

Computing Biological Function:

***Bioinformatics approach to the analysis and prediction of
protein function***

Yanay Ofran

Marco Punta

Columbia University

Outline

Prediction of protein function is one of the major challenges of computational biology today. Some studies have attempted to predict function of individual proteins ((Whisstock and Lesk 2003), while others have offered ideas, tools and methods for high throughput, automated function prediction (Rost, Liu et al. 2003). In this tutorial we will present some of these tools and ideas, and discuss many of the open challenges in the field.

It is divided into two parts: First, we define "function", discuss the approaches used to study it *in-silico* and survey the available tools for function prediction. In this part we elaborate on open challenges and un-tackled research questions.

In the second part (see the tutorial Slides), we use the recent literature to illustrate, using real-life examples, the power, as well as some of the pitfalls, of current function prediction methods.

Automatic prediction methods

Gap between available and annotated sequences

The relative simplicity with which researchers can obtain the sequences of biological macromolecule led to the launching of several large-scale genome-sequencing projects. Combined with thousands of individual sequencing laboratories all over the world, these projects supply tremendous number of sequences. The pace at which these sequences accumulate far exceeds the ability of experimental biologists to process them and decipher their biochemical traits and biological functions. This pace is so rapid that it outgrows even computer integrated circuits, often considered to be the most rapidly advancing technological frontier. However, traditional methods for analyzing protein function deal with a single protein at a time. Expressing and purifying a protein, and studying its activity *in vitro* and *in vivo*, is laborious and may take a long time. Therefore, most new sequences remain without annotation. A tremendous effort is therefore invested in developing high throughput methods for the analysis and prediction of protein function. The goal of bioinformatics, in this context, is to devise computational tools that will help decipher the information that is encoded in these sequences, thus enabling the prediction of their structure and function.

The pressing need to annotate large numbers of newly sequenced proteins is not the only *raison d'être* of *in silico* function analysis. While exploring the molecular minutia of a single protein may reveal its function, comparing and analyzing thousands or even millions of proteins could sometimes be as, if not more, revealing. Large-scale analysis of this sort could arguably be performed only using computerized tools. In the next sections we will discuss some of the challenges of function prediction and survey the tools and approaches that were developed to explore protein function *in silico*.

How is this gap treated for protein structures?

A key notion of the Central Dogma of Molecular Biology is the hierarchical structure in which biological information flows. The DNA and RNA sequences include the information that is encoded into the sequence of proteins. This sequence largely determines to what three dimensional structure the protein will fold (Anfinsen 1973). The most successful approach in structure prediction is based on the simple notion of using a known structure to predict the structure of proteins with similar sequences.

The first entire genome (DNA) sequence of a free-living organism, *Haemophilus influenzae*, was published in 1995 (Fleischmann, Adams et al. 1995). Currently, we know the entire genomic sequence for over 100 organisms; for over 60 of these the data is publicly available and contributes about 250,000 protein sequences, i.e. about one fourth of all currently known protein sequences (Liu and Rost 2001; Liu and Rost 2002; Carter, Liu et al. 2003; Pruess, Fleischmann et al. 2003). Only for a small fraction of them is there an experimentally determined structure available. Computational biology plays a central role in bridging this gap (Fleischmann, Moller et al. 1999; Holm and Sander 1999; Luscombe, Laskowski et al. 2001; Thornton 2001; Valencia 2002; Valencia and Pazos 2002): For about 40% of all sequences, we can deduce structure from homology to known structures (Gerstein and Levitt 1997; Teichmann, Chothia et al. 1999; Wolf, Brenner et al. 1999; Moult and Melamud 2000; Liu and Rost 2001; Vitkup, Melamud et al. 2001; Liu and Rost 2002).

Protein sequence determines, to a large extent, where in the cell it will reside, with which other molecules it will interact, what biochemical and physiological tasks it will be able to carry out, and eventually when and how it will be broken down and reduced to its building blocks. In short - the function, or in the case of a disease, the malfunction, of every protein is encoded in its sequence of amino acids. Can we use annotation transfer between similar sequences to predict function, the same way this is done to predict structure?

Structure prediction methods applicable to function?

Annotation transfer - Sequence comparison

For about 60% of all sequences from current genome projects sequence homology suggests some aspects of function (Bork, Ouzounis et al. 1992; Andrade, Brown et al. 1999; Iliopoulos, Tsoka et al. 2001; Koonin 2001). However, drawing a firm conclusion about function is not always possible. Various analyses have established that sequence similarity above a certain cutoff ascertains similarity in structure. Even though the opposite is not always true, and proteins of similar structure may sometime lack any detectable similarity, this relationship between sequence and structure makes structural homology modeling a reliable way to predict structure of newly sequenced proteins. However, the relationship between sequence and function, and even structure and function, is more complicated. It is rather simple to establish a sequence similarity cutoff that would ascertain that a pair of proteins folds into the same structure. It is not yet clear how to determine a similar cutoff for function.

Several studies have shown that the precise values for thresholds of significant sequence similarity (T) that would imply function similarities are specific to particular aspects of function and have to be re-established for any given task (Shah and Hunter 1997; Ouzounis, Perez-Irratxeta et al. 1998; Devos and Valencia 2000; Pawlowski, Jaroszewski et al. 2000; Wilson, Kreychman et al. 2000; Todd, Orengo et al. 2001; Nair and Rost 2002; Rost 2002; Wrzeszczynski and Rost 2003). The problem of annotating function was illustrated immediately after the release of the first genome: 148 amendments were published a few weeks after the original publication (Casari, Andrade et al. 1995). Similar amendments followed most papers presenting entirely sequenced genomes (Ouzounis, Casari et al. 1996; Kyripides and Ouzounis 1998; Kyripides and Ouzounis 1999). Several pitfalls in transferring annotations of function have been reported, e.g. inadequate knowledge of thresholds for 'significant sequence similarity', or using only the best database hit, or ignoring the domain organization of proteins (Galperin and Koonin 1998; Kyripides and Ouzounis 1998; Brenner 1999; Kyripides and Ouzounis 1999; Mushegian 2000; Devos and Valencia 2001; Tamames, Gonzalez-Moreno et al. 2001). However, Eugene Koonin and colleagues turned the issue of annotation transfer errors around by collecting a few examples for which subsequent experiments showed that theoretical predictions had been more accurate than previous experiments (Iyer, Aravind et al. 2001).

What is function?

Structural similarity is a measurable magnitude, making structure prediction and its assessment a fairly rigorous domain. It is easy to score a prediction according to its similarity to the experimentally determined structure, and it is relatively straightforward to group proteins according to their structure in order to create a training set for a machine learning algorithm. Function, however, is a fuzzy term. When referring to protein function, different people mean different things. While biophysicists refer to physico-chemical characteristics of a protein as its function, biochemists are more likely refer to its biochemical traits or to the biosynthetic pathways in which it is instrumental. Molecular biologists may refer to its cellular role, and others, such as physiologists, developmental biologists or neurobiologists, may refer to its role in the context of the tissue or even the entire organism.

Several groups and associations have ventured to solve this problem by introducing rigorous schemata to define function. One of the first attempts was the introduction of Enzyme Classification (EC (Webb 1992)); this classification uses four digits to classify enzymatic activity (Todd, Orengo et al. 1999). The MIPS database attempts

to extend this idea to a wider spectrum of more proteins and more roles through its classification catalogue (Mewes, Frishman et al. 2002). Another characterization of protein function originates from the Gene Ontology (GO) consortium (Ashburner, Ball et al. 2000). GO distinguishes three levels of protein function. (1) *Molecular function*: at the molecular level, the protein can, for example, catalyze a metabolic reaction or transmit a signal. (2) *Biological process*: a set of many co-operating proteins is responsible for achieving broad biological goals, for example, mitosis or purine metabolism, or signal transduction cascades (3) *Cellular component*: this category includes the structure of sub-cellular compartments, the localization of proteins, and macromolecular complexes. Examples include *nucleus*, *telomere*, and *origin recognition complex*. The sub-cellular localization of a protein is an essential attribute for this level. Although not complete, GO constitutes the best set of definitions available today.

Problem: functional information not machine-readable

Nearly all databases present the protein sequence and structures are in formats that are more or less straightforward to parse by computers. Thus, researchers can construct large data sets of protein structure and use them to train machine-learning algorithms and develop structure prediction tools. However, functional annotations are mostly written in free text using a rich biological vocabulary that often varies in different areas of research. Such annotations are primarily meant for the eyes of human experts, hence, they are not machine-readable (Eisenhaber and Bork 1998). Another problem that hampers automatic annotations is the quality of database annotations: only a few database groups attempt quality control of curated annotations (Tsoka and Ouzounis 2001).

The reliability of transfer by homology depends on the particular feature of function/structure considered. In order to estimate the accuracy in transferring function given a particular threshold in sequence similarity, we have to complete the following three steps:

1. Build data sets that have experimental annotations about the presence (true, e.g. all proteins experimentally known to be nuclear) and absence (false, e.g. all proteins experimentally known NOT to be nuclear) of a certain aspect of function.
2. In order to avoid estimates that are incorrectly biased by the distribution of today's experimental information (Rost 2002), a representative sub-set of sequence-unique proteins from the true data has to be extracted and aligned

against all proteins in the true set (minus the representative sub-set) and false set.

3. For all alignments, we then have to count how many true and false we find at every given threshold for sequence similarity.

How should one measure sequence similarity? The most popular way is the level of pairwise sequence identity, i.e. the percentage of residues that are identical in an alignment of two proteins (R on R \rightarrow 1, R on K \rightarrow 0). The major problem with such a score in the context of automatic annotations is that it does not reflect the length of the alignment. For example, a stretch of 11 identical residues may differ in both function and structure (Rost 1999; Nair and Rost 2002; Rost 2002). On the other hand, levels of pairwise sequence identity of around 33% for alignments longer than 100 residues, or 22% for alignments longer than 250 residues imply similarity in structure (Sander and Schneider 1991). This observation is used to compile an empirical threshold for significant sequence similarity as a function of alignment length (Nielsen, Engelbrecht et al. 1996). We refer to this threshold as the HSSP-value; it is empirically chosen such that any pair of proteins A, B have similar structure if HSSP-value (A,B)>0. Another measure of sequence similarity is the expectation value built into the popular PSI-BLAST (Altschul, Madden et al. 1997) alignment program. An important point to realize for BLAST and PSI-BLAST users is that the expectation value depends on the database used to search for related proteins. This implies the following: assume we align proteins A and B by pairwise BLAST in two ways: (i) by searching with A against SWISS-PROT, and (ii) by searching with A against SWISS-PROT + PDB (Berman, Westbrook et al. 2000) . Even if the resulting alignments between A and B are identical, the expectation values may differ significantly due to the difference in size of the two databases.

Using these measurements of sequence identity it is possible to determine whether two aligned proteins are likely to share the same structure. However, when it comes to functional annotation transfer things get more complicated. Unfortunately, the accuracy of transferring different aspects of function differs substantially.

From sequence to structure, from structure to function

The attempt to extract biologically important information from protein sequence has been dominated in the last few decades by structure prediction. Since sequence and structure are so tightly interconnected, high throughput function prediction can benefit from the automated methods for structure prediction. Two kinds of tools for structure prediction are of particular interest in this context: prediction of solvent

accessibility and prediction of transmembrane segments. Active sites in proteins are most often exposed to the solvent so as to enable the interaction between the protein and its target. Hence identifying the solvent accessible residues in a protein is an important step in zooming in on the functional residues.

Solvent accessibility

Most of the methods that predict solvent accessibility combine searches for sequence homologues, which are used to construct a sequence profile, and a machine learning algorithm, which uses the profile to predict the solvent accessibility of a residue.

Prediction methods

PHDacc and PROFacc: These two sister methods (Rost, Casadio et al. 1996), which are part of the PredictProtein service, are based on the same concept, the second being an improvement of the first. When a query sequence is submitted to the server the program perform a database search and constructs a sequence alignment using MaxHom (Sander and Schneider 1991). A neural network then assigns one of ten possible levels of exposure to each residue in the query sequence. These states could be translated into relative solvent accessibility, describing for each residue its accessibility to the solvent as a percentage of its surface area. Alternatively, the ten states could be grouped into a two state scheme: if more than 16% of the surface area is accessible to solvent it is defined as exposed; otherwise, the residue is considered to be buried. The 10-state scheme could also be used to predict solvent accessibility in terms of square Angstroms.

Jpred: Jpred (Cuff and Barton 2000) is a prediction service that uses profiles that are produced by HMM and by PSI-BLAST. A neural network uses these profiles to predict one of three categories of exposure: 0%, 5% and 25%. The output of predictions from two different Networks is combined to give an average relative solvent accessibility.

Transmembrane segments

The communication between a cell and its surroundings, be it a unicellular cell sensing its medium, or an animal or plant cell interacting with other cells in its vicinity and in other tissues, is based almost exclusively on proteins that are embedded in the cell's membrane and interact with molecules on both the inter-cellular and the extra-cellular sides. Integral membrane proteins compose, according to some estimates, 25% of the proteomes sequenced thus far (Melen, Krogh et al. 2003). Identifying these transmembrane proteins and deciphering their

molecular mechanisms is, therefore, of high interest in many fields of biomedicine. Typically, the transmembrane segments are classified into one of two classes according to their secondary structures: helix or strands. It is reasonable to assume that all transmembrane segments share common biophysical features. These common features are probably reflected in the protein sequence, and hence many bioinformaticians are attempting to develop methods that identify transmembrane segments *in-silico*. The basic biophysical requirement for a residue to be buried in the membrane is hydrophobicity. High hydrophobicity is what enables most of the transmembrane segments to remain in the membrane and avoid the solvent on either of its sides. Hence, the first and most basic methods focused on a search for long hydrophobic stretches of sequence. In 1982 Kyte and Doolittle (Kyte and Doolittle 1982) proposed a simple method to identify transmembrane segments of proteins, based on the analysis of hydropathy. Since that pioneering work, many groups have suggested improvements to Kyte and Doolittle's idea, as well as novel approaches to the problem. Some of them also offer prediction of the overall topology of the transmembrane protein. The fact that it is difficult to decipher the structure of transmembrane segments experimentally makes *in-silico* prediction both a greater challenge and a more valuable tool.

Prediction methods

TopPred: Combining hydrophobicity analysis with the analysis of electrical charges, TopPred (von Heijne 1992) was one of the first methods for the prediction of transmembrane segments and the topology of transmembrane proteins. When a protein sequence is submitted to TopPred, the program calculates a hydrophobicity profile for it. Sequence stretches that are found to be rich in hydrophobic residues are marked as transmembrane helices. Stretches that are hydrophobic but fail to surmount a predefined cutoff of hydrophobicity are considered "putative transmembrane helices". Finally, based on various rules, the predominant of which is the distribution of positive residues between the transmembrane helices, the overall topology of the protein is predicted, with and without the putative helices.

PHDhtm: PHDhtm (Rost, Casadio et al. 1996) is the part of the PredictProtein service dedicated to prediction of transmembrane helices. As in the cases of secondary structure and solvent accessibility prediction, the method first constructs a profile based on a database search and a multiple sequence alignment. Then, a neural network predicts for each residue whether it is likely to be part of a transmembrane helix. Another neural network is then used to decide whether the protein as a whole is a helix bundle integral membrane protein. Finally, the system

predicts the topology of the protein based on its similarity to known topologies of transmembrane proteins.

ProfTMB: Specializing in β -strands, this method (Bigelow, Petery et al. 2004) uses multiple sequence alignments to produce a profile that is fed to an HMM. The HMM is trained on examples from one group of membrane proteins known as beta-barrels - proteins that reside in the outer membrane of gram-negative bacteria, mitochondria and chloroplasts - thus is particularly potent in finding sequences that belong to this family. The service predicts whether the query sequence belongs to this class, and identifies the transmembrane β -strands.

SOSUI: This server (Hirokawa, Boon-Chieng et al. 1998) bases its predictions on four parameters: First it calculates the hydropathy based on the Kyte-Doolittle index. Then it calculates charges of the residues and the amphiphilicity, namely the distribution of electric charges around the helix. Finally, the length of the sequence is incorporated into the calculation. The output of the program includes a graph with the hydropathy profile of the query sequence, and a helical wheel diagram of the predicted transmembrane segments. This representation shows the different features of the helix residues and enables the visualizations of the biophysical traits of the helix as a whole.

TMHMM: TMHMM (Krogh, Larsson et al. 2001) uses hidden Markov models (HMM) to predict transmembrane segments and the topology of the transmembrane proteins. Many machine learning algorithms are designed to identify patterns in ostensibly irregular sequences. Among these, HMM is particularly useful in matching a sequence to a predefined “grammar”. Transmembrane proteins tend to obey a relatively strict “grammar” – with alternating segments of membrane and non-membrane segments and a well-defined organization of positively charged residues. Using HMM, TMHMM tries to match the query sequence to this “grammar”, derived from a set of well-characterized transmembrane proteins. By searching a known transmembrane “grammars” to which the query protein obeys, TMHMM predicts the segments that are most likely to be transmembrane and the most likely topology of the whole protein.

DAS: The Dense Alignment Surface (DAS) method (Cserzo, Wallin et al. 1997) assesses the sequence similarities between segments of the query proteins and known transmembrane segments. Thus, it identifies those sequence stretches in the query sequence that are likely to be transmembrane by virtue of their biophysical similarity to stretches that were shown experimentally to be integrated in the membrane.

Functional residues, active sites and interaction sites

The attempt to develop automated tools for function prediction includes the development of various methods to identify functionally important residues based on their conservation throughout evolution (Casari, Sander et al. 1995). If a residue is highly conserved it is likely to be functionally important. Several methods offer tools that vary from identifying residues that have functional importance (such as SequenceSpace, ConSurf and ISIS, see slides for details) to the identification of highly specialized functional elements such as DNA or protein binding sites, or regulatory elements in DNA sequence.

Motifs and Patterns

Another way to identify functional elements, or sequence signatures that are associated with a certain function, is through sequence motifs and patterns. Sometimes, the divergence between the sequence of a newly discovered protein and any other annotated protein is too wide to establish relatedness based on simple pairwise sequence alignment. But the existence of a relatively short sequence motif that is highly conserved evolutionarily and highly specific functionally within this newly discovered sequence might surrender the function of this protein. For example, if we find in a newly-sequenced protein a sequence element that appears in many known DNA binding sites, we can predict that the function of our new protein involves an interaction with DNA. A few databases are dedicated to this idea. They offer a large library of sequence motifs that have been collected either manually by experts, or automatically by pattern-searching algorithms. Many of these libraries include a searching tool. When a query sequence is submitted to these tools, it is compared with all the known motifs in search of a match. Finding one of these well-characterized motifs in a newly discovered sequence could offer some insights into its structure, function and even mechanisms of action.

Prediction methods

PROSITE: Developed and maintained by the team that maintains SWISS-PROT, PROSITE (Falquet, Pagni et al. 2002) is a large collection of biologically important motifs that is curated manually. The database contains three types of motifs: patterns, rules and profiles; each represents a different automated method of searching for motifs. These methods were applied to SWISS-PROT to construct a large database of motifs. Every entry in PROSITE includes a description of the proteins that it is designed to detect and the reason for including it in the database. The close relationship between SWISS-PROT and PROSITE, is most beneficial when it comes to annotations. The wealth of information included in each database

also benefits the annotations of its sister database as they are often updated together by their developers. It is possible to search PROSITE using free text to mine the annotation, or using a SWISS-PROT / TrEMBL ID. It is also possible to use ScanPro, a search tool for scanning a sequence against PROSITE.

Blocks: Blocks (Petrokovski, Henikoff et al. 1996) is a database of motifs that has been produced automatically, from an ungapped multiple sequence alignment of the most conserved regions of proteins. Blocks offer a large database of motifs that have been gleaned from InterPro (Apweiler, Attwood et al. 2000) - a database of protein families, domains and functional sites, that is an integration of many motif libraries. The Blocks that are produced from InterPro can be searched using a search tool called Blocks searcher. Blocks maker, the tool that was used to produce the Blocks database, is offered to the users who want to produce their own Blocks from a dataset they have constructed.

Pfam: Using hidden Markov models (HMM) Pfam (Sonnhammer, Eddy et al. 1997), offers a powerful tool for producing motifs from alignments and for finding them in a query sequence. Based on this tool the developers built a manually curated database of protein families. Pfam was used by several genome projects (including the human and the fly) for high throughput annotation of the function of newly discovered genes (Bateman, Birney et al. 2002). Each protein family is represented in Pfam by a set of well-characterized proteins, that are used to train the HMM, and additional sequences that are obtained when the trained model is used to search for new members of the family. The annotation in Pfam includes a description of each family and links to other resources and literature references.

Sub-cellular localization

The methods we covered so far have all been based on the notion of annotation transfer. That is, finding similarities between a query sequence and other proteins that have been thoroughly characterized experimentally. However, in recent years attempts have been made to develop tools that will decipher the function of a protein from its sequence even when the most sophisticated tools for annotation transfer yield no results.

When the pioneers of structure prediction launched their enterprise, one of their first steps was to break down the somewhat fuzzy concept of “structure”, into well-defined structural features like “secondary structures” or “topologies”, which we discussed above, and other concepts like “structural family” or “fold” which will be discussed in the next chapter. If “structure” is a fuzzy concept that requires a meticulous set of subcategories, then “function”, as we explained above, is even

more so. What we usually refer to as the “function” of a protein could be purely biochemical (such as “phosphorylation”), cellular (e.g. “cytoskeletal protein”), physiological or pertaining to the organism as a whole (e.g. “developmental”). Each of these implications of “function” depends on different biophysical and biochemical features of the proteins, and hence is probably encoded differently in its sequence. Therefore, if we want to predict function from sequence, a first step would be to define which aspect of function we attempt to predict. The eukaryotic cell has many compartments, each of which host very different biochemical and biological processes, carried out by different proteins. Identifying the sub-cellular localization of a newly discovered sequence is a crucial step in finding the process in which it partakes and what its function may be. A few groundbreaking works in recent years have shown that, in many cases, it is possible to predict the sub-cellular localization of a protein from its sequence.

Prediction methods

SUBLOC: Using the amino acid composition alone, SUBLOC (Hua and Sun 2001) applies Support Vector Machine (SVM) to predict in which sub-cellular locale a protein resides. It offers one of three localizations for prokaryotes (extracellular, periplasmic, cytoplasmic) and four for eukaryotes (extracellular, mitochondrial, cytoplasmic, nuclear).

PSORT: PSORT (Nakai and Horton 1999) receives as an input the amino acid sequence of a protein and the type of organism from which it was obtained (gram positive bacteria, gram-negative bacteria, yeast, animal or plant). Based on the origin of the protein the system checks for a few sub-cellular localizations (e.g. chloroplast for plant cells). The program then searches for several features that may reflect the sub-cellular localization of the protein. For instance, it has been found that the trafficking of proteins to some sub-cellular compartments is dictated by short signal peptides at the N or C terminal of the protein. PSORT employs a library of the known signals peptides and searches for them in the query sequence. It also checks predicted structural features (such as topology that may indicate that protein is transmembrane, amino acid composition, and PROSITE motifs).

TargetP: TargetP (Emanuelsson, Nielsen et al. 2000) focuses on signal peptides at the N-terminal end of a protein. It uses a series of machine learning algorithms including neural networks and SVM to identify signal peptides of three types: chloroplast transit peptides, mitochondrial targeting peptides and secretory pathway signal peptides. The interface of TargetP enables the user to define the desired specificity of the prediction.

LOC3D: LOC3D (Nair and Rost 2003) is a database of predicted sub-cellular localization for eukaryotic proteins of known 3-D structure. It uses three methods: predicNLS, which searches for a known nuclear localization signal, LOChom, which uses homology to determine localization, and LOC3D, which is a neural network based prediction method. LOCkey, a related service uses keywords in SWISS-PROT annotation to predict the sub-cellular localization. Altogether, this suite offers comprehensive coverage of the methods and approaches suggested and implemented so far for prediction of sub-cellular localization.

Functional class

Monica Riley introduced the most widely used schema for classes of cellular function to annotate *E. coli* (Riley 1993). TIGR (The Institute for Genome Research) and many other genome centers have adopted this schema with minor modifications. Transferring annotations of cellular function by homology has for long been almost the only field in which methods were developed. In fact, many researchers exclusively consider such methods when referring to the prediction of protein function. However, recently groups have begun developing methods that predict functional classes in the absence of experimental annotations.

Functional classes can be predicted from sequence. An interesting hybrid system uses inductive logic programming to predict functional classes with and without homology to experimentally annotated proteins (Clare and King 2002). While it is not clear how successful the system is in *ab initio* prediction, the levels of accuracy published on average appear promising. Genes located in a close neighborhood on the genome may have some functional commonalities. While such neighborhood relations sometimes enable prediction of aspects such as classes of cellular function, the average signal is very weak, i.e. most often neighbors are not related in function (Tamames, Casari et al. 1997; Overbeek, Fonstein et al. 1999; Galperin and Koonin 2000). The most recent breakthrough in the field of predicting protein function came through a collaboration of the groups from Soren Brunak (CBS Copenhagen) and Alfonso Valencia (CNB Madrid). Their ends are to predict cellular function from sequence alone. Their means are complex, elaborate, and hierarchical systems of neural networks (Jensen, Gupta et al. 2002). A first group of networks is used to identify 'sequence features' (like protein length or amino acid composition) that optimally separate between any two types of functional classes. These basic predictions are then combined into a final prediction step, again through neural networks. The authors applied their method to annotating functional classes for all human proteins. For example, the prion protein is predicted to belong to the

'transport and binding category' and to 'not have enzymatic activity'. This appears compatible with the observation that the prion binds and transports copper while no catalytic activity has ever been observed (Brown 2002). Recently, the Brunak group have applied their new concepts to identifying novel enzymes in archae (Jensen, Skovgaard et al. 2002) and to predicting the functional type of all human proteins according to the GO classification (Jensen, Gupta et al. 2003). The most impressive news from these ground-breaking methods is that aspects of function can be predicted without homology, i.e. for completely uncharacterised proteins.

Prediction methods

EUCLID: This method (Tamames, Ouzounis et al. 1998) uses the keywords in SWISS-PROT to assign a protein to one of Reilly's functional classes. The algorithm at the heart of this method is a basic Machine learning algorithm that learns, based on a manually curated training-set, which composition of keywords is most likely to indicate that the protein belongs to a certain functional type. The developers report that in more than 90% of the cases the functional type that was determined by the automated method was identical to the one that was assigned to it by a human experts. However, EUCLID requires that some annotation, namely SWISS-PROT keywords, would already be assigned to the sequence. Thus, it is not really a method for prediction from sequence. Having only the sequence of a newly discovered protein would not allow one to use EUCLID.

ProtFun: ProtFun (Jensen, Gupta et al. 2003) represents a recent and promising step towards the prediction of function from sequence. To define a functional type, ProtFun uses Gene Ontology (GO). Each protein could be assigned to a certain molecular function, a certain biological process and certain cellular component. GO is attempting to assign a number to each protein that will represent these three types of functional description. Currently there are many hundreds of GO categories. ProtFun focuses on 347 of them and uses complex systems of neural networks to predict the GO functional classification of a protein from its sequence. The developers report an impressive accuracy – in most cases more than 90% of their predictions are correct. However, currently, their coverage is only partial, and many of the query proteins are returned without any prediction. Yet, when it does give a prediction, in most cases it is correct.

Conclusion

The ability to analyze large amounts of data simultaneously enables computational biologists to compile large datasets of functionally similar proteins and use them to predict function. By comparing many proteins of a similar function one could identify

typical characteristics in the sequence or the structure of these proteins. The characteristics features could then be used to search among vast numbers of unannotated sequences for other proteins that may have the same function. A large number of methods, some of them surveyed above, already offer predictions based on this concept. However, the field is still in its infancy and each of the stages in this process could be enhanced and improved.

Sub-cellular localization and some functional sites can be predicted with high accuracy from sequence. One challenge in the field is to define other functional aspects that could be predicted from sequence or structure.

Thousands of sequence motifs and patterns are available in different databases and could be used for prediction. Improving the methods for finding motifs and patterns automatically and associating them with functions remains a major challenge.

Structural motifs and patterns are hard to identify in annotated proteins and harder to search for in unannotated ones. Improved tools that will combine structural alignment with biophysical and spatial analysis may constitute a breakthrough in this arena.

Finally determining the sequence similarity threshold for each function is a continuous effort that requires a wise choice of sequence alignment parameters and a cautious utilization of available sequences.

Almost all of these methods depend on large veritable datasets for training. Better methods for data mining that would lead to larger and cleaner datasets are one of the major keys for the progress of the field. Tools that are based on the conceptual approach we described can automatically predict function, or some aspects of it, for a large number of proteins in a relatively short time. Hence they are very useful for a high throughput annotation of whole genomes or large datasets. The results of these tools could also be used for the analysis of single proteins by theoreticians or experimentalists. These points are illustrated in the tutorial second part (Part II: Practical Examples, see the tutorial Slides).

REFERENCES

- Altschul, S., T. Madden, et al. (1997). "Gapped Blast and PSI-Blast: a new generation of protein database search programs." Nucleic Acids Research **25**: 3389-3402.
- Andrade, M. A., N. P. Brown, et al. (1999). "Automated genome sequence analysis and annotation." Bioinformatics **15**(5): 391-412.
- Anfinsen, C. B. (1973). "Principles that govern the folding of protein chains." Science **181**: 223-230.
- Apweiler, R., T. K. Attwood, et al. (2000). "InterPro--an integrated documentation resource for protein families, domains and functional sites." Bioinformatics **16**(12): 1145-1150.
- Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature Genetics **25**(1): 25-29.
- Bateman, A., E. Birney, et al. (2002). "The Pfam protein families database." Nucleic Acids Res **30**(1): 276-80.
- Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Research **28**: 235-242.
- Bigelow, H., D. Petery, et al. (2004). "ProfTMB: A Profile HMM for two-state residue and whole-protein prediction of transmembrane beta-barrels." proteins: Structure, Function, and Genetics in preparation.
- Bork, P., C. Ouzounis, et al. (1992). "What's in a genome?" Nature **358**: 287.
- Brenner, S. E. (1999). "Errors in genome annotation." Trends in Genetics **15**: 132-133.
- Brown, D. R. (2002). "Copper and prion diseases." Biochem Soc Trans **30**(4): 742-5.
- Carter, P., J. Liu, et al. (2003). "PEP: Predictions for Entire Proteomes." Nucleic Acids Res **31**(1): 410-3.
- Casari, G., M. A. Andrade, et al. (1995). "Challenging times for bioinformatics." Nature **376**: 647-648.
- Casari, G., C. Sander, et al. (1995). "A method to predict functional residues in proteins." Nature Structural Biology **2**: 171-178.
- Clare, A. and R. D. King (2002). "Machine learning of functional class from phenotype data." Bioinformatics **18**(1): 160-6.
- Cserzo, M., E. Wallin, et al. (1997). "Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method." Protein Eng **10**(6): 673-6.
- Cuff, J. A. and G. J. Barton (2000). "Application of multiple sequence alignment profiles to improve protein secondary structure prediction." Proteins: Structure, Function, and Genetics **40**(3): 502-511.
- Devos, D. and A. Valencia (2000). "Practical limits of function prediction." Proteins: Structure, Function, and Genetics **41**: 98-107.
- Devos, D. and A. Valencia (2001). "Intrinsic errors in genome annotation." Trends in Genetics **17**(8): 429-431.
- Eisenhaber, F. and P. Bork (1998). "Wanted: subcellular localization of proteins based on sequence." Trends in Cell Biology **8**: 169-170.
- Emanuelsson, O., H. Nielsen, et al. (2000). "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence." Journal of Molecular Biology **300**(4): 1005-1016.
- Falquet, L., M. Pagni, et al. (2002). "The PROSITE database, its status in 2002." Nucleic Acids Research **30**(1): 235-238.

- Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd." Science **269**: 496-512.
- Fleischmann, W., S. Moller, et al. (1999). "A novel method for automatic functional annotation of proteins." Bioinformatics **15**(3): 228-233.
- Galperin, M. Y. and E. V. Koonin (1998). "Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption." In Silico Biol **1**(1): 55-67.
- Galperin, M. Y. and E. V. Koonin (2000). "Who's your neighbor? New computational approaches for functional genomics." Nature Biotechnology **18**(6): 609-613.
- Gerstein, M. and M. Levitt (1997). "A structural census of the current population of protein sequences." Proceedings of the National Academy of Sciences **94**(22): 11911-11916.
- Hirokawa, T., S. Boon-Chieng, et al. (1998). "SOSUI: classification and secondary structure prediction system for membrane proteins." Bioinformatics **14**(4): 378-379.
- Holm, L. and C. Sander (1999). "Protein folds and families: sequence and structure alignments." Nucleic Acids Research **27**(1): 244-247.
- Hua, S. and Z. Sun (2001). "Support vector machine approach for protein subcellular localization prediction." Bioinformatics **17**(8): 721-8.
- Iliopoulos, I., S. Tsoka, et al. (2001). "Genome sequences and great expectations." Genome Biology **2**(1): interactions 2000.
- Iyer, L. M., L. Aravind, et al. (2001). "Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences." Genome Biology **2**(12): RESEARCH0051.
- Jensen, L. J., R. Gupta, et al. (2002). "Prediction of human protein function from post-translational modifications and localization features." Journal of Molecular Biology **319**(5): 1257-1265.
- Jensen, L. J., R. Gupta, et al. (2003). "Prediction of human protein function according to Gene Ontology categories." Bioinformatics **19**(5): 635-42.
- Jensen, L. J., M. Skovgaard, et al. (2002). "Prediction of novel archaeal enzymes from sequence-derived features." Protein Sci **11**(12): 2894-8.
- Keller, J. P., P. M. Smith, et al. (2002). "The crystal structure of MT0146/CbiT suggests that the putative precorrin-8w decarboxylase is a methyltransferase." Structure (Camb) **10**(11): 1475-87.
- Koonin, E. V. (2001). "Computational genomics." Curr Biol **11**(5): R155-8.
- Krogh, A., B. Larsson, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." Journal of Molecular Biology **305**(3): 567-580.
- Kyrpides, N. C. and C. A. Ouzounis (1998). "Errors in genome reviews." Science **281**(5382): 1457.
- Kyrpides, N. C. and C. A. Ouzounis (1999). "Whole-genome sequence annotation: 'Going wrong with confidence'." Mol Microbiol **32**(4): 886-7.
- Kyte, J. and R. F. Doolittle (1982). "A simple method for displaying the hydrophathic character of a protein." Journal of Molecular Biology **157**: 105-132.
- Liu, J. and B. Rost (2001). "Comparing function and structure between entire proteomes." Protein Science **10**(10): 1970-1979.
- Liu, J. and B. Rost (2002). "CHOP proteomes into structural domains." Bioinformatics: in preparation.
- Liu, J. and B. Rost (2002). "Target space for structural genomics revisited." Bioinformatics **18**: 922-933.
- Luscombe, N. M., R. A. Laskowski, et al. (2001). "Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level." Nucleic Acids Research **29**(13): 2860-2874.

- Melen, K., A. Krogh, et al. (2003). "Reliability measures for membrane protein topology prediction algorithms." J Mol Biol **327**(3): 735-44.
- Mewes, H. W., D. Frishman, et al. (2002). "MIPS: a database for genomes and protein sequences." Nucleic Acids Research **30**(1): 31-34.
- Moult, J. and E. Melamud (2000). "From fold to function." Current Opinion in Structural Biology **10**(3): 384-389.
- Mushegian, A. R. (2000). "Annotations of biochemically uncharacterized open reading frames (ORFs)." Mol Microbiol **35**(3): 697-8.
- Nair, R. and B. Rost (2002). "Sequence conserved for subcellular localization." Protein Sci **11**(12): 2836-47.
- Nair, R. and B. Rost (2003). "Better prediction of sub-cellular localization by combining evolutionary and structural information." Proteins **53**(4): 917-30.
- Nakai, K. and P. Horton (1999). "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization." Trends Biochem Sci **24**(1): 34-6.
- Nielsen, H., J. Engelbrecht, et al. (1996). "Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site." Proteins: Structure, Function, and Genetics **24**: 165-177.
- Ouzounis, C., G. Casari, et al. (1996). "Computational comparisons of model genomes." Trends in Biotechnology **14**: 280-285.
- Ouzounis, C., C. Perez-Irratxeta, et al. (1998). "Are binding residues conserved?" Pac Symp Biocomput **3**: 399-410.
- Overbeek, R., M. Fonstein, et al. (1999). "Use of contiguity on the chromosome to predict functional coupling." In Silico Biol **1**(2): 93-108.
- Pawlowski, K., L. Jaroszewski, et al. (2000). "Sensitive sequence comparison as protein function predictor." Pac Symp Biocomput **8**: 42-53.
- Petrokovski, S., J. G. Henikoff, et al. (1996). "The Blocks database- a system for protein classification." Nucleic Acids Research **24**: 197-201.
- Pruess, M., W. Fleischmann, et al. (2003). "The Proteome Analysis database: a tool for the in silico analysis of whole proteomes." Nucleic Acids Res **31**(1): 414-7.
- Riley, M. (1993). "Function of the gene products in Escherichia coli." Microbiol. Rev. **57**: 862-952.
- Rost, B. (1999). "Twilight zone of protein sequence alignments." Protein Engineering **12**(2): 85-94.
- Rost, B. (2002). "Enzyme function less conserved than anticipated." Journal of Molecular Biology **318**(2): 595-608.
- Rost, B., R. Casadio, et al. (1996). Refining neural network predictions for helical transmembrane proteins by dynamic programming. Fourth International Conference on Intelligent Systems for Molecular Biology, St. Louis, M.O., U.S.A., Menlo Park, CA: AAAI Press.
- Rost, B., J. Liu, et al. (2003). "Automatic prediction of protein function." Cell Mol Life Sci **60**(12): 2637-50.
- Sander, C. and R. Schneider (1991). "Database of homology-derived structures and the structural meaning of sequence alignment." Proteins: Structure, Function, and Genetics **9**: 56-68.
- Shah, I. and L. Hunter (1997). Predicting enzyme function from sequence: a systematic appraisal. Fifth International Conference on Intelligent Systems for Molecular Biology, Halkidiki, Greece, AAAI Press.
- Sonnhammer, E. L., S. R. Eddy, et al. (1997). "Pfam: a comprehensive database of protein domain families based on seed alignments." Proteins: Structure, Function, and Genetics **28**(3): 405-420.
- Tamames, J., G. Casari, et al. (1997). "Conserved clusters of functionally related genes in two bacterial genomes." J Mol Evol **44**(1): 66-73.

- Tamames, J., M. Gonzalez-Moreno, et al. (2001). "Bringing gene order into bacterial shape." Trends in Genetics **17**(3): 124-126.
- Tamames, J., C. Ouzounis, et al. (1998). "EUCLID: automatic classification of proteins in functional classes by their database annotations." Bioinformatics **14**(6): 542-3.
- Teichmann, S. A., C. Chothia, et al. (1999). "Advances in structural genomics." Current Opinion in Structural Biology **9**: 390-399.
- Thornton, J. M. (2001). "From genome to function." Science **292**(5524): 2095-2097.
- Todd, A. E., C. A. Orengo, et al. (1999). "Evolution of protein function, from a structural perspective." Curr Opin Chem Biol **3**(5): 548-56.
- Todd, A. E., C. A. Orengo, et al. (2001). "Evolution of function in protein superfamilies, from a structural perspective." Journal of Molecular Biology **307**(4): 1113-1143.
- Tsoka, S. and C. A. Ouzounis (2001). "Functional versatility and molecular diversity of the metabolic map of Escherichia coli." Genome Research **11**(9): 1503-1510.
- Valencia, A. (2002). "Bioinformatics: biology by other means." Bioinformatics **18**(12): 1551-2.
- Valencia, A. and F. Pazos (2002). "Computational methods for the prediction of protein interactions." Curr Opin Struct Biol **12**(3): 368-73.
- Vitkup, D., E. Melamud, et al. (2001). "Completeness in structural genomics." Nature Structural Biology **8**(6): 559-566.
- von Heijne, G. (1992). "Membrane protein structure prediction." Journal of Molecular Biology **225**: 487-494.
- Webb, E. C. (1992). Recommendation of the nomenclature committee of the international union of biochemistry and molecular biology. new york, academic press.
- Whisstock, J. C. and A. M. Lesk (2003). "Prediction of protein function from protein sequence and structure." Quarterly Reviews of Biophysics **36**(3): 307-340.
- Wilson, C. A., J. Kreychman, et al. (2000). "Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores." Journal of Molecular Biology **297**(1): 233-249.
- Wolf, Y., S. Brenner, et al. (1999). "Distribution of protein folds in the three superkingdoms of life." Genome Research **9**: 17-26.
- Wrzeszczynski, K. O. and B. Rost (2003). Cataloguing proteins in cell cycle control. Cell cycle Checkpoint Control Protocols. L. H. Totowa, NJ, Humana Press: 219-233.

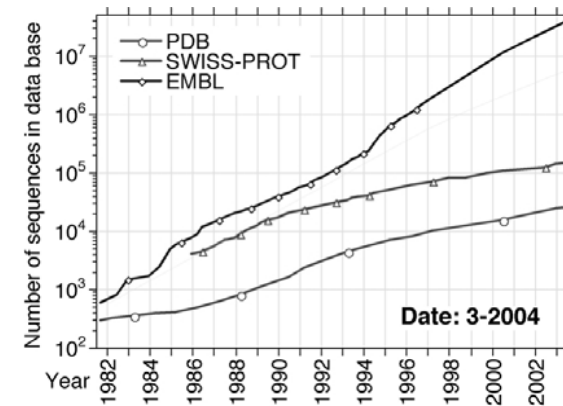


Computing Biological Function:
Bioinformatics approach to the analysis and prediction of protein function

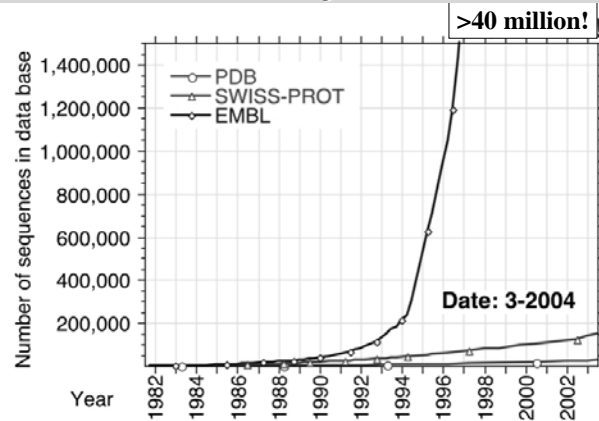
(Part 1)

Yanay Ofra & Marco Punta
Columbia University, New York

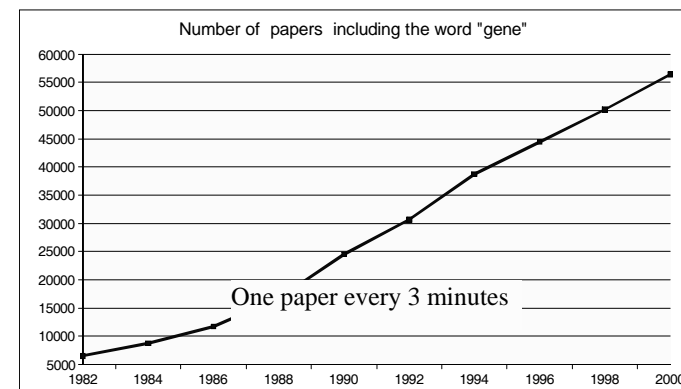
Growth of Biological Databases



Growth of Biological Databases



number of scientific papers in genetic



Part I (Yanay)

Part II (Marco)

Part I.I:The problem

Part I.II:
Solutions

The drill

Introduction

Annotation Transfer

Structure Prediction as a Model

Function by Association

Defining Function

Aspects of Function

Functional Type

High throughput predictions

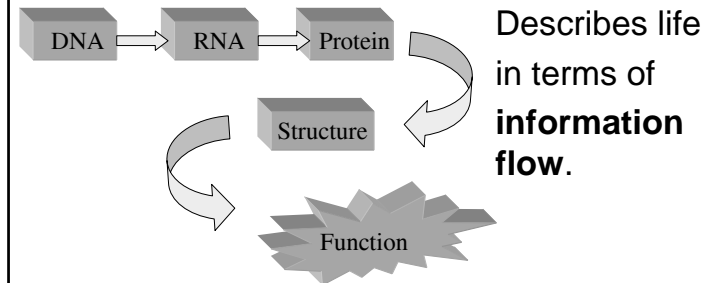
Conclusion



Fortaleza, Brazil
August 6-10, 2006

5

The Central Dogma of Molecular Biology

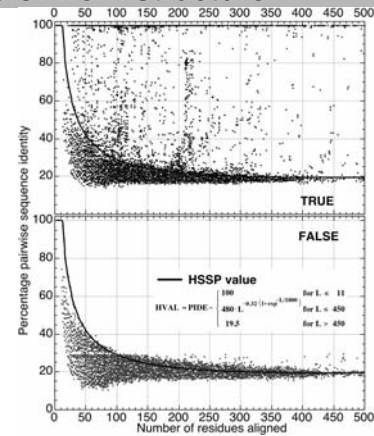


ISMB 2006
Fortaleza, Brazil
August 6-10, 2006

6

Annotation transfer from structure

Similar Structure



Dissimilar Structure

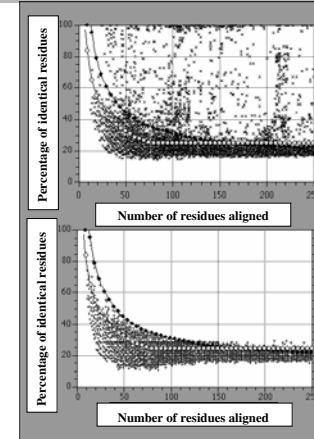


Fortaleza, Brazil
August 6-10, 2006

Rost (1999) *Protein Engineering* 12: 85-94

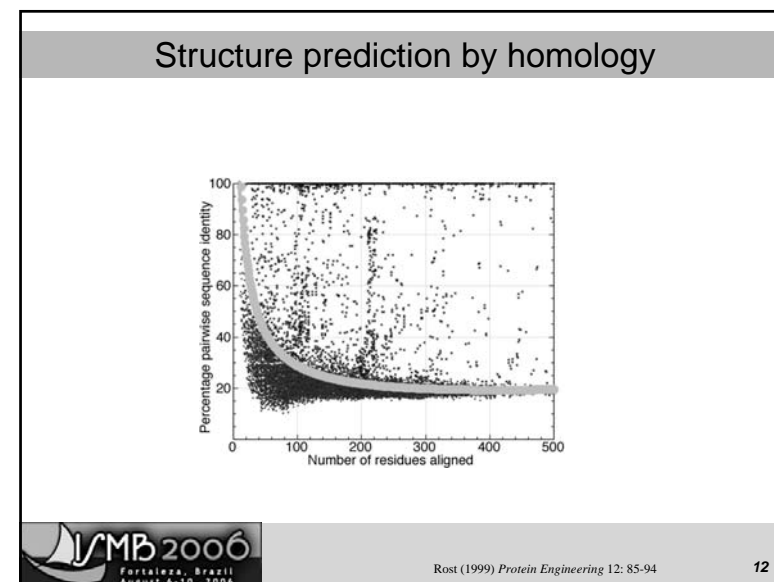
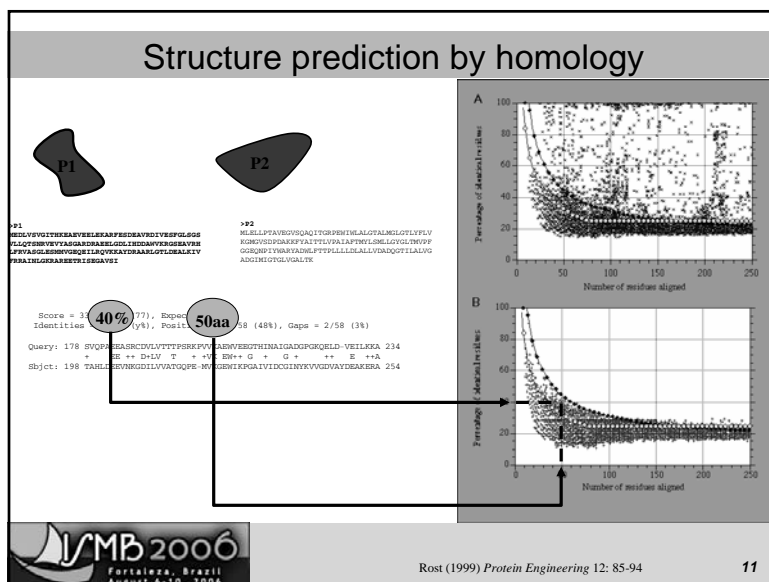
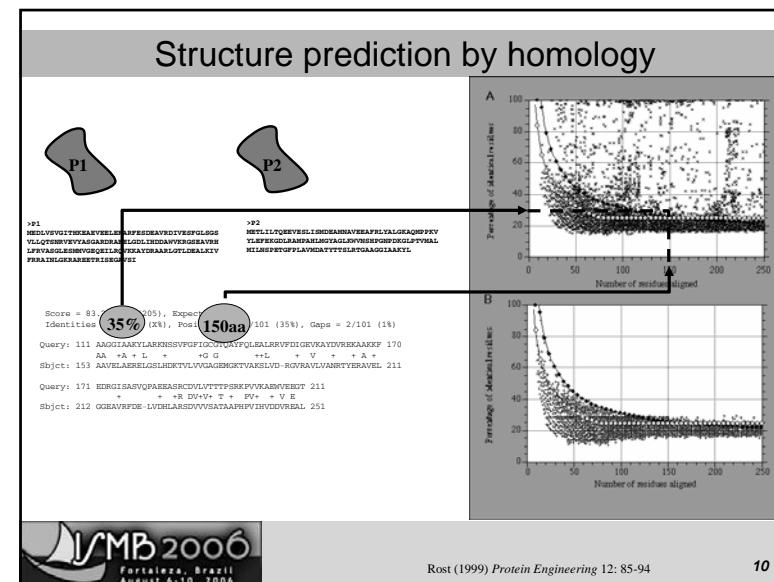
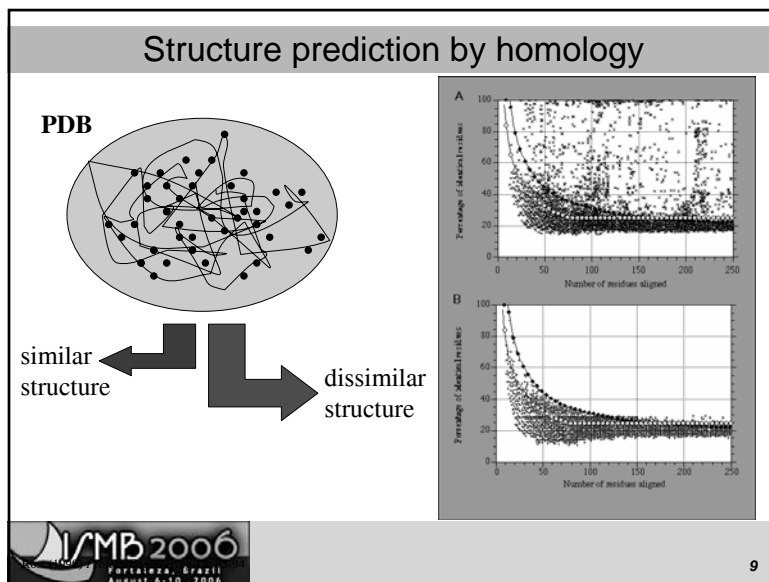
7

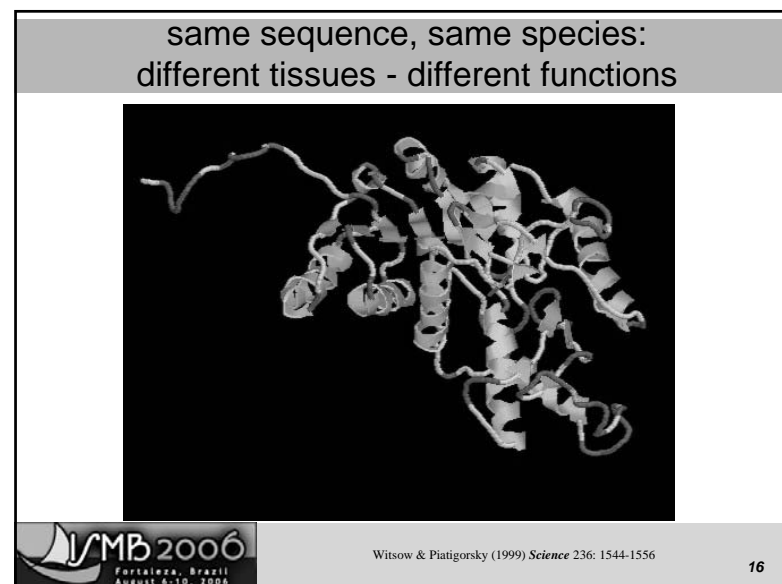
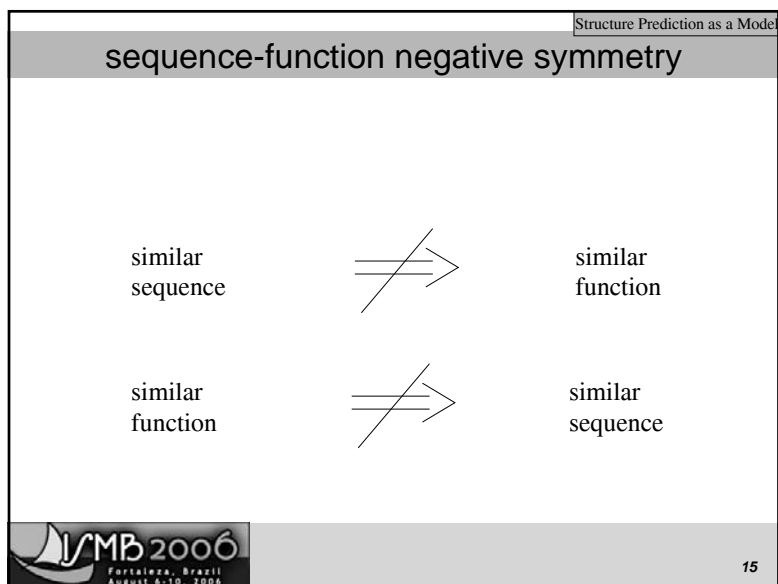
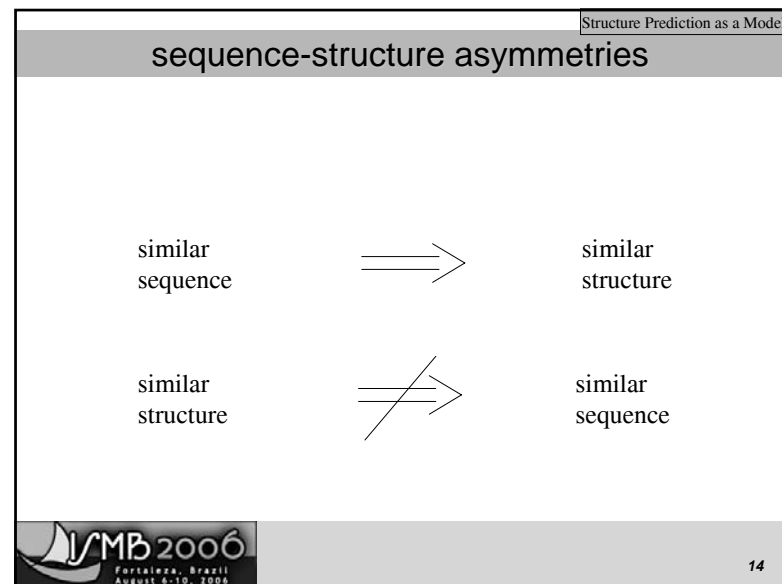
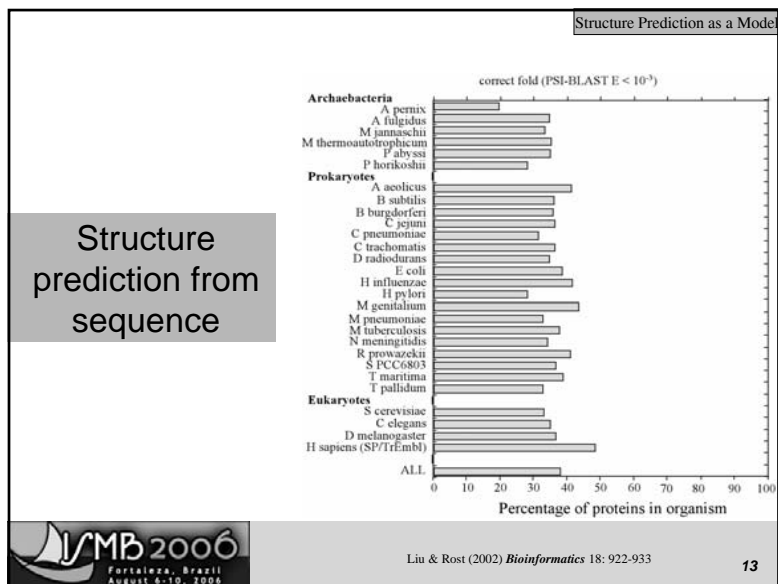
Structure prediction by homology



ISMB 2006
Fortaleza, Brazil
August 6-10, 2006

8





Structure Prediction as a Model

structure-function negative symmetry

similar
structure


⇒

similar
function

similar
function

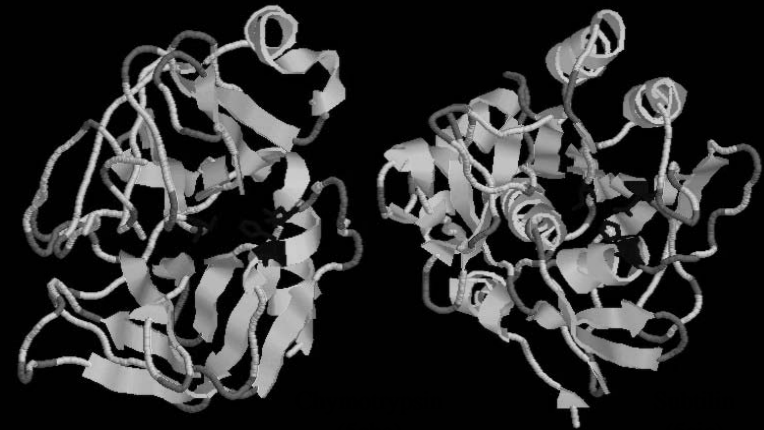
⇒


similar
structure


17

Structure Prediction as a Model


Different structures -> same function




18

Moonlighting proteins


One function	Additional functions
PutA proline dehydrogenase	Transcriptional repressor
Phosphoglucose isomerase	Neuroleukin, autocrine motility factor, differentiation and maturation mediator
Thymidine phosphorylase	Platelet-derived endothelial cell growth factor
Neuropilin (VEGF receptor)	Receptor for semaphorin III (nerve axons)
Uracil-DNA glycosylase	Glyoxaldehyde-3-phosphate dehydrogenase
Aconitase	Iron-responsive-element binding protein (IRE-BP)
Carbonylamine dehydratase	Dimerization enhancer (DeoH)
<i>Escherichia coli</i> thiorodotin	Subunit of T7 DNA polymerase
<i>E. coli</i> aspartate receptor	Maltose-binding-protein receptor
PMH2 mismatch-repair enzyme	Hypermutation of antibody variable chains
Ribosomal proteins	DNA repair, translational regulators, development regulators, etc.
Leuc crystallins	Heat-shock proteins, lactate dehydrogenase, argininosuccinate, retinaldehyde dehydrogenase, lyase, enolase, quinase
C/EBP alpha/cleaved	Regulator of other epithelial union channels
P-glycoprotein (transporter)	Regulator of cell-swelling ion channel
Thrombin protease	Ligand for cell surface receptors
Thymidylate synthase	Translation inhibitor
<i>E. coli</i> <i>bcrA</i> tripartite	Bio operon repressor
Mitochondrial LON protease	Chaperone
Bacterial PsaH chaperone	Metalloprotease
Band-3 anion exchanger	Regulator of glycolysis


Jeffery CJ (1999) Trends Biochem Sci. 24(1):8-11.
19

Defining Function

What is function?


- Biochemical
- Physiological
- Biological
- Behavioral


20

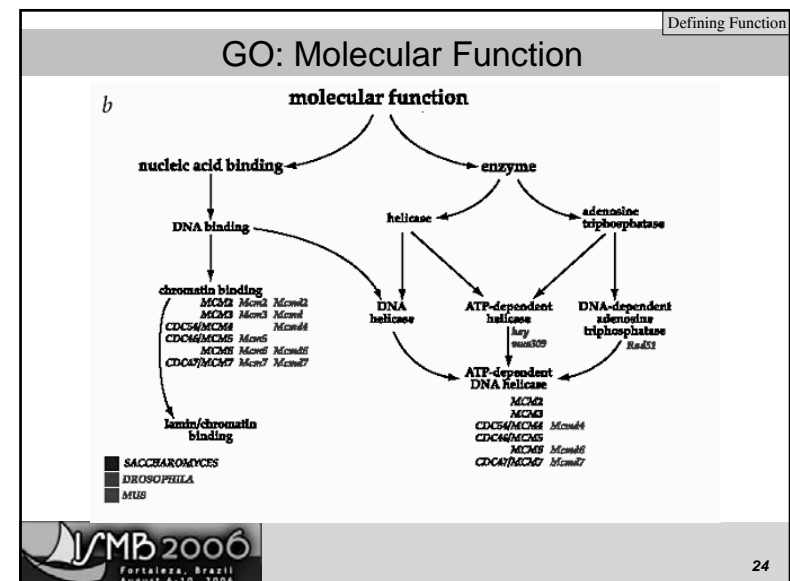
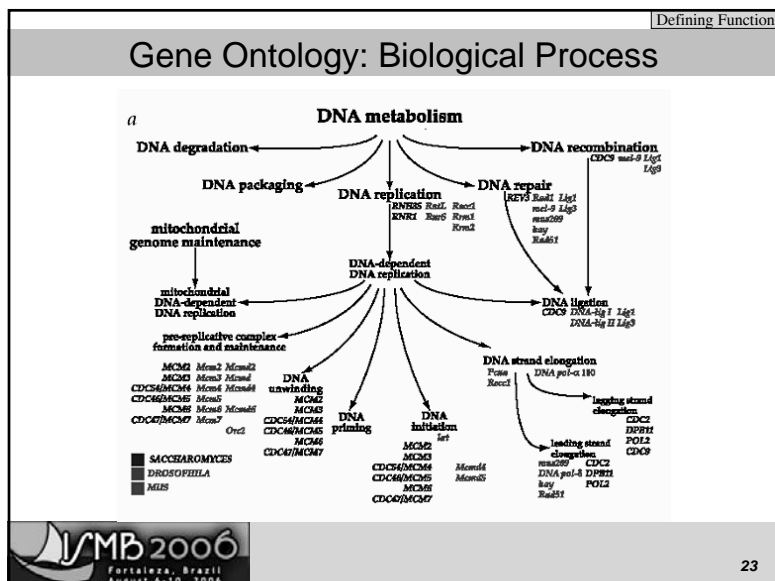
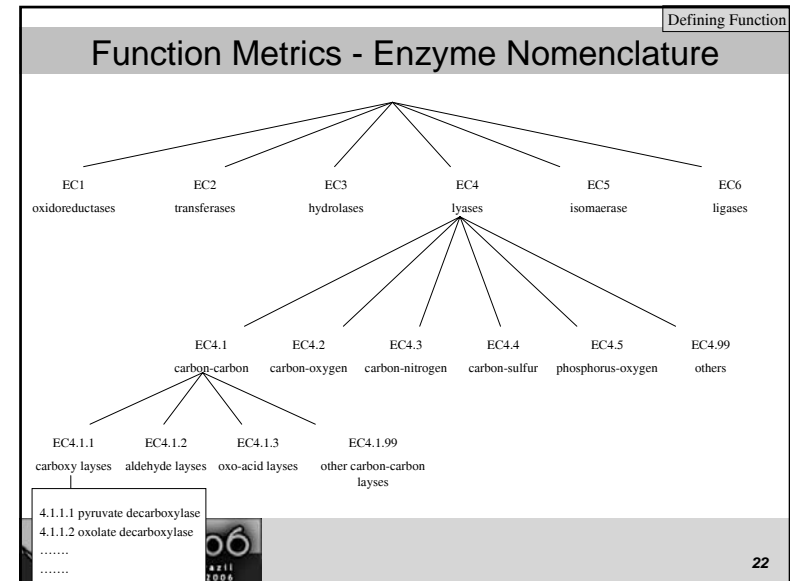
Defining Function

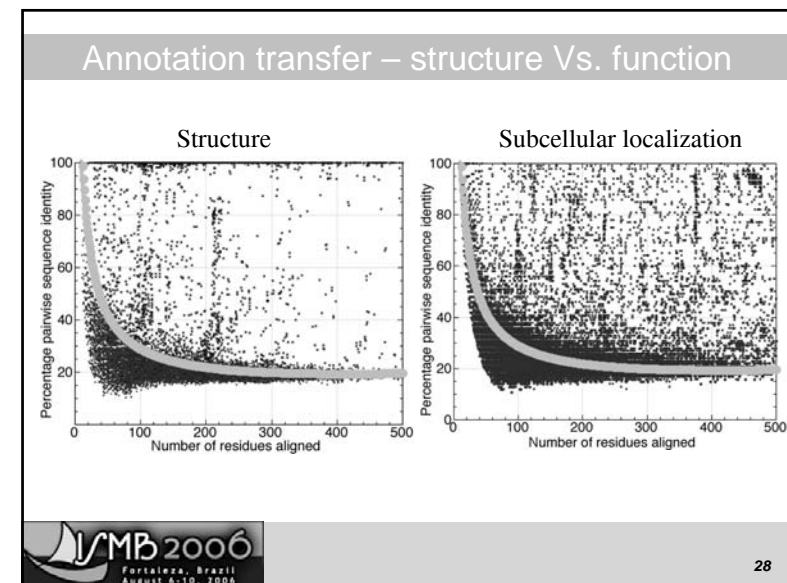
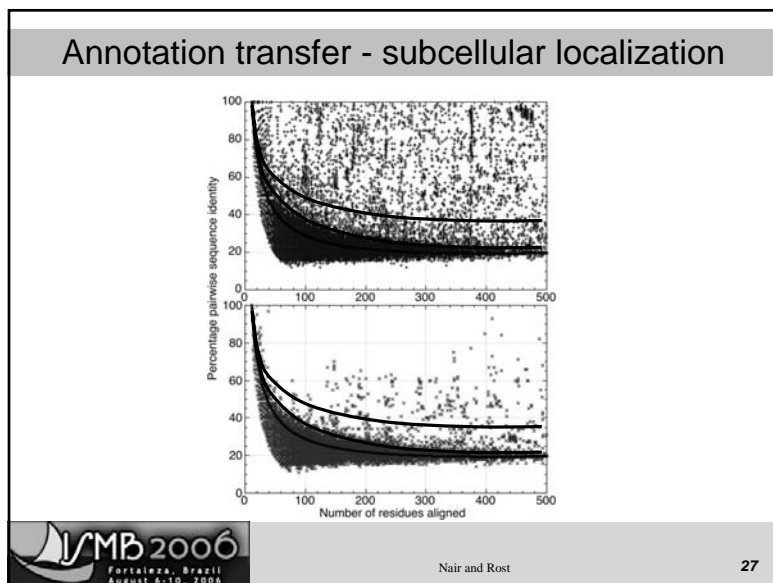
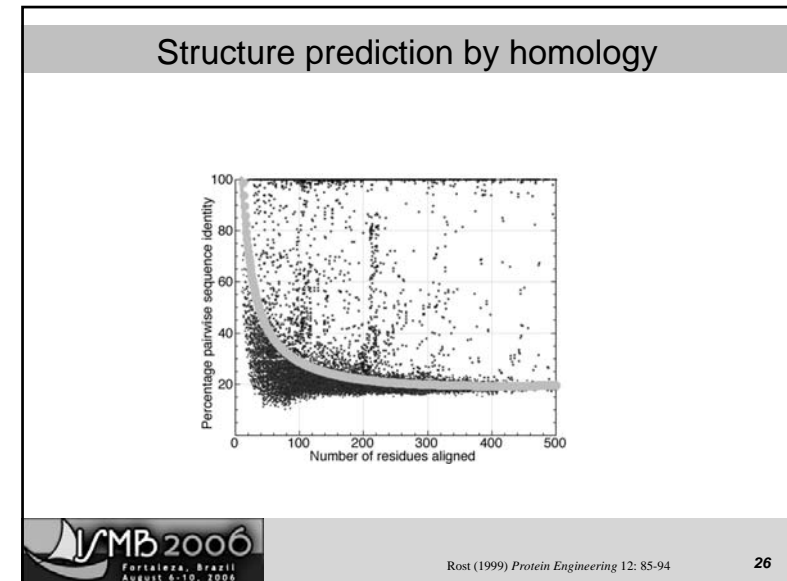
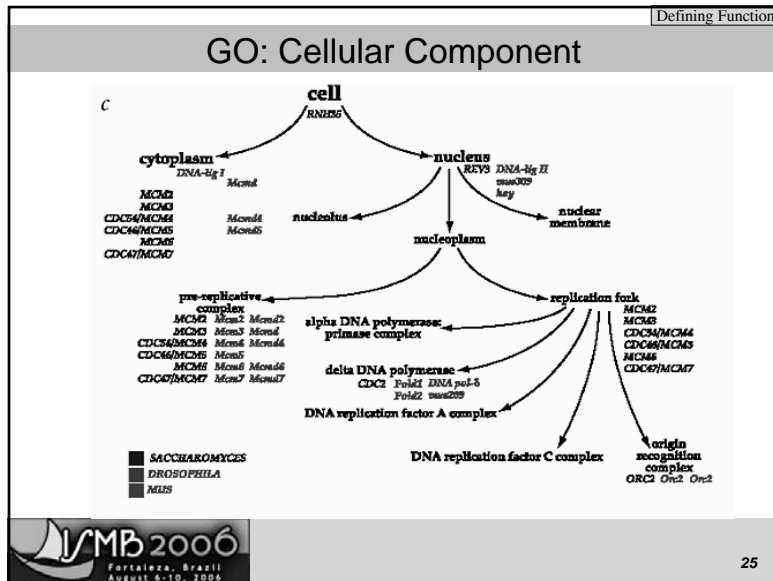
Prediction and analysis are based on metrics

- Sequence:
E-value, Sequence Identity, etc.
- Structure: RMSD
- Function?

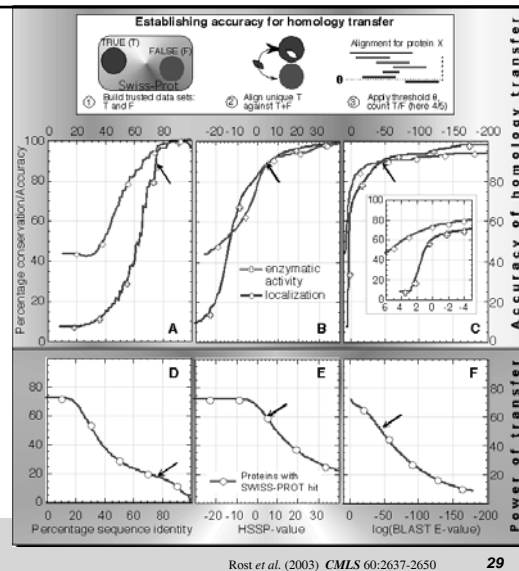

 Fortaleza, Brazil
August 6-10, 2006

21





Annotation Transfer



Rost et al. (2003) CMLS 60:2637-2650

29

Possible solution

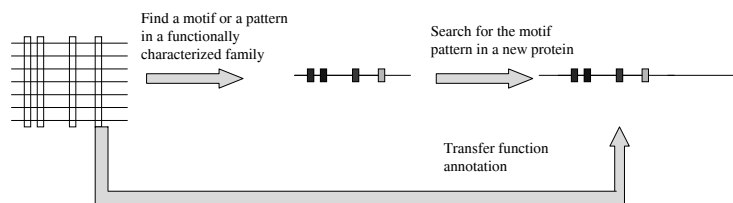
- Establish a family specific similarity cutoff
- Establish a function specific similarity cutoff.



30

Motifs and patterns

- Manual sequence alignment and annotation of patterns.
- Automatic alignment and annotation.



31

Graphical summary of hits (Java applet)

pattern CK2_PPROTEIN_SITE_178-179

58 hits with 9 PROSITE entries

Site map

Search EMBL

Contact us

Swiss-Prot

PROSITE

ProteinWiki tools

August 6-10, 2006

[Home](#)
[Databases](#)
[PubMed](#)
[WEBSTER](#)
[haemid](#)
[vcomp](#)
[13.0x.0m](#)
[Impact Factors](#)
[J. E-Journals](#)
[Citation Index](#)
[http://www.dsl...](#)

[Gefenheimer Johannes Mader](#)
[Natl Academies Press, The Role of ...](#)
[USCIS, SAM-739 query data and re...](#)
[JMI secondary structure prediction](#)
[Plan, Tr...](#)




Figure 1: 128
Signaling proteins
 Crystal structure of bovine rhodopsin at 2.5 angstroms resolution

Key: Domain Start End Residue
 [701... A 54 206
 [701... B 54 206

The DsopsignPDB mapping was provided by NSG.

For additional annotation, see the [PROSITE](#) document PDOC00210. [Trends](#) | [SRS-UK](#) | [SRS-USA](#)

Alignment

☐ Seed (84) ☐ Full (8438)

Formal:

Further alignment options [here](#)
 Help returning to these alignments [here](#)

Domain organisation

☐ Seed (84) ☐ Full (8438) ☐ Colored (1)

As a Graphic:

As a Tree: ☐ Bootstrap tree ☐ NPA Applied

View Graphic:

To find out about the NPAAS tree-viewer, click [here](#)

Species Distribution

☒ View alignments & domain organisation by species

Tree depth:

Phylogenetic tree

☒ Seed (84) ☐ Full (8438)

The trees were generated using [ClustalW](#)
 To find out more about ATV phylogenetic tree-viewer click [here](#)

Databases references

PDB

[Home](#)
[Databases](#)
[PubMed](#)
[WEBSTER](#)
[haemid](#)
[vcomp](#)
[13.0x.0m](#)
[Impact Factors](#)
[J. E-Journals](#)
[Citation Index](#)
[http://www.dsl...](#)

[Gefenheimer Johannes Mader](#)
[Natl Academies Press, The Role of ...](#)
[USCIS, SAM-739 query data and re...](#)
[JMI secondary structure prediction](#)
[Plan, Tr...](#)




Figure 1: 128
Signaling proteins
 Crystal structure of bovine rhodopsin at 2.5 angstroms resolution


Key: Domain Start End Residue
 [701... A 54 206
 [701... B 54 206

The DsopsignPDB mapping was provided by NSG.

For additional annotation, see the [PROSITE](#) document PDOC00210. [Trends](#) | [SRS-UK](#) | [SRS-USA](#)

open challenges - motifs and patterns

- Automate
- Unify
- Remote homologues



IAMB 2006
Fortaleza, Brazil
August 6-10, 2006

34

Structural patterns

- Manual identification of active site
- Automatic structural alignment?

The diagram illustrates a workflow for structural pattern identification and transfer function annotation. It begins with a large, dark, T-shaped structure representing a protein. An arrow points from this structure to a smaller, stylized star-like shape, labeled "Identify active site / functional element". From this star-like shape, another arrow points to a grey oval containing a star, labeled "Search for this structural pattern in a new protein". A long arrow points from the bottom of the first structure to the bottom of the oval, labeled "Transfer function annotation".

Identify active site / functional element

Search for this structural pattern in a new protein

Transfer function annotation


IUMB 2006
Fortaleza, Brazil
August 8-10, 2006

For review see: Jones & Thornton (2004) *Curr Opin Struc Biol* 8:3-7

35

open challenges - structural patterns

- Find
- Search
- Add biophysics of the site to the spatial search



ICMB 2006
Fortaleza, Brazil
August 14-19, 2006

36

Function by Association

Predict protein-protein interactions: fusion

'Fused' and 'separated'

Genome A: Protein 1, Protein 2
Genome B: Protein 3

For example see: Marcotte *et al.* (1999) *Science* 285:751-753

37

Function by Association

Predict protein-protein interaction: Rosetta

	sc	bs	ih	dm	ce	hs	hp
P1	1	0	1	0	0	1	0
P2	1	1	0	0	0	0	1
P3	0	0	0	1	1	1	0
P4	1	0	1	0	0	1	0
P5	0	1	1	0	1	0	1
P6	0	0	0	1	1	1	0
p7	1	1	0	0	0	0	1

P1:1010010
P4:1010010

P2:1100001
P7:1100001

P3:0001110
P6:0001110

For example see: Pellegrini *et al.* (1999) *PNAS* 96:4285-4288

38

Function by Association

Predict protein-protein interaction: Motifs

Sequence signatures

Known to interact

Proteins 1, 3, 5 (left) and Proteins 2, 4, 6 (right)

Legend: \triangle , \square , \circ , \diamond

\triangle	0	1	2	1	0
\square		2	0	2	1
\circ			0	0	0
\diamond				0	1
					1

For example see: Sprinzak & Margalit (2001) *J.Mol. Biol.* 271:511-523

39

Function by Association

Predict protein-protein interaction: correlated mutations

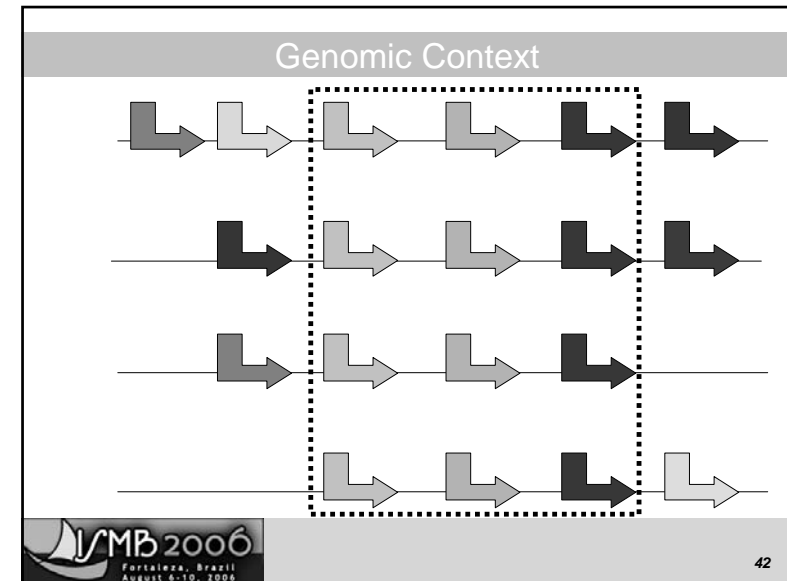
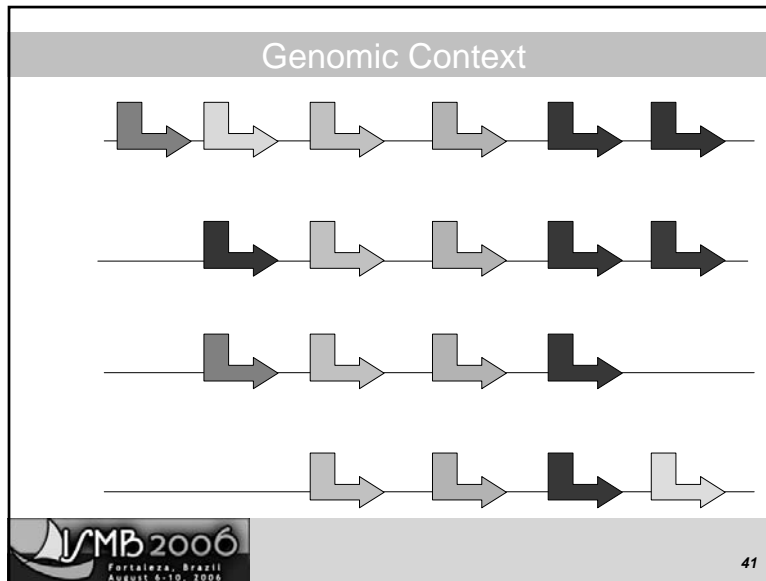
Correlated mutations

Alignment of protein 1

Alignment of protein 2

For example see: Pazos & Valencia (2002) *Proteins*. 47:219-227

40



Function by Association

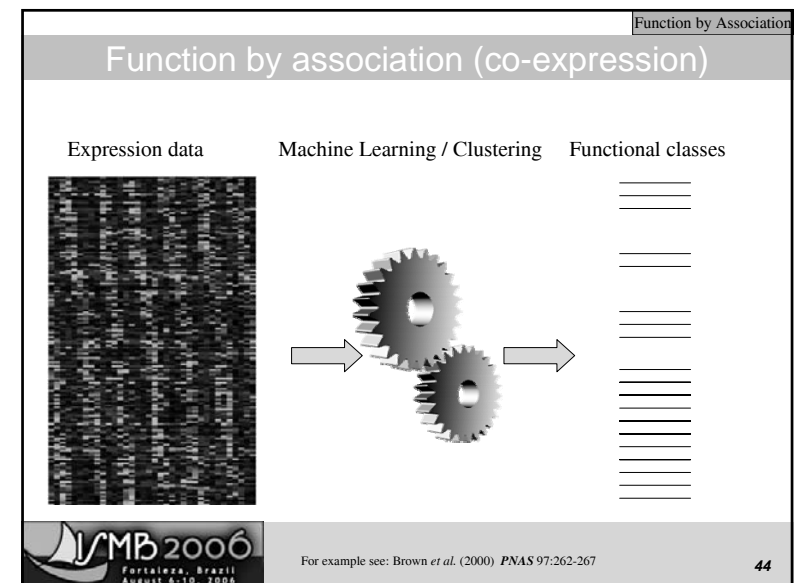
Function by association: open challenges

- Differentiate functional and physical interaction
- Improve accuracy and coverage (data, algorithm)
- Ab-initio prediction

IMB 2006
Fortaleza, Brazil
August 6-10, 2006

Bairoch A. (2000) *Nucleic Acid Research*. 28:304-305

43



How useful are these methods?

On *Mycoplasma genitalium*

Method	% of genes to which criterion is applicable
Gene fusion	6%
Conservation of the local context	45% (37% + 8%)
Genomic profiles	11%
Combined	~50%

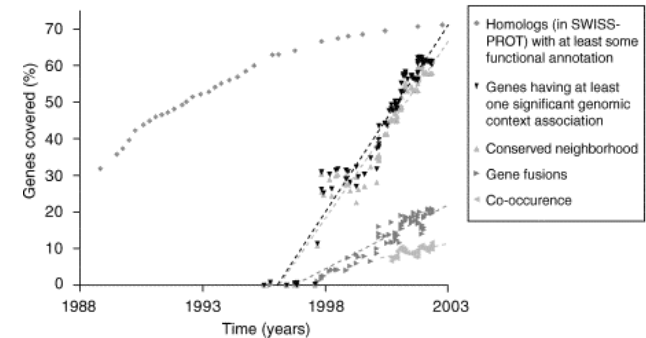


Huynen et al. (2000) *Genome Res* 10: 1204-1210

45

How useful are these methods?

Escherichia coli K12, coverage of functional annotations



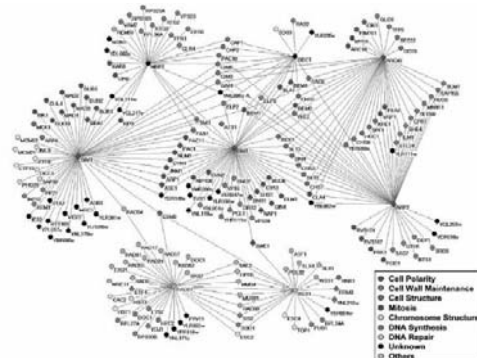
Current Opinion in Cell Biology



Huynen et al. (2003) *Curr Op Cell Biol* 15: 191-198

46

Function by association (interactions / network)



For example see: Tong et al. (2002) *Science* 295:321-324

47

Function by association: open challenges

- Differentiate functional and physical interaction
- Improve accuracy and coverage (data, algorithm)
- Ab-initio prediction



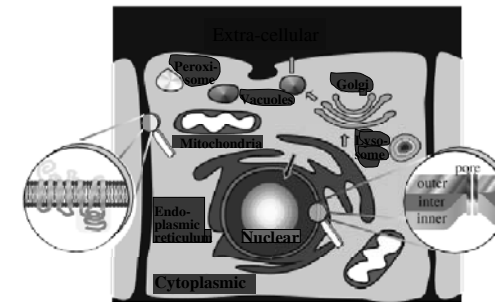
3304-305

48

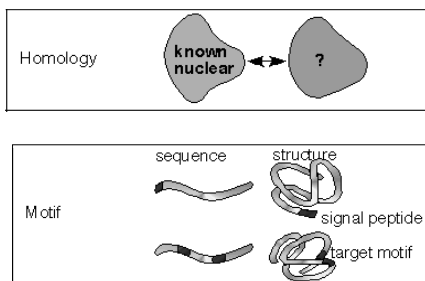
Predict aspects of function

- Sub cellular localization (nucleus, membrane, etc.)
- Post translational modifications
- Functionally important residues
- Interaction sites

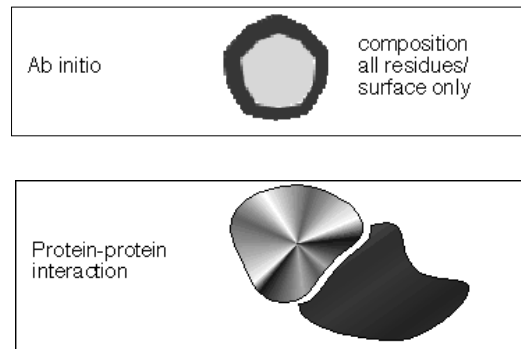
Sub-cellular localization



Prediction of sub cellular localization



Prediction of sub cellular localization



Aspects of Function

Post translational modifications

- N-terminal signal peptide cleavage
- Proteolytic cleavage, proteosome cleavage
- Phosphorylation
- Lipid modification
- N- and O-glycosylations

IUMB 2006
Fortaleza, Brazil
August 6-10, 2006

For review see: Rost *et al.* (2003) *CMLS* 60:2637-2650

53

Aspects of Function

Functionally important residues

Variable Conserved

IUMB 2006
Fortaleza, Brazil
August 6-10, 2006

Glaser *et al.* (2003) *Bioinformatics* 19:163-164

54

Aspects of Function

Functionally important residues - interactions sites

..QIKILGNA.
..--PP--PP.

IUMB 2006
Fortaleza, Brazil
August 6-10, 2006

Ofran & Rost (2003) *FEBS Letters* 544:236-239

55

Aspects of Function

Functionally important residues - interactions sites

..LNDRA.
..---P-

IUMB 2006
Fortaleza, Brazil
August 6-10, 2006

Ofran & Rost (2004) *submitted*

56

Aspects of Function

Functionally important residues - open challenges

- DNA binding
- Antigenic sites
- Metal binding
- Ion binding
- Improve coverage and accuracy

IMB 2006
Fortaleza, Brazil
August 6-10, 2006

57

Functional Type

Functional Type - Data Mining

MIP Class

IMB 2006
Fortaleza, Brazil
August 6-10, 2006

Clare & King (2003) *Bioinformatics* 19:ii42-ii49

58

Functional Type

From Structure: ProKnow

Input **Extract Features** **Associate GO Terms**

Structure
Sequence
ADRTYFGH
NWDERFGH
TYMKLPRS

Feature Extractors
a DALI / DASEY
b RIGOR
c PSI-BLAST
d PROSITE
e DIP

Protein Features
a Fold
b 3D-Motifs
c Sequence
d Motifs
e Functional-linkages

Clues
Each feature may give multiple clues

ProKnow Knowledgebase

Controlled Vocabulary
Functions mapped to protein features by the Annotation Profile

Bayes Theorem to weight function

Weighted Set of Functions

IMB 2006
Fortaleza, Brazil
August 6-10, 2006

Pal & Eisenberg (2005) *Structure* 13(1):121-30

59

Functional Type

Functional Type - ProtFun: Ab initio from sequence

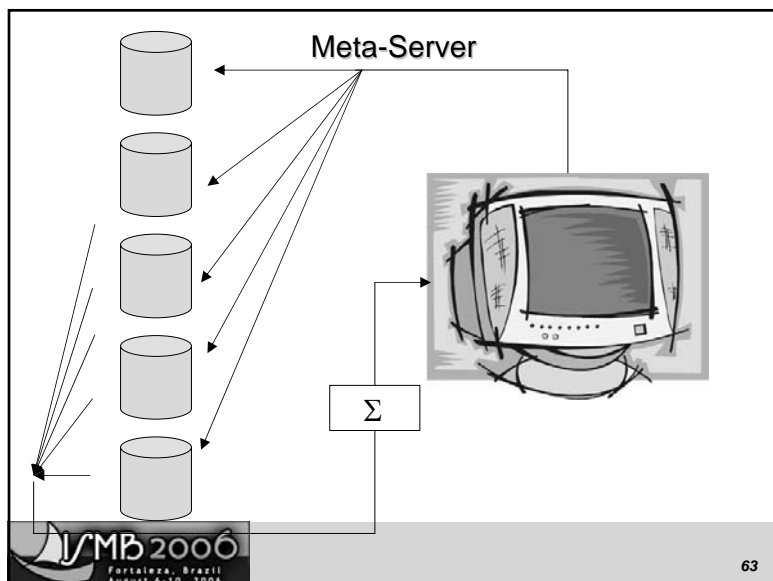
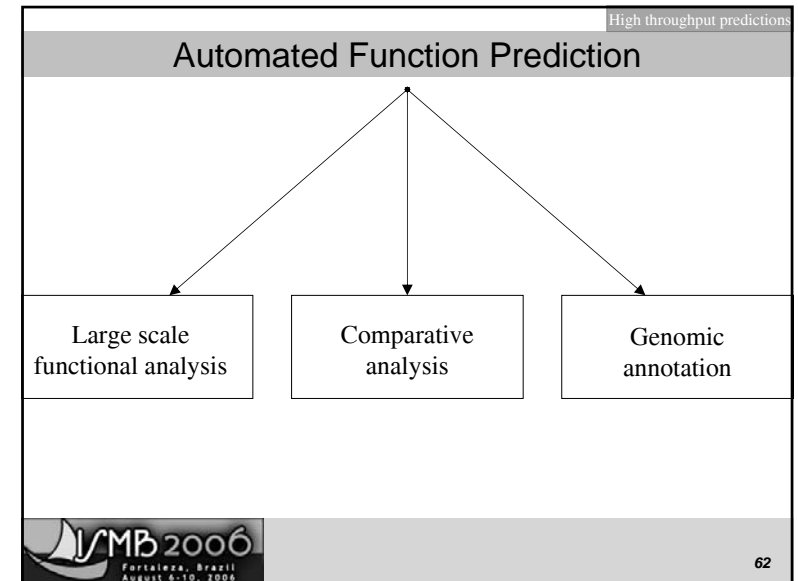
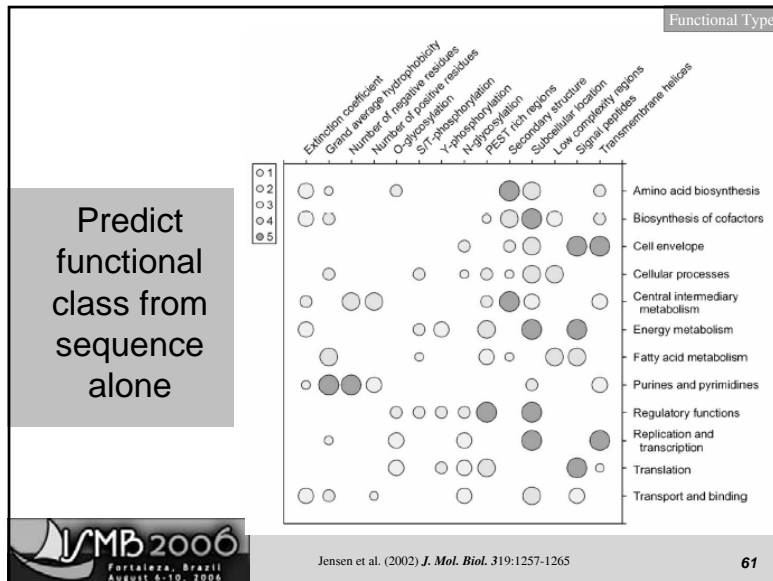
KKVVLGKKGDTVELTCTASQKKSIQFHWKNSNQIKILGNQG

GO class

IMB 2006
Fortaleza, Brazil
August 6-10, 2006

Jensen et al. (2002) *J. Mol. Biol.* 319:1257-1265

60



- Conclusion
- ## Open Challenges
- ☒ Data
 - ☒ Algorithms
 - ☐ Defining the problem
 - ☐ Choice of relevant features
 - ☐ Preparation of data
 - ☐ Assessment of performance
- 64

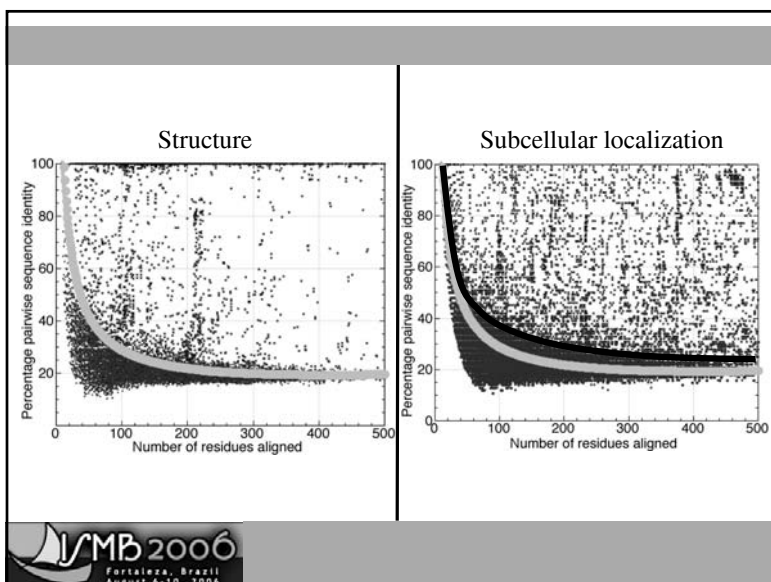
Computing Biological Function: Bioinformatics approach to the analysis and prediction of protein function

(Part 2)

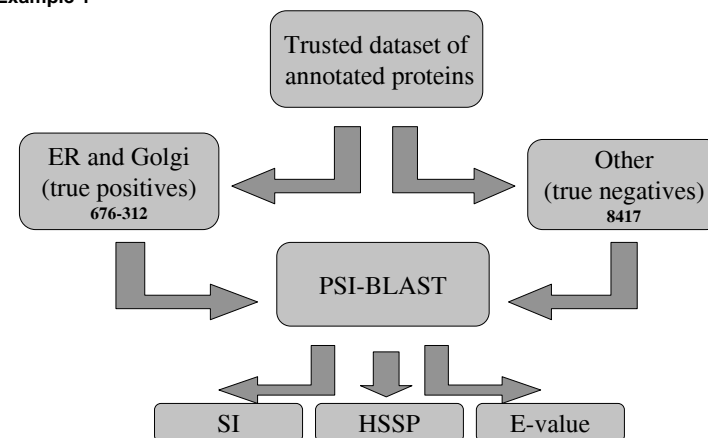
Yanay Ofran & Marco Punta
Columbia University, New York

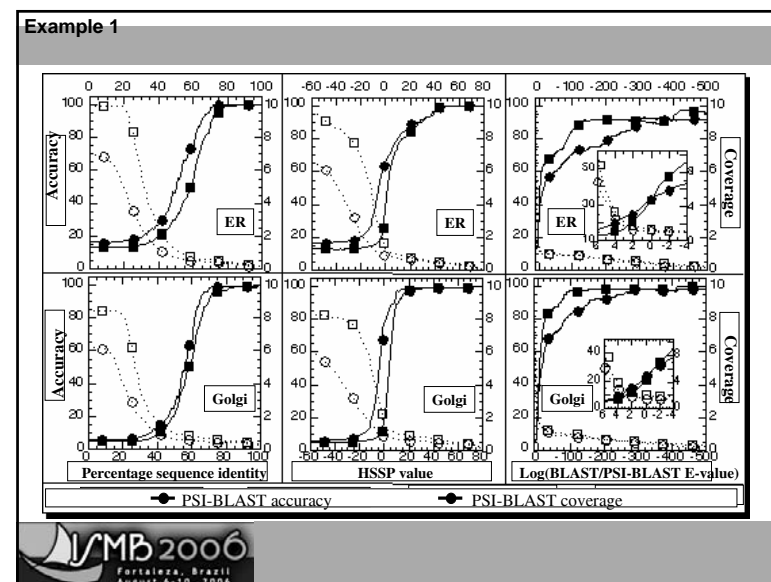
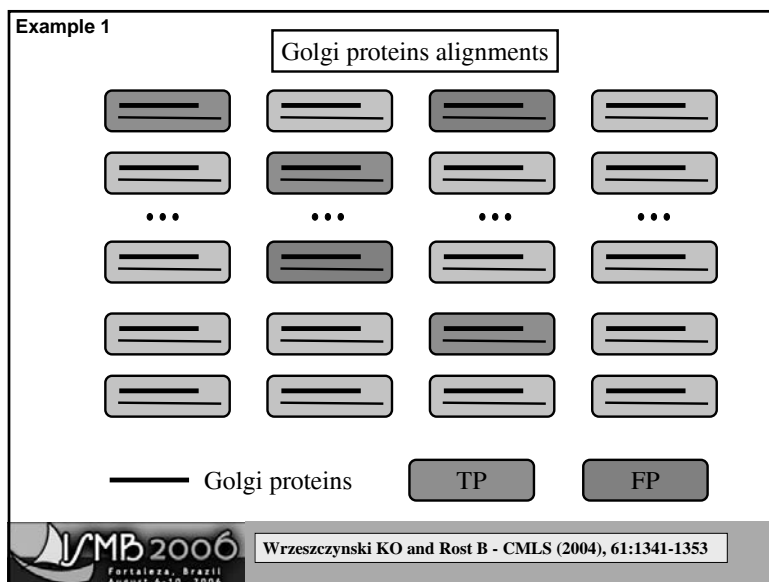
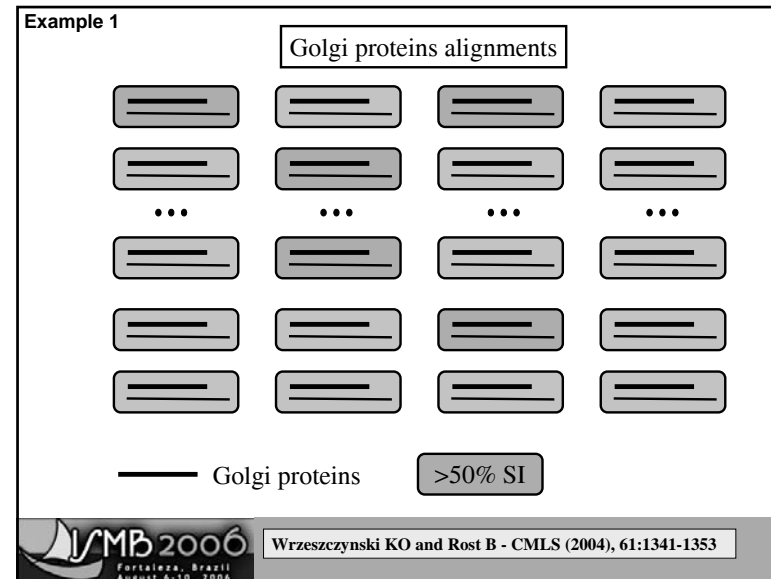
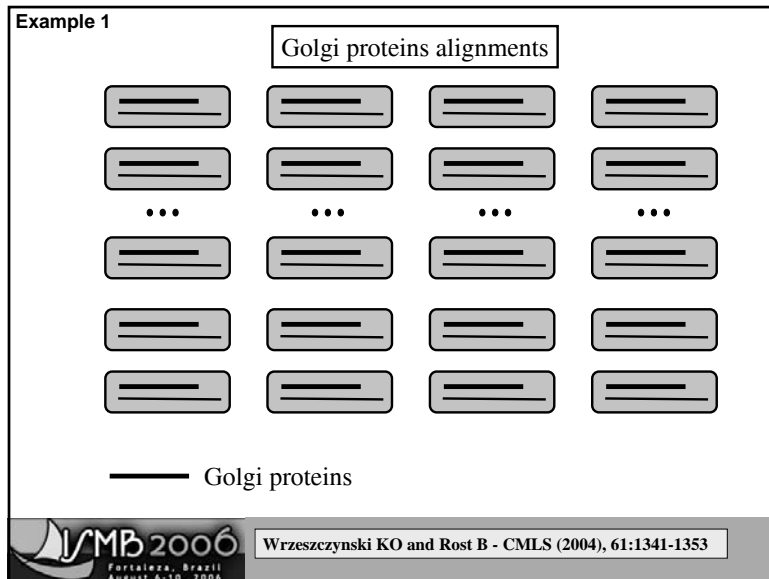
I/MB 2006 Fortaleza, Brazil August 6-10, 2006

Using Homology Transfer

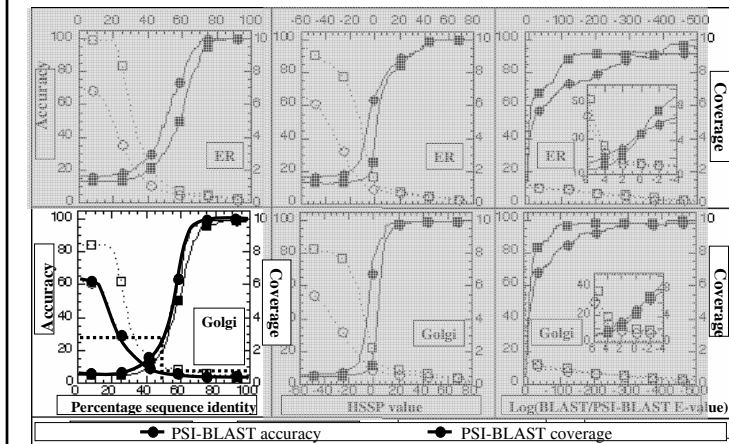


Example 1

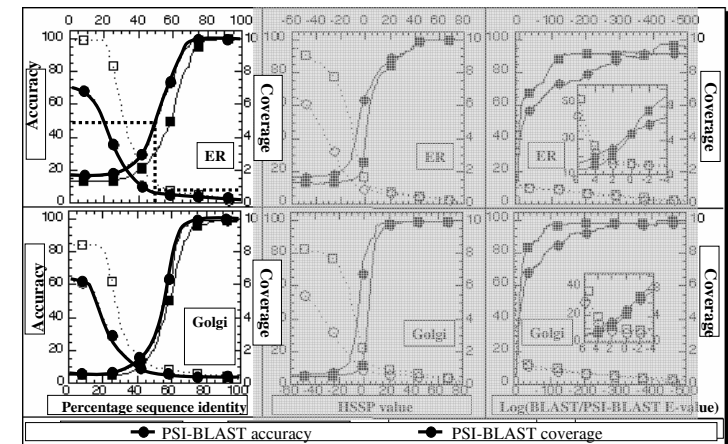




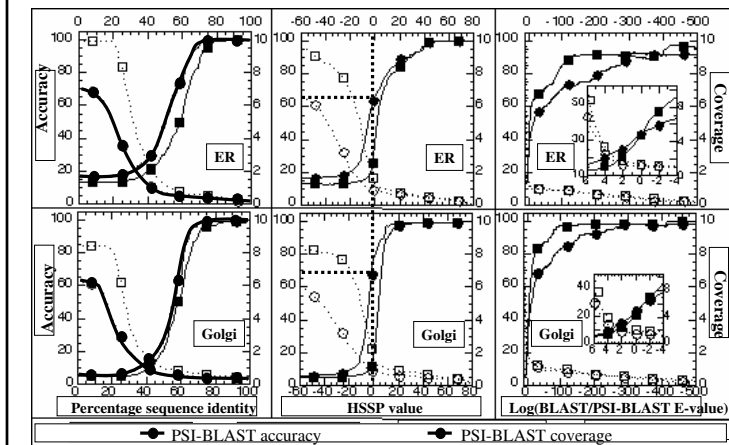
Example 1



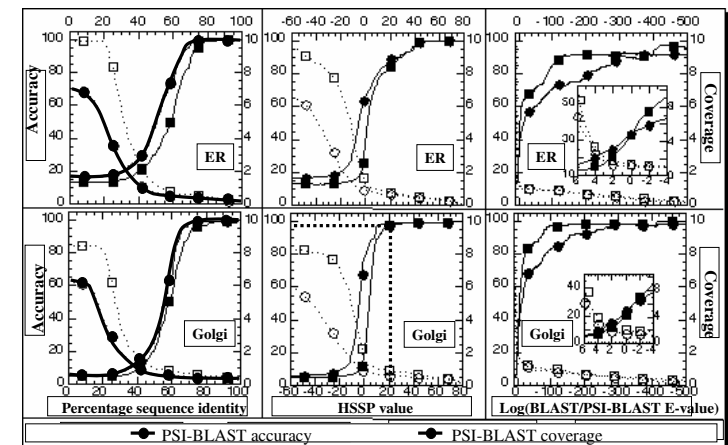
Example 1



Example 1



Example 1



Example 1

HSSP > 23 → 98% estimated accuracy

Proteome	Predicted	Annotated as Golgi in Swiss-Prot	Different Swiss-Prot annotation	Hypothetical protein	estimated # of errors
<i>S.cerevisiae</i>	70	53	17	8	1-2
<i>A.thaliana</i>	70	31	39	5	1-2
<i>C.elegans</i>	57	23	34	27	1
<i>D.melanogaster</i>	61	40	32	0	1
<i>M.musculus</i>	195	145	50	0	4
<i>H.sapiens</i>	347	273	74	0	7
All 6	800	565	235	40	16



Example 1

HSSP > 23 → 98% estimated accuracy

Proteome	Predicted	Annotated as Golgi in Swiss-Prot	Different Swiss-Prot annotation	Hypothetical protein	estimated # of errors
<i>S.cerevisiae</i>	70	53	17	8	1-2



Example 1

HSSP > 23 → 98% estimated accuracy

Proteome	Predicted	Annotated as Golgi in Swiss-Prot	Different Swiss-Prot annotation	Hypothetical protein	estimated # of errors
<i>S.cerevisiae</i>	70	53	17	8	1-2
<i>A.thaliana</i>	70	31	39	5	1-2
<i>C.elegans</i>	57	23	34	27	1
<i>D.melanogaster</i>	61	40	32	0	1
<i>M.musculus</i>	195	145	50	0	4
<i>H.sapiens</i>	347	273	74	0	7
All 6	800	565	235	40	16



Example 1

Proteome	Predicted	Annotated as Golgi in Swiss-Prot	Different Swiss-Prot annotation	Hypothetical protein	estimated # of errors
23 (98%)	800	565	235	40	16
16 (95%)	1110	675	435	66	55
12 (90%)	1358	728	630	99	136
8 (85%)	1726	812	914	125	259
7 (78%)	1853	826	1027	134	407



Beyond homology transfer

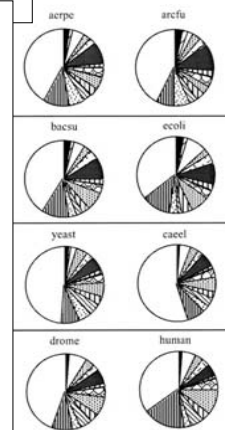


Liu J and Rost B - Protein Science (2001), 10:1970-1979

Example 2

Functional classification of genomes

Homology Transfer?



QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.



Homology transfer with > 30% SI, ~70% accuracy for EUCLID classification
Tamames...Valencia Bioinformatics 14 (1998)

Example 2

% of different protein types in genomes

Eukaryotes

Bacteria

Archaea

% mem % coils % sigp

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

0 5 10 15 20 25 30 35 2 4 6 8 10 12 5 10 15 20 25 30

Percentage of proteins in entire proteome



Liu J and Rost B - Protein Science (2001), 10:1970-1979
Methods used: PHDhtm, COILS, SignalP

Example 2

% of different protein types in genomes

Eukaryotes

Bacteria

Archaea

% mem % coils % sigp

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

0 5 10 15 20 25 30 35 2 4 6 8 10 12 5 10 15 20 25 30

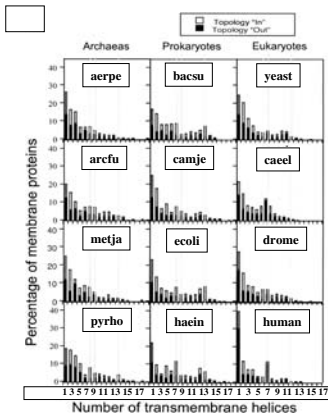
Percentage of proteins in entire proteome



Rost, B., Casadio, R. & Fariselli, P. (1996). Prot. Sci., 5, 1704-1718.

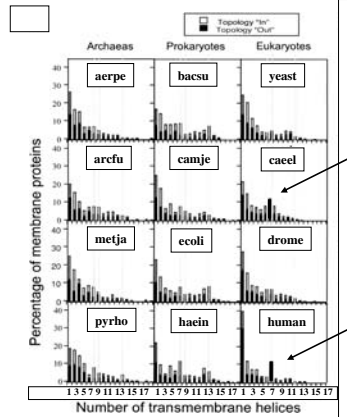
Example 2

% of membrane proteins in genomes



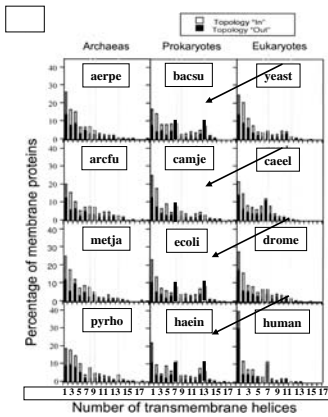
Example 2

% of membrane proteins in genomes



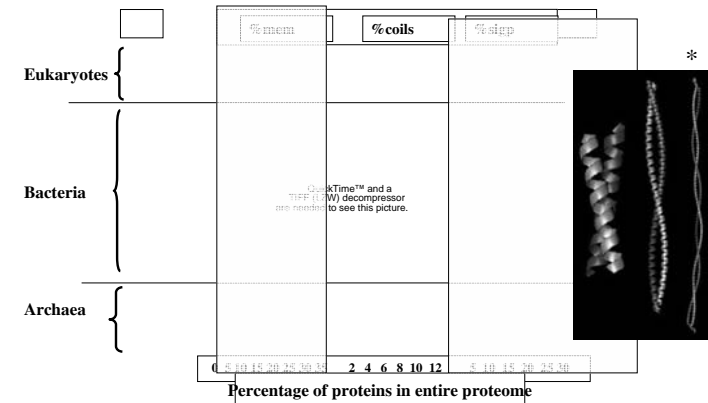
Example 2

% of membrane proteins in genomes



Example 2

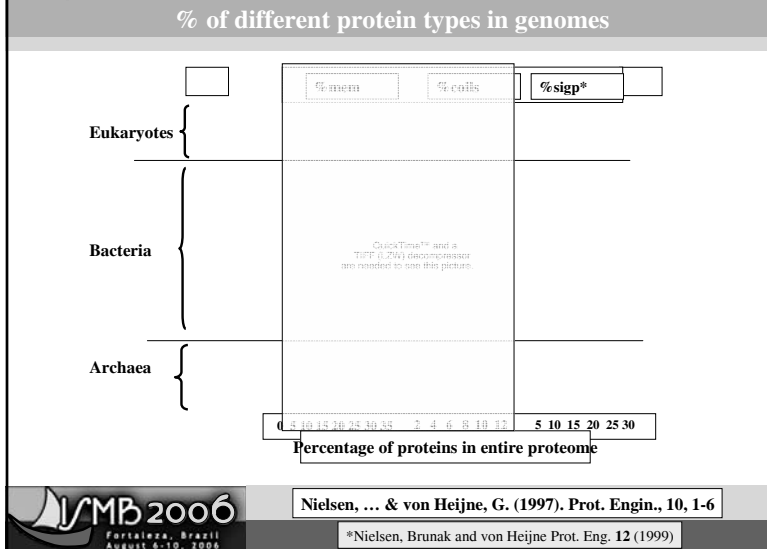
% of different protein types in genomes



Lupas, A. (1996). Meth. Enzymol., 266, 513-525

*From http://membranes.nbi.dk/article_coiled-coil/graphics/coiled-coil_proteins1.jpeg

Example 2



Using interaction maps

Experimental methods for detecting protein-protein interactions

Two-hybrid systems

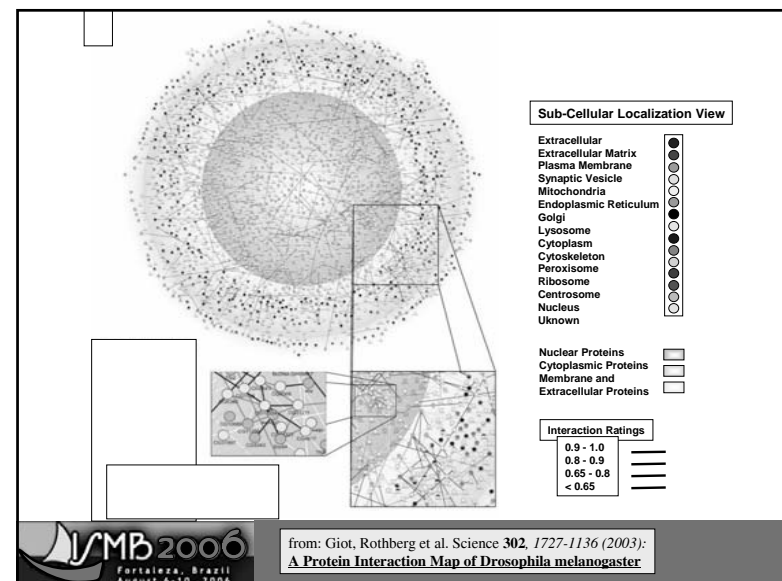
Fields *FEBS* (2005) **262**:5391-5399 - Review

Mass Spectrometry

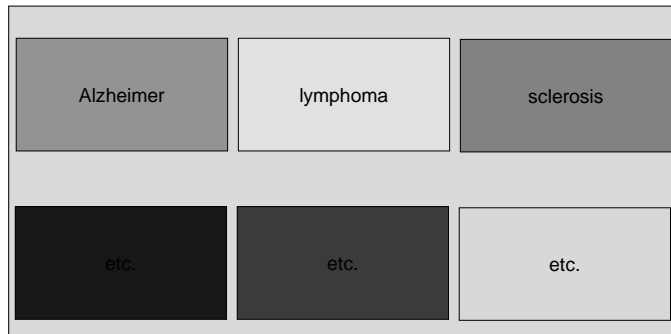
Mann et al. (2001) *Rev Biochem.* **70**:437-473 -Review

Microarrays

ESPEJO et al. (2002) *Biochem. J.* **367** (697-702)



<http://www.ncbi.nlm.nih.gov/omim/>

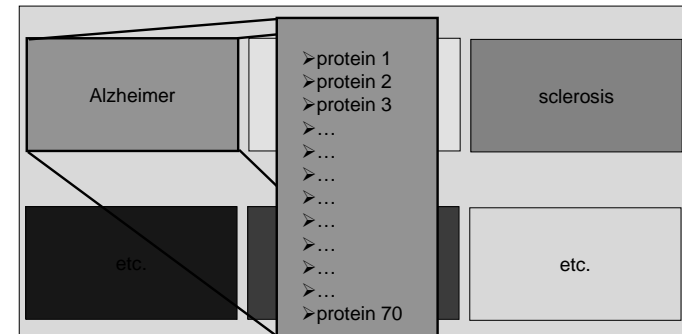


Chen et al. Pacific Symposium on Biocomputing 11:367-378(2006)

Example 3

Online Mendelian Inheritance in Man (OMIM)

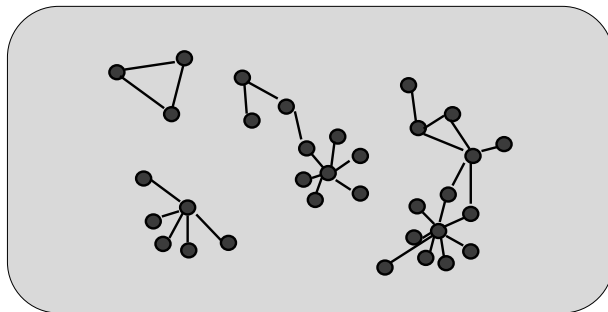
<http://www.ncbi.nlm.nih.gov/omim/>



Example 3

Online predicated human interaction database (OPHID)

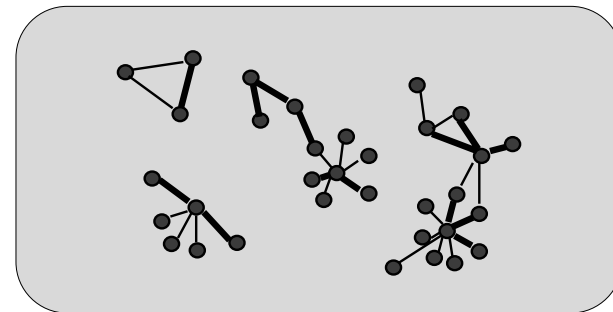
<http://ophid.utoronto.ca/ophid/>



Example 3

Online predicated human interaction database (OPHID)

<http://ophid.utoronto.ca/ophid/>

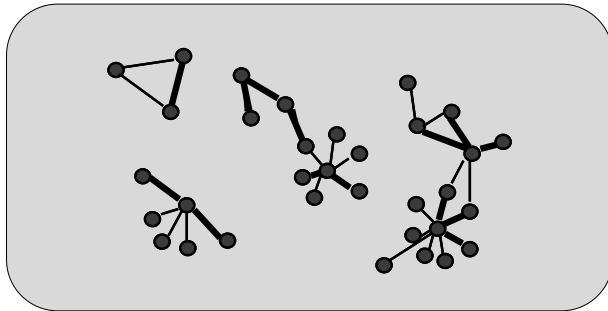


Example 3

Online predicated human interaction database (OPHID)
<http://ophid.utoronto.ca/ophid/>

Alzheimer

- >protein 1
- >protein 2
- >protein 3
- >...
- >...
- >...
- >...
- >...
- >protein 70

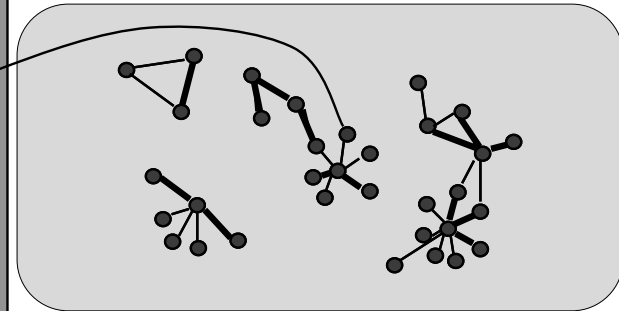


Example 3

Online predicated human interaction database (OPHID)
<http://ophid.utoronto.ca/ophid/>

Alzheimer

- >protein 1
- >protein 2
- >protein 3
- >...
- >...
- >...
- >...
- >...
- >protein 70

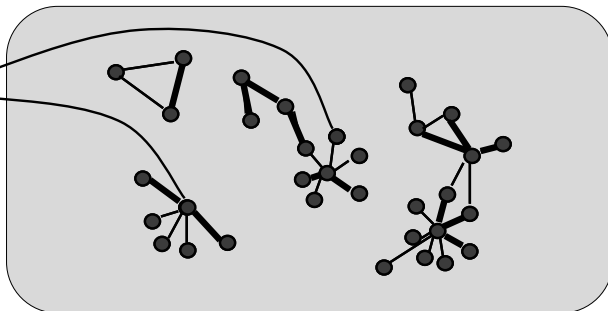


Example 3

Online predicated human interaction database (OPHID)
<http://ophid.utoronto.ca/ophid/>

Alzheimer

- >protein 1
- >protein 2
- >protein 3
- >...
- >...
- >...
- >...
- >...
- >protein 70

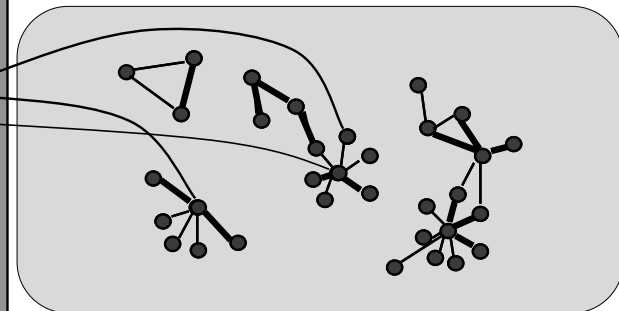


Example 3

Online predicated human interaction database (OPHID)
<http://ophid.utoronto.ca/ophid/>

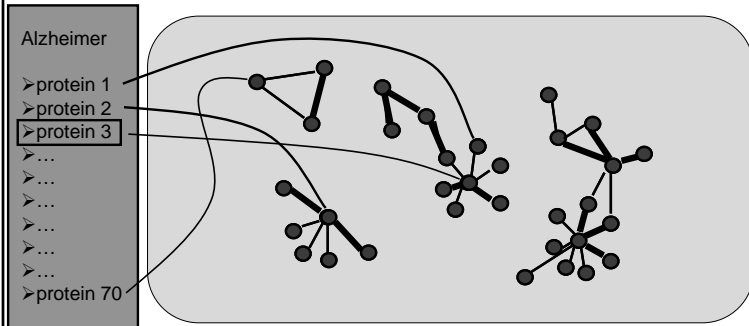
Alzheimer

- >protein 1
- >protein 2
- >protein 3
- >...
- >...
- >...
- >...
- >...
- >protein 70



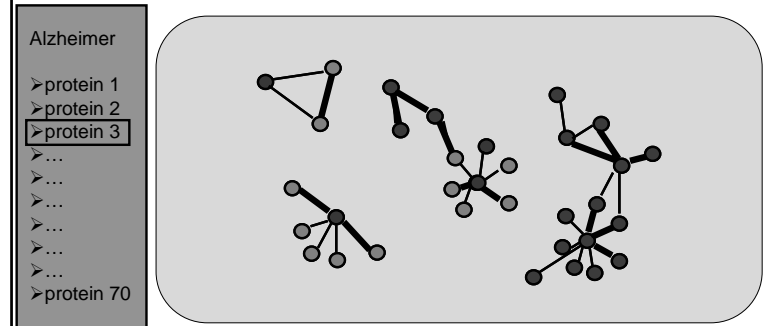
Example 3

Online predicated human interaction database (OPHID)
<http://ophid.utoronto.ca/ophid/>



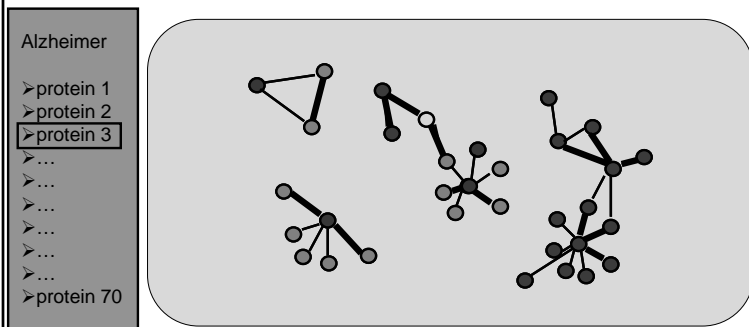
Example 3

Online predicated human interaction database (OPHID)
<http://ophid.utoronto.ca/ophid/>

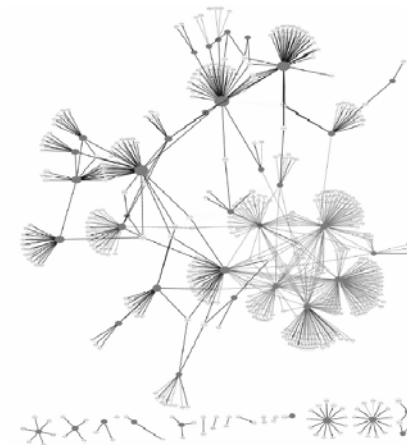


Example 3

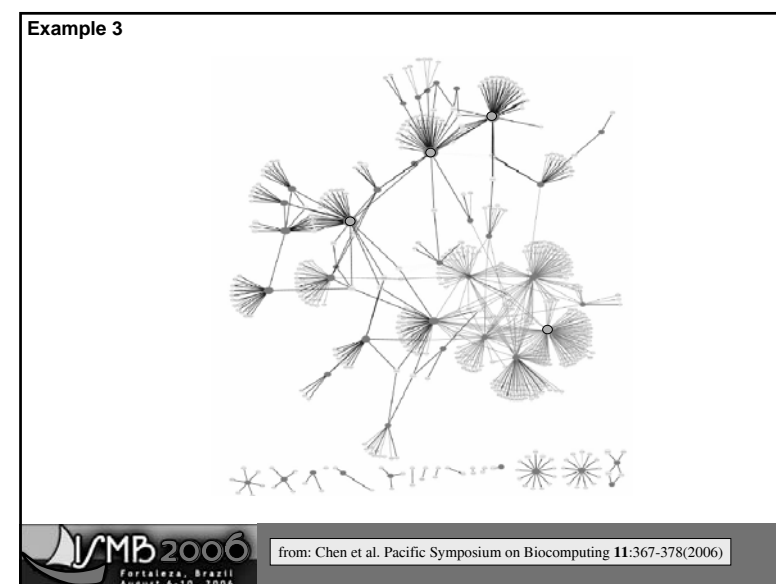
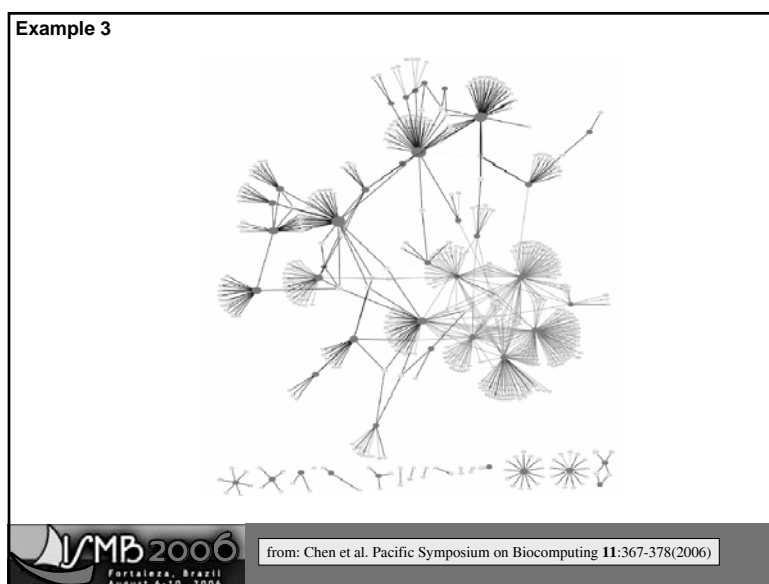
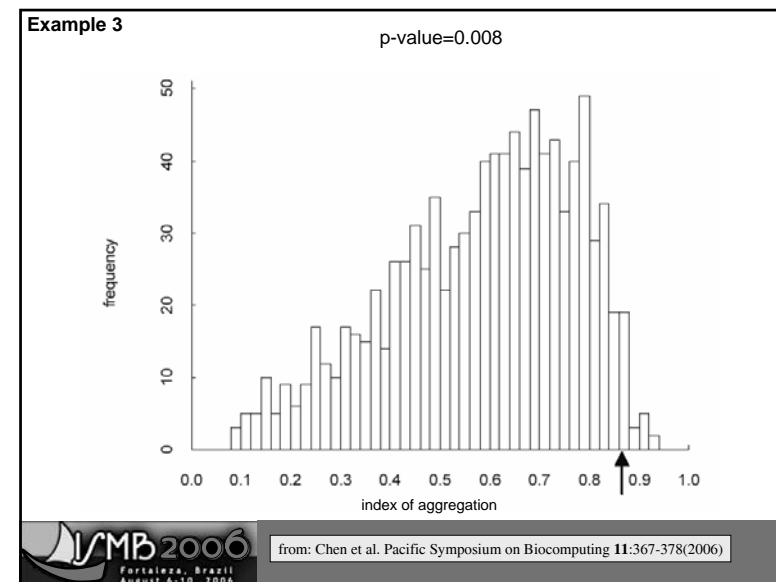
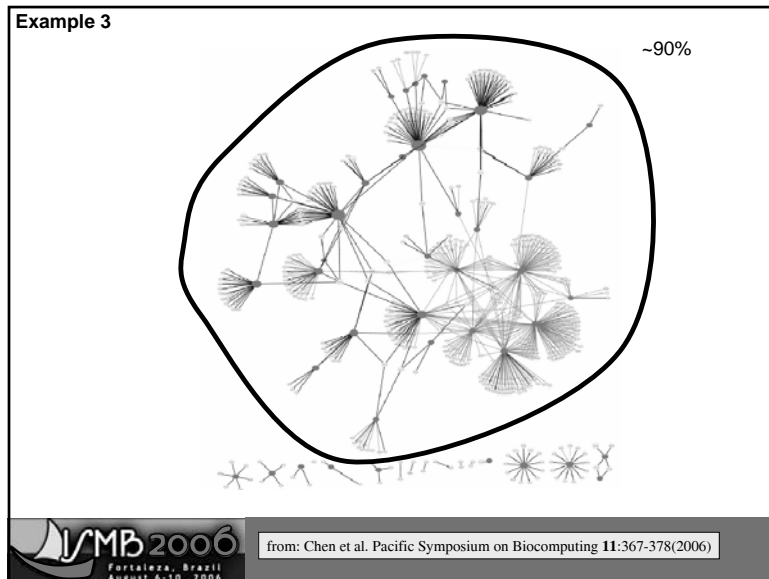
Online predicated human interaction database (OPHID)
<http://ophid.utoronto.ca/ophid/>



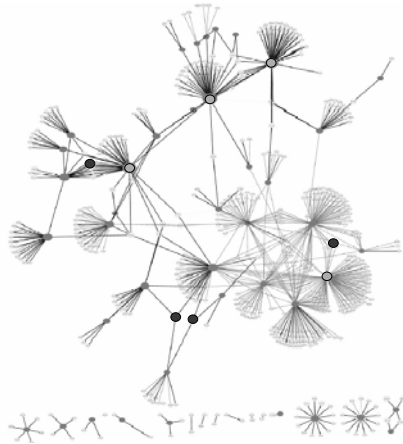
Example 3



from: Chen et al. Pacific Symposium on Biocomputing 11:367-378(2006)



Example 3



from: Chen et al. Pacific Symposium on Biocomputing 11:367-378(2006)

Using Structure



What information from structure?

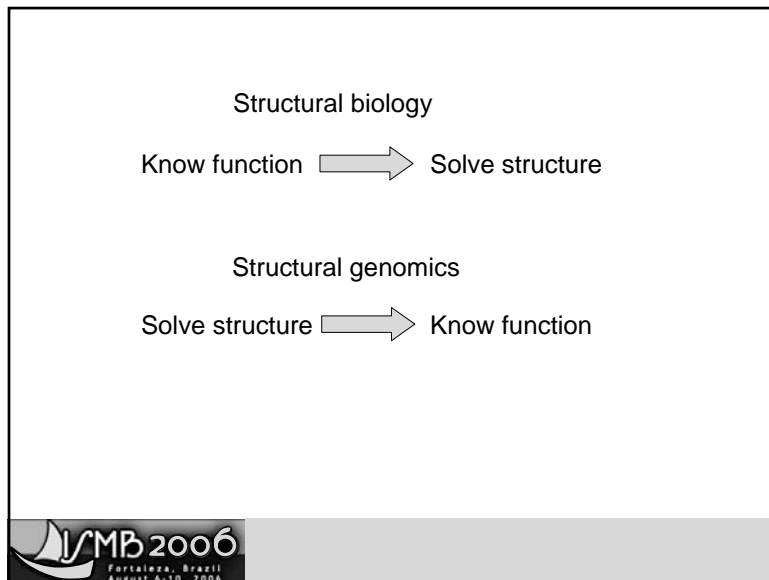
If function is known...it can help us understanding
the underlying molecular mechanisms
(e.g. enzymatic reactions)



What information from structure?


If function is not known...we can use structural
similarity with proteins of known function
(if any) to annotate the protein





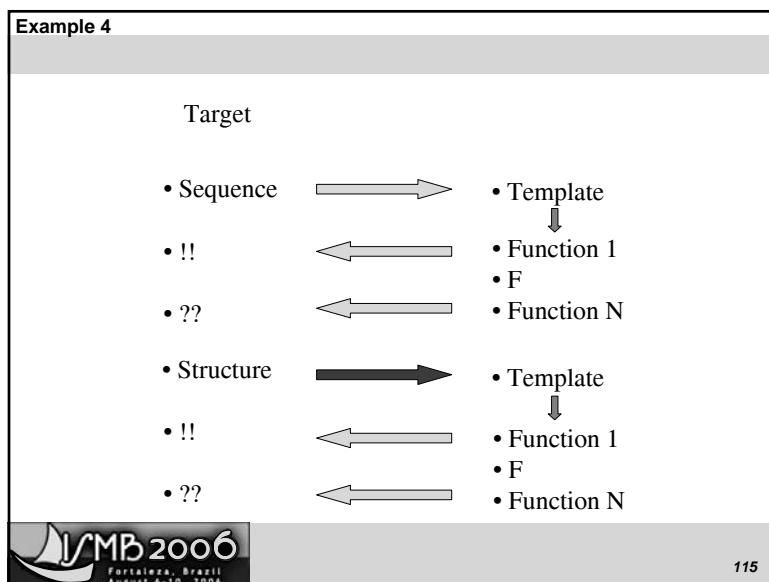
Example 4

Structure better than sequence



[ref] Keller JP, Smith PM, Benach J, Christendat D, deTitta GT, and Hunt JF *Structure* 2002, 10:1475-87

114




Example 4

Target sequence

The MT0146/CbiT sequence

```

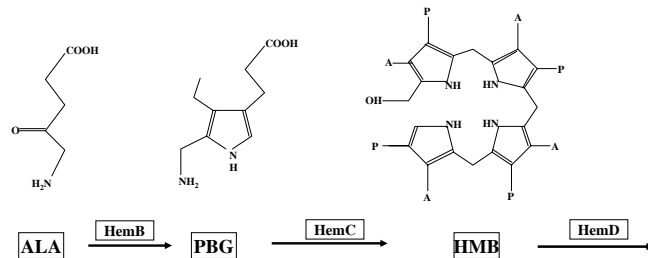
1 MIPDDEFIKNPSVPGPTAMEVRCLIMCLAEPGKNDVAVDVGCCTGGVTLELAGRVRVYA
IDRNPEAISTTEMNLQRHGLGDNVTLMEGDAPEALCKIPDIDIAVVGSGGELQEILRII
KDKLKPGGRIIVTAILLETKEAMECLRDLGFDVNI TELNIARGRALDRGTMVSRNPVA
LIYTGVSHEKND 192
  
```



116

Example 4

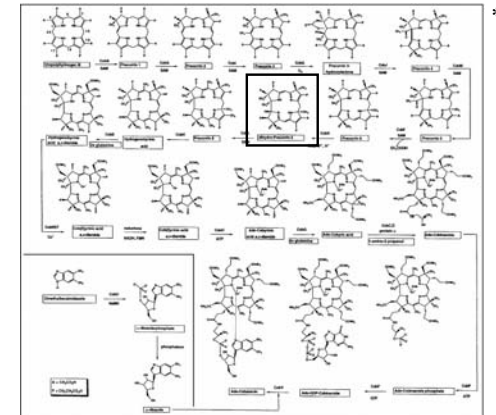
Biosynthesis of cobalamin (vitamin B₁₂)



117

Example 4

Biosynthesis of cobalamin (vitamin B₁₂)

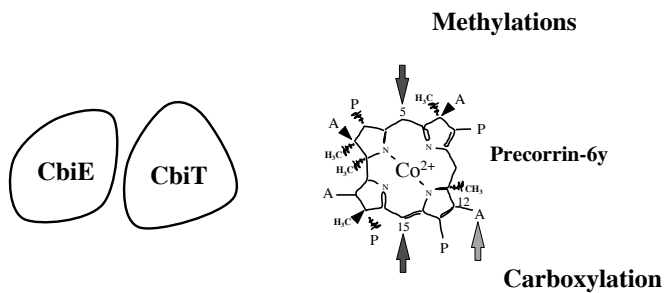


From Scott AI, and Roessner CA *Biochem Soc Trans.* 2002, 30:613-620

118

Example 4

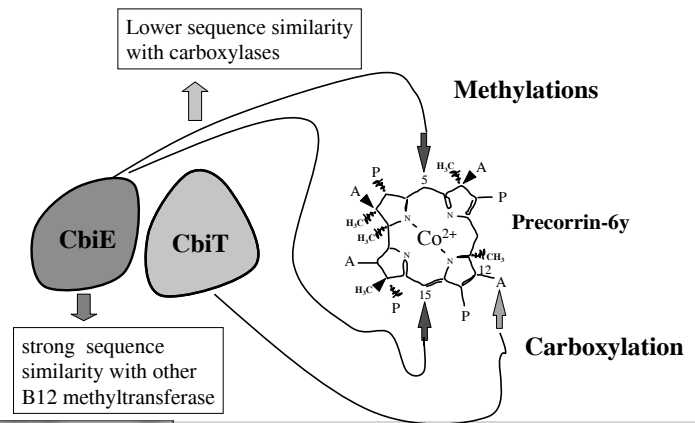
Known functional facts



119

Example 4

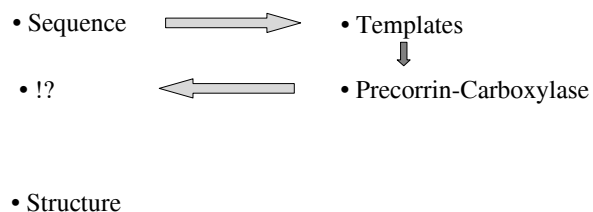
Functional facts and hypothesis



120

Example 4

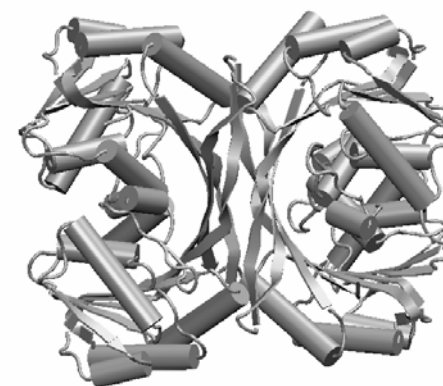
Target



121

Example 4

CbiT structure



Keller JP, Smith PM, Benach J, Christendat D, deTitta GT, and Hunt JF *Structure* 2002, 10:1475-87

122

Example 4

Some of the available programs for structural similarity searches:

DALI: www.ebi.ac.uk/dali/

VAST: www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html

GRATH: www.biochem.ucl.ac.uk/cgi-bin/cath/Grath.pl

CE: <http://cl.sdsc.edu/ce.html>



123

Example 4

Structural Comparisons

DALI output

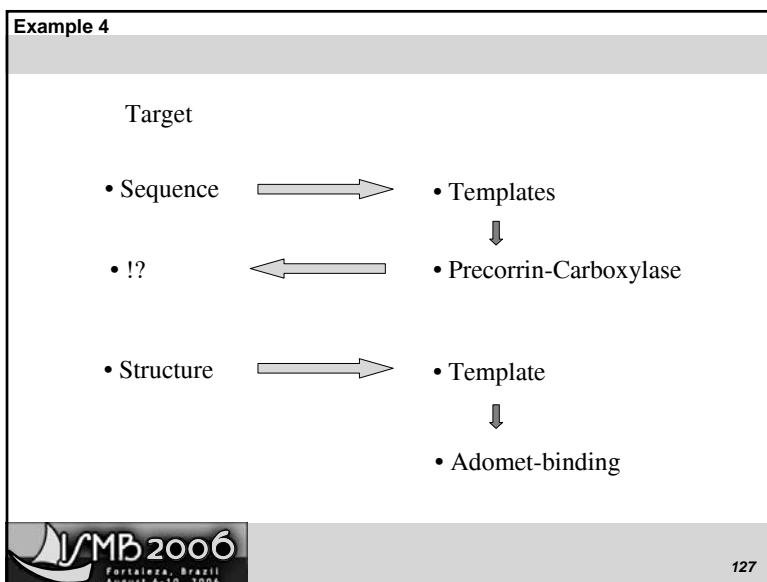
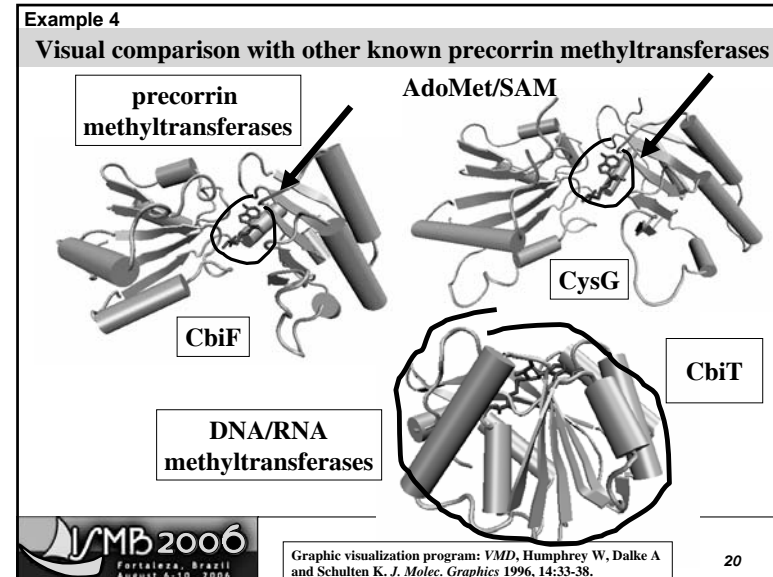
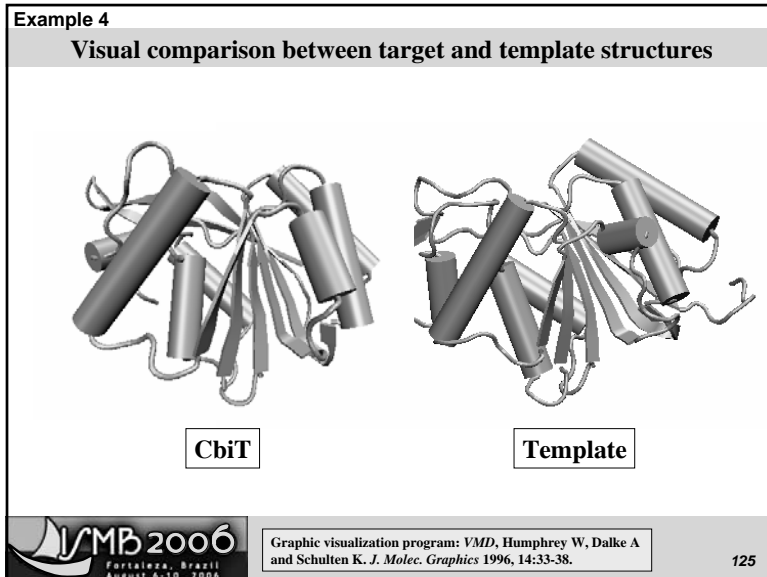
SEQLLENGTH 178
 NALIGN 374
 WARNING pairs with Z<2.0 are structurally dissimilar

```
## SUMMARY: PDB/chain identifiers and structural alignment statistics
NR. STRID1 STRID2 Z RMSD LALI LSEQ2 %IDE REVERS PERMUT NFRAG TOPO PROTEIN
1: 7768-A 1i9g-A 20.7 2.7 170 264 19 0 0 16 S Adomet-dependent methyltransferase
2: 7768-A 1dus-A 17.7 2.6 157 194 21 0 0 12 S mJ0882 - methyltransferase
3: 7768-A 1vid 15.7 2.5 159 214 14 0 0 14 S catechol o-methyltransferase
4: 7768-A 1m6y-A 15.7 2.4 151 289 22 0 0 14 S s-adenosyl-methyltransferase
5: 7768-A 1im8-A 14.5 2.6 152 225 16 0 0 15 S yeco (methyltransferase)
6: 7768-A 1fbn-A 14.2 2.9 148 230 17 0 0 13 S mJ fibrillarlin homologue
7: 7768-A 1nv8-A 14.0 2.9 151 271 17 0 0 16 S hemk protein
8: 7768-A 1khh-A 13.9 3.4 151 193 13 0 0 14 S guanidinacetate methyltransferase
9: 7768-A 1lxk-A 13.5 2.6 148 298 19 0 0 16 S methyltransferase
10: 7768-A 1jg3-A 13.3 2.9 153 295 12 0 0 14 S spermidine synthase (putrescine)
11: 7768-A 1kr5-A 13.2 2.5 139 218 23 0 0 15 S l-isoadipate-methyltransferase
12: 7768-A 1kp9-B 13.1 3.3 156 270 15 0 0 16 S cyclopropane-fatty-acyl-phosphol
13: 7768-A 1ej0-A 13.1 3.0 144 180 17 0 0 14 S ftsj (ftsJ methyltransferase)
14: 7768-A 2erc-A 12.9 3.2 145 235 21 0 0 16 S rRNA methyl transferase fragment
15: 7768-A 1ril-A 12.9 3.0 151 252 19 0 0 15 S mRNA capping enzyme - methyltransferase
..
```



DALI: <http://www.ebi.ac.uk/dali/>

124



Example 4
SAM binding site analysis

Motif	M.HhaI	M.TaqI	COMT	role
I	AGxGG	PSxAxGP	GAxxG	H-bonds
II	E40 W41	E71 I72	E90 M91	H-bond with Ribose hydroxyls VdW with adenine
III	D60	D89	S119	H-bonds to N atoms
V	L100 F18	L142 F146	W143 H142	VdW contacts with adenine

Schluckebier G, O'Gara M, Saenger W, and Cheng X *JMB* 1995, 247:16-20

128

Example 4

SAM binding site analysis

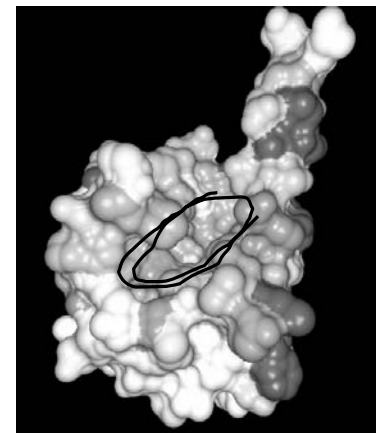
Motif	M.HhaI	M.TaqI	COMT	role
I	AGxGG	PSxAxGP	GAxxG	H-bonds
II	E40	E71	E90	H-bond with Ribose hydroxyls
	W41	I72	M91	VdW with adenine
III	D60	D89	S119	H-bonds to N atoms
V	L100	L142	W143	VdW contacts with adenine
	F18	F146	H142	



129

Example 4

SAM binding site analysis

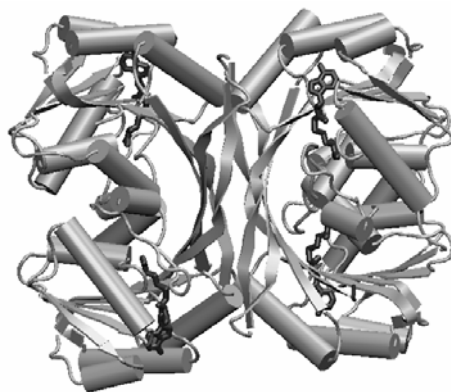


CONSURF: <http://consurf.tau.ac.il/>

130

Example 4

CbiT-SAM crystal structure

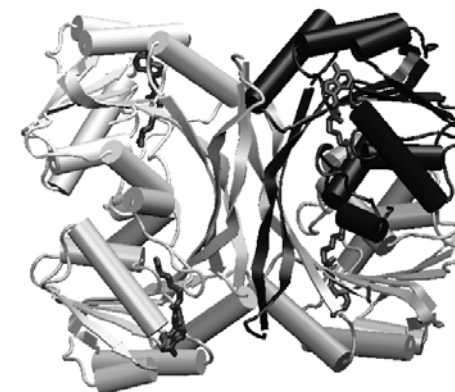


Graphic visualization program: VMD, Humphrey W, Dalke A and Schulten K. J. Molec. Graphics 1996, 14:33-38.

131

Example 4

CbiT-SAM crystal structure

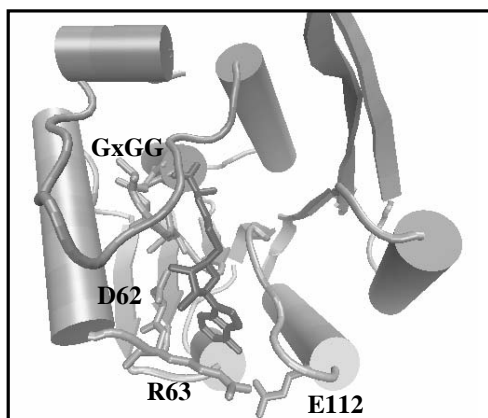


Graphic visualization program: VMD, Humphrey W, Dalke A and Schulten K. J. Molec. Graphics 1996, 14:33-38.

132

Example 4

SAM binding site analysis

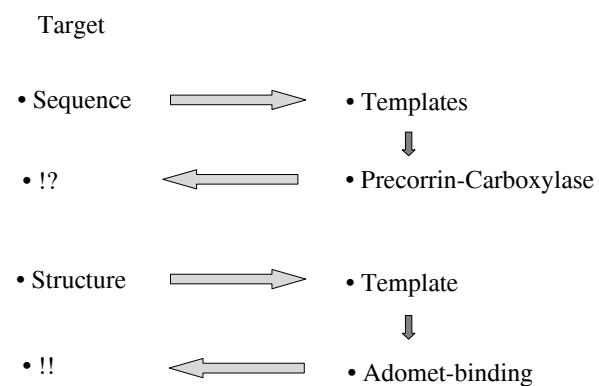


Graphic visualization program: VMD, Humphrey W, Dalke A and Schulten K. *J. Molec. Graphics* 1996, 14:33-38.

133

Example 4

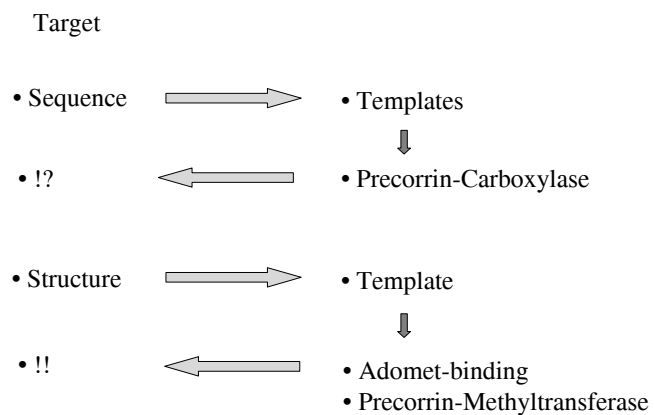
Functional annotation of the target



134

Example 4

Functional annotation of the target

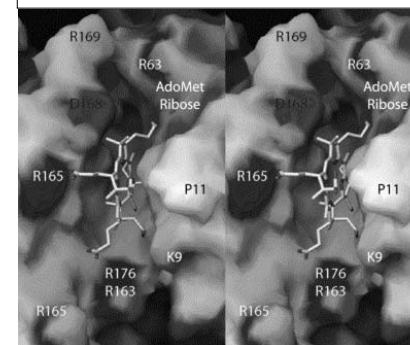


135

Example 4

Precorrin binding site analysis

GRASP: <http://tranter.bioc.columbia.edu/grasp/>

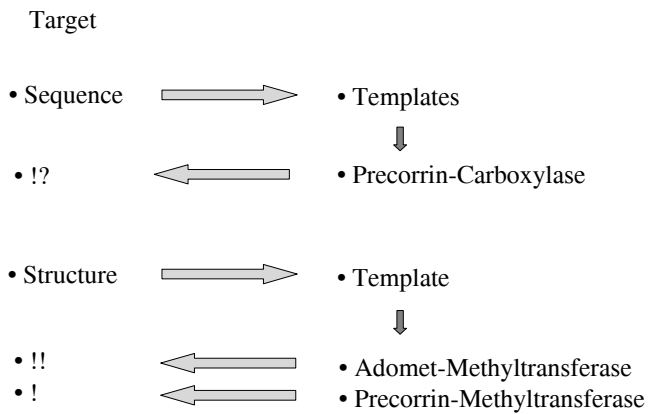


* From Keller JP, Smith PM, Benach J, Christendat D, deTitta GT, and Hunt JF *Structure* 2002, 10:1475-87

136

Example 4

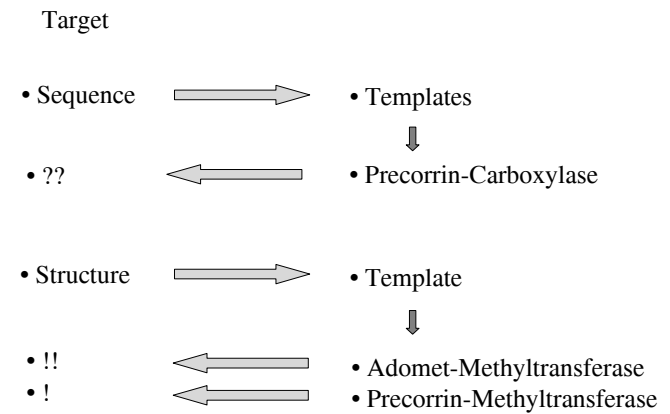
Functional annotation of the target



137

Example 4

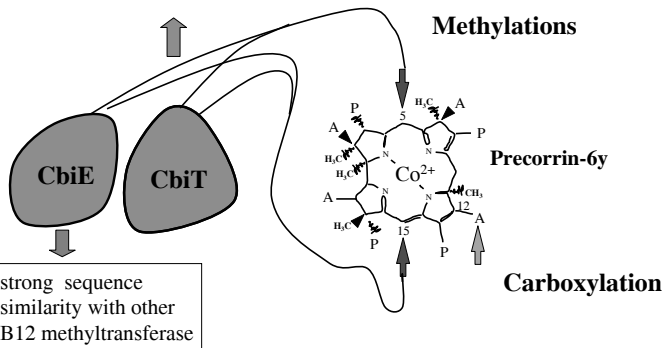
Functional annotation of the target



138

Example 4

Structural similarity
with methyltransferases



139

Example 4

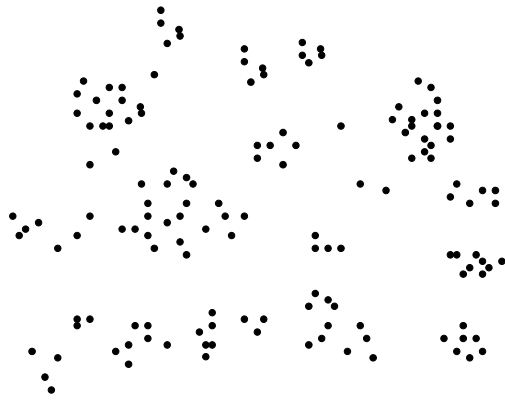
Hypothesis on precorrin **carboxylation**:

- Spontaneous after double methylation
- CbiT protein is also a carboxylase

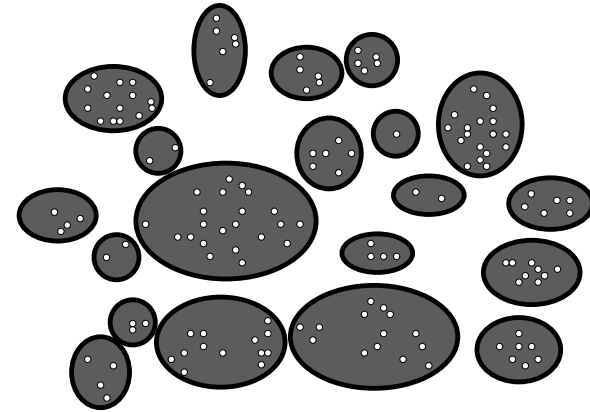


140

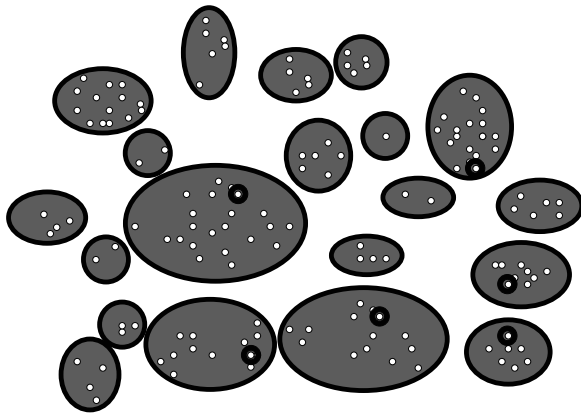
Structural genomics



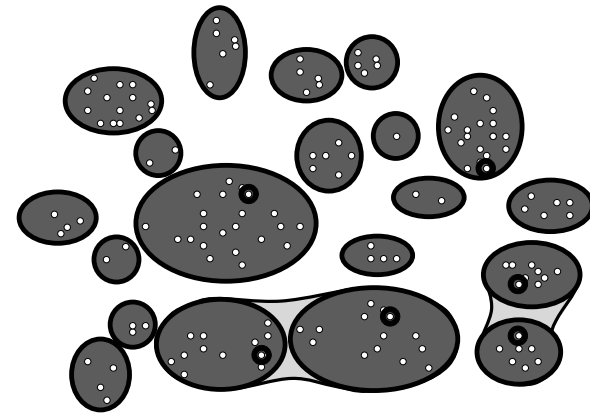
Structural genomics: clustering



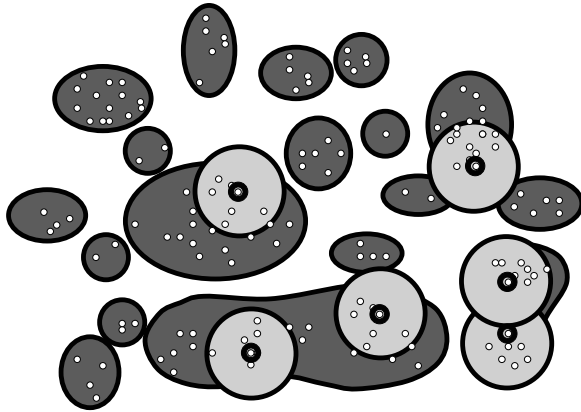
Structural genomics: structure determination



Structural genomics: merging families



Structural genomics: checking leverage



Example 5

Many structures better than one (!)



Example 5

Structural Classification of Proteins



Search the scop database [scop 1.69]

You can use this search engine to search the SCOP database using several access methods (including *sumit*, *sid*, *seccs*, PDB identifiers, and any word that appears in any of the SCOP pages) as well as more sophisticated options. Please read the [release notes](#) for a detailed explanation and examples. This kind of search is internal to a SCOP release and therefore will always provide complete results.

By checking the PDB box, you can also search SCOP using the external MSDlite search engine for words that appear in several *text fields* in the corresponding PDB file (including header, author names, abstract, and MeSH terms from the primary citation). Please refer to [MSDlite](#) for more details.

PUA

- ☒ Search the SCOP database.
☐ Search the PDB database using [MSDlite](#).
[Retrieve information](#)
[Clear](#) the search form.

Copyright © 1994-2005 The scop authors / scop@mrc-lmb.cam.ac.uk
 July 2005



Example 5

Structural Classification of Proteins



Fold: PUA domain-like

pseudobarrel; mixed folded sheet of 5 strands; order 13452; strand 1 and 3 are parallel to each other

Lineage:

1. Root: scop
2. Class: [All beta proteins](#) [48724]
3. Fold: [PUA domain-like](#) [88696]
pseudobarrel; mixed folded sheet of 5 strands; order 13452; strand 1 and 3 are parallel to each other

Superfamilies:

1. [PUA domain-like](#) [88697] (4)
 1. [PUA domain](#) [88698] (6) [ms](#)
RNA-binding domain
 2. [ATP sulfurylase N-terminal domain](#) [63801] (4) [ms](#)
contains extra structures; some similarity to the PK beta-barrel domain
 3. [YggJ N-terminal domain-like](#) [89451] (2) [ms](#)
 4. [Hypothetical protein EF3133](#) [110339] (1) [ms](#)
DUF984; Pfam 06171



See also: Iyer LM, Burroughs AM, and Aravind L. *Bioinformatics*. 2006; 22(3):257-63.

Example 5

Structural Classification of Proteins

Family: **Hypothetical protein EF3133**

DUF984; Pfam 06171

Lineage:

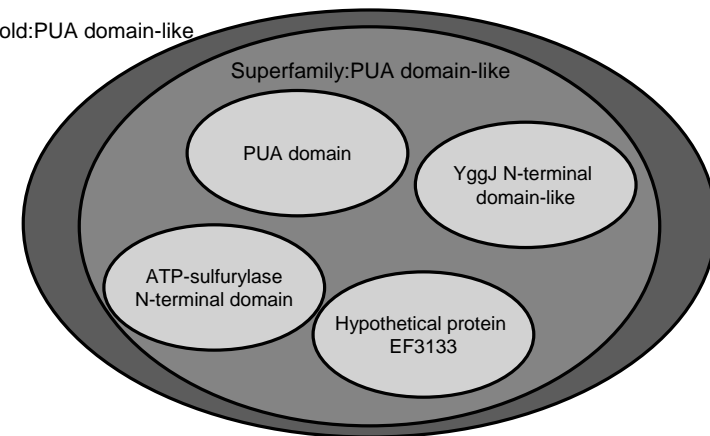
1. Root: scop
2. Class: All beta proteins [48724]
3. Fold: PUA domain-like [88696]
pseudobarrel, mixed folded sheet of 5 strands; order 13452; strand 1 and 3 are parallel to each other
4. Superfamily: PUA domain-like [88697]
5. Family: Hypothetical protein EF3133 [110339]
DUF984; Pfam 06171

Protein Domains:

1. Hypothetical protein EF3133 [110340]
 1. Enterococcus faecalis [110341] (1)
 1. 1062 [50]
SQ 0822D1 # 1 Structural genomics target
 1. chain_a [106541] [50]
 2. chain_b [106542] [50]

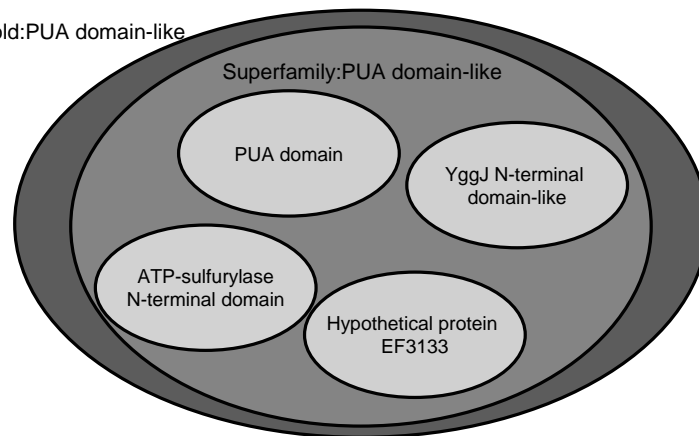
Example 5

Fold: PUA domain-like



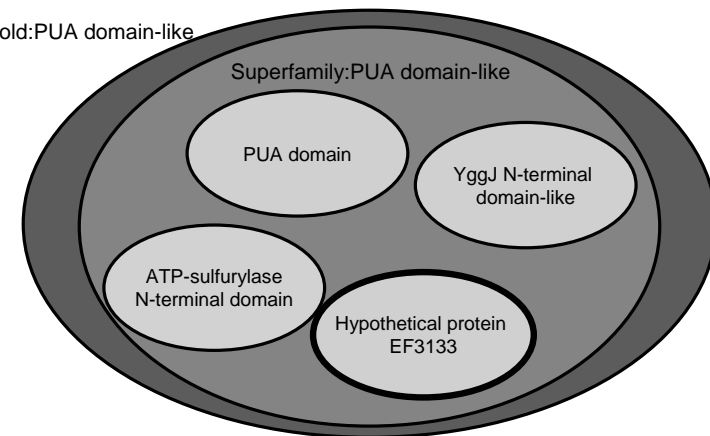
Example 5

Fold: PUA domain-like

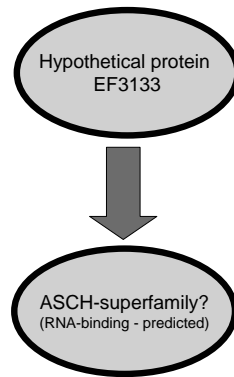


Example 5

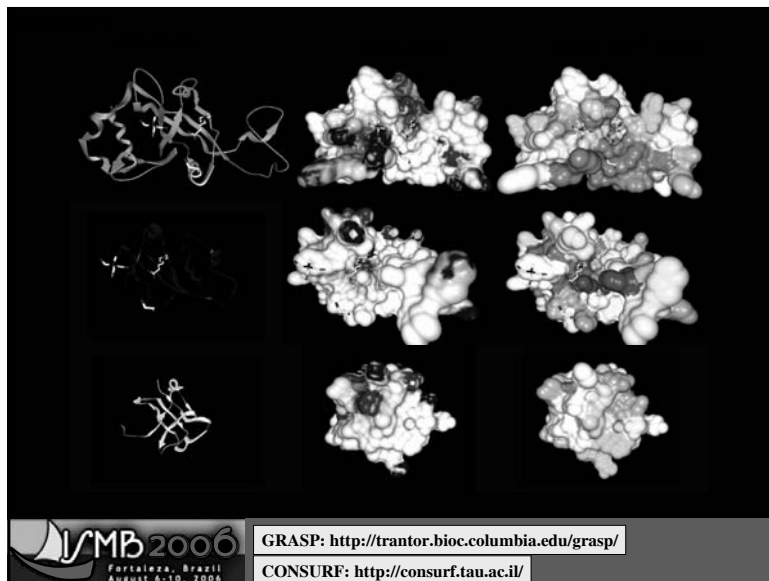
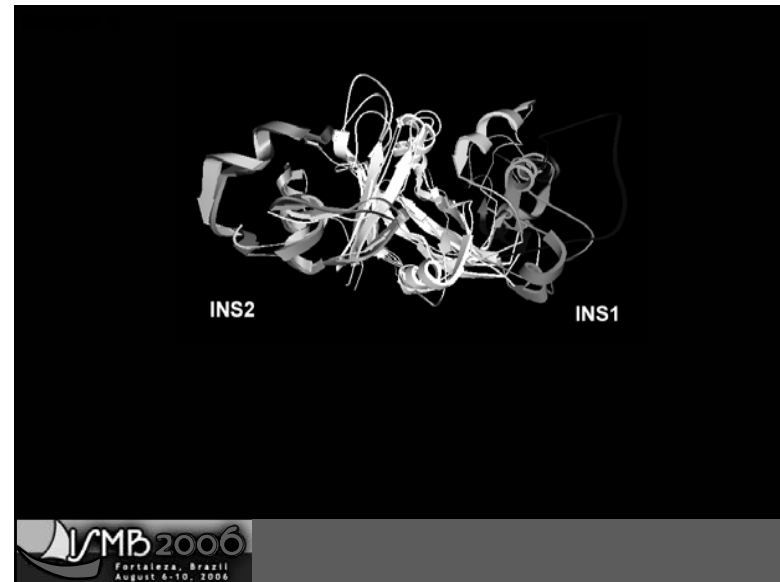
Fold: PUA domain-like



Example 5



Iyer L.M., Burroughs A.M., and Aravind L. Bioinformatics. 2006; 22(3):257-63.



GRASP: <http://trantor.bioc.columbia.edu/grasp/>

CONSURF: <http://consurf.tau.ac.il/>