# CHAPTER 3

# Prediction of Protein Function:
## Two Basic Concepts and One Practical Recipe

**Frank Eisenhaber***

## Abstract

The analysis of uncharacterized biomolecular sequences obtained as a result of genetic screens, expression profile studies, etc. is a standard task in a life science research environment. The understanding of protein function is typically the main difficulty. This chapter intends to give practical advise to students and researchers that have only introductory knowledge in the field of protein sequence analysis.

Applicable theoretical approaches range from (1) textual analyses, interpretation in terms of patterns of physical properties of amino acid side chains and (2) the extrapolation of empirically established relationships between local sequence motifs with known structural and functional properties to the collection of sequence segment families with sequence distance metrics and protein function derivation with annotation transfer (concept of homologous families). Here, the impact of different techniques for the biological interpretation of targets is discussed from the practitioner's point of view and illustrated with examples from recent research reports. Although sequence similarity searching techniques are the most powerful instruments for the analysis of high-complexity regions, other techniques can supply important additional evaluations including the assessment of applicability of the sequence homology concept for the given target segment.

## Introduction

The genome has become the integrating principle for the various fields of biology and the clarification of pathways that lead to the realization of genome information into phenotypes under varying environmental conditions has become the central task for life sciences. As a first step, it is critical to understand the function of genes at least in qualitative terms; i.e., to name the molecular function of encoded proteins and to uncover the topology of interactions of networks involving them. Given that, currently, the molecular function of at least two thirds of all genes in completely sequenced eukaryote genomes remains more or less clouded, this would represent a dramatic progress. At the same time, it should be noted that real theoretical predictability of biological systems above the level of educated guesses (for example, for drug engineering) typically requires quantitative characterization of gene and protein activity and modeling of biological networks, which will be, in most cases, not a matter of the coming handful of years. Possibly, this is even an optimistic assessment.

With the central role of the genome in the functioning of biological systems, it is not surprising that experimental screens for genes relevant for the processes investigated are a standard approach in today's experimental biology; for example, expression profiling with DNA

*Frank Eisenhaber—Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, A-1030 Vienna, Republic Austria. Email: Frank.Eisenhaber@imp.univie.ac.at

*Discovering Biomolecular Mechanisms with Computational Biology*,
edited by Frank Eisenhaber. 

microarrays, yeast two hybrid screens, etc. If the biological phenomenon has not been well described in already published research, the screens lead typically to sequence tags of yet uncharacterized genes. Their sequence information has then to be interpreted in functional terms within the given physiological context. Stereotypically, the sequence is submitted to a similarity search in sequence databases. As a rule, the amount of insight produced by such a direct approach is indirectly proportional to the novelty of the gene target. In this tractate, we want to discuss the few fundamental principles that underlie state-of-the-art protein sequence analysis approaches. Then, we propose a general recipe for the practitioner who looks for research hints in his target sequences. We will give interpretation guides for sequence analytic findings and emphasize limitations where appropriate.

## The Beginning: Deriving the Protein Sequence and the Definition of Protein Function

Typically, the starting point is a partial nucleic acid sequence representing a piece of mRNA. Whereas the experimental extension of the sequence to a full transcript was mandatory before the era of large-scale sequencing, this step can often be avoided now. In this case, it is necessary to find (1) a longer expressed sequence tag (EST), (2) a cluster of ESTs with a consensus sequence or, luckily, (3) a complete cDNA in the databases that obviously contains the reliably sequenced segment of the partial sequence obtained in the screen. The completeness of the putative transcript sequence can be investigated by mapping relevant ESTs onto the genome sequence. Especially in the case of incomplete transcripts involving only 3' untranslated regions, searching for the closest predicted gene upstream in the genome might yield the desired gene.[1,2] Searches for ESTs that bridge the distance between the detected gene and the mapped site are a possible reliability check and can also discriminate cases of alternative splicing. Further, the possibility of stumbling onto a pseudogene must be ruled out.[3,4]

Whereas all the steps leading to the protein sequences possibly encoded in the given transcript (in this essay, we do not consider untranslated RNAs) are sometimes complicated by sequencing errors (frameshifts, single point exchanges, genome fusion errors) but, in most cases, are just a technical exercise, the insufficient understanding of biological function for proteins known only as conceptual translations has become the major bottleneck in sequence data interpretation.

A few words on protein function: Protein function requires a hierarchical concept for the description of its many aspects that reflects the complexity of living systems.[5] The protein's function at the molecular level is rather a list of potential capabilities determined by its primary and tertiary structure. *Molecular function* description includes qualitative and quantitative aspects of diffusion properties in solution and membrane environments, conformational flexibility, allosteric conformational changes, possible ligand-binding (or catalytic) activities and ability for posttranslational modifications. Depending on cellular context (subcellular localization), different features of the molecular function may become important. A set of many cooperating proteins is responsible for a *cellular function* (metabolic pathway, signal transduction cascade, cytoskeletal complex, etc.). Since gene expression is regulated in a time- and tissue-dependent manner, regulatory sequences in the genomic environment of the gene considered come additionally into play at this level.[2] Finally, the presence and activity of a gene product may be directly associated with a *phenotypic function* at the organism or population level. Typically, only some aspects of molecular or cellular function are in the reach of sequence analytic studies.

## Concept No. 1: Function Inheritance from a Common Ancestor Gene

The most widely known, the evolutionary (historic) approach for inferring protein function with nonexperimental means is based on the frequent observation of similarity between biomolecular sequences coding proteins with similar molecular function. Since the early examples were typically metabolic enzymes or transporters (such as hemoglobin) for which the 3D structure was available, the insight materialized soon in the paradigm of both equal/similar

three-dimensional structural fold and molecular function as a consequence of similarity of protein sequence. Within this concept, a family of homologous gene/protein sequences is hypothesized to appear evolutionary during radiation of species (rarely via horizontal gene transfer) from an ancestor gene in the founding species via multiple mutations and, sometimes, gene duplications. In this context, the closest homologue of a gene in another organism ("the same gene") with most likely the same function is called orthologue, more distantly related homologues that, probably, arise from gene duplications and might assume new functions are named paralogues. Nevertheless, distant sequence similarity as a result of functional pressure or physicochemical constraints (analogous sequences in a scenario of convergent evolution) cannot always be excluded but, from the viewpoint of protein function prediction, the evolutionary pathway is not the major issue.

Functional annotation available from experimental studies of one family member is thought to be fully or partially transferable to all other members in the family. Therefore, considerable research effort has been focused on method development for more and more distantly related homologue detection to increase the likelihood of having experimentally studied family members. Except for obvious alignments with high sequence identity, it is not trivial to decide whether the similarity between sequences is significant in a statistical sense. The sequence homology approach is unthinkable without a mathematical function for measuring the similarity of two sequences quantitatively; i.e., a distance metric for the sequence space.

At the level of nucleic acids (genes and transcripts), the only possible measure is the count of identical positions in an optimal alignment. In this way, only relatively close sequence neighbors can be detected. Whereas the transcript sequence itself is just a redundant four-letter text, the translation into an amino acid sequence yields a more informative 20-letter message that often can be directly interpreted in physical and structural terms. Matrices of likelihood of amino acid type exchanges have been determined from experimentally established sequence families of globular proteins including some representatives with known tertiary structure. For example, amino acid type exchanges without changes of residue polarity/hydrophobicity or secondary structural preference impair protein structures less and are, therefore, more likely. Typically, such an exchange matrix enters the pairwise sequence similarity score function together with an empirical expression for the evaluation of evolutionary costs of deletions/insertions. For convenience of statistical evaluation, the score is recalculated into the probability (E-value) of incidentally reaching an alignment with the same or better score with a sequence taken randomly from a database of the same size. If this E-value is low, the predicted alignment is considered statistically significant. As probabilities, E-values should be always smaller than or equal to unity but analytically simplified computations of E-values, for example in the BLAST suite,[6] may lead to meaningless results above one for nonsignificant alignments with a low similarity score.

When a group of related proteins is known, then profiles that describe the likelihood of amino acid type occurrence at alignment positions can be extracted (see Step 5 in the Recipe below for detail). In turn, they allow the determination of ever more distantly related homologues in iterative cycles of profile extraction from growing alignments. Modern sequence profile techniques are the 'super-weapon' for collecting families of distantly related homologues and for assigning functions to globular domains via annotation transfer. Application of this technique lead to a number of breakthroughs in biology essentially with theoretical data analysis alone; (e.g., see refs. 7-14).

## Limitations of the Homology Search Concept

The deduction of the sequence distance metric has consequences for the applicability of homology searches in databases, for example with the BLAST/PSI-BLAST suite:[6,15]

1. The sequence distance metrics have been derived from alignments of globular proteins; more accurately, from alignments of secondary structural elements (e.g., BLOSUM62[16,17]). Obviously, such similarity functions may fail for other types of sequences; for example, for

cases having *amino acid compositions* that differ drastically from those of globular proteins. For example, long hydrophobic stretches with many transmembrane regions regardless of origin have a general tendency to appear similar. The same problem create long polar runs, sequences with systematic periodicities (coiled coils, collagen, etc.) as well as sequence segments with many cysteines, prolines or tryptophanes, amino acid types that are typically rare in globular proteins and the match of which is given high weight in the similarity measure. Thus, a sequence needs to be preprocessed to filter out all probable nonglobular segments before its submission to homology searches in sequence databases. Essentially, the term "distantly related sequence homologue" is not really applicable for nonglobular regions.

2. Each alignment position contributes a summand in the total score independent of all other position. Thus, the *mutual independence of sequence positions* in their mutation ability is assumed in contrast to well-known examples of correlated mutations not only in globular proteins[18-20] but also in some shorter motifs.[21,22] Thus, sequences that fit alignments somehow at all positions but do not comply with yet hidden inter-positional constraints may nevertheless pass the sequence similarity significance criterion. This effect is practically not important for long regions of homology since the number of correlating sites is small compared with the length. In contrast, this is one of the reasons why hits with shorter alignment length are often false.
3. Yet another problem is created by the modular structure of proteins that results from *sequence segment recombination* at the genomic level. Often, the homology relationship exists rather at the level sequence segments than for whole proteins. Therefore, it becomes important to delineate these homology segments and collect their families individually.
4. *Alignment length* and *sequence identity* are of critical importance for the transferability of functional annotation. Only about 50 positions and more allow reliably assuming similarity in 3D structure.[23] With decreasing sequence identity (especially below 40%), attributes such as enzyme class, binding sites or cellular function can be transferred only with caution.[24]

## Concept No. 2: Lexical Analysis, Physical Interpretation and Sequence Motif-Function Correlations

A biomolecular sequence may be analyzed in the same way as a text in a foreign language by studying occurrences/absences of certain letters (amino acid types) in the total sequence and in subsegments, by analyzing combinations of letters as well as their relative order, especially the repetitions of clusters of letters. As simple as the arithmetics of pure letter occurrences may appear, important conclusions can be drawn from such a study. The results receive a biological interpretation with the knowledge of physicochemical properties of amino acids and oligopeptides. For example, long stretches of hydrophobic amino acids may indicate secondary structural elements buried intramolecularly, within protein complexes or in lipid membranes. Runs with many polar residues are likely not to have the potential to form a hydrophobic core for a tertiary, native structure. The general relationship of hydrophobic and hydrophilic residues in larger segments might be, at least qualitatively, informative with respect to solubility and total charge. Such information can be helpful for the design of deletion mutants since those consisting mostly of hydrophobic segments are likely to produce false positive hits in a yeast two-hybrid screen and to aggregate after over-expression.

The concept of compositional bias towards certain amino acid types can be generalized with the notion of sequence complexity (information content, sequence entropy) as implemented, for example, in the SEG program.[25] Low complexity regions (LCRs) are common in sequence database proteins (~25% of all residues in sequence databases).[26,27] Sometimes, LCRs compose almost the whole protein as in the case of brakeless, a protein important for optical axon guidance in *D. melanogaster*.[28] Despite their wide spread and expected functional importance, the characterization of many LCRs, especially of those with many polar residues, still remains poor.[26]

LCRs are almost absent in known 3D structures of globular proteins (~0.5% of all residues in the protein structure database).[26,27] Thus, the concept of sequence complexity is a powerful quantitative measure for the distinction between globular (typically high complexity) and nonglobular (low complexity) regions (see ref. 29 for review). Only the high complexity regions represent good targets for sequence homology searches in database.

Many biological properties (helical transmembrane regions, coiled coils, N-terminal targeting signals, several posttranslational modifications, etc.; see Step 3 in the Recipe) are predicted from sequence with knowledge-based predictors: From a learning set of protein sequences, which are known to possess a biological feature, the encoding sequence pattern is extracted in a mathematically formalized way. Then, this pattern is searched for in query sequences, a concordance score is calculated and, in the most advanced techniques, the probability of false positive prediction is calculated. The quality of the predictor depends, first of all, on the learning set. Sometimes, it is small and does not reflect the true sequence variability in the pattern. Also, the various proteins in the learning set are typically not of the same quality with respect to their experimental verification status.

When the number of known sequences was small, a number of properties encoded in protein sequences could be associated with short amino acid type motifs ('sequence words'), which have been collected in databases, for example in PROSITE.[30] Today's sequence databases populate the available sequence space much more evenly. Therefore, short sequence motifs have a dramatically reduced predictive power (for example, the N-terminal myristoylation,[31] see also Step 2 in the Recipe).

## A Recipe for Analyzing Protein Sequences

The following section is a description of a series of steps that, if executed sequentially, will typically lead to insight into structural and functional features associated with an otherwise uncharacterized protein sequence if this is achievable with existing techniques at all. With our comments below, we want to show what is generally possible but also where are the today's limits and where we have to settle for lesser goals until methodical advances move the horizons further. As practical illustration of the recipe, we invite the reader to repeat the analysis of the pds5p sequence[32] together with us (see also Fig. 1). To avoid spoiling of the text with many WWW links that change anyhow with time, this information has been collected in a regularly maintained WWW-page associated with this article (http://mendel.imp.univie.ac.at/RECIPE/).

The basic paradigm in protein sequence analysis requires the dissection of the total sequence into segments (regions, domains), each of which has its own molecular functional features. The function of the whole protein is then obtained as superposition of the segments' elementary functions.

Functional sequence regions of a protein can be classified with respect to their intrinsic structural preference in a physiological environment. Some segments have a native structure (globular domains, nonglobular helical regions in coiled coil and transmembrane regions, collagens etc.); others have not. This distinction is critical for assessing interaction capabilities: Segments with intrinsic structural preference can supply specific, stable surface recognition sites for interactions with ligands (therefore, they have a large variety of specific functions); unstructured regions cannot. As we have seen above, various types of segments require different methods for their analysis. First, nonglobular regions (phase one, steps 1-3) and, then, segments belonging to already known families of globular domains (phase two, step 4) are determined. Finally, the remaining segments are expected to represent yet unknown globular domains and are subjected to sequence family search procedures (phase three, step 5). The final step involves analysis and synthesis of the sequence analytic findings.

### *Step 1: Linguistic Analysis*

At the beginning, it is necessary to check for linguistic particularities in the query protein sequence or its fragments. Such a textual analysis can be carried out by visual inspection or with

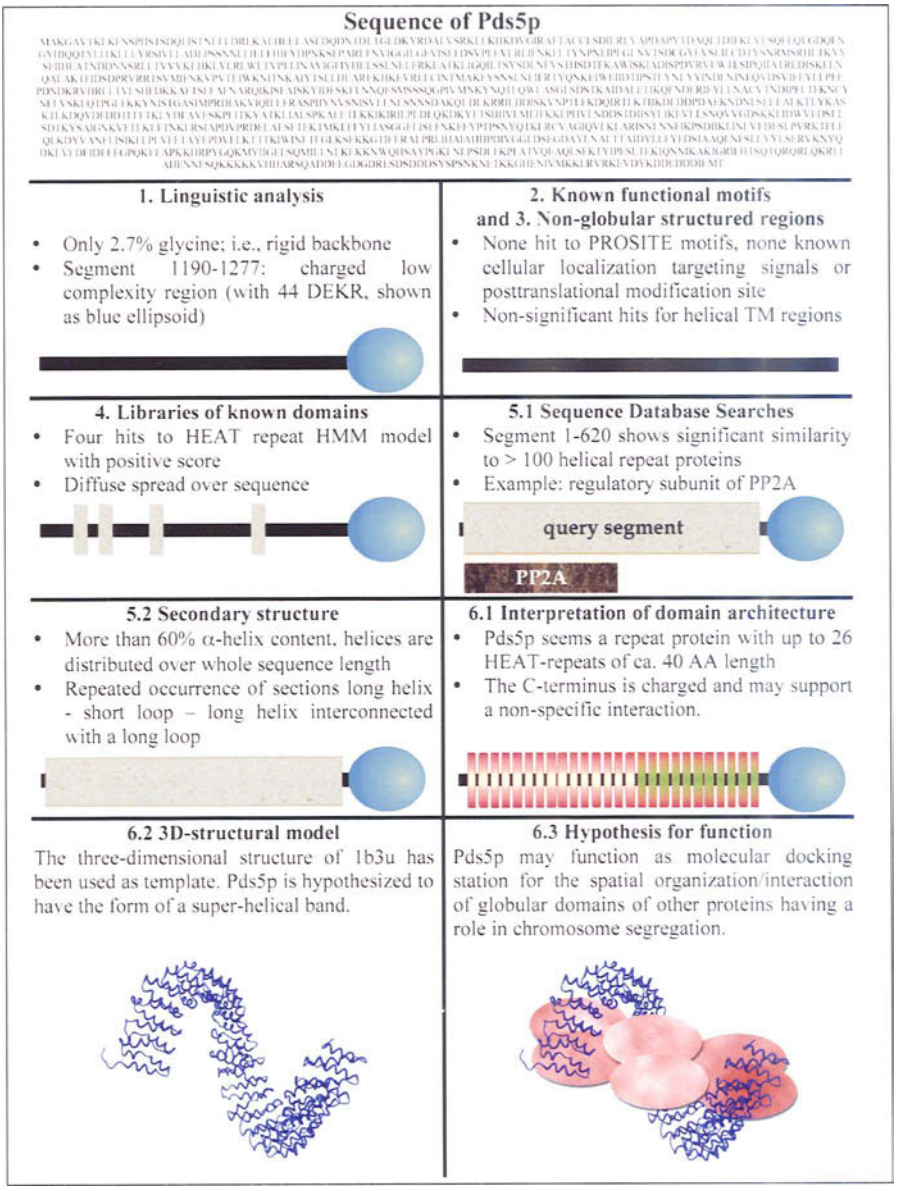


Figure 1. Sequence analysis of yeast pds5p. When the sequence-analytic study of yeast pds5p was started, only its sequence (top of the figure, 1277 amino acid residues) and its knock-out phenotype in mitosis were known.[32] Searches for nonglobular regions detected only a strongly charged region at the C-terminus. Compositional studies revealed a surprisingly low content of glycines indicating a generally rigid backbone. Three arguments (comparison with known domain profiles of helical repeats, distant similarity with the regulatory subunit of PP2A and predicted helical secondary structure including also the pattern of two helices interconnected by a short loop and a long loop between helix pairs) support the view that HEAT-repeats occupy the major part of the sequence. The reliability of these predictions decreases towards the C-terminal part. The HEAT-repeat region is suggested to fold into a super-helical band with interaction sites for other proteins, the charged C-terminal region has, apparently, a role for unspecific amplification of some binding reaction.[32]

computerized tools such as SAPS.[33] This program incorporates also rigorous statistical criteria for finding significant differences of the query's lexical properties from averages of SWISS-Prot sequences.[34]

Regions of low sequence complexity, another important lexical property, can be determined with tools such as SEG[25] or CAST.[35] The SEG program has three recommended parametrizations with sequence windows of w = 12, 25 or 45 residues. In standard applications, only the smallest window, the most stringent criterion, is applied. Personal experience shows that the larger window (w = 25) helps detecting less obvious LCRs, although SEG marks sometimes also globular regions as LCRs if applied with maximal window size (w = 45). The final output of SEG should be preprocessed for further analysis: (1) Sometimes, SEG leaves a small segment (with length below window size) between two neighboring LCRs unassigned. Such a segment can often be fused with the two LCRs into a single larger LCR. (2) Evaluation of polarity of LCRs is helpful for their functional assessment. Hydrophobic LCRs (rarely longer than 30 residues) often have a role in membrane attachment or are buried internally in protein complexes. Functional assignment of polar LCRs, especially those with more than 100 residues length, is more problematic. Polar LCRs are thought to be intrinsically unstructured and in contact with the aqueous phase. Some serve as mechanical linkers between domains, have a role in electrostatic interactions or carry sites for posttranslational modifications. The specific molecular function of polar LCRs is typically unclear except for rare cases.[26,28,29]

## *Step 2: Motifs for Subcellular Targeting and Posttranslational Modifications*

A number of functional motifs for posttranslational modifications or targeting to subcellular localizations are located within sequence regions without intrinsic structural preference. Specialized predictors can test the occurrence of these motifs. Several N-terminal signals involving typically 20-40 residues encode targeting to organelles: SIGNALP[36] recognizes the signal leader peptide for export to the endoplasmic reticulum. CHLOROP[37] searches for chloroplast- and another tool[38] for mitochondrion-targeted proteins. SIGNALP in its recent version has very reasonable prediction accuracy above 80% for true predictions for large sequence sets and a low rate (~14-19%) for false-positive hits and compares favorably with alternative tools.[39] Prediction of chloroplast- and mitochondrial targeting are not comparable in this respect, first of all, because the available sets of experimentally learning sequences are less comprehensive and reliable. TARGETP[40] represents a unified version of all three predictors. A new predictor for the C-terminal PTS1 signal (with a length of about 12 residues) that encodes pex5-dependent peroxysomal localization has a sensitivity >95% and a selectivity below 0.5%.[41]

Several lipid posttranslational modifications of proteins can now be reliably predicted from sequence. (1) N-terminal N-myristoylation is encoded by a signal of about 17 residues. It is recognized with >95% for true sites and with less than 0.5% for unrelated sequences by a recently developed tool.[22,31] In some cases of posttranslational processing, internal glycines become N-terminal and myristoylated. This program analyzes also a number of such scission patterns. N-terminal N-myristylation with subsequent palmytoylation (if there are cysteines close to the N-terminus) might hint at a noncanonical export mechanism.[42] (2) Glycosylphosphatidylinositol (GPI) lipid anchoring is a posttranslational modification of protein C-termini carrying the respective recognition signal of ca. 40 residues. The anchor is attached after proteolytic scission of a propeptide. The big-II predictor predicts GPI lipid anchor attachment (~80% accuracy for truly anchored animal proteins with ~0.2% false positives) and computes also the one or two most probable attachment sites.[21,43-47] (3) A recently released predictor for farnesylation and geranylgeranylation, the two types of prenylation at protein C-termini, is accessible from the WWW-page associated with this article.

The localization and lipid modification signals discussed above involve 12-50 residues from the respective termini. Typically, they are not characterized by amino acid type preferences alone but also by sequence context involving a strong pattern of physical properties and, partially, by some inter-positional correlations within the motif. Only this additional information

allows reliable motif detection in uncharacterized sequences and the assessment of the possible prediction error in tests involving dozens of thousands of sequences.[48]

It should be emphasized that conservation of a handful of residues in a short motif alone does not imply function correlation and, barely, supports more than a working hypothesis. Typically, short, polar oligopeptides do not have intrinsic structural preferences;[49] they cannot supply a stable interface for intermolecular interactions. Even in the case of true function embedding into an unstructured region of a protein that interacts with a globular domain of another protein, a functional motif requires a sequential environment involving residues for less specific interactions and linker function.[21,22,48]

To illustrate, a number of short PROSITE motifs[30] are also used for characterizing post-translational modification sites (for example, for phosphorylation, N-glycosylation and myristoylation) but with a high rate of false hits.[31] Other arguments (e.g., experimental data) are needed to support the relevance of predicted sites. There are alternative neural network based predictors for phosphorylation,[50] O- and N-glycosylation[51,52] but their prediction accuracy is not yet sufficient for unsupervised sequence annotation.

Similarly, many other, scarcely described and yet insufficiently understood sequence signals, e.g., for nuclear import[53] and export[54] or the PEST degradation signature,[55] circulate widely in the literature but their predictive significance for sequence analysis is still low since the correlation between protein sequence variability and function remains ambiguous. Often, the biological mechanisms for read-out of these signals are poorly understood.

### *Step 3: Nonglobular Regions with Intrinsic Structural Preference*

At early stages of sequence studies, it is important to recognize α-helical transmembrane regions and coiled coil segments. Both have compositional bias, which is often not recognized by sequence complexity computing programs, and, consequently, these segments should also be removed from the sequence before submission to searches for distant relatives in sequence databases.

Coiled coil regions can be predicted from sequence with the updated COILS algorithm of Lupas.[56] Typically, WWW-server versions run COILS only with standard parametrization and, sometimes, predict coiled coils wrongly in regions with many polar residues without any hydrophobic amino acids in 'a' and 'd' positions of the heptade repeat. A second COILS run with a changed weighting for polar residues as recommended in the manual diagnoses many of those doubtful assignments. To notify, there are also versions of COILS in the public domain erroneously deviating from the original implementation of algorithm and resulting in fewer and shorter predicted coiled coil segments for some proteins.

There may be other fibrillar segments in proteins. For example, collagen segments are recognized by typical glycine- and proline-rich repeats and this property is incorporated in an HMM of the PFAM domain PF01391.[57]

The prediction of membrane attachment of integral membrane proteins via protein segments immersed into the lipid bilayer is still problematic. If transmembrane helical regions are present, they are readily recognized by prediction tools like TMHMM[58] or DAS-TMfilter, a recent update of DAS,[59] as well by a number of other programs.[60] With less accuracy, the protein topology with respect to the membrane is predicted (mostly based on the positive-inside-rule[61]). Since the motif description rests almost entirely on the requirement of long hydrophobic stretches (except for a minimum length), false positive prediction, especially of single membrane-pass proteins is frequent. TMHMM and DAS-TMfilter have a better selectivity than the competing programs but they also fail for proteins with long helical, hydrophobic repeats (for example, ARM/HEAT repeat proteins such as tis7 (gi321269) or inscuteable (gi1079094)).

The architectural diversity of proteins attached to membranes involves more than just transmembrane helical regions but these configurations cannot be predicted with available TM region prediction tools. For example, there is an interesting class of amphipatic helices

embedded into the membrane parallel to the bilayer surface (monotopic membrane proteins).[62-64] Further, transmembrane helix formation is not entirely determined locally by the hydrophobic stretch itself but may depend on the rest of the protein sequence[65] or even complex formation.

### *Step 4: Known Sequence Families of Globular Domains*

Globular domains are the main structural and functional building blocks of proteins. Various definitions of the notion 'domain' differ but their content is overlapping. From the viewpoint of three-dimensional structure, a domain is a compact, spatially distinct unit with its own hydrophobic core, the fundament of its native tertiary structure. In the kinetic sense, a native structure implies that conformational fluctuations are locally confined (i.e., are smaller than the size of the three-dimensional structure). Thus, globular domains can supply stable interfaces and recognition sites for other molecules, even for those without intrinsic structural preference. Thermodynamically, a domain is melting independently. Often, a domain is considered an autonomous folding unit. At the same time, a structural unit might not be continuous in the sequence. In the evolutionary perspective and in sequence comparisons, a domain is a family of significantly similar sequences that are related by their mutational history. From the functional viewpoint, domains may be promiscuous with different active sites and binding capabilities for various sequence family members but the degree of diversity is uneven among domains. A typical globular domain involves 100-150 amino acid residues;[66-68] thus, much longer segments can be supposed to involve several independent domains. To avoid confusions, it is advised to use the term "domain" in the sense of globular domain and to apply sequence region or segment in other context.

At this stage of analysis, it is a good decision to compare the target sequence with entries in public domain databases. There are traditional profile-based (PROSITE,[30] BLOCKS,[69] PRINTS[70]); hidden Markov model (HMM)-based (PFAM,[57] SMART,[71] significance threshold typically E~0.1); combined tools (PANAL[72]) and RPS-BLAST profile-based (CDD search,[73] significance threshold typically E~0.01) collections. There are at least two reasons: The given sequence might be so distantly related to a known family that a simple pairwise similarity search with the query or any of the family members would not detect that relationship. Profiles describing whole families are much more sensitive. Second, one domain in multi-domain targets may have so many close relatives in the sequence database that the output list from a BLAST search with the full sequence would be obliterated with those hits alone. It makes sense to compare a query with all available domain libraries since definitions of even actually the same domain may slightly differ and numerical noise can lead to hits in one but not in another library.

Currently, there are two major primary domain libraries. PFAM is unprecedented in sequence coverage.[57] At the same time, the domain definitions may contain slight inconsistencies mostly concerning boundaries of domains. Sometimes, signal peptides, fibrillar protein segments or helical transmembrane regions are included into the profile or the domain definition contains actually several domains. SMART is a very carefully curated but much smaller domain databases that focuses on certain classes of signaling, nuclear and extracellular proteins.[71] SMART domain boundaries typically define the core of a single globular domain.[74]

There are two modes for searching the occurrence of domains in query sequences with HMMs and profiles. In the so-called global mode, the presence of only complete domains is assumed and the optimal alignment of a query segment with the complete domain profile is searched. This mode is typically more sensitive that the fragmented domain search where also partial hits of the domain profile in the query are reported. In the ideal case, both regimes deliver the same result. Most hits from the fragmented domain search are meaningless in the absence of full-domain matches but if they coincide with known binding sites for ligands or otherwise functionally relevant parts of the domain, careful sequence inspection may lead to a discovery of very distantly related sequence homologues.

A fragmented domain search with the profile of the histone acetyltransferase family has hit eco1p, a yeast protein for the establishment of cohesion between chromatids during mitosis, in the region of the acetyl-CoA binding site.[9] The partial hit was extended with arguments based on secondary structure prediction and the conservation of a hydrophobic pattern. This finding stimulated experimental analysis and finally led to the discovery of a new family of acetyl-CoA binding and acetyl-transferring enzymes with a role in cohesion.[9]

The domains with multiple internal structural repeats are difficult to detect; therefore, this domain class requires special attention.[75] Such repeats are known as closed structures (e.g., β-propellers) or as semi-closed forms, for example the superhelical armadillo or heat repeats. Many repeat proteins have scaffolding functions for protein-protein interactions. For repeat detection, the query should be cleaned from compositionally biased regions in accordance with steps 1-3 of the recipe. The PROSPERO tool[76] is designed for recognizing even subtle internal sequence repeats. Since it operates with rigorous statistical criteria, the validity of the finding can be assessed in probabilistic terms. The REP tool[77] compares the query sequence with an HMM library of known repeats. Unfortunately, the evolutionary pressure for sequence conservation within repeats is typically low and reduced to the requirement of packing and maintenance of the hydrophobic core. Therefore, even hits with low statistical significance deserve attention.

### *Step 5: Sequence Database Searches*

Searches for similar sequences in databases can be applied in two different contexts. Full sequence searches are reasonably aimed only at finding closely related sequential neighbors where the methodical details of deriving the sequence distance metric do not have a major impact on the search result (typically, log (*E - value*) < - 10 for BLAST).

A search in sequence databases for similar but distantly related proteins with the target under study is in fact the last step of sequence analysis. Only sequence segments without low complexity, transmembrane and coiled coil regions, peptide segments for posttranslational modifications and cellular targeting, and known domains can routinely be subjected to such searches. Now, the effort is aimed at collecting the complete sequence family. The larger the family, the higher is the probability of hitting a functionally annotated family member. Additionally, it is necessary to understand the sequence variability within the sequence.

Traditionally, this a process of repeated application of pairwise sequence comparison techniques such as BLAST and general profile-searching techniques relying on manually or automatically constructed alignments (PSI-BLAST[78] with inclusion E-values up to ~0.01, SAM-T99,[79,80] or a combination of Clustalx[81,82] with a profile searching technique). Both the primary query as well as any new family members is subjected to such searches. The optimal search heuristics are a matter of continued scientific discussion.[74] Large sequence families have an internal structure consisting of clusters of sequentially (and, often, functionally) more similar proteins with statistically significant links between them.

Three aspects deserve additional comments: First, borderline hits require visual inspection before inclusion into the family or their final rejection. An excellent review of physical and structural criteria for nonstatistical evaluation of alignment significance (based on considerations of protein structural architecture) has been supplied by Bork and Gibson.[83] Reoccurrence of some motif conserved within the family might indicate correct assignment. Finally, the correct inclusion into the family should be verified by a reciprocal database search (started with the doubtful sequence segment) that collects already verified family members with statistical significance. It must be noted that many database search programs are not 100% commutative with respected to starting and hit sequences due to algorithmic simplifications that save computing time. Second, manually constructed alignments may be superior over those automatically generated, especially if 3D structural information for at least one family member is available. In the case of the pleckstrin homology (PH) domain sequences, sequence identities had been very low but reliable alignments applicable for further rounds of profile searches were obtained with manual adjustment emphasizing the conserved hydrophobic patterns and a

conserved tryptophane position.[84,85] Third, the probability of finding hits can be increased if EST and genome databases are six-frame translated on the fly and included into the search for relatives.[86] In a few cases, some relaxation of search thresholds leads to the necessary intermediate sequence hits during family collection. Fourth, since most amino acid substitution matrices give high weights to matches of rare residues such as cysteine, database searches with such a sequence segment to database searches may result in spurious hits with underestimated E-values, which may become close to standard selection thresholds. This has happened in the case of the C-terminal domain of wingless/wnt-1 that was incorrectly suggested to be related to the lipid-binding domain of phospholipase A2.[87] This possibility was later ruled out by structural arguments (completeness of the hydrophobic core, satisfaction of disulphid bonds).[88]

Until recently, it was very difficult to find routinely so distantly related family members with known 3D structures that have no recognizable sequence similarity with pure sequence-based approaches but, nevertheless, have the same fold. Higher sensitivity is achieved in comparisons of two profiles, one extracted from the query's sequence family and the other from a family of proteins of similar 3D structure and their sequential homologues. In addition to information from amino acid letter comparison, some structural information can be mobilized: The alignment of query sequences with structural templates, the mapping of sequence positions to structural positions, allows, for example, scoring of the agreement between predicted secondary structure of the query with the secondary structure of the template or the polarity of amino acid residues of the query with the accessibility of template sites. Different strategies have been implemented in 3D-PSSM,[89] bioinbqu,[90] DOE FOLD predictor,[91] FFAS,[92] PSIPRED,[93] SAM,[79] SDSC1[94] and SUPERFAMILY,[95] which are available as WWW-servers. Generally, their predictions have to be viewed with caution. Similar predictions for various sequence family members are indicative for higher significance. Some of these techniques have been equipped with methods for assessing the probability of false positive prediction. There are cases where the prediction of the 3D-structure with fold predictors has produced the decisive hint. For example, the predicted β-propeller structure of the globular domain of PIG-T can explain its molecular function as gate mechanism for protein substrates of the transamidase PIG-K in the GPI lipid anchor biosynthesis pathway.[96]

Yet another approach for enlarging the sequence family focuses on sequence architecture, the linear order of functional segments in a protein. Sub-threshold similarity in some sequence segment combined with similar length and order of other architectural elements can indicate on the existence of homologues in other species, even if the evolutionary divergence has become high.[97,98]

After having the sequence family completed, the family sequence alignment, known structures of family members, the available sequence annotation and the scientific literature for all family members have to be studied. First, conservation patterns of hydrophilic/hydrophobic residues and of secondary structural elements (indicating fold conservation), or of motifs with functional residues (giving a hint at conserved ligand binding and active sites) have to be taken into account.[99] The secondary structure predicted with JPRED[100] or PSIPRED[93] for the sequence family can help in the interpretation of the data. Second, details of known structures of family members that do not depend on the sequence-variable positions should be searched for. For example, the distribution of the electrostatic potential at the protein surface is sometimes invariant within a family and may explain the binding behavior. Searches for proteins with the same fold[101] can give lead to functional information on other proteins with the same fold. Third, the taxonomic distribution[102] of the family is informative with respect to the evolution of the cellular processes involving the sequence domain studied. Sometimes, evolutionary trees constructed from all family members may yield additional insight.

The scientific literature must be searched for experimental evidence of biological function that can be linked with the sequence segment in some family members. The degree of possible annotation transfer from family members to the target under consideration depends on many aspects. As a rule, the similarity with respect to the 3D fold can be determined with greater

reliability but molecular and, the more, cellular functional descriptors cannot always be transferred with the same confidence due to considerable plasticity of protein function.[5,103] For example, a large family of proteins has in common a domain responsible for ras-binding in the case of many family members.[104] This information was extrapolated to the whole family including the Rho-GTPase-activating protein myr-5. For the latter one, it turned out that the presence of this domain and its fold was predicted correctly, but the actual function was not.[105]

### *Step 6: Analysis of Sequence Analytic Findings and Synthesis of Molecular Function*

First, it is necessary to evaluate the reliability of predictions and annotations for overlapping sequence segments and to resolve possible contradictions. Then, the prediction results should allow segmentation of the query sequence into sequence regions, to which the collected structural and functional annotation can be attached. Often, some experimental data for the protein analyzed is available from the cooperating experimental researchers, which has to be discussed now in context with the sequence-analytic findings. Synthesis of the segments' functions into the protein function is the most creative step in the whole procedure where the biological knowledge of the researcher and his experience in using sequence analytic methods come together. It is possible that the collected evidence is so strong that there is no doubt (see ref. 106 for discussion). In most cases, the thought concentrates on consequences for the further experimental strategy. For example, clear directives can be given for mutant design: Deletion mutants should follow the derived segmental structure; point mutation should focus on conserved sequence positions.

Protein structure and function are encrypted in the protein sequence; thus, they can be predicted relying on amino acid sequence information in principle. Sometimes, molecular and cellular properties can be predicted. Phenotypic functions are usually outside the predictive power of sequence analytic studies (only in cases of clear homology). It should be emphasized that there are aspects of molecular function that strongly resist theoretical treatment. It is highly unlikely that theoretical methods will predict biological features without any analogy to experimentally studied cases since all procedures finally rely on observed sequence-function correlation.

Even if the 3D structures of two individual subunit proteins are known, it is still not possible to reliably predict the specific protein-protein interaction in a complex.[107] In the general case, there is no way to predict even the fact of complex formation from sequence alone. Potential hints can be obtained from homology considerations but, as in the case of the putative ras-binding activity of myr-5,[104,105] with low reliability. Sometimes, conservation of gene order or regulatory genomic neighborhood, gene fusion events or the conserved cooccurrence of genes in different genomes might be supportive for interaction[108-110] but not more. With large-scale mass spectrometric analysis, list of proteins in complexes have been compiled that can be looked up as well as interactions from two-hybrid screens.[111-114]

## Concluding Remarks

The development of high-throughput experimental technologies and its first major breakthrough, the complete sequencing of the genomes of organisms ranging from viruses over bacteria, lower eukaryotes to human, has changed life science research qualitatively. For the first time, the biological object can be studied in its totality at the molecular level. The immediate task for the coming decade consists in assigning functions to all genes known by sequence. Since the new data are so large and their the biological interpretation require complex approaches, theoretical science can and must contribute decisively to the research progress. The research success in life sciences depends increasingly on the ability of researchers in experimental and theoretical biology to jointly focus on relevant questions.

Modern protein sequence analysis relies on two major approaches: protein homology searches based on the concept of statistically significant sequence similarity and textual analysis with physical interpretation and the extrapolation of empirical relationships established between local sequence motifs and patterns with structural and functional properties.

***Acknowledgements***

The author is grateful for generous support from Boehringer Ingelheim and from the Bioinformatics Network within the Genome Research Austria program (Gen-AU BIN). F. Eisenhaber thanks Peer Bork (EMBL Heidelberg) who attracted him to the field of biomolecular sequence analysis during 1996 and readily shared his practical experience. This chapter benefited from conversations with Chris Ponting (MRC Oxford) and Eugene Koonin (NCBI Bethesda) as well as from interactions with wet lab IMP scientists that kindly shared their findings with the IMP bioinformatics group on a continuous basis. Special thanks are to B. Eisenhaber, W. Kubina, S. Maurer-Stroh, G. Neuberger, M. Novatchkova, A. Schleiffer, G. Schneider, S. Washietl, M. Wildpaner for discussions of various aspects of this work and S. Maurer-Stroh for support in creating parts of Figure 1.

## References

1. Novatchkova M, Eisenhaber F. Can molecular mechanisms of biological processes be extracted from expression profiles? Case study: Endothelial contribution to tumor-induced angiogenesis. Bioessays 2001; 23:1159-1175.
2. Zhang MQ. Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet 2002; 3:698-709.
3. Fickett JW. ORFs and genes: How strong a connection? J Comput Biol 1995; 2:117-123.
4. Harrison PM, Hegyi H, Balasubramanian S et al. Molecular fossils in the human genome: Identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 2002; 12:272-280.
5. Bork P, Dandekar T, Diaz-Lazcoz Y et al. Predicting function: From genes to genomes and back. J Mol Biol 1998; 283:707-725.
6. Altschul S, Boguski M, Gish W et al. Issues in searching molecular sequence databases. Nature Genetics 1994; 6:119-129.
7. Yuan YP, Schultz J, Mlodzik M et al. Secreted fringe-like signaling molecules may be glycosyltransferases. Cell 1997; 88:9-11.
8. Rea S, Eisenhaber F, O'Carroll D et al. Regulation of chromatin structure by site-specific histone h3 methyltransferases. Nature 2000; 406:593-599.
9. Ivanov D, Schleiffer A, Eisenhaber F et al. Eco1 is a novel acetyltransferase that can acetylate proteins involved in cohesion. Curr Biol 2002; 12:323-328.
10. Dlakic M. Chromatin silencing protein and pachytene checkpoint regulator dot1p has a methyltransferase fold. Trends Biochem Sci 2001; 26:405-407.
11. van Leeuwen F, Gafken PR, Gottschling DE. Dot1p modulates silencing in yeast by methylation of the nucleosome core. Cell 2002; 109:745-756.
12. Aravind L, Koonin EV. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol 2001; 2:RESEARCH0007.
13. Trewick SC, Henshaw TF, Hausinger RP et al. Oxidative demethylation by escherichia coli AlkB directly reverts DNA base damage. Nature 2002; 419:174-178.
14. Falnes PO, Johansen RF, Seeberg E. AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia Coli. Nature 2002; 419:178-182.
15. Altschul SF, Madden TL, Schaffer AA et al. Gapped blast and PSI-blast: A new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389-3402.
16. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Nat Acad Sci USA 1992; 89:10915-10919.
17. Henikoff S, Henikoff JG. Amino acid substitution matrices. Adv Protein Chem 2000; 54:73-97.
18. Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: Maximum likelihood identification and relationship to structure. J Mol Biol 1999; 287:187-198.
19. Cootes AP, Curmi PM, Cunningham R et al. The dependence of Amino acid pair correlations on structural environment. Proteins 1998; 32:175-189.
20. Chelvanayagam G, Eggenschwiler A, Knecht L et al. An analysis of simultaneous variation in protein structures. Protein Eng 1997; 10:307-316.
21. Eisenhaber B, Bork P, Eisenhaber F. Sequence properties of GPI-anchored proteins near the Ω-site: Constraints for the polypeptide binding site of the putative transamidase. Protein Eng 1998; 11:1155-1161.
22. Maurer-Stroh S, Eisenhaber B, Eisenhaber F. N-terminal N-myristoylation of proteins: Prediction of substrate proteins from Amino acid sequence. J Mol Biol 2002; 317:541-557.
23. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991; 9:56-68.

24. Devos D, Valencia A. Practical limits of function prediction. Proteins 2000; 41:98-107.
25. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 1996; 266:554-571.
26. Wootton JC. Sequences with 'Unusual' Amino acid compositions. Curr Op Struct Biol 1994; 4:413-421.
27. Saqi M. An analysis of structural instances of low complexity sequence segments. Protein Eng 1995; 8:1069-1073.
28. Senti K, Keleman K, Eisenhaber F et al. Brakeless is required for lamina targeting of R1-R6 axons in the Drosophila visual system. Development 2000; 127:2291-2301.
29. Eisenhaber B, Eisenhaber F. Sequence complexity of proteins and its significance in annotation. In: Subramaniam S, ed. Bioinformatics in the Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. New York: Wiley Interscience, 2005:4, (DOI:10.1002/047001153X.g403313).
30. Falquet L, Pagni M, Bucher P et al. The PROSITE database, its status in 2002. Nucleic Acids Res 2002; 30:235-238.
31. Maurer-Stroh S, Eisenhaber B, Eisenhaber F. N-terminal N-myristoylation of proteins: Refinement of the sequence motif and its taxon-specific differences. J Mol Biol 2002; 317:523-540.
32. Panizza S, Tanaka T, Hochwagen A et al. Pds5 cooperates with cohesin in maintaining sister chromatid cohesion. Curr Biol 2000; 10:1557-1564.
33. Brendel V, Bucher P, Nourbakhsh IR et al. Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci USA 1992; 89:2002-2006.
34. Karlin S, Brendel V. Chance and statistical significance in protein and DNA sequence analysis. Science 1992; 257:39-49.
35. Promponas VJ, Enright AJ, Tsoka S et al. CAST: An iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. Bioinformatics 2000; 16:915-922.
36. Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng 1999; 12:3-9.
37. Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 1999; 8:978-984.
38. Emanuelsson O, von Heijne G, Schneider G. Analysis and prediction of mitochondrial targeting peptides. Methods Cell Biol 2001; 65:175-187.
39. Menne KM, Hermjakob H, Apweiler R. A comparison of signal sequence prediction methods using a test set of signal peptides. Bioinformatics 2000; 16:741-742.
40. Emanuelsson O, von Heijne G. Prediction of organellar targeting signals. Biochim Biophys Acta 2001; 1541:114-119.
41. Neuberger G, Maurer-Stroh S, Eisenhaber B et al. Prediction of PTS1 signal dependent peroxysomal targeting from protein sequences. submitted 2002.
42. Denny PW, Gokool S, Russell DG et al. Acylation-dependent protein export in leishmania. J Biol Chem 2000; 275:11017-11025.
43. Eisenhaber B, Bork P, Eisenhaber F. Prediction of potential GPI-modification sites in proprotein sequences. J Mol Biol 1999; 292:741-758.
44. Eisenhaber B, Bork P, Yuan Y et al. Automated annotation of GPI anchor sites: Case study C. Elegans. Trends Biochem Sci 2000; 25:340-341.
45. Eisenhaber B, Bork P, Eisenhaber F. Post-translational GPI lipid anchor modification of proteins in kingdoms of life: Analysis of protein sequence data from complete genomes. Protein Eng 2001; 14:17-25.
46. Eisenhaber B, Schneider G, Wildpaner M et al. A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for aspergillus nidulans, candida albicans, neurospora crassa, Saccharomyces Cerevisiae and schizosaccharomyces pombe. J Mol Biol 2004; 337:243-253.
47. Eisenhaber B, Wildpaner M, Schultz CJ et al. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for arabidopsis and rice. Plant Physiol 2003; 133:1691-1701.
48. Eisenhaber B, Eisenhaber F, Maurer-Stroh S et al. Prediction of sequence signals for lipid post-translational modifications: Insights from case studies. Proteomics 2004; 4:1614-1625.
49. Minor Jr DL, Kim PS. Context-dependent secondary structure formation of a designed protein sequence. Nature 1996; 380:730-734.
50. Blom N, Gammeltoft S, Brunak S. Sequence and structurebased prediction of eukaryotic protein phosphorylation sites. J Mol Biol 1999; 294:1351-1362.
51. Hansen JE, Lund O, Tolstrup N et al. NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. Glycoconj J 1998; 15:115-130.
52. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. Pac Symp Biocomput 2002; 310-322.

53. Cokol M, Nair R, Rost B. Finding nuclear localization signals. EMBO Rep 2000; 1:411-415.
54. Yoneda Y. Nucleocytoplasmic protein traffic and its significance to cell function. Genes Cells 2000; 5:777-787.
55. Rechsteiner M, Rogers SW. PEST sequences and regulation by proteolysis. Trends Biochem Sci 1996; 21:267-271.
56. Lupas A. Predicting coiled-coil regions in proteins. Curr Opin Struct Biol 1997; 7:388-393.
57. Bateman A, Birney E, Cerruti L et al. The Pfam protein families database. Nucleic Acids Res 2002; 30:276-280.
58. Krogh A, Larsson B, von Heijne G et al. Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. J Mol Biol 2001; 305:567-580.
59. Cserzo M, Wallin E, Simon I et al. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: The dense alignment surface method. Protein Eng 1997; 10:673-676.
60. Moller S, Croning MD, Apweiler R. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 2001; 17:646-653.
61. von Heijne G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 1992; 225:487-494.
62. Picot D, Garavito RM. Prostaglandin H synthase: Implications for membrane structure. FEBS Lett 1994; 346:21-25.
63. Wendt KU, Lenhart A, Schulz GE. The structure of the membrane protein squalene-hopene cyclase at 2.0 a resolution. J Mol Biol 1999; 286:175-187.
64. Sukumar N, Xu Y, Gatti DL et al. Structure of an active soluble mutant of the membrane-associated (S)- mandelate dehydrogenase. Biochem 2001; 40:9870-9878.
65. Goder V, Spiess M. Topogenesis of membrane proteins: Determinants and dynamics. FEBS Lett 2001; 504:87-93.
66. Trifonov EN. Segmented structure of protein sequences and early evolution of genome by combinatorial fusion of DNA elements. J Mol Evol 1995; 40:337-342.
67. Wheelan SJ, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. Bioinformatics 2000; 16:613-618.
68. Xu D, Nussinov R. Favorable domain size in proteins. Fold Des 1998; 3:11-17.
69. Henikoff JG, Pietrokovski S, McCallum CM et al. Blocks-based methods for detecting protein homology. Electrophoresis 2000; 21:1700-1706.
70. Attwood TK, Beck ME, Flower DR et al. The PRINTS protein fingerprint database in its fifth year. Nucleic Acids Res 1998; 26:304-308.
71. Letunic I, Goodstadt L, Dickens NJ et al. Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res 2002; 30:242-244.
72. Silverstein KA, Kilian A, Freeman JL et al. PANAL: An integrated resource for protein sequence ANALysis. Bioinformatics 2000; 16:1157-1158.
73. Marchler-Bauer A, Panchenko AR, Shoemaker BA et al. CDD: A database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 2002; 30:281-283.
74. Ponting CP, Schultz J, Copley RR et al. Evolution of domain families. Adv Protein Chem 2000; 54:185-244.
75. Chelvanayagam G, Knecht L, Jenny T et al. A combinatorial distance-constraint approach to predicting protein tertiary models from known secondary structure. Fold Des 1998; 3:149-160.
76. Mott R. Accurate formula for P-values of gapped local sequence and profile alignments. J Mol Biol 2000; 300:649-659.
77. Andrade MA, Ponting CP, Gibson TJ et al. Homology-based method for identification of protein repeats using statistical significance estimates. J Mol Biol 2000; 298:521-537.
78. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem Sci 1998; 23:444-447.
79. Karplus K, Hu B. Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. Bioinformatics 2001; 17:713-720.
80. Karplus K, Karchin R, Barrett C et al. What is the value added by human intervention in protein structure prediction? Proteins 2001; 45(Suppl 5):86-91.
81. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22:4673-4680.
82. Higgins D, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignment. Meth Enzymol 1996; 266:383-402.
83. Bork P, Gibson TJ. Applying motif and profile searches. Meth Enzymol 1996; 266:162-184.
84. Musacchio A, Gibson TJ, Rice P et al. The PH-domain: A common piece in the structural patchwork of signalling proteins. Trends Biochem Sci 1993; 18:343-348.

85. Gibson TJ, Hyvönen M, Musacchio A et al. PH domain: The first anniversary. Trends Biochem Sci 1994; 19:349-353.
86. Aravind L, Koonin EV. Classification of the caspase-hemoglobinase fold: Detection of new families and implications for the origin of the eukaryotic separins. Proteins 2002; 46:355-367.
87. Reichsman F, Moore HM, Cumberledge S. Sequence homology between wingless/Wnt-1 and a lipid-binding domain in secreted phospholipase A2. Curr Biol 1999; 9:R353-R355.
88. Barnes MR, Russell RB, Copley RR et al. A lipid-binding domain in Wnt: A case of mistaken identity? Curr Biol 1999; 9:R717-R719.
89. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000; 299:499-520.
90. Fischer D. Hybrid fold recognition: Combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000; 5:119-130.
91. Mallick P, Goodwill KE, Fitz-Gibbon S et al. Selecting protein targets for structural genomics of pyrobaculum aerophilum: Validating automated fold assignment methods by using binary hypothesis testing. Proc Natl Acad Sci USA 2000; 97:2450-2455.
92. Rychlewski L, Jaroszewski L, Li W et al. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000; 9:232-241.
93. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000; 16:404-405.
94. Shindyalov IN, Bourne PE. Improving alignments in HM protocol with intermediate sequences. Forth Meeting on the Critical Assessment of Techniques for Protein Structure Prediction 2000; A92.
95. Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res 2002; 30:268-272.
96. Novatchkova M, Eisenhaber F. A CH domain-containing N terminus in NuMA? Protein Sci 2002; 11:2281-2284.
97. Lorenz A, Wells JL, Pryce DW et al. Pombe meiotic linear elements contain proteins related to synaptonemal complex components. J Cell Sci 2004; 117:3343-3351.
98. Rabitsch KP, Gregan J, Schleiffer A et al. Two fission yeast homologs of Drosophila mei-S332 are required for chromosome segregation during meiosis I and II. Curr Biol 2004; 14:287-301.
99. Ponting CP. Issues in predicting protein function from sequence. Brief Bioinform 2001; 2:19-29.
100. Cuff JA, Clamp ME, Siddiqui AS et al. JPred: A consensus secondary structure prediction server. Bioinformatics 1998; 14:892-893.
101. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998; 11:739-747.
102. Wildpaner M, Schneider G, Schleiffer A et al. Taxonomy workbench. Bioinformatics 2001; 17:1179-1182.
103. Devos D, Valencia A. Intrinsic errors in genome annotation. Trends Genet 2001; 17:429-431.
104. Ponting CP, Benjamin DR. A novel family of Ras-binding domains. Trends Biochem Sci 1996; 21:422-425.
105. Kalhammer G, Bahler M, Schmitz F et al. Ras-binding domains: Predicting function versus folding. FEBS Lett 1997; 414:599-602.
106. Iyer LM, Aravind L, Bork P et al. Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. Genome Biol 2001; 2, (RESEARCH0051).
107. Strynadka NCJ, Eisenstein M, Katchalski-Katzir E et al. Molecular docking programs successfully predict the binding of a B-lactamase inhibitory protein to TEM-1 \BETA\-lactamase. Nature Struct Biol 1996; 3:233-239.
108. Dandekar T, Snel B, Huynen M et al. Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biochem Sci 1998; 23:324-328.
109. Marcotte EM, Pellegrini M, Ng HL et al. Detecting protein function and protein-protein interactions from genome sequences. Science 1999; 285:751-753.
110. Enright AJ, Iliopoulos I, Kyrpides NC et al. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999; 402:86-90.
111. Gavin AC, Bosche M, Krause R et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002; 415:141-147.
112. von Mering C, Krause R, Snel B et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nature 2002; 417:399-403.
113. Ho Y, Gruhler A, Heilbut A et al. Systematic identification of protein complexes in Saccharomyces Cerevisiae by mass spectrometry. Nature 2002; 415:180-183.
114. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol 2000; 18:1257-1261.