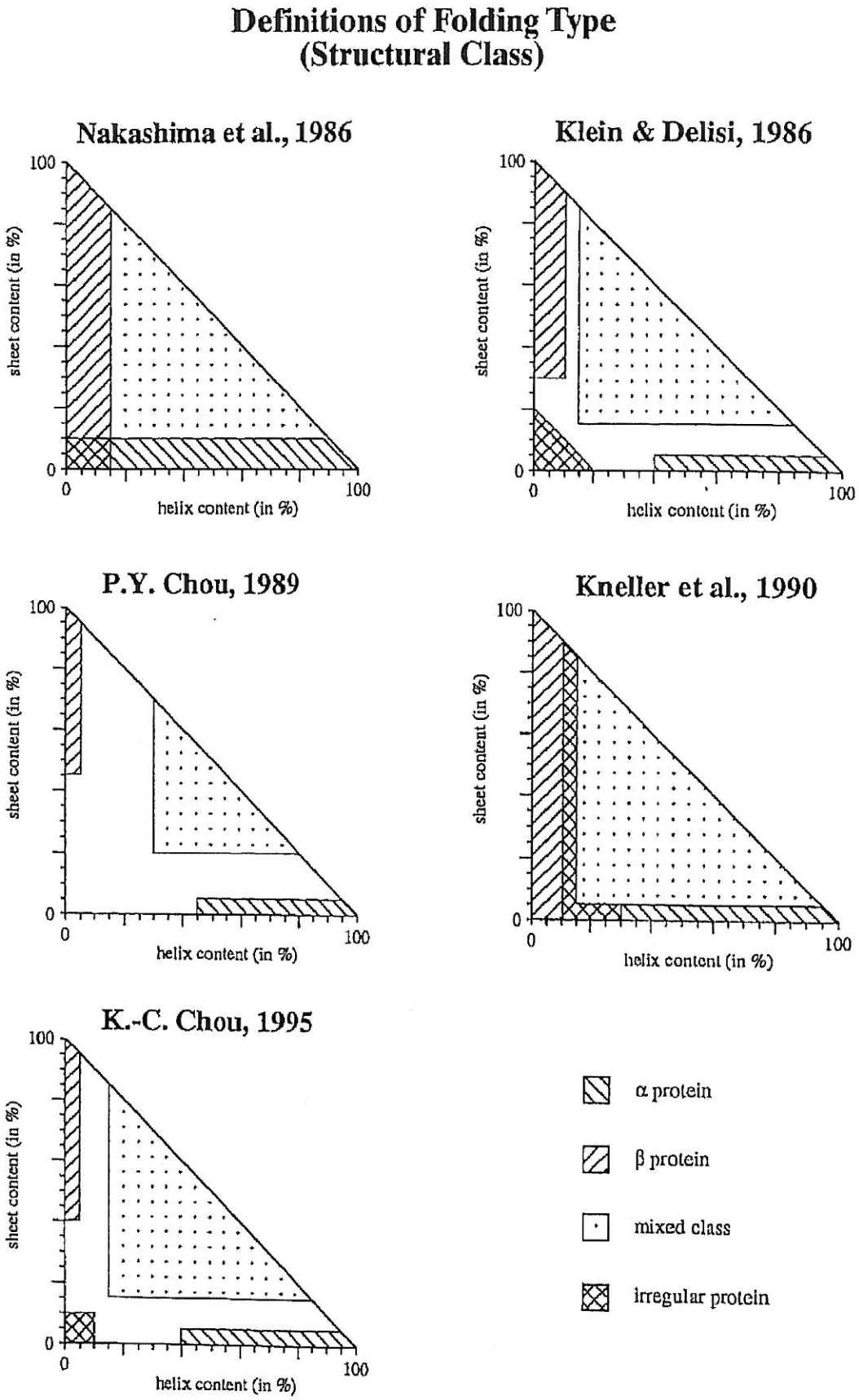

Fig. 6. Secondary structural class (folding type) definitions. The definitions of NAKASHIMA et al. (1986), KLEIN and DELISI (1986), CHOU (1989), KNELLER et al. (1990), and CHOU (1995) are illustrated in form of α vs. β content trigonal plots. The white regions correspond to proteins which are not assignable to any secondary structural class with the definition of the given authors.

example, both the acylphosphatase, PDB entry 1APS (PASTORE et al., 1992), and the B-chain of the regulatory domain of the aspartate carbamoyltransferase, PDB entry 8ATC (STEVENS et al., 1990), have a 2-layered structure consisting of an antiparallel β-sheet and two parallel α-helices. The existence of an antiparallel sheet is characteristic for $\alpha+\beta$ structures. But a more detailed investigation of the structures reveals a high degree of secondary structural alternation and $\beta\alpha\beta\beta\alpha\beta$ and a doublet of the $\beta\alpha\beta$ motif, both observations pointing to class α/β. It was argued by EISENHABER et al. (1996b) that the term "secondary structural class" may be applied today, if at all, only as a classification in α, β, mixed and irregular proteins. The best definition in thresholds of secondary structural content is that of NAKASHIMA et al. (1986) since

(1) it is applicable to all proteins,
(2) it is compatible with intuitive definitions of irregular and mixed proteins, and
(3) the secondary structural content thresholds are at minima of the occurrence vs. secondary structure contents plot for large selections of PDB structures (EISENHABER et al., 1996b).

Nevertheless, the concept of structural class, based on the secondary structural content of the protein and the directionality of β-strands, is useful from the experimental as well as the theoretical point of view. The folding type of a protein can be directly determined by relatively simple spectroscopic methods. With a sufficient quantity of the protein available, circular dichroism (CD) spectroscopy in the UV absorption range can be used to obtain reliable measures of secondary structure content, especially for α-helices, but also for parallel and antiparallel β-strands (JOHNSON, Jr., 1990; PERCZEL et al., 1991; SREERAMA and WOODY, 1994).

Secondary structural class restrictions have a high impact for secondary and tertiary structure prediction (EISENHABER et al., 1995b). The accuracy of secondary structure prediction from the amino acid sequence with methods designed for all-α proteins is larger than 80% compared with maximally ~70% in the general case (KNELLER et al., 1990; MUGGLETON et al., 1992, 1993). The effect of knowledge of structural class alone in improving secondary structure prediction is comparable with the use of the extra information contained in multiple alignments of homologous sequences (LEVIN et al., 1993; ROST and SANDER, 1993b). The secondary structural class is related to various properties of a protein such as its location in extra- or intracellular compartments, biological function (being an enzyme or not), or the existence of disulfide bonds (NISHIKAWA and OOI, 1982; NISHIKAWA et al., 1983a, b).

4.4 Prediction of Secondary Structural Features

All secondary structure prediction methods currently in use are knowledge-based. A learning set of protein structures from the PDB is utilized to derive a prediction rule; e.g., for the calculation of propensities (the relation of the frequency of a given amino acid type in a certain secondary structure with the frequency of any residue to be observed in the same secondary structural state). The prediction rule is applied to protein sequences from a test set of PDB structures to estimate the expected rate of successful prediction.

Such an approach has two types of problems. Difficulties of the first type are of technical nature. Researchers often invent prediction functions with such a large number of parameters that the number of structurally non-homologous proteins in the learning set is not sufficient to determine all parameters unambiguously. Also, the protein structure in the test set should not be contained in the learning set to exclude the information to be predicted from the rule. The latter difficulty can be avoided with the so-called "jackknife"-scheme: learning is done on N-1 proteins of a set and prediction is obtained for the Nth protein. This procedure is repeated for all proteins in the data set (EISENHABER et al., 1996a).

The second class of difficulties has its roots in the current state of the protein Data Bank.

Generally, these prediction techniques will work only for sequences representing globular proteins. All sequences with compositional bias constituting about 40% of SWISS-PROT (WOOTTON, 1994) cannot be well considered since an insufficient number of the structures of such proteins is known. Additionally, it is not clear whether the learning sets of proteins used now explore the available sequence and structure space for proteins comprehensively. Therefore, protein structures available in the PDB in future years will probably result in a decrease of the prediction power of today's algorithms and the success rates discussed in this work should be considered upper estimates.

4.4.1 Secondary Structural Class and Content Prediction

In the hierarchy of prediction methods, secondary structural class prediction [secondary structure above or below a threshold with classification into folding types all-α, all-β, mixed types (sometimes subclassified in $\alpha+\beta$ α/β) and irregular forms] corresponds to a lower level and appears a simpler task compared with secondary structural content prediction (fraction of residues in the 3 states helix, sheet, and coil) or even traditional secondary structure prediction (state of every residue among the 3 alternatives helix, sheet, and coil). Since the discovery of NISHIKAWA and OOI (1982) that the amino acid composition is tightly related to secondary structural class, both analytical distance criteria in the amino acid composition space and neural network methods have been applied for the jury decision between 3–5 folding types. A detailed review has been given by EISENHABER et al. (1995b, 1996b). Especially after 3 recent publications (CHOU, 1995; ZHANG and CHOU, 1995; CHOU and ZHANG, 1995), the paradoxical situation emerged that folding type prediction appears solved (reported prediction accuracies up to 100%) whereas secondary structure prediction even with multiple alignments approaches only about 70% accuracy, and the success rate of its class prediction is only near 75% (ROST and SANDER, 1993a; LEVIN et al., 1993; ROST and SANDER, 1994b). This paradox has now been solved (EISENHABER et al., 1996b). It was shown (1) that certain structural class definitions leave many proteins with intermediate secondary structural content without assignment to any class (predictions are made only for proteins with extreme contents); (2) that various analytical distance-based jury decision methods yield only prediction accuracies up to 55% for representative test sets even of extreme proteins; and (3) that the real impact of amino acid composition on secondary structural class is only about 60%. The amino acid composition determines the secondary structural content with an error of about 13% (EISENHABER et al., 1996a). Secondary structural content and class prediction based on the knowledge of only amino acid composition of the query protein is available on the WWW (program SSCP with the URL http://www.embl-heidelberg.de/~eisenhab/).

4.4.2 Traditional Secondary Structure Prediction

The prediction whether a residue in a protein sequence is in helix, sheet, or coil state is a classical problem in protein structure prediction. It was found early that different amino acid types have various preponderance for particular secondary structural environment. Consequently, methods of the first generation were propensity-based. The principal achievments are represented by the Chou-Fasman (CHOU and FASMAN, 1974a, b; YANG, 1996), GORIII (GARNIER et al., 1978) and COMBI (GIBRAT et al., 1987) methods [for detailed review see EISENHABER et al. (1995b) pp. 10–18]. The GORIII method relies not only on single residue propensities but also on statistically significant pairwise residue interactions. The prediction accuracy achieved was 60–63% (GIBRAT et al., 1987; GARNIER and LEVIN, 1991). A further improvement of about 4% (BIOU et al., 1988) was attained by combining the GORIII method with 2 other prediction schemes: one based on hydrophobicity patterns which are often observed in regular secondary structures (bit pattern method),

and the other using structural similarity between short, sequentially homologous peptides (LEVIN and GARNIER, 1988). As was shown by GIBRAT et al. (1991), the predictive power of methods relying only on sequentially local structure information is limited to about 65%. With today's databases, the estimate would be even lower. The inability to find properly helical and strand segments is an even more important deficiency of the classical methods (ZHU, 1995).

Further development in secondary structure prediction occurred in two directions (EISENHABER et al., 1995b):

(1) It led to the involvement of multiple alignments of database sequences with the query sequence giving information about the mutability of sequence positions. The prediction accuracy increased to values of about 70% (LEVIN et al., 1993; MEHTA et al., 1995). Combined use of evolutionary information and capping rules was described by WAKO and BLUNDELL (1994a).
(2) Neural networks were applied to find a complex regression between diverse types of input information (e.g., data from sequence windows with respect to single sequences and multiple alignments, amino acid compositions and sequence length of the query sequence, etc.) and the secondary structural state of a residue in the center of the sequence window. The most prominent algorithm of this type is PHDIII (ROST and SANDER, 1994b).

Direct inclusion of non-local interactions proved to increase the prediction success significantly even for single sequence predictions. FRISHMAN and ARGOS (1996) used the statistics of hydrogen bonds for pairs of amino acid types in α-helices and β-strands and achieved correct predictions of about 68%.

Several engines for secondary structure prediction are available at http://www.embl-heidelberg.de [among which are PHD (ROST and SANDER, 1994b), SSPRED (MEHTA et al., 1995), PREDATOR (FRISHMAN and ARGOS, 1996)]. The server at http://www.genebee.msu.su provides a modification of the GOR algorithm implemented by BRODSKY and coworkers. Secondary structure prediction service is also provided by the servers at

- http://kiwi.imgen.bcm.tmc.edu [nearest neighbor method of SALAMOV et al., 1995),
- http://www.ibcp.fr/predict.html [SOPM method of GEOURJON and DELEAGE (1994)], and
- http://www.cmpharm.ucsf.edu [a neural network technique of KNELLER et al. (1990)].

4.4.3 Prediction of Transmembrane Regions, Coiled-Coil Segments, and Antigenic Sites

Until recently, *transmembrane segments* were known only as α-helices consisting of stretches of 21 hydrophobic residues. A sequence database analysis of putative transmembrane segments without any assumption of secondary structure has indeed shown that maximal sequence correlation is observed at a periodicity of 3.6 residues characteristic for α-helices (SAMATEY et al., 1995). Other types of transmembrane segments have been found only recently. The porins (WEISS and SCHULZ, 1992) have a build-up consisting of 16 β-sheets arranged as a complete transmembrane barrel giving rise to a big central hole. Recent three-dimensional structures also show further variants. The helices can be tilted against the perpendicular plane, like the case in a light harvesting complex (KÜHLBRANDT et al., 1994). Presence of helices parallel to the membrane plane has also been shown (KÜHLBRANDT et al., 1994; PICOT et al., 1994). There are also indications that the membrane-spanning segments can consist of single extended β-strand like structures, thus making it possible to span the membrane with fewer residues than in the case of an α-helix (HUCHO et al., 1994). All this implies that prediction of membrane-spanning regions might be a more difficult task than has hitherto been anticipated.

Nevertheless, recent prediction algorithms are still tuned to find only α-helical transmembrane segments. The algorithms rely on hydrophobicity scales and information from multiple alignments (EISENHABER et al., 1995b). The prediction function is sometimes found in form of a neural network (DOMBI and LAWRENCE, 1994; ROST et al., 1995). In other cases, explicit expressions are optimized for a given learning set (PERSSON and ARGOS, 1994). Profile techniques have also been employed (EFREMOV and VERGOTEN, 1996a, b). The topology prediction, the derivation of the intra- and extracellular sides of the helices (PERSSON and ARGOS, 1996; ROST et al., 1996), is based on the so-called "positive-inside rule" of VON HEIJNE (1986, 1995). It was found that internal loops between transmembrane segments are richer in positively charged residues. Although the authors usually claim very high success rates for prediction with their technique, the results should be considered with caution given the small number of known transmembrane segments. WWW servers for the prediction of transmembrane regions are available from http://www.embl-heidelberg.de [TMAP and TTOP (PERSSON and ARGOS, 1994, 1996), PredictProtein (ROST et al., 1995, 1996)].

Coiled-coils are intertwined α-helices. Due to their docking under a small angle, the residues at the docking side need to be large and hydrophobic; for example, such as leucine (WALTHER et al., 1996). It is this position-dependent pattern which is analyzed and searched for by the algorithm of LUPAS et al. (1991, 1996).

Transmembrane and coiled-coil regions are special examples of compositional bias. General software is available to diagnose sequence segments with low complexity and information content (WOOTTON, 1994). The analysis of compositional bias is an important initial step in protein sequence analysis.

An important step in the biochemical characterization of a protein is the detection of *antigenic sites*, responsible for specific antibody binding. Since epitopes are usually located in loop structures, this issue is discussed here as a secondary structural feature. Prediction algorithms attempt to locate antigenic sites indirectly as hydrophilic and malleable loops at the protein surface. Antigenic epitopes are also known as mutation hotspots. All these properties have been employed in sophisticated prediction algorithms, albeit with limited success. A detailed review has been given elsewhere (see pp. 8–9 of EISENHABER et al., 1995b).

5 Tertiary Protein Structures

This section is dedicated to the principles of tertiary structure construction and methods for prediction and modeling of tertiary structural features. The analysis and classification of known tertiary structures have helped in the derivation of rules followed by nature in the design of proteins. At the same time, we do not understand the structure formation sufficiently well as demonstrated by the small success of tertiary structure prediction just from amino acid sequence.

5.1 Phenomenology of Tertiary Protein Structures

Today, the three-dimensional structures of several thousand proteins are known from X-ray crystallographic or NMR studies, albeit many of the proteins have similar amino acid sequences [e.g., the PDB contains about 250 mutants of T4 lysozyme (ABOLA et al., 1987)]. Analysis and comparison of this structural information is and will continue to be our main source on protein tertiary structures since they are very complex. This view on tertiary structures has also its drawbacks: The protein is mainly seen as static entity with a fixed structure whereas other experimental techniques but with lower resolution (as well as the comparative analysis of the same protein in different crystallographic environments) deliver much information on small and large scale conformational fluctuations and transitions.

5.1.1 Construction Principle No. 1 – Close Packing

As known from statistical physics, heteropolymers with primarily attractive forces between monomers form globular structures with locally confined conformational fluctuations. This is the case also for tertiary structures of small proteins. Their main characteristic is the close packing of atoms (RICHARDS and LIM, 1994) inside a volume of generally spherical shape with an irregular surface. The dense packing is achieved by contacts between residues with large sequence separation. Larger proteins appear to consist of several domains, each having its own densely packed core. Most likely, domains are connected with a single segment of the polypeptide chain and each domain consists of a single stretch. In some cases, protein domains are not so heavily segregated. For example, in pyruvat kinase, phosphofructokinase, and arabinose-binding protein, there are 2 or 3 links between domains. Much effort has been concentrated on elaborating objective criteria and automatic algorithms for domain recognition in large proteins (HOLM and SANDER, 1994a; NICHOLS et al., 1995; ISLAM et al., 1995; SWINDELLS, 1995a; SOWDHAMINI and BLUNDELL, 1995; SIDDIQUI and BARTON, 1995).

Valuable information can be obtained from studying the close packing of secondary structural elements in so-called supersecondary structures. Just the condition of packing optimization was sufficient to obtain a rigorous mathematical model and to derive equations for the parameters describing α-helix-α-helix docking (WALTHER et al., 1996). The authors showed:

(1) the existence of 3 different packing cell systems resulting in 5 types of docking angles,
(2) the dependence of the packing cells on helix radius and, therefore, on amino acid composition of the helix, and
(3) the hierarchy of optimal and suboptimal "knobs-into-holes" and "knobs-onto-knobs" packing schemes.

Similar studies obtained global conformational constraints for β-sheets (MURZIN et al., 1994a, b; VTYURIN, 1993; VTYURIN and PANOV, 1995). Typical supersecondary structures with β-strands are the greek key and the meander. The $\beta\alpha\beta$-motif is a common mixed supersecondary structure where an α-helix is packed against 2 β-strands forming a parallel sheet (CREIGHTON, 1992).

The search of dense clusters of residues in a protein tertiary structure is yet another perspective for studying close packing. HERINGA and ARGOS (1993) have developed a strategy for locating such groups of residues. This condition has also been used to identify core regions in proteins (SWINDELLS, 1995b). Close packing is often only possible if some substructures accept strained conformation; for example, side chains may adopt non-rotameric torsion angle combinations (SCHRAUBER et al., 1993).

Often, packing is not optimal and does not exhaust the internal space of a protein. Cavities are wide-spread in protein structures. They are usually more readily tolerated than actively favored and can sometimes, be considered as a type of conformational strain since many cavities destabilize proteins (HUBBARD et al., 1994; HUBBARD and ARGOS, 1995).

The compactization of a protein is thought to proceed co-translationally; i.e., some part of the polypeptide folds whereas the remaining chain is still synthesized. Therefore, it is difficult for topological knots to appear in tertiary structures. Nevertheless, knots do happen. In the case of carbonic anhydrase (MANSFIELD, 1994), a C-terminal knot exists. (S)-adenosylmethionine synthetase has a real N-terminal knot (TAKUSAGAWA and KAMITORI, 1996). LIANG and MISLOW (1994a, b) and MAO (1993) have described topological chiralities formed through disulfide bonds, hydrogen bonds and coordination bonds.

5.1.2 Construction Principle No. 2 – Hydrophobic Interior and Hydrophilic Exterior

The uneven distribution of hydrophilic and hydrophobic residues between the interior and the surface is the second basic property of globular proteins. This is the result of interaction with water and ions of the solvent (hydrophobic effect). As a result, a protein core region with low dielectric permissiveness is embedded in an environment with high dielectric permittivity. Special restraints apply to atomic groups capable of hydrogen bonding. At the protein surface, they can have contact with water molecules. All hydrogen bond donors and acceptors buried inside must also have a hydrogen partner supplied by the protein itself. Unsatisfied hydrogen bonding abilities in the protein core are extremely destabilizing for tertiary structures.

For the analysis of surface properties of proteins, specialized and fast software tools are required. Traditionally, the van der Waals, the solvent-accessible (LEE and RICHARDS, 1971), and the molecular (CONNOLLY, 1983) surface are studied (Fig. 7). Recently, efficient techniques for computing both van der Waals and solvent-accessible surfaces of proteins have been published (EISENHABER and ARGOS, 1993; EISENHABER et al., 1995a). Programm ASC and NSC are available via WWW (http://www.embl-heidelberg.de/~eisenhab/).

Regions of the protein surface formed by hydrophobic atoms are unable to interact with surrounding water molecules in a similarly strong manner as polar atomic groups. Therefore, exposed hydrophobic surface patches are generally destabilizing for the protein and are often sites of binding subunits or other ligands. Hence, it is desirable to have objective criteria for locating hydrophobic surface patches. But paradoxically, the solvent-accessible surface as defined by LEE and RICHARDS (1971) which is traditionally used for the analysis of solvation properties of proteins is not informative for the determination of hydrophobic surface clusters. The hydrophobic part of the solvent-accessible surface of a typical monomeric globular protein consists of a single and large interconnected region formed from faces of apolar atoms and constituting about 60% of the solvent-accessi-

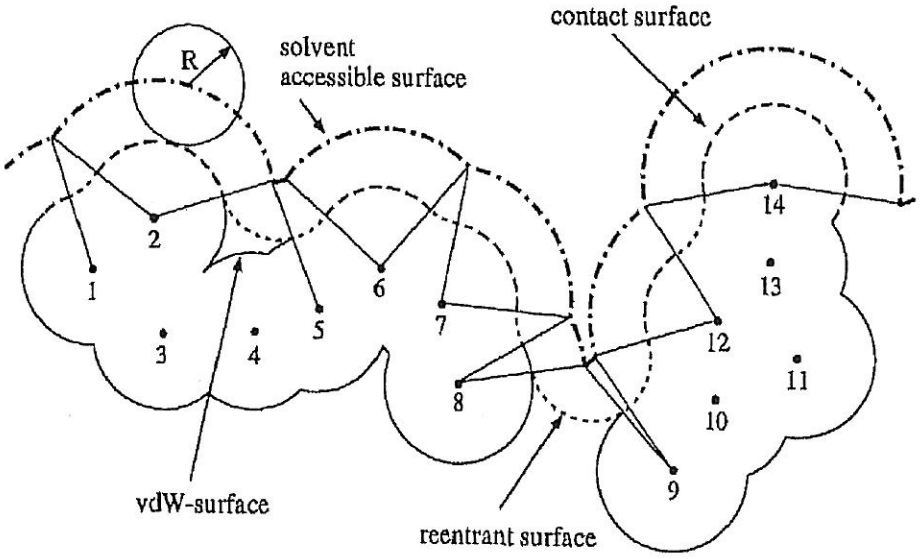

Fig. 7. Protein surfaces. The van der Waals surface (– – –) and the solvent-accessible surface (–·–·–) consist of pieces of spheres centred at atomic positions that are not occluded by neighboring spheres. In the first case, the radii are the so-called van der Waals radii, in the second case, they are incremented by the radius of a probe sphere modeling the solvent (usually the probe radius is 1.4 Å for water). The molecular surface is only in part identical with the van der Waals surface; the reentrant surface (– – – –) is formed by probe spheres at invaginations which are too small for allowing the probe to enter. The molecular surface surrounds the solvent-excluded volume.

ble surface area. Therefore, the direct delineation of hydrophobic surface patches on an atom-wise basis is impossible. Experimental data indicate that, in a 2-state hydration model, a protein can be considered as unified with its first hydration shell in its interaction with bulk water. It has been shown (EISENHABER and ARGOS, 1996) that, if the surface area occupied by water molecules bound at polar protein atoms is removed, only about two thirds of the hydrophobic part of the protein surface remain accessible to bulk solvent. Moreover, the organization of the hydrophobic part of the solvent-accessible surface experiences a drastic change such that the single interconnected hydrophobic region disintegrates into many smaller patches; i.e., the physical definition of a hydrophobic surface region as non-occupied by 1st hydration shell water molecules distinguishes between hydrophobic surface clusters and small interconnecting channels. The formation of hydrophobic surface regions owing to the structure of the first hydration shell can be computationally simulated by a small radial increment of solvent-accessible polar atoms (0.35–0.5 Å), followed by calculation of the remaining exposed hydrophobic patches.

Based on the area distribution of hydrophobic surface regions, a surface energy value of 18 ± 2 cal mol^{-1} $Å^{-2}$ was obtained which compares favorably with the parameters for carbon obtained by other authors who use the crystal geometry of succinic acid or energies of transfer from hydrophobic solvent to water for small organic compounds (EISENHABER, 1996). Thus, the transferability of atomic solvation parameters for hydrophobic atoms to macromolecules has been directly demonstrated.

The solvation energy of a protein in an aqueous environment includes several components:

- the hydrogen bond formation with polar groups of the macromolecule,
- the entropy change of water molecules due to binding with polar groups or their release into bulk water,
- the cavity formation due to the solute excluded volume,
- non-valent interaction of non-hydrogen bonded water molecules with protein atoms at the surface, and
- the polarization of bulk water (and of the volume inside the macromolecule) and the changes of salt density.

Whereas the first 4 components can, in a first approximation, be linearly related to surface properties of the protein (short-range part), the last contribution is of electrostatic nature (long-range part) and requires, in principle, an integration over the whole volume of solute and solvent which is the equivalent solution of the Poisson–Boltzmann differential equation (PBE). It is not possible to reduce the whole solvation energy to a simple surface term (JUFFER et al., 1995, 1996).

5.1.3 Protein Structure Comparison and Structural Families

Protein tertiary structure comparison is necessary to elucidate topologically equivalent regions, to determine structural differences in space and to find insertions/deletions in one structure relative to others. To date, superimposing 3D protein structures provide the most sensitive technique for recognizing very distant relationships between amino acid sequences with low residue identities up to only 2% (HOLM and SANDER, 1996). It was one of the surprising news after a larger number of protein structures has been resolved with X-ray crystallography that many protein domains were visually similar above the regularities due to secondary structural constraints. The need for objective clarification of this similarity was felt by many researchers, and a variety of definitions of structural similarity and algorithms for protein structure comparison was proposed. The methods differ in the following:

1. Is the main emphasis put on global, overall coincidence or on the search of only local structural similarities?
2. Does the algorithm require that all structural elements occur in the same sequential

order or is it sensitive to sequential interchanges?
3. Are the structures compared up to atomic detail or is the algorithm oriented on a comparison of higher order structural blocks?

As shown by GODZIK (1996), the quantitative results of structural comparisons depend highly on the criterion applied. Indeed, the three best known classifications of protein structures

- SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/index.html),
- CATH (http://www.biochem.ucl.ac.uk/bsm/cath/CATHintro.html), and
- FSSP (http://swift.embl-heidelberg.de/fssp/)

differ drastically even at higher hierarchical levels. This is just a reflection of the fact that the principles of tertiary structure formation are not well understood, and subjective choices such as mutual Cα distance metrics or scaffolds of secondary structural elements are used as basis of the classifications. The situation is similar to that of Linné when he started to introduce systematics into botany based on the number of anthers in flowers.

Comparison techniques have been reviewed in detail by EISENHABER et al. (1995b). Here, we give a short outline of the variety of algorithms. The classical measure of distance between two structures is the global r.m.s.d. of the distance between equivalent Cα atoms after spatial superposition. Early comparison techniques involve rigid body superpositions with the subjective assignment of initial equivalences. For largely divergent structures, window comparison techniques relying on successive oligopeptide superpositions along the sequence have been applied to locate similar substructures. Structure alignment methods based on dynamic programing have been developed to handle rigorously variable length gaps in the alignment (TAYLOR and ORENGO, 1989; ZUKER and SOMORJAI, 1989; SALI and BLUNDELL, 1990). The sensitivity of the technique has been enhanced by including hydrogen bonding, solvent exposure, torsional angles, secondary structural assignments and the like in addition to α-carbon distances. Dynamic programing cannot directly incorporate sequentially nonlocal effects, consequently, multiple levels of dynamic programing (TAYLOR and ORENGO, 1989; ORENGO and TAYLOR, 1990), self-consistency tests for suboptimal alignments (LUO et al., 1993), or stochastic optimization via genetic algorithms (MAY and JOHNSON, 1994) have been used.

Algorithms relying on two-dimensional plots of pairwise Cα distances (or hydrogen bonds or main chain dihedral angle matches) do not depend on the sequential order of structural blocks. Parts of these plots can be compared to search for similar substructures or patterns (BARTON and STERNBERG, 1988; RICHARDS and KUNDROT, 1988; VRIEND and SANDER, 1991; HOLM and SANDER, 1993). The direct comparison of substructures (e.g., hexapeptides) and the search for the langest common homologous domain is a variant of this approach (BACHAR et al., 1993; ALEXANDROV and GO, 1994).

On the other hand, emphasis can be put on overall topological equivalence of secondary structural blocks rather than on detailed atomic correspondence (EISENHABER et al., 1995b). Characteristic patterns of secondary structure are much more robust to structural changes than individual amino acid positions, and even mutations often do not destroy the overall topology of the main chain. Simplified representations of building blocks in the form of vectors, e.g., along the axes of secondary structural elements, can be compared. In graph theoretic approaches, protein structural elements and their relations are coded in the form of nodes and edges.

Analyses of both known three-dimensional protein structures and amino acid sequences revealed that proteins are clustered into families whose members may have evolved from a common ancestor, share a characteristic fold and, sometimes, have a similar function (PASCARELLA and ARGOS, 1992; HOLM et al., 1992; YEE and DILL, 1993; ORENGO et al., 1993; HOLM and SANDER, 1994b; LESSEL and SCHOMBURG, 1994; SOWDHAMINI et al., 1996). Some authors even think that the total number of different folds may be in the range of a few thousand (for critical review, see EI-

SENHABER et al., 1995b). It must be emphasized that the number of structural families depends critically on the value of the homology threshold applied in the routine comparison of protein structures, and structural similarity is not structural identity. For pairs of distantly related proteins (residue identity ~20%), the regions with the same general fold comprise less than half of each molecule and the r.m.s. deviation between equivalent main chain atoms is 1.8–2.5 Å (CHOTHIA and LESK, 1986). The distance error can be in the range of up to 5 Å for Cα atoms of proteins with about 25% residue identity (CHELVANAYAGAM et al., 1994). The equivalent secondary structural units may be shifted relative to each other by as much as 7 Å with rotations up to 30° (LESK and CHOTHIA, 1980; CHELVANAYAGAM et al., 1994). Pairwise residue–residue contacts may be conserved to only 12%, solvent accessibilities and secondary structures can be maintained to less than 40% (RUSSEL and BARTON, 1994). Thus, even if the scaffold of secondary structures is similar, the physical nature of stabilizing forces can be entirely different.

5.2. Prediction and Modeling of Tertiary Structure

Theoretical approaches to tertiary structure modeling are classified into 3 main streams.

(1) Attempts to predict the protein topology directly from sequence data alone based on the knowledge of intramolecular interactions are unified under the name *ab initio* approach. Since these techniques have not yet proven successful, other methods involving additional structural data on similar proteins have been developed.
(2) Threading tries to identify a suitable fold for the query sequence among those already known.
(3) Homology modeling includes techniques for fitting a new sequence into the known structure of another protein.

Solvent accessibility of residues is an important characteristic for the location of residues in the tertiary structure. Methods developed for secondary structure prediction have also found application for this purpose.

5.2.1 Computation of Protein Structures Based on Fundamental Physical Principles (*ab initio* Approach)

The fundamental physical approach to the protein folding problem (predicting the tertiary fold from the amino acid sequence alone) relies on the hypothesis on the native protein structure as a minimum of free energy; i.e., the native protein structure corresponds to a system at thermodynamic equilibrium with a minimum of free energy. *In vitro* renaturation experiments strongly support this view (ANFINSEN, 1973; CREIGHTON, 1992) since they imply that the complete information necessary for protein folding is comprised in the amino acid sequence. Thus, it would be sufficient to compute an ensemble of conformations representative for the state of lowest free energy. The conformational invariants of this ensemble (e.g., the densely packed protein core) are the characteristics of the native structure. In a more simplified approach, a unique conformation with the lowest sum of intramolecular potential energy, conformational entropy term and solvation free energy is considered to represent the native state. This view is not unchallenged.

The computational problem of finding the lowest energy conformation of a polypeptide chain from an energy function containing pairwise terms and possibly other expressions is NP-complete (NGO and MARKS, 1992; UNGER and MOULT, 1994; FRAENKEL, 1993). The contention of LEVINTHAL (1968) that proteins search only a tiny fraction of the conformational space and move into the lowest kinetically accessible free energy minimum appears much more likely in this context. First experimental evidence in support of this view has been provided recently. The α-lytic protease was shown to exist in two forms: an inactive, metastable intermediate and an active native structure. Both conformations are

separated by a barrier with activation energy of about 27 kcal mol^{-1}. A catalyst which is normally covalently attached to the protein is necessary to complete folding of the intermediate "molten globule", a less compact state compared with the native conformation (BAKER et al., 1992a, b). In the case of serpins, a metastable kinetically trapped 5-stranded β-sheet conformation was found which only slowly rearranges to the native 6-stranded form (MOTTONEN et al., 1992). During the last years, the role of biological factors such as the peptidyl–prolyl–isomerase, disulfide isomerase and molecular chaperonins in controlling the kinetics of protein folding and subunit assembly has been discovered (GETHING and SAMBROOK, 1992; HARTL, 1994). Often, the protein tertiary structure is incorporated into a system of a higher organizational level, and requirements of the latter may shift the conformation from a low-energy state.

In the "weak thermodynamic" hypothesis of UNGER and MOULT (1994), evolutionary arguments are taken into consideration. An originally functional structure of a protein corresponding to a local minimum will drift towards the global free energy minimum due to the combined effect of random mutations and the constant selective pressure of evolution. Hiding the native structure behind a large energy barrier may also be a sophisticated variant of enzyme activity regulation. Probably, only those sequences fold into unique conformations, fold fast and fold *via* the all-or-none transition (with the release of substantial latent heat) that have a pronounced energy minimum which is sufficiently distinguished from all other conformational states in the energy spectrum (SALI et al., 1994).

Hence, the search for low-energy conformations of polypeptides is still a promising plan for prediction of 3D structure and function. Two prerequisites for this approach are necessary:

- an energy function for discriminating the weight of different conformations in the native ensemble and
- a procedure for efficient searching of the conformational space.

A detailed review of both aspects has been given by EISENHABER et al. (1995b).

Although 3 decades of enormous scientific efforts have been concentrated on the *ab initio* folding problem, a solution to compute the structure from sequence has not yet been developed. The hardest problem waiting for solution is an appropriate energy function for the discrimination of native and non-native conformations. Electrostatic interactions in aqueous solutions, the solvation energy, hydrogen bonding, and entropic effects are not sufficiently well described. The identification of low-energy conformations in the highly dimensional conformation space is also not a trivial task, although different techniques for conformational searches such as Monte Carlo, genetic algorithms, and molecular dynamics have achieved a high level of maturity.

The main applications of the *ab initio* approach are conformational searches in combination with experimental restraints. Packing requirements and covalent strain are sufficiently well modeled by existing energy functions. Together with experimental data about pairwise distances (from NMR, cross-linking studies, and the like) or with X-ray diffraction data, the search techniques are applied for structure refinement and the generation of conformations satisfying restraints.

5.2.2 Threading Amino Acid Sequences Through Structural Motifs

The "threading" approach makes the far-reaching assumption that the query sequences under study might accept one of the protein folds already studied by X-ray crystallography or NMR. The structure prediction problem is thus greatly simplified since the allowable conformational space is reduced to about 100–300 unique protein topologies presently known. The primary goal of a "threading" method is to establish relationships between amino acid sequences and folding patterns, i.e., to select the most probable fold for a given sequence or to recognize suitable sequences that might fold into a given structure.

This approach has been stimulated by 3 observations:

(1) The number of different folds in the PDB grows more slowly than the number of new protein structures (notion of structural families).
(2) Distant relationships between sequences may be found by alignment to property profiles (see Sect. 3.3.2).
(3) Empirical potential functions for estimating solvation can distinguish incorrect folds.

As introduced by BRYANT and LAWRENCE (1993), "threading" a sequence through a fold implies a specific alignment between the amino acids of the sequence under consideration and the residue positions of the folding motif. The known structure establishes a set of possible amino acid positions in the three-dimensional space (the tertiary template) characterized physically by solvent accessibility, types and number of residue–residue contacts, backbone conformation and the like. The query sequence is made similar to the structure by placing its amino acids into their aligned positions and by taking into account the propensity of different amino acid types, oligopeptide fragments or residue pairs for a given physical environment. The recognition of sequence and structure is mediated by a suitable score or potential function for the evaluation of each alignment. The methods described in the literature vary

(1) in the derivation of the score function and
(2) in the alignment procedure for a single sequence with a single structure.

The technical details have been reviewed (EISENHABER et al., 1995b).

Threading has seen only a few real cases of competent application. Its efficiency in recognizing new distantly related homologues is low. Standard multiple sequence alignment methods or profile analysis are computationally cheaper and most often have the same predictive power. Threading methods will probably fail if the evolutionary divergence has removed most of the sequence similarity, if parts of the backbone have significantly moved and if secondary structural elements are inserted or deleted despite preservation of a similar basic fold pattern. This happens even within protein families (RUSSEL and BARTON, 1994). Such relative ineffectiveness is unexpected since the consideration of additional structural information should favor threading compared with a simple sequence pattern search. The crude formulation of the potential function compared even with that used in *ab initio* techniques and the NP-completeness of the alignment procedure appear responsible for this result. If threading is considered a problem of statistical hypothesis testing, it can be shown that the parameters currently used for structure description such as pairwise potentials differ if the learning set of tertiary structures is varied (S. SUNYAEV and F. EISENHABER, unpublished results).

5.2.3 Homology Modeling

The so-called modeling by homology can be applied if a protein with a given amino acid sequence is known (or supposed) to have a three-dimensional structure very similar to that of other proteins from the structural database. The unknown tertiary structure is produced by copying conserved parts of the structure (usually secondary structural elements) and fitting loop regions relying on the construction principles of close packing and hydrophobic-hydrophilic discrimination at the protein surface. The algorithm of homology modeling involves the following principal steps:

(1) Structurally conserved regions (SCR, the "tertiary template") are found on the basis of 3D structural comparisons and/or multiple sequence alignments within the protein family.
(2) The tertiary template (set of spatial positions of residues) must be aligned with the amino acid sequence that is a putative member of the same family. This step represents usually a multiple alignment with other sequences of the family or a profile analysis.

(3) Given the 3D-1D alignment, the new backbone of the protein being modeled is constructed. *Ab initio* modeling techniques are applied for constructing the loops which are usually the structurally variable regions (SVR).
(4) The conformations of the side chains anchored at the new backbone are placed with *ab initio* or knowledge-based methods.
(5) The structure proposal is finally subjected to several cycles of energy refinement and the check of verification criteria (control of strain, packing, and solvation energy based on stereochemical knowledge and tertiary structure database statistics).

The simple procedures in the early years requiring repeated human intervention have been gradually replaced by more sophisticated and generally automated techniques. In fact, there is not so much to predict. Since the fold is known, only local structural details need be tuned to comply with energetic and/or database criteria.

Another, conceptionally different approach is based on distance geometry and related algorithms. In this perspective, the tertiary template restrictions are translated into distance restraints which are used as input for distance geometry programs (HAVEL and SNOW, 1991; TAYLOR, 1993; SALI and BLUNDELL, 1993). This technique allows to integrate a variety of experimental information that can be used to formulate conformational restraints (SALI, 1995). The number of distance restraints sufficient for reproducing the protein structure correctly was also studied thoroughly (YCAS, 1990; OSHIRO et al., 1991; SIBBALD, 1995).

Technical details of homology modeling have been reviewed in great detail by JOHNSON et al. (1994) and EISENHABER et al. (1995b). The modeling error depends largely on the sequence identity between the query sequence and the known protein. If it is 40% or more, about 90% of the backbone atoms can be expected at a root-meansquare deviation of ≈ 1 Å. Side chain placement is usually worse. Below 40% sequence identity, misalignments with the target sequence become a major problem as well as the positioning of large structurally variable regions. As a result, the error rate increases drastically.

Site directed mutagenesis aimed at changing physical and chemical properties of proteins (e.g., engineering enhanced thermostability) is a specific application for homology modeling methods since, as a rule, only a few amino acids are changed. The conformational space to be searched is, therefore, not very large and enumeration techniques can be applied. With a well refined, closely homologous structure as a starting point, the model can achieve accuracies in the range of 1 Å (VRIEND and EIJSINK, 1993; DE FILLIPIS et al., 1994).

5.2.4 Prediction of Solvent Accessibility

The solvent accessibility of residues is a major tertiary property. The knowledge which residues form the protein core and which residues are located at the surface significantly reduces the possible conformations accessible for a query sequence. YCAS (1990) has estimated the distance of an amino acid residue from the protein midpoint from its hydrophobicity.

Methods similar to those developed for secondary structure prediction and based on multiple sequence alignments have been applied to this problem. Neural network approaches to the prediction of amino acid accessibility have been described (HOLBROOK et al., 1990; ROST and SANDER, 1994a). Another method is based on environment specific amino acid substitution tables (WAKO and BLUNDELL, 1994b). Because of the lower conservation of accessibility within protein families (RUSSEL and BARTON, 1994), the improvement in prediction accuracy from the use of multiply aligned sequences is not as large as in the case of secondary structure prediction, and the correlation between the predicted and observed accessibilities is only in the range of 0.36–0.77 for different sets of sequences (ROST and SANDER, 1994a).

6 Quarternary Structures of Proteins

6.1 Phenomenology of Quarternary Structural Features

Many proteins exist in form of complexes of several polypeptide chains (called subunits or protomers), since the regulation of their function probably requires many types of interactions and binding sites. The subunits may be identical or different in sequence. The protein may be dimeric, trimeric or even a higher order aggregate, though dimers and tetramers are the most frequent combinations (JONES and THORNTON, 1996). Generally, each subunit is expected to fold into an apparently independent tertiary structure and to have its own hydrophobic core. This is not always the case; e.g., two polypeptide chains are intimately intertwined in the dimeric *trp*-repressor and in the *met*-aporepressor. The quarternary structure as a higher organizational level introduces its own requirements on the tertiary structures which may again result in conformational strain. For example, metabolic energy was found to be necessary for accurate folding, for correct disulphide bond formation and for maintaining influenza hemagglutinin in its oligomerization-competent state (BRAAKMAN et al., 1992). Sometimes, the quarternary structure is not even unique and depends on pH and salt concentration (HUANG et al., 1996).

It is known that a significant part of the subunit surface in multimeric proteins and complexes is shielded from contact with the solvent (ARGOS, 1988; JANIN et al., 1988; MILLER, 1989; JANIN and CHOTHIA, 1990). The typical surface buried by one partner in a subunit contact is about 600–1000 Å^2 with 55–70% non-polar (JANIN and CHOTHIA, 1990). The interfaces are generally more similar to the interior of proteins than to water-exposed surfaces and involve often large hydrophobic surface regions.

Contacting surfaces between subunits show a high degree of geometrical complementarity (on the level of van der Waals or molecular surfaces) and are closely packed, although intersubunit (and interdomain) cavities (packing defects) are commonly larger than inside single-domain proteins (HUBBARD and ARGOS, 1994). In some cases, these packing defects have a functional role and allow relative motions of domains and, probably, also subunits (HUBBARD and ARGOS, 1996). The types of interactions at protein–protein interfaces have been a subject of detailed investigations (JONES and THORNTON, 1996), and a database of protein interfaces is also available (TSAI et al., 1996). Complementarity between docking partners has also been observed with respect to the electrostatic potential energy (HONIG and YANG, 1995).

The hydration shell structure changes during ligand association. Bound water molecules have to be removed from the interface before the macromolecular contact can happen. This effect is responsible for repulsion at a distance of about 10 Å between the docking partners (LEICKBAND et al., 1994).

6.2 Prediction of Protein–Protein Docking

The prediction of protein–protein docking is one aspect of the general problem of ligand binding by proteins. The various docking algorithms proposed in the literature try to utilize the properties of docking complexes described in the previous section and can be classified as follows:

(1) shape complementarity based techniques,
(2) approaches using solvation properties of interfaces, and
(3) methods developed for *ab initio* structure simulation.

The first group of algorithms puts major emphasis on close packing at the subunit interface. Both molecules are considered as rigid bodies and the level of complementarity at different mutual orientations is computed in a systematic or heuristic manner (LASKOWSKI et al., 1996; SOBOLEV et al., 1996). Alternatively, sections of the surface with

pronounced shape complementarity can be searched (CONNOLLY, 1992).

Since the hydrophobic effect may be considered a driving force in protein association (DILL, 1990), it is desirable to elaborate techniques for the consideration of solvation in protein aggregation (KORN and BURNETT, 1991; YOUNG et al., 1994; COVELL et al., 1994; JACKSON and STERNBERG, 1995). Besides purely surface oriented algorithms aiming at burying as many hydrophobic patches as possible and making polar groups accessible to solvent (NAUCHITEL et al., 1995), electrostatic energy calculations have been applied (JACKSON and STERNBERG, 1995; WENG et al., 1996). Since the evaluation of the Poisson-Boltzmann equation is computationally time-consuming, effective charges for each of the macromolecules are proposed. The volume of the docking partner having also a low dielectric permittivity is ignored at the initial calculation (GABDOULLINE and WADE, 1996).

Whereas *ab initio* simulation techniques with full atomic detail and variable side chain conformations are extremely computer-time consuming if applied to docking problems (TOTROV and ABAGYAN, 1994), low-resolution studies relying even only on Cα-atom positions are already sufficient to predict the correct side of contact between subunits (VAKSER, 1996a, b).

The crystal structure of the complex of TEM-1 β-lactamase (262 residues) with one of its inhibitors (BLIP, 165 residues) was used as a large-scale test for various docking algorithms (STRYNADKA et al., 1996). The structures of both individual molecules were made known to the researchers. It is remarkable that all algorithms produced a solution with the correct overall mode of BLIP binding to the TEM-1 β-lactamase (association at the active site of the enzyme). At the same time, even a search with full atomic detail and an energy function of *ab initio* techniques as attempted by TOTROV and ABAGYAN (1994) as well as more simple approaches were not able to predict details of the interface like side chain rearrangements, correct residue–residue contacts or hydrogen bonds. Thus, the gross matching of molecular shapes is sufficient to yield an approximate docking solution. Further details are outside the scope of recently developed methods mainly due to the weakness of the energy function which has to discriminate between correct and wrong types of docking complexes.

7 Concluding Remark

The field of protein structure analysis and prediction has received an exciting development during the last three decades. The major challenge, cracking the protein folding puzzle, is still unsolved. Nevertheless, a wealth of valuable information in form of sequences and structures of proteins has been accumulated in databases and many algorithms able to predict structural and functional features of proteins have been developed. The necessity to rely on prediction techniques will even grow in the future with the successful realization of genome projects since many proteins will be known only in form of amino acid sequences. A wide field of activity has opened for applied research of practitioners who attempt the selection and modification of proteins for medical or biotechnological applications.

Acknowledgement

The authors thank JOACHIM SELBIG (GMD St. Augustin-Bonn) for critical reading of the manuscript.

8 References

ABOLA, E. E., BERNSTEIN, F. C., BRYANT, S. H., KOETZLE, T. F., WENG, J. (1987), Protein data bank, in: *Crystallographic Databases – Information Content, Software Systems, Scientific Applications* (ALLEN, F. H., BERGERHOFF, G., SIEVERS, R., Eds.), pp. 107–132. Bonn, Cambridge, Chester: Data Commission of the International Union of Crystallography.

ADZHUBEI, A. A., STERNBERG, M. J. E. (1993), Left-handed polyproline II helices commonly occur in globular protein, *J. Mol. Biol.* **229**, 472–493.

ADZHUBEI, A. A., STERNBERG, M. J. E. (1994), Conservation of polyproline II helices in homologous proteins: implications for structure prediction by model building, *Protein Sci.* **3**, 2395–2410.

ADZHUBEI, A. A., LAUGHTON, C. A., NEIDLE, S. (1995), An approach to protein homology modelling based on an ensemble of NMR structures: application to the Sox-5 HMG-box protein, *Protein Eng.* **8**, 615–625.

ALEXANDROV, N. N., GO, N. (1994), Biological meaning, statistical significance, and classification of local spatial similarities in non-homologous proteins, *Protein Sci.* **3**, 866–875.

ALTSCHUL, S., BOGUSKI, M., GISH, W., WOOTTON, J. C. (1994), Issues in searching molecular sequence databases, *Nature Genetics* **6**, 119–129.

ANFINSEN, C. B. (1973), Principles that govern the folding of protein chains, *Science* **181**, 223–230.

ARGOS, P. (1988), An investigation of protein subunit and domain interfaces, *Protein Eng.* **2**, 101–113.

BACHAR, O., FISCHER, D., NUSSINOV, R., WOLFSON, H. (1993), A computer vision based technique for 3-D sequence-independent structural comparison of proteins, *Protein Eng.* **6**, 279–288.

BAIROCH, A., BUCHER, P. (1994), PROSITE: recent developments, *Nucleic Acids Res.* **22**, 3583–3859.

BAKER, D., SOHL, J. L., AGARD, D. A. (1992a), A protein-folding reaction under kinetic control, *Nature* **356**, 263–265.

BAKER, D., SOHL, J. L., AGARD, D. A. (1992b), Protease Pro region required for folding is a potent inhibitor of the mature enzyme, *Proteins* **12**, 339–344.

BARLOW, D. J., THORNTON, J. M. (1988), Helix geometry in proteins, *J. Mol. Biol.* **201**, 601–619.

BARTON, G. J., STERNBERG, M. J. E. (1988), LOPAL and SCAMP: techniques for the comparison and display of protein structures, *J. Mol. Graph.* **6**, 190–196.

BERMAN, A. L., KOLKER, E., TRIFONOV, E. N. (1994), Underlying order in protein sequence organization, *Proc. Natl. Acad. Sci. USA* **91**, 4044–4047.

BIOU, V., GIBRAT, J. F., LEVIN, J. M., ROBSON, B., GARNIER, J. (1988), Secondary structure prediction: combination of three different methods, *Protein Eng.* **2**, 185–191.

BIRNEY, E., THOMPSON, J. D., GIBSON, T. J. (1996), Pairwise and Searchwise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames, *Nucleic Acids Res.* **24**, 2730–2739.

BLUNDELL, T. L., ZHU, Z.-Y. (1995), The α-helix as seen from the protein tertiary structure: a 3-D structural classification, *Biophys. Chem.* **55**, 167–184.

BORK, P. (1996), Sperm-egg binding protein or proto-oncogene? *Science* **271**, 1431–1432.

BORK, P., BAIROCH, A. (1996), Go hunting in sequence databases but watch out for the traps, *Trends Genet.* **12**, 425–427.

BORK, P., GIBSON, T. J. (1996), Applying motif and profile searches, *Methods Enzymol.* **266**, 162–184.

BORK, P., GELLERICH, J., GROTH, H., HOOFT, R., MARTIN, F. (1995), Divergent evolution of a β/α-barrel subclass: detection of numerous phosphate-binding sites by motif search, *Protein Sci.* **4**, 268–274.

BORK, P., DOWNING, A. K., KIEFER, B., CAMPBELL, I. D. (1996), Structure and distribution of modules in extracellular proteins, *Q. Rev. Biophys.* **29**, 119–167.

BÖRNER, G. V., PÄÄBO, S. (1996), Evolutionary fixation of RNA editing, *Nature* **383**, 225.

BRAAKMAN, I., HELENIUS, J., HELENIUS, A. (1992), Role of ATP and disulphide bonds during protein folding in the endoplasmic reticulum, *Nature* **356**, 260–262.

BRUNET, A. P., HUANG, E. S., HUFFINE, M. E., LOEB, J. E., WELTMAN, R. J., HECHT, M. H. (1993), The role of turns in the structure of an α-helical protein, *Nature* **364**, 355–358.

BRYANT, S. H., LAWRENCE, C. E. (1993), An empirical energy function for threading protein sequence through the folding motif, *Proteins* **16**, 92–112.

BRYSON, J. W., BETZ, S. F., LU, H. S., SUICH, D. J., ZHOU, H. X., O'NEIL, K. T., DEGRADO, W. F. (1995), Protein design: a hierarchic approach, *Science* **270**, 935–941.

BULT, C. J., WHITE, O., OLSON, G. J., ZHOU, L., FLEISCHMANN, R. D., SUTTON, G. G., BLAKE, J. A., FITZGERALD, L. M., CLAYTON, R. A., GOCAYNE, J. D., KERVELAGE, A. R., DOUGHERTY, B. A., TOMB, J.-F., ADAMS, M. D., REICH, C. I., OVERBEEK, R., KIRKNESS, E. F., WEINSTOCK, K. G., MERRICK, J. M., GLODEK, A., SCOTT, J. L., GEOGHAGEN, N. S. M., WEIDMAN, J. F., FUHRMANN, J. L., NGYUEN, D., UTTERBACK, T. R., KELLEY, J. M., PETERSON, J. D., SADOW, P. W., HANNA, M. C., COTTON, M. D., ROBERTS, K. M., HURST, M. A., KAINE, B. P., BORODOWSKY, M., KLENK, H.-P., FRASER, C. M., SMITH, H. O., WOESE, C. R., VENTER, J. C. (1996), Complete genome sequence of the methanogenic archeon, *Methanococcus jannaschii*, *Science* **273**, 1058–1073.

BURSET, M., GUIGO, R. (1996), Evaluation of gene structure prediction programs, *Genomics* **34**, 353–367.

CARREL, R. W., STEIN, P. E., FERMI, G., WARDELL, M. R. (1994), Biological implications of a 3 Å structure of dimeric antithrombin, *Structure* **2**, 257–270.

CASARI, G., ANDRADE, M. A., BORK, P., BOYLE, J., DARUVAR, A., OUZOUNIS, C., SCHNEIDER, R., TAMAMES, J., VALENCIA, A., SANDER, C. (1994), Challenging times for bioinformatics, *Nature* **376**, 647–648.

CERPA, R., COHEN, F. E., KUNTZ, I. D. (1996), Conformational switching in designed peptides: the helix/sheet transition, *Folding Design* **1**, 91–101.

CHAN, A. W. E., HUTCHINSON, E. G., HARRIS, D., THORNTON, J. M. (1993), Identification, classification, and analysis of beta-bulges in proteins, *Protein Sci.* **2**, 1574–1590.

CHELVANAYAGAM, G., ROY, G., ARGOS, P. (1994), Easy adaptation of protein structure to sequence, *Protein Eng.* **7**, 173–184.

CHOTHIA, C., LESK, A. M. (1986), The relation between the divergence of sequence and structure in proteins, *EMBO J.* **5**, 823–826.

CHOU, P. Y. (1989), Prediction of protein structural classes from amino acid composition, in: *Prediction of Protein Structure* (FASMAN, G. D., Ed.), pp. 549–586. New York: Plenum Press.

CHOU, K.-C. (1995), A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, *Proteins* **21**, 319–344.

CHOU, P. Y., FASMAN, G. (1974a), Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins, *Biochemistry* **13**, 211–222.

CHOU, P. Y., FASMAN, G. (1974b), Prediction of protein conformation, *Biochemistry* **13**, 222–245.

CHOU, K.-C., ZHANG, C.-T. (1995), Prediction of protein structural classes, *CRC Crit. Rev. Biochem. Mol. Biol.* **30**, 275–349.

COLLOCH, N., ETCHEBEST, C., THOREAU, E., HENRISSAT, B., MORNON, J.-P. (1993), Comparison of three algorithms for the assignment of secondary structure in proteins: the advantage of a consensus assignment, *Protein Eng.* **6**, 377–382.

CONNOLLY, M. L. (1983), Analytical molecular surface calculation, *J. Appl. Cryst.* **16**, 548–558.

CONNOLLY, M. L. (1992), Shape distributions of protein topography, *Biopolymers* **32**, 1215–1236.

COOPER, A. A., STEVENS, T. H. (1995), Protein splicing: self-splicing of genetically mobile elements at the protein level, *Trends Biochem. Sci.* **20**, 351–356.

COVELL, D. G., SMYTHERS, G. W., GRONENBORN, A. M., CLORE, M. G. (1994), Analysis of hydrophobocity in the a and b chemokine families and its relevance to dimerization, *Protein Sci.* **3**, 2064–2072.

CREIGHTON, T. E. (1992), Protein Folding. New York: Freeman.

DAFFNER, C., CHELVANAYAGAM, G., ARGOS, P. (1994), Structural characteristics and stabilizing principles of bent beta-strands in protein tertiary structures, *Protein Sci.* **3**, 876–882.

DE FILLIPIS, V., SANDER, C., VRIEND, G. (1994), Predicting local structural changes that result from point mutations, *Protein Eng.* **7**, 1203–1208.

DILL, K. A. (1990), Dominant forces in protein folding, *Biochemistry* **29**, 7133–7155.

DOIG, A. J., BALDWIN, R. L. (1995), N- and C-capping preferences for all 20 amino acids in α-helical peptides, *Protein Sci.* **4**, 1325–1336.

DOMBI, G. W., LAWRENCE, J. (1994), Analysis of protein transmembrane helical regions by a neural network, *Protein Sci.* **3**, 557–566.

DORAN, J. D., CAREY, P. R. (1996), α-helix dipoles and catalysis: absorption and raman spectroscopic studies of acyl cysteine proteases, *Biochemistry* **35**, 12495–12502.

DUBCHAK, I., HOLBROOK, S. R., KIM, S.-H. (1993), Prediction of protein folding class from amino acid composition, *Proteins* **16**, 79–91.

EFREMOV, R. G., VERGOTEN, G. (1996a), Recognition of transmembrane α-helical segments with environmental profiles, *Protein Eng.* **9**, 253–263.

EFREMOV, R. G., VERGOTEN, G. (1996b), Hydrophobic organization of α-helix membrane bundle in bacteriorhodopsin, *J. Protein Chem.* **15**, 63–76.

EISENHABER, F. (1996), Hydrophobic regions on protein surfaces. Derivation of the solvation energy from their area distribution in crystallographic protein structures, *Protein Sci.* **5**, 1676–1686.

EISENHABER, F., ARGOS, P. (1993), Improved strategy in analytic surface calculation for molecular systems: handling of singularities and computational efficiency, *J. Comp. Chem.* **14**, 1272–1280.

EISENHABER, F., ARGOS, P. (1996), Hydrophobic regions on protein surfaces. Definition based on hydration shell structure and a quick method for region computation, *Protein Eng.* **9**, 1121–1133.

EISENHABER, F., LIJNZAAD, P., ARGOS, P., SANDER, C., SCHARF, M. (1995a), The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and for generating dot surfaces of molecular assemblies, *J. Comp. Chem.* **16**, 273–284.

EISENHABER, F., PERSSON, B., ARGOS, P. (1995b), Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence, *CRC Crit. Rev. Biochem. Mol. Biol.* **30**, 1–94.

EISENHABER, F., IMPERIALE, F., ARGOS, P., FRÖMMEL, C. (1996a), Prediction of secondary structural content of proteins from their amino acid composition alone. I. New vector decomposition methods, *Proteins* **25**, 157–168.

EISENHABER, F., FRÖMMEL, C., ARGOS, P. (1996b), Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class, *Proteins* **25**, 169–179.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R., KIRKNESS, E. F., KERLAGE, A. R., BULT, C. J., TOMB, J.-F., DOUGHERTY, B. A., MERRICK, J. M., MCKENNEY, K., SUTTON, G., FITZHUGH, W., FIELDS, C., GOCAYNE, J. F., SCOTT, J., SHIRLEY, R., LIU, L.-I., GLODEK, A., KELLEY, J. M., WEIDMAN, J. F., PHILLIPS, C. A., SPRIGGS, T., HEDBLOM, E., COTTON, M. D., UTTERBACK, T. R., HANNA, M. C., NGUYEN, D. T., SAUDEK, D. M., BRANDON, R. C., FINE, L. D., FRITCHMAN, J. L., FUHRMANN, J. L., GEOGHAGEN, N. S. M., GNEHM, C. L., MCDONALD, L. A., SMALL, K. V., FRASER, C. M., SMITH, H. O., VENTER, J. C. (1995), Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, *Nature* **269**, 496–512.

FRAENKEL, A. S. (1993), Complexity of protein folding, *Bull. Math. Biol.* **55**, 1199–1210.

FRISHMAN, D. I., ARGOS, P. (1992), Recognition of distantly related protein sequences using conserved motifs and neural networks, *J. Mol. Biol.* **228**, 951–962.

FRISHMAN, D., ARGOS, P. (1995), Knowledge-based protein secondary structure assignment, *Proteins* **23**, 566–579.

FRISHMAN, D., ARGOS, P. (1996), Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence, *Protein Eng.* **9**, 133–142.

GABDOULLINE, R. R., WADE, R. C. (1996), Effective charges for macromolecules in solvent, *J. Phys. Chem.* **100**, 3868–3878.

GARNIER, J., LEVIN, J. M. (1991), The protein structure code: what is its present status? *Comput. Appl. Biosci.* **7**, 133–142.

GARNIER, J., OSGUTHORPE, D. J., ROBSON, B. (1978), Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *J. Mol. Biol.* **120**, 97–120.

GEOURJON, C., DELÉAGE, G. (1994), SOPM: a self-optimized method for protein secondary structure prediction, *Protein Eng.* **7**, 157–164.

GESTELAND, R. F., ATKINS, J. F. (1996), Recoding: dynamic reprogramming of translation, *Annu. Rev. Biochem.* **65**, 741–768.

GETHING, M. J., SAMBROOK, J. (1992), Protein folding in the cell, *Nature* **355**, 33–45.

GIBRAT, J.-F., GARNIER, J., ROBSON, B. (1987), Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs, *J. Mol. Biol.* **198**, 425–443.

GIBRAT, J.-F., ROBSON, B., GARNIER, J. (1991), Influence of the local amino acid sequence upon the zones of the torsional angles phi and psi adopted by residues in proteins, *Biochemistry* **30**, 1578–1586.

GODZIK, A. (1996), The structural alignment between two proteins: is there a unique answer? *Protein Sci.* **5**, 1325–1338.

GOLDSMITH, E. J., MOTTONEN, J. (1994), Serpins: the uncut version, *Structure* **2**, 241–244.

GONG, Y., ZHOU, H. X., GUO, M., KALLENBACH, N. R. (1995), Structural analysis of the N- and C-termini in a peptide with consensus sequence, *Protein Sci.* **4**, 1446–1456.

HAN, K.-K., MARTINAGE, A. (1992), Possible relationship between coding recognition of amino acid sequence motif or residue(s) and posttranslational chemical modification of proteins, *Int. J. Biochem.* **24**, 1349–1363.

HANKE, J., BECKMANN, G., BORK, P., REICH, J. G. (1996), Self-organizing hierarchic networks for pattern recognition in protein sequence, *Protein Sci.* **5**, 72–82.

HARTL, F. U. (1994), Secrets of a double-doughnut, *Nature* **371**, 557–559.

HAVEL, T., SNOW, M. E. (1991), A new method for building protein conformations from sequence alignments with homologues of known structure, *J. Mol. Biol.* **217**, 1–7.

HERINGA, J., ARGOS, P. (1993), A method to recognize distant repeats in protein sequences, *Proteins* **17**, 391–411.

HIGGINS, D., THOMPSON, J. D., GIBSON, T. J. (1996), Using CLUSTAL for multiple sequence alignment, *Methods Enzymol.* **266**, 383–402.

HOLBROOK, S. R., MUSKAL, S. M., KIM, S.-H. (1990), Predicting surface exposure of amino acids from protein sequence, *Protein Eng.* **3**, 659–665.

HOLM, L., SANDER, C. (1993), Protein structure comparison by alignment of distance matrices, *J. Mol. Biol.* **233**, 123–138.

HOLM, L., SANDER, C. (1994a), Searching protein structure databases has come of age, *Proteins* **19**, 165–173.

HOLM, L., SANDER, C. (1994b), The FSSP database of structurally aligned protein fold families, *Nucleic Acids Res.* **22**, 3600–3609.

HOLM, L., SANDER, C. (1996), Mapping the protein universe, *Science* **273**, 595–602.

HOLM, L., OUZOUNIS, C., SANDER, C., TUPAREV, G., VRIEND, G. (1992), A database of protein structure families with common folding motifs, *Protein Sci.* **1**, 1691–1698.

HONIG, B., YANG, A.-S. (1995), Free energy balance in protein folding, *Adv. Protein Chem.* **46**, 27–58.

HUANG, K., STRYNADKA, N. C. J., BERNARD, V. D., PEANASKY, R. J., JAMES, M. N. G. (1994), The molecular structure of the complex of *Ascaris* chymotrypsin/elastase inhibitor with porcine elastase, *Structure* **2**, 679–689.

HUANG, D.-B., AINSWORTH, C. F., STEVENS, F. J., SCHIFFER, M. (1996), Three quarternary structures for a single protein, *Proc. Natl. Acad. Sci. USA* **93**, 7017–7021.

HUBBARD, S. J., ARGOS, P. (1994), Cavities and packing at protein interfaces, *Protein Sci.* **3**, 2194–2206.

HUBBARD, S. J., ARGOS, P. (1995), Detection of internal cavities in globular proteins, *Protein Eng.* **8**, 1011–1015.

HUBBARD, S. J., ARGOS, P. (1996), A functional role for protein cavities in domain:domain motions, *J. Mol. Biol.* **261**, 289–300.

HUBBARD, S. J., GROSS, K.-H., ARGOS, P. (1994), Intramolecular cavities in globular proteins, *Protein Eng.* **7**, 613–626.

HUCHO, F., GÖRNE-TSCHELNOKOW, U., STRECKER, A. (1994), β-structure in the membrane-spanning part of the nicotinic acetylcholine receptor (or how helical are transmembrane helices?), *Trends Biochem. Sci.* **19**, 383–387.

ISLAM, S. A., LUO, J., STERNBERG, M. J. E. (1995), Identification and analysis of domains in proteins, *Protein Eng.* **8**, 513–525.

JACKSON, R. M., STERNBERG, M. J. E. (1995), A continuum model for protein-protein interactions: application to the docking problem, *J. Mol. Biol.* **250**, 258–275.

JANIN, J., CHOTHIA, C. (1990), The structure of protein-protein recognition site, *J. Biol. Chem.* **265**, 1627–1630.

JANIN, J., MILLER, S., CHOTHIA, C. (1988), Surface, subunit interfaces and interior of oligomeric proteins, *J. Mol. Biol.* **204**, 155–164.

JOHNSON, W. C., JR. (1990), Protein secondary structure and circular dichroism: a practical guide, *Proteins* **7**, 205–214.

JOHNSON, M. S., SRINIVASAN, N., SOWDHAMINI, R., BLUNDELL, T. L. (1994), Knowledge-based protein modelling, *CRC Crit. Rev. Biochem. Mol. Biol.* **29**, 1–68.

JONES, S., THORNTON, J. M. (1996), Principles of protein-protein interactions, *Proc. Natl. Acad. Sci. USA*, **93**, 13–20.

JUFFER, A., EISENHABER, F., HUBBARD, S. J., WALTHER, D., ARGOS, P. (1995), Comparison of atomic solvation parameter sets: applicability and limitations in protein folding and binding, *Protein Sci.* **4**, 2499–2509.

JUFFER, A. H., EISENHABER, F., HUBBARD, S. J., WALTHER, D., ARGOS, P. (1996), Erratum: Comparison of atomic solvation parameter sets: applicability and limitations in protein folding and binding, *Protein Sci.* **5**, 1748–1749.

KABSCH, W., SANDER, C. (1983), Dictionary of protein secondary structures: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**, 2577–2637.

KARPLUS, P. A. (1996), Experimentally observed conformation-dependent geometry and hidden strain in proteins, *Protein Sci.* **5**, 1406–1420.

KLEIN, P., DELISI, C. (1986), Prediction of protein structural class from the amino acid sequence, *Biopolymers* **25**, 1659–1672.

KNELLER, D. G., COHEN, F. E., LANGRIDGE, R. (1990), Improvements in protein secondary structure prediction by enhanced neural networks, *J. Mol. Biol.* **214**, 171–182.

KOLKER, E., TRIFONOV, E. N. (1995), Periodic recurrence of methionines: fossil of gene fusion? *Proc. Natl. Acad. Sci. USA* **92**, 557–560.

KOONIN, E. V., MUSHEGIAN, A. R., BORK, P. (1996), Non-orthologous gene displacement, *Trends Genet.* **12**, 334–336.

KORN, A. P., BURNETT, R. M. (1991), Distribution and complementarity of hydropathy in multisubunit proteins, *Proteins* **9**, 37–55.

KROGH, A., BROWN, M., MIAN, I. S., SJÖLANDER, K., HAUSSLER, D. (1994), Hidden Markov models in computational biology, *J. Mol. Biol.* **235**, 1501–1531.

KÜHLBRANDT, W., WANG, D. N., FUJIYOSHI, Y. (1994), Atomic model of plant light-harvesting complex by electron crystallography, *Nature* **367**, 614–621.

KWASIGROCH, J.-M., CHOMILIER, J., MORNON, J.-P. (1996), A global taxonomy of loops in globular proteins, *J. Mol. Biol.* **259**, 855–872.

LASKOWSKI, R. A., THORNTON, J. M., HUMBLET, C., SINGH, J. (1996), X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins, *J. Mol. Biol.* **259**, 175–201.

LECKBAND, D. E., SCHMITT, F.-J., ISRAELACHVILI, J. N., KNOLL, W. (1994), Direct force measurements of specific and nonspecific protein interactions, *Biochemistry* **33**, 4611–4624.

LEE, B., RICHARDS, F. M. (1971), The interpretation of protein structures: estimation of static accessibility, *J. Mol. Biol.* **55**, 379–400.

LESK, A. M., CHOTHIA, C. (1980), How different amino acid sequences determine similar protein

structures: the structure and evolutionary dynamics of the globins, *J. Mol. Biol.* **136**, 225–270.

LESSEL, U., SCHOMBURG, D. (1994), Similarities between protein 3-D structures, *Protein Eng.* **7**, 1175–1187.

LEVIN, J. M., GARNIER, J. (1988), Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool, *Biochim. Biophys. Acta* **955**, 283–295.

LEVIN, J. M., PASCARELLA, S., ARGOS, P., GARNIER, J. (1993), Quantification of secondary structure prediction improvement using multiple alignment, *Protein Eng.* **6**, 849–854.

LEVINTHAL, C. (1968), Are there pathways for protein folding? *J. Chem. Phys.* **65**, 44–45.

LEVITT, M., CHOTHIA, C. (1976), Structural patterns in globular proteins, *Nature* **261**, 552–558.

LIANG, C., MISLOW, K. (1994a), Topological chirality of proteins, *J. Am. Chem. Soc.* **116**, 3588–3592.

LIANG, C., MISLOW, K. (1994b), Topological features of chorionic gonadotropin, *Biopolymers* **35**, 343–345.

LUO, Y., LAI, L., XU, X., TANG, Y. (1993), Defining topological equivalences in protein structures by means of a dynamic programming algorithm, *Protein Eng.* **6**, 373–376.

LUPAS, A. (1996), Coiled coils: new structures and new functions, *Trends Biochem. Sci.* **21**, 375–382.

LUPAS, A., VAN DYKE, M., STOCK, J. (1991), Predicting coiled coils from protein sequence, *Science* **252**, 1162–1164.

MAKEEV, V. J., TUMANYAN, V. G. (1996), Search of periodicities in primary structure of biopolymers: a general Fourier approach, *Comput. Appl. Biosci.* **12**, 49–54.

MANSFIELD, M. L. (1994), Are there knots in proteins? *Nature Struct. Biol.* **1**, 213–214.

MAO, B. (1993), Topological chirality of proteins, *Protein Sci.* **2**, 1057-1059.

MARTIN, A. C. R., TODA, K., STIRK, H. J., THORNTON, J. M. (1995), Long loops in proteins, *Protein Eng.* **8**, 1093–1101.

MAY, A. C. W., JOHNSON, M. S. (1994), Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization, *Protein Eng.* **7**, 475–485.

MAYO, K. H., ILYINA, E., PARK, H. (1996), A recipe for designing water-soluble, β-sheet-forming peptides, *Protein Sci.* **5**, 1301–1315.

MCLACHLAN, A. D. (1983), Gene duplications in the structural evolution of chymotrypsin, *J. Mol. Biol.* **128**, 49–79.

MEHTA, K. P., HERINGE, J., ARGOS, P. (1995), A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%, *Protein Sci.* **4**, 2517–2525.

MILLER, S. (1989), The structure of interfaces between subunits of dimeric and tetrameric proteins, *Protein Eng.* **3**, 77–83.

MILLHAUSER, G. L. (1995), Views of helical peptides: a proposal for the position of 3_{10}-helix along the thermodynamic folding pathway, *Biochemistry* **34**, 3873–3877.

MOTTONEN, J., STRAND, A., SYMERSKI, J., SWEET, R. M., DANLEY, R. E., GEOGHEGEN, K. F., GERARD, R. D., GOLDSMITH, E. J. (1992), Structural basis of latency in plasminogen activator inhibitor-1, *Nature* **355**, 270–273.

MUGGLETON, S., KING, R. D., STERNBERG, M. J. E. (1992), Protein secondary structure prediction using logic-based machine learning, *Protein Eng.* **5**, 647–657.

MUGGLETON, S., KING, R. D., STERNBERG, M. J. E. (1993), Corrigenda: protein secondary structure prediction using logic-based machine learning, *Protein Eng.* **6**, 549.

MURZIN, A. G., LESK, A. M., CHOTHIA, C. (1994a), Principles determining the structure of β-sheet barrels in proteins. I. A theoretical analysis, *J. Mol. Biol.* **236**, 1369–1381.

MURZIN, A. G., LESK, A. M., CHOTHIA, C. (1994b), Principles determining the structure of β-sheet barrels in proteins. II. The observed structures, *J. Mol. Biol.* **236**, 1382–1400.

NAKASHIMA, H., NISHIKAWA, K., OOI, T. (1986), The folding type of a protein is relevant to the amino acid composition, *J. Biochem.* **99**, 153–162.

NAUCHITEL, V., VILLAVERDE, M. C., SUSSMAN, F. (1995), Solvent accessibility as a predictive tool for the free energy of inhibitor binding to the HIV-1 protease, *Protein Sci.* **4**, 1356–1364.

NGO, T. J., MARKS, J. (1992), Computational complexity of a problem in molecular structure prediction, *Protein Eng.* **5**, 313–321.

NICHOLS, W. L., ROSE, G. D., TEN EYCK, L. F., ZIMM, B. H. (1995), Rigid domains in proteins: an algorithmic approach to their identification, *Proteins* **23**, 38–48.

NISHIKAWA, K., OOI, T. (1982), Correlation of the amino acid composition of a protein to its structural and biological characters, *J. Biochem.* **91**, 1821–1824.

NISHIKAWA, K., KUBOTA, Y., OOI, T. (1983a), Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types, *J. Biochem.* **94**, 997–1007.

NISHIKAWA, K., KUBOTA, Y., OOI, T. (1983b), Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution, *J. Biochem.* **94**, 981–995.

ORENGO, C. A., TAYLOR, W. R. (1990), A rapid method of protein structure alignment, *J. Theor. Biol.* **147**, 517–551.

ORENGO, C. A., FLORES, T. P., TAYLOR, W. R., THORNTON, J. M. (1993), Identification and classification of protein fold families, *Protein Eng.* **6**, 485–500.

OSAWA, S., JUKES, T. H., WATANABE, K., MUTO, A. (1992), Recent evidence for evolution of the genetic code, *Microbiol. Rev.* **56**, 229–264.

OSHIRO, C. M., THOMASON, J., KUNTZ, I. D. (1991), The effects of limited input distance constraints upon the distance geometry algorithm, *Biopolymers* **31**, 1049–1064.

PASCARELLA, S., ARGOS, P. (1992), A data bank merging related protein structures and sequences, *Protein Eng.* **5**, 121–137.

PASTORE, A., SAUDEK, V., RAMPONI, G., WILLIAMS, R. J. P. (1992), Three-dimensional structure of acylphosphatase, *J. Mol. Biol.* **224**, 427–440.

PERCZEL, A., HOLLÓSI, M., TUSNÁDY, G., FASMAN, G. D. (1991), Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins, *Protein Eng.* **4**, 669–679.

PERSSON, B., ARGOS, P. (1994), Prediction of transmembrane regions in proteins utilising multiple sequence alignments, *J. Mol. Biol.* **237**, 182–192.

PERSSON, B., ARGOS, P. (1996), Topology prediction of membrane proteins, *Protein Sci.* **5**, 363–371.

PICOT, D., LOLL, P. J., GARAVITO, M. (1994), The X-ray crystal structure of the membrane protein prostaglandin H_2 synthase-1, *Nature* **367**, 243–249.

PRESTA, L. G., ROSE, G. D. (1988), Helix signals in proteins, *Science* **240**, 1632–1641.

QIAN, H., CHAN, S. I. (1996), Interactions between a helical residue and tertiary structures: helix propensities in small peptides and in native proteins, *J. Mol. Biol.* **261**, 279–288.

RAJASHANKAR, K. R., RAMAKUMAR, S. (1996), p-turns in proteins and peptides: classification, conformation, occurrence, hydration and sequence, *Protein Sci.* **5**, 932–946.

RECZKO, M., BOHR, H. (1994), The DEF data base of sequence based protein fold class predictions, *Nucleic Acids Res.* **22**, 3616–3619.

RECZKO, M., BOHR, H., SUBRAMANIAM, S., PAMIGIGHANTAM, S., HATZIGEORGIOU, A. (1994), Fold-class prediction by neural networks, in: *Protein Structure by Distance Analysis* (BOHR, H., BRUNAK, S., Eds.), pp. 277–286. Amsterdam, Tokyo: IOS Press, Ohmsha.

RESH, M. D. (1994), Myristylation and palmitylation of Src family members: the fats of the matter, *Cell* **76**, 411–413.

RICHARDS, F. M., KUNDROT, C. E. (1988), Identification of structural motifs from protein coordinate data: secondary and first-level supersecondary structure, *Proteins* **3**, 71–84.

RICHARDS, F. M., LIM, W. A. (1994), An analysis of packing in the protein folding problem, *Q. Rev. Biophys.* **26**, 423–498.

RIEK, R., HORNEMANN, S., WIDER, G., BILLETER, M., GLOCKSHUBER, R., WÜTHRICH, K. (1994), NMR structure of the mouse prion protein domain PrP (121–231), *Nature* **382**, 180–182.

ROST, B., SANDER, C. (1993a), Secondary structure prediction of all-helical proteins in two states, *Protein Eng.* **6**, 831–836.

ROST, B., SANDER, C. (1993b), Prediction of protein secondary structure at better than 70% accuracy, *J. Mol. Biol.* **232**, 584–599.

ROST, B., SANDER, C. (1994a), Combining evolutionary information and neural networks to predict protein secondary structure, *Proteins* **19**, 55–72.

ROST, B., SANDER, C. (1994b), Conservation and prediction of solvent accessibility in protein families, *Proteins* **20**, 216–226.

ROST, B., CASADIO, R., FARISELLI, P., SANDER, C. (1995), Transmembrane helices predicted at 95% accuracy, *Protein Sci.* **4**, 521–533.

ROST, B., FARISELLI, P., CASADIO, R. (1996), Topology prediction for helical transmembrane proteins at 86% accuracy, *Protein Sci.* **5**, 1704–1718.

RUSSEL, R. B., BARTON, G. J. (1994), Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts, secondary structure and accessibility, *J. Mol. Biol.* **244**, 332–350.

SALAMOV, A. A., SOLOVYEV, V. V. (1995), Prediction of protein secondary structure by combining nearest neighbor algorithms and multiple sequence alignments, *J. Mol. Biol.* **247**, 11–15.

SALI, A. (1995), Modelling mutations and homologous proteins, *Curr. Opin. Struct. Biol.* **6**, 437–451.

SALI, A., BLUNDELL, T. L. (1990), Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming, *J. Mol. Biol.* **212**, 403–428.

SALI, A., BLUNDELL, T. L. (1993), Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.* **234**, 779–815.

SALI, A., SHAKHNOVICH, E., KARPLUS, M. (1994), Kinetics of protein folding. A lattice model study of the requirements for folding to the native state, *J. Mol. Biol.* **235**, 1614–1636.

SAMATEY, F. A., XU, C., POPOT, J.-L. (1995), On the distribution of amino acid residues in transmembrane α-helix bundles, *Proc. Natl. Acad. Sci. USA* **92**, 4577–4581.

SCHRAUBER, H., EISENHABER, F., ARGOS, P. (1993), Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins, *J. Mol. Biol.* **230**, 592–612.

SERRANO, L., SANCHO, J., HIRSHBERG, M., FERSHT, A. R. (1992), α-Helix stability in proteins. I. Empirical correlations concerning substitutions of side chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces, *J. Mol. Biol.* **227**, 544–559.

SHCHERBAK, V. I. (1988), To co-operative symmetry of the genetic code, *J. Theor. Biol.* **132**, 121–124.

SHCHERBAK, V. I. (1989), The "START" and the "STOP" of the genetic code: why exactly ATG and TAG. TAA? *J. Theor. Biol.* **139**, 283–286.

SIBANDA, B. L., THORNTON, J. M. (1993), Accomodating sequence changes in b-hairpins in proteins, *J. Mol. Biol.* **229**, 428–447.

SIBBALD, P. R. (1995), Deducing protein structures using logic programming: exploiting minimum data of diverse types, *J. Theor. Biol.* **173**, 361–375.

SIDDIQUI, A. S., BARTON, G. J. (1995), Continuous and discontinuous domains: an algorithm for the automatic generation of reliable domain definitions, *Protein Sci.* **4**, 872–884.

SKLENAR, H., ETCHEBEST, C., LAVERY, R. (1989), Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis, *Proteins* **6**, 46–60.

SOBOLEV, V., WADE, R. C., VRIEND, G., EDELMAN, M. (1996), Molecular docking using surface complementarity, *Proteins* **25**, 120–129.

SOWDHAMINI, R., BLUNDELL, T. L. (1995), An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins, *Protein Sci.* **4**, 506–520.

SOWDHAMINI, R., RUFINO, S. D., BLUNDELL, T. L. (1996), A database of globular protein structural domains: clustering of representative family members into similar folds, *Folding Design* **1**, 209–220.

SREERAMA, N., WOODY, R. W. (1994), Protein secondary structure from circular dichroism spectroscopy, *J. Mol. Biol.* **242**, 497–507.

STADTMAN, T. C. (1996), Selenocysteine, *Annu. Rev. Biochem.* **65**, 83–100.

STEVENS, R. C., GOUAUX, J. E., LIPSCOMB, W. N. (1990), Structural consequences of effector binding to the T state of aspartate carbamoyltransferase: crystal structure of the unligated and ATP- and CTP-complexed enzymes at 2.6-Å resolution, *Biochemistry* **29**, 7691–7701.

STRYNADKA, N. C. J., EISENSTEIN, M., KATCHALSKI-KATZIR, E., SHOICHET, B. K., KUNTZ, I. D., ABAGYAN, R., TOTROV, M., JANIN, J., CHERFILS, J., ZIMMERMANN, F., OLSON, A., DUNCAN, B., RAO, M., JACKSON, R., STERNBERG, M., JAMES, M. N. G. (1996), Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1β-lactamase, *Nature Struct. Biol.* **3**, 233–239.

SWINDELLS, M. B. (1995a), A procedure for the automatic determination of hydrophobic cores in protein structures, *Protein Sci.* **4**, 93–102.

SWINDELLS, M. B. (1995b), A procedure for detecting structural domains in proteins, *Protein Sci.* **4**, 103–112.

TAKUSAGAWA, F., KAMITORI, S. (1996), A real knot in protein, *J. Am. Chem. Soc.* **118**, 8945–8946.

TATUSOV, R. L., MUSHEGIAN, A. R., BORK, P., BROWN, N. P., HAYES, W. S., BORODOVSKY, M., RUDD, K. E., KOONIN, E. V. (1996), Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*, *Curr. Biol.* **6**, 279–291.

TAYLOR, W. R. (1993), Protein fold refinement: building models from idealized folds using motif constraints and multiple sequence data, *Protein Eng.* **6**, 593–604.

TAYLOR, W. R., ORENGO, C. A. (1989), Protein structure alignment, *J. Mol. Biol.* **208**, 1–22.

TOTROV, M., ABAGYAN, R. (1994), Detailed *ab initio* prediction of the lysozyme-antibody complex with 1.6 Å accuracy, *Nature Struct. Biol.* **1**, 259–263.

TSAI, C.-J., LIN, S. L., WOLFSON, H. J., NUSSINOV, R. (1996), Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarity and differences, *CRC Crit. Rev. Biochem. Mol. Biol.* **31**, 127–152.

UNGER, R., MOULT, J. (1994), Finding the lowest free energy conformation is an NP-hard problem: proof and implications, *Bull. Math. Biol.* **55**, 1183–1198.

VAKSER, I. A. (1996a), Main-chain complementarity in protein-protein recognition, *Protein Eng.* **9**, 741–744.

VAKSER, I. A. (1996b), Low-resolution docking: prediction of complexes for underdetermined structures, *Biopolymers* **39**, 455–464.

VAN HEEL, M. (1992), A new family of powerful multivariate statistical sequence analysis (MSSA) techniques, *J. Mol. Biol.* **216**, 877–887.

VON HEIJNE, G. (1986), The distribution of positively charged residues in bacterial inner membrane proteins correlates with transmembrane topology, *EMBO J.* **5**, 3021–3027.

VON HEIJNE, G. (1995), Membrane protein assembly: rules of the game, *Bioessays* **17**, 25–30.

VRIEND, G., EIJSINK, V. (1993), Prediction and analysis of structure, stability and unfolding of thermolysin-like proteases, *J. Comput. Aided Mol. Des.* **7**, 367–396.

VRIEND, G., SANDER, C. (1991), Detection of common three-dimensional substructures in proteins, *Proteins* **11**, 52–58.

VTYURIN, N. (1993), The role of tight packing of hydrophobic groups in β-structure, *Proteins* **15**, 62–70.

VTYURIN, N., PANOV, V. (1995), Packing constraints of hydrophobic side chains in $(\alpha/\beta)_8$ barrels, *Proteins* **21**, 256–260.

WAKO, H., BLUNDELL, T. L. (1994a), Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologues proteins. I. Solvent accessibility classes, *J. Mol. Biol.* **238**, 682–692.

WAKO, H., BLUNDELL, T. L. (1994b), Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures, *J. Mol. Biol.* **238**, 693–708.

WALTHER, D., EISENHABER, F., ARGOS, P. (1996), Principles of helix-helix packing in proteins: the helical lattice superposition model, *J. Mol. Biol.* **255**, 536–553.

WEISS, M. S., SCHULZ, G. E. (1992), Structure of porin refined at 1.8 Å resolutiom, *J. Mol. Biol.* **227**, 493–509.

WENG, Z., VAJDA, S., DELISI, C. (1996), Prediction of protein complexes using empirical free energy functions, *Protein Sci.* **5**, 614–626.

WILMOT, C. M., THORNTON, J. M. (1990), β-Turns and their distortions: a proposed new nomenclature, *Protein Eng.* **3**, 479–493.

WOOTTON, J. C. (1994), Sequences with 'unusual' amino acid compositions, *Curr. Opin. Struct. Biol.* **4**, 413–421.

YANG, J. T. (1996), Prediction of protein secondary structure from amino acid sequence, *J. Protein Chem.* **15**, 185–191.

YCAS, M. (1990), Computing tertiary structures of proteins, *J. Protein Chem.* **9**, 177–200.

YEE, D. P., DILL, K. A. (1993), Families and the structural relatedness among globular proteins, *Protein Sci.* **2**, 884–899.

YOUNG, L., JERNIGAN, R. L., COVELL, D. G. (1994), A role for surface hydrophobicity in protein-protein recognition, *Protein Sci.* **3**, 717–729.

ZHANG, C.-T., CHOU, K.-C. (1995), An eigenvalue-eigenvector approach to predicting protein folding types, *J. Protein Chem.* **14**, 309–326.

ZHU, Z.-Y. (1995), A new approach to the evaluation of protein secondary structure predictions at the level of elements of secondary structure, *Protein Eng.* **8**, 103–108.

ZUKER, M., SOMORJAI, R. L. (1989), The alignment of protein structures in three dimensions, *Bull. Math. Biol.* **51**, 55–78.