Combining Protein Structure Prediction with Experiments and Functional Information

Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
im Fach Bioinformatik

eingereicht im Februar 2006 an der
Naturwissenschaftlich-Technischen Fakultät I
der Universität des Saarlandes

von

Mario Albrecht

aus Altenstadt a. d. Waldnaab

Mario Albrecht: *Combining Protein Structure Prediction with Experiments and Functional Information*. Ph.D. thesis for obtaining the academic degree of a doctor of the natural sciences in bioinformatics, submitted to the Faculty I of Natural Sciences and Technology (mathematics and computer science) of the Saarland University, Saarbrücken, Germany, 2006.

Dekan (Dean of the Faculty):	Prof. Dr. Jörg Eschmeier / Prof. Dr.-Ing. Thorsten Herfet
Einreichung (Submission):	22. Februar 2006
Kolloquium (Colloquium):	07. Juni 2006
Vorsitzender (Chairman):	Prof. Dr. Joachim Weickert
Gutachter (Reviewers):	1. Prof. Dr. Thomas Lengauer, Ph.D.
	2. Prof. Dr. Hans-Peter Lenhof
	3. Dr. Sylvia Krobitsch
Protokollant (Minute taker):	Dr. Ingolf Sommer

Contents

Abstract

Proteins are key players in all cells of living organisms. In particular, knowledge of the spatial protein structure may give fundamental insights into protein function and disease processes. For many years, the successful prediction of the structural and functional properties of proteins has been a major research field in bioinformatics. This field is also addressed in this work, which comprises an applied biomedical and a methodological part.

Comprehensive application studies of bioinformatics approaches were performed, which primarily targeted autoinflammatory and neurodegenerative diseases. A variety of computational tools was used to analyze medically relevant proteins and to evaluate experimental data. Many bioinformatics methods were applied to predict the molecular structure and function of proteins. The results provided a rationale for the design, prioritization, and interpretation of experiments performed by cooperation partners. Some of the generated biological hypotheses were tested and confirmed by experiments.

In addition, the application studies revealed limitations of current bioinformatics techniques, which led to suggestions for novel approaches. Three new computational methods were developed to support the prediction of the secondary and tertiary structure of proteins and the investigation of their interaction networks. First, consensus formation between three different methods for secondary structure prediction was shown to considerably improve the prediction quality and reliability. Second, in order to utilize experimental measurements in tertiary structure prediction, scoring functions were implemented that incorporate distance constraints into the alignment evaluation, thus increasing the fold recognition rate. Third, an automatic procedure for decomposing protein networks into interacting domains was designed to obtain a more detailed molecular view of protein-protein interactions, facilitating further functional and structural analyses.

Kurzfassung

Proteinen kommt in allen Zellen lebender Organismen eine Schlüsselrolle zu. Insbesondere die Kenntnis der Raumstruktur von Proteinen kann fundamentale Einsichten in ihre Funktion und in Krankheitsprozesse liefern. Seit vielen Jahren ist die erfolgreiche Vorhersage struktureller und funktioneller Eigenschaften von Proteinen ein wichtiges Forschungsgebiet in der Bioinformatik. Dieses Gebiet ist auch Gegenstand der vorliegenden Arbeit, welche einen angewandten biomedizinischen und einen methodischen Teil umfasst.

Es wurden umfangreiche Applikationsstudien von bioinformatischen Verfahren durchgeführt, die sich vornehmlich mit autoinflammatorischen und neurodegenerativen Erkrankungen befassten. Verschiedene Computerwerkzeuge wurden verwendet, um medizinisch relevante Proteine zu analysieren und experimentelle Daten auszuwerten. Es kamen viele Bioinformatikmethoden zur Anwendung, um die molekulare Struktur und Funktion von Proteinen vorherzusagen. Die Ergebnisse dienten als Grundlage für die Planung, Priorisierung und Interpretation von Experimenten, die von Kooperationspartnern durchgeführt wurden. Einige der generierten biologischen Hypothesen wurden durch Experimente überprüft und bestätigt.

Zusätzlich deckten die Applikationsstudien Grenzen von Bioinformatikmethoden auf, was zu Vorschlägen für neuartige Verfahren führte. So wurden drei neue rechnerbasierte Methoden entwickelt, um die Vorhersage der Sekundär- und Tertiärstruktur von Proteinen sowie die Untersuchung ihrer Interaktionsnetzwerke zu unterstützen. Erstens wurde gezeigt, dass die Bildung eines Konsensus zwischen drei verschiedenen Methoden der Sekundärstrukturvorhersage die Vorhersagequalität und -verlässlichkeit erheblich verbessert. Zweitens wurden zur Nutzung experimenteller Messungen in der Tertiärstrukturvorhersage Bewertungsfunktionen implementiert, die Distanzbeschränkungen in die Alignmentevaluation einbinden, um die Faltungserkennungsrate zu erhöhen. Drittens wurde eine automatische Prozedur zur

Dekomposition von Proteinnetzwerken in interagierende Domänen entworfen, um eine detailliertere molekulare Sicht von Interaktionen zwischen Proteinen zu erhalten. Hierdurch werden weitere Analysen zu Funktion und Struktur erleichtert.

Dedicated to my parents

Acknowledgements

First and foremost, I would like to express my gratitude to Thomas Lengauer for introducing me to the field of bioinformatics and for providing his expertise and excellent guidance during my research studies. I am also grateful to Hans-Peter Lenhof and Sylvia Krobitsch for their kind willingness to be further reviewers of this thesis.

I especially wish to thank Francisco Domingues, Oliver Sander, Ingolf Sommer, and Silvio Tosatto, whose discussion on scientific issues and beneficial comments on manuscripts I have greatly appreciated. Additionally, I am indebted to many other colleagues, cooperation partners, and coauthors of joint publications for their help.

Finally, yet importantly, my family and friends merit a heartfelt mention for their encouragement and considerable support in many respects. In particular, I want to thank Dorothea and my brother and parents for their invaluable advice and understanding throughout this work.

1 Introduction

This chapter introduces the topic of the dissertation. It first addresses some issues that motivate the research into combining protein structure prediction with experiments and functional information. After the description of the research objectives, a brief overview of the performed work is provided and the structure of the thesis is outlined.

1.1 Motivation

Computer science is generally known as the science of information processing. Accordingly, bioinformatics may be defined as the computational science of processing biological information. Since it is the biologist who primarily deals with biological knowledge, bioinformatics particularly aims at supporting his or her work. To this end, numerous algorithms have been implemented in the last 20-25 years in order to analyze, annotate, curate, integrate, search, store, transform or validate biological data.

However, experimentally working biologists will often lack the time to keep abreast with the quickly progressing field of bioinformatics with hundreds of new methods and databases published every year. An unfortunate consequence of this rapid development is that many bioinformatics computer programs are never used by biologists in practice. This may also be due to the fact that judging the performance of bioinformatics methods and the quality of their results requires interdisciplinary expertise in informatics and statistics as well as in biology and medicine.

Therefore, it is useful that bioinformaticians do not only develop novel and advanced approaches to solve problems motivated by biomedicine, but also closely cooperate with bench biologists in applying computational methods. This collaboration is crucial for the accurate interpretation of bioinformatics findings and their effective incorporation into biomedical research, yielding integrative models containing experimental and computational knowledge for biology and medicine. In return,

bioinformaticians gain insights into the biological and medical aspects involved and obtain feedback for improvements and future extensions of their applications.

Joint studies of bioinformaticians and experimentalists are especially worthwhile when analyzing large volumes of experimental data in order to address the right biological questions and to find reliable and meaningful answers. The first considerable amounts of experimental high-throughput data have consisted of genomic sequences and gene expression profiles. They are still being accumulated at an increasing rate to be processed and integrated with further biological information. Additionally, large molecular data sets produced by novel metabolomics and proteomics techniques during cell-wide measurements of metabolites and proteins, respectively, have recently attracted much attention from bioinformatics research.

Metabolomics focuses on quantifying and modeling biochemical pathways of small molecule metabolites such as nucleotides, lipids, and saccharides. These substrates are catalyzed in chemical reactions, whose enzyme kinetics are determined for pathway simulations. In contrast, proteomics deals with ensembles of proteins, the proteomes, contained as gene products in cellular compartments of a given organism and tissue type at certain time points. Therefore, proteomics researches the structure and function of proteins and their interactions. These efforts are complemented by structural genomics efforts to provide spatial structure models of proteins and their binding complexes.

Generally, proteins are key players in dynamic processes inside and between cells and form complex interaction networks. They may fulfill essential functions as antibodies, enzymes, transmembrane channels, molecular motors, signal transducers, structural building blocks, substrate transporters, and transcription factors. In particular, since proteins are fundamental to life, defects of their structure and function often cause severe human diseases. Fortunately, many computational methods have already been devised to support molecular protein analyses performed by biologists and medical researchers.

Examples are sophisticated database search algorithms that are often able to detect distantly homologous protein sequences and thus discover interesting evolutionary and functional relationships between proteins. Other state-of-the-art bioinformatics methods are capable of delineating the functional protein domain architecture and of recognizing the correct structural fold of protein domains. In addition, structure predictions create models of the secondary and tertiary protein structure with sufficient accuracy for further molecular investigations. For instance, it may be possible to map the location of genetic variations found with patients onto reliable three-dimensional structural models to elucidate functional defects causative of an illness.

Considering the great importance of a beneficial cooperation and information exchange between bioinformaticians and experimentalists for successful joint biological and medical investigations, the objective of this work is two-fold. First, vital problems in biology and medicine are selected to explore the value of bioinformatics support for experiment evaluation and hypothesis formation. Importantly, the application of computational tools does not only advance the understanding of molecular disease processes, but it also reveals limitations of current bioinformatics methods. Therefore, the second, and no less important, aim of this dissertation is to address some of the

encountered problems with methodological improvements concerning analyses of protein structures and interactions.

1.2 Overview

Part of the research in the course of this dissertation has involved comprehensive application studies of bioinformatics approaches targeted primarily at autoinflammatory and neurodegenerative disorders. A great variety of computational tools were applied to analyze medically relevant proteins and to evaluate experimental results. Bioinformatics techniques were also used to predict structural and functional properties of proteins, some of which have been tested and confirmed by experiments. In other words, prediction methods for protein structures were used in combination with experimental results and further functional information. Importantly, this work has also led to the particular development of three novel computational approaches supporting the biological and medical investigation of proteins.

These three methodological contributions generally improve the prediction of protein structures and facilitate the exploration of proteins and their interaction networks. First, building the consensus between predicted secondary structures is demonstrated to increase the prediction quality and reliability. Second, a novel method utilizing experimental distance constraints is introduced to improve the recognition of structural protein domain folds and to validate tertiary structure predictions. Third, an automatic decomposition of protein networks into interacting domains is developed, which provides a more detailed molecular view of protein-protein interactions for further functional and structural examinations.

Overall, this thesis is based on about 200 pages (and 90 supplementary pages available online) of 25 coauthored publications in important scientific journals and conference proceedings (a short summary in numbers and paper abstracts are given in the Appendix). It is noteworthy that most studies focusing on specific diseases have been conducted in cooperation with experimental partners at biological and medical institutes in Germany, Italy, the Netherlands, Spain, and the USA. Several joint publications also include contributions from former and current colleagues at the Fraunhofer Institute for Scientific Computing and Algorithmics (SCAI, formerly German National Research Center for Information Technology, GMD) in St. Augustin and at the Max Planck Institute for Informatics (MPI-INF) in Saarbrücken, Germany.

The work has been performed in the context of several bioinformatics research projects with financial support from the German Research Foundation (DFG) for the projects PROSEQO and PROSTFUN on the structure and function prediction of proteins, from the Federal Ministry of Education and Research for projects within the German National Genome Research Network (NGFN) including the genome networks on diseases of the nervous system and due to environmental factors, and from the European Commission funding the BioSapiens Network of Excellence for genome annotation.

1.3 Outline

The remainder of this thesis is organized into six chapters followed by a Summary and an Appendix. Chapter 2 describes the application of bioinformatics methods to biological and medical questions, which gives rise to further methodological ideas. Chapters 3, 4 and 5 particularly explicate novel and improved computational approaches to solve biomedical problems concerning protein structure and function. Chapter 6 and the succeeding Summary conclude the accomplished work. The Appendix contains the abstracts of 25 published journal articles providing additional experimental details and biological implications. The contents of each chapter are briefly summarized in the following:

Chapter 2 gives a comprehensive account of numerous computational analyses of medically relevant proteins. It focuses on the bioinformatics methods applied and the results obtained regarding protein structures and functions, and it introduces biological and clinical aspects of the studied autoinflammatory diseases and neurodegenerative disorders. This presentation is complemented by a comparison of formerly predicted and now experimentally solved structures and by the description of methodological limitations identified during the bioinformatics application studies. The chapter closes with perspectives of computational systems biology for modeling disease processes.

Chapter 3 demonstrates a new method for predicting consensus secondary structure. Although this method is simple to implement, it is quite successful in improving the performance of secondary structure prediction. It forms a consensus prediction using the results of three different prediction methods. The benchmarking analysis performed also provides valuable insights into the similarity of the prediction results and the higher confidence in consistently predicted secondary structure.

Chapter 4 deals with improving tertiary structure prediction using additional distance constraints. The latter may be obtained by experimental techniques such as mass spectrometry or NMR spectroscopy. Significant improvements of the recognition rate of structural domain folds were observed by combining prediction results with a novel post-filtering procedure utilizing distance constraints. Novel scoring functions are applied to the computed alignments and incorporate measures of constraint satisfaction.

Chapter 5 approaches the task of how to automate the decomposition of protein networks into domain-domain interactions. It explains the design of a new plugin for Cytoscape, a software platform for the visualization and analysis of protein networks, to facilitate the exploration of protein-protein interactions at a more detailed molecular level. The plugin subdivides interacting proteins into their respective domains to compute a putative network of the corresponding domain-domain interactions.

Chapter 6 draws conclusions from the conducted research studies and summarizes the main achievements. It also evaluates the work accomplished and discusses future methodological perspectives.

2
Analyzing Medically Relevant Proteins

Integrative approaches combining the results of bioinformatics methods with additional biological information from experiments support the elucidation of molecular protein structures and functions. This chapter covers more than twenty coauthored publications on the application of computational techniques to the analysis of medically relevant proteins. Most studies have been performed in close cooperation with experimental research groups from biological and medical institutes investigating autoinflammatory or neurodegenerative disorders. The following sections summarize the joint work with focus on the involved bioinformatics work.

The first section describes various bioinformatics methods applied for the different molecular analyses. The next two sections provide biomedical background knowledge on the studied diseases and report bioinformatics findings for relevant proteins using several illustrative figures. Two distinct types of diseases have been in the center of research: autoinflammatory diseases and neurodegenerative disorders. Examples for diseases underlying autoinflammation are Crohn's, an inflammatory bowel disease, and sarcoidosis, primarily affecting the lung. Neurodegeneration may be caused by spinocerebellar ataxias, Huntington's and Parkinson's disorders. Recently, some of the diseases have attracted much attention in scientific magazines such as *Scientific American* (Cattaneo et al., 2002; Lozano and Kalia, 2005; O'Neill, 2005) and *The Scientist* (Lewis, 2003; Roberts, 2003; Anderson, 2004; Constans, 2005).

Additionally, a comparison of structural models with recent, experimentally solved, structures verifies some of the former bioinformatics predictions. Finally, remarks on identified limitations of the applied bioinformatics methods and on the perspectives of computational systems biology for modeling disease processes conclude the chapter.

2.1 Bioinformatics Methods

2.1.1 Protein Structure and Function

Each protein sequence can consist of several functionally distinct regions of varying length and structure. Protein regions may encompass signal peptides, transmembrane α-helices or β-barrels composed of β-strands, low-complexity and intrinsically unstructured regions including short interaction motifs, and evolutionarily conserved globular domains containing binding sites for other proteins and ligands. Thus, biologists often construct fragments of proteins or mutate sequences to delineate the boundaries and the function of selected regions and amino acids by experiments.

Numerous computational methods that predict the structure and the function of specific protein regions are already available to support such investigations. We applied different, and often complementary, bioinformatics tools in order to characterize medically relevant proteins structurally and functionally. Our findings then provided a rationale for the design and the interpretation of experimental studies conducted by our biomedical cooperation partners. Frequently, we also discovered novel sequence motifs and new protein family members including orthologs in the same or paralogs in another species. These discoveries provided additional insight into protein functions.

The succeeding sections describe bioinformatics methods that have been applied successfully to advance the understanding of disease-associated aspects of protein structure and function. The methods have usually not been exercised in a pipeline fashion, but rather in an integrative manner guided by the current biological questions. The various applications concern the identification and alignment of homologous sequences, the characterization of the primary protein architecture consisting of domains and binding motifs, the prediction of secondary and tertiary structures of proteins, the analysis of binding sites for proteins and other ligands, the structural localization of disease-associated sequence variants and the functional interpretation of their effects, and the exploration of protein interaction networks.

2.1.2 Functional Protein Architecture

Diverse databases and predictive methods were used to explore the functional architecture of proteins. The following sections describe the computational tools applied for sequence database searches, the delineation of protein domain boundaries, the identification of sequence motifs, the detection of putative transmembrane regions, and the computation of multiple sequence alignments.

Sequence database searches

Protein sequences were retrieved from NCBI (Wheeler *et al.*, 2004), Ensembl (Birney *et al.*, 2004), and UniProt (formerly SPTrEMBL) (Apweiler *et al.*, 2004) databases. To search for homologous sequences, we commonly used the standard BLAST and PSI-BLAST (E-value cut-off 0.005) programs (Altschul *et al.*, 1997). Alternatively, we sometimes used the FASTA search for full-length pairwise sequence alignments (Pearson, 2000) or the HMMER suite of programs for searches with our own HMMs

constructed from manually curated multiple sequence alignments (Eddy, 1996). To determine as yet unidentified mammalian orthologs of some human genes and their corresponding protein sequences, we used the genomic synteny views and other tools such as gene prediction programs, which are offered by the Ensembl browser (Birney *et al.*, 2004), the VISTA browser (Frazer *et al.*, 2004), or the UCSC genome browser (Karolchik *et al.*, 2003). Expressed sequence tags and other sequence fragments gave evidence for the expression of the identified genes displayed in the browsers.

Protein domain delineation

Protein domain architectures known so far, but often incomplete, were mainly obtained from the Pfam (Bateman *et al.*, 2004) and SMART (Letunic *et al.*, 2004) databases. Both databases are also contained in the NCBI conserved domain database CDD (Marchler-Bauer and Bryant, 2004). Pfam and SMART define protein domain families based on hidden Markov models (HMMs) derived from multiple sequence alignments. In contrast, domain searches in CDD use position-specific scoring matrices (PSSMs) (Gribskov *et al.*, 1987) derived from Pfam and SMART. Those CDD searches are significantly faster than the HMM-based searches in Pfam and SMART. Additional sources for domain delineations were ProDom (Servant *et al.*, 2002), which is also contained in Pfam, and InterPro (Mulder *et al.*, 2003), an integrated resource of major domain/motif databases such as Pfam, SMART, ProDom, and PROSITE. In contrast to the domain databases, PROSITE is a collection of biologically meaningful sequence motifs (Hulo *et al.*, 2004).

Sequence motif identification

Searches for sequence patterns of functional relevance for protein and ligand interactions or posttranslational modifications like glycosylation and phosphorylation (Yang, 2005) were performed mainly in the PROSITE database, sometimes in the eMOTIF database (Huang and Brutlag, 2001), and on four other prediction servers: PSORT II (Nakai and Horton, 1999) for potential signals of ER retention or nuclear localization, and ELM (Puntervoll *et al.*, 2003), iSPOT (Brannetti and Helmer-Citterich, 2003) and ScanSite (Obenauer *et al.*, 2003) for polyproline and other peptides that may constitute binding sites, for instance, of SH3 or WW domains (Zarrinpar *et al.*, 2003). A sequence profile for nuclear export signals was taken from NESbase (la Cour *et al.*, 2003). The SignalP server (Bendtsen *et al.*, 2004) and PSORT II were used to predict the existence of possible cleavage sites for N-terminal signal peptides. Such signal peptides were often supported by the actually incorrect prediction of a single transmembrane helix at the N-terminus due to hydrophobic amino acids. Sequence patterns contained in different proteins were also analyzed using the TEIRESIAS web service (Rigoutsos and Floratos, 1998). Repeats within the same protein sequence were discovered by means of the online tool RADAR (Heger and Holm, 2000).

Transmembrane region detection

To find transmembrane protein domains, we plotted the Kyte-Doolittle hydropathy index using the ExPASy ProtScale online service (Gasteiger *et al.*, 2005), whose positive hydrophobicity values over the common threshold 1.6 indicate transmembrane regions (Kyte and Doolittle, 1982). We also detected transmembrane regions using

more advanced prediction methods implemented in web servers such as DAS, HMMTOP2, MEMSAT, PHDhtm, PRED-TMR2, PSORT II, SOSUI, SPLIT, TMAP, TMHMM2, TMpred and TopPred2. The web links to all servers are listed in the supplementary online material of our publication (Albrecht *et al.*, 2003b). Also, the performance of seven of the used methods (DAS, HMMTOP2, PHDhtm, PRED-TMR2, SOSUI, TMHMM2, Top-Pred2) has been benchmarked comprehensively (Chen *et al.*, 2002). It reaches an accuracy between 80 and 99%, and at least one out of the seven methods is normally able to detect all transmembrane helices of some protein.

Sequence alignment computation

Multiple sequence alignments were assembled using CLUSTAL W (Chenna *et al.*, 2003) for closely related homologs and the programs T-COFFEE (Poirot *et al.*, 2003) or MUSCLE (Edgar, 2004) for sets of more diverse sequences. T-COFFEE appears to achieve a slightly better alignment quality than MUSCLE under difficult conditions of distant evolutionary relationships, but its runtime is quite long in comparison to MUSCLE. In many cases, the computed alignments could be improved manually by minor modifications, in particular, based on structure prediction results. The sequence alignments depicted in figures were prepared in the GeneDoc (Nicholas *et al.*, 1997) or SEAVIEW (Galtier *et al.*, 1996) editors and illustrated by the online web service ESPript (Gouet *et al.*, 2003).

2.1.3 Protein Structure Analysis

The analysis of the protein architecture based on the primary sequence alone already provides valuable functional information to interpret experimental results as well as to devise further experiments. In addition, globular protein domains adopt a folded three-dimensional (3D) structure, which provides a complementary structural view of biological processes. For instance, known protein structures and predicted structural models can help to verify suggested binding mechanisms of interacting proteins and ligands. Protein structures can also aid in the explanation of functional changes that may be caused by disease-associated mutations.

In the following sections, only computational tools are considered that were used in our own bioinformatics work. This concerns available protein structure databases, programs for 3D structure superpositions, methods for secondary and tertiary structure prediction and visualization, and online services for binding site analysis. Notably, if the structure of a protein has not been solved yet experimentally, 3D structure prediction methods are valuable in two respects. On the one hand, they can be applied as powerful fold recognition methods to detect distant evolutionary relationships between protein domain structures and functions if normal sequences database searches such as PSI-BLAST fail. On the other hand, the predicted 2D and 3D structure may give novel clues to the purpose of conserved sequence regions and may unveil potential binding sites.

Protein structure databases

Experimentally determined 3D protein structures were retrieved from the PDB database (Bourne *et al.*, 2004). The DSSP database (Kabsch and Sander, 1983) contains the secondary structure assignments for PDB structures. The SCOP database (Andreeva *et*

al., 2004) provides a hierarchical classification of PDB structures based on structural and evolutionary relationships of their 3D domain folds. Therefore, the SCOP database also assists in finding structural neighbors of the investigated proteins with possibly related functions.

3D structure superposition

The DALI server (Holm and Sander, 1993) was used to search the PDB for similar structures. Pairwise superpositions for structural comparisons and modeling purposes were computed by means of the program CE (Combinatorial Extension) (Shindyalov and Bourne, 1998) or DaliLite (Holm and Park, 2000). Alternative superpositions were occasionally computed using the ProSup server (Lackner *et al.*, 2000) or the online FATCAT method for flexible structural alignments (Ye and Godzik, 2003). The root mean square deviations (RMSDs) between protein structure backbones were always taken from the superposition results. Superpositions of short peptides were calculated in the DeepView/Swiss-PdbViewer (Guex and Peitsch, 1997).

Secondary structure analysis

To predict the secondary structure of proteins, we applied one or more of the following advanced methods through web servers: PROFsec (Rost and Eyrich, 2001), PSIPRED (McGuffin *et al.*, 2000), SAM-T99 (Karplus *et al.*, 1998), and SSpro2 (Pollastri *et al.*, 2002). All of them are based on neural networks or HMM techniques and reach an average three-state Q_3 prediction accuracy close to 80% (Koh *et al.*, 2003). We also formed consensus predictions by majority voting (see Chapter 3) using three selected secondary structure predictions (Albrecht *et al.*, 2003e). In addition, the NCOILS (Lupas *et al.*, 1991) and MultiCoil (Wolf *et al.*, 1997) online servers were applied to predict coiled coils in proteins. To identify intrinsically unstructured and disordered regions in proteins, we explored the prediction results returned by the online services DisEMBL (Linding *et al.*, 2003a), DISOPRED (Ward *et al.*, 2004), GlobPlot (Linding *et al.*, 2003b), NORSp (Liu and Rost, 2003) and PONDR (Romero *et al.*, 2001). The lack of pronounced secondary structure prediction of α-helices or β-strands in certain sequence regions was also indicative of putative intrinsic disorder.

Tertiary structure prediction

To obtain suggestions for globular 3D protein domain folds of amino acid sequences without a known structure, we usually investigated the results of all state-of-the-art fold recognition methods that are available via the online meta-server BioInfo.PL (Bujnicki *et al.*, 2001). This BioInfo.PL web server contacts a dozen other state-of-the-art prediction servers, the names of which are listed on the web site. This server is also coupled to the online 3D-Jury system that allows for the comparison and evaluation of the predicted 3D models in a consensus view (Ginalski and Rychlewski, 2003). We often compared these 3D predictions with the results of the in-house fold recognition server Arby (von Öhsen *et al.*, 2004). While the 3D-Jury system assesses the quality of the structure predictions based on a sophisticated scoring scheme (Ginalski *et al.*, 2003), Arby provides statistically derived confidence scores for protein fold predictions (Sommer *et al.*, 2002).

To model protein structures, we mostly submitted a sequence-structure alignment, which was commonly extracted from a manually improved structure-based multiple sequence alignment, to the 3D modeling server WHAT IF (Rodriguez *et al.*, 1998). Sometimes we applied the side chain placement program SCWRL (Canutescu *et al.*, 2003) to the resulting 3D model to increase the atomic model accuracy; the side chain conformation of amino acids that were identical in both aligned sequences was preserved. Alternatively, we used the completely automatic 3D-JIGSAW modeling server (Bates *et al.*, 2001) to obtain full-atom 3D models of protein sequences that are closely related to PDB domain structures. The protein structure images were drawn in the Accelrys Discovery Studio ViewerLite.

To create structural models of protein complexes, we superimposed modeled protein structures with crystallographically determined complexes (Albrecht *et al.*, 2003a). In one case, we also used the protein docking program HADDOCK (Dominguez *et al.*, 2003) to re-compute a published protein complex derived from NMR studies (Nicastro *et al.*, 2005).

Binding site analysis

To identify possible interatomic contacts of amino acids and ligands in 3D structures, we used the LPC and CSU online tools (Sobolev *et al.*, 1999) Strongly conserved columns of multiple sequence alignments were determined with the ConSurf online service (Glaser *et al.*, 2003) and mapped onto the corresponding known or predicted 3D structure. The electrostatic potential shown in protein surface pictures was generated using GRASP2 (Petrey and Honig, 2003). Both ConSurf and GRASP2 supported the visual localization of potential binding sites characterized by conserved or charged surface patches.

2.1.4 Protein Interaction Data

Novel high-throughput proteomics-based approaches have generated enormous amounts of protein-protein interaction data (Cusick *et al.*, 2005). They can now be mined for additional information on the functions and interrelationships of proteins (Bork *et al.*, 2004). The interaction network of disease-associated proteins and their homologs (orthologs or paralogs) can be explored to gain insight into their cellular roles and to direct further experiments. Most interaction data is currently available for yeast, but two large human networks have also been published recently (Rual *et al.*, 2005; Stelzl *et al.*, 2005).

Several interaction databases and visualization tools are available to facilitate bioinformatics work with the large data sets. The BIND (Bader et al., 2003), DIP (Salwinski and Eisenberg, 2003), GRID (Breitkreutz *et al.*, 2003a), MINT (Zanzoni et al., 2002), IntAct (Hermjakob et al., 2004), and SGD (Christie et al., 2004) resources as well as the GeneDB database (Hertz-Fowler et al., 2004) provided information on yeast proteins and their interactions for *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. To visualize and edit protein interaction networks, we used the Cytoscape platform (Shannon et al., 2003) and the Osprey software (Breitkreutz *et al.*, 2003b), the latter of which is linked to the GRID database.

2.2 Autoinflammatory Diseases

2.2.1 Biological and Clinical Background

Autoinflammation of human tissue appears to originate from a malfunctioning immune system. Human defense mechanisms against invasive pathogens can be broadly categorized as innate and adaptive immune systems, but this dichotomy has started to blur because of the recent discovery of links between them (Flajnik and Du Pasquier, 2004; Hoebe *et al.*, 2004). Extensive knowledge has already been accumulated on the sophisticated adaptive immune system, which is mainly mediated by B and T lymphocytes. In contrast, research on innate immunity is quite a recent development and focuses mainly on rapid responses induced by surveillance receptors of pathogens. These guard proteins recognize a large variety of distinct pathogen-associated molecular patterns (PAMPs) such as bacterial cell-wall components and viral carbohydrates or nucleotides.

Three important groups of PAMP receptors are as follows: membrane-bound or secreted C-type lectin-like receptors (CLRs) (McGreal *et al.*, 2004; van Kooyk *et al.*, 2004; Cambi *et al.*, 2005; McGreal *et al.*, 2005), cytoplasmic NACHT-LRR domain receptors (NLRs) (Inohara *et al.*, 2005; Kufer *et al.*, 2005; Martinon and Tschopp, 2005; Ting and Davis, 2005), and transmembrane Toll-like receptors (TLRs) (Dunne and O'Neill, 2005; Hopkins and Sriskandan, 2005; Liew *et al.*, 2005; Takeda and Akira, 2005). NLRs are the members of the so-called CATERPILLER protein family, which can be further subdivided into evolutionarily related neuronal apoptosis-inhibiting proteins (NAIPs), NACHT, LRR, and PYD domain-containing proteins (NALPs, also known as PYPAFs), and nucleotide-binding oligomerization domain-containing proteins (NODs).

Well-known CLRs are the dendritic cell-specific DC-SIGN family receptors for envelope glycoproteins of HIV-1 (human immunodeficiency virus 1) and HCV (hepatitis C virus) (Cambi and Figdor, 2003; van Kooyk and Geijtenbeek, 2003). In contrast, the ligand specificity of many NLRs are not known yet, but NALP3 and NOD1/2 have been shown to work as intracellular sensors of bacterial peptidoglycan (PGN) (Boneca, 2005; McDonald *et al.*, 2005). However, most of the 11 TLRs have already been studied intensively, each of which detect specific microbial molecules such as lipoprotein, lipopolysaccharide, flagellin, zymosan, and DNA/RNA that are derived from pathogens including bacteria, fungi, protozoa, and viruses (Akira and Takeda, 2004; O'Neill, 2004). Intriguingly, the surveillance function of mammalian PAMP receptors is also exercised by similar proteins found in the innate immune systems of flies and plants (Dangl and Jones, 2001; Girardin *et al.*, 2002; Ausubel, 2005).

CLRs, NLRs, and TLRs form part of complex signaling pathways with intricate cross-talk (Athman and Philpott, 2004; Geijtenbeek *et al.*, 2004; Hopkins and Sriskandan, 2005). After the recognition of certain PAMPs, the receptors trigger specific gene expression patterns by the activation of important transcription factors such as NF-κB and interferon-regulatory factors (Bonizzi and Karin, 2004; Moynagh, 2005). Responsive genes range from proinflammatory cytokines and interferons to co-stimulatory molecules. They mount an immune response resulting in the removal and

destruction of the invading pathogen. Therefore, the impairment of essential signaling cascades by mutant proteins can lead to the dysregulation of human immunity, causing acute or chronic diseases of autoimmunity or immunodeficiency (Beutler, 2004; Ting and Davis, 2005). However, the modification of innate immune responses by therapeutic targeting of the implicated cellular mechanisms may provide new opportunities for clinical treatment of patients (Karin *et al.*, 2004; Ulevitch, 2004).

The NLR family members NALP3 and NOD2, also known as CIAS1/PYPAF1 and CARD15, respectively, are the research focus of our medical cooperation partners because they have been associated with several autoimmune disorders. Sequence variants in both proteins are causative of inherited autoinflammatory diseases with clinically distinct phenotypes, but similar inflammatory pathophysiology (Albrecht *et al.*, 2003a; Van Duist *et al.*, 2005). In addition, some NOD2 variants also confer susceptibility to Crohn's disease (CD), a chronic inflammatory bowel disease with a high lifetime prevalence of up to 0.15% in Western Europe and North America (Macdonald and Monteleone, 2005; Schreiber *et al.*, 2005). CD belongs to a group of complex, polygenic, barrier disorders such as asthma, atopic eczema, and sarcoidosis, which affect either mucosal surfaces or the skin and exhibit a multifactorial etiology involving environmental factors. The frequent concordance in monozygotic twins, which is not seen in dizygotic twins, points to the strong contribution of genetic susceptibility to the overall risk for CD.

In the gut mucosa, NOD2 like NALP3 senses muropeptides, which are cell wall components of pathogenic bacteria, and appears to be responsible for the maintenance of epithelial barrier integrity and the immune defense in interplay with TLRs (Yuan and Walker, 2004; Mueller and Podolsky, 2005). On the molecular level, it has been discovered that NALP3 and NOD2 assemble into large signaling complexes named inflammasome and noddosome, respectively, after the recognition of microbial products. Such complexes activate inflammatory caspases and are assumed to function similarly to the apoptosome of APAF-1, the apoptotic protease-activating factor 1 (Martinon and Tschopp, 2004; Riedl *et al.*, 2005; Yu *et al.*, 2005). However, many details on the exact biological roles of NALP3, NOD2 and other NLRs and their participation in signal transduction processes within the immune system are still unclear (Martinon and Tschopp, 2004; Eckmann and Karin, 2005; Murray, 2005; Strober, 2006).

Besides PAMP receptors, we also investigated other players without as yet well-characterized function in the innate and adaptive immune systems: the interferon-inducible p200 (also known as IFI-200/HIN-200) family of proteins and butyrophilin-like members of the immunoglobulin superfamily. The former regulate cell growth and differentiation, and confer resistance to the development of tumors and virus infections (Asefa *et al.*, 2004). The latter are closely related to B7 immune-regulatory ligands of antigen-presenting cells (Greenwald *et al.*, 2005) and include the co-stimulatory receptor BTNL2 on the cell surface, which has been associated with the multisystemic immune disorder sarcoidosis (Valentonyte *et al.*, 2005). The clinical presentation of sarcoidosis varies in patients, but its inflammatory manifestation is predominantly the lung. Sarcoidosis probably results from disproportionate immune responses to some airborne antigen (Rybicki *et al.*, 2005).

2.2.2 Bioinformatics Findings

To analyze the locations of sequence variations associated with autoinflammatory diseases, we extensively surveyed the protein domain architecture of CATERPILLER family members such as the NLRs NALP3/PYPAF1 and NOD2/CARD15 (Figure 1). We used large sequence alignments of NLRs to identify homologous sequence positions and variants (Figure 2) and their functional relevance (Albrecht *et al.*, 2003a, 2003d; Schreiber *et al.*, 2005; Albrecht and Takken, 2006). Our bioinformatics work also supported the classification of newly discovered sequence variants of patients as putative disease-causing mutations. This was based on the close localization of the mutations near functionally relevant amino acids linked to other inherited autoinflammatory disorders (Figure 2 and Figure 3) (Van Duist *et al.*, 2005).

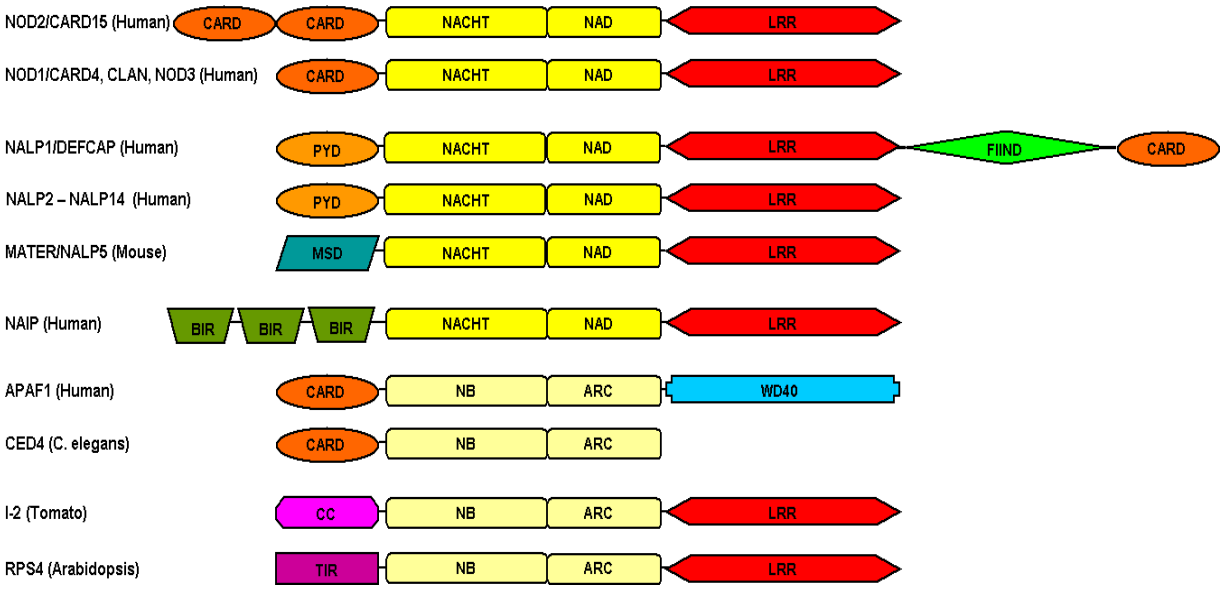


Figure 1. Protein domain architectures of selected CATERPILLER-related gene products from different eukaryotes (Albrecht *et al.*, 2003a; Schreiber *et al.*, 2005). The N-terminal effector-binding CARD (caspase recruitment) and PYD (pyrin) domains of NLRs are evolutionarily related and adopt the same structural fold. The CARD and PYD counterparts in plants are coiled coils (CC) and Toll/interleukin-1 receptor domains (TIR) of numerous disease resistance proteins (R proteins) such as I-2 (confers resistance to race 2 isolates of *Fusarium oxysporum*) and RPS4 (confers resistance to *Pseudomonas syringae*). The central nucleotide-binding domains are designated NACHT and NB domains and belong to a recently defined family of P-loop NTPases, which is distantly related to AAA+ ATPases and named STAND domain family (Hanson and Whiteheart, 2005). The NTPase domains are proposed to work as switches regulating signal transduction by conformational changes (Albrecht *et al.*, 2003a; Leipe *et al.*, 2004). The structural extensions of the NACHT and NB domains are homologous, named NAD and ARC, respectively, and consist of three subdomains NAD1-3 and ARC1-3 (Albrecht and Takken, 2006). The number of leucine-rich repeats (LRRs) of the C-terminal sensor domain varies within the CATERPILLER family, with NALP10 containing the least number of LRRs.

Furthermore, we recognized the identity of a protein named AVR reported first in 1995 with the cytoplasmic NLR protein NALP6/PYPAF5 described in 2002 (Albrecht *et al.*, 2003b). However, our study also showed how the incorrect application of bioinformatics methods, in this case, the ignorance of the recommended threshold for the Kyte-Doolittle transmembrane prediction method (Figure 4), led to wrong results on AVR in the original report published in *Nature Medicine* (Ruiz-Opazo *et al.*, 1995). This publication falsely assumed the discovery of a novel membrane-anchored angiotensin II and vasopressin receptor, although NALP6 is clearly cytosolic.

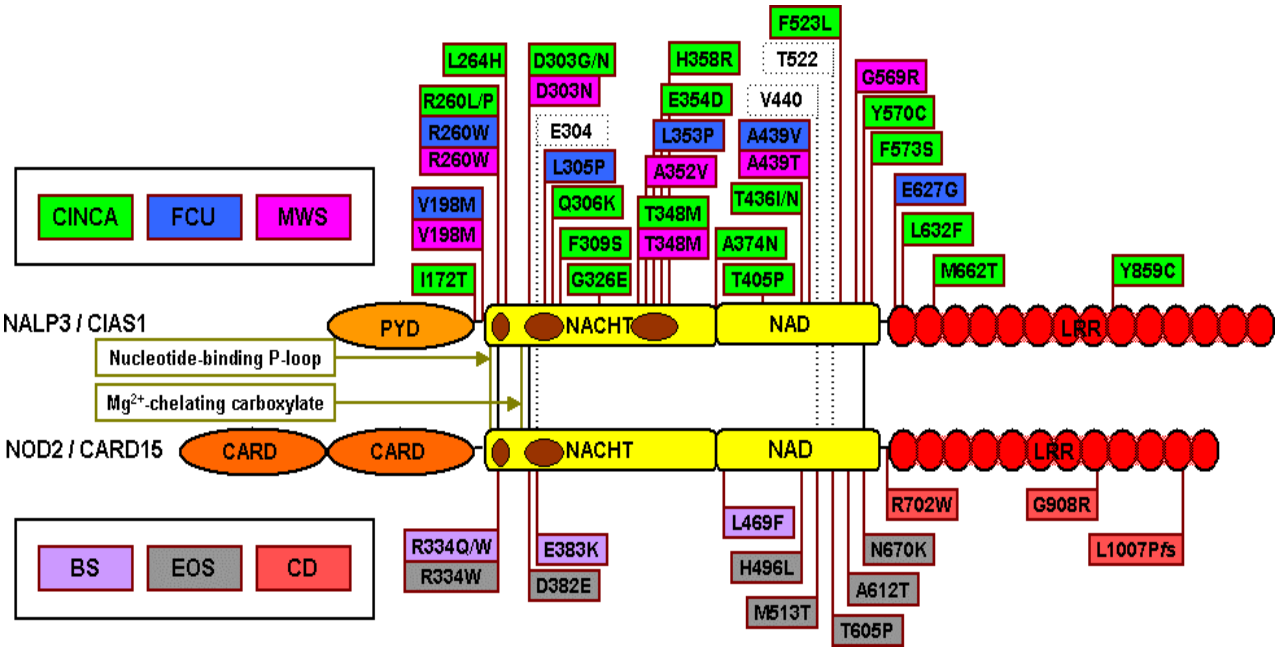


Figure 2. Sequence variations in the homologous NLR family members NALP3 and NOD2 contribute to protein plasticity and give rise to various autoinflammatory diseases (Albrecht *et al.*, 2003a; Schreiber *et al.*, 2005). Crohn's disease-associated sequence variants are mainly found within the LRR domain, whereas mutations linked to other inflammatory diseases are predominantly situated in the nucleotide-binding domain consisting of the NACHT and NAD subdomains. This distinct domain localization might partially explain phenotypic differences between the disorders. Interestingly, several mutations in NALP3 and NOD2 are located at equivalent sequence positions (black vertical lines), some of which form mutational hot spots (brown ovals) near the binding site of the magnesium-nucleotide complex in the NACHT domain. The annotated autoinflammatory diseases besides Crohn's disease (CD) are as follows: BS, Blau syndrome (also known as ACUG, arthrocutaneouveal granulomatosis); CINCA, chronic infantile neurological cutaneous and articular syndrome (also known as NOMID, neonatal-onset multisystem inflammatory disease); EOS, early-onset sarcoidosis; FCU, familial cold urticaria (also known as FCAS, familial cold autoinflammatory syndrome); MWS, Muckle-Wells syndrome.

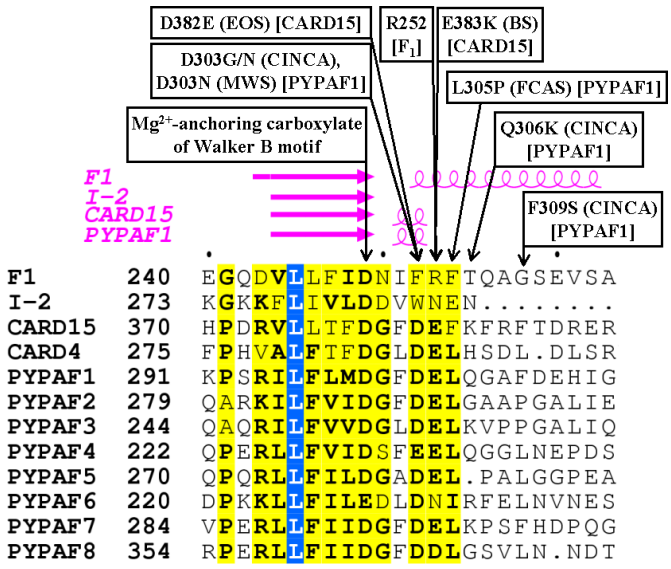


Figure 3. Structure-based multiple sequence alignment of the nucleotide-binding site termed Walker B motif (Van Duist *et al.*, 2005). It contains the Mg^{2+}-anchoring aspartate in the NACHT domains of human NOD2/CARD15, NOD1/CARD4, NALPs/PYPAFs, in the ATPase domain of tomato disease resistance protein I-2, and in the β-subunit of bovine F_1-ATPase. The known secondary structure of the F_1-ATPase and the corresponding consensus predictions for I-2, CARD15, and PYPAF1 are depicted in the upper pink part (α-helices as curled lines and β-strands as horizontal arrows). Alignment columns in which more than 60% of all residues are physicochemically equivalent are shown in yellow boxes. A strictly conserved leucine is highlighted in blue. Text labels point to functionally relevant residues and disease-associated missense variants. Importantly, the new mutation E383K found with an Italian patient of Blau syndrome (BS) is next to other already known variants.

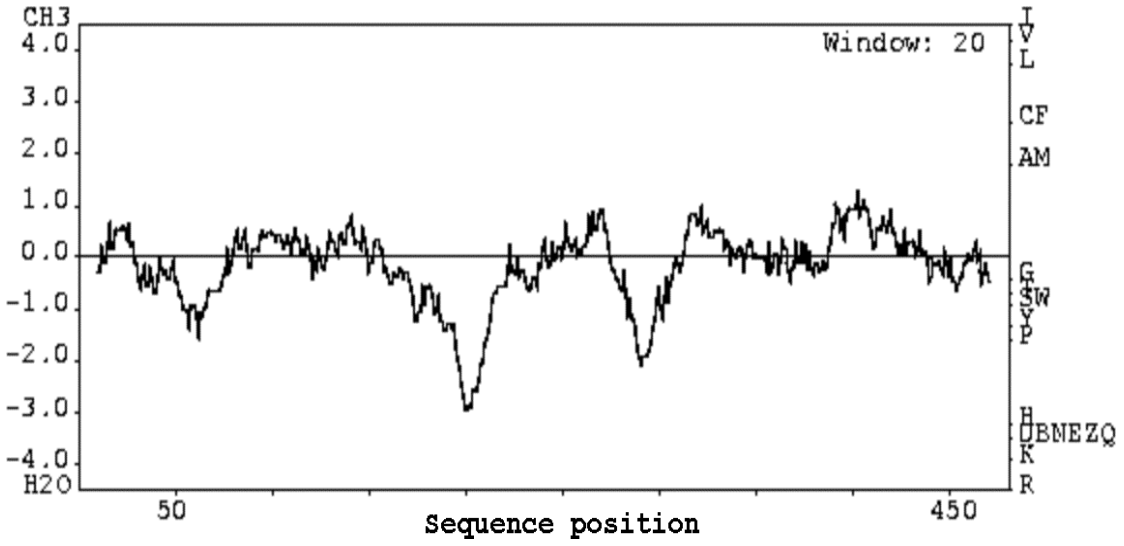


Figure 4. Plot of the Kyte-Doolittle hydropathy index (using an averaging-window size of 20 residues) for AVR, constituting the LRR domain of a truncated NALP6/ PYPAF5 protein. It does not indicate any transmembrane regions because the hydropathy values remain below the recommended threshold 1.6 (Albrecht *et al.*, 2003b).

To visualize and study conserved and functionally relevant sequence regions and residues in 3D, we constructed various structural models of domains contained in disease-relevant proteins. Using a structure-based multiple sequence alignment of homologous proteins, we mapped NALP3 and NOD2 sequence variants associated with different autoinflammatory diseases into 3D domain models (Figure 5). In particular, our analyses led to the intriguing hypothesis that nucleotide binding of NALP3 and NOD2 may be impaired by sequence mutations, causing a constitutively active protein inducing inflammatory immune responses (Albrecht *et al.*, 2003a, 2003d; Van Duist *et al.*, 2005; Albrecht and Takken, 2006).

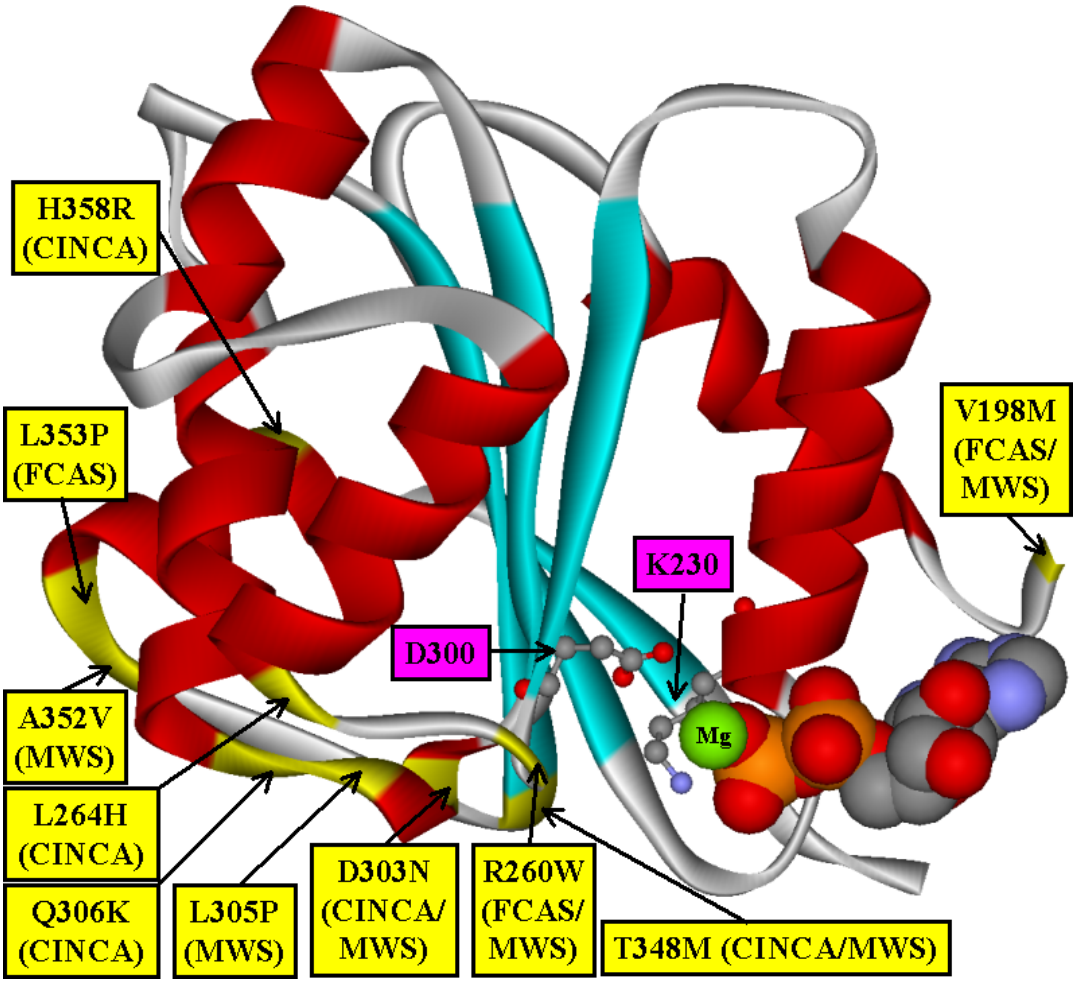


Figure 5. 3D structure model of the nucleotide-binding NACHT domain of NALP3/CIAS1/PYPAF1 based on the AAA+ ATPase Cdc6 from *Pyrobaculum aerophilum* (PDB identifier 1fnn, chain A) (Albrecht *et al.*, 2003a). While α-helices are colored in red and β-strands in blue, locations of selected sequence variants associated with autoinflammatory diseases are marked in yellow. Many of them are found near the C-termini of the β-strands. Other functional residues interacting with the bound magnesium-nucleotide complex are indicated in pink: the phosphate-binding lysine K230 and the Mg^{2+}-anchoring aspartate D300.

In another study, we performed a thorough bioinformatics analysis of IFI-200 proteins (Albrecht *et al.*, 2005a). IFI-200 family members are thought to exert their biological effects by modulation of the transcriptional activities of numerous factors and interaction with other proteins through the C-terminal HIN domains. The HIN domain structure and function had remained obscure, but our multiple sequence alignment and the application of fold recognition methods revealed that the HIN domain consists of two consecutive OB domains (Figure 6). Therefore, this structural model of a DNA-binding HIN domain afforded long-sought interpretations for many previous experimental observations of IFI-200 proteins working as transcriptional regulators.

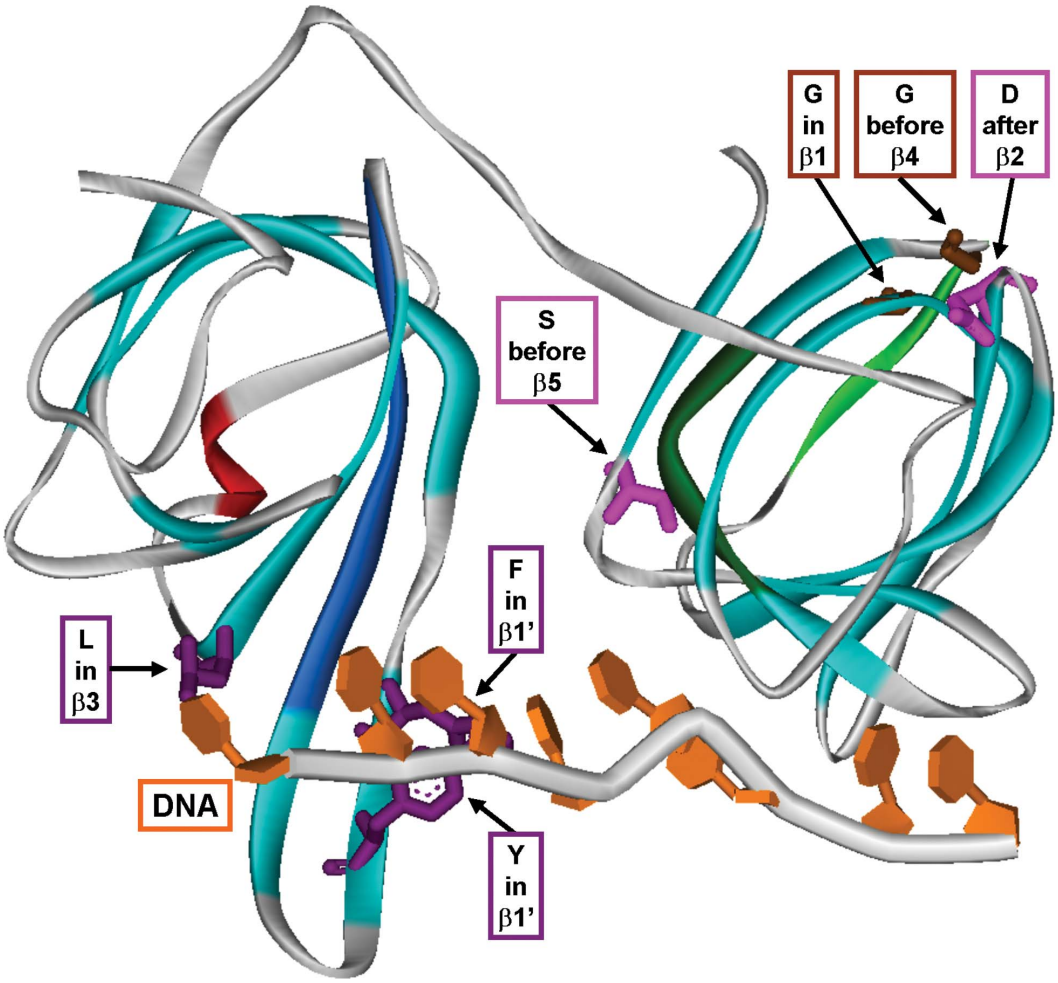


Figure 6. 3D structure model of the HIN-N and HIN-C domains (shown on the left and right, respectively) of the IFI-200 family member AIM2 based on the DNA-binding OB domain structures DBD-A and DBD-B (PDB identifier 1jmc, chain A) of the human replication protein A (Albrecht *et al.*, 2005a). All β-strands are colored in cyan, and the α-helix after the β3-strand is red. Positions of conserved sequence motifs associated with previous experimental observations are indicated: MFHATVAT in the β2-strand (blue), IxCxE in the loop preceding the β4-strand (light green), and LxCxR corresponding to LxCxE of IFI202 and IFI204 in the β4-strand (dark green). Text labels annotate functionally and structurally relevant residues.

In projects conducted within large research networks, we could also contribute to functional analysis of gene expression profiles (Costello *et al.*, 2005) and of another potential susceptibility gene termed DLG5 for inflammatory bowel disease (IBD) (Stoll *et al.*, 2004; Newman and Siminovitch, 2005), an as yet uncharacterized membrane-associate guanylate (Funke *et al.*, 2005). We delineated the protein domain architecture of DLG5 and mapped genetic variations into conserved sequence regions. However, we could find only weak bioinformatics support for the IBD association of DLG5 sequence variants (Stoll *et al.*, 2004). This low support was in contrast to the strong evidence based on structural predictions that we could obtain for the BTNL2 protein truncation (Figure 7) causing sarcoidosis (Valentonyte *et al.*, 2005). Therefore, our bioinformatics observations agreed well with the fact that an IBD association of DLG5 could not be replicated by other medical research groups (Schreiber *et al.*, 2005), whereas the BTNL2 link to sarcoidosis could recently be reproduced by another group (Rybicki *et al.*, 2005). Our cooperation partners could also confirm the predicted truncation of BTNL2 experimentally. It causes the loss of the C-terminal transmembrane region of BTNL2, which is essential for anchoring the N-terminal, extracellular immunoglobulin-like, domains of BTNL2 into the cell membrane (Valentonyte *et al.*, 2005).

2.3 Neurodegenerative Disorders

2.3.1 Biological and Clinical Background

Neurodegeneration is often caused by death of specific neuron populations in the brain after formation of intracellular protein aggregates. Some biological and clinical characteristics of the investigated diseases are given in the following. Two autosomal-dominant hereditary neurodegenerative disorders are spinocerebellar ataxia types 2 (SCA2) and type 3 (SCA3, also known as Machado-Joseph disease) (Kawaguchi *et al.*, 1994; Pulst *et al.*, 1996). Both belong to a heterogeneous group of trinucleotide repeat disorders, which includes Huntington's disease (HD), dentatorubral-pallidoluysian atrophy (DRPLA), and other spinocerebellar ataxia types such as SCA1, SCA6, SCA7 and SCA17 (Zoghbi and Orr, 2000; Gatchel and Zoghbi, 2005; Manto, 2005). They result from progressive neurodegenerative processes, which affect the cerebellum, brainstem and spinal cord (Schols *et al.*, 2004). Clinical main features are ataxia and dementia, which can also resemble parkinsonism (Taroni and DiDonato, 2004). The age of patients at the onset of SCA2 and SCA3 lies in the third to fourth decade (Margolis, 2002). The neurodegenerative disorders share common phenotypic features, in particular, toxic accumulation of mutant misfolded proteins in affected neurons and cellular degeneration causing apoptosis (Soto, 2003; Ross and Poirier, 2004). In contrast, the expression of the disease-associated genes occurs in a great variety of tissues and is not restricted to neuronal cells.

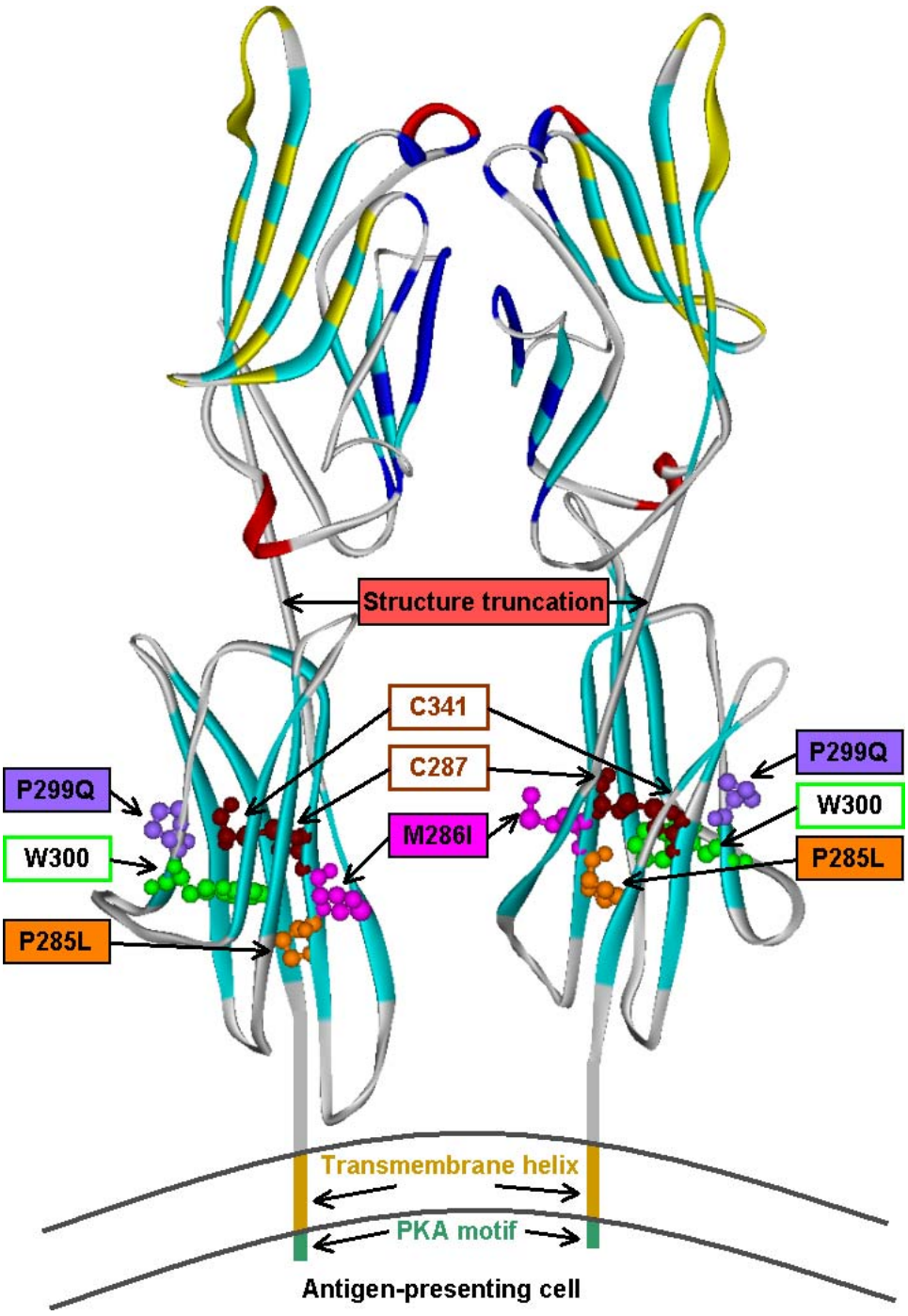


Figure 7. 3D structure model of the second IgV domain (top) and the following IgC domain (bottom) of the BTNL2 homodimer based on the B7-1 template structure (PDB identifier 1i8l) (Valentonyte *et al.*, 2005). A transmembrane helix near the C-terminus anchors the extracellular domains into the cell membrane. The locations of the sarcoidosis-associated C-terminal truncation (red box) and several other sequence variants found in BTNL2 are indicated. An adjacent disulfide bond between C287 and C341 is depicted in brown.

The SCA2 and SCA3/MJD genes have been mapped to chromosomes 12q24.1 and 14q32.1, respectively (Kawaguchi *et al.*, 1994; Pulst *et al.*, 1996). The common underlying genetic basis of the diseases SCA2 and SCA3 is the expansion of a CAG repeat region beyond a certain threshold (Pearson *et al.*, 2005). These CAG repeats encode a polyglutamine (polyQ) tract in the respective proteins ataxin-2 and ataxin-3 (Everett and Wood, 2004). The polyQ stretch in ataxin-2 lies near the N-terminus at the end of exon 1, but the polyQ region of ataxin-3 is contained in exon 10 close to the C-terminus (Albrecht *et al.*, 2004). While ataxin-2 is located predominantly in the cytoplasm, ataxin-3 is found in both the nucleus and the cytoplasm of cells.

Each polyglutamine expansion protein is causative of a specific neurodegenerative disorder, whereas different mutant proteins, which are evolutionarily unrelated and differ in their cellular function, predispose to Parkinson's disease (PD) (Bonifati *et al.*, 2004). The majority of PD cases are sporadic with a complex etiology due to interactions between environmental factors and the individual genetic constitution (Huang *et al.*, 2004). However, the identification of point mutations in single genes responsible for rare familial forms of PD provides important insights into the underlying disease mechanisms (Moore *et al.*, 2005). One recently discovered PD-associated gene product is the receptor-interacting protein RIP7, also known as leucine-rich repeat kinase LRRK2 or dardarin (Gasser, 2005). RIP7 encodes a large protein with multiple domains containing several mutations found in familial PD cases (Albrecht, 2005; Meylan and Tschopp, 2005). Phenotypically, PD is characterized by the loss of dopaminergic neurons primarily in the substantia nigra and the presence of cytoplasmic protein inclusions known as Lewy bodies (von Bohlen und Halbach *et al.*, 2004). Clinical manifestations include motor abnormalities, autonomic disturbances, psychiatric depressions, and cognitive impairments (Greenamyre and Hastings, 2004). However, the neuropathology and the age of onset is very variable even within the same family (Brice, 2005).

2.3.2 Bioinformatics Findings

Concerning the neurodegenerative disorders, we primarily investigated conserved domains and sequence motifs of proteins related to the disorders ataxia type 2 and type 3. We characterized the protein architectures of ataxin-2 (Figure 8), which contains two proline-rich sequence motifs (Figure 9) and one Lsm domain (Albrecht *et al.*, 2004; Ralser *et al.*, 2005a, 2005b). Both proline-rich motifs were predicted with high confidence scores as characteristic of SH3 domain binding (Zarrinpar *et al.*, 2003). Later on, they could be experimentally verified as SH3 binding sites of endophilins (Landgraf *et al.*, 2004; Ralser *et al.*, 2005b). During our work on ataxin-2, we also discovered novel Lsm domain proteins Lsm12-16 (Figure 10) related to ataxin-2 and as yet uncharacterized methyltransferases (Albrecht and Lengauer, 2004a). Furthermore, we studied the domain structure of ataxin-3 (Figure 11) extensively (Albrecht *et al.*, 2003c; Albrecht *et al.*, 2004).

Lsm domains occur in a number of vital RNA-processing proteins conserved in many organisms (Khusial *et al.*, 2005; Wilusz and Wilusz, 2005). Since several Lsm domain structures have been determined experimentally, we assembled a structure-

based multiple sequence alignment for the Lsm domain using ataxin-2 homologs inclusive the yeast homolog Pbp1 (Figure 12). Based on that, we could explore the functional implications of conserved residues in RNA binding and complex formation. Moreover, our detailed analysis of Lsm14-16 homologs revealed a conspicuous, highly conserved, sequence motif consisting of aspartates and phenylalanines near the C-terminus (Figure 13). Thus, it may be worthwhile mutating those amino acids experimentally to find clues on a molecular function.

In addition, we built a 3D structure model for the RNA-binding Lsm domain of ataxin-2 (Figure 14) to gain insight into the putative RNA-binding mode. Similarly, we used a structure-based multiple sequence alignment of the Josephin domain of ataxin-3 homologs to derive an illustrative 3D model of this domain (Figure 15). This was also shown on the front cover of the European Journal of Biochemistry (now FEBS Journal) issue containing the corresponding publication (Albrecht *et al.*, 2004). The ataxin-3 model enabled the identification of specific amino acids in the Josephin domain that are assumed to be involved in protease activity necessary for de-ubiquitinylation. Generally, protein ubiquitinylation (also known as ubiquitination) is the reversible process of conjugating ubiquitin molecules to proteins (Ciechanover and Brundin, 2003). The removal of ubiquitin can be mediated by various de-ubiquitinating enzymes (DUBs), which are cysteine proteases like ataxin-3 or metalloproteases (Guterman and Glickman, 2004; Soboleva and Baker, 2004; Nijman *et al.*, 2005).

Concerning Parkinson's disease (PD), a homology model of the protein kinase RIP7/LRRK2 (Figure 16) revealed that the two adjacent PD-associated mutations G2019S and I2020T presumably impair kinase activity because they are contained in the well-studied kinase activation segment (Nolen *et al.*, 2004; Albrecht, 2005). Recently, this hypothesis was confirmed experimentally using the corresponding mutants of RIP7/LRRK2 or its human homolog RIP6/LRRK1 (West *et al.*, 2005; Gloeckner *et al.*, 2006; Korr *et al.*, 2006).

Regarding interaction networks, it has already been shown for the Huntington's and Parkinson's diseases that the outcome of yeast experiments can also help in elucidating human disease processes (Krobitsch and Lindquist, 2000; Outeiro and Lindquist, 2003). Therefore, we studied direct and indirect interaction partners of the yeast homolog Pbp1 of ataxin-2 in addition to the 3D structure model of its Lsm domain (Ralser *et al.*, 2005a). The interaction network around Pbp1 (Figure 17), which includes further Lsm domain proteins such as Lsm12 and Lsm13 and the RNA-helicase Dhh1, lends further support to the hypothesis that ataxin-2 is involved in RNA metabolism and may bind RNA within its Lsm domain. This assumption was additionally corroborated by experimental evidence that ataxin-2 and its yeast homolog Pbp1 interact with the poly(A)-binding protein. The relevant binding motif predicted in ataxin-2 and named PAM2 (Albrecht and Lengauer, 2004b) could be confirmed experimentally by our collaboration partners (Ralser *et al.*, 2005a).

In general, our bioinformatics findings provided a useful basis for the evaluation and prioritization of experiments. More details on biological insights obtained during these comprehensive studies and on complementary lab experiments performed by our cooperation partners can be found in the original publications (Albrecht *et al.*, 2003c; Albrecht *et al.*, 2004; Albrecht and Lengauer, 2004a, 2004b; Albrecht, 2005; Ralser *et*

al., 2005a, 2005b). Interestingly, shortly after our publication on the characterization of novel Lsm12-16 proteins (Albrecht and Lengauer, 2004a), identical results supporting our work were reported independently by another research group (Anantharaman and Aravind, 2004). Recently, both Lsm14 in *Caenorhabditis elegans* (known as CAR-1) and Lsm16 in yeast (known as EDC3) were shown to be involved in RNA processing, confirming our function predictions (Badis *et al.*, 2004; Kshirsagar and Parker, 2004; Audhya *et al.*, 2005; Boag *et al.*, 2005; Squirrell *et al.*, 2005).

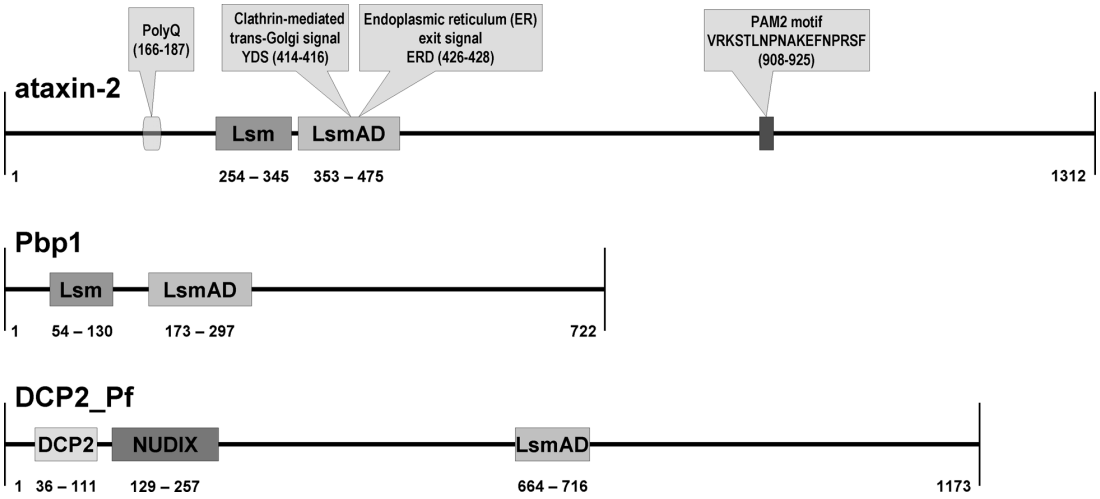


Figure 8. Protein architectures of human ataxin-2, its yeast homolog Pbp1, and the *Plasmodium falciparum* homolog PF13_0048 of the mRNA-decapping enzyme DCP2 (Albrecht *et al.*, 2004; Ralser *et al.*, 2005a). The appearance of the as yet uncharacterized LsmAD domain in DCP2 suggests a functional relationship of LsmAD to the RNA-binding Lsm domain.

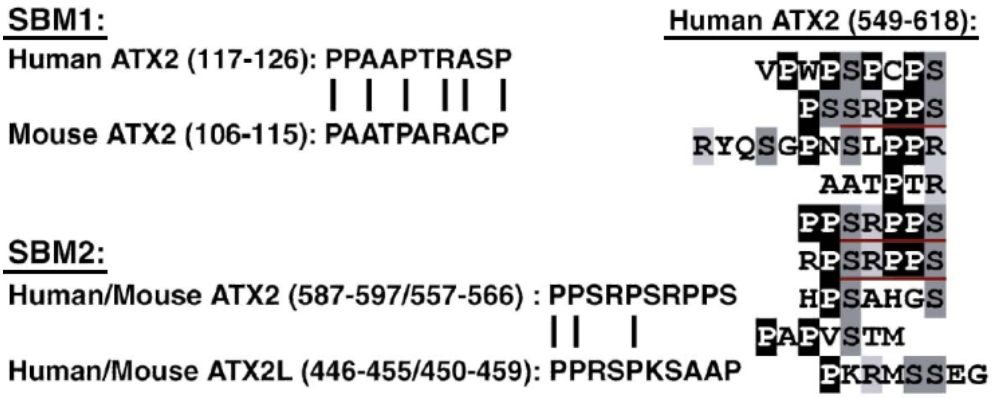


Figure 9. Two conserved proline-rich sequence motifs (SBM1 and SBM2) are shown on the left in human and mouse ataxin-2 (ATX2) as well as in human and mouse ATX2 and ATX2L paralogs (Ralser *et al.*, 2005b). Such motifs are characteristic of SH3 domain binding sites (Zarrinpar *et al.*, 2003). Three identical SRPPS motifs in sequence repeats containing SBM2 of human ATX2 are underlined in red. Frequently occurring prolines, serines, and arginines are highlighted in the motifs.

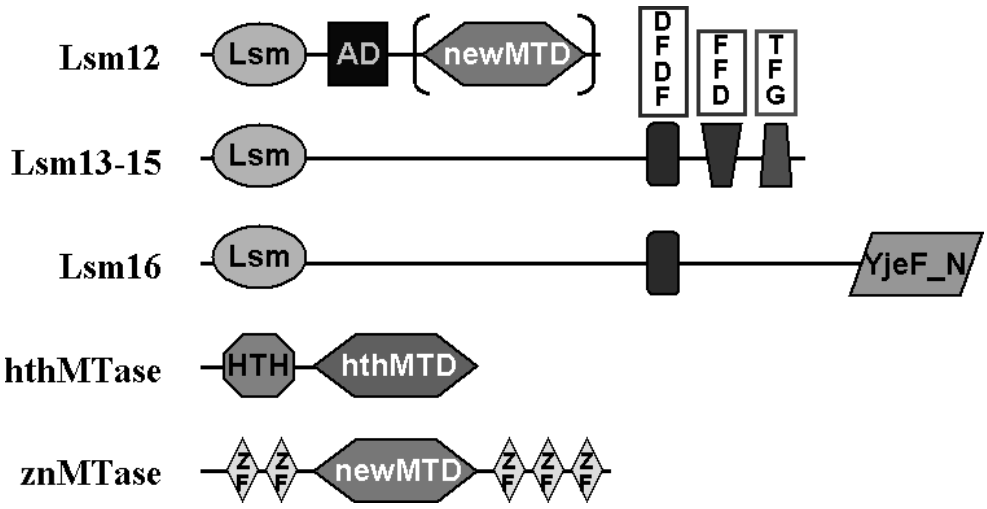


Figure 10. Protein architectures of novel Lsm domain proteins and methyltransferases (MTases) (Albrecht and Lengauer, 2004a). The latter contain not only the catalytic domain (MTD), but also zinc fingers (ZF) and helix-turn-helix (HTH) motifs, both of which may contribute to RNA binding via the Lsm domain.

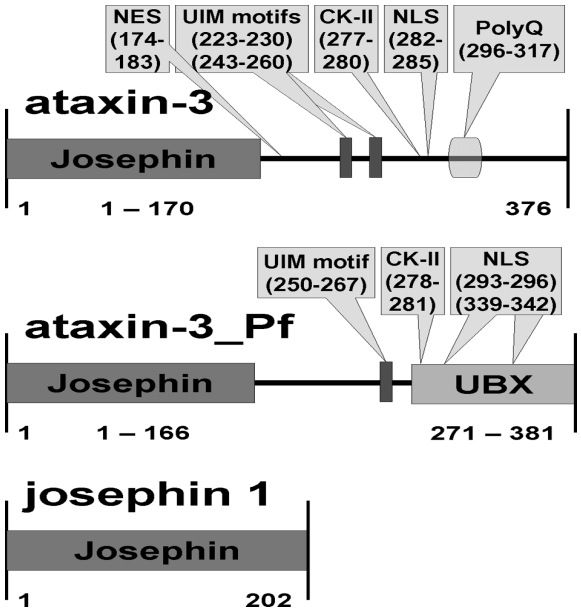


Figure 11. Protein architectures of human ataxin-3, its *Plasmodium falciparum* homolog PFL1295w (ataxin-3_Pf), and human josephin 1 (Albrecht *et al.*, 2004). The Josephin domain has de-ubiquitinating activity, UBX is a ubiquitin-like domain of unknown function, UIMs are ubiquitin-interaction motifs, CK-II motifs are putative phosphorylation sites of casein kinase II, NLS are potential nuclear localization motifs, and NES is proposed to be a nuclear export signal.

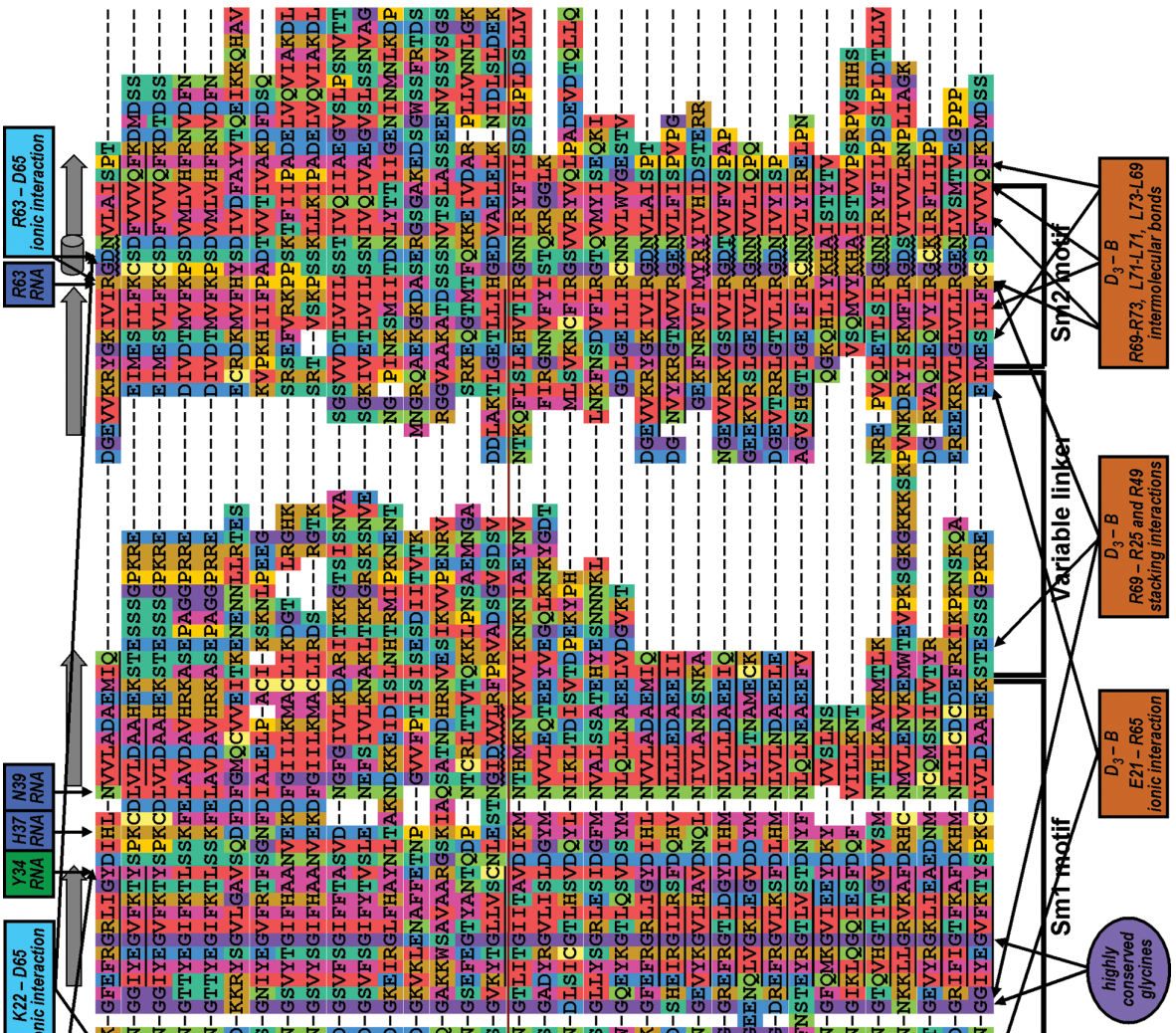


Figure 12. Structure-based multiple sequence alignment of the Lsm domains of ataxin-2 homologs including the yeast homolog Pbp1 (upper part) with Sm and Sm-like proteins (lower part) (Albrecht *et al.*, 2004). *Legend is continued on next page.*

Figure 12 *(legend continued)***.** The known DSSP secondary structure assignment of the Sm1 protein from *Pyrococcus abyssi* is shown at the top of the alignment (cylinder for α-helix, arrow for β-strand), and the amino acid sequences of crystallographically determined PDB structures of Lsm domains are underlined accordingly (curled line for α-helix, straight line for β-strand). The secondary structure predictions for the Lsm domains of ataxin-2 and Pbp1 are also given. Physicochemically similar amino acids are colored identically. Highly conserved glycines are indicated that are characteristic of Lsm domains and form β-sheet stabilizing hydrogen bonds. In the upper part, blue text boxes point to functionally relevant amino acids forming an internal binding site for uridine heptamers bound to Sm1 from *Pyrococcus abyssi*, while green text boxes mark amino acids of the external RNA binding site. In the lower part, brown text labels annotate how the dimerization of the snRNPs D_3 and B is stabilized by intermolecular interactions.

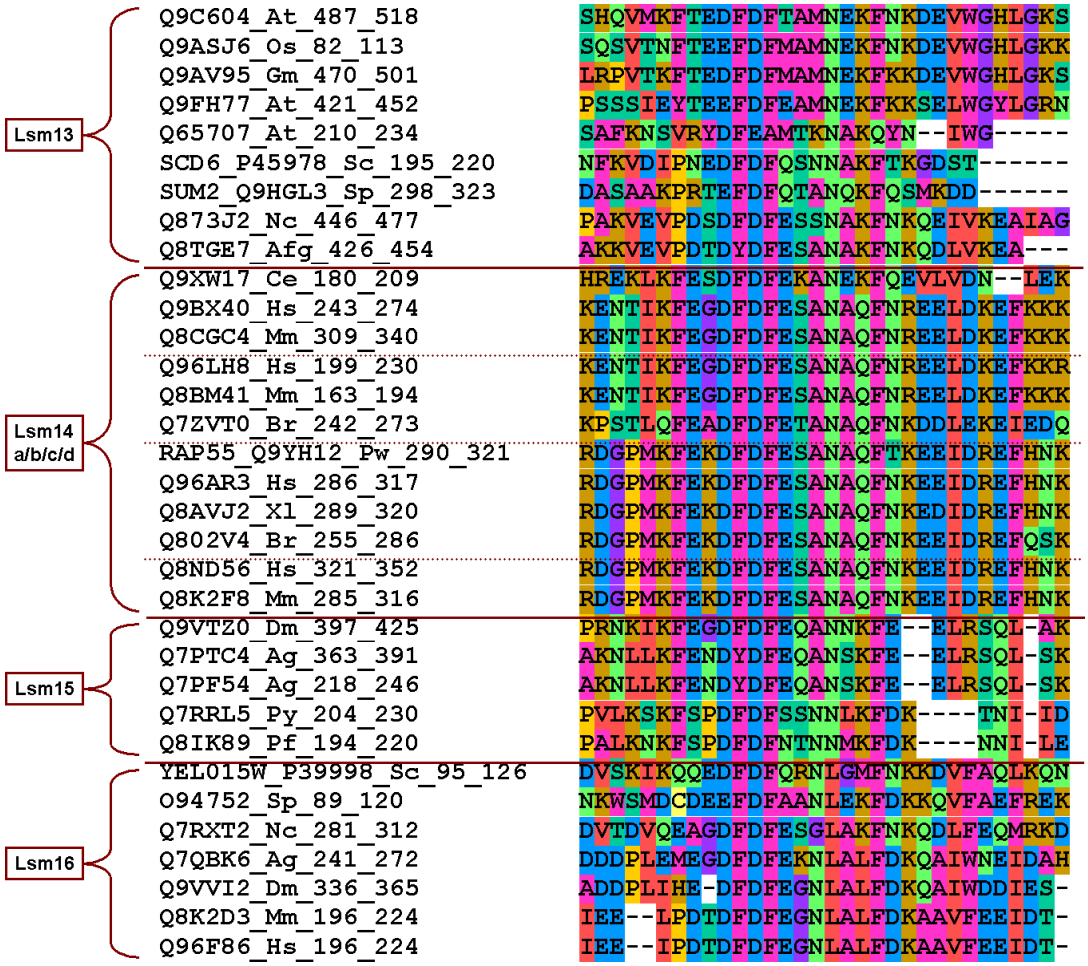


Figure 13. Multiple sequence alignment of a striking sequence region containing a strongly conserved DFDF box of unknown function near the C-terminus of Lsm13-16 proteins (Albrecht and Lengauer, 2004a). The Lsm14 group is exemplarily subdivided into four putative subgroups a-d by dotted horizontal lines. Physicochemically similar amino acids are colored identically.

Figure 14 *(see opposite page)***.** Predicted 3D structure model of three oligomerized Lsm domains of ataxin-2 using three protomers of the Sm1 protein complex from *Pyrococcus abyssi* as template (PDB identifier 1m8v, chains A, B and G) (Albrecht *et al.*, 2004). The model illustrates predicted internal (blue) and external (green) binding sites of ataxin-2 to RNA (gray). While α-helices are shown in red, β-strands are in cyan. Only functionally relevant residues of the central ataxin-2 protomer are annotated as follows: the dark blue box points to residues forming the internal site, and light blue boxes mark amino acids stabilizing the RNA binding area; the dark green box highlights residues involved in the external site, and the light green box indicates stabilizing hydrogen bonds.

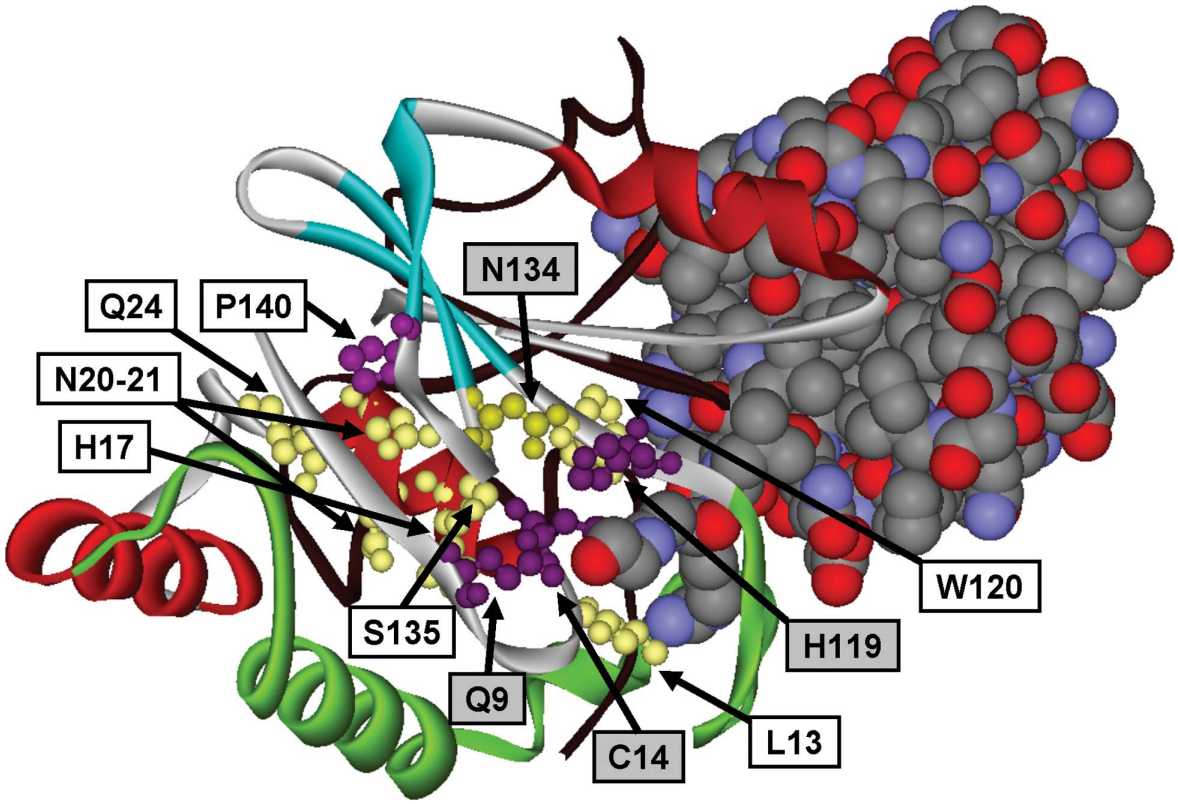


Figure 15. Predicted 3D structure model of the de-ubiquitinating Josephin domain of ataxin-3 using the template structure of the yeast ubiquitin-hydrolase YUH1 (Albrecht *et al.*, 2004). Ataxin-3 (left) is bound to the ubiquitin-like inhibitor Ubal, ubiquitin-aldehyde (right), taken from the binding complex of YUH1 with Ubal (PDB identifier 1cmx, chains A and B, respectively). Gray-shaded text labels indicate the four catalytic residues (violet) forming the active site of ataxin-3, a cysteine protease. The remaining text boxes point to other residues that are highly conserved in the Josephin domain. The N-terminal YUH1 extension missing in ataxin-3 homologs is depicted in the background as thin protein backbone (brown). The less conserved central part of ataxin-3 is colored green; this part could not be modeled reliably using YUH1 as template because of low sequence similarity.

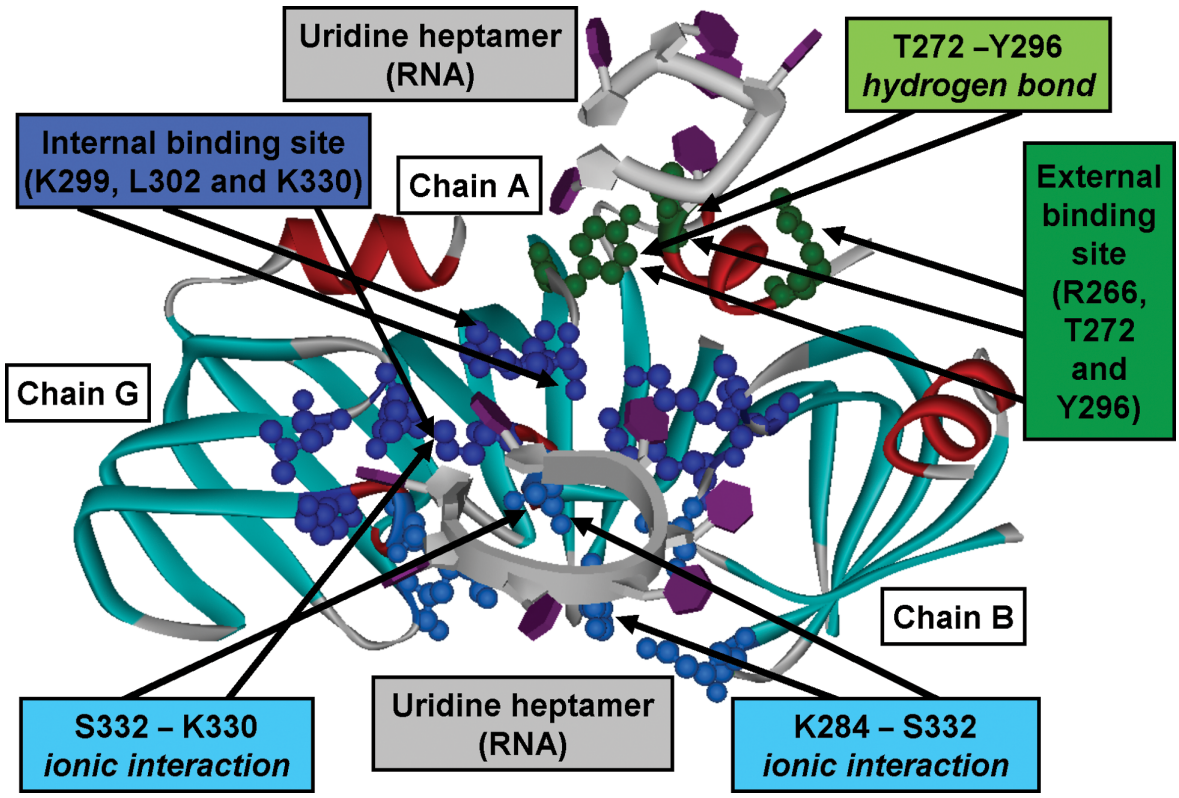
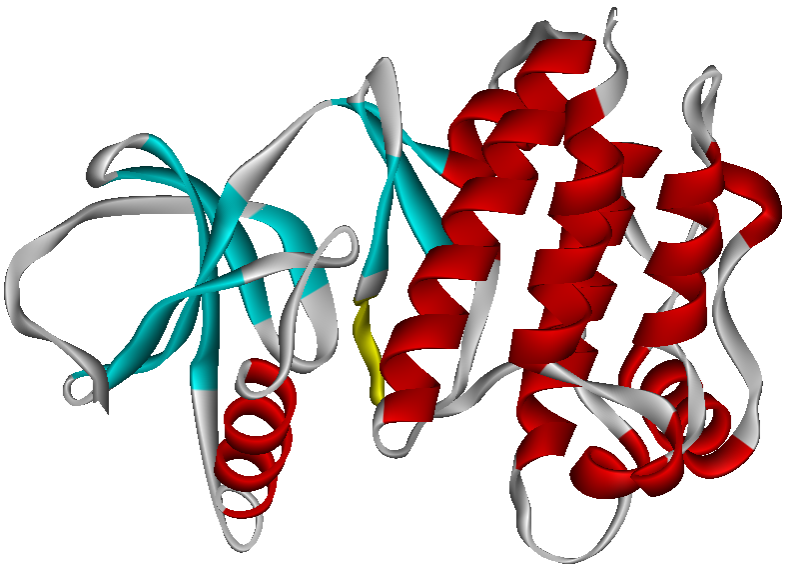


Figure 16. Predicted 3D structure model of the protein kinase RIP7/ LRRK2 based on the B-RAF kinase domain template (PDB identifier 1uwh) (Albrecht, 2005). Two mutations G2019S and I2020T associated with Parkinson's disease are located in the kinase activation loop (yellow). Protein kinase activity is regulated by phosphorylation of specific amino acids within this loop, which may be impaired by the mutations. RIP7 G2019 corresponds to G595 of B-RAF, whose mutation is associated with cancer (Wan *et al.*, 2004).



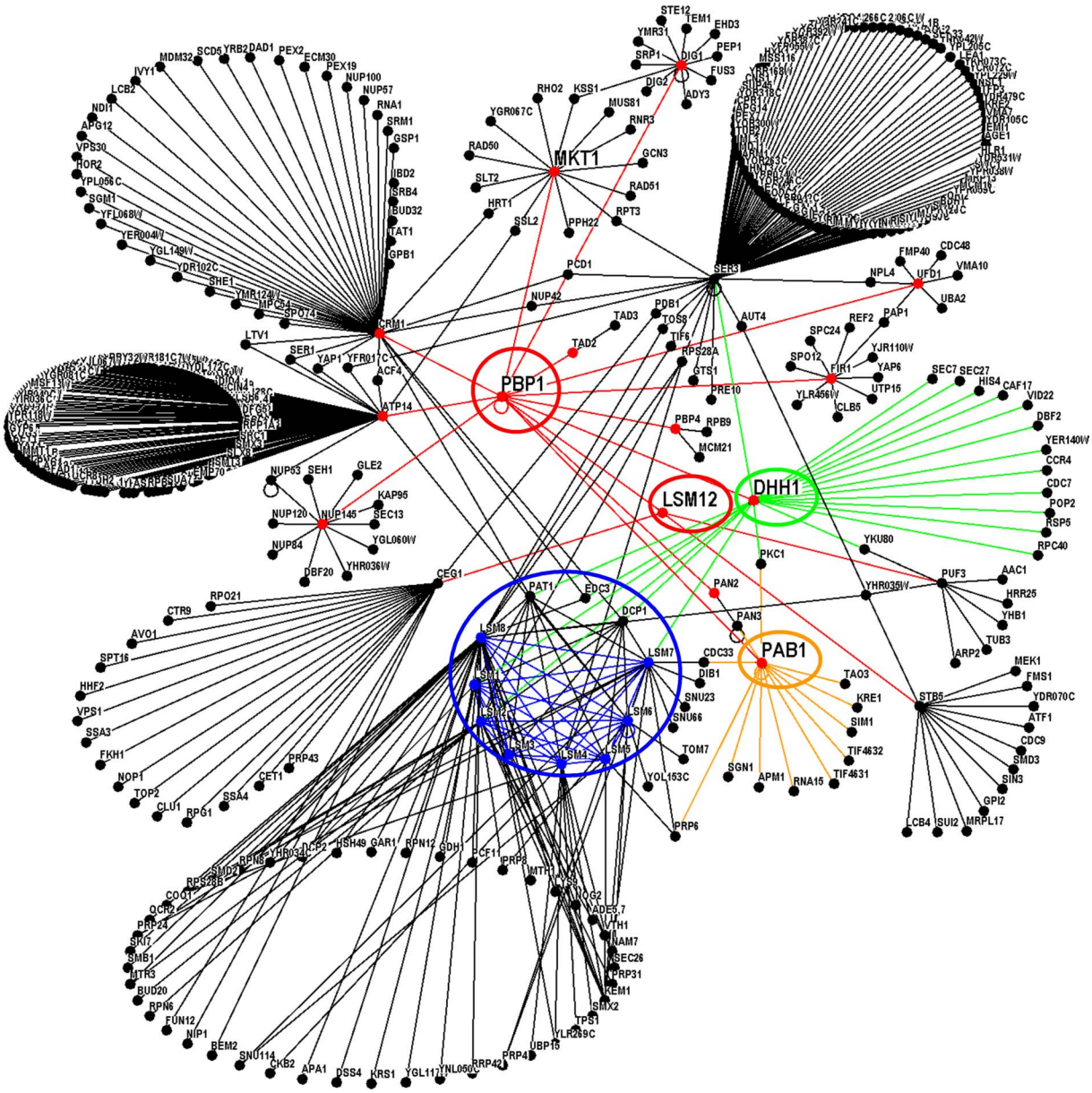


Figure 17. Yeast protein interaction network with focus on the yeast homolog Pbp1 of ataxin-2 and its interaction partner Lsm12 (Ralser *et al.*, 2005a). Each node represents a different *Saccharomyces cerevisiae* protein, whose association with another protein is indicated by an edge. All interaction partners of each protein directly linked to Pbp1 or Lsm12 are shown together with some selected additional proteins. However, not all interaction edges between depicted nodes could be drawn without cluttering the figure, and thus many peripheral edges had to be omitted. The nodes for the Lsm domain proteins Pbp1 and Lsm12 are marked by circles and colored in red together with the connecting edges. The dense interaction network of the Lsm1-8 proteins is encircled and depicted in blue. Edges emanating from the RNA-helicase Dhh1 and the poly(A)-binding protein Pab1 are colored in green and orange, respectively.

2.4 Predictions and Solved Protein Structures

Recently, other research groups determined experimental solution structures of predicted protein models using X-ray crystallography or NMR spectroscopy. This provides the unique opportunity to verify some of the structure and function predictions obtained by bioinformatics methods. In the following, we compare predicted and solved 3D structures concerning NLRs and ataxin-3. In addition, we briefly describe our accurate fold recognition of pyranose oxidase, an enzyme of biomedical and biotechnological relevance.

2.4.1 NLRs and the APAF-1 Crystal Structure

Our bioinformatics analyses of NLR homologs such as NALP3 and NOD2 revealed (Albrecht *et al.*, 2003a) that those proteins share a common nucleotide-binding region consisting of at least three structurally distinct subdomains named NACHT-NAD1-NAD2. We also showed that these subdomains are closely related to the NB-ARC1-ARC2 subdomains of human APAF-1 (apoptotic protease-activating factor 1) and plant disease resistance (R) proteins (Albrecht *et al.*, 2003a; Albrecht and Takken, 2006). Our observations were corroborated further in independent work by Leipe, Koonin, and Aravind. They assembled the so-called STAND family of NTPases including NACHT-NAD and NB-ARC proteins and designated the NAD1/ARC1 and NAD2/ARC2 subdomains GxP module and HETHS domain, respectively (Leipe *et al.*, 2004).

Initially, the 3D structure of APAF-1 and its *Caenorhabditis elegans* ortholog CED-4 was modeled (Cardozo and Abagyan, 1998) based on the GTP-binding domain of G proteins such as Ras (Paduch *et al.*, 2001). Thereafter, other researchers modeled APAF-1 and CED-4 structures similar to those of AAA+ ATPases (Jaroszewski *et al.*, 2000). Indeed, AAA+ ATPases are the closest evolutionary neighbors of STAND NTPases in contrast to G proteins with a quite different structural topology (Albrecht *et al.*, 2003a; Leipe *et al.*, 2004). Therefore, we used the AAA+ ATPase Cdc6 as template to construct 3D models of NALP3 and NOD2 (Albrecht *et al.*, 2003a). A few weeks later, another research group independently published a NALP3 model based on the same template structure (Neven *et al.*, 2004). Like in our work (Figure 5), this NALP3 model was annotated with the locations of disease-associated sequence variants, leading to functional interpretations that have been identical to our conclusions about a potential impairment of ATP hydrolysis caused by the variants.

Eventually, the X-ray crystal structures of ADP-binding APAF-1 (Riedl *et al.*, 2005) and ATP-binding CED-4 (Yan *et al.*, 2005) were determined experimentally. They unveiled that G protein structures were not the best choice of modeling templates for APAF-1 and CED-4. In contrast, the very similar crystal structures of APAF-1 and CED-4 confirmed that the AAA+ ATPase structure of Cdc6 was the best available template structure for modeling the nucleotide-binding region of APAF-1/CED-4 and thus of NALP3 and NOD2 as well. The superposition of the APAF-1 and Cdc6 structures using the FATCAT method (Ye and Godzik, 2003) yields a small RMSD of 3.0Å over all subdomains NB-ARC1-ARC2 (Figure 18). Based on this superposition, we also found that the previous sequence-structure alignments used for 3D modeling

were highly accurate and that the APAF-1 and CED-4 crystal structures validated our original sequence-based division of the NACHT-NAD region into three structural subdomains NACHT-NAD1-NAD2.

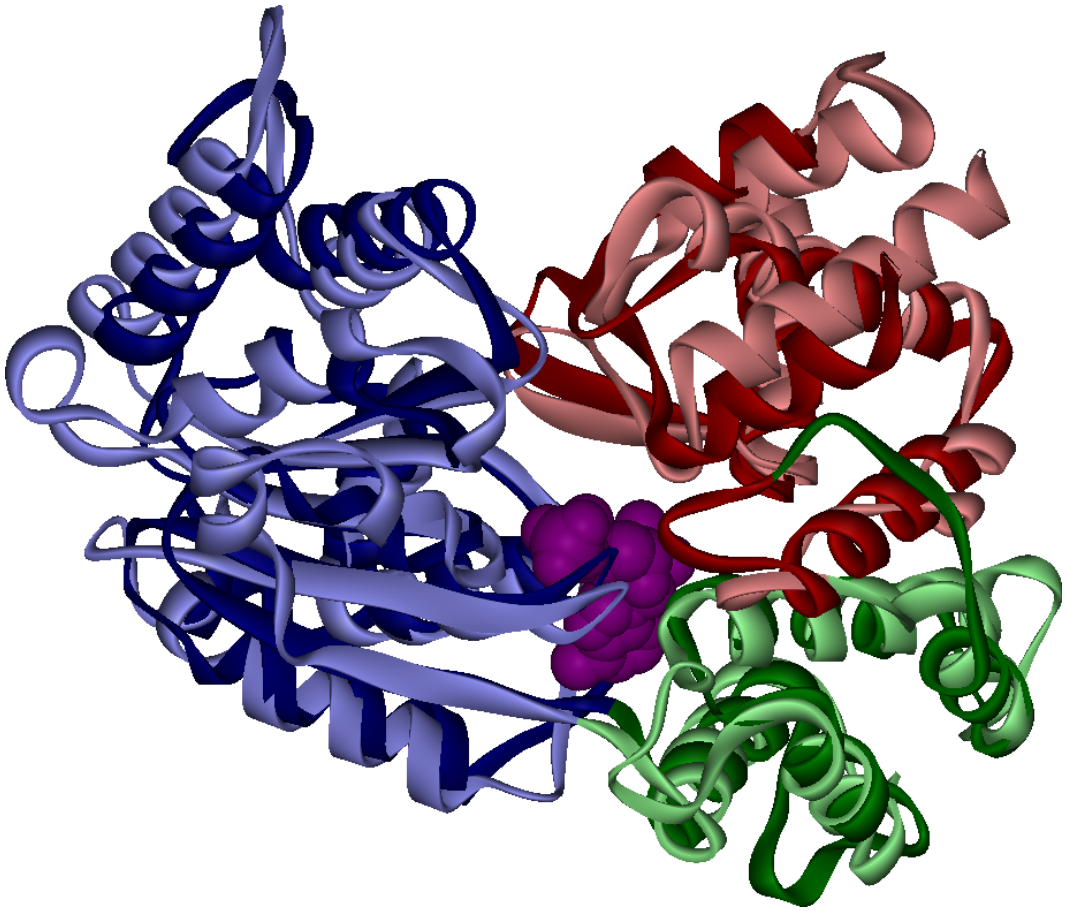


Figure 18. 3D FATCAT superposition of the nucleotide-binding region of the NALP3 and NOD2 template structure Cdc6 (PDB identifier 1fnn, chain A) with the experimentally determined APAF-1 structure in complex with ADP (PDB identifier 1z6t, chain A). The three structural subdomains of Cdc6/APAF-1 are colored light/dark blue, green, and red from the N-terminus to the C-terminus. The ADP molecule bound to APAF-1 is depicted with violet spheres.

2.4.2 Solution Structure of the Josephin Domain

Recently, the NMR solution structure of the de-ubiquitinating Josephin domain of ataxin-3 was determined by two independent research groups (Mao *et al.*, 2005; Nicastro *et al.*, 2005) and confirmed our structural and functional predictions (Figure 19). The solution structure adopts the fold of our modeling template YUH1 with a very low RMSD of 2.8Å over 89 amino acids and a sequence identity of only 11% as measured by structural superposition using the DALI method. In particular, the active site of the de-ubiquitinating protease is conserved well.

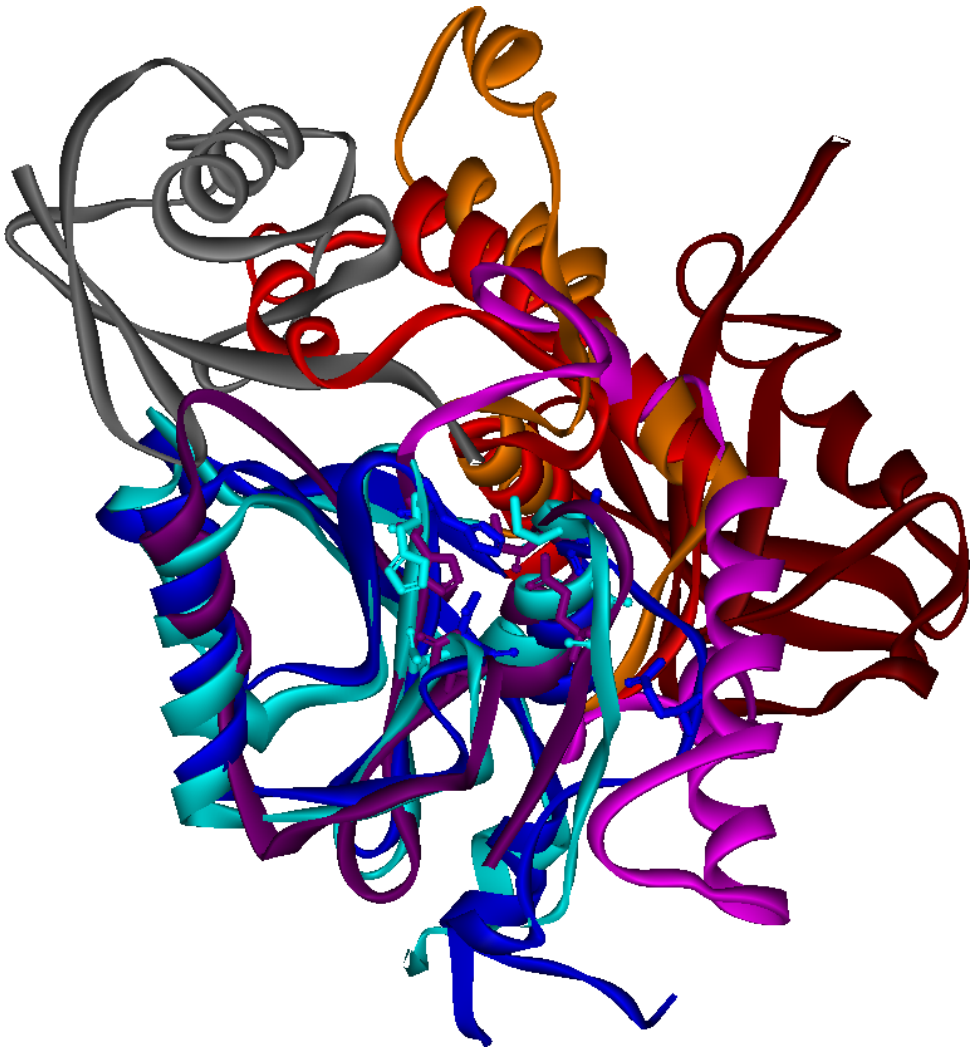


Figure 19. 3D model of the protein complex between one NMR structure (PDB identifier 1yzb) of the Josephin domain of ataxin-3 (cyan) and the ubiquitin-like domain UBL (PDB identifier 1p1a) of the ataxin-3 binding partner hHR23 (brown). This Josephin-UBL complex was re-computed according to (Nicastro *et al.*, 2005) using HADDOCK. We also superimposed this NMR structure both with the other NMR structure (PDB identifier 2aga) of the Josephin domain and with our previous model complex (Figure 15) of ataxin-3 (blue and purple, respectively) bound to the ubiquitin-substitute Ubal (gray). In the picture center, the four catalytic residues of the ubiquitin-hydrolase ataxin-3 are shown as sticks for the experimental and predicted structures. The catalytic triad of cysteine (top right), histidine (top left), and asparagine (bottom left) is conserved well, while the position of the glutamine (bottom right) seems to be variable. The less conserved central part of ataxin-3 contains a helical insertion, which we could not model reliably due to low sequence similarity. Indeed, the structural arrangement of the respective helices is quite flexible and differs significantly between both NMR structures (orange and red) and our structural model (pink) of the Josephin domain. Interestingly, the helix of one NMR structure (red) would collide with Ubal (gray) binding the Josephin domain, which suggests drastic conformational changes during ubiquitin binding.

The NMR structures of ataxin-3 also unveiled a novel helical insertion that is completely different from all other known protease structures. Its conformation varies significantly between the calculated NMR structures (Figure 19). This flexibility contributes much to the overall RMSD of 2.7Å over 158 amino acids between the independently determined NMR structures (computed by DALI using the first PDB model of each NMR research group). In comparison, we originally predicted the sequence region of this helical insertion to consist of helices, but already noted that their structural arrangement may differ. The reason for our caution was that comparative modeling of this structure part was difficult due to insufficient sequence similarity (Albrecht *et al.*, 2004).

Finally, our suggested model of a complex between the Josephin domain and ubiquitin based on the template complex of YUH1 and Ubal (Figure 15) can readily be transferred to the NMR solution structure by superimposing the structures of the Josephin domain and YUH1 (Figure 19). Based on NMR data, it appears quite likely that this modeled complex is approximately correct (Mao *et al.*, 2005) and may be stabilized by the described helical insertion (Nicastro *et al.*, 2005).

2.4.3 Crystal Structure of Pyranose Oxidase

Pyranose oxidase (POX), also known as glucose 2-oxidase (EC 1.1.3.10), catalyzes the oxidation of monosaccharides such as D-glucose and other aldopyranoses at carbon-2 (or sometimes at carbon-3) in order to yield keto-aldoses while reducing O_2 to H_2O_2 (Giffhorn *et al.*, 2000). This flavoenzyme is produced by several lignin-degrading basidiomycete fungi and is assumed to play an important functional role by supplying H_2O_2 as co-substrate to lignin-decomposing peroxidases. The preferred substrate of POX is both α- and β-D-glucose, whereas the functionally related enzyme glucose oxidase (GOX) acts solely on β-D-glucose at carbon-1 (EC 1.1.3.4). POX like GOX has received remarkable attention both in biotechnological applications for high yields of D-fructose and in food technology as a tool for glucose measurement (Giffhorn, 2000). This tool is also important for analytical purposes, for instance, in clinical chemistry as a diagnostic marker of diabetes.

In contrast to GOX, no X-ray crystal structure had been determined for POX. However, after a thorough bioinformatics study of conserved sequence features in homologous proteins, we could classify POX into the GMC (glucose-methanol-choline) oxidoreductase family containing GOX (Albrecht and Lengauer, 2003). Based on our structure-based multiple sequence alignment, we could predict the structural fold of POX to be similar to GOX, which enabled the reliable identification of potentially catalytic residues. Meanwhile, our predictions have been confirmed by mass spectrometry concerning the correct localization of the FAD-binding site (Halada *et al.*, 2003) and especially by the determination of two X-ray crystal structures of POX (Bannwarth *et al.*, 2004; Hallberg *et al.*, 2004). These experimental structures revealed that our original sequence alignment was of high quality and that we had identified most functionally relevant amino acids in the active site. Thus, our biotechnological partners could have used this information for site-directed mutagenesis experiments to improve enzymatic properties.

2.5 Methodological Limitations

The bioinformatics application studies detailed in the preceding sections have not only revealed structural and functional properties of medically relevant proteins, but also some limitations of bioinformatics methods. Therefore, the following list describes encountered problems and provides suggestions for the development of novel and improved computational techniques to support further biological and medical investigations:

(1) Since numerous prediction methods for protein secondary structure are readily available online and provide swift results, it is difficult to know the best method to apply. However, instead of relying on a single prediction result only, one should consider the use of several methods at the same time. Still, it has been unclear so far whether a consensus prediction formed from the different results of state-of-the-art prediction methods is likely to improve the prediction quality and reliability.

(2) Various bioinformatics methods are offered by web services for the prediction of transmembrane helices. Accordingly, this variety raises similar questions concerning consensus formation like secondary structure prediction as explained in (1). It would also be interesting to learn whether the maximum or the minimum or the average of the number of helices predicted by each method may be the best estimate for the transmembrane topology. In addition, it is not obvious how to compute the start and end positions of the predicted helices for a consensus.

(3) When comparing previously predicted secondary structures with protein structures solved later in experiment, we could frequently observe that the entire secondary structure had been predicted with very high accuracy. Sometimes, no appropriate 3D modeling template had been detected by any fold recognition method although several templates have possessed very similar secondary structure and the correct tertiary fold. This observation suggests possible improvements in the utilization of predicted secondary structure for recognizing appropriate template structures, particularly in cases of distant evolutionary relationships. Additionally, if fold recognition methods based on sequence similarities do not return useful and reliable results, secondary structure strings of target and template proteins could be compared directly with each other without considering the primary amino acid sequences.

(4) When comparing previously modeled 3D structures with recently determined solution structures, we have recognized that regions of secondary structure prediction differing considerably from the structure of the modeling templates may indicate unreliable parts of the resulting 3D model. Consequently, this observation might lead to new model quality measures using secondary structure prediction for assessing the reliability of specific model parts.

(5) Frequently, additional experimental or predicted data are known for target proteins, for instance, posttranslational modifications of protein residues (glycosylation, phosphorylation, ubiquitinylation, sumoylation, etc.), the primary domain architecture, the location of binding sites, secondary structure motifs, residue burial or exposition to solvent, positions of disulfide bonds, and spatial distances between amino acids. This information could be utilized in structure prediction methods to select structural templates, to improve target-template alignment quality, and to validate 3D structure models of the target protein. It could also be valuable in improving and accelerating computational fold recognition approaches by discriminating between alternative templates and by filtering ranking lists of possible templates.

(6) Careful template selection for 3D structure modeling currently involves mainly manual work to identify and compare structurally and functionally relevant amino acids of target and template homologs using multiple sequence alignments and structure superpositions. Although this work is essential for assessing the quality and reliability of the modeling alignments and the resulting tertiary structure models, the necessary data can often be extracted solely by time-consuming reading of dozens of biological journal articles. Unfortunately, no bioinformatics databases are really established for experimental information on single amino acids of proteins. UniProt contains such annotations, but much more effort on maintaining regular updates and comprehensive descriptions would be needed. Moreover, biologists are not required yet to deposit related biochemical and mutational data in databases like UniProt after the acceptance of their publication. For this functional annotation could additionally be used in structure prediction methods as detailed in (5) and displayed in aligned sequences to support the verification of target-template alignments.

(7) Many web services and command line programs support different semi-automatic stages of protein structure modeling, a process coupled with 3D structure viewing and involving the following basic modules: sequence profile searches, multiple sequence alignment computations and editing, secondary structure predictions, template selection and analysis, structural superpositions, protein backbone generation, side chain and loop modeling, model quality evaluation and energy calculations. However, those process modules are usually not provided within a uniform and flexible framework offering a user-friendly graphical interface as it is nowadays standard for software programming projects based on integrated development environments (IDE). The usability of bioinformatics methods and databases could be enhanced considerably if research groups provide their computational tools simply by implementing appropriate IDE plugins. For the analysis of gene or protein interaction networks, this is already possible with the extensible and open-source IDE-like software platform Cytoscape.

(8) When investigating protein-protein interactions, it is first necessary to determine the protein domains responsible for the interactions before being able to model a protein domain complex and to analyze the corresponding binding sites in 3D. To this end, the development of bioinformatics methods for the prediction of domain interactions and for the automatic decomposition of protein-protein interactions into domain-domain interactions is important. In this context, new databases containing information on experimentally observed interactions of protein fragments/domains and splice variants might facilitate bioinformatics research.

(9) After modeling a protein domain complex based on the 3D structure prediction of each domain, it is reasonable to assess the overall quality of the modeled complex. For this purpose, confidence measures to assess domain-domain interaction interfaces could be developed. They could indicate the likelihood of the predicted domain interaction to occur *in vivo*. Moreover, advanced bioinformatics methods to predict flexible protein regions moving during domain-domain binding are desirable.

(10) Homologous proteins often share not only 3D structure, but also similar functions and interaction partners. Knowing the domains of proteins interacting with the target protein may aid in the selection of the best modeling template for predicting the structure of the target. More details on this idea to improve fold recognition approaches can be found in the last section of Chapter 5.

Solutions to the three methodological limitations (1), (5) and (8) form part of this dissertation and will be addressed in the succeeding Chapters 3, 4 and 5, respectively. While two more issues (3) and (7) have recently been tackled by other research groups and are reported below, the remaining problems (2), (4), (6), (9) and (10) with bioinformatics methods still appear to be unsolved to date.

The difficulties explained in (3) have been acknowledged by our cooperation partner Silvio Tosatto and his coworkers at the University of Padova. They implemented a publicly available web service named SSEA for the computation of protein secondary structure alignments (Fontana *et al.*, 2005). It supports both performing pairwise alignments and searching a given secondary structure against a library of domain folds derived from the PDB.

Furthermore, the lack of a generic modeling IDE as discussed in (7) has been approached by Roland Dunbrack and his colleagues. They designed a free and open-source molecular integrated development environment (MollIDE) that combines the most frequent modeling steps and guides the user from the target protein sequence to the final 3D structure model (Canutescu and Dunbrack, 2005). Another new and versatile graphical IDE like MollIDE is BALLView, which focuses on visualization and energy computation of 3D structures including molecular dynamics simulations (Moll *et al.*, 2006). Both IDEs can be regarded as first innovative steps towards a more

comprehensive and adaptable modeling environment providing richer functionality as outlined in (7).

2.6 Systems Biology Perspectives

In various application studies described in this chapter, we have investigated the structure and function of seemingly disparate proteins associated with autoinflammatory and neurodegenerative diseases of different etiologies and phenotypes. However, on the molecular level, surprising functional similarities between medically relevant proteins and their binding partners have been observed recently. In future work, it will be beneficial to take such relationships into account when inferring cellular disease models.

For instance, recent research findings on neurodegenerative disorders point to the involvement of causative, evolutionarily unrelated, proteins in identical or at least very similar signaling pathways. Examples are the polyglutamine proteins ataxin-2 and ataxin-3 causing ataxia types 2 and 3, both of which were reported to be associated with Parkinson's disease as well (Morris, 2005). Regarding Huntington's disease, we particularly found that both ataxin-2 and huntingtin interact with endophilins (Harjes and Wanker, 2003; Ralser *et al.*, 2005b). In the near future, such molecular similarities may facilitate the identification of common cellular dysfunctions for therapeutic targeting.

Remarkably, further functional interrelationships even between autoinflammatory and neurodegenerative diseases have become apparent on the molecular level despite of prominent phenotypic differences. Concrete examples are the evolutionarily related protein kinases RIP2 (RICK/CARDIAK) and RIP7 (LRRK2/dardarin) (Meylan and Tschopp, 2005), which have been found to act in Crohn's, Huntington's and Parkinson's diseases. NOD2 requires RIP2 to induce ubiquitinylation of NEMO, a key component of the NF-κB signaling complex affected in Crohn's disease (Abbott *et al.*, 2004). Also, dysregulation of RIP2 is associated with Huntington's disease progression (Wang *et al.*, 2005), whereas mutations of RIP7 are causative of Parkinson's disease (Singleton, 2005). Therefore, detailed molecular knowledge of the same signaling proteins and their pathways may facilitate research on distinct diseases.

Generally, the small protein ubiquitin plays an important role not only in the regulation of NOD2 and the modulation of immune responses (Liu *et al.*, 2005b), but also in neurodegeneration-associated pathways involving diverse proteins linked to Parkinson's and Huntington's disease as well as ataxias (Ciechanover and Brundin, 2003; Johnston and Madura, 2004; Ross and Pickart, 2004). In addition, the transcription factor NF-κB is crucial for immune responses and apoptotic processes in many diseases (Li and Verma, 2002; Vila and Przedborski, 2003; Bonizzi and Karin, 2004). Apoptosis is regulated by complicated biological mechanisms with intertwining signaling cascades, caspase-regulated processes, and ubiquitin-mediated degradation (Riedl and Shi, 2004; Vaux and Silke, 2005; Yan and Shi, 2005).

In conclusion, the consideration of intricate molecular relationships will become increasingly important in order to advance the understanding of disease processes and the consequences of defects in multifunctional proteins. It will no longer suffice to

analyze single proteins, but their functional context will need to be explored intensively as well. This will necessitate a systems biology approach, which integrates rapidly growing experimental knowledge on molecular interactions and their spatiotemporal changes into cellular models (Cusick *et al.*, 2005).

This endeavor will include the comparative analysis of large interaction networks and the quantitative simulation of complex pathways for disease-associated proteins (Bork *et al.*, 2004). For this purpose, protein interaction maps of substantial size have recently been generated for NF-κB (Bouwmeester *et al.*, 2004), huntingtin (Goehler *et al.*, 2004), and many other human proteins (Rual *et al.*, 2005; Stelzl *et al.*, 2005). Eventually, comprehensive biological models of varying granularity including spatial information on protein structures and complexes (Aloy *et al.*, 2005) as well as on cellular communication will be needed to explain physiological phenomena (Xia *et al.*, 2004). Fortunately, it appears that the same model may aid the study of several diseases. For example, the discovery of similar molecular reactions underlying clinically distinct autoinflammatory diseases may facilitate the search for anti-inflammatory drugs (Schreiber *et al.*, 2005). In case of neurodegeneration, the bioinformatics-supported distinction of neuroprotective and toxic modulations of aberrant protein interactions triggering pathogenic cascades may point to common drug targets (Ryan and Matthews, 2005).

3
Predicting Consensus Secondary Structure

This chapter shows that, in contrast to previously published, more sophisticated, methods, simple consensus procedures are effective and sufficient in improving secondary structure prediction. This research arose from the use of multiple prediction methods in application studies on medically relevant proteins (Chapter 2). It was conducted in cooperation with Silvio Tosatto and his colleagues at the University of Padova in Italy. They participated in the fifth round of the Critical Assessment of Structure Prediction (CASP5 in 2002) and submitted the results of the consensus procedure closely following my suggestions. Since this procedure was surprisingly successful in CASP5 (Aloy *et al.*, 2003b), we then analyzed the performance of our approach more comprehensively (Albrecht *et al.*, 2003e).

After some more background information on secondary structure prediction, the following sections detail our consensus method and benchmarking results. Basically, our consensus prediction is obtained by majority voting on minimal combination sets of three state-of-the-art prediction methods. Using large data sets for benchmarking, we demonstrate that our method achieves a significant improvement in the average Q_3 prediction accuracy of up to 1.5 percentage points by consensus formation. Interestingly, the application of an additional trivial filtering procedure for predicted secondary structure elements that are too short, does not significantly affect the prediction accuracy. Our analysis also provides valuable insight into the similarity of the results obtained by prediction methods that we combine. Additionally, we observe a higher confidence in consistently predicted secondary structure.

3.1 Introduction

The prediction of secondary structure is a frequent task in sequence analysis of globular proteins. It provides information on the putative number and positions of α-helices and β-strands, which is particularly useful for 3D structure prediction (Przybylski and Rost, 2002). Recent improvements in prediction accuracy have been accomplished not only by incorporating evolutionary information from homologous sequences into the prediction algorithms, but also by combining the results of single, independent, secondary structure prediction methods into a consensus prediction (Rost, 2001).

The first implementation of the prediction server Jpred computed a consensus of prediction results simply by majority voting (Cuff *et al.*, 1998). Minor method variations such as different weights added to the results did not lead to significantly higher prediction accuracies (Cuff and Barton, 1999; King *et al.*, 2000). However, the Jpred server has been improved recently (Cuff and Barton, 2000) and now employs a complex combination of neural networks, a method that has also been applied successfully for consensus formation by other groups (Chandonia and Karplus, 1999; King *et al.*, 2000; Petersen *et al.*, 2000). Similar sophisticated approaches use multivariate linear regression (Guermeur *et al.*, 1999) or decision trees (Selbig *et al.*, 1999) trained for optimal method selection. Other method variants apply either cascaded multiple classifiers of secondary structure (Ouali and King, 2000) or a composite secondary structure assembled from the results of several methods (An and Friesner, 2002). The common feature of all these consensus approaches is the use of results from usually more than three secondary structure prediction methods.

We found that a set of only three state-of-the-art methods combined by straightforward majority voting is sufficient to achieve similar improvements in prediction accuracy (Figure 20). This simple approach runs at low computational cost and uses the currently best prediction servers.

Figure 20. Consensus prediction formed by majority voting using the results of three secondary structure prediction methods (SSPred1-3).

In order to test our approach, we participated in the critical assessment of structure prediction, the CASP5 experiment of the year 2002 (Tramontano, 2003). We combined the prediction results of the three web servers PSIPRED, SAM-T02, and SSpro2, which are based on different prediction approaches using neural networks and hidden Markov models. The three servers have shown top performance in former CASP experiments and the continuous automatic evaluation (EVA) of protein structure prediction servers (Eyrich *et al.*, 2001; Rost and Eyrich, 2001). The prediction methods implemented by the three web servers have higher overall accuracy than older combination procedures such as Jpred.

Astonishingly, our method named CaspIta (group number 108) significantly outperformed almost all other methods participating in CASP5 and reached the second rank below a manual expert submission according to the SOV score, normalized with respect to the total number of all 78 target protein domain sequences (Aloy *et al.*, 2003b). In particular, our method ranked first regarding the SOV accuracy measure for a subset of 21 target domains unrelated by sequence and with low sequence similarity to known protein structures. Regarding the alternative Q_3 score for prediction accuracy, our combination method ranks first for the set of all targets and the subset of sequence unrelated targets (for details see *http://www.russell.embl.de/casp5/*).

Encouraged by the CASP5 results, we decided to investigate our approach on larger benchmark sets obtained from EVA. In particular, we show that our approach always improves the prediction accuracy over the best single method of the three methods combined to form the consensus. Comparing the frequencies for the occurrence of certain majority situations, we are able to draw interesting conclusions on the degree of similarity between results of single prediction methods and on the increased confidence in consistently predicted secondary structure.

3.2 Materials and Methods

3.2.1 Benchmark Sets

In our evaluation, we used the three benchmark sets ‘common2’, ‘common5’, and ‘common6’ from 22 September 2002 with sequences of low identity, as provided by the EVA web site (*http://cubic.bioc.columbia.edu/~eva/*). The set ‘common2’ contains 121 sequences with 16,858 amino acids, ‘common5’ contains 214 sequences with 44,871 amino acids, and ‘common6’ contains 539 sequences with 98,308 residues. Because not all methods have returned predictions for every sequence requested by EVA, not every benchmark set could be combined with the same three methods used for consensus computation (see legend of Table I).

3.2.2 Consensus Formation

For each benchmark set, three single methods of top performance in EVA are selected in order to compute the consensus secondary structure sequence by majority voting (using Perl scripts). Specifically, we processed the results of the following seven prediction methods: PSIPRED (Jones, 1999; McGuffin *et al.*, 2000), SAM-T99 (Karplus *et al.*, 1998), SSpro1 (Baldi *et al.*, 1999), SSpro2 (Pollastri *et al.*, 2002), PHDpsi (Przybylski and Rost, 2002), PROFsec (Rost and Eyrich, 2001), Jpred (Cuff and Barton, 2000).

Three cases need to be distinguished when forming the consensus sequence per amino acid according to the three possible secondary states α-helix (H), β-strand (E), and other/loop (L): 3:0 votes means consistent prediction among all three methods. 2:1 votes result in the majority decision. The rare case of a tie 1:1:1 is resolved by assuming the L state. Each consensus sequence is annotated with a confidence array, which

contains values ranging from 1 to 3 according to the maximum number of identical votes per residue.

3.2.3 Prediction Accuracy

To determine the prediction accuracy, we compared the predicted consensus sequence to the true three-state sequence derived from the DSSP secondary structure assignment of known 3D structures (Kabsch and Sander, 1983). Each of the three possible states H, E, and L per residue results from the eight possible DSSP states according to the following standard transformation schema (Rost and Eyrich, 2001): {G,H,I} $\rightarrow$ α-helix (H), {B,E} $\rightarrow$ β-strand (E), {S,T,'.'} $\rightarrow$ other (L).

For each benchmark set, we computed average Q_3 and SOV percentage values (Rost *et al.*, 1994; Zemla *et al.*, 1999; Rost and Eyrich, 2001) as well as the separate percentages Q_H, Q_E, and Q_L of residues predicted correctly in the observed H, E, and L states, respectively. For each accuracy measure, we calculated the standard error by dividing the standard deviation of the measure by the square root of the benchmark set size. Assuming a normal distribution of the accuracy measures, the accuracy difference between two distinct prediction methods may be assumed to be statistically significant if it is larger than the maximum of the standard errors (Rost and Eyrich, 2001).

3.3 Results and Discussion

3.3.1 Accuracy Improvement

The results of the consensus formation by majority voting using three different benchmark sets are summarized in Table I. The comparison of our consensus approach to the respective best single method demonstrates that the total average Q_3 accuracy is increased considerably by 1.45, 1.50, and 0.41 percentage points for each set 'common2', 'common5', and 'common6', respectively. In particular, the accuracy increase is statistically significant (in the sense described above) in case of the sets 'common2' and 'common5'. In addition, the SOV measure is improved by 0.68 percentage points for the 'common5' set, while it does not change substantially for the other two sets. Table I also contains the results of the consensus prediction method Jpred as available for the sets 'common2' and 'common5', but its accuracy is generally clearly below those of other methods. For comparison, we included the results of PROFsec, another top-performing single prediction method, into Table I for the 'common2' set. Its Q_3 prediction accuracy shows a significantly lower performance reduced by 1.27-1.57 percentage points in contrast to the very similar consensus results of any three single methods combined out of the four available methods PSIPRED, SAM-T99, SSpro, and PROFsec.

If one compares the accuracy measures of each of the methods that have been combined to form the consensus, based on a separation according to the true H, E, and L states observed, it appears that the consensus formation always improves the Q_3 value of the L state class by 0.51-1.55 percentage points. This finding could mean that single methods tend to underpredict the L state.

Table I.

Results of the consensus secondary structure prediction for the benchmark sets (a) 'common2', (b) 'common5', and (c) 'common6'. The first three prediction methods shown in (a)-(c) are combined by consensus formation with majority voting. For each method, the means (μ) and standard errors (err_σ) of the accuracy measures Q_3, Q_H, Q_E, Q_L, and SOV are given in percent. For comparison, we included the results of the prediction methods PROFsec and Jpred. The consensus results for every possible combination of PROFsec with two of the first three methods are also included in (a).

(a)

	Q_3		Q_H		Q_E		Q_L		SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
PSIPRED (PS)	74.24	1.08	76.34	2.20	69.61	2.82	74.73	1.23	70.23	1.63
SAM-T99 (SA)	73.97	0.96	77.02	2.29	69.58	2.43	71.84	1.55	67.67	1.56
SSpro2 (SS)	73.71	1.00	74.57	2.44	69.64	2.39	74.55	1.26	67.43	1.58
PROFsec (PR)	74.42	0.90	70.72	2.77	70.70	2.63	73.87	1.33	69.80	1.50
Jpred	72.03	1.13	62.07	2.94	62.43	2.95	82.52	1.24	66.58	1.91
Cons. of PS, SA, SS	75.69	0.98	77.58	2.31	70.28	2.50	75.79	1.32	70.18	1.63
Cons. of PS, SA, PR	75.98	0.89	75.99	2.27	70.91	2.52	76.39	1.33	70.97	1.59
Cons. of PS, SS, PR	75.84	0.93	76.01	2.35	71.18	2.51	75.89	1.30	70.46	1.56
Cons. of SA, SS, PR	75.94	0.90	75.60	2.37	70.92	2.43	76.15	1.35	70.45	1.56

(b)

	Q_3		Q_H		Q_E		Q_L		SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
PSIPRED	76.16	0.69	77.78	1.45	68.63	1.85	75.62	0.88	71.58	1.03
SSpro1	76.03	0.68	77.51	1.52	65.43	1.81	76.24	0.86	70.38	1.04
PROFsec	76.33	0.63	75.83	1.55	69.42	1.79	74.75	0.94	72.23	0.93
Jpred	74.63	0.63	68.24	1.76	60.82	1.89	82.51	0.83	69.08	1.04
Consensus	77.83	0.65	78.13	1.48	68.24	1.83	77.79	0.87	72.91	1.00

(c)

	Q_3		Q_H		Q_E		Q_L		SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
PSIPRED	75.60	0.48	77.95	0.95	69.07	1.20	75.51	0.58	71.51	0.70
PROFsec	75.68	0.44	74.21	1.12	69.98	1.15	74.81	0.61	71.48	0.63
PHDpsi	73.48	0.46	72.53	1.16	65.76	1.20	73.40	0.65	68.52	0.64
Consensus	76.09	0.44	74.87	1.10	68.73	1.17	76.02	0.61	71.58	0.63

3.3.2 Filtering of Prediction Results

Furthermore, we found that the application of a trivial filtering procedure that eliminates α-helices and β-strands that are too short generally neither deteriorates nor ameliorates the prediction accuracy significantly, be it before and/or after the consensus formation. In detail, this procedure converts the secondary structure states of residues in secondary structure elements of impossible length (α-helices shorter than three residues and β-

strands shorter than two residues) to the L state. As can be seen from Table II and Table III, the application of the filtering procedure generally does not affect the prediction accuracy significantly, be it before and/or after the consensus formation. An interesting exception is the SOV value of the SSpro1 and SSpro2 methods, which is improved by 0.98 and 1.76 percentage points, respectively, after the application of the filter.

In summary, this kind of structural filtering can be employed without disadvantages in order to clean up the secondary structure predictions before further processing. For example, this procedure may be particularly useful in 3D structure prediction, where impossible secondary structure elements could complicate the selection of an appropriate template structure.

Table II.

Results of applying the filtering procedure before and/or after consensus formation (pre-/post-filter) for the benchmark sets (a) 'common2', (b) 'common5', and (c) 'common6'. The first three prediction methods shown in (a)-(c) of Table I are combined by consensus formation with majority voting. The means (μ) and standard errors (err_σ) of the accuracy measures Q_3, Q_H, Q_E, Q_L, and SOV are given in percent.

(a)

	Q_3		Q_H		Q_E		Q_L		SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
Consensus	75.69	0.98	77.58	2.31	70.28	2.50	75.79	1.32	70.18	1.63
Prefilter + Cons.	75.58	1.00	77.37	2.34	70.10	2.51	75.75	1.35	70.36	1.66
Cons. + Postfilter	75.66	1.00	77.53	2.35	69.94	2.52	75.77	1.36	71.22	1.66
Pref.+Cons.+Postf.	75.63	1.00	77.50	2.35	69.94	2.52	75.77	1.36	70.99	1.67

(b)

	Q_3		Q_H		Q_E		Q_L		SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
Consensus	77.83	0.65	78.13	1.48	68.24	1.83	77.79	0.87	72.91	1.00
Prefilter + Cons.	77.79	0.65	78.14	1.49	67.99	1.85	77.70	0.88	73.33	1.00
Cons. + Postfilter	77.76	0.65	78.13	1.49	67.78	1.87	77.71	0.88	73.17	1.03
Pref.+Cons.+Postf.	77.75	0.65	78.14	1.50	67.77	1.87	77.77	0.88	73.13	1.02

(c)

	Q_3		Q_H		Q_E		Q_L		SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
Consensus	76.09	0.44	74.87	1.10	68.73	1.17	76.02	0.61	71.58	0.63
Prefilter + Cons.	76.04	0.44	74.97	1.10	68.45	1.18	75.87	0.62	71.50	0.65
Cons. + Postfilter	76.06	0.44	74.89	1.11	68.35	1.19	75.95	0.62	71.50	0.65
Pref.+Cons.+Postf.	76.05	0.44	74.93	1.11	68.31	1.19	75.91	0.62	71.51	0.65

Table III.

Results of applying the filtering procedure to single methods for the benchmark sets (a) 'common2', (b) 'common5', and (c) 'common6'. For each method, the means (μ) and standard errors (err_σ) of the accuracy measures Q_3 and SOV before and after filtering the prediction results are given in percent.

(a)

	Q_3		filtered Q_3		SOV		filtered SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
PSIPRED	74.24	1.08	74.16	1.08	70.23	1.63	70.13	1.64
SAM-T99	73.97	0.96	73.85	0.97	67.67	1.56	67.69	1.64
SSpro2	73.71	1.00	73.70	1.02	67.43	1.58	69.19	1.58
PROFsec	74.42	0.90	74.42	0.90	69.80	1.50	70.34	1.55
Jpred	72.03	1.13	71.98	1.14	66.58	1.91	66.15	1.92

(b)

	Q_3		filtered Q_3		SOV		filtered SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
PSIPRED	76.16	0.69	76.15	0.69	71.58	1.03	71.77	1.03
SSpro1	76.03	0.68	75.99	0.69	70.38	1.04	71.36	1.06
PROFsec	76.33	0.63	76.31	0.64	72.23	0.93	72.30	0.94
Jpred	74.63	0.63	74.60	0.64	69.08	1.04	68.84	1.05

(c)

	Q_3		filtered Q_3		SOV		filtered SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
PSIPRED	75.60	0.48	75.57	0.49	71.51	0.70	71.51	0.71
PROFsec	75.68	0.44	75.65	0.44	71.48	0.63	71.67	0.65
PHDpsi	73.48	0.46	73.47	0.47	68.52	0.64	68.33	0.67

3.3.3 Frequency of Majority Situations

The additional analysis of the overall frequency of the three possible types of majority situations 3:0, 2:1, and tie 1:1:1 uncovers that the problematic case of a tie with each of the three single methods predicting a different secondary structure occurs in at most 1% of all cases (Table IV(a)). Thus, the tie case can be neglected when applying our consensus approach.

In contrast, 3:0 consistency appears about three times as often as 2:1 majority. Here, some methods resemble each other more than others. For instance, based on the benchmark set 'common6', PROFsec is much more similar to PHDpsi than to PSIPRED: the pair (PROFsec, PHDpsi) has a higher 2:1 frequency of 12.6% than the pair (PSIPRED, PROFsec) and the pair (PSIPRED, PHDpsi) with 6.2% and 3.2%, respectively. In contrast, the 2:1 frequencies of the three methods employed for the sets 'common2' and 'common5' are not far from each other. Thus, the respective three methods that are combined seem to be equally dissimilar to each other. Together with

the higher improvement in prediction accuracy observed for the same two benchmark sets compared to the 'common6' set, the rule of thumb may be deduced that the best performance of our approach can be obtained by combining three single methods of top prediction accuracy with approximately equal dissimilarity of their results.

Table IV.

(a) Overall frequencies of the three types of majority situations for the three benchmark sets 'common2', 'common5', and 'common6'. The 2:1 majority situation is additionally subdivided into the three possible combinations of a pair of two methods voting unanimously versus a third method. Each combination consists of the first three methods (1)-(3) as listed in (a), (b) or (c) of Table I: PSIPRED (1), SAM-T99 (2) and SSpro2 (3) for (a); PSIPRED (1), SSpro1 (2) and PROFsec (3) for (b); and PSIPRED (1), PROFsec (2) and PHDpsi (3) for (c).

Benchmark Set	*3:0*	*2:1*				*1:1:1*
		total	*1,2 vs. 3*	*1,3 vs.2*	*2,3 vs. 1*	
common2	75.1%	24.2%	8.2%	8.5%	7.5%	0.7%
common5	76.3%	22.7%	6.4%	8.9%	7.4%	1.0%
common6	77.2%	22.0%	6.2%	3.2%	12.6%	0.8%

(b) Improvement of the Q_3, Q_H, Q_E, Q_L, and SOV accuracy measures in percent if their computation is restricted to secondary structure states that are consistently predicted by all three methods for the benchmark sets (a) "common2", (b) "common5", and (c) "common6".

	Q_3		Q_H		Q_E		Q_L		SOV	
Method	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
Consensus (a)	75.69	0.98	77.58	2.31	70.28	2.50	75.79	1.32	70.18	1.63
Restricted Cons. (a)	82.01	1.01	81.70	2.49	73.16	2.85	81.88	1.38	75.24	1.83
Consensus (b)	77.83	0.65	78.13	1.48	68.24	1.83	77.79	0.87	72.91	1.00
Restricted Cons. (b)	84.26	0.61	83.42	1.50	72.55	2.01	83.62	0.94	78.37	1.07
Consensus (c)	76.09	0.44	74.87	1.10	68.73	1.17	76.02	0.61	71.58	0.63
Restricted Cons. (c)	82.71	0.44	80.90	1.12	72.69	1.28	82.22	0.63	76.76	0.70

3.3.4 Prediction Confidence

We also verified the intuitive expectation that the confidence in the correctness of the prediction is increased by consensus formation. We found that the Q_3 and SOV values computed solely for secondary structure states that are consistently predicted by all three methods are much higher than overall values with an increase of 6.32-6.62 and 5.06-5.46 percentage points for Q_3 and SOV, respectively (Table IV(b)). Similar results are obtained after separating the Q_3 value into the three secondary structure classes: Q_H and Q_E are increased by 4.12-6.03 and 2.88-4.31 percentage points, respectively, while Q_L is increased on average by 5.83-6.20 percentage points.

3.4 Conclusions

In summary, we recognized that a simple consensus approach based on the majority voting of solely three prediction methods can be superior to each of the three methods as well as to complex combinations of more than three single prediction methods as employed in Jpred. Our method proved to work with distinct combinations of different prediction methods on large benchmark sets. Presumably, the success of the method is mainly due to the use of three of the currently best single methods and the noise-removing properties of a consensus approach, which helps to ignore the errors of single methods. We believe that any three state-of-the-art prediction methods can be used for the consensus. The method is computationally less expensive than other consensus approaches and has the advantage of not requiring the calibration of parameters.

4
Improving Structure Prediction by Distance Constraints

This chapter describes a comprehensive analysis of methods for improving the success rate of protein fold recognition, also known as threading in protein structure prediction. The methods utilize a small number of additional distance constraints between protein residues, which can be obtained by experimental techniques such as mass spectrometry or NMR spectroscopy. As detailed below, a post-filtering step with novel scoring functions incorporating measures of constraint satisfaction is applied to improve ranking lists of threading alignments. This new approach combining structure prediction and experiments can be especially valuable for rapid structure determination and the validation of protein models. It partially originated from the collaboration with experimental partners studying the structures of specific proteins such as ataxin-3 causative of neurodegeneration (Chapter 2).

In the following, the computed results show that, based on a small representative benchmark set, the fold recognition rate can be improved significantly by up to 30% from about 54%-65% to 77%-84%, approaching the maximally attainable performance of 90% estimated by structural superposition alignments. This gain in performance adds about 10% to the recognition rate already achieved with cross-link constraints only. Notably, this work was first presented at the German Conference on Bioinformatics (GCB) in 2001, and an extended version was subsequently published as journal article (Albrecht *et al.*, 2002).

4.1 Introduction

The threading approach predicts the three-dimensional protein structure by comparing and aligning representative template protein structures to the amino acid sequence of the target protein (Eisenberg, 1997; Finkelstein, 1997; Jones, 1997; Rost *et al.*, 1997; Zhang *et al.*, 1997; Koehl and Levitt, 1999; Sternberg *et al.*, 1999). This is in contrast to *ab initio*/*de novo* structure prediction endeavors that aim at the construction of structural protein models primarily based on physicochemical interactions between amino acids (Schonbrun *et al.*, 2002). The computation and subsequent evaluation of threading alignments usually gives a sequence-structure similarity score, which is the result of applying a scoring function for each alignment. According to the fold recognition protocol, the alignments obtained are then ranked by their respective scores. Hereby this procedure yields a ranking list of target-template alignments. The best-scoring alignment should identify the correct template structure and its corresponding fold class. It is assumed to be most compatible with the target sequence and to constitute a meaningful model for the yet unknown structure of the target sequence.

However, the problem of developing an accurate scoring function is still unsolved for distantly related target and template proteins sharing the same fold. Especially, making the scoring scheme reflect diverse biological constraints seems to be a difficult task. Thus, threading methods based solely on sequence information of the target protein often fail. To remedy the inherent shortcomings of the scoring function and, at the same time, to enhance the credibility of the proposed models, it becomes necessary to exploit more biological knowledge on the target protein in the prediction process. The additional information to include into the threading procedure could consist of specific constraints for the computed alignments. Such constraints can be obtained from experimental data, for instance, as distances between atoms of protein residues. They may be measured by protein cross-linking reagents functioning as molecular rulers (Figure 21) in mass spectrometry (MS) or by NOE (nuclear Overhauser effect) restraints in NMR spectroscopy. The utilization of additional constraints is expected to lead to more accurate fold recognition results and an improvement in prediction and alignment quality.

This combined approach can be particularly beneficial for structural genomics projects that intend to determine many protein structures in short time. Experimental results that would give insufficient data for the complete structure determination if taken alone may already yield enough constraints to support protein structure prediction considerably using the threading method. The constraints applied to the set of predicted structures can help in selecting and validating the more plausible models of the target protein. This procedure may accelerate the overall structure determination because the amount of data that is usually required can be reduced. Additionally, the threading models may render feasible the structure determination of proteins of larger size, for instance, by NMR spectroscopy, which is still limited to protein sizes up to about 30kD.

There have been few publications on threading with experimental constraints. Recently, Xu and colleagues described how to incorporate NOE restraints as distance constraints in their threading program PROSPECT (Xu *et al.*, 2000). They use a larger number of constraints than in our study described below in order to compute better

protein models and to support NMR structure determination. They also give some examples how the threading performance can be improved by the incorporation of partial structural information about the position of disulfide bonds and active sites in the target protein (Xu and Xu, 2000). In contrast, Young and coworkers demonstrated the feasibility of rapid structure determination by the combination of intramolecular cross-links measured by mass spectrometry with threading results, using solely a single target as model protein (Young *et al.*, 2000). Distance constraints derived from the measured cross-links are used to rank the set of computed threading models by the constraint error. This error is the sum of the distance deviations between the target structure and the models.

In the following, we present a comprehensive analysis of methods for improving the fold recognition rates in the threading approach by utilizing a small number of additional distance constraints from experiments. Here, in contrast to our previous paper focusing solely on simulated cross-link constraints (Hoffmann *et al.*, 2002), we included sets of NOE constraints and used different scoring functions.

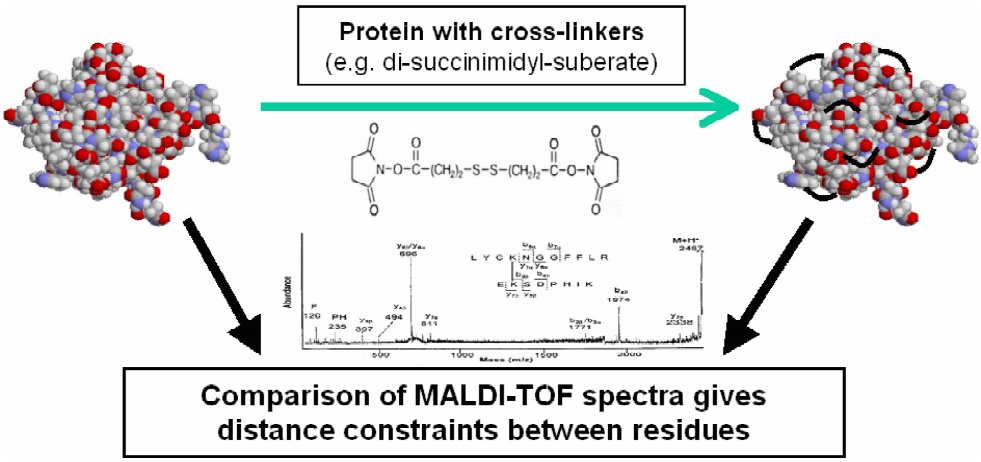


Figure 21. Schematic view of the determination of experimental distance constraints by mass spectrometry.

4.2 Materials and Methods

4.2.1 Benchmark Set

To evaluate the performance of our methods, we used a standard template library of representative protein structures. This comprehensive benchmark set taken from the Hobohm96-25 database (Hobohm and Sander, 1994) consists of 251 single-domain proteins, whose pairwise sequence identity is below 25%. This makes it hard to compute biologically reasonable alignments based solely on sequence information. The SCOP annotation (Murzin *et al.*, 1995) of the proteins is used to divide the library into structural fold classes, which results in 11 classes containing 5 to 11 members. The minimum of 5 members was chosen to allow a reasonable analysis of ranking results for members of the same fold class. These 11 folds contain altogether 81 target proteins

(Table V), which are threaded against the complete template library. As described in (Thiele *et al.*, 1999), this benchmark set represents a demanding task for threading methods because fewer than 50% of the residues in proteins with identical folds can be superposed within 3.0Å in most cases.

Table V.

Each fold class is described by the count of members, the α/β-type class, the SCOP name, and the minimum and maximum sequence length of all proteins contained in the class.

fold	count	type	SCOP name	min	max
1	8	α	4-helical cytokines	119	172
2	7	α	Four-helical up-and-down bundle	106	154
3	5	-	Cystine-knot cytokines	85	112
4	6	$\alpha+\beta$	Ferredoxin-like	81	143
5	5	α/β	Flavodoxin-like	128	302
6	11	α	Globin-like	136	172
7	6	α/β	α/β-Hydrolases	265	534
8	5	β	Lipocalins	131	176
9	7	β	OB-fold	98	280
10	11	α/β	TIM β/α-barrel	228	490
11	10	β	Viral and capsid proteins	175	548

4.2.2 Constraint Filter

The constraint filter that checks the violation of target distance constraints requires both target residues related by the constraint to be aligned to template residues within the given distance (Figure 22).

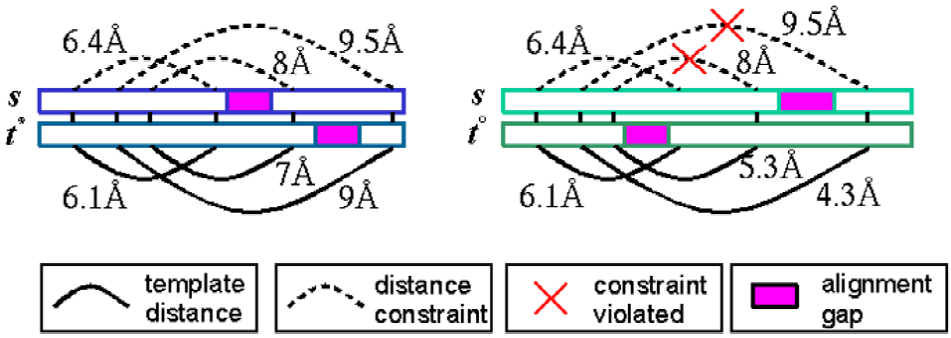


Figure 22. Alignments of a target sequence *s* with selected distance constraints to two templates t^* and t° annotated with the corresponding distances in their structures. Assuming a distance tolerance δ = 2.5Å and a position tolerance σ = 0, all distance constraints in the left alignment are satisfied, whereas two distance constraints are violated in the right alignment.

Formally, let the target sequence s be represented by the amino acids $s_1, s_2, ..., s_n$ and the sequence of the template structure t by the amino acids $t_1, t_2, ..., t_m$. We are given a set of distance constraints $C = \{(s_i, s_j, d_{s_i s_j}) \mid s_i, s_j \in s, d_{s_i s_j} \in \Re^+\}$, where $d_{s_i s_j}$ is the atomic distance between the residues s_i and s_j. We then regard some alignment $A(s, t)$ of length z that maps the target segment $s' = s'_1, s'_2, ..., s'_z$ onto the template segment $t' = t'_1, t'_2, ..., t'_z$. Every residue of s contained in s' is either mapped to some residue of t in t' or to a gap. A distance constraint $c = (s_i, s_j, d_{s_i s_j})$ is said to be aligned to two template positions t'_k and t'_l if the two target residues s_i and s_j are aligned to t'_k and t'_l, respectively. If t'_k or t'_l represents an alignment gap, the distance $d_{t'_k t'_l}$ is defined to be $+\infty$.

The stringency of the distance constraint filter in terms of specificity and sensitivity can be adjusted by two tolerance parameters:

- Distance tolerance: the maximum deviation δ from the given distance value.
- Position tolerance: the number σ of sequence positions searched left and right of the aligned template position for matching distances.

The distance tolerance accounts mainly for the inaccuracy of the experimental measurements, but also for small local structural deviations between the target and the template. Because the alignments computed by the threading method do usually not reflect structural similarity perfectly and may be shifted with respect to the standard-of-truth alignments, we set the position tolerance to $\sigma = 4$. This particular parameter setting yields a search interval of size 9 around the aligned residue positions. It allows α-helices to be misaligned by one turn, and β-strands by one or two shifts of two consecutive residues. Generally, the position tolerance can be set to small values for high-quality alignments, while larger values should be chosen for alignments of lower quality.

Now we define that some distance constraint $c \in C$ aligned to t'_k and t'_l is satisfied if the distance between the mapped residues in the template structure closely matches the measured distance constraint in the target structure:

$$\exists p \in \{k-\sigma, ..., k, ..., k+\sigma\},\ q \in \{l-\sigma, ..., l, ..., l+\sigma\}:$$
$$(1 \le p, q \le z) \wedge (|d_{s_i s_j} - d_{t'_p t'_q}| \le \delta)$$

Otherwise the constraint c is said to be violated.

4.2.3 Distance Constraints

We constructed three different sets of distance constraints for each target protein. Since experimental data is not yet available for a large number of proteins, we simulate the distances by simply reading them off resolved target structures contained in the Protein Data Bank PDB (Berman *et al.*, 2000). To store the data for each protein, we used the XML-based ProML specification language (Hanisch *et al.*, 2002).

We always selected distances between C_β atoms in order to be able to compare the results of the three sets (for glycine residues, we computed pseudo C_β atoms). However, it would be readily possible to extract similar distance constraints from real

experimental data measured between sulfur or hydrogen atoms, as it is often the case in MS or NMR, respectively.

For the first set C_m of distance constraints, we mimic the measurement process of intramolecular homobifunctional cross-linkers by MALDI-TOF mass spectra. To this end, we select distances in the range of 8Å to 14Å between aspartate, glutamate, and lysine residues (Green *et al.*, 2001; Hoffmann *et al.*, 2002). The distance tolerance was chosen as $\delta = 2.5$Å, which corresponds to the maximum standard deviation observed in reagent manufacturer and simulated data (Green *et al.*, 2001). Additionally, we included distance constraints derived from disulfide bridges as "natural" cross-linking reagents into our set. The number of distance constraints in C_m was set to 0.1 constraints per residue.

For the second and third constraint set C_{n1} (C_{n2}), we randomly picked 0.1 (0.2) artificial long-range NOE restraints per residue of distances 4Å to 6Å with a minimum sequential separation of three adjacent residues. This corresponds to sparse data that is acquired early in the NMR structure determination process (Skolnick *et al.*, 1997; Standley *et al.*, 1999; Bowers *et al.*, 2000). The distance tolerance of NOE enhancements was chosen to be more stringent with $\delta = 1.0$Å.

4.2.4 Scoring Functions

We apply the distance constraint filter to the ranking list of target-template alignments computed by the threading program. We investigated four different scoring functions that validate the structure models given by the target-template alignments by checking the experimental distance constraints for violation. Two functions simply count the number of satisfied and violated distance constraints. In contrast, the other two functions use the idea that distance constraints conserved among the members of a fold class should receive more weight in the summation.

For this purpose, we extracted structurally conserved cores of each fold class contained in the template library by means of structural superpositions and multiple alignments. We assigned a higher weight to a distance constraint if its both ends lie within a core region of the template. In practice, we assume a weight of 5, which amounts to scoring the distance constraint inside a core region five times as high as a non-core constraint.

In the following, let n_a be the size of a set containing all experimental distance constraints of a target *s,* and let n_f be the number of fulfilled distance constraints in some target-template alignment $A(s, t)$. The threading score of $A(s, t)$ is denoted by r. For structural cores, the weighted sum computed from n_f and n_a is given by w_f and w_a, respectively. Then the four scoring terms are defined as follows:

(1) $s_{fa} = \mathrm{n_f} / (\mathrm{n_a} + 1)$

(2) $s_{rfa} = r + \rho_{fa} \cdot s_{fa}$

(3) $s_{wfa} = \mathrm{w_f} / (\mathrm{w_a} + 1)$

(4) $s_{rwfa} = r + \rho_{wfa} \cdot s_{wfa}$

The parameters ρ_{fa} and ρ_{wfa} in the linearly combined terms s_{rfa} (2) and s_{rwfa} (4), respectively, were chosen such that the threading score r and the constraint scores s_{fa} (1) and s_{wfa} (3) contribute roughly equally to the combined value if about 50% of all potential constraints are fulfilled in the top-ranked threading alignments. As further investigations into the optimal choice of these parameters revealed, this assumption usually maximizes the fold recognition rate in case of our sets C_m and C_{n1} with 0.1 constraints/residue. However, the exact choice is not that relevant as maximum performance is achieved over a relatively wide interval of parameter values around the chosen values ρ_{fa} = 1600 and ρ_{wfa} = 1600. Interestingly, maximum performance for the set C_{n2} with 0.2 constraints/residue was reached with the double parameter values ρ_{fa} = 3200 and ρ_{wfa} = 3200. This may be due to the fact that, in contrast to the 0.1-sets, the same absolute amount of fulfilled constraints is already obtained if only about 25% of all possible constraints of the 0.2-set are satisfied.

4.2.5 Threading Alignments

We used the threading program 123D (Alexandrov *et al.*, 1996) and its extension 123D* (Sommer *et al.*, 2002) incorporating profile methods to predict the structure of all 81 target proteins in our benchmark set by threading them against the template library. The result of a 123D run for each target is a list of all 250 template proteins (excluding the target protein itself), which is ranked by the target-template alignment scores. We employed three different parameter sets to compute global alignments of increasing quality:

P_1: Standard parameter set with gap insertion and extension cost 20 and 0.8.

P_2: Optimized parameter set as published in (Zien *et al.*, 2000).

P_3: Improved threading with frequency profiles.

In addition, we generated standard-of-truth alignments using structural superpositions computed by the program SARF2 (Alexandrov, 1996).

4.3 Results and Discussion

4.3.1 Post-Filter for Alignments

According to the fold recognition protocol, all target proteins are classified and their predicted folds compared with the true folds in order to calculate a recognition rate. The fold of a target structure is correctly recognized if the best-scoring template in the ranking list (not containing the target protein) belongs to the target fold class. In the special case that more than one template reaches the identical best score, the target protein is counted as correctly recognized only if all best-scoring templates share the same target fold.

In order to improve the fold recognition rate, we apply the distance constraint filter to the 123D results in a post-filtering step. This procedure amounts to a re-evaluation and validation of the computed threading alignments with one of the four scoring

functions s_{fa}, s_{rfa}, s_{wfa}, and s_{rwfa}. In this way, the aligned templates are checked for the violation of distance constraints contained in one of the three sets C_m, C_{n1}, and C_{n2}.

Table VI shows the fold recognition rate for each pair of scoring function and constraint set, and compares it with the performance of the 123D scoring function r for every threading parameter set P_1, P_2, and P_3.

Table VI.

Recognition rates for 123D alignments, dependent on the scoring functions r, s_{fa}, s_{rfa}, s_{wfa}, s_{rwfa}, the distance constraint sets C_m, C_{n1}, C_{n2}, and the threading parameter sets P_1, P_2, P_3.

		r 123D	s_{fa} constr.	s_{rfa} 123D+constr.	s_{wfa} core-constr.	s_{rwfa} 123D+core-const
C_m	P_1	54%	21%	47%	49%	58%
MS	P_2	60%	23%	63%	54%	73%
0.1-set	P_3	65%	20%	56%	43%	68%
C_{n1}	P_1	54%	46%	57%	64%	67%
NOE	P_2	60%	52%	72%	65%	79%
0.1-set	P_3	65%	57%	67%	75%	77%
C_{n2}	P_1	54%	52%	57%	72%	73%
NOE	P_2	60%	62%	75%	73%	84%
0.2-set	P_3	65%	64%	73%	79%	80%

Apparently, the combined scoring functions s_{rfa} and s_{rwfa} outperform both the 123D scoring function r and the simpler functions s_{fa} and s_{wfa}. The reason may be that the distance constraint score exploits an orthogonal measure (the fraction of satisfied constraints) to complement the threading score r. While the 123D score gives an empirical estimate of the alignment quality, number of gaps, sequence and structural similarity between the target and template, the distance constraint score serves as a high-quality indicator of structural accuracy and thus of indispensable properties of a correct template structure.

Furthermore, the scoring functions s_{wfa} and s_{rwfa}, which are based on structural cores, also lead to a substantial increase of the recognition rate when compared to the functions s_{fa} and s_{rfa}. Moreover, the data shows that alignment quality is crucial for the improvement potential of the scoring function.

Apart from that, the comparison of the performance of the distance constraint sets C_{n1} and C_{n2} reveals the expected relationship that the recognition rate increases with the number of distance constraints. More specifically, it appears that the distance constraints C_{n1} obtained by NMR perform slightly better than the MS constraint set C_m.

This observation may be explained by the fact that the maximum of the atomic distance distribution for some protein is often attained towards values larger than 8Å. Thus, the satisfaction of a typical NOE distance constraint of 4Å to 6Å is more significant for the correctness of the chosen template structure. Another, but only minor, effect (as revealed in further tests not detailed here) is the different accuracy of the MS and NMR measurements expressed by the distance tolerance parameter. A third reason might be that the distance constraints from cross-linkers are biased towards certain amino acids, which is not the case with NOEs.

Figure 23 depicts how the increase of the recognition rate also depends on the fold class. It is particularly interesting that the targets with ferredoxin-like fold (column number 4 in Figure 23) are much better recognized by the combined function with 79%. In contrast, the 123D scoring function usually fails completely.

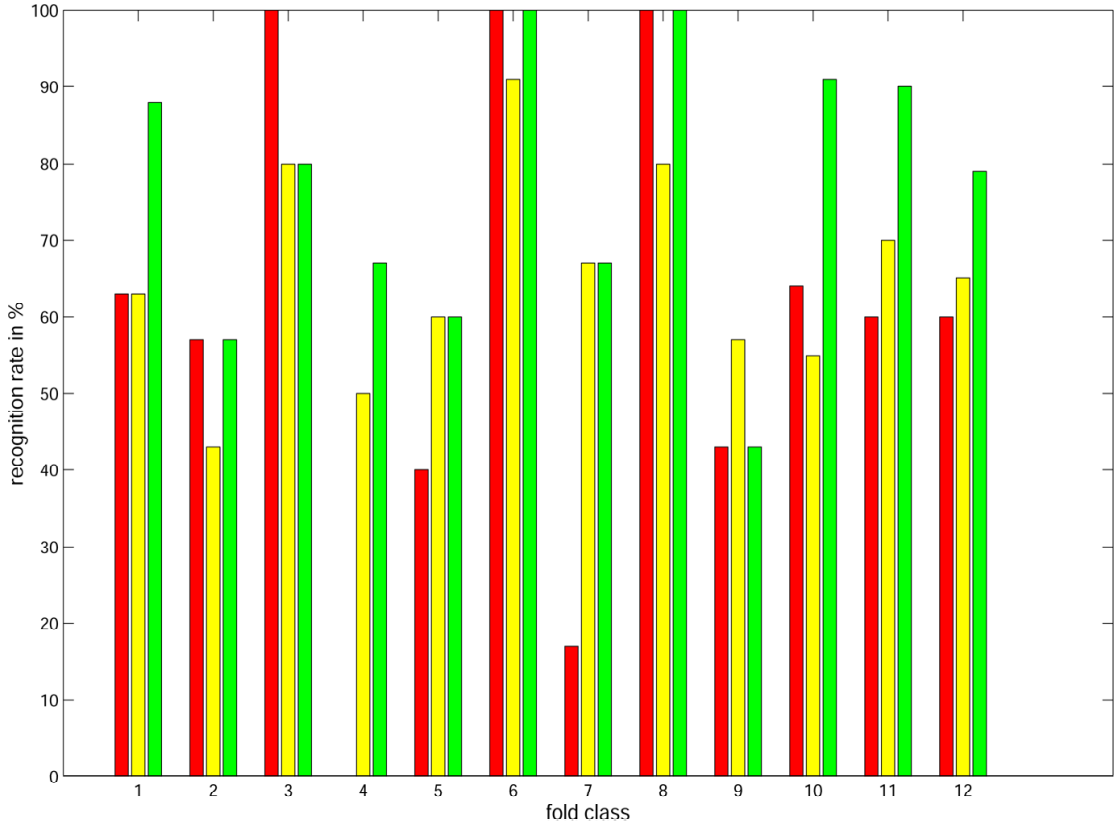


Figure 23. Recognition rates per fold class, using the 123D parameter set P_2 and the NOE distance constraint set C_{n1}. Each triple of bars shows the performance of the 123D scoring function r (red) and of the constraint scoring functions s_{wfa} (yellow) and s_{rwfa} (green), both of which are based on structural cores. The 12th triple of bars depicts the overall recognition rates averaged over all 11 fold classes.

4.3.2 Comparison with Superpositions

In order to determine an upper bound on the recognition rate that can be maximally achieved, we evaluated the standard-of-truth SARF superposition alignments for our benchmark set with the scoring functions s_{fa} and s_{wfa}. In addition, we used the number of aligned residues as an artificial scoring function a to obtain an estimate of the best achievable recognition rate. This is possible because SARF aligns only those residues that are structurally superposable within a predefined threshold of 3.0Å. In general, the more residues are aligned, the closer a structural relationship between the target and template protein can be assumed.

The results are given in Table VII, and some of them are illustrated in Figure 24. As shown, we cannot hope to reach a recognition rate much higher than 90% on our particular protein benchmark set because some structural similarities are stronger between different SCOP fold classes than within the same class. In agreement with the observations described in the previous section, we notice again an increased fold recognition rate among the distance constraint sets C_m, C_{n1}, C_{n2}, and that the scoring function s_{wfa} based on structural cores performs better than the simpler function s_{fa}. It is also striking that the improved alignment quality produced by the SARF program yields higher recognition rates.

Table VII.

Recognition rates for SARF alignments, dependent on the scoring functions a, s_{fa}, s_{wfa} and the distance constraint sets C_m, C_{n1}, C_{n2}.

		a aligned res.	s_{fa} constr.	s_{wfa} core-constr.
C_m	MS 0.1-set	90%	65%	85%
C_{n1}	NOE 0.1-set	90%	79%	90%
C_{n2}	NOE 0.2-set	90%	85%	95%

In summary, it is remarkable that the fold recognition rates of our combination methods already approach the maximal fold recognition rate of about 90%. The best fold recognition rates shown in Table VI lie between 68% and 73% for MS (in agreement with (Hoffmann *et al.*, 2002)) and between 77% and 84% for NMR constraints. This is a significant improvement of up to 30% as compared to the best 123D* threading method with 65% recognition rate.

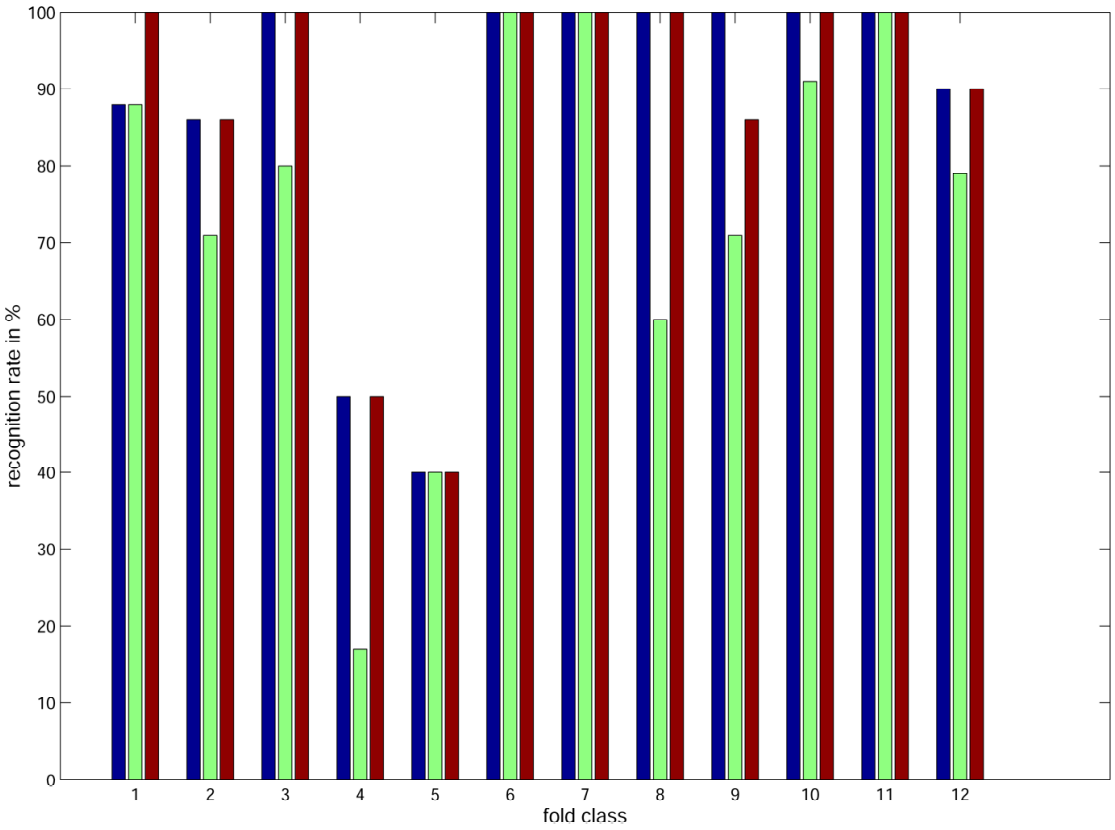


Figure 24. Recognition rates per fold class, using SARF alignments and the NOE distance constraint set C_{n1}. Each triple of bars shows the performance of the scoring function *a* (blue) and of the constraint scoring functions s_{fa} (green) and s_{wfa} (red), the latter of which is based on structural cores. The 12th triple of bars depicts the overall recognition rates averaged over all 11 fold classes.

4.3.3 Larger Benchmark Sets

Our findings in the analysis detailed above are supported by additional results on a larger benchmark set. The template library, which is also used as the set *PDB40D* in (Sommer *et al.*, 2002), consists of 2808 SCOP domains (Murzin *et al.*, 1995) in 540 fold classes. The maximal sequence identity as determined by the ASTRAL server (Brenner *et al.*, 2000) is 40%. We selected all 1613 single-domain target protein chains covering 283 fold classes from the set *PDB40C,* in which each target chain has at least another template partner of the same fold class in the library.

We computed global alignments between the target and template sequences with the threading parameter set P_2, which performed well on the smaller benchmark set as described above. We applied the distance constraint post-filter for the three sets C_m, C_{n1}, and C_{n2} to the threading results, using the scoring functions s_{fa} and s_{rfa}. We did not use scoring functions based on structural cores because the cores were not available for each fold class. Generally, the construction of structural cores for fold classes of template

domains is difficult because of large structural variations between certain superfamilies of the same fold class.

Furthermore, we included the use of a confidence function that computes the "raw score gap" as described in (Sommer *et al.*, 2002) for each threading prediction by the program 123D. This confidence function gives a quantitative estimate on the reliability of the target fold prediction. The raw score gap is the difference between the best threading score of the top-ranking template fold to the score of the next template fold in the ranking list. The introduction of a confidence threshold T discriminating between reliable and non-reliable predictions yields a modified fold recognition protocol. In this protocol, the fold of a target chain is regarded as reliably predicted if the computed raw score gap is at least as large as a given threshold. In that case, the ranking list of threading alignments for the target chain is not subject to post-filtering procedures. If the confidence value is below the threshold, the computed ranking list is re-evaluated by a constraint scoring function. This approach reduces the number of re-evaluations and thus avoids unnecessary experiments to obtain distance constraints.

The fold recognition rates on the larger benchmark set are shown in Table VIII. As already observed for the smaller benchmark set, the NOE distance constraint sets C_{n1} and C_{n2} perform better than the MS constraint set C_m and improve the recognition rate by about 5%. The application of the confidence function is important for increasing the recognition rates substantially for the set C_m. However, further improvements of the recognition rate could be expected by constraint scoring based on structural cores. In the following, the confidence thresholds T_m, T_{n1}, and T_{n2} assumed for the constraint sets C_m, C_{n1}, and C_{n2}, respectively, are chosen such that their application results in a maximum recognition rate after post-filtering. If several confidence thresholds lead to the same maximum recognition rate, the smallest threshold is used.

Table VIII.

Recognition rates for 123D alignments on a larger benchmark set with threading parameter set P_2, using the scoring functions r, s_{fa}, s_{rfa} and the confidence thresholds T_m = 139, T_{n1} = 340, T_{n2} = 340 for the distance constraint sets C_m, C_{n1}, C_{n2}, respectively.

		r 123D	s_{fa} constr.	s_{rfa} 123D+constr.	s_{rfa} with confidence thresholds
C_m	MS 0.1-set	71.73%	16.21%	71.61%	73.16%
C_{n1}	NOE 0.1-set	71.73%	35.47%	76.26%	76.63%
C_{n2}	NOE 0.2-set	71.73%	44.77%	76.88%	77.12%

As depicted in Figure 25 and Figure 26, a lower confidence threshold decreases the number n_l of target chains that are 'lost', i.e., that were correctly recognized by threading but not any more by post-filtering. In contrast, the number n_g of targets that are 'gained', i.e., that can be solely recognized by post-filtering, is sustained on a very high level over a large interval of threshold values. As can be seen by the comparison of the plots of the overall number $n_d = n_g - n_l$ of additional recognized target chains, it is advisable to choose a threshold T_m below the thresholds T_{n1} and T_{n2} because post-filtering with MS distance constraints does not seem to work as well as post-filtering with NOE restraints. Thus, only predictions for targets with rather low confidence values should be included in the MS re-evaluation.

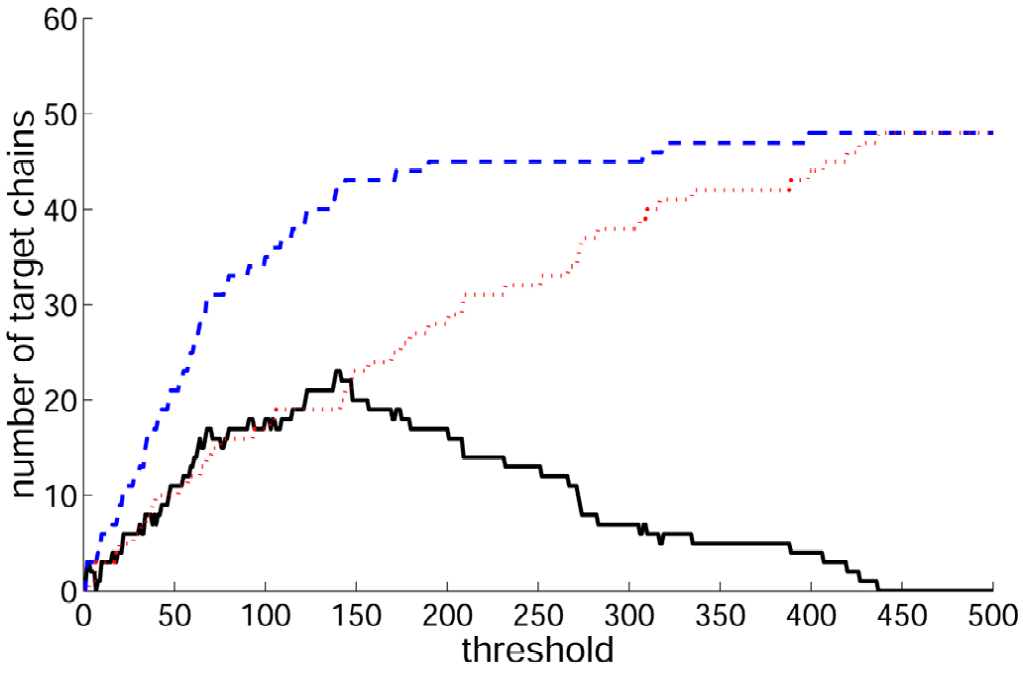


Figure 25. Numbers n_l (red dotted line) and n_g (blue dashed line) of 'lost' and 'gained' target chains together with the overall number n_d (black solid line) in dependence of the chosen confidence threshold T_m after post-filtering with the distance constraint set C_m.

In general, the confidence measure appears to be a very good indicator of the threading prediction quality despite its apparent simplicity. Thus, this indicator helps to avoid post-filtering procedures for target protein chains whose fold is already predicted reliably. At the same time, the number of additional time-consuming and expensive experiments to collect distance constraints is decreased in practice. In particular, only 29.20% (40.55%) of all target chains are included into the re-evaluation in case of an optimally chosen confidence threshold $T_m = 139$ ($T_{n1} = T_{n2} = 340$). This threshold gives the highest overall numbers of additionally recognized targets after post-filtering.

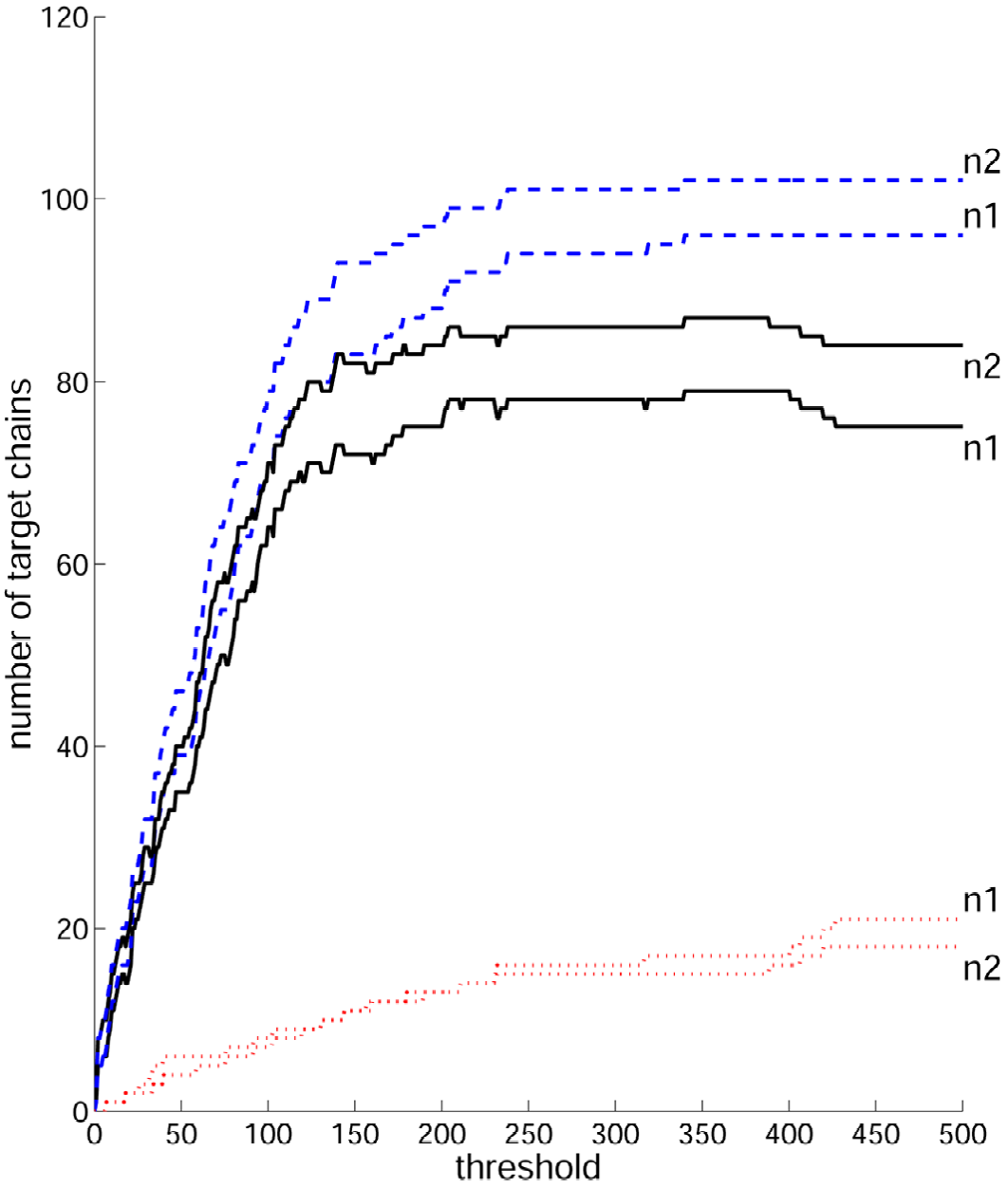


Figure 26. Numbers n_l (red dotted line) and n_g (blue dashed line) of 'lost' and 'gained' target chains together with the overall number n_d (black solid line) in dependence of the chosen confidence thresholds T_{n1} and T_{n2} after post-filtering with distance constraint sets C_{n1} and C_{n2}, respectively.

4.4 Conclusions

We demonstrated that a small number of experimental distance constraints is already sufficient to improve the fold recognition rate considerably. The distance constraints can be collected rapidly by experimental techniques such as mass spectrometry and

NMR spectroscopy. The introduction of a confidence measure on the reliability of threading predictions helps to reduce the number of necessary experiments. Our observations should be particularly beneficial for current efforts in high-throughput structure determination on a genomic scale.

Even in the hard case of low sequence similarity or missing homologs, suitable template structures can be selected reliably by post-filtering and re-ranking the list of threading alignments. Better performance was achieved with combined scoring functions that consider a small number of additional constraints and their location in structural cores. Our comprehensive analysis also indicated that the application of NOE constraints improves upon the recognition rate achieved by exploitation of MS cross-link constraints. In addition, we showed that the success of the post-filtering approach depends considerably on the alignment quality. Thus, the use of advanced threading programs is recommended to generate high-quality alignments.

In future work, real experimental data should be applied and other scoring functions may be explored together with some significance value for the satisfaction of distance constraints. Alternatively, a new threading algorithm that allows the direct incorporation of distance constraints into the alignment computation could be developed to improve the alignment quality and, at the same time, to reduce the overall computation time.

5 Decomposing Protein Networks into Domain Interactions

The application of novel experimental techniques has produced enormous amounts of protein interaction data. Important information on the structure and cellular function of protein-protein interactions in regulatory and metabolic pathways forming complex networks are often obtained from the evolutionarily conserved domains contained in interacting proteins. Therefore, this chapter presents the design of an elaborate plugin for Cytoscape, a free software platform for gene and protein network visualization. This plugin decomposes interacting proteins into their respective domains in order to compute a putative network of the corresponding domain-domain interactions. To this end, the Cytoscape network graph of proteins has been extended by additional node and edge types for domain interactions, including different node and edge shapes as well as coloring schemes for changing the visualization according to user preferences.

The development of this automatic decomposition procedure stems from the comprehensive, necessarily manual, analysis of the protein interaction network around the yeast homologs of ataxin-2 and ataxin-7, both of which are causative of neurodegeneration (Chapter 2). The design of the plugin has been published in the recent proceedings of the European Conference of Computational Biology (ECCB) (Albrecht *et al.*, 2005b). It is also available for download from our web site. Part of the implementation work has been performed by Carola Huthmacher for her diploma thesis. Further software extensions and future applications of domain-domain interactions in combination with protein structure prediction are outlined at the end of the chapter.

5.1 Introduction

Frequently, protein binding is characterized by specific interactions of evolutionarily conserved domains (Bornberg-Bauer *et al.*, 2005), which are incorporated into different proteins by genetic duplications and rearrangements (Vogel *et al.*, 2004). Globular domains are defined as structural units of fifty and more amino acids that usually fold independently of the remaining polypeptide chain to form stable, compact structures (Orengo and Thornton, 2005). Even if the structure of a domain is unknown, it is still possible to define domain boundaries in many cases based on homology criteria using sequence data (Bateman *et al.*, 2004). In general, domains can be found alone or in conjunction with other domains and intrinsically disordered, mainly unstructured, protein regions connecting globular domains (Dunker *et al.*, 2005).

Novel high-throughput techniques have generated large networks of protein-protein interactions (Cusick *et al.*, 2005), which need to be analyzed further using additional functional and structural data (Bork *et al.*, 2004). Important information on the cellular function of specific protein interactions and complexes can often be gained from the known functions of the interacting protein domains (Pawson and Nash, 2003). Domains may contain binding sites for proteins and ligands such as metabolites, DNA/RNA, and drug-like molecules (Xia *et al.*, 2004). Therefore, it is useful and often even necessary to decompose protein-protein interactions into their constituent domains to answer the following questions: Why and how do two proteins interact (Figure 27)? Which domains are responsible for this interaction or the binding of ligands? To address these issues, our approach allows to functionally characterize protein interactions further on the domain level. In Figure 27, it also becomes apparent that this view supports modeling and investigating the spatial structure of protein domain complexes (Park *et al.*, 2001; Aloy *et al.*, 2005).

Notably, it may be confusing that the term ‘domain’ is commonly used in two slightly different meanings. In the context of domain databases such as Pfam (Bateman *et al.*, 2004) and InterDom (Ng *et al.*, 2003), a domain basically consists of a set of homologous sequence regions. In contrast, a single protein may contain one or more domains, which are concrete sequence regions within its amino acid sequence. To draw a parallel to the object-oriented programming paradigm, domain databases provide domain classes, whose instances, the objects, occur in specific proteins. In the past, the topological properties of domain interaction graphs have been studied intensively (Wuchty, 2002; Ye and Godzik, 2004). In such graphs, nodes represent domain classes. Two domain classes are linked by edges if there is at least one protein-protein interaction known in which some protein contains one domain instance of the two domain classes and the interacting protein contains the other domain instance. However, this approach is different from ours as detailed in the following. Our application links two domain instances (and not two domain classes) based on the assumption that both domain instances are responsible for a specific protein-protein interaction.

In order to facilitate research on the molecular basis of an observed or predicted protein-protein interaction, we have designed a tool named DomainNetworkBuilder. It works as Java plugin for Cytoscape, a free open-source software platform for the visualization and analysis of biomolecular networks (Shannon *et al.*, 2003). This plugin

DomainNetworkBuilder decomposes protein networks into domain-domain interactions and generates a new network of interacting domains.

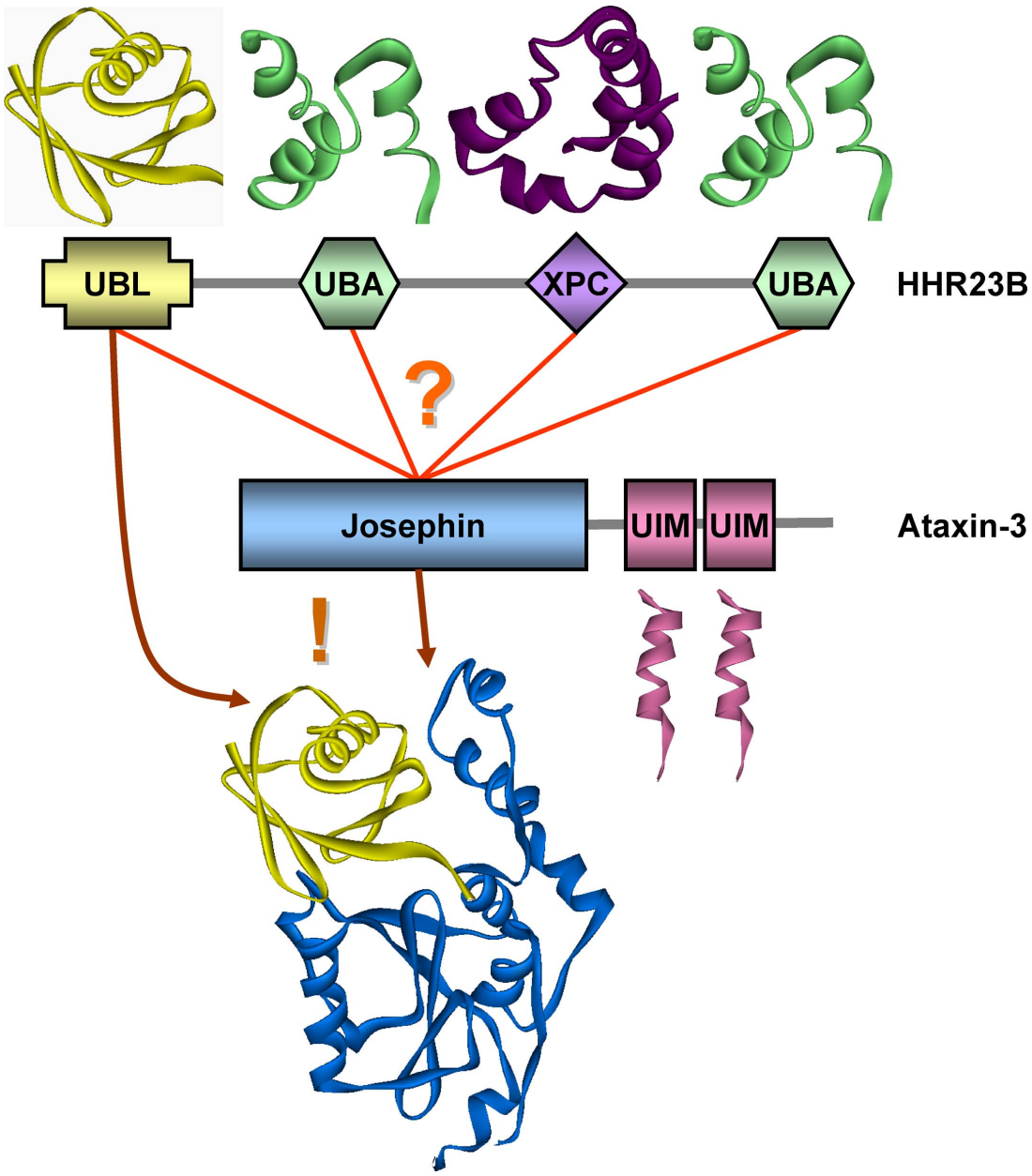


Figure 27. Exemplary interaction between the two human proteins HHR23B and ataxin-3 (cf. Chapter 2). Each protein domain commonly adopts a particular 3D structure and may fulfill a specific molecular function. Generally, the domains responsible for an observed protein-protein interaction need to be determined before further functional characterizations are possible. Here, it is known from experiments that the ubiquitin-like domain UBL of HHR23B (yellow) forms a complex with de-ubiquitinating Josephin domain of ataxin-3 (blue) (Nicastro *et al.*, 2005).

We have also implemented another Cytoscape plugin named DomainWebLinks that provides additional context-dependent web links to Internet resources on domain function and structure. It links protein/domain nodes to InterPro and Pfam, databases of domain families (Mulder *et al.*, 2003; Bateman *et al.*, 2004), to InterDom, a database of putative interacting domains (Ng *et al.*, 2003), and to iPfam and 3did, databases of 3D interacting domains for known, experimentally solved, structures (Finn *et al.*, 2005; Stein *et al.*, 2005). Further web links lead to the Dasty and SPICE viewer of external annotations for protein sequences and structures (Jones *et al.*, 2005; Prlic *et al.*, 2005).

5.2 Materials and Methods

We have established a client-server architecture with the Cytoscape plugin DomainNetworkBuilder working as client. It queries an in-house MySQL database through our Apache web server using a simple XML-RPC protocol and processes the received data through PHP/SQL scripts to create a network of interacting domains (Figure 28).

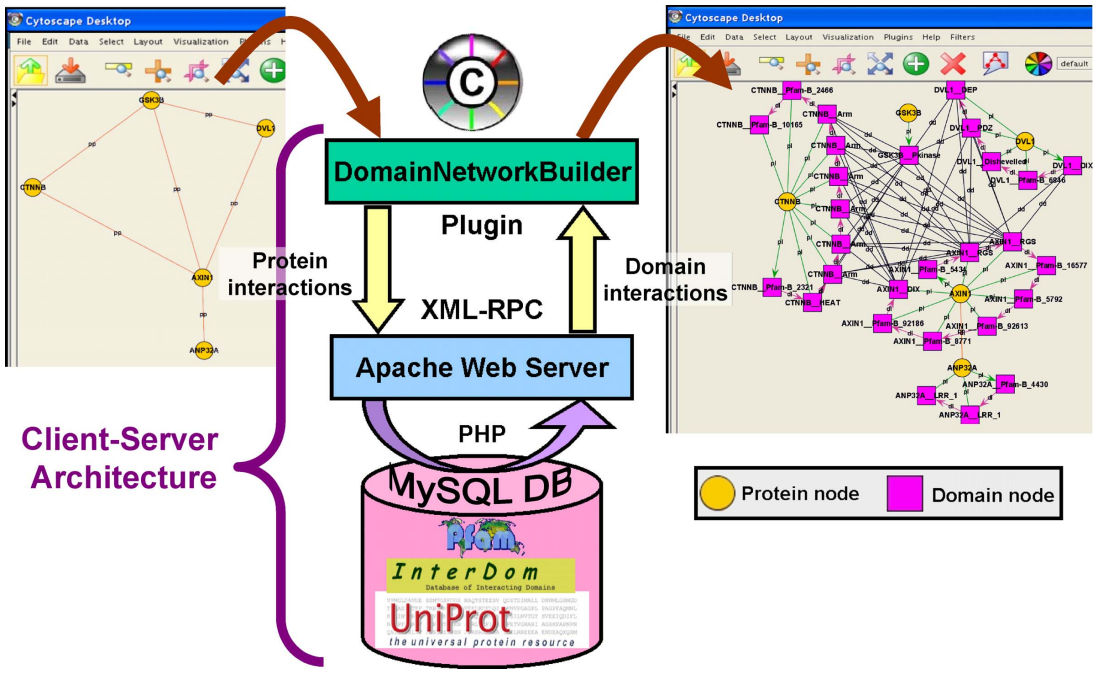


Figure 28. Client-server architecture for the plugin DomainNetworkBuilder in Cytoscape. A protein-protein interaction network loaded into Cytoscape serves as input. After sending these input data to our web server via XML-RPC and retrieving domain information from the MySQL database using PHP/SQL, the plugin creates a new network of interacting domains to output the returned database query results.

The database stores synonyms for each gene/protein name, all protein domains from Pfam (Bateman *et al.*, 2004), a special list of short repetitive Pfam domain motifs, and domain-domain interactions with reliability scores from InterDom, a database of putatively interacting Pfam domains (Ng *et al.*, 2003). Our database already covers all human, fly, worm, and yeast proteins taken from UniProt (Apweiler *et al.*, 2004) and can easily be extended to other species.

After a protein network has been loaded as a graph consisting of nodes and edges, the DomainNetworkBuilder plugin can be executed in Cytoscape. It uses the given protein labels in the network to retrieve the respective Pfam domain architectures and InterDom domain-domain interactions from our MySQL database. It then generates and outputs a domain-domain network as described in the next section. Each protein label needs to be identified in our database and thus should consist of either the standard gene/protein name or the corresponding UniProt accession number of the protein sequence. If two or more proteins share the same label, one of the proteins is arbitrarily selected by our system and a warning message is shown. Another warning message appears if the protein label is not found in our database. In this case, the protein will be handled like a protein that does not contain any Pfam domains.

It is possible to use other known or predicted domain-domain interactions alternatively or additionally to InterDom if a reliability score accompanies each interaction. Thus, we will also provide more recent sets of predicted domain-domain interactions (Liu *et al.*, 2005a; Riley *et al.*, 2005). A manually curated list of repetitive domain motifs was compiled based on the Pfam database field TP containing the keyword ‘repeat’. This word indicates tandem sequence motifs such as HEAT or leucine-rich repeats forming one structural domain.

5.3 Results and Discussion

If a protein contains one or more domains, each domain is represented by a separate node labeled by the Pfam domain name and optionally by the protein name and the start and end position of the domain in the respective protein sequence. Like the interaction type ‘pp’ used by Cytoscape for a protein-protein interaction edge, we have introduced three new edge types for domain nodes (Figure 29 and Figure 30): ‘dl’ for a domain linker between domain nodes of the same protein, ‘pl’ for a protein linker between a protein and domain node of the same protein, and ‘dd’ for a domain-domain interaction between different proteins. All domain nodes of the same protein are linearly connected in a chain of nodes by directed edges (arrows pointing from the N-terminus to the C-terminus). The user can choose whether this chain of domains is linked by a single directed edge to the protein node, which serves as N-terminal anchor, or each domain node belonging to a protein is connected directly to the protein node. The latter alternative may result in a closer local placement of the protein node to its domain nodes if appropriate graph drawing algorithms are applied.

Domain-domain interaction edges between different proteins are created only if the respective interaction score exceeds the overall threshold set by the user. If no domain-domain interaction edge can be established between two interacting proteins, the protein nodes remain connected. Otherwise, the protein-protein interaction edge is removed and

replaced by domain-domain interaction edges. Alternatively, the user can choose to keep the protein-protein interaction edge besides the additional domain-domain interaction edges. If more than one domain-domain interaction edge is possible between two proteins, the user can choose either always to select the edge between two domains with the largest, most reliable, interaction score or to use all possible domain-domain edges (because two proteins could indeed interact through more than two domains). Apart from that, the user can disable the display of protein and/or domains nodes that do not possess any edges of protein-protein and/or domain-domain interactions, respectively.

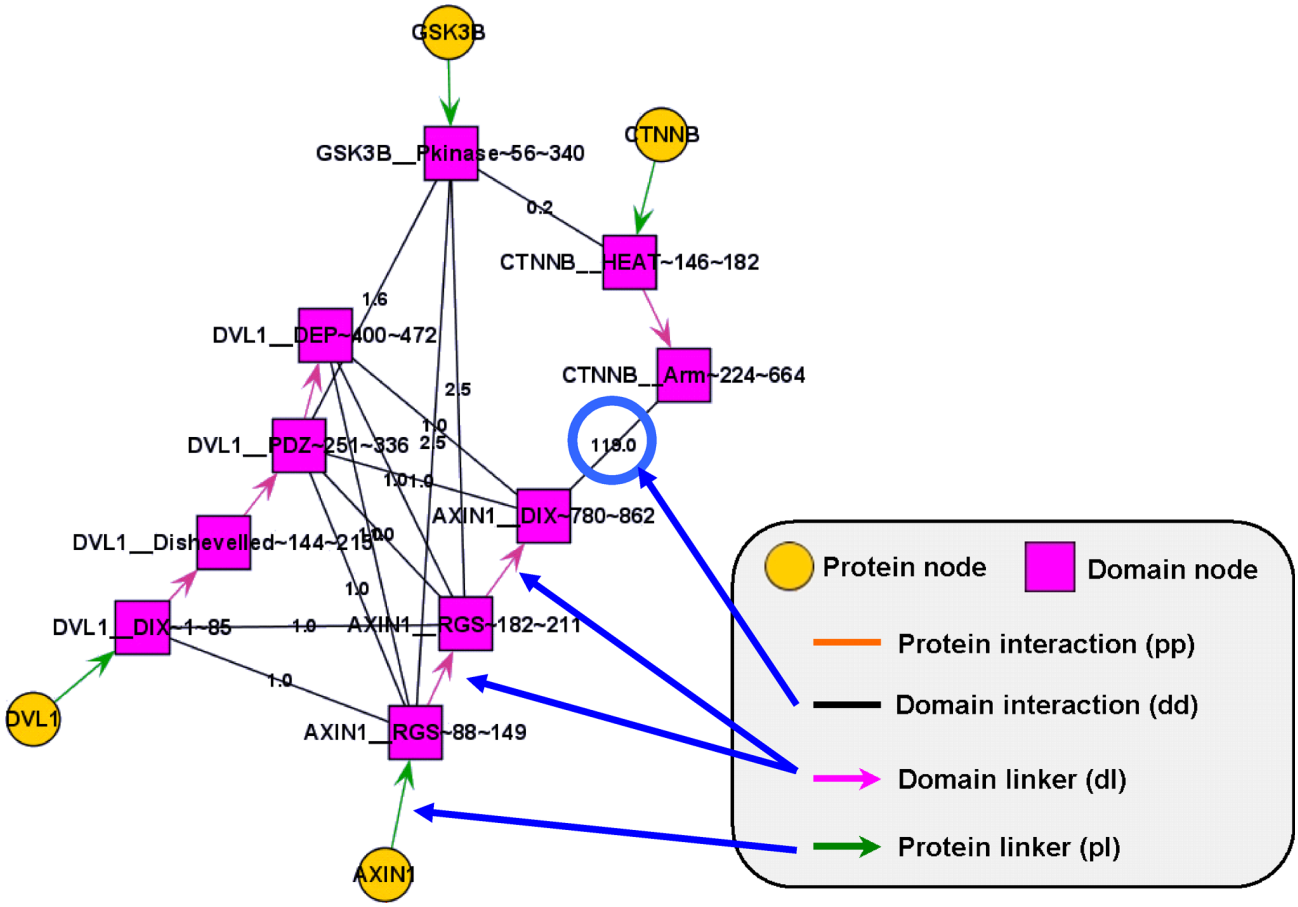


Figure 29. Three new edge types are introduced in a domain-domain interaction network: protein linkers from the protein node to the first (or, alternatively, all) domain nodes, domain linkers between domain nodes of the same protein, and domain-domain interaction edges between different proteins. Here, the latter edges are annotated with the respective InterDom domain-domain interaction score.

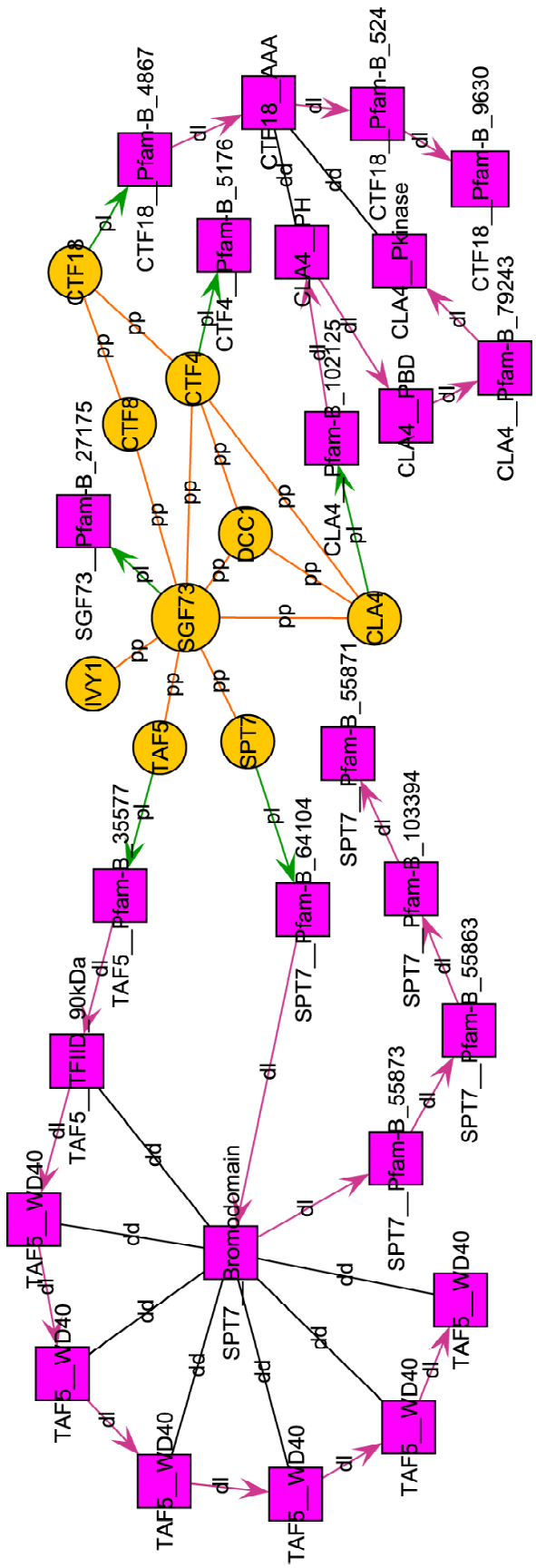


Figure 30. Domain–domain interaction network around SGF73, the yeast homolog of ataxin-7 causative of the neurodegenerative disorder ataxia type 7 (Helmlinger *et al.*, 2004). It is contained in transcriptional SAGA complexes that include TAF5 and SPT7 and show histone acetyltransferase activity. It may also play an important role in sister-chromatid cohesion, which involves an alternative replication factor C complex (CTF4, CTF8, CTF18 and DCC1) and presumably the protein kinase CLA4. Domain nodes are depicted as pink squares, protein nodes as orange circles. Edges are annotated by their respective interaction types.

Adjacent repetitive domain motifs constituting one structural domain (Andrade *et al.*, 2001) need special treatment to avoid confusion of the network image (Figure 31). To select a subset from our manually curated list of ~100 repetitive domain motifs of length up to ~60, the user can set a threshold for the maximum motif length. Consecutive nodes of the same domain motif shorter than this threshold are merged into a single domain node if the distances between the motifs (measured by the number of amino acids) are not larger than a user-defined maximum distance.

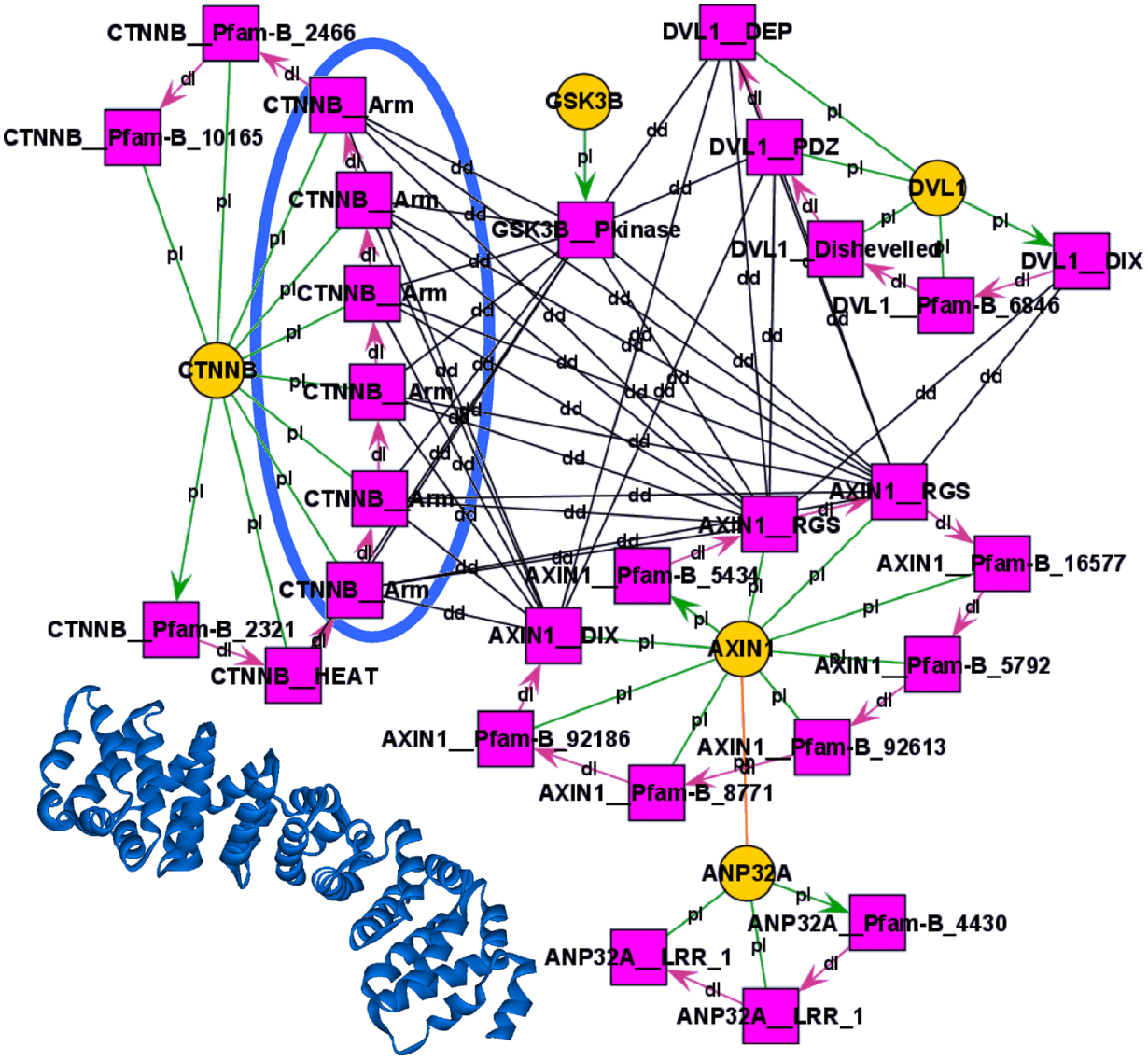


Figure 31. Exemplary repetitive domain motifs named CTNNB_Arm are encircled in blue. Each motif constitutes one bundle of three helices known as armadillo repeat (Coates, 2003). Since these repeats form one structural protein domain (see blue helical structure), the user can choose merging them automatically into a single domain node.

Further options offered to the user are that uncharacterized Pfam-B domains are ignored and not depicted and that edge labels can be changed. Edge labels can consist of the interaction type or, in case of domain-domain interactions, of the InterDom interaction score. Moreover, the coloring schema as well as the different shapes of protein and domain nodes and interaction edges can easily be changed using the visualization tools of Cytoscape. The generated domain network can also be saved in file formats supported by Cytoscape.

5.4 Conclusions and Outlook

Our Cytoscape plugin DomainNetworkBuilder together with the supplementary plugin DomainWebLinks provides tools for investigating and visualizing protein interactions on the more detailed molecular level of domains. It decomposes a given protein-protein interaction network into a network of interacting protein domains. This approach assists in the validation and functional analysis of observed and predicted protein interactions based on domain-domain interactions. It particularly supports the evaluation, planning, and prioritization of further experiments, which are often conducted with fragments of proteins to determine the exact location of binding sites.

Importantly, many human diseases can be traced to aberrant protein-protein interactions, leading to loss or gain of unfavorable protein functions. The molecular cause of a disease may be due to a severe defect of an essential interaction or the formation of a protein complex fulfilling its function at an inappropriate cellular location or time (Ryan and Matthews, 2005). Therefore, the specific inhibition of protein-protein interactions necessitate the accurate determination of their binding domains (Santonico *et al.*, 2005). For instance, disease-associated mutations in human NOD2 cause a constitutively active protein whose N-terminal caspase recruitment domain CARD homodimerizes with another CARD domain contained in the binding partner, the protein kinase RIP2 (see Chapter 2 for more biological details). Thus, it could be interesting to find a drug molecule that blocks this specific CARD-CARD domain interaction to deactivate the disease-causing pathway. Apart from that, knowing the domains responsible for a protein-protein interaction is also a crucial prerequisite for 3D modeling of domain interactions and protein complexes.

Besides minor additions to the plugins that are already in preparation and mentioned in the preceding sections, several major extensions of this work are planned that aim at a more complete description of cellular processes consisting of protein interactions. Proteins do not only interact through structural domains, but may also bind to specific segments in disordered protein regions (Dunker *et al.*, 2005), many of which are known as linear motifs (Neduva and Russell, 2005). Examples are short proline-rich peptides such as those contained in human ataxin-2 outside globular domains (see Chapter 2 for more biological details) that serve as binding sites for domains like SH3 or WW (Zarrinpar *et al.*, 2003). Other proteins working as enzymes may modify certain amino acids (by glycosylation, phosphorylation, ubiquitinylation, sumoylation, etc.) of their interaction partners (Yang, 2005).

Therefore, it is reasonable to generalize the representation of proteins as chains of domains to chains of interacting regions and binding sites. To this end, the

computational deduction of concrete interaction sites needs to be improved. This necessitates novel methods to detect and link the respective sites accurately involving statistical interaction scores computed from available protein interaction data. Increased confidence into the existence of a protein-protein interaction may be justified if putative interactions of corresponding sequence regions contained in the two proteins can be derived reliably. In this context, new databases containing information on experimentally observed interactions of protein fragments and splice variants as well as on posttranslational protein modifications would be useful. This knowledge could basically be extracted from the literature, but automatic text mining approaches for this purpose are still to be devised.

Another problem to solve by updated and refined predictions is that the currently used InterDom database of predicted domain-domain interactions is outdated and not based on all experimental interaction data contained in rapidly growing databases such as BIND (Bader et al., 2003), DIP (Salwinski and Eisenberg, 2003), GRID (Breitkreutz *et al.*, 2003a), MINT (Zanzoni et al., 2002), and IntAct (Hermjakob et al., 2004). Furthermore, InterDom does not take into account the respective interactions of adjacent domains within the same protein and biological differences of interaction patterns between species. Moreover, it would be desirable that a network of interaction sites integrates a more detailed functional annotation of the interaction purpose (activation, inhibition, etc.) as well as spatial and temporal information (cellular localizations, 3D structures, gene expression time points, etc.).

5.5 Further Applications

Several bioinformatics approaches exist that utilize protein interaction data for the prediction of protein function (Huynen *et al.*, 2003). In the past, interaction patterns of proteins were also used either for ranking SCOP folds (Qian *et al.*, 2001) or for assigning them to proteins while ignoring sequence similarity (Lappe *et al.*, 2001). In addition to these applications, a novel approach outlined in the following could exploit protein interaction data to improve fold recognition methods for 3D structure prediction (von Öhsen *et al.*, 2004; Ginalski *et al.*, 2005). Its biological assumption is that homologous proteins share similar structures as well as functions, that is, interaction partners (Espadaler *et al.*, 2005).

This idea requires that a target protein, whose structure is to predict, or its full-length homolog, which can be found by FASTA search (Pearson, 2000), is contained in a protein interaction network. We then regard the SCOP domain folds (Andreeva *et al.*, 2004) of the first-level interaction partners, which are directly linked to the target protein, and of the second-level interaction partners, which are connected to the first-level proteins (Figure 32).

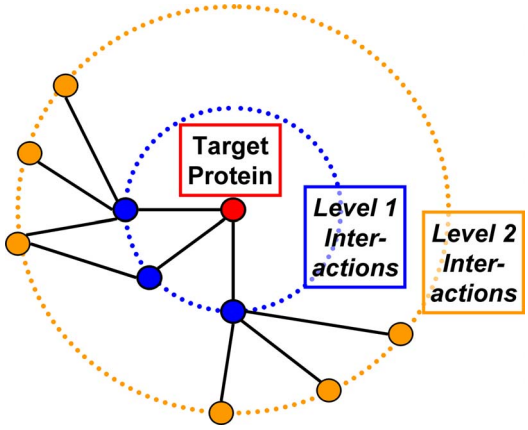


Figure 32. Target protein interactions with direct and indirect binding partners (level 1 and level 2, respectively). The interaction network suggests a set of interacting proteins and/or domains for each target protein.

Protein homologs often interact in a similar fashion and interaction domains may adopt identical folds (Aloy *et al.*, 2003a; Yu *et al.*, 2004). Therefore, the fold of the target protein may be among the SCOP domain folds contained in the first-level interaction partners if the target protein forms homomers or, in particular, in the second-level interaction partners because they interact, like the target protein, with the first-level proteins. Using a joint scoring function, the resulting list of possible domain folds as derived from interaction data for the target protein could be combined with the ranking list of all domain folds computed by a fold recognition method. Alternatively, the resulting fold list could be used to increase the confidence into a fold prediction. Both approaches are closely related to the notion of improving protein structure prediction using experimental data as presented in Chapter 4.

In detail, one of at least four different strategies (1)-(4) described in the following could be applied to obtain a list of SCOP folds for the domains involved in second-level interactions:

(1) Retrieve the Pfam domain architecture of the second-level interaction partners and obtain the SCOP folds of the Pfam domains using a Pfam-to-SCOP domain mapping (protein-protein-protein strategy).

(2) Retrieve the Pfam domain architecture of the first-level interaction partners, then use the InterDom database to collect Pfam domains interacting with the Pfam domains contained in the first-level interaction partners, and obtain the SCOP folds of the collected Pfam domains (protein-protein/domain-domain strategy).

(3) Retrieve the Pfam domain architecture of the target protein, then use InterDom to collect interacting Pfam domains, and obtain the SCOP folds of the Pfam domains that interact with these collected Pfam domains according to InterDom (domain-domain-domain strategy).

(4) Retrieve the Pfam domain architecture of the target protein, then use InterDom to collect interacting Pfam domains, and obtain the SCOP folds of the Pfam domains that are contained in all proteins interacting with these collected Pfam domains (domain-domain/protein-protein strategy).

Of course, the Pfam and InterDom databases could be replaced with other similar databases. The Pfam-to-SCOP mapping could be retrieved from iPfam or computed by an appropriate BLAST search. The performance of the new fold recognition approach on the different SCOP levels could be analyzed using a comprehensive benchmark set of target proteins with mutual sequence identity below 40% (von Öhsen *et al.*, 2004). Different scoring functions and parameter settings including the FASTA E-values and the InterDom confidence score for domain interactions could be tested. In addition, the effects of decreasing or increasing the number of available interactions could be evaluated while excluding hub proteins forming many interactions or including additional predicted protein-protein interactions.

6 Conclusions

This chapter closes the thesis, recapitulating the accomplished work on bioinformatics approaches and outlining future directions. It begins with summarizing remarks on the performed application studies and the development of novel computational methods to support the investigations of biologists and medical researchers. Considering the increasing amounts of heterogeneous biological data produced by novel experimental techniques, bioinformatics perspectives for methodological improvements are briefly discussed at the end of the chapter.

6.1 Summarizing Remarks

Proteins, some of the most important molecules in nature and crucial for the functioning of all living organisms, have been the main focus of this thesis. The molecular study of their structural and functional features using bioinformatics and experimental methods has been the objective of collaborative research on biological and medical questions. Since knowledge of the spatial protein structure provides fundamental insights into protein function and disease processes, successful prediction of protein structure has been a main research field in bioinformatics for many years. Therefore, the work in the course of this dissertation has basically proceeded along two lines:

On the one hand, in joint projects with experimental cooperation partners, comprehensive analyses of medically relevant proteins have been performed to interpret experimental results and to provide a rationale for the design of further experiments. Two distinct types of frequently occurring diseases have been studied extensively, namely, autoinflammatory diseases and neurodegenerative disorders. Using bioinformatics methods, we explored alignments of homologous sequences, characterized domain architectures of proteins, predicted secondary and tertiary structures, located sequence variants and binding sites of proteins or ligands in the 3D

structure, and analyzed protein interaction networks around disease-associated proteins. Hereby, we could often suggest molecular causes of protein defects underlying diseases. We have also discovered novel sequence motifs of functional relevance and additional protein family members. By comparing formerly predicted and now experimentally solved protein structures, we have been able to verify our predictions. This underlined the great value of bioinformatics support in guiding experimentation.

On the other hand, the intensive utilization of bioinformatics tools has revealed current methodological deficits, giving rise to useful ideas for innovative computational approaches. Three of them concerning protein structure have been tackled during this dissertation. First, we showed that a simple consensus formation based on only three secondary structure prediction methods is capable of increasing the prediction accuracy and reliability significantly. Second, we introduced a new approach combining tertiary structure prediction with a small number of additional distance constraints produced by fast experimental techniques like mass spectrometry or NMR spectroscopy. The employed post-filtering procedure for scoring the constraints in computed alignments led to considerable increases of the fold recognition rates. Third, we developed a method for automating the decomposition of protein networks into domain-domain interactions. This method implemented as Cytoscape plugin particularly facilitates the exploration of protein-protein interactions at a more detailed molecular level. However, it is also a useful tool for 3D structure modeling of protein domain complexes.

In conclusion, the cooperation of bioinformatics and experimental groups regarding vital biological and medical problems has created a valuable cross-fertilization. This has proven to be quite effective in successfully answering biomedical questions and in deepening the molecular understanding of disease processes. The joint work has also motivated the development of novel bioinformatics approaches, especially in combining protein structure prediction with further experimental and functional information.

6.2 Methodological Perspectives

When reviewing research developments related to the methodological work, it becomes apparent that the three bioinformatics methods presented in this thesis currently have distinct perspectives:

Since 2001 it has been stated repeatedly in the literature that existing secondary structure prediction algorithms are almost optimal due to the naturally occurring variability of secondary structure and the observed flexibility of protein backbones (Rost, 2001; Huang and Wang, 2002; Crooks and Brenner, 2004). Therefore, dramatic improvements in prediction accuracy may not be expected any more. Notably, the better performance of our approach that combines three secondary structure predictions into a more accurate consensus has also been observed by others later on (Simossis and Heringa, 2004). For instance, the group headed by David Jones has used a combination of a new classifier based on support vector machines (SVMs) with two established methods (Ward *et al.*, 2003). Other studies of the same group with more prediction methods gave very similar results (McGuffin and Jones, 2003).

Regarding the combination of tertiary structure prediction with additional experimental distance constraints, the lack of high-throughput data from mass spectrometry has curbed scientific progress. Insufficient data from our own experimental collaboration partners using this technique has also been the main reason why we have discontinued our project in 2001. In recent years, other research groups have published related bioinformatics approaches, but none of them appears to be applied on large scale (Back *et al.*, 2003). For instance, since their initial work in 2000 (Young *et al.*, 2000), Friedman and coworkers have applied their methods to only two target proteins and developed a sophisticated probabilistic framework for planning mass spectrometry experiments and discriminating 3D structure models on the basis of cross-linking data (Ye *et al.*, 2004).

In contrast, other approaches not only combining structural models with distance constraints such as NOE restraints, but also with additional proteomics data from NMR spectroscopy, have been quite successful meanwhile. They utilize residual dipolar couplings (RDCs) and unassigned chemical shifts, which provide information on atomic bond angles and protein secondary structure, respectively. For instance, Baker's group has concentrated on NMR-based data for rapid protein fold determination in combination with *de novo* structure predictions (Meiler and Baker, 2003; Kim *et al.*, 2004). Similarly, members of Skolnick's lab and Xu's lab have also used NOE restraints and RDCs in the *ab initio* method TOUCHSTONE (Haliloglu *et al.*, 2003; Li *et al.*, 2004) and the protein threading algorithm PROSPECT (Xu *et al.*, 2000; Qu *et al.*, 2004).

Our very recent bioinformatics work, the decomposition of protein networks into interacting domains, has led into the emerging field of interactome analysis with rapidly increasing amounts of data. As described in Chapter 5, several avenues can be followed in future research. For instance, the representation of protein interactions based on globular domains may be generalized to all kinds of binding regions for proteins and other ligands. Another application may be the combination of domain-domain interactions with structural domain fold recognition. Furthermore, additional advances in the reliable derivation of domain-domain interactions and confidence values are required. To this end, more spatial and temporal information could be included into the deduction procedure.

In summary, recent results of computational approaches are quite promising for combining 3D structure prediction with NMR spectroscopy data and for analyzing protein-protein interactions on domain level. However, it may not be rewarding to develop more complicated bioinformatics methods for secondary structure prediction without evidence that the prediction accuracy can be increased significantly in practice. The same may hold true for tertiary structure prediction using distance constraints from mass spectrometry as long as the prevalent experimental obstacles are not being resolved.

Finally, the investigation of static interactomes is topical because it serves as intriguing gateway into computational systems biology (Cusick *et al.*, 2005). This new branch of molecular physiology aims at the quantitative description of dynamic biological processes and their disease-causing malfunctioning at varying levels of cellular detail (Bork and Serrano, 2005). The ultimate goal is a computational zoom lens

for organisms reaching from the physiological interplay of organs to an atomic resolution of molecular interactions (Aloy and Russell, 2005). In biomedical studies, the integration of heterogeneous data using computational means will form the basis for moving from genotype to phenotype (Uetz and Finley Jr., 2005). In this context, recent projects of other research groups have coupled genetic and physical interactions with gene expression and phenotypic profiles (de Lichtenberg *et al.*, 2005; Gunsalus *et al.*, 2005).

Certainly, such complex endeavors will necessitate the close collaboration of bioinformaticians and experimentally working biologists and medical researchers. It will be beneficial to establish interdisciplinary research teams similar to our clinical research group on hepatitis C, involving experts in biology, medicine, computer science and mathematics (Sarrazin *et al.*, 2005). For successful experimentation, computational approaches will be crucial that assist in hypothesis formation and the prioritization and evaluation of experiments. Interesting examples in this context are our integrative studies of ataxin-2 (Ralser *et al.*, 2005a, 2005b), BTNL2 (Valentonyte *et al.*, 2005), and, most recently, selenoproteins (Castellano *et al.*, 2005; Stillwell and Berry, 2005). They encompassed the bioinformatics-supported analysis of as yet uncharacterized proteins and their experimental investigation. However, additional methodological refinements will be required to handle various types of experimental data. Taking quality issues into account, these data need to be incorporated into reliable cellular models together with computational results. Eventually, the design and simulation of those integrated models will enable substantial contributions of bioinformatics to biomedical research.

Summary

Proteins fulfill essential functions in living organisms and are key players in complex dynamic processes inside and between cells. In particular, since proteins are fundamental to life, defects of their structure and function often cause severe diseases. Many computational methods already exist to support molecular protein analyses of interest to experimentally working biologists and medical researchers. However, judging the performance of bioinformatics methods and the quality of their results requires interdisciplinary expertise in informatics and statistics as well as in biology and medicine.

Therefore, it is useful that bioinformaticians do not only develop novel and advanced approaches to solve problems motivated by biomedicine, but also cooperate with bench biologists in applying computational methods. This collaboration is crucial for the accurate interpretation of bioinformatics findings and their effective incorporation into biomedical research, which may yield integrative models containing experimental and computational knowledge for biology and medicine. In return, bioinformaticians gain beneficial insights into the biological and medical aspects and obtain feedback for future improvements and extensions of their applications.

Considering the great importance of a close cooperation between bioinformaticians and experimentalists for successful biological and medical investigations, the objective of the dissertation was two-fold. On the one hand, vital problems in biology and medicine were selected to explore the value of bioinformatics support for experiment evaluation and hypothesis formation. On the other hand, some of the encountered limitations of bioinformatics approaches were addressed with methodological improvements concerning analyses of protein structures and interactions. Accordingly, the application of computational tools did not only advance the understanding of molecular disease processes, but it also indicated suitable starting points for improving bioinformatics methods.

Part of the research published in over twenty journal articles has involved comprehensive application studies of bioinformatics approaches targeted primarily at autoinflammatory and neurodegenerative disorders like Crohn's, Huntington's and Parkinson's disease. A variety of computational techniques was used to analyze medically relevant proteins and to evaluate experimental data. Examples of investigated proteins are the pathogen receptors NALP3 and NOD2 regulating inflammatory immune responses and the polyglutamine proteins ataxin-2 and ataxin-3 causing inherited neurodegeneration.

Numerous bioinformatics methods were applied to predict structural and functional properties of the proteins and to explore their interaction networks. The methods supported the identification and alignment of homologous sequences and the characterization of the primary protein architecture consisting of domains and binding motifs. This included the discovery of novel sequence motifs with functional relevance and new protein family members. In addition, the secondary and tertiary structures of proteins were predicted and the binding sites for proteins and other ligands were analyzed. Disease-associated sequence variants were localized in three-dimensional structure models to suggest potential functional effects and molecular mechanisms defective in diseases.

The computational results have often provided a rationale for the design, prioritization, and interpretation of experimental studies conducted by biomedical cooperation partners. Some of the generated hypotheses on protein function were also tested and confirmed by experiments. The comparison of predicted structural models with recent, experimentally solved, structures validated the bioinformatics predictions. This underlines the great value of structural and functional predictions in guiding experimentation.

Importantly, the conducted bioinformatics application studies have also led to the identification of methodological limitations. In particular, the recognized problems gave rise to the development of three novel computational approaches supporting the biological and medical investigation of proteins. These new methods generally advance the prediction of the secondary and tertiary structures of proteins and facilitate the exploration of their functions and interaction networks.

A new method for predicting secondary structure was introduced to increase the accuracy of prediction results. Although this method is simple to implement, it is quite successful in improving the performance of secondary structure prediction. It forms a consensus prediction using the results of three different prediction methods and raises the prediction quality and reliability significantly. Further analyses also provided valuable insights into the similarity of the prediction results and the higher confidence in consistently predicted secondary structure.

To utilize experimental measurements of molecular distances in tertiary structure prediction, a new approach was developed that combines protein structure predictions with a small number of additional distance constraints. The latter may be obtained by fast experimental techniques like mass spectrometry or NMR spectroscopy. For the evaluation of the computed alignments, novel scoring functions were applied that incorporated measures of constraint satisfaction to validate structural models. The

employed post-filtering procedure for scoring the distance constraints in sequence-structure alignments results in considerable increases of the recognition rates for domain folds.

Another new method to automate the decomposition of protein networks into domain-domain interactions was designed and implemented as a plugin for Cytoscape. Cytoscape is a software platform for the visualization and analysis of protein interaction networks. The plugin subdivides interacting proteins into their respective domains to compute a putative network of the corresponding domain interactions. Hereby, it facilitates the exploration of protein-protein interactions at a more detailed molecular level. Also, it is a useful tool for 3D structure modeling of protein domain complexes.

In conclusion, the cooperation of bioinformaticians and experimentally working groups that study frequently occurring human diseases has created a valuable cross-fertilization. The collaboration has proven to be quite effective in successfully answering biological and medical questions and in deepening the molecular understanding of disease processes. The joint work has also motivated the development of novel bioinformatics approaches, especially in combining protein structure prediction with experimental and functional information. Future methodological work will focus on the integration of the growing amounts of heterogeneous biological data produced by high-throughput proteomics technologies into cellular disease models.

Zusammenfassung

Proteine erfüllen essentielle Funktionen in lebenden Organismen und spielen eine Schlüsselrolle in komplexen dynamischen Prozessen innerhalb und außerhalb von Zellen. Gerade weil Proteine so wichtig fürs Leben sind, verursachen Defekte ihrer Struktur und Funktion oft ernsthafte Erkrankungen. Es gibt bereits eine Vielzahl von Computermethoden zur Unterstützung molekularer Proteinanalysen, die für experimentell arbeitende Biologen und forschende Mediziner von Interesse sind. Jedoch benötigt die Einschätzung der Performanz von Bioinformatikmethoden und der Qualität ihrer Ergebnisse interdisziplinäre Expertise in Informatik und Statistik ebenso wie in Biologie und Medizin.

Daher ist es nützlich, dass Bioinformatiker nicht nur neuartige und verbesserte Ansätze zur Lösung von biomedizinischen Problemen entwickeln, sondern auch mit Laborbiologen bei der Anwendung rechnerbasierter Methoden kooperieren. Diese Zusammenarbeit ist entscheidend für die akkurate Interpretation von bioinformatischen Ergebnissen und für ihre effektive Einbindung in die biomedizinische Forschung, denn daraus können sich integrative Modelle ergeben, die experimentelle und rechnergestützte Erkenntnisse für die Biologie und Medizin beinhalten. Im Gegenzug gewinnen Bioinformatiker nützliche Einblicke in die biologischen und medizinischen Aspekte und erhalten Feedback für künftige Verbesserungen und Erweiterungen ihrer Anwendungen.

Da somit enge Kooperationen zwischen Bioinformatikern und experimentell arbeitenden Partnern für erfolgreiche biologische und medizinische Studien von großer Bedeutung sind, verfolgte diese Dissertation zwei Ziele: Zum einen wurden wichtige Probleme aus der Biologie und Medizin ausgewählt, um den Wert bioinformatischer Unterstützung bei der Auswertung von Experimenten und der Bildung von Hypothesen zu erkunden. Zum anderen wurden methodische Verbesserungen für einige der aufgedeckten Beschränkungen bioinformatischer Verfahren entwickelt, die Analysen von Proteinstrukturen und -interaktionen betreffen. Demgemäß führte die Anwendung

von Computerwerkzeugen nicht nur zu Fortschritten im Verständnis molekularer Krankheitsprozesse, sondern zeigte auch geeignete Ausgangspunkte für Verbesserungen von Bioinformatikmethoden auf.

Teile der in über zwanzig Zeitschriftenartikeln veröffentlichten Forschung betrafen umfangreiche Applikationsstudien von Bioinformatikmethoden, die sich vornehmlich mit schweren autoinflammatorischen und neurodegenerativen Erkrankungen wie Morbus Crohn, Huntington und Parkinson befassten. Es wurden verschiedene Computermethoden verwendet, um medizinisch relevante Proteine zu analysieren und experimentelle Daten auszuwerten. Beispiele für untersuchte Proteine sind die Pathogenrezeptoren NALP3 und NOD2, welche inflammatorische Immunantworten regulieren, und die Polyglutaminproteine Ataxin-2 und Ataxin-3, die erbliche Formen von Neurodegeneration verursachen.

Zahlreiche Bioinformatikmethoden kamen zur Anwendung, um strukturelle und funktionelle Eigenschaften von Proteinen vorherzusagen und ihre Interaktionsnetzwerke zu erkunden. Die Methoden unterstützten die Identifizierung und das Alignment von homologen Sequenzen und die Charakterisierung der primären Proteinarchitektur, die aus Domänen und Bindungsmotiven besteht. Dies schloss die Entdeckung von neuartigen Sequenzmotiven mit funktioneller Relevanz und neuen Proteinfamilienmitgliedern ein. Zusätzlich wurden die Sekundär- und Tertiärstrukturen von Proteinen vorhergesagt und die Bindungsstellen für Proteine und andere Liganden analysiert. In dreidimensionalen Strukturmodellen wurden krankheitsassoziierte Sequenzvarianten lokalisiert, um potentielle funktionelle Effekte vorherzusagen und molekulare Mechanismen aufzuzeigen, die bei Erkrankungen defekt sind.

Die Ergebnisse aus dem Rechner dienten oft als Grundlage für die Planung, Priorisierung und Interpretation von experimentellen Studien, die von Kooperationspartnern aus der Biomedizin durchgeführt wurden. Auch wurden einige der generierten Hypothesen zur Proteinfunktion durch Experimente überprüft und bestätigt. Der Vergleich vorhergesagter Strukturmodelle mit vor kurzem experimentell bestimmten Strukturen validierte die Bioinformatikvorhersagen. Dies untermauert den hohen Stellenwert der Struktur- und Funktionsvorhersagen für die Durchführung von Experimenten.

Wichtig ist es festzuhalten, dass die durchgeführten bioinformatischen Applikationsstudien auch zur Identifizierung methodischer Grenzen führten. Insbesondere gaben die erkannten Probleme den Anstoß für die Entwicklung von drei neuartigen rechnerbasierten Verfahren zur Unterstützung der biologischen und medizinischen Untersuchung von Proteinen. Diese neuen Methoden verbessern die Vorhersage der Sekundär- und Tertiärstrukturen von Proteinen und erleichtern die Untersuchung ihrer Funktionen und Interaktionsnetzwerke.

Eine neue Methode zur Vorhersage von Sekundärstrukturen wurde eingeführt, um die Genauigkeit der vorhergesagten Ergebnisse zu erhöhen. Obwohl diese Methode einfach zu implementieren ist, ist sie recht erfolgreich und verbessert die Performanz der Sekundärstrukturvorhersage. Sie bildet eine Konsensusvorhersage aus den Ergebnissen dreier verschiedener Vorhersagemethoden und verbessert die Vorhersagequalität und Verlässlichkeit signifikant. Weitere Analysen ergaben zudem wertvolle

Einsichten in die Ähnlichkeit der Vorhersageergebnisse und die höhere Konfidenz in konsistent vorhergesagter Sekundärstruktur.

Um experimentelle Messungen molekularer Distanzen in der Tertiärstrukturvorhersage zu nutzen, wurde ein neues Verfahren entwickelt, das Proteinstrukturvorhersagen mit einer kleinen Anzahl zusätzlicher Distanzbeschränkungen kombiniert. Letztere können durch schnelle experimentelle Techniken wie Massenspektrometrie oder NMR-Spektroskopie gewonnen werden. Für die Auswertung von berechneten Alignments wurden neuartige Bewertungsfunktionen angewendet, welche Maßzahlen über die Einhaltung von Beschränkungen einbinden, um strukturelle Modelle zu validieren. Die angewandte Prozedur enthält einen nachgeschalteten Filter, der die Distanzbeschränkungen in Alignments von Sequenz und Struktur bewertet, woraus eine erhebliche Erhöhung der Erkennungsrate von Domänfaltungen resultiert.

Eine weitere neue Methode wurde als Plugin für Cytoscape entworfen und implementiert, um die Dekomposition von Proteinnetzwerken in Interaktionen von Domänen zu automatisieren. Cytoscape ist eine Softwareplattform für die Visualisierung und Analyse von Proteininteraktionsnetzwerken. Das Plugin unterteilt interagierende Proteine in ihre jeweiligen Domänen, um ein mutmaßliches Netzwerk von entsprechenden Domäneninteraktionen zu berechnen. Hierdurch erleichtert es die Erkundung von Interaktionen zwischen Proteinen auf einer detaillierteren molekularen Ebene. Zudem ist es ein nützliches Werkzeug für die 3D-Modellierung von Proteindomänenkomplexen.

Zusammenfassend kann man festhalten, dass die Kooperation von Bioinformatikern und experimentell arbeitenden Gruppen, die häufig auftretende Erkrankungen untersuchen, für beide Seiten eine wertvolle Bereicherung war. Die Zusammenarbeit erwies sich als sehr effektiv zur erfolgreichen Beantwortung biologischer und medizinischer Fragestellungen und zum Vertiefen des Verständnisses von Krankheitsprozessen auf molekularer Ebene. Die gemeinsame Arbeit motivierte auch die Entwicklung neuartiger Bioinformatikverfahren, besonders bei der Einbindung von experimentellen und funktionellen Informationen zur Vorhersage von Proteinstrukturen. Künftige methodische Arbeiten werden sich auf die Integration der anwachsenden Mengen heterogener biologischer Daten, die von Proteomiktechnologien mit hohem Durchsatz erzeugt werden, in zelluläre Modelle von Erkrankungen konzentrieren.

Bibliography

Abbott, D.W., Wilkins, A., Asara, J.M. and Cantley, L.C. (2004) The Crohn's disease protein, NOD2, requires RIP2 in order to induce ubiquitinylation of a novel site on NEMO. *Curr Biol*, **14**, 2217-2227.

Akira, S. and Takeda, K. (2004) Toll-like receptor signalling. *Nat Rev Immunol*, **4**, 499-511.

Albrecht, M., Hanisch, D., Zimmer, R. and Lengauer, T. (2002) Improving fold recognition of protein threading by experimental distance constraints. *In Silico Biol*, **2**, 325-337.

Albrecht, M. and Lengauer, T. (2003) Pyranose oxidase identified as a member of the GMC oxidoreductase family. *Bioinformatics*, **19**, 1216-1220.

Albrecht, M., Domingues, F.S., Schreiber, S. and Lengauer, T. (2003a) Structural localization of disease-associated sequence variations in the NACHT and LRR domains of PYPAF1 and NOD2. *FEBS Lett*, **554**, 520-528.

Albrecht, M., Domingues, F.S., Schreiber, S. and Lengauer, T. (2003b) Identification of mammalian orthologs associates PYPAF5 with distinct functional roles. *FEBS Lett*, **538**, 173-177.

Albrecht, M., Hoffmann, D., Evert, B.O., Schmitt, I., Wüllner, U. and Lengauer, T. (2003c) Structural modeling of ataxin-3 reveals distant homology to adaptins. *Proteins*, **50**, 355-370.

Albrecht, M., Lengauer, T. and Schreiber, S. (2003d) Disease-associated variants in PYPAF1 and NOD2 result in similar alterations of conserved sequence. *Bioinformatics*, **19**, 2171-2175.

Albrecht, M., Tosatto, S.C., Lengauer, T. and Valle, G. (2003e) Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng*, **16**, 459-462.

Albrecht, M., Golatta, M., Wüllner, U. and Lengauer, T. (2004) Structural and functional analysis of ataxin-2 and ataxin-3. *Eur J Biochem*, **271**, 3155-3170.

Albrecht, M. and Lengauer, T. (2004a) Novel Sm-like proteins with long C-terminal tails and associated methyltransferases. *FEBS Lett*, **569**, 18-26.

Albrecht, M. and Lengauer, T. (2004b) Survey on the PABC recognition motif PAM2. *Biochem Biophys Res Commun*, **316**, 129-138.

Albrecht, M. (2005) LRRK2 mutations and Parkinsonism. *Lancet*, **365**, 1230.

Albrecht, M., Choubey, D. and Lengauer, T. (2005a) The HIN domain of IFI-200 proteins consists of two OB folds. *Biochem Biophys Res Commun*, **327**, 679-687.

Albrecht, M., Huthmacher, C., Tosatto, S.C. and Lengauer, T. (2005b) Decomposing protein networks into domain-domain interactions. *Bioinformatics*, **21 Suppl 2**, ii220-ii221.

Albrecht, M. and Takken, F.L.W. (2006) Update on the domain architectures of NLRs and R proteins. *Biochem Biophys Res Commun*, **339**, 459-462.

Alexandrov, N.N. (1996) SARFing the PDB. *Protein Eng*, **9**, 727-732.

Alexandrov, N.N., Nussinov, R. and Zimmer, R.M. (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac Symp Biocomput*, 53-72.

Aloy, P., Ceulemans, H., Stark, A. and Russell, R.B. (2003a) The relationship between sequence and interaction divergence in proteins. *J Mol Biol*, **332**, 989-998.

Aloy, P., Stark, A., Hadley, C. and Russell, R.B. (2003b) Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins*, **53 Suppl 6**, 436-456.

Aloy, P., Pichaud, M. and Russell, R.B. (2005) Protein complexes: structure prediction challenges for the 21st century. *Curr Opin Struct Biol*, **15**, 15-22.

Aloy, P. and Russell, R.B. (2005) Structure-based systems biology: a zoom lens for the cell. *FEBS Lett*, **579**, 1854-1858.

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

An, Y. and Friesner, R.A. (2002) A novel fold recognition method using composite predicted secondary structures. *Proteins*, **48**, 352-366.

Anantharaman, V. and Aravind, L. (2004) Novel conserved domains in proteins with predicted roles in eukaryotic cell-cycle regulation, decapping and RNA stability. *BMC Genomics*, **5**, 45.

Anderson, M.W. (2004) Amending the amyloid hypothesis. *Scientist*, **18**, 28-29.

Andrade, M.A., Perez-Iratxeta, C. and Ponting, C.P. (2001) Protein repeats: structures, functions, and evolution. *J Struct Biol*, **134**, 117-131.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, **32**, D226-229.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L.S. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*, **32**, D115-119.

Asefa, B., Klarmann, K.D., Copeland, N.G., Gilbert, D.J., Jenkins, N.A. and Keller, J.R. (2004) The interferon-inducible p200 family of proteins: a perspective on their roles in cell cycle regulation and differentiation. *Blood Cells Mol Dis*, **32**, 155-167.

Athman, R. and Philpott, D. (2004) Innate immunity via Toll-like receptors and Nod proteins. *Curr Opin Microbiol*, **7**, 25-32.

Audhya, A., Hyndman, F., McLeod, I.X., Maddox, A.S., Yates, J.R., 3rd, Desai, A. and Oegema, K. (2005) A complex containing the Sm protein CAR-1 and the RNA helicase CGH-1 is required for embryonic cytokinesis in Caenorhabditis elegans. *J Cell Biol*, **171**, 267-279.

Ausubel, F.M. (2005) Are innate immune signaling pathways in plants and animals conserved? *Nat Immunol*, **6**, 973-979.

Back, J.W., de Jong, L., Muijsers, A.O. and de Koster, C.G. (2003) Chemical cross-linking and mass spectrometry for protein structural modeling. *J Mol Biol*, **331**, 303-313.

Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, **31**, 248-250.

Badis, G., Saveanu, C., Fromont-Racine, M. and Jacquier, A. (2004) Targeted mRNA degradation by deadenylation-independent decapping. *Mol Cell*, **15**, 5-15.

Baldi, P., Brunak, S., Frasconi, P., Soda, G. and Pollastri, G. (1999) Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937-946.

Bannwarth, M., Bastian, S., Heckmann-Pohl, D., Giffhorn, F. and Schulz, G.E. (2004) Crystal structure of pyranose 2-oxidase from the white-rot fungus *Peniophora sp*. *Biochemistry*, **43**, 11683-11690.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) The Pfam protein families database. *Nucleic Acids Res*, **32**, D138-141.

Bates, P.A., Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, **Suppl 5**, 39-46.

Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**, 783-795.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.

Beutler, B. (2004) Inferences, questions and possibilities in Toll-like receptor signalling. *Nature*, **430**, 257-263.

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, X.M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iyer, V., Kahari, A., Jekosch, K., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, C., Clamp, M. and Hubbard, T. (2004) Ensembl 2004. *Nucleic Acids Res*, **32**, D468-470.

Boag, P.R., Nakamura, A. and Blackwell, T.K. (2005) A conserved RNA-protein complex component involved in physiological germline apoptosis regulation in C. elegans. *Development*, **132**, 4975-4986.

Bojunga, J., Welsch, C., Antes, I., Albrecht, M., Lengauer, T. and Zeuzem, S. (2005) Structural and functional analysis of a novel mutation of CYP21B in a heterozygote carrier of 21-hydroxylase deficiency. *Hum Genet*, **117**, 558-564.

Boneca, I.G. (2005) The role of peptidoglycan in pathogenesis. *Curr Opin Microbiol*, **8**, 46-53.

Bonifati, V., Oostra, B.A. and Heutink, P. (2004) Unraveling the pathogenesis of Parkinson's disease - the contribution of monogenic forms. *Cell Mol Life Sci*, **61**, 1729-1750.

Bonizzi, G. and Karin, M. (2004) The two NF-kappaB activation pathways and their role in innate and adaptive immunity. *Trends Immunol*, **25**, 280-288.

Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I. and Marcotte, E.M. (2004) Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, **14**, 292-299.

Bork, P. and Serrano, L. (2005) Towards cellular systems in 4D. *Cell*, **121**, 507-509.

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S.K., Teichmann, S.A. and Weiner, J., 3rd (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*, **62**, 435-445.

Bourne, P.E., Addess, K.J., Bluhm, W.F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J.C., Townsend-Merino, W., Weissig, H., Westbrook, J. and Berman, H.M. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res*, **32**, D223-225.

Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P.O., Bergamini, G., Croughton, K., Cruciat, C., Eberhard, D., Gagneur, J., Ghidelli, S., Hopf, C., Huhse, B., Mangano, R., Michon, A.M., Schirle, M., Schlegl, J., Schwab, M., Stein, M.A., Bauer, A., Casari, G., Drewes, G., Gavin, A.C., Jackson, D.B., Joberty, G., Neubauer, G., Rick, J., Kuster, B. and Superti-Furga, G. (2004) A physical and functional map of the human TNF-alpha/NF-kappaB signal transduction pathway. *Nat Cell Biol*, **6**, 97-105.

Bowers, P.M., Strauss, C.E. and Baker, D. (2000) De novo protein structure determination using sparse NMR data. *J Biomol NMR*, **18**, 311-318.

Brannetti, B. and Helmer-Citterich, M. (2003) iSPOT: A web tool to infer the interaction specificity of families of protein modules. *Nucleic Acids Res*, **31**, 3709-3711.

Breitkreutz, B.J., Stark, C. and Tyers, M. (2003a) The GRID: the General Repository for Interaction Datasets. *Genome Biol*, **4**, R23.

Breitkreutz, B.J., Stark, C. and Tyers, M. (2003b) Osprey: a network visualization system. *Genome Biol*, **4**, R22.

Brenner, S.E., Koehl, P. and Levitt, M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, **28**, 254-256.

Brice, A. (2005) How much does dardarin contribute to Parkinson's disease? *Lancet*, **365**, 363-364.

Bujnicki, J.M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001) Structure prediction meta server. *Bioinformatics*, **17**, 750-751.

Cambi, A. and Figdor, C.G. (2003) Dual function of C-type lectin-like receptors in the immune system. *Curr Opin Cell Biol*, **15**, 539-546.

Cambi, A., Koopman, M. and Figdor, C.G. (2005) How C-type lectins detect pathogens. *Cell Microbiol*, **7**, 481-488.

Canutescu, A.A., Shelenkov, A.A. and Dunbrack, R.L., Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, **12**, 2001-2014.

Canutescu, A.A. and Dunbrack, R.L., Jr. (2005) MollDE: a homology modeling framework you can click with. *Bioinformatics*, **21**, 2914-2916.

Cardozo, T.J. and Abagyan, R. (1998) Molecular modeling of the domain shared between CED-4 and its mammalian homologue Apaf-1: a structural relationship to the G-proteins. *J Mol Model*, **4**, 83-93.

Castellano, S., Lobanov, A.V., Chapple, C., Novoselov, S.V., Albrecht, M., Hua, D., Lescure, A., Lengauer, T., Krol, A., Gladyshev, V.N. and Guigo, R. (2005) Diversity and functional plasticity of eukaryotic selenoproteins: Identification and characterization of the SelJ family. *Proc Natl Acad Sci U S A*, **102**, 16188-16193.

Cattaneo, E., Rigamonti, D. and Zuccato, C. (2002) The enigma of Huntington's disease. *Sci Am*, **287**, 92-97.

Chandonia, J.M. and Karplus, M. (1999) New methods for accurate prediction of protein secondary structure. *Proteins*, **35**, 293-306.

Chen, C.P., Kernytsky, A. and Rost, B. (2002) Transmembrane helix predictions revisited. *Protein Sci*, **11**, 2774-2791.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*, **31**, 3497-3500.

Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E., Hong, E.L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D. and Cherry, J.M. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res*, **32**, D311-314.

Ciechanover, A. and Brundin, P. (2003) The ubiquitin proteasome system in neurodegenerative diseases: sometimes the chicken, sometimes the egg. *Neuron*, **40**, 427-446.

Coates, J.C. (2003) Armadillo repeat proteins: beyond the animal kingdom. *Trends Cell Biol*, **13**, 463-471.

Constans, A. (2005) Giving a Nod2 the right target. *Scientist*, **19**, 22.

Costello, C.M., Mah, N., Häsler, R., Rosenstiel, P., Waetzig, G.H., Hahn, A., Lu, T., Gurbuz, Y., Nikolaus, S., Albrecht, M., Hampe, J., Lucius, R., Klöppel, G., Eickhoff, H., Lehrach, H., Lengauer, T. and Schreiber, S. (2005) Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays. *PLoS Med*, **2**, e199.1-17.

Crooks, G.E. and Brenner, S.E. (2004) Protein secondary structure: entropy, correlations and prediction. *Bioinformatics*, **20**, 1603-1611.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892-893.

Cuff, J.A. and Barton, G.J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**, 508-519.

Cuff, J.A. and Barton, G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502-511.

Cusick, M.E., Klitgord, N., Vidal, M. and Hill, D.E. (2005) Interactome: gateway into systems biology. *Hum Mol Genet*, **14 Suppl 2**, R171-181.

Dangl, J.L. and Jones, J.D. (2001) Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826-833.

de Lichtenberg, U., Jensen, L.J., Brunak, S. and Bork, P. (2005) Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724-727.

Dominguez, C., Boelens, R. and Bonvin, A.M. (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, **125**, 1731-1737.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M. and Uversky, V.N. (2005) Flexible nets. *Febs J*, **272**, 5129-5148.

Dunne, A. and O'Neill, L.A. (2005) Adaptor usage and Toll-like receptor signaling specificity. *FEBS Lett*, **579**, 3330-3335.

Eckmann, L. and Karin, M. (2005) NOD2 and Crohn's disease: loss or gain of function? *Immunity*, **22**, 661-667.

Eddy, S.R. (1996) Hidden Markov models. *Curr Opin Struct Biol*, **6**, 361-365.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792-1797.

Eisenberg, D. (1997) Into the black of night. *Nat Struct Biol*, **4**, 95-97.

Espadaler, J., Aragues, R., Eswar, N., Marti-Renom, M.A., Querol, E., Aviles, F.X., Sali, A. and Oliva, B. (2005) Detecting remotely related proteins by their interactions and sequence similarity. *Proc Natl Acad Sci U S A*, **102**, 7151-7156.

Everett, C.M. and Wood, N.W. (2004) Trinucleotide repeats and neurodegenerative disease. *Brain*, **127**, 2385-2405.

Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Fiser, A., Pazos, F., Valencia, A., Sali, A. and Rost, B. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, **17**, 1242-1243.

Finkelstein, A.V. (1997) Protein structure: what is it possible to predict now? *Curr Opin Struct Biol*, **7**, 60-71.

Finn, R.D., Marshall, M. and Bateman, A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410-412.

Flajnik, M.F. and Du Pasquier, L. (2004) Evolution of innate and adaptive immunity: can we draw a line? *Trends Immunol*, **25**, 640-644.

Fontana, P., Bindewald, E., Toppo, S., Velasco, R., Valle, G. and Tosatto, S.C. (2005) The SSEA server for protein secondary structure alignment. *Bioinformatics*, **21**, 393-395.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. and Dubchak, I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res*, **32**, W273-279.

Funke, L., Dakoji, S. and Bredt, D.S. (2005) Membrane-associated guanylate kinases regulate adhesion and plasticity at cell junctions. *Annu Rev Biochem*, **74**, 219-245.

Galtier, N., Gouy, M. and Gautier, C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci*, **12**, 543-548.

Gasser, T. (2005) Genetics of Parkinson's disease. *Curr Opin Neurol*, **18**, 363-369.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D. and Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. In: *The Proteomics Protocols Handbook*, pp. 571-607 (Walker, J.M., Ed.) Humana Press.

Gatchel, J.R. and Zoghbi, H.Y. (2005) Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet*, **6**, 743-755.

Geijtenbeek, T.B., van Vliet, S.J., Engering, A., t Hart, B.A. and van Kooyk, Y. (2004) Self- and nonself-recognition by C-type lectins on dendritic cells. *Annu Rev Immunol*, **22**, 33-54.

Giffhorn, F. (2000) Fungal pyranose oxidases: occurrence, properties and biotechnical applications in carbohydrate chemistry. *Appl Microbiol Biotechnol*, **54**, 727-740.

Giffhorn, F., Kopper, S., Huwig, A. and Freimund, S. (2000) Rare sugars and sugar-based synthons by chemo-enzymatic synthesis. *Enzyme Microb Technol*, **27**, 734-742.

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

Ginalski, K. and Rychlewski, L. (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res*, **31**, 3291-3292.

Ginalski, K., Grishin, N.V., Godzik, A. and Rychlewski, L. (2005) Practical lessons from protein structure prediction. *Nucleic Acids Res*, **33**, 1874-1891.

Girardin, S.E., Sansonetti, P.J. and Philpott, D.J. (2002) Intracellular vs extracellular recognition of pathogens - common concepts in mammals and flies. *Trends Microbiol*, **10**, 193-199.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163-164.

Gloeckner, C.J., Kinkl, N., Schumacher, A., Braun, R.J., O'Neill, E., Meitinger, T., Kolch, W., Prokisch, H. and Ueffing, M. (2006) The Parkinson disease causing LRRK2 mutation I2020T is associated with increased kinase activity. *Hum Mol Genet*, **15**, 223-232.

Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., Droege, A., Lindenberg, K.S., Knoblich, M., Haenig, C., Herbst, M., Suopanki, J., Scherzinger, E., Abraham, C., Bauer, B., Hasenbank, R., Fritzsche, A., Ludewig, A.H., Buessow, K., Coleman, S.H., Gutekunst, C.A., Landwehrmeyer, B.G., Lehrach, H. and Wanker, E.E. (2004) A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell*, **15**, 853-865.

Gouet, P., Robert, X. and Courcelle, E. (2003) ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res*, **31**, 3320-3323.

Green, N.S., Reisler, E. and Houk, K.N. (2001) Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers. *Protein Sci*, **10**, 1293-1304.

Greenamyre, J.T. and Hastings, T.G. (2004) Biomedicine. Parkinson's - divergent causes, convergent mechanisms. *Science*, **304**, 1120-1122.

Greenwald, R.J., Freeman, G.J. and Sharpe, A.H. (2005) The B7 family revisited. *Annu Rev Immunol*, **23**, 515-548.

Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, **84**, 4355-4358.

Guermeur, Y., Geourjon, C., Gallinari, P. and Deleage, G. (1999) Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, **15**, 413-421.

Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714-2723.

Gunsalus, K.C., Ge, H., Schetter, A.J., Goldberg, D.S., Han, J.D., Hao, T., Berriz, G.F., Bertin, N., Huang, J., Chuang, L.S., Li, N., Mani, R., Hyman, A.A., Sonnichsen, B., Echeverri, C.J., Roth, F.P., Vidal, M. and Piano, F. (2005) Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. *Nature*, **436**, 861-865.

Guterman, A. and Glickman, M.H. (2004) Deubiquitinating enzymes are in/(trinsic to proteasome function). *Curr Protein Pept Sci*, **5**, 201-211.

Halada, P., Leitner, C., Sedmera, P., Haltrich, D. and Volc, J. (2003) Identification of the covalent flavin adenine dinucleotide-binding region in pyranose 2-oxidase from *Trametes multicolor*. *Anal Biochem*, **314**, 235-242.

Haliloglu, T., Kolinski, A. and Skolnick, J. (2003) Use of residual dipolar couplings as restraints in ab initio protein structure prediction. *Biopolymers*, **70**, 548-562.

Hallberg, B.M., Leitner, C., Haltrich, D. and Divne, C. (2004) Crystal structure of the 270 kDa homotetrameric lignin-degrading enzyme pyranose 2-oxidase. *J Mol Biol*, **341**, 781-796.

Hanisch, D., Zimmer, R. and Lengauer, T. (2002) ProML - the protein markup language for specification of protein sequences, structures and families. *In Silico Biol*, **2**, 313-324.

Hanson, P.I. and Whiteheart, S.W. (2005) AAA+ proteins: have engine, will work. *Nat Rev Mol Cell Biol*, **6**, 519-529.

Harjes, P. and Wanker, E.E. (2003) The hunt for huntingtin function: interaction partners tell many different stories. *Trends Biochem Sci*, **28**, 425-433.

Heger, A. and Holm, L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, **41**, 224-237.

Helmlinger, D., Hardy, S., Sasorith, S., Klein, F., Robert, F., Weber, C., Miguet, L., Potier, N., Van-Dorsselaer, A., Wurtz, J.M., Mandel, J.L., Tora, L. and Devys, D. (2004) Ataxin-7 is a subunit of GCN5 histone acetyltransferase-containing complexes. *Hum Mol Genet*, **13**, 1257-1265.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. and Apweiler, R. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*, **32**, D452-455.

Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., Parkhill, J., Ivens, A.C., Rajandream, M.A. and Barrell, B. (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res*, **32**, D339-343.

Hobohm, U. and Sander, C. (1994) Enlarged representative set of protein structures. *Protein Sci*, **3**, 522-524.

Hoebe, K., Janssen, E. and Beutler, B. (2004) The interface between innate and adaptive immunity. *Nat Immunol*, **5**, 971-974.

Hoffmann, D., Schnaible, V., Wefing, S., Albrecht, M., Hanisch, D. and Zimmer, R. (2002) A new method for the fast solution of protein-3D-structures, combining experiments and bioinformatics. In: *Coupling of biological and electronic systems: Proceedings of the 2nd Caesarium, Bonn, November 1-3, 2000*, pp. 59-78 (Hoffmann, K.-H., Ed.) Springer-Verlag.

Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol*, **233**, 123-138.

Holm, L. and Park, J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566-567.

Hopkins, P.A. and Sriskandan, S. (2005) Mammalian Toll-like receptors: to immunity and beyond. *Clin Exp Immunol*, **140**, 395-407.

Huang, J.T. and Wang, M.T. (2002) Secondary structural wobble: the limits of protein prediction accuracy. *Biochem Biophys Res Commun*, **294**, 621-625.

Huang, J.Y. and Brutlag, D.L. (2001) The eMOTIF database. *Nucleic Acids Res*, **29**, 202-204.

Huang, Y., Cheung, L., Rowe, D. and Halliday, G. (2004) Genetic contributions to Parkinson's disease. *Brain Res Brain Res Rev*, **46**, 44-70.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res*, **32**, D134-137.

Huynen, M.A., Snel, B., von Mering, C. and Bork, P. (2003) Function prediction and protein networks. *Curr Opin Cell Biol*, **15**, 191-198.

Inohara, N., Chamaillard, M., McDonald, C. and Nunez, G. (2005) NOD-LRR proteins: role in host-microbial interactions and inflammatory disease. *Annu Rev Biochem*, **74**, 355-383.

Jaroszewski, L., Rychlewski, L., Reed, J.C. and Godzik, A. (2000) ATP-activated oligomerization as a mechanism for apoptosis regulation: fold and mechanism prediction for CED-4. *Proteins*, **39**, 197-203.

Johnston, J.A. and Madura, K. (2004) Rings, chains and ladders: ubiquitin goes to work in the neuron. *Prog Neurobiol*, **73**, 227-257.

Jones, D.T. (1997) Progress in protein structure prediction. *Curr Opin Struct Biol*, **7**, 377-387.

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

Jones, P., Vinod, N., Down, T., Hackmann, A., Kahari, A., Kretschmann, E., Quinn, A., Wieser, D., Hermjakob, H. and Apweiler, R. (2005) Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics*, **21**, 3198-3199.

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

Karin, M., Yamamoto, Y. and Wang, Q.M. (2004) The IKK NF-kappaB system: a treasure trove for drug development. *Nat Rev Drug Discov*, **3**, 17-26.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. and Kent, W.J. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*, **31**, 51-54.

Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846-856.

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I. and et al. (1994) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat Genet*, **8**, 221-228.

Khusial, P., Plaag, R. and Zieve, G.W. (2005) LSm proteins form heptameric rings that bind to RNA via repeating motifs. *Trends Biochem Sci*, **30**, 522-528.

Kim, D.E., Chivian, D. and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res*, **32**, W526-531.

King, R.D., Ouali, M., Strong, A.T., Aly, A., Elmaghraby, A., Kantardzic, M. and Page, D. (2000) Is it better to combine predictions? *Protein Eng*, **13**, 15-19.

Koehl, P. and Levitt, M. (1999) A brighter future for protein structure prediction. *Nat Struct Biol*, **6**, 108-111.

Koh, I.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A. and Rost, B. (2003) EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res*, **31**, 3311-3315.

Korr, D., Toschi, L., Donner, P., Pohlenz, H.D., Kreft, B. and Weiss, B. (2006) LRRK1 protein kinase activity is stimulated upon binding of GTP to its Roc domain. *Cell Signal*, **18**, 910-920.

Krobitsch, S. and Lindquist, S. (2000) Aggregation of huntingtin in yeast varies with the length of the polyglutamine expansion and the expression of chaperone proteins. *Proc Natl Acad Sci U S A*, **97**, 1589-1594.

Kshirsagar, M. and Parker, R. (2004) Identification of Edc3p as an enhancer of mRNA decapping in Saccharomyces cerevisiae. *Genetics*, **166**, 729-739.

Kufer, T.A., Fritz, J.H. and Philpott, D.J. (2005) NACHT-LRR proteins (NLRs) in bacterial infection and immunity. *Trends Microbiol*, **13**, 381-388.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**, 105-132.

la Cour, T., Gupta, R., Rapacki, K., Skriver, K., Poulsen, F.M. and Brunak, S. (2003) NESbase version 1.0: a database of nuclear export signals. *Nucleic Acids Res*, **31**, 393-396.

Lackner, P., Koppensteiner, W.A., Sippl, M.J. and Domingues, F.S. (2000) ProSup: a refined tool for protein structure alignment. *Protein Eng*, **13**, 745-752.

Landgraf, C., Panni, S., Montecchi-Palazzi, L., Castagnoli, L., Schneider-Mergener, J., Volkmer-Engert, R. and Cesareni, G. (2004) Protein interaction networks by proteome peptide scanning. *PLoS Biol*, **2**, 94-103.

Lappe, M., Park, J., Niggemann, O. and Holm, L. (2001) Generating protein interaction maps from incomplete data: application to fold assignment. *Bioinformatics*, **17 Suppl 1**, S149-156.

Leipe, D.D., Koonin, E.V. and Aravind, L. (2004) STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. *J Mol Biol*, **343**, 1-28.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. and Bork, P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res*, **32**, D142-144.

Lewis, R. (2003) Huntington disease pathology unfolds. *Scientist*, **17**, 32-33.

Li, Q. and Verma, I.M. (2002) NF-kappaB regulation in the immune system. *Nat Rev Immunol*, **2**, 725-734.

Li, W., Zhang, Y. and Skolnick, J. (2004) Application of sparse NMR restraints to large-scale protein structure prediction. *Biophys J*, **87**, 1241-1248.

Liew, F.Y., Xu, D., Brint, E.K. and O'Neill, L.A. (2005) Negative regulation of toll-like receptor-mediated immune responses. *Nat Rev Immunol*, **5**, 446-458.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003a) Protein disorder prediction: implications for structural proteomics. *Structure (Camb)*, **11**, 1453-1459.

Linding, R., Russell, R.B., Neduva, V. and Gibson, T.J. (2003b) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*, **31**, 3701-3708.

Liu, J. and Rost, B. (2003) NORSp: Predictions of long regions without regular secondary structure. *Nucleic Acids Res*, **31**, 3833-3835.

Liu, Y., Liu, N. and Zhao, H. (2005a) Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, **21**, 3279-3285.

Liu, Y.C., Penninger, J. and Karin, M. (2005b) Immunity by ubiquitylation: a reversible process of modification. *Nat Rev Immunol*, **5**, 941-952.

Lozano, A.M. and Kalia, S.K. (2005) New movement in Parkinson's. *Sci Am*, **293**, 68-75.

Lupas, A., Van Dyke, M. and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162-1164.

Macdonald, T.T. and Monteleone, G. (2005) Immunity, inflammation, and allergy in the gut. *Science*, **307**, 1920-1925.

Manto, M.U. (2005) The wide spectrum of spinocerebellar ataxias (SCAs). *Cerebellum*, **4**, 2-6.

Mao, Y., Senic-Matuglia, F., Di Fiore, P.P., Polo, S., Hodsdon, M.E. and De Camilli, P. (2005) Deubiquitinating function of ataxin-3: insights from the solution structure of the Josephin domain. *Proc Natl Acad Sci U S A*, **102**, 12700-12705.

Marchler-Bauer, A. and Bryant, S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res*, **32**, W327-331.

Margolis, R.L. (2002) The spinocerebellar ataxias: order emerges from chaos. *Curr Neurol Neurosci Rep*, **2**, 447-456.

Martinon, F. and Tschopp, J. (2004) Inflammatory caspases: linking an intracellular innate immune system to autoinflammatory diseases. *Cell*, **117**, 561-574.

Martinon, F. and Tschopp, J. (2005) NLRs join TLRs as innate sensors of pathogens. *Trends Immunol*, **26**, 447-454.

McDonald, C., Inohara, N. and Nunez, G. (2005) Peptidoglycan signaling in innate immunity and inflammatory disease. *J Biol Chem*, **280**, 20177-20180.

McGreal, E.P., Martinez-Pomares, L. and Gordon, S. (2004) Divergent roles for C-type lectins expressed by cells of the innate immune system. *Mol Immunol*, **41**, 1109-1121.

McGreal, E.P., Miller, J.L. and Gordon, S. (2005) Ligand recognition by antigen-presenting cell C-type lectin receptors. *Curr Opin Immunol*, **17**, 18-24.

McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404-405.

McGuffin, L.J. and Jones, D.T. (2003) Benchmarking secondary structure prediction for fold recognition. *Proteins*, **52**, 166-175.

Meiler, J. and Baker, D. (2003) Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci U S A*, **100**, 15404-15409.

Meylan, E. and Tschopp, J. (2005) The RIP kinases: crucial integrators of cellular stress. *Trends Biochem Sci*, **30**, 151-159.

Moll, A., Hildebrandt, A., Lenhof, H.P. and Kohlbacher, O. (2006) BALLView: a tool for research and education in molecular modeling. *Bioinformatics*, **22**, 365-366.

Moore, D.J., West, A.B., Dawson, V.L. and Dawson, T.M. (2005) Molecular pathophysiology of Parkinson's disease. *Annu Rev Neurosci*, **28**, 57-87.

Morris, H.R. (2005) Genetics of Parkinson's disease. *Ann Med*, **37**, 86-96.

Moynagh, P.N. (2005) TLR signalling and activation of IRFs: revisiting old friends from the NF-kappaB pathway. *Trends Immunol*, **26**, 469-476.

Mueller, T. and Podolsky, D.K. (2005) Nucleotide-binding-oligomerization domain proteins and toll-like receptors: sensors of the inflammatory bowel diseases' microbial environment. *Curr Opin Gastroenterol*, **21**, 419-425.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J., Vaughan, R. and Zdobnov, E.M. (2003) The InterPro database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, **31**, 315-318.

Murray, P.J. (2005) NOD proteins: an intracellular pathogen-recognition system or signal transduction modifiers? *Curr Opin Immunol*, **17**, 352-358.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**, 536-540.

Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, **24**, 34-36.

Neduva, V. and Russell, R.B. (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett*, **579**, 3342-3345.

Neven, B., Callebaut, I., Prieur, A.M., Feldmann, J., Bodemer, C., Lepore, L., Derfalvi, B., Benjaponpitak, S., Vesely, R., Sauvain, M.J., Oertle, S., Allen, R., Morgan, G., Borkhardt, A., Hill, C., Gardner-Medwin, J., Fischer, A. and de Saint Basile, G. (2004) Molecular basis of the spectral expression of CIAS1 mutations associated with phagocytic cell-mediated autoinflammatory disorders CINCA/NOMID, MWS, and FCU. *Blood*, **103**, 2809-2815.

Newman, B. and Siminovitch, K.A. (2005) Recent advances in the genetics of inflammatory bowel disease. *Curr Opin Gastroenterol*, **21**, 401-407.

Ng, S.K., Zhang, Z., Tan, S.H. and Lin, K. (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res*, **31**, 251-254.

Nicastro, G., Menon, R.P., Masino, L., Knowles, P.P., McDonald, N.Q. and Pastore, A. (2005) The solution structure of the Josephin domain of ataxin-3: Structural determinants for molecular recognition. *Proc Natl Acad Sci U S A*, **102**, 10493-10498.

Nicholas, K., Nicholas, H. and Deerfield, D. (1997) GeneDoc: Analysis and visualization of genetic variation. *EMBNEW.NEWS*, **4**, 14.

Nijman, S.M., Luna-Vargas, M.P., Velds, A., Brummelkamp, T.R., Dirac, A.M., Sixma, T.K. and Bernards, R. (2005) A genomic and functional inventory of deubiquitinating enzymes. *Cell*, **123**, 773-786.

Nolen, B., Taylor, S. and Ghosh, G. (2004) Regulation of protein kinases: controlling activity through activation segment conformation. *Mol Cell*, **15**, 661-675.

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

O'Neill, L.A. (2004) TLRs: Professor Mechnikov, sit on your hat. *Trends Immunol*, **25**, 687-693.

O'Neill, L.A. (2005) Immunity's early-warning system. *Sci Am*, **292**, 24-31.

Orengo, C.A. and Thornton, J.M. (2005) Protein families and their evolution - a structural perspective. *Annu Rev Biochem*, **74**, 867-900.

Ouali, M. and King, R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci*, **9**, 1162-1176.

Outeiro, T.F. and Lindquist, S. (2003) Yeast cells provide insight into alpha-synuclein biology and pathobiology. *Science*, **302**, 1772-1775.

Paduch, M., Jelen, F. and Otlewski, J. (2001) Structure of small G proteins and their regulators. *Acta Biochim Pol*, **48**, 829-850.

Park, J., Lappe, M. and Teichmann, S.A. (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol*, **307**, 929-938.

Pawson, T. and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445-452.

Pearson, C.E., Edamura, K.N. and Cleary, J.D. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet*, **6**, 729-742.

Pearson, W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol*, **132**, 185-219.

Petersen, T.N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G.P. and Lund, O. (2000) Prediction of protein secondary structure at 80% accuracy. *Proteins*, **41**, 17-20.

Petrey, D. and Honig, B. (2003) GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol*, **374**, 492-509.

Poirot, O., O'Toole, E. and Notredame, C. (2003) Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res*, **31**, 3503-3506.

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

Prlic, A., Down, T.A. and Hubbard, T.J. (2005) Adding some SPICE to DAS. *Bioinformatics*, **21 Suppl 2**, ii40-ii41.

Przybylski, D. and Rost, B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197-205.

Pulst, S.M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X.N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A., DeJong, P., Rouleau, G.A., Auburger, G., Korenberg, J.R., Figueroa, C. and Sahba, S. (1996) Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat Genet*, **14**, 269-276.

Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M., Ausiello, G., Brannetti, B., Costantini, A., Ferre, F., Maselli, V., Via, A., Cesareni, G., Diella, F., Superti-Furga, G., Wyrwicz, L., Ramu, C., McGuigan, C., Gudavalli, R., Letunic, I., Bork, P., Rychlewski, L., Kuster, B., Helmer-Citterich, M., Hunter, W.N., Aasland, R. and Gibson, T.J. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*, **31**, 3625-3630.

Qian, J., Stenger, B., Wilson, C.A., Lin, J., Jansen, R., Teichmann, S.A., Park, J., Krebs, W.G., Yu, H., Alexandrov, V., Echols, N. and Gerstein, M. (2001) PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res*, **29**, 1750-1764.

Qu, Y., Guo, J.T., Olman, V. and Xu, Y. (2004) Protein structure prediction using sparse dipolar coupling data. *Nucleic Acids Res*, **32**, 551-561.

Ralser, M., Albrecht, M., Nonhoff, U., Lengauer, T., Lehrach, H. and Krobitsch, S. (2005a) An integrative approach to gain insights into the cellular function of human ataxin-2. *J Mol Biol*, **346**, 203-214.

Ralser, M., Nonhoff, U., Albrecht, M., Lengauer, T., Wanker, E.E., Lehrach, H. and Krobitsch, S. (2005b) Ataxin-2 and huntingtin interact with endophilin-A complexes to function in plastin-associated pathways. *Hum Mol Genet*, **14**, 2893-2909.

Riedl, S.J. and Shi, Y. (2004) Molecular mechanisms of caspase regulation during apoptosis. *Nat Rev Mol Cell Biol*, **5**, 897-907.

Riedl, S.J., Li, W., Chao, Y., Schwarzenbacher, R. and Shi, Y. (2005) Structure of the apoptotic protease-activating factor 1 bound to ADP. *Nature*, **434**, 926-933.

Rigoutsos, I. and Floratos, A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, **14**, 55-67.

Riley, R., Lee, C., Sabatti, C. and Eisenberg, D. (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol*, **6**, R89.

Roberts, J.P. (2003) An immunological role in the CARDs. *Scientist*, **17**, 29-30.

Rodriguez, R., Chinea, G., Lopez, N., Pons, T. and Vriend, G. (1998) Homology modeling, model and software evaluation: three related resources. *Bioinformatics*, **14**, 523-528.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. and Dunker, A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38-48.

Ross, C.A. and Pickart, C.M. (2004) The ubiquitin-proteasome pathway in Parkinson's disease and other neurodegenerative diseases. *Trends Cell Biol*, **14**, 703-711.

Ross, C.A. and Poirier, M.A. (2004) Protein aggregation and neurodegenerative disease. *Nat Med*, **10 Suppl**, S10-17.

Rost, B., Sander, C. and Schneider, R. (1994) Redefining the goals of protein secondary structure prediction. *J Mol Biol*, **235**, 13-26.

Rost, B., Schneider, R. and Sander, C. (1997) Protein fold recognition by prediction-based threading. *J Mol Biol*, **270**, 471-480.

Rost, B. (2001) Review: protein secondary structure prediction continues to rise. *J Struct Biol*, **134**, 204-218.

Rost, B. and Eyrich, V.A. (2001) EVA: large-scale analysis of secondary structure prediction. *Proteins*, **Suppl 5**, 192-199.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P. and Vidal, M. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173-1178.

Ruiz-Opazo, N., Akimoto, K. and Herrera, V.L. (1995) Identification of a novel dual angiotensin II/vasopressin receptor on the basis of molecular recognition theory. *Nat Med*, **1**, 1074-1081.

Ryan, D.P. and Matthews, J.M. (2005) Protein-protein interactions in human disease. *Curr Opin Struct Biol*, **15**, 441-446.

Rybicki, B.A., Walewski, J.L., Maliarik, M.J., Kian, H. and Iannuzzi, M.C. (2005) The BTNL2 gene and sarcoidosis susceptibility in African Americans and Whites. *Am J Hum Genet*, **77**, 491-499.

Salwinski, L. and Eisenberg, D. (2003) Computational methods of analysis of protein-protein interactions. *Curr Opin Struct Biol*, **13**, 377-382.

Santonico, E., Castagnoli, L. and Cesareni, G. (2005) Methods to reveal domain networks. *Drug Discov Today*, **10**, 1111-1117.

Sarrazin, C., Mihm, U., Herrmann, E., Welsch, C., Albrecht, M., Sarrazin, U., Traver, S., Lengauer, T. and Zeuzem, S. (2005) Clinical significance of in vitro replication-enhancing mutations of the hepatitis C virus (HCV) replicon in patients with chronic HCV infection. *J Infect Dis*, **192**, 1710-1719.

Schols, L., Bauer, P., Schmidt, T., Schulte, T. and Riess, O. (2004) Autosomal dominant cerebellar ataxias: clinical features, genetics, and pathogenesis. *Lancet Neurol*, **3**, 291-304.

Schonbrun, J., Wedemeyer, W.J. and Baker, D. (2002) Protein structure prediction in 2002. *Curr Opin Struct Biol*, **12**, 348-354.

Schreiber, S., Rosenstiel, P., Albrecht, M., Hampe, J. and Krawczak, M. (2005) Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat Rev Genet*, **6**, 376-388.

Selbig, J., Mevissen, T. and Lengauer, T. (1999) Decision tree-based formation of consensus protein secondary structure prediction. *Bioinformatics*, **15**, 1039-1046.

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

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498-2504.

Shindyalov, I.N. and Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, **11**, 739-747.

Simossis, V.A. and Heringa, J. (2004) Integrating protein secondary structure prediction and multiple sequence alignment. *Curr Protein Pept Sci*, **5**, 249-266.

Singleton, A.B. (2005) Altered alpha-synuclein homeostasis causing Parkinson's disease: the potential roles of dardarin. *Trends Neurosci*, **28**, 416-421.

Skolnick, J., Kolinski, A. and Ortiz, A.R. (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J Mol Biol*, **265**, 217-241.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E. and Edelman, M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327-332.

Soboleva, T.A. and Baker, R.T. (2004) Deubiquitinating enzymes: their functions and substrate specificity. *Curr Protein Pept Sci*, **5**, 191-200.

Sommer, I., Zien, A., von Öhsen, N., Zimmer, R. and Lengauer, T. (2002) Confidence measures for protein fold recognition. *Bioinformatics*, **18**, 802-812.

Soto, C. (2003) Unfolding the role of protein misfolding in neurodegenerative diseases. *Nat Rev Neurosci*, **4**, 49-60.

Squirrell, J.M., Eggers, Z.T., Luedke, N., Saari, B., Grimson, A., Lyons, G.E., Anderson, P. and White, J.G. (2005) CAR-1, a protein that localizes with the mRNA decapping component DCAP-1, is required for cytokinesis and ER organization in Caenorhabditis elegans embryos. *Mol Biol Cell*, **17**, 336-344.

Standley, D.M., Eyrich, V.A., Felts, A.K., Friesner, R.A. and McDermott, A.E. (1999) A branch and bound algorithm for protein structure refinement from sparse NMR data sets. *J Mol Biol*, **285**, 1691-1710.

Stein, A., Russell, R.B. and Aloy, P. (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, **33 Database Issue**, D413-417.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. and Wanker, E.E. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957-968.

Sternberg, M.J., Bates, P.A., Kelley, L.A. and MacCallum, R.M. (1999) Progress in protein structure prediction: assessment of CASP3. *Curr Opin Struct Biol*, **9**, 368-373.

Stillwell, R.J. and Berry, M.J. (2005) Expanding the repertoire of the eukaryotic selenoproteome. *Proc Natl Acad Sci U S A*, **102**, 16123-16124.

Stoll, M., Corneliussen, B., Costello, C.M., Waetzig, G.H., Mellgard, B., Koch, W.A., Rosenstiel, P., Albrecht, M., Croucher, P.J., Seegert, D., Nikolaus, S., Hampe, J., Lengauer, T., Pierrou, S., Foelsch, U.R., Mathew, C.G., Lagerstrom-Fermer, M. and Schreiber, S. (2004) Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nat Genet*, **36**, 476-480.

Strober, W., Murray, P.J., Kitani, A. and Watanabe, T. (2006) Signalling pathways and molecular interactions of NOD1 and NOD2. *Nat Rev Immunol*, **6**, 9-20.

Takeda, K. and Akira, S. (2005) Toll-like receptors in innate immunity. *Int Immunol*, **17**, 1-14.

Taroni, F. and DiDonato, S. (2004) Pathways to motor incoordination: the inherited ataxias. *Nat Rev Neurosci*, **5**, 641-655.

Thiele, R., Zimmer, R. and Lengauer, T. (1999) Protein threading by recursive dynamic programming. *J Mol Biol*, **290**, 757-779.

Ting, J.P. and Davis, B.K. (2005) CATERPILLER: a novel gene family important in immunity, cell death, and diseases. *Annu Rev Immunol*, **23**, 387-414.

Tramontano, A. (2003) Of men and machines. *Nat Struct Biol*, **10**, 87-90.

Uetz, P. and Finley Jr., R.L. (2005) From protein networks to biological systems. *FEBS Lett*, **579**, 1821-1827.

Ulevitch, R.J. (2004) Therapeutics targeting the innate immune system. *Nat Rev Immunol*, **4**, 512-520.

Valentonyte, R., Hampe, J., Huse, K., Rosenstiel, P., Albrecht, M., Stenzel, A., Nagy, M., Gaede, K.I., Franke, A., Haesler, R., Koch, A., Lengauer, T., Seegert, D., Reiling, N., Ehlers, S., Schwinger, E., Platzer, M., Krawczak, M., Müller-Quernheim, J., Schürmann, M. and Schreiber, S. (2005) Sarcoidosis is associated with a truncating splice site mutation in BTNL2. *Nat Genet*, **37**, 357-364.

Van Duist, M.M., Albrecht, M., Podswiadek, M., Giachino, D., Lengauer, T., Punzi, L. and De Marchi, M. (2005) A new CARD15 mutation in Blau syndrome. *Eur J Hum Genet*, **13**, 742-747.

van Kooyk, Y. and Geijtenbeek, T.B. (2003) DC-SIGN: escape mechanism for pathogens. *Nat Rev Immunol*, **3**, 697-709.

van Kooyk, Y., Engering, A., Lekkerkerker, A.N., Ludwig, I.S. and Geijtenbeek, T.B. (2004) Pathogens use carbohydrates to escape immunity induced by dendritic cells. *Curr Opin Immunol*, **16**, 488-493.

Vaux, D.L. and Silke, J. (2005) IAPs, RINGs and ubiquitylation. *Nat Rev Mol Cell Biol*, **6**, 287-297.

Vila, M. and Przedborski, S. (2003) Targeting programmed cell death in neurodegenerative diseases. *Nat Rev Neurosci*, **4**, 365-375.

Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C. and Teichmann, S.A. (2004) Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol*, **14**, 208-216.

von Bohlen und Halbach, O., Schober, A. and Krieglstein, K. (2004) Genes, proteins, and neurotoxins involved in Parkinson's disease. *Prog Neurobiol*, **73**, 151-177.

von Öhsen, N., Sommer, I., Zimmer, R. and Lengauer, T. (2004) Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics*, **20**, 2228-2235.

Wan, P.T., Garnett, M.J., Roe, S.M., Lee, S., Niculescu-Duvaz, D., Good, V.M., Jones, C.M., Marshall, C.J., Springer, C.J., Barford, D. and Marais, R. (2004) Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*, **116**, 855-867.

Wang, X., Wang, H., Figueroa, B.E., Zhang, W.H., Huo, C., Guan, Y., Zhang, Y., Bruey, J.M., Reed, J.C. and Friedlander, R.M. (2005) Dysregulation of receptor interacting protein-2 and caspase recruitment domain only protein mediates aberrant caspase-1 activation in Huntington's disease. *J Neurosci*, **25**, 11645-11654.

Ward, J.J., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2003) Secondary structure prediction with support vector machines. *Bioinformatics*, **19**, 1650-1655.

Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. and Jones, D.T. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20**, 2138-2139.

West, A.B., Moore, D.J., Biskup, S., Bugayenko, A., Smith, W.W., Ross, C.A., Dawson, V.L. and Dawson, T.M. (2005) Parkinson's disease-associated mutations in leucine-rich repeat kinase 2 augment kinase activity. *Proc Natl Acad Sci U S A*, **102**, 16842-16847.

Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Suzek, T.O., Tatusova, T.A. and Wagner, L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*, **32**, D35-40.

Wilusz, C.J. and Wilusz, J. (2005) Eukaryotic Lsm proteins: lessons from bacteria. *Nat Struct Mol Biol*, **12**, 1031-1036.

Wolf, E., Kim, P.S. and Berger, B. (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci*, **6**, 1179-1189.

Wuchty, S. (2002) Interaction and domain networks of yeast. *Proteomics*, **2**, 1715-1723.

Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D., Zhao, H. and Gerstein, M. (2004) Analyzing cellular biochemistry in terms of molecular networks. *Annu Rev Biochem*, **73**, 1051-1087.

Xu, Y. and Xu, D. (2000) Protein threading using PROSPECT: design and evaluation. *Proteins*, **40**, 343-354.

Xu, Y., Xu, D., Crawford, O.H. and Einstein, J.R. (2000) A computational method for NMR-constrained protein threading. *J Comput Biol*, **7**, 449-467.

Yan, N., Chai, J., Lee, E.S., Gu, L., Liu, Q., He, J., Wu, J.W., Kokel, D., Li, H., Hao, Q., Xue, D. and Shi, Y. (2005) Structure of the CED-4-CED-9 complex provides insights into programmed cell death in *Caenorhabditis elegans*. *Nature*, **437**, 831-837.

Yan, N. and Shi, Y. (2005) Mechanisms of apoptosis through structural biology. *Annu Rev Cell Dev Biol*, **21**, 35-56.

Yang, X.J. (2005) Multisite protein modification and intramolecular signaling. *Oncogene*, **24**, 1653-1662.

Ye, X., O'Neil, P.K., Foster, A.N., Gajda, M.J., Kosinski, J., Kurowski, M.A., Bujnicki, J.M., Friedman, A.M. and Bailey-Kellogg, C. (2004) Probabilistic cross-link analysis and experiment planning for high-throughput elucidation of protein structure. *Protein Sci*, **13**, 3298-3313.

Ye, Y. and Godzik, A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19 Suppl 2**, ii246-ii255.

Ye, Y. and Godzik, A. (2004) Comparative analysis of protein domain organization. *Genome Res*, **14**, 343-353.

Young, M.M., Tang, N., Hempel, J.C., Oshiro, C.M., Taylor, E.W., Kuntz, I.D., Gibson, B.W. and Dollinger, G. (2000) High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc Natl Acad Sci U S A*, **97**, 5802-5806.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*, **14**, 1107-1118.

Yu, X., Acehan, D., Menetret, J.F., Booth, C.R., Ludtke, S.J., Riedl, S.J., Shi, Y., Wang, X. and Akey, C.W. (2005) A structure of the human apoptosome at 12.8 Å resolution provides insights into this cell death platform. *Structure (Camb)*, **13**, 1725-1735.

Yuan, Q. and Walker, W.A. (2004) Innate immunity of the gut: mucosal defense in health and disease. *J Pediatr Gastroenterol Nutr*, **38**, 463-473.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. and Cesareni, G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*, **513**, 135-140.

Zarrinpar, A., Bhattacharyya, R.P. and Lim, W.A. (2003) The structure and function of proline recognition domains. *Sci STKE*, **2003**, RE8.

Zemla, A., Venclovas, C., Fidelis, K. and Rost, B. (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220-223.

Zhang, B., Jaroszewski, L., Rychlewski, L. and Godzik, A. (1997) Similarities and differences between nonhomologous proteins with similar folds: evaluation of threading strategies. *Fold Des*, **2**, 307-317.

Zien, A., Zimmer, R. and Lengauer, T. (2000) A simple iterative approach to parameter optimization. *J Comput Biol*, **7**, 483-501.

Zoghbi, H.Y. and Orr, H.T. (2000) Glutamine repeats and neurodegeneration. *Annu Rev Neurosci*, **23**, 217-247.

Appendix

The subsequent pages present the abstracts of 25 articles published in the course of this dissertation (Albrecht *et al.*, 2002; Hoffmann *et al.*, 2002; Albrecht and Lengauer, 2003; Albrecht *et al.*, 2003a, 2003b, 2003c, 2003d, 2003e; Albrecht *et al.*, 2004; Albrecht and Lengauer, 2004a, 2004b; Stoll *et al.*, 2004; Albrecht, 2005; Albrecht *et al.*, 2005a, 2005b; Bojunga *et al.*, 2005; Castellano *et al.*, 2005; Costello *et al.*, 2005; Ralser *et al.*, 2005a, 2005b; Sarrazin *et al.*, 2005; Schreiber *et al.*, 2005; Valentonyte *et al.*, 2005; Van Duist *et al.*, 2005; Albrecht and Takken, 2006). The contents of the publications may be summarized briefly in numbers as follows:

The coauthored papers consist of about 200 pages plus 90 online supplemental pages. They contain 25 printed tables (plus ~30 supplementary online tables) and ~90 figures (plus ~30 online figures) consisting of ~130 subfigures (plus ~40 online subfigures). The own contributions comprise 12 tables plus 14 online tables and 57 figures (73 subfigures) plus 24 online figures (29 online subfigures). Overall, the published articles cite more than 1300 references, however, some of which may be counted several times because their citations occur in more than one article.

1. Publication (Albrecht *et al.*, 2002)

Albrecht, M., Hanisch, D., Zimmer, R. and Lengauer, T. (2002) Improving fold recognition of protein threading by experimental distance constraints. *In Silico Biol*, **2**, 325-337.

Abstract

We present a comprehensive analysis of methods for improving the fold recognition rate of the threading approach to protein structure prediction by the utilization of few additional distance constraints. The distance constraints between protein residues may be obtained by experiments such as mass spectrometry or NMR spectroscopy. We applied a post-filtering step with new scoring functions incorporating measures of constraint satisfaction to ranking lists of 123D threading alignments. The detailed analysis of the results on a small representative benchmark set show that the fold recognition rate can be improved significantly by up to 30% from about 54%-65% to 77%-84%, approaching the maximal attainable performance of 90% estimated by structural superposition alignments. This gain in performance adds about 10% to the recognition rate already achieved in our previous study with cross-link constraints only. Additional recent results on a larger benchmark set involving a confidence function for threading predictions also indicate notable improvements by our combined approach, which should be particularly valuable for rapid structure determination and validation of protein models.

2. Publication (Hoffmann *et al.*, 2002)

Hoffmann, D., Schnaible, V., Wefing, S., Albrecht, M., Hanisch, D. and Zimmer, R. (2002) A new method for the fast solution of protein-3D-structures, combining experiments and bioinformatics. In: *Coupling of biological and electronic systems: Proceedings of the 2nd Caesarium, Bonn, November 1-3, 2000*, pp. 59-78 (Hoffmann, K.-H., Ed.) Springer-Verlag.

Abstract

Proteins can be considered molecular machines, and protein 3D-structures are key to the understanding of these machines and to many applications in biotechnology and medicine. We are developing a method to speed up the time consuming process of structure determination significantly. The method closely couples bioinformatics for protein structure prediction with fast experiments (chemical cross-linking, specific proteolysis, mass spectrometry) for structure validation. For a given protein, the method iterates over cycles of bioinformatics and experiments to collect more and more information on the protein structure, finally resulting in an Experimentally Validated Model (EVAM) of the structure.

3. Publication (Albrecht and Lengauer, 2003)

Albrecht, M. and Lengauer, T. (2003) Pyranose oxidase identified as a member of the GMC oxidoreductase family. *Bioinformatics*, **19**, 1216-1220.

Abstract

Fungal pyranose oxidase is a flavoenzyme whose preferred substrate among several monosaccharides is D-glucose. After a comprehensive analysis of conserved features in a structure-based multiple sequence alignment of homologous proteins, we could classify this enzyme into the GMC oxidoreductase family. The identified homology also suggests a three-dimensional protein structure similar to the functionally related glucose oxidase.

4. Publication (Albrecht *et al.*, 2003a)

Albrecht, M., Domingues, F.S., Schreiber, S. and Lengauer, T. (2003) Structural localization of disease-associated sequence variations in the NACHT and LRR domains of PYPAF1 and NOD2. *FEBS Lett*, **554**, 520-528.

Abstract

Several autoinflammatory diseases with distinct clinical manifestations have been associated with sequence variations in the gene products PYPAF1/CIAS1 and NOD2/CARD15. Both proteins belong to the PYD/CARD-containing family of apoptosis regulators and activators of pro-inflammatory caspases. To gain insight into the dysfunctional role of sequence alterations, we assembled a structure-based multiple sequence alignment of family members and related proteins. This allowed us to analyze the putative effect of the alterations on the function of nucleotide-binding (NACHT) and leucine-rich repeat (LRR) domains shared by the family members. In support of this analysis, we carefully selected template structures for the NACHT and LRR domains and mapped the genetic variations onto 3D domain models. Additionally, we propose a model of the NACHT and LRR domain complex. Our study revealed that many of the disease-associated sequence variants are located close to highly conserved sequence regions of functional relevance and are spatially adjacent in the predicted 3D structure. The implications on the domain functions such as NTP-hydrolysis or oligomerization are discussed.

5. Publication (Albrecht *et al.*, 2003b)

Albrecht, M., Domingues, F.S., Schreiber, S. and Lengauer, T. (2003) Identification of mammalian orthologs associates PYPAF5 with distinct functional roles. *FEBS Lett*, **538**, 173-177.

Abstract

PYRIN- and CARD-containing proteins belong to a recently identified protein family involved in the regulation of apoptosis and inflammatory processes. Variations in the gene products of the family members PYPAF1 and NOD2/CARD15 have been associated with several autoinflammatory diseases. We could identify the mouse orthologs of PYPAF1, PYPAF5, NOD1, NOD2 and the rat ortholog of PYPAF5. Intriguingly, we found that PYPAF5 has been reported previously not only as regulator of NF-κB and caspase-1, but also as angiotensin II and vasopressin receptor. In particular, based on a comprehensive sequence analysis, we propose a structural model for this hormone receptor that is different from the model suggested previously.

6. Publication (Albrecht *et al.*, 2003c)

Albrecht, M., Hoffmann, D., Evert, B.O., Schmitt, I., Wüllner, U. and Lengauer, T. (2003) Structural modeling of ataxin-3 reveals distant homology to adaptins. *Proteins*, **50**, 355-370.

Abstract

Spinocerebellar ataxia type 3 (SCA3) is a polyglutamine disorder caused by a CAG repeat expansion in the coding region of a gene encoding ataxin-3, a protein of yet unknown function. Based on a comprehensive computational analysis, we propose a structural model and structure-based functions for ataxin-3. Our predictive strategy comprises the compilation of multiple sequence and structure alignments of carefully selected proteins related to ataxin-3. These alignments are consistent with additional information on sequence motifs, secondary structure, and domain architectures. The application of complementary methods revealed the homology of ataxin-3 to ENTH and VHS domain proteins involved in membrane trafficking and regulatory adaptor functions. We modeled the structure of ataxin-3 using the adaptin AP180 as a template and assessed the reliability of the model by comparison with known sequence and structural features. We could further infer potential functions of ataxin-3 in agreement with known experimental data. Our database searches also identified an as yet uncharacterized family of proteins, which we named josephins because of their pronounced homology to the Josephin domain of ataxin-3.

7. Publication (Albrecht *et al.*, 2003d)

Albrecht, M., Lengauer, T. and Schreiber, S. (2003) Disease-associated variants in PYPAF1 and NOD2 result in similar alterations of conserved sequence. *Bioinformatics*, **19**, 2171-2175.

Abstract

Sequence variations in the gene products PYPAF1/CIAS1 and NOD2/CARD15 have been associated with several autoinflammatory diseases that, although clinically different, share a similar inflammatory pathophysiology. A multiple sequence alignment of homologous proteins demonstrates that some of the missense variants are located in highly conserved regions of the NTPase domain and possibly impair NTP-hydrolysis. Intriguingly, one of the variations, which is found identically in PYPAF1 and NOD2, is located at the same alignment position. Our findings suggest that evolutionary gene duplication can give rise to disease families because variants affect conserved sequence in a similar fashion.

8. Publication (Albrecht *et al.*, 2003e)

Albrecht, M., Tosatto, S.C., Lengauer, T. and Valle, G. (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng*, **16**, 459-462.

Abstract

We have analyzed the performance of majority voting on minimal combination sets of three state-of-the-art secondary structure prediction methods in order to obtain a consensus prediction. Using three large benchmark sets from the EVA server, our results show a significant improvement in the average Q_3 prediction accuracy of up to 1.5 percentage points by consensus formation. The application of an additional trivial filtering procedure for predicted secondary structure elements that are too short, does not significantly affect the prediction accuracy. Our analysis also provides valuable insight into the similarity of the results of the prediction methods that we combine as well as the higher confidence in consistently predicted secondary structure.

9. Publication (Albrecht *et al.*, 2004)

Albrecht, M., Golatta, M., Wüllner, U. and Lengauer, T. (2004) Structural and functional analysis of ataxin-2 and ataxin-3. *Eur J Biochem*, **271**, 3155-3170.

Abstract

Spinocerebellar ataxia types 2 (SCA2) and 3 (SCA3) are autosomal-dominantly inherited, neurodegenerative diseases caused by CAG repeat expansions in the coding regions of the genes encoding ataxin-2 and ataxin-3, respectively. To provide a rationale for further functional experiments, we explored the protein architectures of ataxin-2 and ataxin-3. Using structure-based multiple sequence alignments of homologous proteins, we investigated domains, sequence motifs, and interaction partners. Our analyses focused on presumably functional amino acids and the construction of tertiary structure models of the RNA-binding Lsm domain of ataxin-2 and the deubiquitinating Josephin domain of ataxin-3. We also speculate about distant evolutionary relationships of ubiquitin-binding UIM, GAT, UBA and CUE domains and helical ANTH and UBX domain extensions.

10. Publication (Albrecht and Lengauer, 2004a)

Albrecht, M. and Lengauer, T. (2004) Novel Sm-like proteins with long C-terminal tails and associated methyltransferases. *FEBS Lett*, **569**, 18-26.

Abstract

Sm and Sm-like proteins of the Lsm (like Sm) domain family are generally involved in essential RNA-processing tasks. While recent research has focused on the function and structure of small family members, little is known about Lsm domain proteins carrying additional domains. Using an integrative bioinformatics approach, we discovered five novel groups of Lsm domain proteins (Lsm12-16) with long C-terminal tails and investigated their functions. All of them are evolutionarily conserved in eukaryotes with an N-terminal Lsm domain to bind nucleic acids followed by as yet uncharacterized C-terminal domains and sequence motifs. Based on known yeast interaction partners, Lsm12-16 may play important roles in RNA metabolism. Particularly, Lsm12 is possibly involved in mRNA degradation or tRNA splicing, and Lsm13-16 in the regulation of the mitotic G2/M phase. Lsm16 proteins have an additional C-terminal YjeF_N domain of as yet unknown function. The identification of an additional methyltransferase domain at the C-terminus of one of the Lsm12 proteins also led to the recognition of three new groups of methyltransferases, presumably dependent on S-adenosyl-L-methionine. Further computational analyses revealed that some methyltransferases contain putative RNA-binding helix-turn-helix domains and zinc fingers.

11. Publication (Albrecht and Lengauer, 2004b)

Albrecht, M. and Lengauer, T. (2004) Survey on the PABC recognition motif PAM2. *Biochem Biophys Res Commun*, **316**, 129-138.

Abstract

The PABP-interacting motif PAM2 has been identified in various eukaryotic proteins as an important binding site for the PABC domain. This domain is contained in homologs of the poly(A)-binding protein PABP and the ubiquitin-protein ligase HYD. Despite the importance of the PAM2 motif, a comprehensive analysis of its occurrence in different proteins has been missing. Using iterated sequence profile searches, we obtained an extensive list of proteins carrying the PAM2 motif. We discuss their functional context and domain architecture, which often consists of RNA-binding domains. Our list of PAM2 motif proteins includes eukaryotic homologs of eRF3/GSPT1/2, PAIP1/2, Tob1/2, ataxin-2, RBP37, RBP1, Blackjack, HELZ, TPRD, USP10, ERD15, C1D4.14, and the viral protease P29. The identification of the PAM2 motif in as yet uncharacterized proteins can give valuable hints with respect to their cellular function and potential interaction partners and suggests further experimentation. It is also striking that the PAM2 motif appears to occur solely outside globular protein domains.

12. Publication (Stoll *et al.*, 2004)

Stoll, M., Corneliussen, B., Costello, C.M., Waetzig, G.H., Mellgard, B., Koch, W.A., Rosenstiel, P., Albrecht, M., Croucher, P.J., Seegert, D., Nikolaus, S., Hampe, J., Lengauer, T., Pierrou, S., Foelsch, U.R., Mathew, C.G., Lagerstrom-Fermer, M. and Schreiber, S. (2004) Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nat Genet*, **36**, 476-480.

Abstract

Crohn disease and ulcerative colitis are two subphenotypes of inflammatory bowel disease (IBD), a complex disorder resulting from gene-environment interaction. We refined our previously defined linkage region for IBD on chromosome 10q23 and used positional cloning to identify genetic variants in DLG5 associated with IBD. DLG5 encodes a scaffolding protein involved in the maintenance of epithelial integrity. We identified two distinct haplotypes with a replicable distortion in transmission ($P = 0.000023$ and $P = 0.004$ for association with IBD, $P = 0.00012$ and $P = 0.04$ for association with Crohn disease). One of the risk-associated DLG5 haplotypes is distinguished from the common haplotype by a nonsynonymous single-nucleotide polymorphism 113G→A, resulting in the amino acid substitution R30Q in the DUF622 domain of DLG5. This mutation probably impedes scaffolding of DLG5. We stratified the study sample according to the presence of risk-associated CARD15 variants to study

potential gene-gene interaction. We found a significant difference in association of the 113A DLG5 variant with Crohn disease in affected individuals carrying the risk-associated CARD15 alleles versus those carrying non-risk-associated CARD15 alleles. This is suggestive of a complex pattern of gene-gene interaction between DLG5 and CARD15, reflecting the complex nature of polygenic diseases. Further functional studies will evaluate the biological significance of DLG5 variants.

13. Publication (Albrecht, 2005)

Albrecht, M. (2005) LRRK2 mutations and Parkinsonism. *Lancet*, **365**, 1230.

Abstract

None.

14. Publication (Albrecht *et al.*, 2005a)

Albrecht, M., Choubey, D. and Lengauer, T. (2005) The HIN domain of IFI-200 proteins consists of two OB folds. *Biochem Biophys Res Commun*, **327**, 679-687.

Abstract

The interferon-inducible p200 (IFI-200/HIN-200) family of proteins regulates cell growth and differentiation, and confers resistance to the development of tumors and virus infections. IFI-200 family members are thought to exert their biological effects by modulation of the transcriptional activities of numerous factors and interaction with other proteins through the C-terminal HIN domains. However, the HIN domain structure and function have remained obscure. Therefore, we performed a comprehensive bioinformatics analysis and assembled a structure-based multiple sequence alignment of IFI-200 proteins. The application of fold recognition methods revealed that the HIN domain consists of two consecutive OB domains. Our structural models of DNA-binding HIN domains afford the long-sought interpretations for many previous experimental observations. Our results also raise the possibility of as yet unexplored functional roles of IFI-200 proteins as transcriptional regulators and as interaction partners of proteins involved in immunomodulatory and apoptotic processes.

15. Publication (Albrecht *et al.*, 2005b)

Albrecht, M., Huthmacher, C., Tosatto, S.C. and Lengauer, T. (2005) Decomposing protein networks into domain-domain interactions. *Bioinformatics*, **21 Suppl 2**, ii220-ii221.

Abstract

The application of novel experimental techniques has generated large networks of protein-protein interactions. Frequently, important information on the structure and cellular function of protein-protein interactions can be gained from the domains of interacting proteins. We have designed a Cytoscape plugin that decomposes interacting proteins into their respective domains and computes a putative network of corresponding domain-domain interactions. To this end, the network graph of proteins has been extended by additional node and edge types for domain interactions, including different node and edge shapes and coloring schemes used for visualization. An additional plugin provides supplementary web links to Internet resources on domain function and structure.

16. Publication (Bojunga *et al.*, 2005)

Bojunga, J., Welsch, C., Antes, I., Albrecht, M., Lengauer, T. and Zeuzem, S. (2005) Structural and functional analysis of a novel mutation of CYP21B in a heterozygote carrier of 21-hydroxylase deficiency. *Hum Genet*, **117**, 558-564.

Abstract

Congenital adrenal hyperplasia (CAH) due to 21-hydroxylase deficiency is one of the most common autosomal recessive disorders and occurs in its non-classical form in up to 6% of hirsute women. We report on a young woman with the clinical diagnosis of non-classical CAH and a novel, heterozygous missense mutation CTG→GTG in exon 8, codon 317, of the steroid 21-hydroxylase CYP21B and complete loss of pseudogenes. Protein sequences of closely related P450 cytochromes and a homology-based 3D model of CYP21B were used for further functional analyses. We found that the mutated residue is part of a large cluster of hydrophobic residues. This cluster has three important features: (1) it is located directly next to the binding pocket, in close vicinity of the heme-cofactor, (2) all amino acids of the cluster are directly connected to two important binding regions, and (3) the packing within the cluster is very dense. Due to the tight packing in the cluster and its direct connection to the binding pocket region, any changes induced by the mutation of residue 317 can be expected to lead to structural shifts within the binding pocket and can explain the clinically observed impairment of 21-hydroxylase activity. In conclusion, the novel mutation L317V of the steroid 21-hydroxylase gene is associated with reduced steroid 21-hydroxylase activity

probably due to structural shifts within the binding pocket and a mild phenotype of steroid 21-hydroxylase deficiency. In addition, the results support previous findings in which heterozygous CYP21 mutations are associated with symptoms of hyperandrogenism in susceptible individuals.

17. Publication (Castellano *et al.*, 2005)

Castellano, S., Lobanov, A.V., Chapple, C., Novoselov, S.V., Albrecht, M., Hua, D., Lescure, A., Lengauer, T., Krol, A., Gladyshev, V.N. and Guigo, R. (2005) Diversity and functional plasticity of eukaryotic selenoproteins: Identification and characterization of the SelJ family. *Proc Natl Acad Sci U S A*, **102**, 16188-16193.

Abstract

Selenoproteins are a diverse group of proteins that contain selenocysteine (Sec), the 21st amino acid. In the genetic code, UGA serves as a termination signal and a Sec codon. This dual role has precluded the automatic annotation of selenoproteins. Recent advances in the computational identification of selenoprotein genes have provided a first glimpse of the size, functions, and phylogenetic diversity of eukaryotic selenoproteomes. Here, we describe the identification of a selenoprotein family named SelJ. In contrast to known selenoproteins, SelJ appears to be restricted to actinopterygian fishes and sea urchin, with Cys homologues only found in cnidarians. SelJ shows significant similarity to the jellyfish J1-crystallins and with them constitutes a distinct subfamily within the large family of ADP-ribosylation enzymes. Consistent with its potential role as a structural crystallin, SelJ has preferential and homogeneous expression in the eye lens in early stages of zebrafish development. A structural role for SelJ would be in contrast to the majority of known selenoenzymes. The unusually highly restricted phylogenetic distribution of SelJ, its specialization, and the comparative analysis of eukaryotic selenoproteomes reveal the diversity and functional plasticity of selenoproteins and point to a mosaic evolution of the use of Sec in proteins.

18. Publication (Costello *et al.*, 2005)

Costello, C.M., Mah, N., Häsler, R., Rosenstiel, P., Waetzig, G.H., Hahn, A., Lu, T., Gurbuz, Y., Nikolaus, S., Albrecht, M., Hampe, J., Lucius, R., Klöppel, G., Eickhoff, H., Lehrach, H., Lengauer, T. and Schreiber, S. (2005) Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays. *PLoS Med*, **2**, e199.1-17.

Abstract

The differential pathophysiologic mechanisms that trigger and maintain the two forms of inflammatory bowel disease (IBD), Crohn disease (CD), and ulcerative colitis (UC) are only partially understood. cDNA microarrays can be used to decipher gene regulation events at a genome-wide level and to identify novel unknown genes that might be involved in perpetuating inflammatory disease progression. High-density cDNA microarrays representing 33,792 UniGene clusters were prepared. Biopsies were taken from the sigmoid colon of normal controls (n = 11), CD patients (n = 10) and UC patients (n = 10). ^{33}P-radiolabeled cDNA from purified poly(A)$^{+}$ RNA extracted from biopsies (unpooled) was hybridized to the arrays. We identified 500 and 272 transcripts differentially regulated in CD and UC, respectively. Interesting hits were independently verified by real-time PCR in a second sample of 100 individuals, and immunohistochemistry was used for exemplary localization. The main findings point to novel molecules important in abnormal immune regulation and the highly disturbed cell biology of colonic epithelial cells in IBD pathogenesis, e.g., CYLD (cylindromatosis, turban tumor syndrome) and CDH11 (cadherin 11, type 2). By the nature of the array setup, many of the genes identified were to our knowledge previously uncharacterized, and prediction of the putative function of a subsection of these genes indicate that some could be involved in early events in disease pathophysiology. A comprehensive set of candidate genes not previously associated with IBD was revealed, which underlines the polygenic and complex nature of the disease. It points out substantial differences in pathophysiology between CD and UC. The multiple unknown genes identified may stimulate new research in the fields of barrier mechanisms and cell signalling in the context of IBD, and ultimately new therapeutic approaches.

19. Publication (Ralser *et al.*, 2005a)

Ralser, M., Albrecht, M., Nonhoff, U., Lengauer, T., Lehrach, H. and Krobitsch, S. (2005) An integrative approach to gain insights into the cellular function of human ataxin-2. *J Mol Biol*, **346**, 203-214.

Abstract

Spinocerebellar ataxia type 2 (SCA2) is a hereditary neurodegenerative disorder caused by a trinucleotide expansion in the SCA2 gene, encoding a polyglutamine stretch in the gene product ataxin-2 (ATX2), whose cellular function is unknown. However, ATX2 interacts with A2BP1, a protein containing an RNA-recognition motif, and the existence of an interaction motif for the C-terminal domain of the poly(A)-binding protein (PABC) as well as an Lsm (Like Sm) domain in ATX2 suggest that ATX2 like its yeast homolog Pbp1 might be involved in RNA metabolism. Here, we show that, similar to Pbp1, ATX2 suppresses the petite (pet^-) phenotype of Δmrs2 yeast strains lacking mitochondrial group II introns. This finding points to a close functional relationship between the two homologs. To gain insight into potential functions of ATX2, we also generated a comprehensive protein interaction network for Pbp1 from publicly available databases, which implicates Pbp1 in diverse RNA-processing pathways. The functional relationship of ATX2 and Pbp1 is further corroborated by the experimental confirmation of the predicted interaction of ATX2 with the cytoplasmic poly(A)-binding protein 1 (PABP) using yeast-2-hybrid analysis as well as co-immunoprecipitation experiments. Immunofluorescence studies revealed that ATX2 and PABP co-localize in mammalian cells, remarkably, even under conditions in which PABP accumulates in distinct cytoplasmic foci representing sites of mRNA triage.

20. Publication (Ralser *et al.*, 2005b)

Ralser, M., Nonhoff, U., Albrecht, M., Lengauer, T., Wanker, E.E., Lehrach, H. and Krobitsch, S. (2005) Ataxin-2 and huntingtin interact with endophilin-A complexes to function in plastin-associated pathways. *Hum Mol Genet*, **14**, 2893-2909.

Abstract

Spinocerebellar ataxia type 2 is an inherited neurodegenerative disorder that is caused by an expanded trinucleotide repeat in the SCA2 gene, encoding a polyglutamine stretch in the gene product ataxin-2. Although evidence has been provided that ataxin-2 is involved in RNA metabolism, the physiological function of ataxin-2 remains unclear. Here, we demonstrate that ataxin-2 interacts with two members of the endophilin family, endophilin-A1 and endophilin-A3. To elucidate the physiological implications of these interactions, we exploited yeast as a model system and discovered that expression of ataxin-2 as well as both endophilin proteins is toxic for yeast lacking the

SAC6 gene product fimbrin, a protein involved in actin filament organization and endocytotic processes. Intriguingly, expression of huntingtin, another polyglutamine protein interacting with endophilin-A3, was also toxic in Δsac6 yeast. These effects can be suppressed by simultaneous expression of one of the two human fimbrin orthologs, L- or T-plastin. Moreover, we have discovered that ataxin-2 associates with L- and T-plastin and that overexpression of ataxin-2 leads to accumulation of T-plastin in mammalian cells. Thus, our findings suggest an interplay between ataxin-2, endophilin proteins and huntingtin in plastin-associated cellular pathways.

21. Publication (Sarrazin *et al.*, 2005)

Sarrazin, C., Mihm, U., Herrmann, E., Welsch, C., Albrecht, M., Sarrazin, U., Traver, S., Lengauer, T. and Zeuzem, S. (2005) Clinical significance of in vitro replication-enhancing mutations of the hepatitis C virus (HCV) replicon in patients with chronic HCV infection. *J Infect Dis*, **192**, 1710-1719.

Abstract

Mutations in nonstructural (NS) hepatitis C virus (HCV) proteins enhance replication in HCV-1a/b replicons. The prevalence of such mutations and their clinical significance in vivo are unknown. Parts of HCV NS3 and NS4B-NS5B genes that included 31 in vitro replication-enhancing sites were sequenced for 26 patients with chronic HCV genotype 1 infection. Five patients showed specific mutations within NS3 at sites enhancing replication in the replicon. Those mutations were associated with a slower decrease in HCV RNA concentration during interferon(IFN)-α-based therapy (P = .007). Neither specific nor other mutations within NS3 and NS4B-NS5B were associated with baseline HCV RNA concentrations. Within NS5A, fewer mutations in the major HCV strain (P = .001) and increased quasi-species complexity (P = .02) and diversity (P = .02) correlated with increasing baseline HCV RNA concentrations. *In silico* analyses of NS3 protein structures suggested that the majority of observed mutations did not lead to major conformational changes. Specific mutations leading to enhanced replication in the replicon system were detected in 5 of 26 patients in vivo and were not associated with baseline HCV RNA concentrations but were associated with a slower decrease in HCV RNA concentration during IFN-α-based therapy. Quasi-species heterogeneity of NS5A correlated with baseline HCV RNA concentrations.

22. Publication (Schreiber *et al.*, 2005)

Schreiber, S., Rosenstiel, P., Albrecht, M., Hampe, J. and Krawczak, M. (2005) Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat Rev Genet*, **6**, 376-388.

Abstract

Chronic inflammatory disorders such as Crohn disease, atopic eczema, asthma and psoriasis are triggered by hitherto unknown environmental factors that function on the background of some polygenic susceptibility. Recent technological advances have allowed us to unravel the genetic aetiology of these and other complex diseases. Using Crohn disease as an example, we show how the discovery of susceptibility genes furthers our understanding of the underlying disease mechanisms and how it will, ultimately, give rise to new therapeutic developments. The long-term goal of such endeavours is to develop targeted prophylactic strategies. These will probably target the molecular interaction on the mucosal surface between the products of the genome and the microbial metagenome of a patient.

23. Publication (Valentonyte *et al.*, 2005)

Valentonyte, R., Hampe, J., Huse, K., Rosenstiel, P., Albrecht, M., Stenzel, A., Nagy, M., Gaede, K.I., Franke, A., Haesler, R., Koch, A., Lengauer, T., Seegert, D., Reiling, N., Ehlers, S., Schwinger, E., Platzer, M., Krawczak, M., Müller-Quernheim, J., Schürmann, M. and Schreiber, S. (2005) Sarcoidosis is associated with a truncating splice site mutation in BTNL2. *Nat Genet*, **37**, 357-364.

Abstract

Sarcoidosis is a polygenic immune disorder with predominant manifestation in the lung. Genome-wide linkage analysis previously indicated that the extended major histocompatibility locus on chromosome 6p was linked to susceptibility to sarcoidosis. Here, we carried out a systematic three-stage SNP scan of 16.4 Mb on chromosome 6p21 in as many as 947 independent cases of familial and sporadic sarcoidosis and found that a 15-kb segment of the gene butyrophilin-like 2 (BTNL2) was associated with the disease. The primary disease-associated variant (rs2076530; P(TDT) = 3×10^{-6}, P(case-control) = 1.1×10^{-8}; replication P(TDT) = 0.0018, P(case-control) = 1.8×10^{-6}) represents a risk factor that is independent of variation in HLA-DRB1. BTNL2 is a member of the immunoglobulin superfamily and has been implicated as a costimulatory molecule involved in T-cell activation on the basis of its homology to B7-1. The G→A transition constituting rs2076530 leads to the use of a cryptic splice site located 4 bp upstream of the affected wild-type donor site. Transcripts of the risk-associated allele have a premature stop in the spliced mRNA. The resulting protein lacks the C-terminal

IgC domain and transmembrane helix, thereby disrupting the membrane localization of the protein, as shown in experiments using green fluorescent protein and V5 fusion proteins.

24. Publication (Van Duist *et al.*, 2005)

Van Duist, M.M., Albrecht, M., Podswiadek, M., Giachino, D., Lengauer, T., Punzi, L. and De Marchi, M. (2005) A new CARD15 mutation in Blau syndrome. *Eur J Hum Genet*, **13**, 742-747.

Abstract

The caspase recruitment domain gene CARD15/NOD2, encoding a cellular receptor involved in an NF-κB-mediated pathway of innate immunity, was first identified as a major susceptibility gene for Crohn's disease (CD), and more recently, as responsible for Blau syndrome (BS), a rare autosomal-dominant trait characterized by arthritis, uveitis, skin rash and granulomatous inflammation. While CARD15 variants associated with CD are located within or near the C-terminal leucine-rich repeat domain and cause decreased NF-κB activation, BS mutations affect the central nucleotide-binding NACHT domain and result in increased NF-κB activation. In an Italian family with BS, we detected a novel mutation E383K, whose pathogenicity is strongly supported by cosegregation with the disease in the family and absence in controls, and by the evolutionary conservation and structural role of the affected glutamate close to the Walker B motif of the nucleotide-binding site in the NACHT domain. Interestingly, substitutions at corresponding positions in another NACHT family member cause similar autoinflammatory phenotypes.

25. Publication (Albrecht and Takken, 2006)

Albrecht, M. and Takken, F.L.W. (2006) Update on the domain architectures of NLRs and R proteins. *Biochem Biophys Res Commun*, **339**, 459-462.

Abstract

None.