Genome-wide structural and functional protein characterization by *ab initio* protein structure prediction

Lars Malmström

Department of Electrical Measurements
Lund Institute of Technology
Lund University

ISBN 91-628-6689-3
Report ???/05
ISSN 0346-6221
ISRN LUTEDX/TEEM - - 10?? - - SE

LUND UNIVERSITY

Genome-wide structural and functional protein characterization by ab initio protein structure prediction

Lars Malmström

Genome-wide structural and functional protein characterization by *ab initio* protein structure prediction

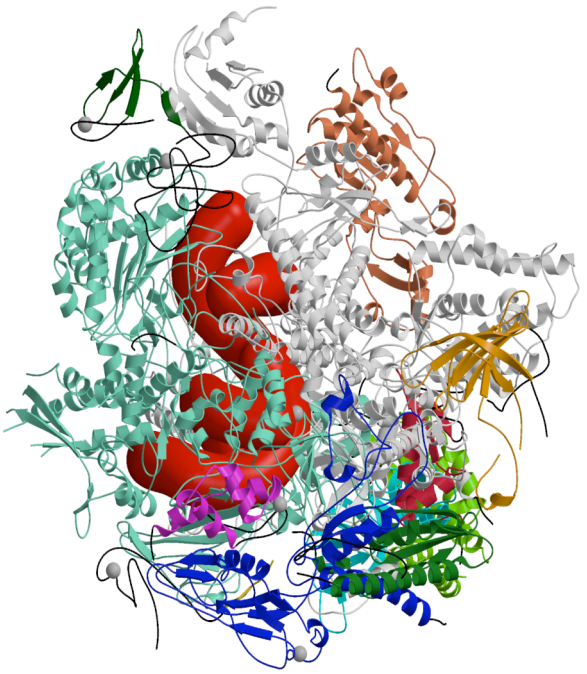
Lars Malmström

Department of Electrical Measurements
Lund Institute of Technology
Lund University

Report 4/05
ISSN 0346-6221
ISRN LUTEDX/TEEM—1084—SE
ISBN 91-628-6689-3

Akademisk avhandling som för avläggande av teknologie doktorsexamen vid tekniska faktulteten vid Lunds Universitet kommer att offentligen försvaras fredagen den 16 december 2005, kl. 10.15 hörsal E:1406 i E-huset, Lund
Fakultetsopponent: David Fenyö, The Rockefeller University, New York, New York, USA

Academic thesis which, by due permission of the faculty of Engineering at Lund University, will be publicly defended on Friday 16th of December, 2005, at 10.15 am in lecture hall E:1406, at the Electrical Enginering building, Lund
Faculty opponent: David Fenyö, The Rockefeller University, New York, New York, USA



A RNA polymerase protein complex

"We are drowning in information
but are starved for knowledge"

Table of Content

Abbreviations

CASP	Critical assessment of methods of protein structure prediction
DAG	Directed Acyclic Graph
GO	Gene Ontology
HMM	Hidden Markov Model
MS	Mass spectrometry
NTP	nucleoside triphosphate
PDB	The Protein Data Bank
ORF	Open Reading Frame
RMSD	root mean square deviation
SCOP	Structural Classification of Protein
SQL	Structured Query Language
TMHMM	Software to identify transmembrane helices from amino acid sequence
Y2H	Yeast two Hybrid

List of publications

I. Philip Bradley[†], **Lars Malmström**[†], Bin Qian[†], Jack Schonbrun[†], Dylan Chivian, David Kim, Jens Meiler, Kira Misura and David Baker[*]. Free Modeling with Rosetta in CASP6. *Proteins* (2005), Accepted

II. Tony Hazbun[†], **Lars Malmström**[†], Scott Anderson, Beth Graczyk, Bethany Fox, Michael Riffle, Bryan Sundin, Jennifer Aranda, Hayes McDonald, Chun-Hwei Chiu, Brian Snydsman, Philip Bradley, Eric Muller, Stanley Fields, David Baker, John Yates III and Trisha Davis. Assigning function to yeast proteins by integration of technologies. *Mol Cell.* (2003) **12**:1353-65

III. **Lars Malmström**, Johan Malmström, György Marko-Varga. and Gunilla Westergren-Thorsson Proteomic 2DE database for spot selection, automated annotation, and data analysis. *J Proteome Res*. (2002) **1**:135-8

IV. **Lars Malmström**, György Marko-Varga, Gunilla Westergren-Thorsson, Thomas Laurell and Johan Malmström. 2DDB - a Bioinformatics Solution for Analysis of Quantitative Proteomics Data. Submitted to BMC bioinformatics

V. **Lars Malmström**[†]**,** Michael Riffle[†], Richard Bonneau, Charlie Strauss and David Baker. Genome-wide de novo structure prediction for Saccharomyces cerevisiae and integration of a structural compendium with the Gene Ontology database. Manuscript; to be submitted to PLoS Biology

[†)] Contributed equally to this work

Publications not included in this thesis

Bonneau, R., Strauss, C., Rohl, C., Chivian, D., Bradley, P., **Malmström**, L., Robertson, T. and Baker, D. De Novo Prediction of Three-dimensional Structures for Major Protein Families. *J Mol Biol*. (2002) **322**:65-78

Chivian,D., Kim, DE., **Malmström**, L., Bradley, P., Robertson, T., Murphy, P., Strauss, CE., Bonneau, R., Rohl, CA. and Baker, D. Automated prediction of CASP-5 structures using the Robetta server. *Proteins*. (2003) **53**:524-33

Malmstrom,J., Larsen, K., **Malmström**, L., Tufvesson, E., Parker, K., Marchese, J., Williamson, B., Patterson, D., Martin, S., Juhasz, P., Westergren-Thorsson, G. and Marko-Varga, G.
Nanocapillary liquid chromatography interfaced to tandem matrix-assisted laser desorption/ionization and electrospray ionization-mass spectrometry: Mapping the nuclear proteome of human fibroblasts. *Electrophoresis.* (2003) **24**:3806-14

Malmström,J., Larsen, K., **Malmström**, L., Tufvesson, E., Parker, K., Marchese, J., Williamson, B., Hattan, S., Patterson, D., Martin, S., Graber, A., Juhasz, HP., Westergren-Thorsson, G. and Marko-Varga, G. Proteome annotations and identifications of the human pulmonary fibroblast. *J Proteome Res*. (2004) **3**:525-37

Riffle,M., **Malmström**, L. and Davis, TN. The yeast resource center public data repository. *Nucleic Acids Res*. (2005) **33**:D378-82

Chivian,D., Kim, DE., **Malmström**, L., Schonbrun, J., Rohl, CA. and Baker, D. Prediction of CASP-6 structures using automated Robetta protocols. *Proteins*. (2005) accepted

Kim,DE., Chivian, D., **Malmström**, L. and Baker, D. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. *Proteins*. (2005) accepted

1. Summary

Very little is known about a considerable part of all proteins and it is time consuming and expensive to study each individual protein to determine its function, structure and cellular role. Proteins retain structural, functional and sequential characteristics from ancestral proteins and hence two proteins that share a common ancestor, i.e. are homologs, will to some extent have similar sequence, structure and function. One way to learn something about a protein is to identify its homologous and use information from those homologs to annotate the protein of interest. Close homologs with a common ancestor can be detected using sequence alone, but more distant homologs cannot. Structure is more conserved than sequence and enables detection of a common ancestor between more distantly related proteins and thereby also enabling transfer of information to a larger fraction of the uncharacterized proteins. This thesis covers my efforts to develop a method to use ab initio protein structure prediction to detect distant homologs and use the homologs to annotate proteins from the genome of Saccharomyces cerevisiae.

The ab initio protein structure prediction software used in this thesis, Rosetta, can predict a protein's tertiary structure using the amino acid sequence alone. Rosetta works by reducing the search space by approximating the local conformation with conformations from the protein data bank, and judging the over all fitness of the simulated protein structure through a statistically derived energy function. The program has been successful in the last three Critical assessment of techniques for protein structure prediction (CASP) and the results from the last

CASP is reported in Paper I. Distant homologs can be detected by comparing the structures generated by Rosetta with structures from the Protein Data Bank (PDB). In general, however, such a comparison is noisy, that is, gives many answers, of which only a few are correct. The noise can be filtered out by utilizing the fact that there is a strong relationship between protein function and protein structure, and either use functional information from a database or infer functional information from one or more experimental high-throughput technologies. This idea was tested in Paper II were 100 proteins were investigated using protein structure prediction, yeast two hybrid, fluorescent microscopy and mass spectrometry. The data from all four technologies was integrated and 77% of the proteins were assigned a function.

Data integration is very labor-intensive when done by hand, and the amount of information generated for each protein investigated is substantial. Everything needs to be automated and all data have to be stored and managed in an efficient way to be able to apply this technology on a genome-wide scale. Paper III and Paper IV cover information management, that is, how the data used and produced in the project is organized and stored. Paper V reports both how we automated the integration process using the software described in Paper I and II and the application of the technology to the genome of Saccharomyces cerevisiae.

2. Introduction

This thesis is about what can be learned about proteins using a state of the art program, Rosetta. It involves developing means to organize and interpret the data, integrate the data with information about the proteins and making everything accessible to the research community. It touches a number of different research areas, including biology, systems biology, biophysics, computer science, information science, mathematics and statistics. My goal is to try to generate information useful to biologist who are trying to gain a better understand how we and other organisms around us function, and hence, this thesis is written from a biological perspective. The first section will give a brief introduction to proteins, protein structure and protein function. After that I will give a short overview of the current state of protein structure prediction and protein function prediction. The final section covers my contribution. I have gathered data by running a large number of programs, collected data from the experimental procedures and organized everything in a relational database. This information resource is overwhelming when presented in a raw format, and to alleviate the problem, I created statistical models and information integration schema. All the data generated will be available to the public by Spring 2006.

Proteins constitute most of the dry mass of a cell and perform nearly all of the thousands of tasks a cell carries out. Proteins catalyze chemical reactions, build up the cells cytoskeleton, are involved in signaling, DNA replication and transportation of metabolites. Proteins are strings of 20 different kinds of amino acids, typically between 50 and 2000 amino acids long and are put together according

to a DNA template, a gene, by a large molecular structure, the ribosome. As the nascent protein emerges from the ribosome, it rapidly folds to an energy minimum, a specific tertiary structure referred to as the proteins native state or native fold [1,2]. Sanger and colleagues sequenced the first protein, Insulin, in 1955 [3], work taking years of painstaking experiments. Since then the efficiency of sequencing proteins has increases many orders of magnitude. The development of the PCR [4,5] and the sequential advent of fast nucleotide sequencing made it possible to indirectly identify millions of protein sequences by translating the nucleotide sequence into amino acid sequence. This has led to an ever-increasing production of biological sequences. A considerable part of the 37.3 million publicly available sequences from 165 000 organisms are uncharacterized [6,7,8], meaning that nothing is known about the protein itself, and no characterized homologous protein (see Section 3.3.1 for definition) can be detected with confidence using sequence alone. This hinders us from understanding biology from a global perspective. Many of the known proteins are related, both in an evolutionary perspective and in a functional, structural and sequential perspective and the further identification of such relations can greatly speed up the characterization of uncharacterized proteins.

3. Background

Understanding the cell on a molecular level and being able to predict outcomes of perturbations with high accuracy will allow us to prolong our lifespan and at the same time improve the quality of life and health. We are beginning to understand the flow of information in the cell how the long-term information storage, DNA, gets translated into protein via a temporary messenger, the messenger RNA or mRNA for short. These proteins self-organize into complex systems of interaction an regulation to build up a cell together with a number of other chemical classes, such as lipids. The next level of organization is how these cells interact, with each other or with the environment.

Technology to sequence DNA is fast and reliable and in 1995 the first full genomic sequence a living organism, Haemophilus influenzae, was finished [9]. This feat was followed by the full genomic sequence of yeast in 1996 [10] and the human genome was sequences in 2001 [11,12]. Several million genes are sequences and new genomes are sequenced every month [13]. There is a large gap between the number of proteins we known the structure of since it is possible to derive a proteins sequence from the sequence of the gene that codes for that protein and sequencing is much faster than solving protein structures experimentally. Also, the majority of these proteins remain uncharacterized, i.e. we have no information about what this proteins function is, where it is located in the cell and with what other proteins it interact. Much can be learned of a protein by knowing its three dimensional structure and since solving protein structures is a difficult problem, generating protein structures through computational means

would allow us to bridge the gap between the number of known protein sequences and known protein structures. Once we have some information about the proteins, the parts that build up the cell, the next step is to elucidate how these proteins interact with each other and now the cell is regulated.

The next sections will cover protein structure, the relation between a structure and function. The last part of this background will cover the fundamentals of bioinformatics and the current state of protein structure prediction.

3.1. Protein Structure

Proteins have evolved to become highly efficient doing what they do, working in a crowded environment [14,15]. Proteins species are present in vastly different concentrations as the number of ranges from just a few copies per cell to several millions [16]. Some of them are small, globular structures catalyzing a chemical reaction in the cytosol, some are embedded in a membrane transporting molecules from one compartment to another and yet others are parts of large molecular machines, or complexes, capable of transcribing DNA to mRNA. Proteins can build up the cytoskeleton by polymerizing into fibers and others make muscles contract. Clearly, a group of chemicals that can perform such diverse set of functions have to be versatile. This versatility is achieved by combining amino acids in a long string, each amino acid having a specific chemical property and size. In addition to perform all these functions, proteins have to interact with each other, either regulating function, or building up a molecular machine. These interactions have to be specific since incorrect interactions lead to diseases, for review see [17]. Cells have to be able to dispose proteins easily and to regulate their

function, for example, by adding a phosphate group. For this to work, proteins cannot be too stable. Too stable proteins will not be affected by minor modifications, or be easily degraded by the cell. All these things influence the proteins structure.

A proteins three-dimensional structure, or tertiary structure, with the lowest energy is referred to as its native state or fold and the process that starts with the unfolded protein and ends with its native state is called protein folding. Although proteins native states differ extensively (see Figure 1) there are common features that are repeated over and over again. The dominating features, the alpha helix and the beta sheet, are called secondary structure elements, and these elements come together to form the tertiary structure. All information needed about the tertiary structure is encoded in the primary sequence as Anfinsen elegantly showed studying the enzymatic properties and the disulfide bonds of ribonuclease [1,18]. This finding gave biologist hope that they one day would be able to determine a proteins structure given only its primary sequence.

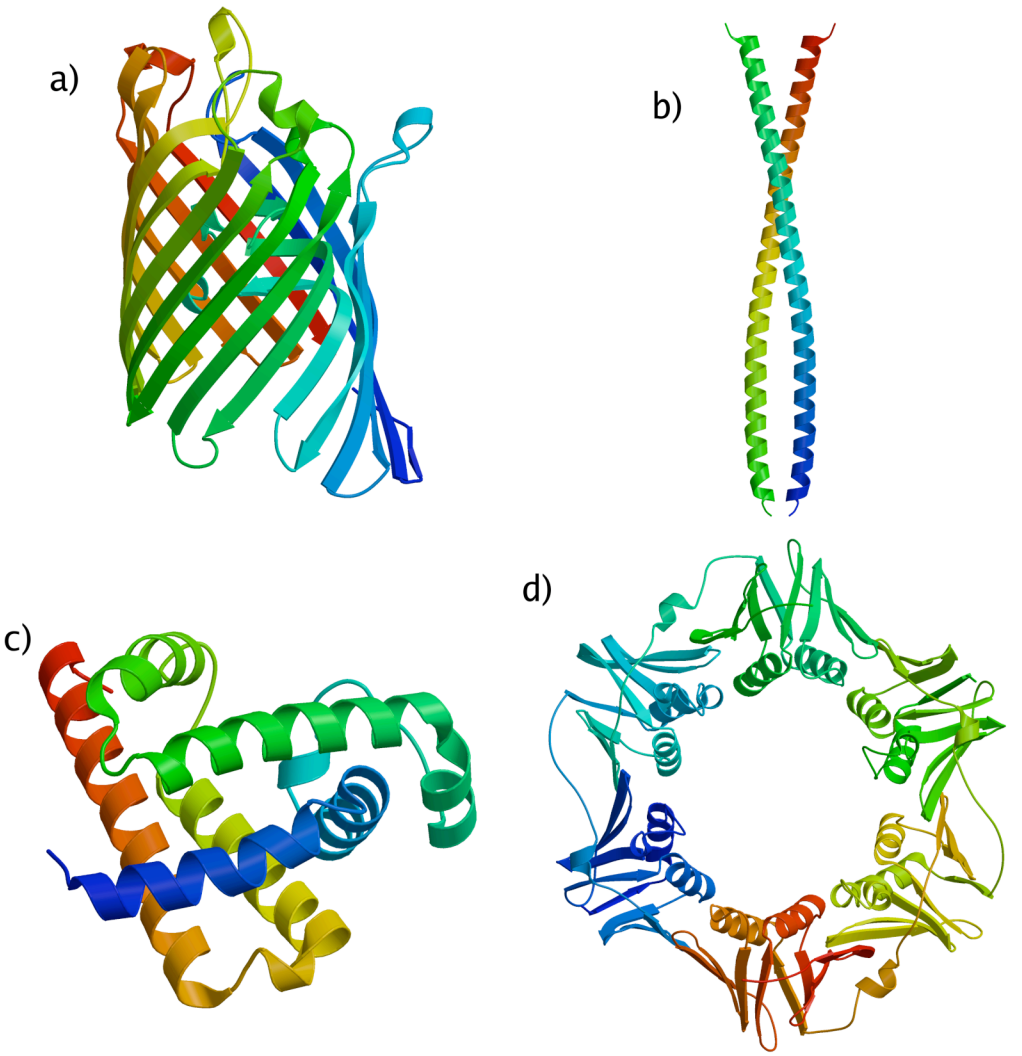


Figure 1. Protein Structures

Proteins come in many different shapes and sizes. This image illustrates some examples: (a) a porin, a beta barrel embedded in the plasma membrane (PDB ID: 1HXT); (b) myoglobin, a globular protein (PDB ID: 1A6N); (c) lamin coil from human, biological unit (PDB ID: 1X8Y) and (d) an E.coli DNA polymerase Beta subunit (PDB ID: 1MMI).

3.1.1. What Can We Learn from Protein Structures?

When the structure of DNA was solved [19], insight into how genetic information is passed along from mother cell to daughter cell was learned. The atom de-

tails of the DNA molecule offered explanations to how it could be replicated with high fidelity and knowing this enabled the development of tools to create DNA, copy DNA and sequence DNA. Scientist knew that proteins played an important role and tried to solve protein structures in hope that it would give a similar revolution that the structure of DNA had done. When Kendrew solved the protein structure of myoglobin, it was clear that a proteins structure is much less regular than DNA and much more complex to utilize. Nevertheless, knowing the structure of a protein at the atomic level can provide powerful means by explain how they function. As an example, I will introduce an RNA polymerase (pol) II complex (see Figure 2). This complex consists of 12 protein subunits and two DNA strands and one RNA strand are also visible in the crystal structure. One of the important things learned from this structure is how the polymerase separates the DNA chains prior to transcription. This is essentially done with six amino acids, three positively charged amino acids, R326, K330 and R337 (green amino acids in Figure 2b), that pulls the negatively charged DNA strand away from the other DNA strand. Three negatively charged amino acids (E1403, E1404 and E1407; blue amino acids in Figure 2b) on the other side repel the DNA strand, resulting in a separation of the two strands. Evolutionary information was also learned from this structure In addition to the functional insight. Residues at the nucleoside triphosphate (NTP) site are universally conserved, and hence the suggesting that the NTP selection mechanism is universal. The amino acids in the NTP site are not sequential and to identify them without a structure is not possible.

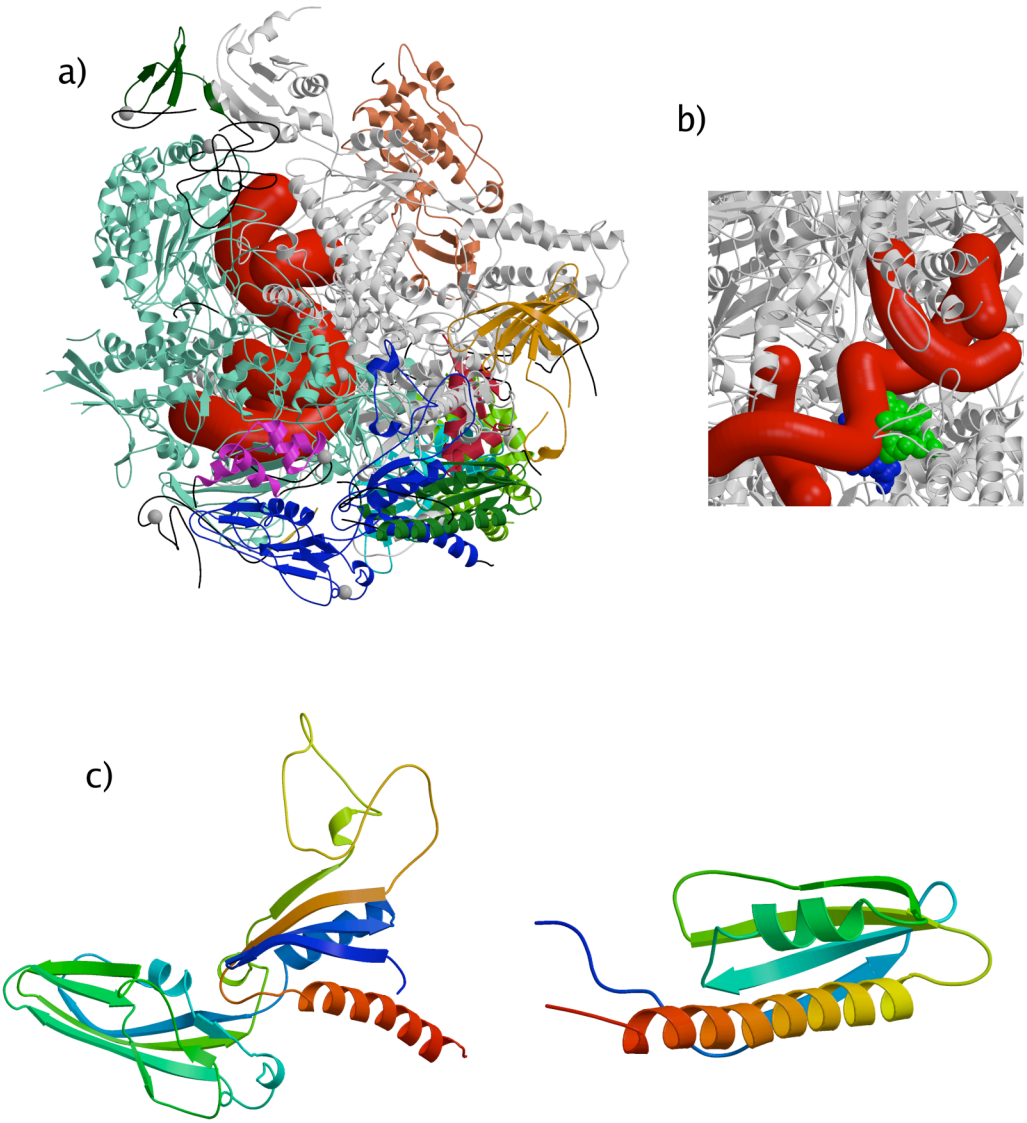


Figure 2. Protein Complexes

a) This protein complex consists of 12 distinct proteins that together build up this RNA polymerase [20]. This RNA polymerase transcribes DNA to RNA and the individual proteins are color in different colors to demonstrate how these proteins come together for form the complex. b) The DNA is red, the proteins is gray. The amino acids responsible for the separation of DNA is displayed as space-fill, there the negatively charged amino acids are colored in blue and the positively charge amino acids in green. c) Two of the subunits are shown isolated from the rest of the complex to demonstrate that these subunits are compact domains that probably fold independ-

ently of the complex and assembled after they have assumed their native fold.

A number of diseases are directly related to protein structure and protein folding. Sickle cell anemia is a well-known disease in which a single amino acid substitution in the beta-chain in hemoglobin renders hemoglobin insoluble, which in turn deforms the red blood cells as the insoluble molecules crystallize. Another group to diseases are caused by accumulation of plaques in cells. Examples of diseases caused by protein plaques are Creutzfeld-Jakob's disease and Alzheimer's disease [21]. The hope is that understanding why these proteins become misfolded and subsequently accumulates, will allow to either prevent the misfolding or stop the accumulation of more proteins, and thereby stop the progression of the disease.

3.1.2. The free energy of a protein

Protein sequences have evolved to fold into a reproducible stable structure [22,23]. According to the widely accepted "thermodynamic hypothesis" the native conformation of a protein corresponds to the global free energy minimum of the protein/solvent system [1,2]. Naturally, the most interesting case is the free energy of the native conformation in an aqueous solution, and to some extent the free energy of a completely unfolded chain in an aqueous solution and the larger the gap between the two states are, the more stable the protein is. Many factors influence the free energy of a protein. The hydrophobic effect [24] plays a crucial role in protein stability and folding. This effect is the result of amino acids that are non-polar are hidden inside the protein, and hence shield away from the polar water. Charged amino acids interact both with the solvent on the surface and in cavities, but also interact with one another. Electrostatics contributes to

the overall stability of proteins, especially the so-called salt bridges. van der Waal forces is the favorable interaction between atoms in the molecule. The forces are weak, but the great number of them makes their contribution important. Hydrogen bonds are important for the overall stability of protein structures. Last, covalent bonds between amino acid residues are stabilizing. The most well known covalently bound residues are the disulfide bridge. See [25] for a review of forces and protein structure.

3.1.3. Secondary Structure

The protein backbone contains polar groups that, if they are not part of a hydrogen bond, are energetically unfavorable inside the protein. There are two major modes utilized to hydrogen-bond all polar groups of the backbone, the first one is local in sequence - the alpha helix, and in the second, the donor-group and the acceptor-group can be separated in sequence - the beta strand. Alpha helices and beta-strands are the two major forms of secondary structure, and there are a number of smaller forms as well, such as the 3-10 helix, the beta-turn and the pi-helix. For an illustration of the different secondary structures, see Figure 3. The helices are stable by them selves, but the beta-strands have to align them selves to another beta-strand to hydrogen bond with. These secondary structure elements come together to form the tertiary structure.

3.1.4. Tertiary Structure

The tertiary structure, or the three-dimensional conformation or the protein, is the result of the secondary structure elements coming together in an energetically

favorable way. The beta-strands form one or many beta-sheets and the helices pack together on top of the sheets. The tertiary structure is the conformation with lowest energy in an aqueous solution. Anfinsen [18] demonstrated that all the information to specify the three dimensional protein structure is contained in the primary sequence. This has led to the assumption that it is possible to determine a proteins three-dimensional structure knowing only its primary sequence. For an illustration of a tertiary structure, see Figure 3. Proteins resemble organic crystals when looking at average density, but look like liquids when looking at their free volume distribution and many protein structures have cavities, [26].

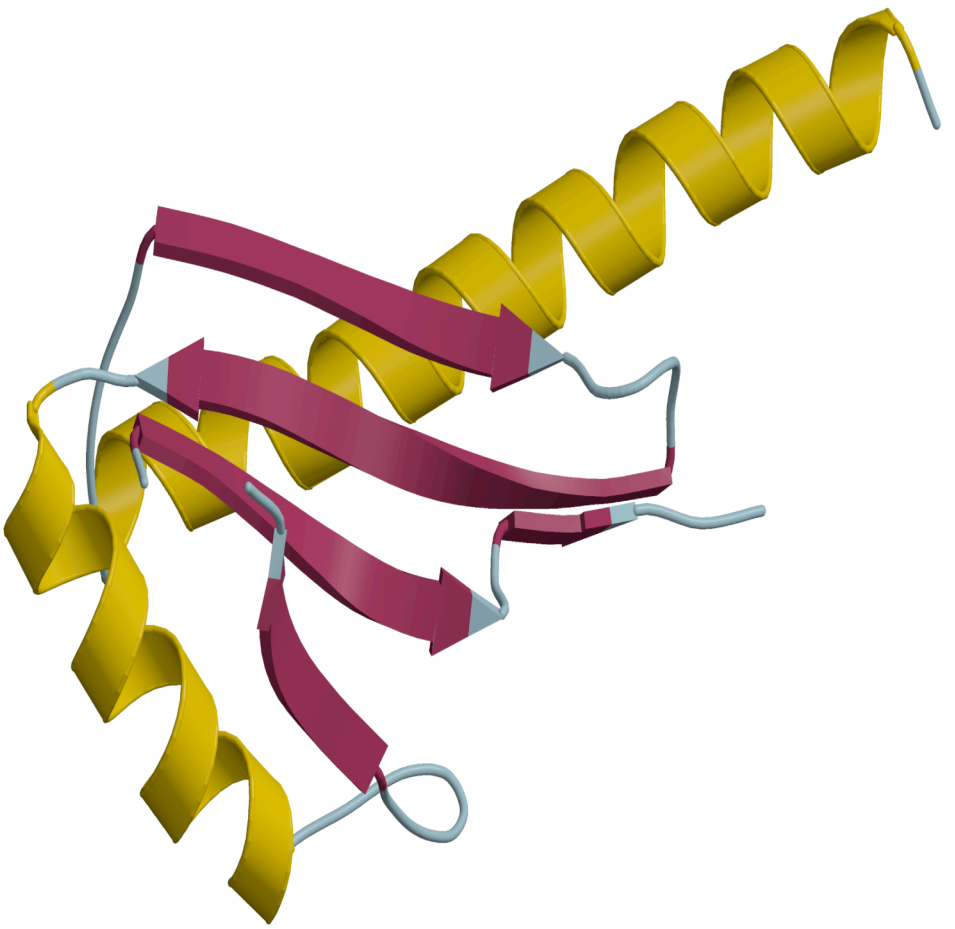


Figure 3. Tertiary Structure

A tertiary structure of a protein. Alpha helices are gold, sheets are maroon, and coils/turns are light blue.

3.1.5. Other structural features

Between 20% and 30% of all proteins are or have some part embedded in the plasma membrane or organelle membranes [27,28]. The parts of the protein that is localized within the membrane are called membrane domains. In the majority of the cases, the trans membrane domain is an alpha helix that spans the mem-

brane, but there are also beta barrels situated in the membrane. The chemical milieu in a membrane is hydrophobic, making the majority of the amino acids exposed to the membrane hydrophobic. The composition differences between trans membrane domains and other secondary structure elements make them relatively easy to detect. One of the best prediction algorithms, TMHMM, is over 95% accurate [27]. Membrane proteins also have an orientation; some have their N-terminal on the inside of the membrane and some have it on the outside. Another prominent feature is the coiled-coil [29], two alpha helices twisted around each other in a super-helix. Other parts of proteins are unstructured or disordered [30,31,32] which means that they are constantly moving and hence not as stable as the rest of the protein. Low complexity regions [33] are areas of the protein where the sequences is very repetitious, or is built up from few amino acid types. All these features have implications for the protein structure and protein structure prediction. They can be detected using software.

3.1.6. Structural Domains

It became clear that there is a structural organization of large proteins into so-called protein domains as the tertiary structure of more proteins got determined [34,35,36,37]. One definition of a protein domain is a polypeptide chain or part of a polypeptide chain that can fold into a stable, tertiary structure. A domain has its own hydrophobic core and the amino acids in a structural domain have less interactions with amino acids that belongs to other domains on the same polypeptide compared to amino acids within the domain. Furthermore, domains are commonly co-linear, and hence are built up by one contiguous peptide chain.

Different regions of a peptide chain or even different peptide chains build up some domains. One domain is often associated with a particular function [35,36,37] and are typical between 40 and 350 amino acids. Some domains are present in numerous proteins and can be looked upon as plug-ins or modules. Two classic examples are immunoglobulin domain [38] and the SH2 domain [39], present in various non-related proteins. For an illustration, see Figure 4 For a review about identifying domains, see [40]. NCBI offers the Conserved Domain Database, CDD, [41] annotating protein sequences with domain information together with Conserved Domain Architecture Retrieval Tool, CDART, [42] classifying proteins with their domain content, and letting the user retrieve proteins with similar architecture in the same fashion Pfam [43,44] does. CDD and CDART are fairly conservative and there coverage is hence limited.

Figure 4. Structural Domains

Sec23/24-Sar1 pre-budding complex, a multi-domain protein (1M2O) [45] is display with the domains colored. The pre-budding complex is an essential part of the COPII vesicular transport system. The 765 amino acid long peptide chain is organized into 5 structural domains, one discontinuous beta sandwich domain in blue between 2-44 and 391-523 (SCOP SCCS: b.2.8.1). The 4 other domains are as follows: Zn-finger domain (red; SCOP SCCS: g.41.10.1; 45-119), a trunk domain (green; SCOP SCCS: c.62.1.2; 120-390), a helical domain, (magenta; SCOP SCCS: a.71.2.1; 524-626) and gelsolin-like domain (gold; SCOP SCCS: d.109.2.1; 627-765).

3.1.7. Protein Folding

The process in which an unfolded protein turns into a folded protein in its native state is called protein folding. Physical forces govern this process and all the information needed to find the native state is encoded in the primary sequence [1,18]. Because of the crowded environment in the cell, chaperons are needed to let protein fold without interacting with other proteins in the cell, reducing the risk of incorrect folding [46]. A simplified view is that there is two components to protein folding, a local, and a global. The local is the interaction between amino acids close in sequence. These interactions give rise to secondary structure, which forms early (and fast) in the process. The global interactions are what make the secondary structures come together in a compact way [47]. There are two methods that dominates the discussion of how proteins folds: Model 1: framework-model and related diffusion-diffusion model in which the secondary structure forms and docking of preformed elements. Model 2: hydrophobic collapse drives compaction so that folding takes place in a confined space. Most proteins seems to be a mixture of the two [48]

3.1.8. Protein Complexes

The majority of proteins join together for form protein complexes. Filamentous proteins, actin and the tubulin monomers assemble together to build up the cells cytoskeleton, over 30 [49] distinct proteins (some are present in multiple copies) build up the nuclear pore transporting molecules in and out of the nucleus, the spliceosome and the DNA repair complexes are also examples of large molecular machines built up by many individual proteins. See Figure 2, an RNA Polym-

erase II complex, for a protein complex determined to a resolution of 4.5Å [20] Protein complexes have received much attention in the last couple of years [50,51] and scientists realize that the context is fundamental in understanding an individual proteins function. It has been suggested that individual protein functions can be thought of as words, and protein complexes as sentences [52]. The implications of this is that the meaning of the word, or the function of the protein is context dependent, and hence the same protein, although performing the same function, might have different outcomes depending on the context, e.g. what protein complex the protein is in.

3.1.9. Determining Protein Structure Experimentally

Two technologies, X-ray and NMR, are by far the two most common technologies used to determine protein structure experimentally. In short, to determine the structure using X-ray, a crystal of the protein has to be produced. This crystal is then X-rayed from a large number of angles, and the resulting scatter of the X-rays can be used to calculate the position of the electron density in space. Once the electron density is known, it is possible to fit the atoms of the protein into these densities. X-ray crystallography method can be broken down into 9 separate steps; over-expressing the protein of interest (commonly in bacteria), purifying it, try to make the protein form a crystal, screening out the best crystals, subjecting them to x-rays, collecting the diffraction data as the rays bounce off the protein atoms and using that data to determine the protein structure [53]. In NMR the protein is in solution so no protein crystal is needed. The spin of various atom types can be aligned in a strong magnetic field. The spin of these atoms can

be influences using radio pulses. One can measure as the atom spin returns to being aligned with the magnetic field. It is possible to determine how close two atoms are in either space or how many bonds are between them moving spin energy from atom to atom using radio pulses. Based on this, it is possible to calculate the proteins structure if enough atom-atom distances are known.

3.1.9.1. Structural Genomic Initiative

Large resources have been spent during the last 5 years to determine many protein structures experimentally in a cost-effective way [54]. By scaling up the effort and developing tools and robotics to streamline the process of going from sequence to structure, each structure can be determined less expensively than it is possible today. The average cost within the Protein Structure Initiative program has dropped from 670000 USD to 180000 USD in the 4th year for each structure solved [53]. This number is expected to drop below 100000 USD for bacterial proteins during 2005. Eukaryotic proteins can cost 10 times as much and the success rate is about 1% compared to 10% for prokaryotes [55]. The goal is to produce structures for 10000 USD each. Traditional structure biology groups spend between 250000 and 300000 USD per protein structure. The Protein Structure Initiative will most likely produce 4000-6000 unique protein structures, i.e. less than 30% sequence identical to any protein already in the PDB [56]. This is done under a common name, structural genomics, and there are many centers that focus on this. The first 5 years have focused on technological development, and was distributed on a fairly large number of centers. The next 5 years the focus will shift from technical development to production, and the number of centers

will be decreased. Traditionally, proteins have been studied extensively before their structure was determined and the structure often was of great help understanding the functions on a molecular level. The structural genomic centers goals are to determine all protein structures so that all known protein sequences have a structure within 30% sequence identity. This cut off was selected because the consensus is that all other proteins then can be modeled using homology modeling methods.

3.2. Protein Function

If it difficult to specify what a protein function is [57]. The reason for this is that function can mean different things. A protein function can be the actual mechanics of how the protein catalyzes a reaction, or it can be defined from a cellular perspective, for example this protein is involved in DNA repair. In order to alleviate some difficulties, a number of structured ontologies has been developed, for example gene ontology (GO), see Section 3.2.1. Proteins that are closely related can be assumed to have the same or related functions. Serine proteases are a textbook example of a large number of sequences with high sequence identity that have the same function. Below 40% sequence identity, the conservation of identical protein functions decays [58,59]. Yet, the functions of more distantly related proteins are similar. It is not possible to link protein function to protein structure directly [58], yet the protein function is dependent on the structure. For example, the catalytic residues of an enzyme are sensitive to their relative positions and even small movements of these residues can compromise the enzymes efficiency.

A proteins function can be defined in numerous, more or less overlapping ways. Because of this, it is very complicated, if not impossible, to fully describe a protein's function. A protein's function can be described as its chemical activity, or it can be described as that partners it interacts with, or where in the cell (or outside the cell) the protein linger. Another abstraction is that proteins can be described by how its regulated or modified.

3.2.1. The Gene Ontology

The Gene Ontology (GO) is an hierarchical ontology developed by Ashburner and colleagues [60,61] to address some of the difficulties mentioned in the previous section. GO describe a proteins function from three perspectives or branches, its localization (cellular component), its biochemical function (molecular function) and the proteins context in the cell (biological process). Each branch is organized in a tree like structure with a single root (i.e. the most basic term) with one or more children. A function is a node in this tree like structure and a relation between functions is called an edge. The lower down in the tree-like structure, the more specific the term, and terms with no children are the most specific functions are called leafs. Each branch is a directed acyclic graph (DAG), which means that each term have one or more parents, and is allowed to have multiple children (this differs from a tree, where terms are only allowed a single parent). The connections between the nodes is directed, that is have a direction, and, by definition, there are no cycles in a DAG, which means you can never get back to the point where you started if you follow the direction of the edges. The simple reason for GO to be a DAG is that the nuclear membrane is

both part of the nucleus and the endomembrane system and hence has two parents. Each protein can have one or many GO-terms ascribed to it, and all parents of the GO-term is implied when a GO-term is assigned.

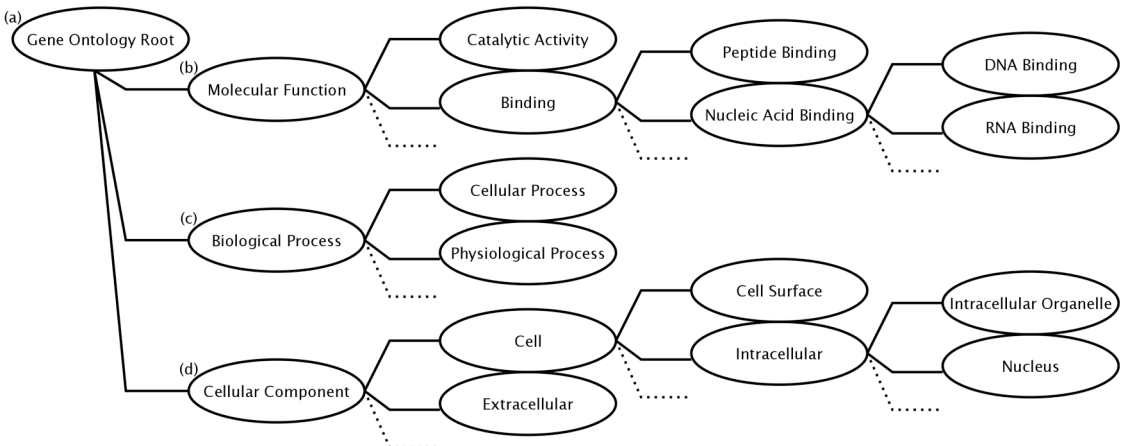


Figure 5. Section of the GO Dag

The gene ontology is a directed acyclic graph, or a DAG. This means that each node in addition to having multiple children also can have multiple parents. Since the edges are directed, and, by definition, the DAG is acyclic, you can never end up at the same node if you are following the direction of the edges.

3.2.2. Protein Interaction, Networks and Systems Biology

It is important to understand the functional organization of cells [62] and how perturbations affect the cell. Every protein interacts with other molecules [63], whether it is other proteins, metabolic molecules or membranes. Protein-protein interactions are of great interest since this is indicative of how the cell is organized. Interactions are not uncommonly mediated through specific structural domains, such as SH2 or SH3 domains [64]. There are many types of interactions, e.g. genetic interactions [65,66], physical interactions [67,68], metabolic networks [69,70,71] or gene regulatory networks [72,73,74]. These networks are

what makes everything work in the cell and small network motifs that can be ascribed a function exists [75], but the emergent properties of the entire network [76] is where knowledge can be gained that is currently not possible to gain by experimental means. Looking at the network topologies, there seems to be modules with very dense interaction networks carrying out some function, with less dense connections to the rest of the network [77]. The field trying to identify and structure all the data is called systems biology [78]. It has to take both genomics and proteomics data in consideration [79].

3.3. Evolution

Evolution is a phenomenon that is central in bioinformatics. Jacob published an article in Science in 1977 where he states that Nature does not invent, nature tinkers [80]. New functions and capabilities gained by a cell is a result of a modification of something already existing and not a product of a de novo invention. Once a functional scaffold has arisen, it is easier to modify that scaffold to perform other related tasks, than to de novo invent a protein scaffold for a new function [80]. Through genetic modifications such as gene duplication [81,82,83] more than one copy of a gene comes into existence. This alleviates the selective pressure to keep the function unchanged and subsequently one or both of the genes can be modified. The result is two similar genes co-existing in a cell performing similar functions. Both retain structural, functional and sequential characteristics from the ancestral gene. Genes and their protein products are said to be homologs if they share a common ancestor. As more mutations changes each gene their common ancestry becomes more difficult to detect. As a rule of

thumb, it is possible to detect homologs using sequence alone down to about 25% sequence identity, e.g. 25% of all the amino acids in the two sequences are identical. By definition, all proteins that belong to a SCOP (see below) superfamily are homologs. Rost showed that the sequence identity of homologs within SCOP superfamilies peak around 9% [84,85], which is below the threshold of sequence detection. Hence, structure is needed to identify all members of superfamilies.

So how do we know that proteins with a certain sequence identity did not get created independently and just look similar by chance? There are 10^{200} (20^{150}) possible sequences for a protein of 150 amino acids [86]. About 10^{38} of these are less than 20% identical. Not all sequences will fold to a stable tertiary structure in an aqueous environment. It has been estimated that about 1 out of a billion will fold. That gives us 10^{29} protein sequences of less than 20% sequence identity that will fold into a stable tertiary structure. Yet, it seems possible to organize 75% of the close to 40 million sequences into 8000 protein family. These numbers indicate that, in fact, most sequences have a great number of homologs.

3.3.1. Homologous Proteins

Two proteins are homologs if they are related by divergence from a common ancestor [87]. There are two kind of homologs; paralogous found in the same organism, for example hemoglobin and myoglobin (Figure 6a,c), and orthologous performing the same task but in a different species - for example human hemoglobin and pig hemoglobin (Figure 6a,b). All three share one common ancestor, but only the orthologous have the exact same function and role in the cell. The

hemoglobin distributes oxygen by carrying it in the blood stream whereas the paralogous myoglobin serve as temporary oxygen storage in the muscles. Yet, they all perform similar tasks and their common ancestry can be detected by sequential and structural similarity (Figure 6). Many tools have been developed to identify homologs, such as BLAST [88]. Genes are at times reorganized; two genes might fuse, there might be an insertion of a domain, or parts get deleted. This results in two homologous proteins only being homologous over some common part, and not over the full sequence [89].

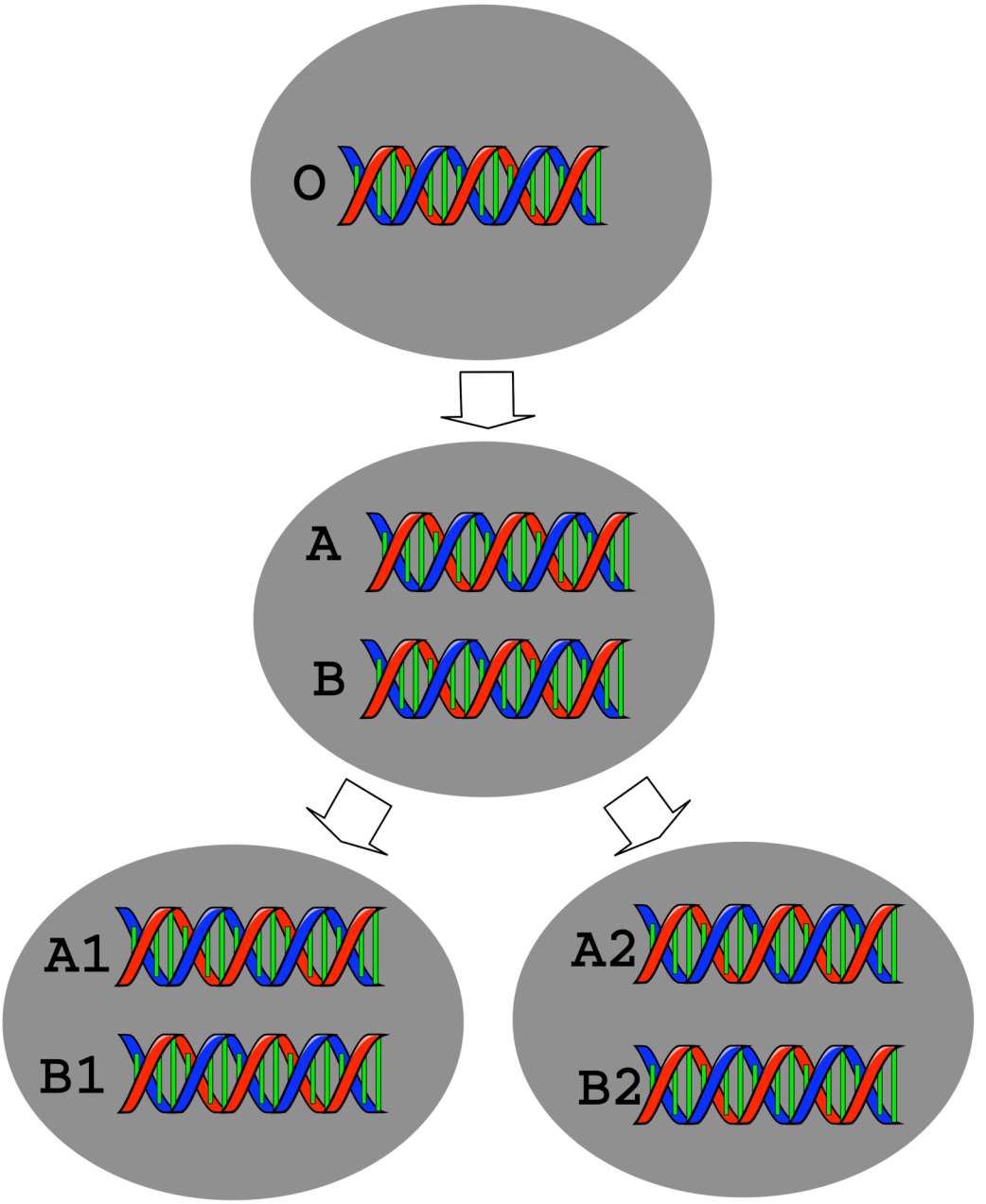


Figure 6. Orthologs vs Paralogs.

An ancestral gene, O, is duplicated into A and B through a gene duplication event. A and B are homologs, but also paralogs since they have a common ancestor, and reside in the same organism. Through speciation, A now becomes two orthologs A1 and A2 and the same holds for B. A1 and B1 are paralogs, which is true for A2 and B2, and A1,B1,A2 and B2 are all homologs [90].

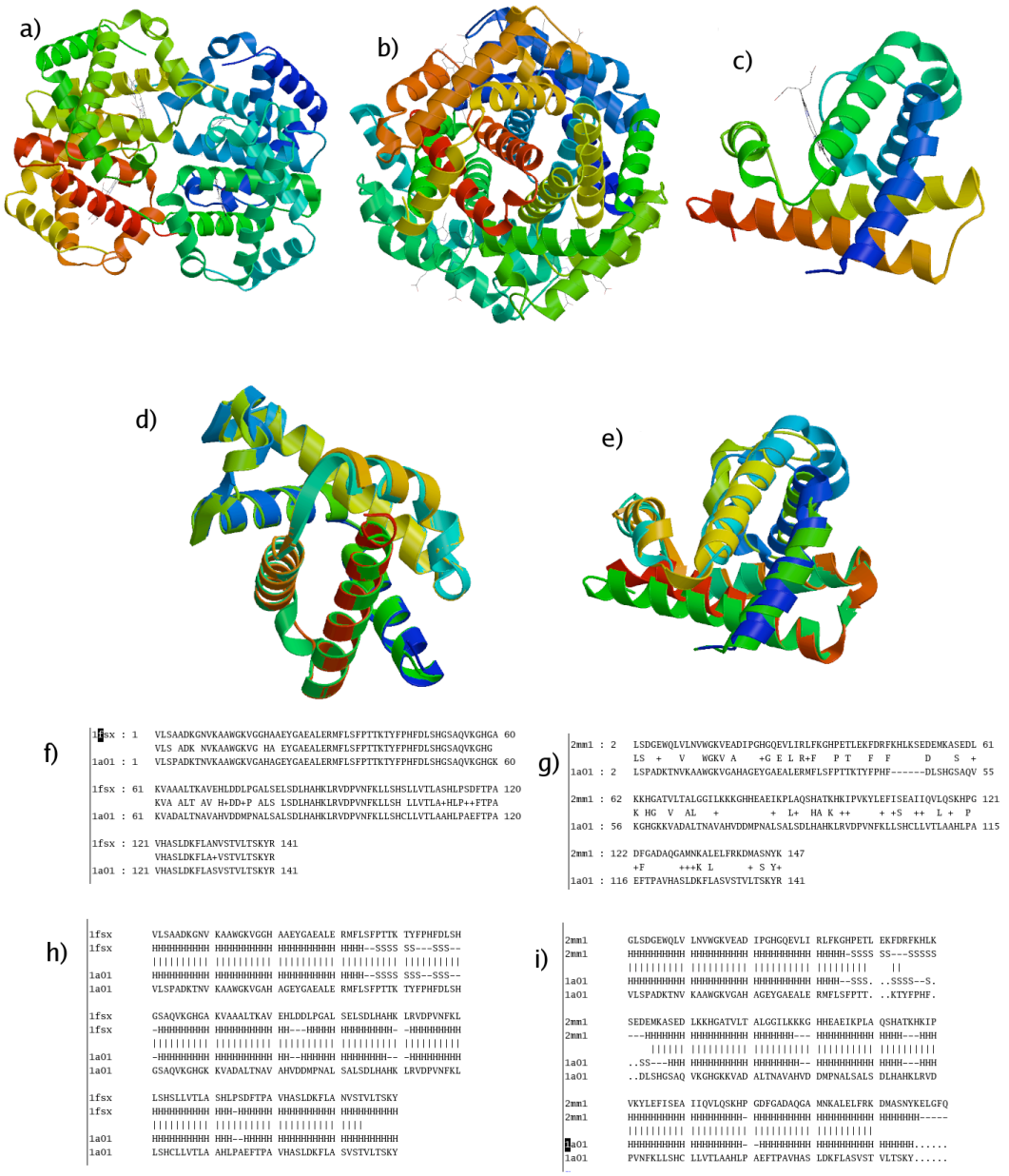
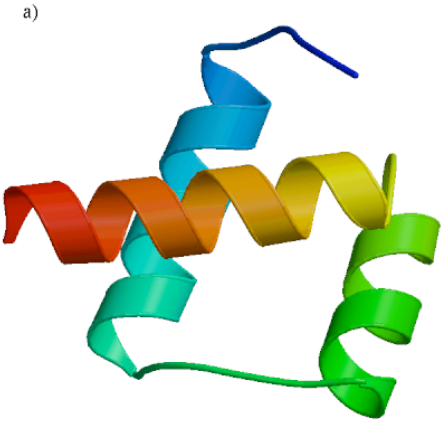


Figure 7. Homologous Proteins.

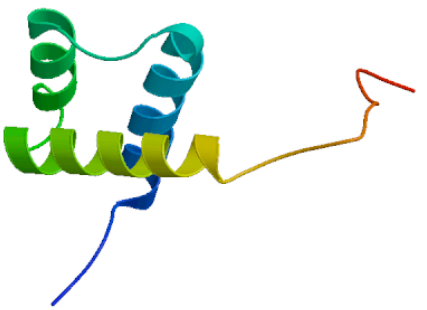
The three homologs, human hemoglobin (a), pig hemoglobin (b) and human myoglobin (c) are compared. Human hemoglobin and human myoglobin are paralogous and share a common ancestor gene and are found in the same organism.

Human hemoglobin and pig hemoglobin are orthologous and share a common ancestor but are found in different organisms. The orthologous proteins share the same function and play the same role in the organism whereas the paralogous have very similar function and play slightly different roles in the organism. Structural super-impositions of the proteins show that they are all very similar, see (d,e) in Figure 7. (f,g) sequence alignments using only sequence and (h,i) sequence alignments using structure. Detecting that two proteins are homologs infers that their structure and function are identical or related (see Figure 7) Hence, if one of the proteins is uncharacterized and the other performs a known function and/or has a known structure, this information can be used to anticipate structure and function for the uncharacterized protein. Detection of homologs for a given protein is not trivial. Finding homologs with a sequence identity of 25-30% and above for a protein of 100 amino acids can be done using only sequence, but below this threshold, in the so-called twilight zone [84], the false positive rate becomes unacceptable. Structure is more conserved than sequence [84,91,92] and homologs in the twilight zone can be detected if the structures of the two homologs are known. Yeast 2 and the drosophila engrailed protein are homologs but share 17 of 60 amino acids which is in the twilight zone [58] (the sequence identity is 28% which is in the twilight zone for shorter proteins). Yet, their function and structure are almost identical, see Figure 8

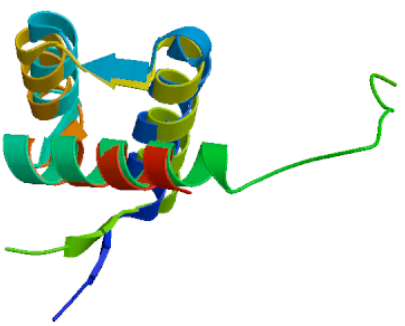
a)



b)



c)



d)

```
1le8b: 26 ENPYLDTKGLENLMKNTSLSRIQIKNWVAARRAK 59
          EN YL  +  + L     L+  QIK W   +RAK
1enh:  20 ENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAK 53

1le8b: 74 SGEPLAKKK 82
          S E LA+ K
1enh:  7  SSEQLARLK 15

~
~
```

e)

```
            ********* *******    ********** ********** **********
1le8b       .RGHRFTKEN VRILESWFAK NIENPYLDTK GLENLMKNTS LSRIQIKNWV
1le8b       .SSSSSS-HH HHHHHHHHHH H---SSSS-H HHHHHHH--- SS-HHHHHH
            ||||||||| ||||||    |||||||||| |||||||||| ||||||||||
1enh        SSSSSS---H HHHHHHH... HH--SSSS-- HHHHHHH--- HHHHHHHHH
1enh        RPRTAFSSEQ LARLKRE... FNENRYLTER RRQQLSSELG LNEAQIKIWF
             ********* *******    ********** ********** **********


            ******
1le8b       AARRAKEKTI TIAPELADLL SGEP
1le8b       HHHHHHHHHH HHHH-HHHHH HHHH

1enh        HHHHHH.... .......... ....
1enh        QNKRAK.... .......... ....
            ******
~
~
```

Figure 8. Identical Structure, Divergent Sequences

17 of 60 amino acids are identical between the D.Melanogester engrailed protein (a) and the S.cerevisiae alpha2 protein (b), yet their structures and functions are identical. This relationship cannot be reliably detected using only sequence. Structural information is needed. The structures are superimposed in (c) and the sequences are aligned using only sequence (d) and using structures (e).

It is important to realize that there are always exceptions. Homologous do not always share function, and similar functions and/or similar structures do not infer homology. The most classical example of close homologs not sharing function is the alpha-lactalbumin and lysozyme [93]. These proteins have a 50% sequence identity, but lysozyme is an enzyme whereas alpha-lactalbumin is a non-enzymatic blood constituent. Homology and structural analogy (i.e. structures resembling each other) means different things. Analogy can be a product of convergence, and homologs per definition do not have to be [58]. In some cases there is no detectable sequence similarity between proteins of similar fold. This suggests a convergent evolution [94].

3.3.2. Protein Families

The logical extension when finding homologous proteins is to organize them in families. This has been done a number of times employing different technologies. One of the more popular, Pfam, [43,44] is based on a hidden Markov model (HMM). A number of seed sequences are identified as belonging to a single sequence. The HMM is constructed from the seed sequences, and used to search all available sequences. In August 2005, there was 7973 protein families, covering 75% percent of all sequences in SwissProt.

3.3.3. Amino Acid Conservation

Talking a close look on these protein families reveals that some amino acids seem to be completely conserved in the families, even though the over all sequence identity is low. These conserved amino acids can in a simplified view be classified into two classes, the functionally conserved and the structurally conserved. The functionally conserved amino acids are amino acids involved in, for example, catalytic reactions or protein interactions that might disrupt the function if mutated. The structurally conserved amino acids have important structural impact. A proline, for example, might be conserved in a tight turn, and since prolines backbone configuration differs from all other amino acids, it cannot be replaced without disrupting the turn. The protein structure might also be disrupted if a small amino acid is substituted for a large one.

3.3.4. Structural Similarity

Before discussing the different prediction strategies available, a more general problem has to be pointed out - what does structurally equivalent mean? If we were able to predict the tertiary structure how do we know that the prediction was correct? Structures are traditionally compared using a metric called root mean square deviation (RMSD), that is, the root of the sum of the squared distance between for example alpha carbons of equivalent amino acids. When structures differ by a mean deviation less than 2Å, they are considered structurally equivalent. RMSD has two fundamental flaws of being dependent on length and sensitive to outliers, that is, if you match two structures perfectly, but the C-terminal alpha helix is packed on the wrong side of the protein, the RMSD will

be high since the parts that are different dominate it. This constitutes a problem in molecular modeling since we are not likely to get near the native structure for RMSD to be a good metric. Which amino acids are equivalent in the two structures is also a difficult problem as the structures become more diverse. There are no clear solutions to these problems. Structurally equivalence is rare between proteins, even within the same protein family. There are a number of algorithms developed to compare protein structures. They are based on different technologies. Two of the most common ones are DALI [95] and VAST [96]. An algorithm, MAMMOTH [97], was developed in 2001 to match structures of lower quality than DALI and VAST are designed to do. See Figure 9 for an illustration. MaxSub alleviate the problems with RMSD in that only aligned residues are considered [98]. MaxSub is normalized between 0 and 1, has no length dependencies, and is not dominated by outliers and considers how many amino acids that can optimally be aligned under a given distance cutoff.

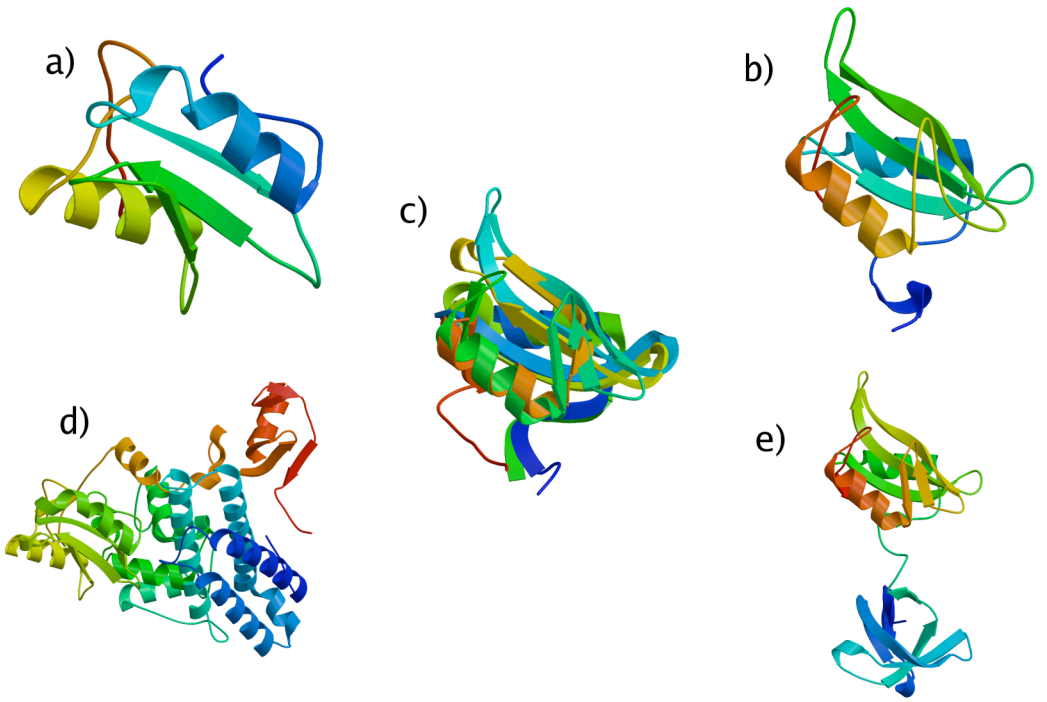


Figure 9. Structure Alignments

This image illustrates two almost structurally identical domains (SH2 domains) found in proteins that are otherwise related. a) This is the A3 domain of PDBID 1FBV, a SH2 domain b) The A2 domain of PDBID 1G83 is also a SH2 domain c) The SH2 domains in a) and b) are overlaid in c) and are surprisingly similar. d) This is the full structure of PDBID 1FBV. This is a 4-domain protein; the SH2 (SCOP SCCS d.93.1.1) domain is located between amino acid 264 and 355. The three other domains are N-cbl (SCOP SCCS a.48.1.1; 47-177), EF-hand (SCOP SCCS a.39.1.7; 178-263 and RING/U-box (SCOP SCCS g.44.1.1; 356-434). e) This is the full structure of PDBID 1G83. The first part is a SH3 domain (SCOP SCCS b.34.2.1) and the second part is a SH2 domain (SCOP SCCS d.93.1.1)

3.3.5. Protein Structure Classifications

The number of potential protein structures is enormous [86] but many of these potential protein structures will resemble each other. By grouping proteins with structures that resembles each other in a tree structure, where groups closer to the leafs are more similar, and closer to the root are less similar, it is possible to reduce the complexity. This is difficult because it is complicated to define a single

metric that describes structural similarity. Murzin and colleagues [99,100] developed a classification system, (SCOP; Structural classification of Proteins) in which most the known structures are classified. According to their definition it has been estimated that there are about 1000 significant folds [101,102,103] and roughly 700 of these are known (as of SCOP version 1.63). The SCOP classification classifies protein structures according to a hierarchical 4 level tree. The levels are, (1) Class, (2), Fold (3) Superfamily (4) Family, and each family contains a number of protein domains. Different parts of a protein structure can have distinct classifications, that is, a multi-domain protein can have numerous classifications. Hence each SCOP classification is tied to a protein structure, a polypeptide chain and sometimes part of a chain. SCOP is a manually curated database and each new structure deposited in the Protein Data Bank [56] is classified by in the SCOP hierarchy by Murzin and colleagues [99], see Figure 10. The current SCOP database, version 1.69, has 25973 protein structures, 70859 domains divided into 945 folds, 1539 superfamilies and 2845 protein families. There are a number of classes in SCOP, 4 of which are more prominent than the others. All alpha proteins consist of mostly alpha helices and beta proteins contain only beta sheets. The alpha+beta proteins contain both alpha helices and beta sheets but the different elements are spatially grouped with secondary structure elements of similar kind. The last group, the alpha/beta group contains alpha helices and beta sheets mixed together. The SCOP classification is used to create the Astral database [104], a domain based database where sequences and structures for all SCOP domains can be downloaded and analyzed. There are more domain databases such as FSSP [105] and CATH [106] but they will not be discussed in this

thesis, see [107].

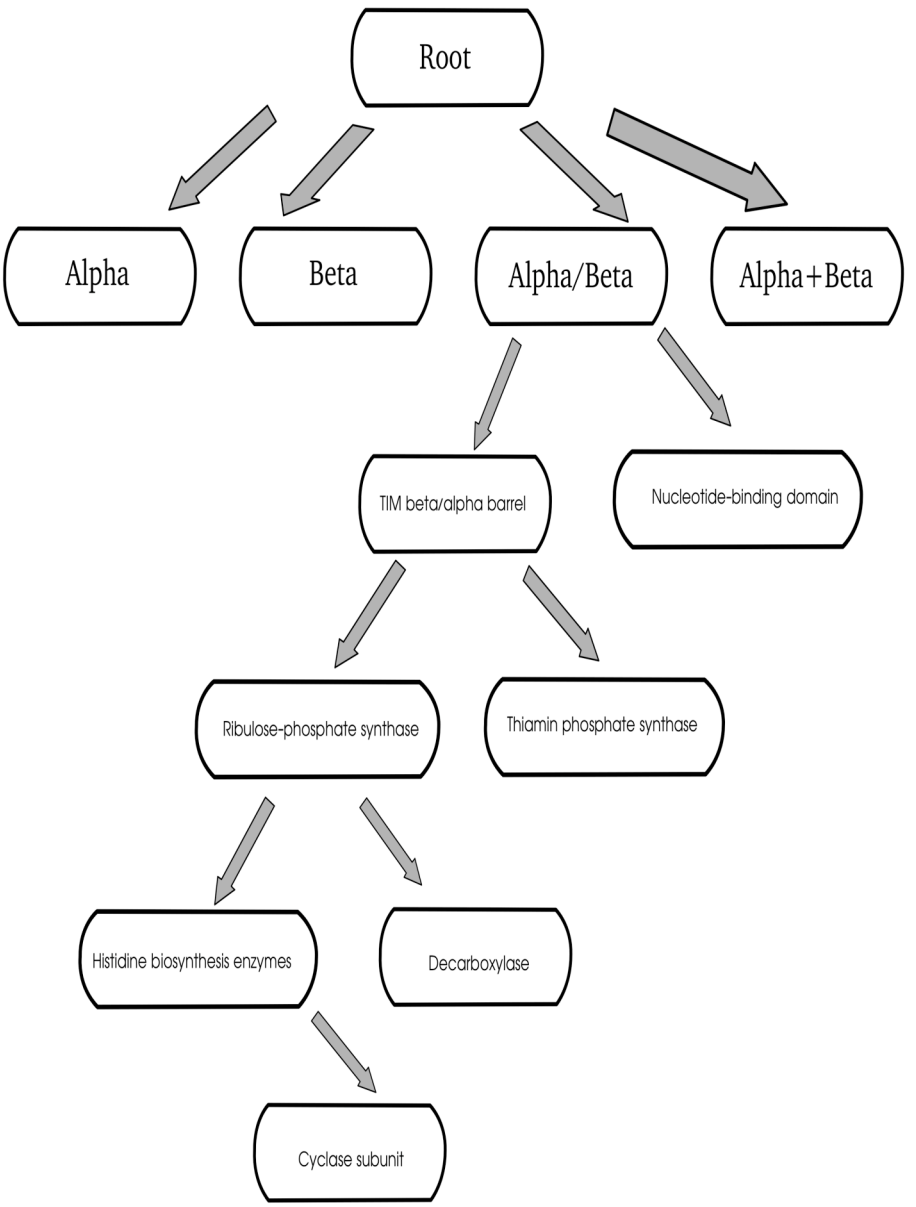


Figure 10. The SCOP classification hierarchy

The SCOP hierarchy classifies protein structures according to their secondary structure composition and their structure.

10 folds occur very frequently, and are referred to as superfolds [108]. Examples are the globin fold, UpDown, TIM barrel and jelly roll. Evidence suggests that these superfolds have been created many times through history since there are no detectable sequence similarities. It has been estimated that 80% of all protein sequences belong to 400 folds [109]. Some estimate that there will be at most 10000 folds [86], most of them so called unifolds, i.e. have only a single member [109].

3.4. Predict Protein Structure

The goal of protein structure prediction is to determine a proteins structure given only the primary sequence. The number of atoms and the number of theoretical conformations of even the smallest peptide is astronomical making the structure prediction difficult. Structure prediction can be divided up into a number of disciplines; secondary structure prediction, domain prediction, homology modeling, fold recognition and ab initio protein structure prediction. Each discipline is described in more detail below but first CASP (the community wide assessment strategy) will be introduced.

3.4.1. CASP and LIVEBENCH

CASP, Critical Assessment of Techniques for Protein Structure Prediction, is a semi-annual event to assess advances in the field of protein structure prediction. Sequences of solved but unpublished protein structures are sent to the participants of CASP. The participants make their best structure prediction and return it to the organizers before the structure is published. The structures are carefully

analyzed and assessed by the organizers [110,111,112]. The query sequences or targets in CASP are divided up into three categories, homology modeling, fold recognition modeling and new fold/ab initio modeling. Rosetta ab initio is applied to the new fold category and the difficult fold recognition category. The number of targets in CASP is not enough to statistically evaluate the progress in the field, and to alleviate this problem a continually ongoing assessment was developed called Livebench. Livebench is an attempt to do the same thing as CASP but in an automated and continuous manner but instead of withholding the structures, the participants are trusted not to use that information. Another difference between Livebench and CASP is that Livebench participants are fully automated whereas there are human interventions in the CASP predictions [113,114].

3.4.2. Secondary Structure Prediction

The objective in secondary structure prediction is to predict whether a given amino acid is part of a helix or a sheet or neither of them. There are many algorithms available and they are based on a variety of different concepts. The best algorithms available today reach an accuracy of 77% and are based on artificial neural nets [115]. At best, secondary structure prediction algorithms are expected to classify up to 80% of the amino acids correctly because secondary structure is not only a local phenomenon, but also depends on long-range interactions [116].

3.4.3. Predict Structural Domains

To accurately predict structural domains and the linking regions between these domains is a very important step in all structural characterization [40]. This is not

trivial even if the proteins structure is known. There are a number of methods developed to parse a structure into structural domains [117]. To parse a protein into structural domains given only the sequences is more difficult, and many different approaches have been attempted [118]. Marsden et al assign putative domains using secondary structure [119] and others have used amino acid composition [120], domain-size distributions [121] or amino acid covariance in the multiple sequence alignments [122]. More computationally expensive algorithms, such as SnapDragon and RosettaDom folds the putative multi-domain protein, and assigns domains using a structure based method [117] and then looks for consensus assignments over many attempts [123]. We use an in-house developed algorithm, Ginzu, [124,125] which in an iterative fashion identifies domains starting with highly confident methods and subsequently applies less confident methods with higher sensitivity. In the first steps, the method is similar to CHOP [126], in that both methods identify homologs with known structure. Since the protein structures almost without exception is over a structural domain this information is very reliable. After this step, Ginzu resorts to more sensitive methods, the fold recognition methods, is applied which again identifies protein structures belonging to the same superfamily. Both methods rely in identifying protein family from Pfam, which also corresponds to structural domains. Ginzu lastly resorts to trying to parse domain information from the Multiple Sequence Alignment, a method that is not very reliable, but is very sensitive.

3.4.4. Homology Modeling

If the sequence for which the structure is to be predicted has a close homolog

(40% sequence identity) of which the structure is known, it is possible to use that structure as a template and drape the structure with the sequence of the protein of known structure. The structure with the new sequence now has to be relaxed and loops rebuilt. Sali and colleagues have automated this process and is serving a large database with homology models [127]. For a review, see [128].

3.4.5. Fold Recognition

A majority of the algorithms presented below are based on secondary structure prediction algorithms. Fold recognition or threading as it is also referred to, becomes increasingly powerful as more structures are solved, (for a review see [129]). There are 10^{29} different amino acid sequences of length 150 of less than 20% sequence identity that will fold into a stable tertiary structure and it has been estimated that the total number of highly populated folds do not exceed 1000 [101]. Hence, it is easier to solve the inverse folding problem, that is, identifying the structure fold that an unknown protein sequences most likely belong to. By draping the query sequence onto all known folds and estimating the compatibility between sequence and the structure, one can estimate the probability of that sequence having that structure. Once a compatible structure is found, using the homologous proteins structure as a template, one can 'drape' that structure with the query sequence and hence come up with a structure for the query protein. Vast arrays of programs are available, based on very different technologies. Some of the programs use 3D profiles where others use only secondary structure prediction and compare to secondary structures of known proteins, such as 3DPSSM [130] and BASIC [131]. Other so called consensus servers or meta

servers are based on a number of other fold recognition servers and tries to find a consensus between them [132]. The PCONS [133] software is a neural net-based consensus server using 5 or 6 (depending on the version of the PCONS server) other servers as input, and produces a consensus structure. This approach works well since the servers PCONS is based on are somewhat orthogonal in their approach to each other. Many genomes have been structurally characterized using homology modeling and fold recognition [134], including human [135].

3.4.6. Ab Initio Protein Structure Prediction

Ab initio protein structure prediction modeling without a template, i.e. a homolog from which you can derive the over all fold. The objective is simple: Given a sequence, predict the tertiary structure. There are two main components of the problem, (1) the energy potential and (2) the search space (the number of possible conformations of the polypeptide). The energy potential guides the search towards the global energy minimum and without an accurate potential this is not possible. The true energy potential is very complicated with numerous different components at play such as solvation, electrostatic interactions, van der Waals interactions, bond lengths and bond angles. Two categories of energy potentials are used in practice, the molecular mechanics potential model and the statistically derived potentials derived from experimental structures [136,137]. Much of the details have to be sacrificed to keep computer models simple and fast and the energy potential used to evaluate the fitness of the model have be modified accordingly. In molecular dynamics simulations one integrates Newton's equation of motion for the polypeptide chain using a physically reasonable

potential function. This approach is computationally expensive (the search problem is enormous) and it is still unclear if the energy function is accurate enough to guide the search. An alternative approach is to reduce the complexity of the search and the energy potential by simplifying the model. Two main categories exist, the lattice models and the non-lattice models. The lattice models have difficulties with describing a protein structure accurately. The non-lattice models have had more success than the other approaches. Common to most of them is that they reduce complexity by fixing bond-lengths and fixing rotamers (the number of possible side-chain conformations) and some use only a few phi and psi angles or a few phi-psi angle combinations. Some use centroids (i.e. represent an amino acids as a single point) instead of the full atom set. The prospects for ab initio modeling is discussed in [138] and [139]. For a review, see [140].

3.4.6.1. Rosetta

Rosetta is a software package developed by David Baker and colleagues. It is written in C++, an object oriented programming language, and it consisted of 290000 lines of code with 275 command line options (July 2005). The Rosetta algorithm [111,112,141,142] is an ab initio protein structure prediction based on a knowledge based potential and that reduces the search space by using local conformations of short amino acid fragments from the PDB [56]. The initial search is done in centroid mode and an optional subsequent model refinement is done in a full atom mode. The Rosetta algorithm was developed under the assumption that a short sequence of amino acids have a finite number of conformations and that these conformations are represented in the PDB. Rosetta is a

Monte Carlo simulated annealing procedure and have had great success in CASP, see Paper I. One can say that Rosetta bridges the gap between sequence and structure, but this bridge goes two ways, that is, it is possible to go from structure to sequence. This is commonly referred to as protein design and two landmark projects have been completed using Rosetta in this realm over the last couple of years. TOP7, designed by Kuhlman and colleagues, is more stable than similar proteins in its size class [143]. This stability leads to a number of properties not readily seen in nature, and has opened a new research field where scientists for the first time can study proteins not optimized for fast and simple folding but just for stability. The other major design project, work by Korkegian and colleagues, was to stabilize an enzyme [144]. A number of key amino acids were replaced and it resulted in a more stable protein that also increased survival of re-engineered bacteria at higher temperatures than bacteria with the wild-type version of the protein.

3.5. Predict Protein Function

Predicting a proteins function from experiments or other properties is difficult. Not only is the notion of function complicated (see above) but the definition when two functions are similar becomes complicated to quantify. Looking at the GO-DAG reveals that functions are described in general terms, (close to the root), and in great detail (close to the leafs), and that the distance between nodes doesn't have much of a correlation of how similar or dissimilar two functions are. In general, there are two ways of looking at similarity/dissimilarity between functional assignments, one is based on the Enzyme classification and the other

is based on GO. The Enzyme Classification is a true tree with four levels, which depend on kind of reaction and type of ligand/product. Belonging to the same branch at some level indicates similar/same function. In the GO DAG, it is possible to trace the two functions to the root, and look at the first common node. Once a common node is found it is possible to calculate some kind of similarity score depending on the total number of gene products in that common node and the fraction of these annotated with one or the other function of interest. Both systems have advantages and disadvantages, most notably that the edges does not correspond to any closeness, i.e. a single edge different might reflect quite different functions or close. The problems of assessing closeness between functions carries over when evaluating success or failure for structure prediction and when analyzing how distant two homologs can be and still retain a similar/identical function.

Bartlett and colleagues did an careful analysis of the relationship between conserved function and sequence identity between two proteins and came they found that above 40% sequence identity, the degree of functional conservation is high [59]. Wilson et al. investigated how much information that can be transferred between homologous protein domains of known structure. [145]. They found that identical protein functions are conserved down to about 40% sequence identity, and broad functional classes conserved down to about 25% sequence identity. On the other hand, Hegyi and Gerstein [146] did a thorough investigation if a stable link between structure and function could be established and they concluded that sometimes such a link could be established. There are some major hurdles that have to be overcome before the relation between structure and func-

tion is truly understood. Functional annotation of unknown genes has been appointed one of the best applications for protein structure prediction. In the majority of cases, a stable link between a structure and a function can be established, and it has been shown that ab initio models have sufficient qualities to identify this function [147].

Many high-throughput technologies aim to place a protein in a context and thereby learn something about the proteins function [50,51,148,149]. These methods do not however elucidate the molecular activity per se. As of today, there are not many high-throughput technologies that target the molecular activity, and some that do are labor-intensive and hence also expensive. For example, protein complex purification followed by MS identification experiments identifies protein complexes. Proteins that belong to the same protein complex are likely to participate in the same biological process and be found in the same cellular compartment. The uncharacterized proteins can be annotated if the biological process or cellular compartment are known for some of the constituents of the protein complex Protein structure on the other hand give hints about the molecular activity, and could potentially be a great complement to other technologies. It is possible to infer function from blind ab initio protein structure prediction [150]. The relation between structure and function is not easily deconvoluted [151]. As the structural genomics centers found out, knowing the structure does not always easily translate to knowing its function Two technologies aim at identifying interacting proteins and assign protein function by guilt by association. These two technologies, Tandem affinity purification (TAG) tag Multidimensional protein identification technology (MudPIT [152]) and yeast

two-hybrid (Y2H) have been applied in genome-wide assays [50,51,148,149], fluorescent microscopy can identify what cellular compartment a protein is found in and ab initio protein structure prediction can give insights in the molecular function of a protein. Transcription factors have been studied on a genome wide scale in yeast by cross-linking myc-tagged transcription factors to the DNA, purifying and sequencing the DNA. 800 genes vary in a periodic fashion during the yeast cell cycle. [153] Proteins can be assigned putative function by co-regulated gene expression [154].

So we know there are proteins with known structure and that we with some accuracy can predict the structure of proteins without close homologs of known structure. We also know that there is a strong relationship between structure and function. Is it possible to use this information and the protein structure prediction tools to more accurately annotate proteins with function or structure?

4. Present Investigation

4.1. Objectives

The objective of this thesis is to use ab initio protein structure prediction to characterize proteins, both functionally and structurally, on a genome-wide scale. I investigated how to integrate information known about proteins, both derived from proteomics experiments and downloaded from on-line databases, to increase the confidence of the structural classification, and assess the quality of the classification by a simple probability measure.

4.2. Method development

4.2.1. Rosetta in CASP6 (Paper I)

This paper illustrates how well Rosetta, the central software, performs in a true blind test. Rosettas performance is fundamental for any of the results in this thesis to be valid.

CASP is a true blind test and the results are indicative what results to expect when applying the methodology on a large scale assuming that the target selection in CASP is unbiased. The structural domain architecture was predicted for all targets, and the longer proteins were parsed in to two or more domains using two domain-parsing technologies, Ginzu and RosettaDom. RosettaDom and Ginzu predict structural domains with 80% accuracy, comparable to human experts and better than other automated servers as assessed in CASP6 [125]. Ro-

setta is one of the most accurate ab initio protein structure prediction algorithms, and the over all performance on smaller structural domains is good. Rosettas performance drops when the domain parse was incorrect, and hence the over all success is dependent on more than Rosetta itself. In conclusion, this experiment reaffirmed that the Rosetta ab initio protocol does in fact predict protein structures well, and the domain parsing algorithms we use, that are essential in the larger studies, are accurate. A few highlights are worth mentioning. T0198 is an alpha-helical protein of close to 200 amino acids, which is at the upper bound of what Rosetta can predict. 90% of the alpha carbon atoms of the best structure prediction are within 4Å when optimally aligned. For target T0281, a 70 amino acid protein, we predicted the structure of a large number of homologs in order to increase the low-resolution sampling. After clustering, a number of decoys were selected for full-atom minimization, a method where all atoms are explicitly represented. This is computationally expensive, but the full atom mode has a more physical energy function, and it seems possible to identify the native topologies using the energy function alone. The core of a protein can be described as a jigsaw puzzle where atoms from different amino acid residues intermingle closely giving favorable van der Waal interactions. This jigsaw puzzle is very difficult to predict because even a small deviation makes the atoms overlap and raises the energy dramatically. The use of multiple homologs in the low-resolution mode and subsequent replacement of the target sequence with loop building gives a more diverse backbone sampling, and increases the likelihood of the minimization step predicting the protein core accurately. Target T0281 was predicted as 1.59Å, the most accurate prediction in CASP ever.

4.2.2. 100 uncharacterized essential Open Reading Frames (Paper II)

This paper is central in this thesis because it describes a proof-of-principle project where we show that integration of protein structure prediction data indeed is useful when studying proteins of which nothing is known. Every protein in this study was manually evaluated.

The project focused on 100 putative uncharacterized essential open reading frame (ORFs) products from Saccharomyces cerevisiae, Baker's yeast. Three proteomics technologies were employed on these 100 ORFs in addition to the domain parsing and ab initio protein structure prediction. The experimental techniques were tandem affinity protein complex purification and subsequent MS analysis (MudPIT) [152], yeast two hybrid (Y2H) and fluorescent microscopy. The MudPIT analysis reveals protein complexes, Y2H identify interacting partners and fluorescent microscopy localized the protein to a cellular compartment [155]. The goal of the experiment was to assign GO-terms from the three branches, molecular function, biological process and cellular compartment, to each of the 100 ORFs. Biological process and in some cases, cellular component can be assigned using MudPIT and Y2H data, cellular component is obtained by fluorescent microscopy and molecular function can be deduced from the protein structure prediction data. All technologies except for the fluorescent localization are noisy, that is, reports many false positives. We utilized the fact that GO-terms are more likely to co-occur with some GO-terms than others, e.g. the biological process term DNA repair is more likely to co-occur with the cellular

compartment term nucleus then the cellular compartment term golgi since DNA is primarily found in the nucleus in eukaryotic organisms. We were able to filter out unlikely GO-term assignments by integrating data from all four technologies. GO-terms could be assigned to 77% of the 100 ORFs and 19 of these could be annotated with gene ontology terms from all three GO branches.

4.3. The DDB information management system (Paper III and Paper IV)

Paper III and Paper IV are in many ways the most important papers in this thesis. These papers describe a software package, DDB, I have developed since the summer of 2001 and the other papers in this thesis have all been analyzed using this software package. The three objectives of this database is to, first store and organize data related to proteins, second, to integrate this data with mainly large on-line data repositories such at NCBI's Entrez system, and third analyze the data. These objectives are difficult to reach in a generic and comprehensive way, and as a result DDB is not developed to be distributed and used by large numbers of non-expert users, but instead, to serve as a framework and a set of standards for an expert developer to fulfill the objectives in a specific project where the incoming data is known, the databases to integrate are determined and the analysis can be developed after the needs of the experimentalist.

This software package was used in Paper I to produce and organize the all the protein targets and their homologs and all the Rosetta prediction generated for them. In Paper II, the system served as bookkeeping software to keep track of the

progress on the selected 100 protein targets and to generate and analyze the protein structure prediction data. The software was utilized in Paper V from downloading the genomes from NCBI to generate statistical models and in the final phase, to generate data structures underlying the public websites. A task of the size of predicting structures for the yeast genome is a too large for a single computer. As a matter of fact, it would talk a single computer one thousand years to complete the yeast genome. Therefore, the DDB system has 4 major sub systems, the web server (front end), a back end server that serves as the main controller together with the database server. The fourth system is a computational unit that can execute large calculations. The core of the software is the data model, served by a relational database by the database server. This database is currently running on a single computer in each of the implementations, see Figure 11, but this can be extended to a whole cluster of computers should the need for capacity increase. The number of tables in this database exceeds 180, and the largest implementation, the hddb system, employs a database server, a back-end server, a web server a file server and about 150 dual-processor computational nodes (computers). The web server displays the data from the database to the user via a web interface. The bulk of the computation is, obviously, done by the computational system, which, in one of the implementations, are done using the world community grid, a service provided by IBM and as of the summer of 2005 involves 130000 computers utilized only when the owner of the machine is not using it. The world community grid was considered one of the 20 largest computers as of the summer of 2005, if measured by operations per second at top-capacity.

Paper III describes the 2DE components of the DDB system. In Paper IV, we report that DDB now handles quantitative data from MS. It can read and analyze MzXML files [156] and hence interfaces with a large number of mass spectrometers. In addition, it readily handles protein domains, protein structure prediction and a number of sequence based prediction software, including Pfam and prosite. Unfortunately, we had to abandon the database-only approach and develop a mixed model, keeping indexes and summary data in the database and the large files on the file system. This is mainly a limitation of the MySQL software, and to some extent a limit of the current hardware. As of the summer of 2005 there are three implementations, the proteomics version, 2ddb, the yeast version, bddb and the human proteome folding project version, hddb. There is a public site (http:///wwww.2ddb.org), derived from the 2ddb implementation, demonstrating some of the published aspects of the software.

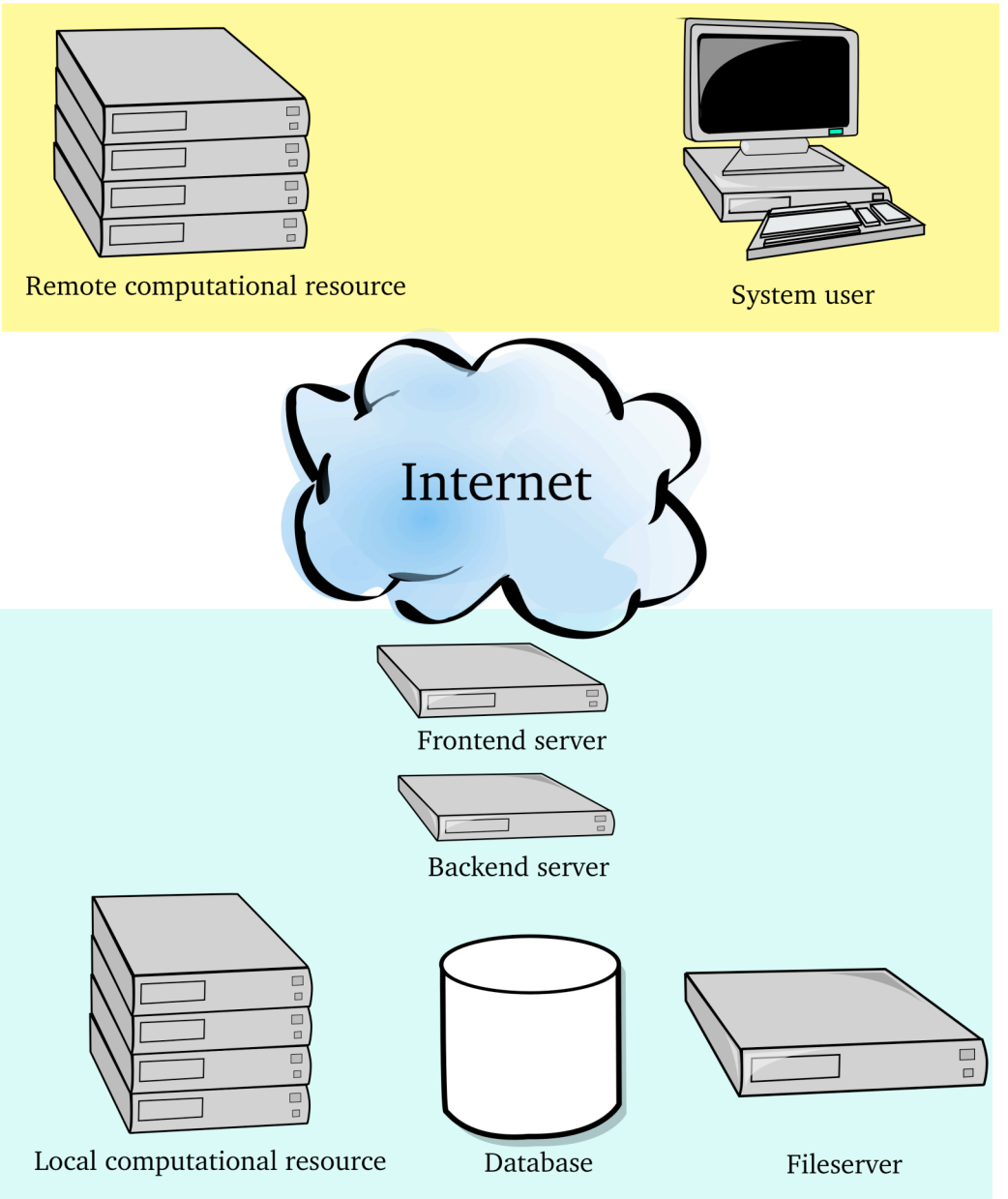


Figure 11. Organization of the DDB System

The DDB system is composed of a front-end, a back-end and a database server. Large computations can be farmed out to condor-based computer facilities or grid-computers. Internally, DDB uses close to 180 relational tables describing proteins and peptides all the way up to describing meta-information about computational resources.

4.4. Structural prediction of yeast (Paper V)

As a continuation of Paper II, we applied the same methodology to a number of genomes, ranging from E.coli to Human. Several hundred hours were poured into the structure prediction part in Paper II, especially in the validation of protein structure matches to mammoth and the subsequent conversion to functional annotations by integrating the data from the other technologies. In Paper V we automated and formalized the steps taken manually in Paper II, and applied the technology on yeast. Yeast has roughly 6200 ORFs with an average length of about 450 amino acids. The average domain is about 175 amino acids long. Hence, it is expected that, on average, yeast proteins have two or three domains. The main challenges in this project was first, data production - how to scale the size of the project sixty times to include all of yeast, secondly, how to deal with large amounts of complex data and last, how to automate the manual steps in Paper II. The first section was solved by collaboration with the ISB and IBM, providing several thousands of computers performing the actual protein structure prediction simulations. Our resources were spend in the preprocessing and post processing. The data management was solved using the DDB software (see Paper III and IV). The absolute bulk of the work put into this project involved the automation procedures. We had to replace the high throughput experiments with an online database, the GO database. We used a Bayesian approach to combine the data, which turned out to work. Our most confident predictions are highlighted in Paper V. Depending on that kind of information was available, we assigned a putative superfamily annotation, and functional annotations from one or

two of the GO-branches.

There are five types of domains that Ginzu assigns, psiblast domains, the most confident, is assigned having a sequence detectable homolog with known structure. The fold recognition domain is very similar to the psiblast domain, but the homolog was detected using a fold recognition algorithm instead of psiblast. The Pfam hidden Markov model identifies Pfam domains and the MSA domain is identified by having a lot of homologs that do not have a known structure and that do not belong to a Pfam family. The last domain, the unassigned domain, is everything else that is not assigned to any of the 4 other types of domains. The 6200 ORFs of the yeast genome was parsed into 15900 domains, and 41% of these were psiblast domains. 11% were fold recognition domains; 6% Pfam and 14% MSA domains. The rest, 29% were unassigned. The MSA, Pfam and unassigned domains were subjected to Rosetta if shorter than 150 amino acids, and we successfully obtained predictions for 3350 domains. Out of these, we could classify 390 domains to a superfamily using the MCM alone, and 1448 were assigned a superfamily using additional data. We also assigned putative functions to 2700 of the 6200 proteins.

5. General Discussion

5.1. Methods

The protein structure prediction technology has improved over the last 10 years and the technology is reaching levels where it is useful from a biological perspective. As with many computational approaches, the false positive rate is high but there are several possibilities to alleviate the problem. We have focused on two ways to increase the accuracy of predictions. The first approach involved increasing the resolution of the prediction, increasing the sampling and evaluating the result using full-atom models. The structure prediction of target T281 in CASP6 (Paper I) indicated that it is possible to predict protein structures at a high resolution and identify good prediction using a full-atom energy function. Structure predictions for T281 and a number of homologs of T281 were generated. The resulting structure predictions were clustered and processed both individually and in combinations. The native T281 sequence war threaded onto the structure predictions of the homologs, and the loops rebuilt. In addition, the centroid-based backbones were converted to full-atom structures, and a more physical energy function was applied to evaluate fitness of the structure. Rosetta was incapable of generating low-rmsd structure predictions for T281, but was capable to do so with one of the homologs. Once the full atom model was built, it was possible to recognize the correct topology by the full atom energy function. This was followed up by P. Bradley who was successful in 6 out of 16 cases [157]. This is about 150 times more expensive computationally compared to the cen-

troid-based structure prediction, and hence, cannot be applied on a genome-wide scale yet. T281 is the most accurate ab initio protein structure prediction to date. Although computationally expensive, this approach is viable, and the most critical step is to select targets of interest. Interesting targets have to be within reach of the protocol, i.e. short enough to be computationally feasible. The second approach to increase the quality of the structure prediction is to filter out the false positives using auxiliary data, either from proteomics experiments (Paper II) or from databases (Paper V). This is much less computationally expensive, but is limited to proteins that belong to a known superfamily.

The Rosetta model evaluation and data integration was done manually in Paper II. In the process, we learned that two long alpha helices in the predicted structure could easily be aligned with two long alpha helices in the experimental domain even though the rest of the proteins are very dissimilar. Hence, two aligned alpha helices are not enough to classify the predicted structure with a superfamily and need to be filtered out. This was achieved automatically by calculating the contact order of the matched region and penalizing local alignments. By comparing the predicted structures to a database of domains, and evaluating the probability that the predicted structure does indeed belong the same, say, SCOP superfamily, is one use of the data. This will give the researcher an opportunity to identify important parts of the sequence and in some cases functional amino acids. In order to assign probabilities on how likely a given predicted structure does match an entry in a domain database. I created a reference dataset, that is, I predicted structures for 1000 proteins of known structure excluding structural information from the protein itself and its sequence detectable homologs. This

dataset will be referred to as the scopFold dataset from here on. By comparing the predicted structures from the scopFold to the domain database (Astral 1.67, reduced to 40% sequence identity), and measuring a number of features, I came up with a list of variables containing information about the likelihood of the proteins belonging to the same protein family or not. From the work in Paper II, we learned that features determining fitness of a match for all alpha proteins are different from the features from all beta proteins or alpha/beta proteins. Logistic regression models were created for all alpha proteins, all beta proteins and alpha/beta proteins using the information carrying variables. The reason for this was a simple analysis of how important individual variables were, and the difference between the different protein classes. Each protein generates several thousand matches to be evaluated, but only a small number of these are correct. This complicates the problem, because just determining "incorrect" will give you right almost all the time, but then, the model would have no predictive power. The four main variables we use are: (1) zscore, which evaluates how similar the predicted structures is; (2) convergence, which is a metric of how confident Rosetta was in the prediction; (3) contact order [158] over the matched region of the predicted structure, which is evaluating the complexity of the match; and (4) the absolute log of the ratio between the two protein lengths. Much time was spent trying to come up with a model that accurately could identify matches between a set of structure predictions and domains. As always, one have to take a stand whether it is most important to maximize the number of true positives or minimize the number of false positives. The first approach will identify more of the correct predictions, but will also allow more incorrect predictions to be reported

as true. The second approach will give you mostly correct predictions, but at the expense of throwing out correct matches to a higher degree. We created several hundred statistical models, mainly utilizing general linear regression, but Gaussian mixture models and k-nearest neighbor approaches were also investigated in collaboration with Luca Cazzanti and Maya Gupta at the Electrical Engineering department, University of Washington (to appear in Proceedings of the IEEE Workshop on Machine Learning for Signal Processing). We produced close to 1100 million structure predictions and thus we were more interested in identifying true positives at the expense of throwing out correct predictions.

In Paper II, the Yeast Resource Center applied four techniques on 100 essential uncharacterized ORFs. These 100 have been of interest to the yeast community since they seem to be essential for survival of the organism under optimal growth conditions. One can speculate that these genes are part of the basic machinery that is the basis of life. The four technologies produced a wealth of data and after months of trying to interpret this data it stood clear that it was possible to quite accurately identify functions for these 100 proteins even though the incoming data is noisy. The reason for this is attributed to the fact that the incorrect information from each technology was incompatible with the incorrect information from other technologies, whereas the correct information in each dataset is compatible. For example, if a protein is implicated in DNA repair and a signal cascade process by MS, found in the nucleus and predicted to be either a DNA binder or involved in electron transfer, it is more probable that DNA repair, DNA binding and nucleus are correct. The main drawback with the approach was the hours spent; three people spent the better part of three months pouring

over the data. This approach hence has a natural limit for how many targets can be processed. The next couple of years were spent trying to rigorously automate what we did in Paper II. To make the approach more generic, we decided to convert interaction data (MS and Y2H) to GO process terms and localization data to GO component terms. Protein structure prediction data can be converted to molecular function, or if structural characterization is the goal, SCOP classification distributions. With this comes the possibility to replace the MS, Y2H and fluorescent microscopy with information from online databases, an approach faster and cheaper than carrying out the experiments. The first thing we wanted to do was do be able to annotate full genomes structurally, i.e. with SCOP superfamily information. The structure prediction data at best gives a distribution of more or less likely superfamilies. If the protein that was folded have functional annotations it is possible to generate a distribution of superfamilies compatible with that function. To do this, a mapping between superfamily and GO is needed. This was generated from 250000 GO annotations from the PDB. We took a Bayesian approach in combining these distributions, and as seen in Paper V, it is a successful way of approaching the automation problem.

5.2. Information Management

Before discussing the protein structure prediction and integration with data, a more technical aspect of this work has to be discussed. Information management, or, how to store, organize, analyze and integrate data has been the focus of attention of a large research community for a long time. Biology has become an information science, and the amount of data available is large, and the rate of pro-

duction is increasing. At some point, too much information becomes a problem not trivial to solve. "We're drowning in information, but starve for knowledge" is an excellent quote that captures the problem accurately. The reason for this is that many information resources, for example, GenBank and PDB, are growing exponentially. The number of protein structures deposited in the last six months outnumbers the number of structures deposited between 1972 and 1992. The amount of data and the complexity of the data create problems when working on full genomes.

Much of the effort in this thesis has been to solve problems related to information management, and both paper III and IV are about information management in Mass Spectrometry but the same software has been utilized in the organization and management of the data in the three other papers. The information management is addressed by a three-tier database/visualization model with a relational database as the data storage tier and a web servers as the visualization tier. This approach has advantages and disadvantages. The advantage of relational databases is the capability of managing large amounts of related data, and making it easily accessible. One of the major disadvantages is that the data model is "square", meaning that all proteins have to have the same attributes. If a subset of your proteins have additional attribute, these attributes have to be stored in either a partly filled column in the database, or as an additional table. I have mainly employed the latter strategy with additional tables, party explaining why the data model is approaching 200 tables.

5.3. Application

Since we cannot guarantee that we will be able to extract useful information from any single target, application on a large scale becomes necessary. This was done in collaboration with the Institute for systems biology and IBM. 80 genomes were processed, ranging from mycoplamsa to human. The main focus for me was on yeast since it is one of the most studied Eukaryotes and is well annotated (Paper V). Yeast is much less complex than the higher Eukaryotes, and is an easier organism to do genome-wide studies on. When there is functional annotations, the success rate is higher, since, not uncommonly, we do have three or four probable superfamilies assigned from the predicted structures. If one of these are favored by the functional data, it will come out as significant. Higher eukaryotes do have another disadvantage - their genes are on average longer, and the biggest constraint for Rosetta is length. The same is true for Ginzu and as the gene gets longer the domain predictions get worse. Circumstantial evidence tells us that domain boundaries that are 10 or 20 amino acids wrong can make subsequent steps more difficult. One can imagine that an extra helix or an extra sheet that in really belongs to one domain get included with another domain. The domain missing and SS element might not be able to construct the sheets it has to have, or will score poorly because a large hydrophobic surface is exposed to the solvent. On the other hand, the second domain now has an extra element that it does not want to leave out in the solvent, but cannot fit within the domain. Both predictions are of lower quality. The goal of the confidence function and the integration is obviously to give these incorrect structures low probabilities and call

them failures and focus on the genes where everything worked out. Because of computational limitations, the proteins were predicted in low resolution, meaning that each amino acid residue was approximated by a centroid, and the energy function is based on features such as hydrophobic burial, radius of gyration and statistically favorable amino acid pairings.

We have generated a rich information resource and made it available to the general public. One of the more frustrating sides of these projects is that we cannot experimentally verify our findings. We are left with presenting the data to the research community and hope that our information is put to use. The quality of our predictions will hopefully become clear as more of our predictions get verified or rejected. From experiences from Paper II, this will take many years and the results are not always easy to interpret.

This thesis investigates whether the recent advances in the protein structure prediction field are applicable in large-scale biology. The approach taken was to search through a large number of protein sequences of unknown structure and identify well-predicted protein structures by statistical means and in combination with other sources of data. There are at least three viable paths to take now when we have a large number of genomes structurally characterized by ab initio protein structure prediction. First, it is now possible to generate high-resolution models with Rosetta, secondly, more experimental constraint data can be used both in increase the quality of all ready predicted structures and extend the scope for Rosetta, and third, do comparative proteomics studies and use potential orthologs to re-cluster and by those means get more reliable structures.

6. Future Perspectives

Rosetta is developed to work on small, globular proteins and it is unclear how well it performs on proteins that do not fit that description. The small solvated globular proteins constitute a significant fraction of all proteins, but proteins of other classes are of great interest. Proteins over 150 amino acids are difficult to model in Rosetta, most likely because the search space becomes too large and the amount of sampling in the current protocol is limited. Many of today's drugs target membrane proteins for example, and although much work has been done to predict structures for membrane proteins, this is still difficult, possibly because we know the structure of only a few membrane proteins making the statistical potentials less reliable. Many proteins are members of complexes, and how these complexes are organized and their dynamics is not well understood. It is crucial that the capability of Rosetta gets extended to larger proteins, proteins in membranes and proteins in complexes. Many of these problems will be difficult to solve given the added complexity of the search space when dealing with larger proteins, protein complexes and protein interactions. To obtain experimental data to reduce the search space seems to be a feasible strategy, one recent example is von Heijnes approach to characterize transmembrane proteins in E.coli. von Heijnes group combined a computational method (TMHMM [27]) with a large scale experiment determining the orientation of membrane proteins in E.coli and achieved a much higher confidence as compared to using TMHMM alone [159]. There are number of considerations that have to be taken into account when dealing with protein structure prediction and experimental data. One is that the ex-

perimental technology used should be applicable on a large scale and hence have to be cheap and fast. The other is that it should be able to say something about organization of proteins, protein interactions or protein complexes. Mass spectrometry is a promising technology that can be applicable in protein structure prediction in a number of ways. Additional possibilities is to cross link inter or intra molecular parts of proteins to find distance constraints. These can be used in protein structure prediction. Another possibility is to modify amino acids found on the surface of the protein that can be detected with the mass spectrometer. It might be possible to exchange the fast-exchanging hydrogens before subjecting the proteins to MS analysis. In general, residues close to the surface have a faster hydrogen exchange, and thus heavily substituted residues are more likely to be close to the surface. This information could be incorporated in the structure prediction process, making the predicted structures explaine more of the data. This would give constraint information effectively reducing the search space, and hence speeding up the prediction process and giving higher quality predictions. Yet another idea would be to modify amino acids from membrane proteins sticking out into the solution to get powerful constraints on predicting membrane protein structures. The list of chemical modifications that could generated constraints detectable by a mass spectrometer is potentially long. A first step would be to predict protein structures with publicly available constraints and compare to a decoy population generated without the constraints. This would give us an estimate of useful constrains are and how much you need to make a significant impact on the structure prediction process.

Some fraction of the proteins in a cell works in protein complexes. Research on

complexes has focused on trying to determine the composition of these complexes. To understand why these complexes look the way they do and what the individual parts are doing will remain a challenge for years. One approach is to identify orthologous complexes. By learning what constitutes a complex in organism A, it is possible to reconstruct the same complex in organism B by homology. Interestingly, some parts of identical complexes are exchanged for non-homologous proteins. One could speculate that these parts need to carry out the same, or very similar functions for the complex to perform the same function. This information can be used in several ways. One way is to identify what protein folds that are compatible with the function of the unknown part in organism B, and then try to find a protein with such a protein fold that has not been implicated in other functions. In the long run, complexes will provide a wealth of information about exchangabililty in evolution and provide insight into the development of the complex systems we find in cells today

Protein structure prediction is gaining traction in the biological community and can provide means to explain functional features observed for a protein of interest. The possibilities to do so will only become greater as the technology matures. The three possible next steps outlined above is just examples for how this technology and the data generated can be used.

7. Populärvetenskaplig Sammanfattning på Svenska

Många av alla kemiska processer som pågår i våra kroppar utförs av proteiner. När de inte fungerar blir vi sjuka. I vissa fall, till exempel vid Alzheimers sjukdom eller Creutzfeld Jacobs sjukdom, beror det på att ett visst protein har fel form. Att förstå hur proteiner antar sin slutliga form och vad det har för inverkan på proteinets funktion är således viktigt. År 2005 kostade det mellan en och ett par miljoner att mäta formen på ett enda protein på grund av att utrustning är dyr och det krävs mycket arbete. På sextiotalet upptäckte Ryle att proteiner verkar ha en ritning för vilken form de antar inbyggt i ordningen på aminosyrorna, proteinernas byggstenar. Sedan dess har det lagts ner mycket tid på att försöka förstå och kunna förutspå vilken form ett protein får när man bara vet ordningen på amino syrorna. Under de senaste tio åren har teknologin blivit bättre. I detta arbete har jag användt mig av Rosetta, ett mjukvaroprogram, som utvecklas av David Baker vid University of Washington. Rosetta kan förutsäga vilken form ett protein har utifrån ordningen av aminosyrorna. Genom att använda Rosetta på alla proteiner i jäst och kombinera resultatet med information både från experimentella tekniker och databaser har vi lyckats öka förståelsen för hur jäst fungerar och vad styrkorna och svagheterna är med den teknologi som vi utvecklat. Förhoppningen är att denna information leder till en ökad förståelse i biologin som helhet.

8. Acknowledgments

Many people have contributed to this thesis, both directly in the form of scientific discussions and collaborations and more indirectly by making my time outside science fulfilling. I would like to specifically thank a few of you. Many thanks to Thomas Laurell, György Marko-Varga, David Baker and Trisha Davis for supervising me, giving me advice and teaching me the art of science. Special thanks Michael Riffle, Richard Bonneau, Philip Bradley, David Kim, Dylan Chivian and Keith Laidig and the rest of the bakerlab for great collaborations and discussions. I would like to thank the YRC and all the great people; Tony Hazbun, Beth Graczyk, Bethany Fox, Bryan Sundin, Brian Snydsman, Scott Anderson, Hayes McDonald, John Yates, Stan Fields, Ruedi Aebersold, Bill Nobel and Michael MacCoss. Thanks ISB and Leroy Hood for making the genome-wide structure predictions possible and thanks IBM, Viktors Berstis, Bill Bovermann and Rick Alther for giving us access to unbelievable amounts of computer power! Gunilla Westergren-Thorsson, thanks for all the great discussions and for letting me use your servers and networks! Thanks Johan, Erik, Nora, Emma, Marianne and Anders for much needed support! Thanks Lisen - you know how important you are to me.

Bibliography

1. Anfinsen, CB. *Principles that govern the folding of protein chains.* Science (1973), *181:* 223-30.

2. Privalov, PL., Gill, SJ. *Stability of protein structure and hydrophobic interaction.* Adv Protein Chem (1988), *39:* 191-234.

3. RYLE, AP., SANGER, F., SMITH, LF., KITAI, R. *The disulphide bonds of insulin.* Biochem J (1955), *60:* 541-56.

4. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., Erlich, H. *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction.* Cold Spring Harb Symp Quant Biol (1986), *51:* 263-73.

5. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., Erlich, H. *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. 1986.* Biotechnology (1992), *24:* 17-27.

6. Benson, DA., Karsch-Mizrachi, I., Lipman, DJ., Ostell, J., Wheeler, DL. *GenBank.* Nucleic Acids Res (2003), *31:* 23-7.

7. Benson, DA., Karsch-Mizrachi, I., Lipman, DJ., Ostell, J., Wheeler, DL. *GenBank: update.* Nucleic Acids Res (2004), *32:* D23-6.

8. Benson, DA., Karsch-Mizrachi, I., Lipman, DJ., Ostell, J., Wheeler, DL. *GenBank.* Nucleic Acids Res (2005), *33:* D34-8.

9. Fleischmann, RD., Adams, MD., White, O., Clayton, RA., Kirkness, EF., Kerlavage, AR., Bult, CJ., Tomb, JF., Dougherty, BA., Merrick, JM.

Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (1995), *269:* 496-512.

10. Goffeau, A. Barrell, BG. Bussey, H. Davis, RW. Dujon, B. Feldmann, H. Galibert, F. Hoheisel, JD. Jacq, C. Johnston, M. Louis, EJ. Mewes, HW. Murakami, Y. Philippsen, P. Tettelin, H., Oliver, SG. *Life with 6000 genes*. Science (1996), *274:* 546, 563-7.

11. Lander, ES., Linton, LM., Birren, B., Nusbaum, C., Zody, MC., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., *Initial sequencing and analysis of the human genome*. Nature (2001), *409:* 860-921.

12. Venter, JC., Adams, MD., Myers, EW., Li, PW., Mural, RJ., Sutton, GG., Smith, HO., Yandell, M., Evans, CA., Holt, RA., Gocayne, JD., Amanatides, P., Ballew, RM., Huson, DH., Wortman, JR., Zhang, Q., Kodira, CD., Zheng, XH., Chen, L., Skupski, M., Subraman *The sequence of the human genome*. Science (2001), *291:* 1304-51.

13. Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J., Curwen, V., Cutts, T., Down, T., Durbin, R., Eyras, E., Fernandez-Suarez, XM., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iy *Ensembl 2004*. Nucleic Acids Res (2004), *32:* D468-70.

14. Dobson, CM. *Chemical space and biology*. Nature (2004), *432:* 824-8.

15. Ellis, RJ. *Macromolecular crowding: an important but neglected aspect of the intracellular environment*. Curr Opin Struct Biol (2001), *11:* 114-9.

16. Ghaemmaghami, S., Huh, WK., Bower, K., Howson, RW., Belle, A., Dephoure, N., O"Shea, EK., Weissman, JS. *Global analysis of protein expression in yeast*. Nature (2003), *425:* 737-41.

17. Dobson, CM. *Protein folding and misfolding*. Nature (2003), *426:* 884-90.

18. ANFINSEN, CB., HABER, E., SELA, M., WHITE, FH. *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain*. Proc Natl Acad Sci U S A (1961), *47:* 1309-14.

19. WATSON, JD., CRICK, FH. *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid*. Nature (1953), *171:* 737-8.

20. Kettenberger, H., Armache, KJ., Cramer, P. *Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS*. Mol Cell (2004), *16:* 955-65.

21. Zhang, Q., Powers, ET., Nieva, J., Huff, ME., Dendle, MA., Bieschke, J., Glabe, CG., Eschenmoser, A., Wentworth, P., Lerner, RA., Kelly, JW. *Metabolite-initiated protein misfolding may trigger Alzheimer"s disease*. Proc Natl Acad Sci U S A (2004), *101:* 4752-7.

22. Scheeff, ED., Fink, JL. *Fundamentals of protein structure*. Methods Biochem Anal (2003), *44:* 15-39.

23. Richardson, JS., Richardson, DC., Tweedy, NB., Gernert, KM., Quinn, TP., Hecht, MH., Erickson, BW., Yan, Y., McClain, RD., Donlan, ME. *Looking*

at proteins: representations, folding, packing, and design. Biophysical Society National Lecture, 1992. Biophys J (1992), *63:* 1185-209.

24. Silverman, BD. *Hydrophobic moments of protein structures: spatially profiling the distribution*. Proc Natl Acad Sci U S A (2001), *98:* 4996-5001.

25. Pace, CN., Shirley, BA., McNutt, M., Gajiwala, K. *Forces contributing to the conformational stability of proteins*. FASEB J (1996), *10:* 75-83.

26. Liang, J., Dill, KA. *Are proteins well-packed?*. Biophys J (2001), *81:* 751-66.

27. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, EL. *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes*. J Mol Biol (2001), *305:* 567-80.

28. Rost, B. *Prediction in 1D: secondary structure, membrane helices, and accessibility*. Methods Biochem Anal (2003), *44:* 559-87.

29. Lupas, A., Van Dyke, M., Stock, J. *Predicting coiled coils from protein sequences*. Science (1991), *252:* 1162-4.

30. Wright, PE., Dyson, HJ. *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm*. J Mol Biol (1999), *293:* 321-31.

31. Li, X., Romero, P., Rani, M., Dunker, AK., Obradovic, Z. *Predicting Protein Disorder for N-, C-, and Internal Regions*. Genome Inform Ser Workshop Genome Inform (1999), *10:* 30-40.

32. Linding, R., Jensen, LJ., Diella, F., Bork, P., Gibson, TJ., Russell, RB. *Protein disorder prediction. Implications for structural proteomics.*

Structure (Camb) (2003), *11:* 1453-9.

33. Romero, P., Obradovic, Z., Li, X., Garner, EC., Brown, CJ., Dunker, AK. *Sequence complexity of disordered protein*. Proteins (2001), *42:* 38-48.

34. Rashin, AA. *Location of domains in globular proteins*. Nature (1981), *291:* 85-7.

35. Janin, J., Wodak, SJ. *Structural domains in proteins and their role in the dynamics of protein function*. Prog Biophys Mol Biol (1983), *42:* 21-78.

36. Rose, GD. *Hierarchic organization of domains in globular proteins*. J Mol Biol (1979), *134:* 447-70.

37. Wetlaufer, DB. *Nucleation, rapid folding, and globular intrachain regions in proteins*. Proc Natl Acad Sci U S A (1973), *70:* 697-701.

38. Go, M. *Modular structural units, exons, and function in chicken lysozyme*. Proc Natl Acad Sci U S A (1983), *80:* 1964-8.

39. Sadowski, I., Stone, JC., Pawson, T. *A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus P130gag-fps*. Mol Cell Biol (1986), *6:* 4396-408.

40. Wernisch, L., Wodak, SJ. *Identifying structural domains in proteins*. Methods Biochem Anal (2003), *44:* 365-85.

41. Marchler-Bauer, A., Anderson, JB., Cherukuri, PF., DeWeese-Scott, C., Geer, LY., Gwadz, M., He, S., Hurwitz, DI., Jackson, JD., Ke, Z., Lanczycki, CJ., Liebert, CA., Liu, C., Lu, F., Marchler, GH., Mullokandov,

M., Shoemaker, BA., Simonyan, V., Song, JS., *CDD: a Conserved Domain Database for protein classification*. Nucleic Acids Res (2005), *33:* D192-6.

42. Geer, LY., Domrachev, M., Lipman, DJ., Bryant, SH. *CDART: protein homology by domain architecture*. Genome Res (2002), *12:* 1619-23.

43. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, SR., Griffiths-Jones, S., Howe, KL., Marshall, M., Sonnhammer, EL. *The Pfam protein families database*. Nucleic Acids Res (2002), *30:* 276-80.

44. Bateman, A., Coin, L., Durbin, R., Finn, RD., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, EL., Studholme, DJ., Yeats, C., Eddy, SR. *The Pfam protein families database*. Nucleic Acids Res (2004), *32:* D138-41.

45. Bi, X., Corpina, RA., Goldberg, J. *Structure of the Sec23/24-Sar1 pre-budding complex of the COPII vesicle coat*. Nature (2002), *419:* 271-7.

46. Hartl, FU., Hlodan, R., Langer, T. *Molecular chaperones in protein folding: the art of avoiding sticky situations*. Trends Biochem Sci (1994), *19:* 20-5.

47. Gross, M. *Protein folding: think globally, (inter)act locally*. Curr Biol (1998), *8:* R308-9.

48. Daggett, V., Fersht, AR. *Is there a unifying mechanism for protein folding?*. Trends Biochem Sci (2003), *28:* 18-25.

49. Wente, SR. *Gatekeepers of the nucleus*. Science (2000), *288:* 1374-7.

50. Gavin, AC. Bösche, M. Krause, R. Grandi, P. Marzioch, M. Bauer, A. Schultz, J. Rick, JM. Michon, AM. Cruciat, CM. Remor, M. Höfert, C.

Schelder, M. Brajenovic, M. Ruffner, H. Merino, A. Klein, K. Hudak, M. Dickson, D. Rudi, T. Gnau, V. Bauch, A. Bastuck *Functional organization of the yeast proteome by systematic analysis of protein complexes*. Nature (2002), *415:* 141-7.

51. Ho, Y. Gruhler, A. Heilbut, A. Bader, GD. Moore, L. Adams, SL. Millar, A. Taylor, P. Bennett, K. Boutilier, K. Yang, L. Wolting, C. Donaldson, I. Schandorff, S. Shewnarane, J. Vo, M. Taggart, J. Goudreault, M. Muskat, B. Alfarano, C. Dewar, D. Lin, Z. Mic *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry*. Nature (2002), *415:* 180-3.

52. Sali, A., Glaeser, R., Earnest, T., Baumeister, W. *From words to literature in structural proteomics*. Nature (2003), *422:* 216-25.

53. Service, R. *Structural biology. Structural genomics, round 2*. Science (2005), *307:* 1554-8.

54. Montelione, GT., Anderson, S. *Structural genomics: keystone for a Human Proteome Project*. Nat Struct Biol (1999), *6:* 11-2.

55. Couzin, J. *Structural biology. Ten centers chosen to decode protein structures*. Science (2005), *309:* 230.

56. Berman, HM., Battistuz, T., Bhat, TN., Bluhm, WF., Bourne, PE., Burkhardt, K., Feng, Z., Gilliland, GL., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, JD., Zardecki, C. *The Protein Data Bank.* Acta Crystallogr D

Biol Crystallogr (2002), *58:* 899-907.

57. Skolnick, J., Fetrow, JS. *From genes to protein structure and function: novel applications of computational approaches in the genomic era*. Trends Biotechnol (2000), *18:* 34-9.

58. Todd, AE., Orengo, CA., Thornton, JM. *Evolution of function in protein superfamilies, from a structural perspective*. J Mol Biol (2001), *307:* 1113-43.

59. Bartlett, GJ., Todd, AE., Thornton, JM. *Inferring protein function from structure*. Methods Biochem Anal (2003), *44:* 387-407.

60. Ashburner, M. Ball, CA. Blake, JA. Botstein, D. Butler, H. Cherry, JM. Davis, AP. Dolinski, K. Dwight, SS. Eppig, JT. Harris, MA. Hill, DP. Issel-Tarver, L. Kasarskis, A. Lewis, S. Matese, JC. Richardson, JE. Ringwald, M. Rubin, GM., Sherlock, G. *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*. Nat Genet (2000), *25:* 25-9.

61. Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., Apweiler, R. *The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro*. Genome Res (2003), *13:* 662-72.

62. Barabasi, AL., Oltvai, ZN. *Network biology: understanding the cell"s functional organization*. Nat Rev Genet (2004), *5:* 101-13.

63. Hazbun, TR., Fields, S. *Networking proteins in yeast*. Proc Natl Acad Sci U S A (2001), *98:* 4277-8.

64. Ng, SK., Zhang, Z., Tan, SH. *Integrative approach for computationally inferring protein domain interactions*. Bioinformatics (2003), *19:* 923-9.

65. Kelley, R., Ideker, T. *Systematic interpretation of genetic interactions using protein networks*. Nat Biotechnol (2005), *23:* 561-6.

66. Tong, AH., Lesage, G., Bader, GD., Ding, H., Xu, H., Xin, X., Young, J., Berriz, GF., Brost, RL., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, DS., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J *Global mapping of the yeast genetic interaction network*. Science (2004), *303:* 808-13.

67. Schwikowski, B. Uetz, P., Fields, S. *A network of protein-protein interactions in yeast*. Nat Biotechnol (2000), *18:* 1257-61.

68. Edwards, A., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., Gerstein, M. *Bridging structural biology and genomics: assessing protein interaction data with known complexes*. Trends Genet (2002), *18:* 529.

69. Jeong, H., Tombor, B., Albert, R., Oltvai, ZN., Barabasi, AL. *The large-scale organization of metabolic networks*. Nature (2000), *407:* 651-4.

70. Ravasz, E., Somera, AL., Mongru, DA., Oltvai, ZN., Barabasi, AL. *Hierarchical organization of modularity in metabolic networks*. Science (2002), *297:* 1551-5.

71. Papp, B., Pal, C., Hurst, LD. *Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast*. Nature (2004), *429:* 661-4.

72. Hasty, J., McMillen, D., Isaacs, F., Collins, JJ. *Computational studies of*

gene regulatory networks: in numero molecular biology. Nat Rev Genet (2001), *2:* 268-79.

73. Shen-Orr, SS., Milo, R., Mangan, S., Alon, U. *Network motifs in the transcriptional regulation network of Escherichia coli*. Nat Genet (2002), *31:* 64-8.

74. Lee, TI., Rinaldi, NJ., Robert, F., Odom, DT., Bar-Joseph, Z., Gerber, GK., Hannett, NM., Harbison, CT., Thompson, CM., Simon, I., Zeitlinger, J., Jennings, EG., Murray, HL., Gordon, DB., Ren, B., Wyrick, JJ., Tagne, JB., Volkert, TL., Fraenkel, E., Giffo *Transcriptional regulatory networks in Saccharomyces cerevisiae*. Science (2002), *298:* 799-804.

75. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U. *Network motifs: simple building blocks of complex networks.* Science (2002), *298:* 824-7.

76. Bhalla, US., Iyengar, R. *Emergent properties of networks of biological signaling pathways*. Science (1999), *283:* 381-7.

77. Hartwell, LH., Hopfield, JJ., Leibler, S., Murray, AW. *From molecular to modular cell biology*. Nature (1999), *402:* C47-52.

78. Ideker, T., Galitski, T., Hood, L. *A new approach to decoding life: systems biology*. Annu Rev Genomics Hum Genet (2001), *2:* 343-72.

79. Tyers, M., Mann, M. *From genomics to proteomics*. Nature (2003), *422:* 193-7.

80. Jacob, F. *Evolution and tinkering*. Science (1977), *196:* 1161-6.

81. Wolfe, KH., Shields, DC. *Molecular evidence for an ancient duplication of the entire yeast genome*. Nature (1997), *387:* 708-13.

82. Long, M., Thornton, K. *Gene duplication and evolution*. Science (2001), *293:* 1551.

83. Lynch, M., Conery, JS. *The evolutionary fate and consequences of duplicate genes*. Science (2000), *290:* 1151-5.

84. Rost, B. *Twilight zone of protein sequence alignments*. Protein Eng (1999), *12:* 85-94.

85. Bourne, PE., Shindyalov, IN. *Structure comparison and alignment*. Methods Biochem Anal (2003), *44:* 321-37.

86. Koonin, EV., Wolf, YI., Karev, GP. *The structure of the protein universe and genome evolution*. Nature (2002), *420:* 218-23.

87. Lichtarge, O. *Getting past appearances: the many-fold consequences of remote homology*. Nat Struct Biol (2001), *8:* 918-20.

88. Altschul, SF., Madden, TL., Schaffer, AA., Zhang, J., Zhang, Z., Miller, W., Lipman, DJ. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res (1997), *25:* 3389-402.

89. Bolten, E. Schliep, A. Schneckener, S. Schomburg, D., Schrader, R. *Clustering protein sequences-structure prediction by transitive homology*. Bioinformatics (2001), *17:* 935-41.

90. Koonin, EV. *An apology for orthologs - or brave new memes*. Genome Biol (2001), *2:* COMMENT1005.

91. Chothia, C., Lesk, AM. *The relation between the divergence of sequence and structure in proteins*. EMBO J (1986), *5:* 823-6.

92. Sander, C., Schneider, R. *Database of homology-derived protein structures and the structural meaning of sequence alignment*. Proteins (1991), *9:* 56-68.

93. Acharya, KR., Ren, JS., Stuart, DI., Phillips, DC., Fenna, RE. *Crystal structure of human alpha-lactalbumin at 1.7 A resolution*. J Mol Biol (1991), *221:* 571-81.

94. Holm, L., Sander, C. *Mapping the protein universe*. Science (1996), *273:* 595-603.

95. Holm, L., Sander, C. *Protein structure comparison by alignment of distance matrices*. J Mol Biol (1993), *233:* 123-38.

96. Gibrat, JF., Madej, T., Bryant, SH. *Surprising similarities in structure comparison*. Curr Opin Struct Biol (1996), *6:* 377-85.

97. Ortiz, AR., Strauss, CE., Olmea, O. *MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison*. Protein Sci (2002), *11:* 2606-21.

98. Siew, N. Elofsson, A. Rychlewski, L., Fischer, D. *MaxSub: an automated measure for the assessment of protein structure prediction quality*. Bioinformatics (2000), *16:* 776-85.

99. Murzin, AG. Brenner, SE. Hubbard, T., Chothia, C. *SCOP: a structural classification of proteins database for the investigation of sequences and*

structures. J Mol Biol (1995), *247:* 536-40.

100. Lo Conte, L., Brenner, SE., Hubbard, TJ., Chothia, C., Murzin, AG. *SCOP database in 2002: refinements accommodate structural genomics*. Nucleic Acids Res (2002), *30:* 264-7.

101. Chothia, C. *Proteins. One thousand families for the molecular biologist*. Nature (1992), *357:* 543-4.

102. Govindarajan, S., Recabarren, R., Goldstein, RA. *Estimating the total number of protein folds*. Proteins (1999), *35:* 408-14.

103. Wolf, YI., Grishin, NV., Koonin, EV. *Estimating the number of protein folds and families from complete genome data*. J Mol Biol (2000), *299:* 897-905.

104. Chandonia, JM., Hon, G., Walker, NS., Lo Conte, L., Koehl, P., Levitt, M., Brenner, SE. *The ASTRAL Compendium in 2004*. Nucleic Acids Res (2004), *32:* D189-92.

105. Holm, L., Ouzounis, C., Sander, C., Tuparev, G., Vriend, G. *A database of protein structure families with common folding motifs*. Protein Sci (1992), *1:* 1691-8.

106. Orengo, CA. Michie, AD. Jones, S. Jones, DT. Swindells, MB., Thornton, JM. *CATH--a hierarchic classification of protein domain structures*. Structure (1997), *5:* 1093-108.

107. Hadley, C., Jones, DT. *A systematic comparison of protein structure classifications: SCOP, CATH and FSSP*. Structure Fold Des (1999), *7:*

1099-112.

108. Salem, GM., Hutchinson, EG., Orengo, CA., Thornton, JM. *Correlation of observed fold frequency with the occurrence of local structural motifs*. J Mol Biol (1999), *287:* 969-81.

109. Söding, J., Lupas, AN. *More than the sum of their parts: on the evolution of proteins from peptides*. Bioessays (2003), *25:* 837-46.

110. Bourne, PE. *CASP and CAFASP experiments and their findings*. Methods Biochem Anal (2003), *44:* 501-7.

111. Simons, KT., Bonneau, R., Ruczinski, I., Baker, D. *Ab initio protein structure prediction of CASP III targets using ROSETTA*. Proteins (1999), *0:* 171-6.

112. Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, CE., Baker, D. *Rosetta in CASP4: progress in ab initio protein structure prediction*. Proteins (2001), *0:* 119-26.

113. Bujnicki, JM., Elofsson, A., Fischer, D., Rychlewski, L. *LiveBench-2: large-scale automated evaluation of protein structure prediction servers*. Proteins (2001), *0:* 184-91.

114. Bujnicki, JM., Elofsson, A., Fischer, D., Rychlewski, L. *LiveBench-1: continuous benchmarking of protein structure prediction servers*. Protein Sci (2001), *10:* 352-61.

115. Jones, DT. *Protein secondary structure prediction based on position-specific scoring matrices*. J Mol Biol (1999), *292:* 195-202.

116. Bonneau, R., Strauss, CE., Baker, D. *Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation*. Proteins (2001), *43:* 1-11.

117. Taylor, WR. *Protein structural domain identification*. Protein Eng (1999), *12:* 203-16.

118. Busetta, B., Barrans, Y. *The prediction of protein domains*. Biochim Biophys Acta (1984), *790:* 117-24.

119. Marsden, RL., McGuffin, LJ., Jones, DT. *Rapid protein domain assignment from amino acid sequence using predicted secondary structure*. Protein Sci (2002), *11:* 2814-24.

120. Suyama, M., Ohara, O. *DomCut: prediction of inter-domain linker regions in amino acid sequences*. Bioinformatics (2003), *19:* 673-4.

121. Wheelan, SJ., Marchler-Bauer, A., Bryant, SH. *Domain size distributions can predict domain boundaries*. Bioinformatics (2000), *16:* 613-8.

122. Rigden, DJ. *Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments*. Protein Eng (2002), *15:* 65-77.

123. George, RA., Heringa, J. *SnapDRAGON: a method to delineate protein structural domains from sequence data*. J Mol Biol (2002), *316:* 839-51.

124. Chivian, D., Kim, DE., Malmström, L., Bradley, P., Robertson, T., Murphy, P., Strauss, CE., Bonneau, R., Rohl, CA., Baker, D. *Automated prediction of CASP-5 structures using the Robetta server*. Proteins (2003),

53: 524-33.

125. Kim, DE., Chivian, D., Malmström, L., Baker, D. *Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM*. Proteins (2005) In print.

126. Liu, J., Rost, B. *CHOP: parsing proteins into structural domains*. Nucleic Acids Res (2004), *32:* W569-71.

127. Pieper, U., Eswar, N., Braberg, H., Madhusudhan, MS., Davis, FP., Stuart, AC., Mirkovic, N., Rossi, A., Marti-Renom, MA., Fiser, A., Webb, B., Greenblatt, D., Huang, CC., Ferrin, TE., Sali, A. *MODBASE, a database of annotated comparative protein structure models, and associated resources.* Nucleic Acids Res (2004), *32:* D217-D222.

128. Krieger, E., Nabuurs, SB., Vriend, G. *Homology modeling*. Methods Biochem Anal (2003), *44:* 509-23.

129. Godzik, A. *Fold recognition methods*. Methods Biochem Anal (2003), *44:* 525-46.

130. Kelley, LA., MacCallum, RM., Sternberg, MJ. *Enhanced genome annotation using structural profiles in the program 3D-PSSM*. J Mol Biol (2000), *299:* 499-520.

131. Rychlewski, L., Jaroszewski, L., Li, W., Godzik, A. *Comparison of sequence profiles. Strategies for structural predictions using sequence information*. Protein Sci (2000), *9:* 232-41.

132. Bujnicki, JM., Elofsson, A., Fischer, D., Rychlewski, L. *Structure*

prediction meta server. Bioinformatics (2001), *17:* 750-1.

133. Lundström, J., Rychlewski, L., Bujnicki, J., Elofsson, A. *Pcons: a neural-network-based consensus predictor that improves fold recognition*. Protein Sci (2001), *10:* 2354-62.

134. Liu, J., Rost, B. *Comparing function and structure between entire proteomes*. Protein Sci (2001), *10:* 1970-9.

135. Muller, A., MacCallum, RM., Sternberg, MJ. *Structural characterization of the human proteome*. Genome Res (2002), *12:* 1625-41.

136. Sippl, MJ. *Knowledge-based potentials for proteins*. Curr Opin Struct Biol (1995), *5:* 229-35.

137. Koppensteiner, WA., Sippl, MJ. *Knowledge-based potentials--back to the roots*. Biochemistry (Mosc) (1998), *63:* 247-52.

138. Simons, KT. Strauss, C., Baker, D. *Prospects for ab initio protein structural genomics*. J Mol Biol (2001), *306:* 1191-9.

139. Bonneau, R., Baker, D. *Ab initio protein structure prediction: progress and prospects*. Annu Rev Biophys Biomol Struct (2001), *30:* 173-89.

140. Chivian, D., Robertson, T., Bonneau, R., Baker, D. *Ab initio methods*. Methods Biochem Anal (2003), *44:* 547-57.

141. Simons, KT. Ruczinski, I. Kooperberg, C. Fox, BA. Bystroff, C., Baker, D. *Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins*. Proteins (1999), *34:* 82-95.

142. Simons, KT. Kooperberg, C. Huang, E., Baker, D. *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol (1997), *268:* 209-25.

143. Kuhlman, B., Dantas, G., Ireton, GC., Varani, G., Stoddard, BL., Baker, D. *Design of a novel globular protein fold with atomic-level accuracy.* Science (2003), *302:* 1364-8.

144. Korkegian, A., Black, ME., Baker, D., Stoddard, BL. *Computational thermostabilization of an enzyme*. Science (2005), *308:* 857-60.

145. Wilson, CA., Kreychman, J., Gerstein, M. *Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.* J Mol Biol (2000), *297:* 233-49.

146. Hegyi, H., Gerstein, M. *The relationship between protein structure and function: a comprehensive survey with application to the yeast genome.* J Mol Biol (1999), *288:* 147-64.

147. Bonneau, R., Strauss, C., Rohl, C., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., Baker, D. *De Novo Prediction of Three-dimensional Structures for Major Protein Families.* J Mol Biol (2002), *322:* 65.

148. Uetz, P. Giot, L. Cagney, G. Mansfield, TA. Judson, RS. Knight, JR. Lockshon, D. Narayan, V. Srinivasan, M. Pochart, P. Qureshi-Emili, A. Li, Y. Godwin, B. Conover, D. Kalbfleisch, T. Vijayadamodar, G. Yang, M.

Johnston, M. Fields, S., Rothberg, JM. *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*. Nature (2000), *403:* 623-7.

149. Ito, T. Chiba, T. Ozawa, R. Yoshida, M. Hattori, M., Sakaki, Y. *A comprehensive two-hybrid analysis to explore the yeast protein interactome*. Proc Natl Acad Sci U S A (2001), *98:* 4569-74.

150. Bonneau, R., Tsai, J., Ruczinski, I., Baker, D. *Functional inferences from blind ab initio protein structure predictions*. J Struct Biol (0), *134:* 186-90.

151. Click, ES., Stearns, T., Botstein, D. *Systematic Structure-Function Analysis of the Small GTPase Arf1 in Yeast*. Mol Biol Cell (2002), *13:* 1652-64.

152. Washburn, MP. Wolters, D., Yates, JR. *Large-scale analysis of the yeast proteome by multidimensional protein identification technology*. Nat Biotechnol (2001), *19:* 242-7.

153. Simon, I., Barnett, J., Hannett, N., Harbison, CT., Rinaldi, NJ., Volkert, TL., Wyrick, JJ., Zeitlinger, J., Gifford, DK., Jaakkola, TS., Young, RA. *Serial regulation of transcriptional regulators in the yeast cell cycle*. Cell (2001), *106:* 697-708.

154. Gasch, AP., Eisen, MB. *Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering*. Genome Biol (2002), *3:* RESEARCH0059.

155. Huh, WK., Falvo, JV., Gerke, LC., Carroll, AS., Howson, RW., Weissman, JS., O"Shea, EK. *Global analysis of protein localization in budding yeast*.

Nature (2003), *425:* 686-91.

156. Pedrioli, PG., Eng, JK., Hubley, R., Vogelzang, M., Deutsch, EW., Raught, B., Pratt, B., Nilsson, E., Angeletti, RH., Apweiler, R., Cheung, K., Costello, CE., Hermjakob, H., Huang, S., Julian, RK., Kapp, E., McComb, ME., Oliver, SG., Omenn, G., Paton, NW. *A common open representation of mass spectrometry data and its application to proteomics research.* Nat Biotechnol (2004), *22:* 1459-1466.

157. Bradley, P., Misura, KM., Baker, D. *Toward high-resolution de novo structure prediction for small proteins.* Science (2005), *309:* 1868-71.

158. Bonneau, R., Ruczinski, I., Tsai, J., Baker, D. *Contact order and ab initio protein structure prediction.* Protein Sci (2002), *11:* 1937-44.

159. Daley, DO., Rapp, M., Granseth, E., Melen, K., Drew, D., von Heijne, G. *Global topology analysis of the Escherichia coli inner membrane proteome.* Science (2005), *308:* 1321-3.