Alternative Splicing and Protein Structure Evolution

Fabian Birzele

SIGILL:UNIVERSITAT:LUDOVIC:MAXIM:

München 2008

Alternative Splicing and Protein Structure Evolution

Fabian Birzele

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Fabian Birzele
aus Würzburg

München, den 09.12.2008

Erstgutachter: Prof. Dr. Ralf Zimmer
Zweitgutachter: Prof. Dr. Dimtrij Frishman
Tag der mündlichen Prüfung: 27.01.2009

Contents

Abstract xiii

Zusammenfassung xv

1 Introduction 1

I Protein Structure Comparison 7

2 Introduction to Protein Structure Analysis 9
- 2.1 Goals and open questions in protein structure analysis 9
- 2.2 Existing approaches to protein structure comparison 10
  - 2.2.1 SCOP and CATH - the standard of truth 10
  - 2.2.2 Automated protein structure alignment and comparison 11

3 A Systematic Comparison of SCOP and CATH 13
- 3.1 Methods 14
  - 3.1.1 Mapping of domain assignments 14
  - 3.1.2 Mapping inner nodes of the hierarchies 15
- 3.2 Results 15
  - 3.2.1 Datasets 15
  - 3.2.2 Detailed Comparison of SCOP and CATH 16
- 3.3 Applications of the SCOP-CATH mapping 20
  - 3.3.1 Benchmarking Structure-Comparison methods 20
  - 3.3.2 Inter-Fold Similarities revealed by Consistency Checks 24
- 3.4 Conclusion 25

4 Protein Structure Alignment considering Phenotypic Plasticity 27
- 4.1 Outline of the method 29
- 4.2 Conclusion 30

5 Vorolign - fast protein structure alignment using Voronoi contacts 33
- 5.1 Methods 34
  - 5.1.1 Voronoi tessellation of protein structures 34

5.1.2 Properties of Voronoi cells 35
5.1.3 Similarity of Voronoi cells 36
5.1.4 Pairwise alignment of protein structures 37
5.1.5 Multiple alignment of protein structures 38
5.1.6 Fast scan for family members: VorolignScan 39
5.1.7 Automatic detection of domains in multi-domain structures 40
5.1.8 Parameter Optimization 40
5.2 Results 41
5.2.1 Family recognition on a testset of 979 query proteins 41
5.2.2 Detailed Evaluation on the SCOP-CATH set 43
5.2.3 Alignment quality 46
5.2.4 Specific examples for multiple alignment of protein structures 47
5.2.5 Large scale evaluation of multiple alignment quality 49
5.3 The AutoPSI database 52
5.3.1 The current database content 52
5.3.2 Database access 53
5.4 Voronoi contact patterns around functionally important sites 55
5.4.1 Outline of the method 56
5.4.2 Results: Case study of the Trypsin-like serine protease family 58
5.4.3 Discussion 60
5.5 Conclusions 61

II Alternative Splicing in the Context of Protein Structure 63

6 Introduction to the Analysis of Alternative Splicing 65
6.1 Fundamentals of alternative splicing 66
6.1.1 The splicing process in the cell 66
6.1.2 Different types of alternative splicing events 66
6.1.3 Regulation of alternative splicing 67
6.2 Biological functions of alternative splicing events 68
6.3 Alternative splicing and disease 69

7 Alternative Splicing and Protein Structure Evolution 71
7.1 Hypotheses and major concepts 72
7.2 Methods 72
7.2.1 Alternative splicing and literature data 73
7.2.2 Protein structure assignment 73
7.2.3 Assignment of protein structures to families 74
7.2.4 Multiple structure alignments and evolutionary ”isoforms” 74
7.2.5 Alternative splicing and alternative structural models 74
7.3 Results 75
7.3.1 Structural complexity of functional isoforms 75

7.3.2 Interpretation of splicing events using evolutionary isoforms 79
7.3.3 Supporting hypotheses on fold evolution via alternative splicing data . . . 80
7.3.4 Alternative splicing: a mechanism to explore the protein fold space? . . . 81
7.4 Discussion 84

8 Alternative Splicing and Protein Repeats 87
8.1 Methods 88
8.1.1 Genomic data and annotated splice variants 88
8.1.2 Splicing events annotated in Swissprot 88
8.1.3 Repeat assignment and data collection 89
8.1.4 Assessing the statistical significance of the results 89
8.2 Results 89
8.2.1 Frequency of repeats in Swissprot and Ensembl 89
8.2.2 Alternative splicing of proteins containing repeats 90
8.2.3 Alternative splicing of C_2H_2-Zinc-finger motifs 94
8.2.4 Sculpting protein binding domains - Ankyrin repeats 95
8.2.5 Alternative splicing of β-propellers 97
8.2.6 Importance of splicing for complex tissue organizations 97
8.3 Discussion 98

III Genome-wide Detection and Analysis of Alternative Splicing 101

9 Detection and prediction of Alternative Splicing Events 103

10 The ProSAS - Database 107
10.1 The ProSAS pipeline 107
10.1.1 Genome and alternative splicing data 108
10.1.2 Protein structure prediction 108
10.1.3 Characterization of splicing events 109
10.1.4 Affymetrix data mapping 110
10.1.5 Further data sources 110
10.2 The ProSAS database content 110
10.3 The ProSAS database web interface 112
10.4 Conclusion 115

11 PASS - Identifying splicing events in Affymetrix Exon Arrays 117
11.1 Outline of the method 117
11.2 Results and Discussion 119

12 Alternative splicing and proteome complexity in mass spectrometry data 123
12.1 Methods 124
12.1.1 Identification of isoforms in precompiled MS peptide datasets 124
12.1.2 Expected versus observed unique isoform peptides 124

12.1.3 Sequence-based analysis of alternative splicing 125
12.1.4 Protein structure prediction pipeline 125
12.2 Results 125
12.2.1 Presence of isoforms in mass spectrometry data 126
12.2.2 Changing signal peptides, localization and domain composition via alternative splicing 128
12.2.3 Structural analysis of isoforms identified in mass spectrometry data . . . 130
12.3 Discussion 134
12.4 Outlook: Identification of novel splice variants in mass spectrometry data 135

IV Conclusions and Outlook 139

13 Conclusion and Outlook 141

Bibliography 147

Acknowledgements 163

Curriculum Vitae 165

List of Figures

3.1 Examples of interfold similarities detected by the systematic comparison of SCOP and CATH 21
3.2 Evaluation of the performance of TM-align on SCOP and the novel benchmark set of consistently classified pairs of structures 22

4.1 Phenotypic plasticity observed in the pheromone binding domain family (a.39.2.1) 28
4.2 Schematic overview on the single steps of the PPM method 29

5.1 Anomaly in a structural alignment due to structural divergence 34
5.2 Construction of a two-dimensional Voronoi tessellation 35
5.3 Basic principle of the Vorolign scoring function to score the similarity of two Voronoi cells 38
5.4 Detailed evaluation of the performance of PPM and Vorolign on the SCOP-CATH consensus set 45
5.5 Structural quality of Vorolign alignments across different sequence identities in more than 8000 pairs of structures from the same SCOP family 46
5.6 Flexible structural alignment of members of the Calmodulin family computed by Vorolign 47
5.7 Pairwise Vorolign alignment quality induced from multiple structure alignments of more than 800 SCOP families 49
5.8 Conservation rate of biologically relevant residues (CSA patterns and cystein bridges) in multiple structure alignments 51
5.9 Conservation rate of biologically relevant residues (PROSITE patterns and Swissprot annotated functional residues) in multiple structure alignments 51
5.10 Detail view in the AutoPSI database web interface 54
5.11 Schematic example of the consistency check in the Vorolign Pattern identification method 57
5.12 PROSITE pattern matches of the catalytic triad on the Trypsin structure (PDB: 1a0j) 58
5.13 Initial Voronoi pattern in the Trypsin structure (PDB: 1a0j) 59
5.14 Final consensus Voronoi pattern for eukaryotic Tyrpsin-like serine proteases family (SCOP: b.47.1.2) 60

6.1 Important consensus splicing signals in intronic sequences 66
6.2 Different types of alternative splicing events 67

7.1 Schematic workflow of our approach for a structure based analysis of splicing events 73
7.2 Distribution of 488 splicing events in domains, large parts (but no domain) as well as variable and conserved regions 75
7.3 Examples for the location of splicing events of different types, known to lead to functional proteins, visualized on the corresponding protein structure 77
7.4 Example how evolutionary isoforms can help to understand the outcome of splicing events 80
7.5 Examples for potential fold changing events mediated by alternative splicing . . . 82

8.1 Evolutionary model for the enhanced generation of new protein variants by coupling repeat expansion (duplication) and alternative splicing 91
8.2 Schematic mechanism how to generate functional diversity through spliced repeat proteins 92
8.3 Examples for splicing events affecting repeats on the protein structure level . . . 96

9.1 Experimental approaches to detect alternative splicing events 104

10.1 Datasources and their interdependency in the ProSAS database 108
10.2 ProSAS gene detail view (Transcript details) in the ProSAS web interface 113
10.3 ProSAS transcript detail view in the ProSAS web interface 114

11.1 Examples for tissue-specific splice variants detected by PASS 121

12.1 Basic principle of our method to detect known isoforms in mass spectrometry data 126
12.2 Evaluation of the structural complexity of splicing events confirmed in mass spectrometry data 131
12.3 Examples of splicing events confirmed by mass spectrometry data visualized on the corresponding protein structure 133
12.4 Novel Vitamin-D binding protein splice junction and isoform identified in mouse plasma mass spectrometry data 138

List of Tables

3.1 Evaluation of the differences in the domain definitions between SCOP and CATH 16
3.2 Results of our mapping of SCOP nodes onto the CATH hierarchy 18
3.3 Pairwise inconsistencies between SCOP and CATH observed in our mapping (with SCOP as reference) 18
3.4 Details on the consistencies and inconsistencies observed in the mapping of domain pairs from SCOP onto CATH 19

5.1 Optimized parameter settings for Vorolign in combination with four different scoring matrices 40
5.2 Comparison of the structure classification accuracy of different methods on a set 974 proteins from SCOP version 1.67 not contained in 1.65 41
5.3 Family recognition rates of Vorolign, PPM and TM-align on the novel benchmark set described in Chapter 3 44
5.4 Evaluation of the recognition errors made by Vorolign, PPM and TM-align on the novel benchmark set described in Chapter 3 44

7.1 Distribution of non-trivial isoforms in eight different event categories 75

8.1 Overview on the most frequent repetitive patterns in the human genome (Ensembl) and the complete SwissProt database 90
8.2 Top 15 spliced patterns in the human genome sorted by their statistical significance (LOD score) 93
8.3 Top 15 spliced patterns in the SwissProt database sorted by their statistical significance (LOD score) 94
8.4 Overview on the number of events changing the number of repeats (in contrast to events removing the complete repetitive domain) 95

10.1 Genomic database content in ProSAS in June 2008 (Release 2.0) 111
10.2 Structural coverage of Ensembl genes and Swissprot entries in ProSAS in June 2008 (Release 2.0) 111
10.3 Coverage of database entries with InterPro features in ProSAS in June 2008 (Release 2.0) 112

11.1 Prediction accuracy of the PASS method on a set of 11 human tissues with respect to Ensembl as gold standard . . . 119

12.1 Basic data content of the mass spectrometry and protein datasets used to study isoforms observed and confirmed in mass spectrometry data . . . 127
12.2 Overview on the number of isoforms confirmed in mass spectrometry dataset of Human, Mouse and Drosophila . . . 127
12.3 Observed vs. expected number of isoforms identified in mass spectrometry data . 128

Summary

The last years have seen a tremendous increase of biological data being available from different sources allowing for detailed insights into the function of cellular components like genes and proteins, their connection to cellular processes and networks as well as their evolution. Consequently, it has led to the need for a computational analysis and interpretation of the large amount of data being produced by experimentalists in order to understand those highly complex systems in more detail.

In this thesis we have focussed on the analysis of two important and intriguing problems in current bioinformatics research, namely the analysis of protein structure evolution and protein structure similarity, as well as the computational analysis of alternative splicing which is an important mechanism contributing to proteome diversity in higher organisms. Moreover, we have introduced a combined analysis of alternative splicing and protein structure evolution which allows for novel insights into both, functional diversity arising from alternative splicing and structural diversity arrising in the course of protein evolution.

In particular, Part I of the thesis describes our work in the field of protein structure comparison.

We have analyzed the two most important resources and gold standard datasets for protein structure classification, SCOP and CATH, and have extracted valuable knowledge from the comparison of both using the orthogonal information contained in those datasets [35].

Furthermore, we will introduce two novel methods, PPM [35] and Vorolign [15], to analyze the similarities and differences in groups of proteins both implementing novel concepts to score and detect structural similarities. While PPM focuses on the computation of highly reliable, conserved cores of protein structure families in the presence of the natural variance observed in those groups (phenotypic plasticity), Vorolign describes a more sequence-based and contact driven similarity function between proteins which allows for a rapid and accurate classification of protein structures and the computation of accurate alignments in terms of structural and sequential quality. Vorolign finds its large-scale application in the AutoPSI database [16]. We further extend the Vorolign idea to the identification of highly conserved, functionally important residue networks in protein families.

The second part of this thesis (Part II) will then deal with the joint analysis of alternative splicing in the light of protein structure evolution.

We will first of all provide a detailed analysis of the complexity of alternative splicing events on the protein structure level [14], introducing a novel approach to understand this complexity by using data on protein structure evolution and the concept of ”evolutionary isoforms”. This analysis shows the large structural complexity of alternative splicing which requires for an explanation why proteins could be able to cope with such major rearrangements. We will then show that the tolerance of structures against major rearrangements can be linked to the evolutionary history of the corresponding protein fold and we can therefore use data on protein structure evolution to predict the outcome of certain splicing events (e.g. via evolutionary isoforms) and vice versa, we can use data on splicing to analyze protein structure evolution. In more detail we will show that we can confirm existing hypotheses on the evolution of specific fold classes via alternative splicing data and furthermore can propose novel hypotheses on fold evolution and the topology of the protein fold space utilizing known splicing events.

We will then extend our ideas to a specific group of proteins, namely proteins made up from repetitive structure motifs, and the variability generated in this group through alternative splicing in very detail. We will show that splicing preferentially makes use of those motifs to increase functional complexity in higher organisms in various functional categories starting from transcriptional control and protein-protein interactions to the organization of complex tissues [17].

In Part III of the thesis we will discuss our approaches and methods for a genome-wide detection and analysis of alternative splicing events in high-throughput experimental datasets.

In more detail we will describe the ProSAS resource [19] which is a comprehensive database for the genome-wide analysis of alternative splicing events annotated in Swissprot and Ensembl as well as identified in mass spectrometry and Affymetrix exon array datasets in the context of protein structures and other functional annotations. ProSAS has been used as the primary data resource for all splicing-related analyses carried out in the course of this thesis.

We will then briefly describe our method for analyzing Affymetrix exon arrays called PASS [89] and discuss examples for its capability to detect tissue-specific splice variants.

Since most data on alternative splicing only provides evidence for the existence of certain isoforms on the mRNA level (via EST or cDNA confirmation) all analyses setting up on this data face the problem of potential nonsense-mediated decay of the final transcripts or a fast degradation of non-folding, unstable proteins after translation which could adulterate the conclusions made. We have therefore analyzed all isoforms observed in mass spectrometry datasets from Human, Mouse and Drosophila in some detail on a sequence and structure basis and find that our conclusions on the structural complexity of alternative splicing and the existence of non-trivial isoforms on the structure level also hold in this set of isoforms confirmed on the protein level [20]. We further extend our search for such mass spectrometry confirmed isoforms to the re-evaluation of raw mass spectrometry datasets and suggest an alternative search strategy to identify novel and so far unknown splice variants.

Zusammenfassung

In den letzten Jahren erleben wir einen dramatischen Anstieg zur Verfügung stehender biologischer Daten in verschiedenen Bereichen der Biologie und Biochemie. Diese erlauben es uns, die Funktion zellulärer Komponenten wie Gene und Proteine, ihre Verknüpfungen in zellulären Prozessen und ihre Evolution im Laufe der Jahrmillionen detailliert zu analysieren. Um aus diesen Daten sinnvolles Wissen zu extrahieren und die hochkomplexen Zusammenhänge in Zellen zu verstehen sind computergestützte Analysen nötig.

Diese Arbeit beschäftigt sich hauptsächlich mit der Analyse zweier wichtiger und faszinierender Probleme der aktuellen, bioinformatischen Forschung, nämlich der Analyse von Proteinstrukturen und ihrer Evolution, sowie der computergestützten Analyse von alternativem Spleißing, einem wichtigen Mechanismus in eukaryotischen Zellen welcher maßgeblich zu Proteomkomplexität in höheren Organismen beiträgt. Desweiteren führt diese Arbeit die kombinierte Analyse von Daten zu alternativem Spleißing und Daten zur Proteinevolution ein, welche neue Einsichten in beide Prozesse erlaubt. Insbesondere ermöglicht sie uns, funktionelle Diversität basierend auf alternativem Spleißing und strukturelle Diversität wie sie im Laufe der Evolution von Proteinen entsteht, miteinander in Beziehung zu setzten und somit detaillierter zu verstehen.

Der erste Teil der Arbeit (Part I) beschäftigt sich dabei mit unseren Arbeiten auf dem Gebiet der strukturellen Bioinformatik und insbesondere dem Vergleich von Proteinstrukturen.

Hierfür haben wir eine detaillierte Analyse der beiden wichtigsten Datenbanken zur Strukturanalyse von Proteinen, SCOP und CATH, durchgeführt und neues Wissen aus dem Vergleich beider Datenbanken und der in ihnen enthaltenen orthogonalen Informationen gezogen [35].

Des Weiteren werden wir zwei neue Methoden, PPM [35] und Vorolign [15], zur Analyse von strukturellen Ähnlichkeiten und Unterschieden in Gruppen von Proteinen einführen, die auch dem wachsenden Bedarf für eine schnelle und zuverlässige automatische Klassifikation von Proteinstrukturen genügen. Beide implementieren neue Ideen und Konzepte um strukturelle Ähnlichkeit zwischen Proteinen zu bewerten und zu erkennen. PPM zielt hierbei hauptsächlich auf die Berechnung qualitativ hochwertiger, konservierter Kerne von Strukturfamilien in der Gegenwart von natürlicher Strukturvariabilität (Phänotypische Plastizität) ab, während Vorolign einen mehr sequenz-basierten und Kontaktnetzwerk-getriebenen Ansatz beschreibt, welcher sich besonders für eine sehr schnelle und zuverlässige Klassifikation von Proteinstrukturen und der Berechnung akkurater Alignments (im Bezug auf strukturelle und sequenzielle Qualität) eignet. Vorolign wird dabei in großem Maßstab in der AutoPSI Datenbank eingesetzt [16].

Der zweite Teil der Arbeit (Part II) befasst dann mit der Analyse von alternativem Spleißing im Kontext von Proteinstrukturen und Proteinstrukturevolution.

Zunächst beschreiben wir eine detaillierte Analyse der Komplexität von Spleißereignissen auf der Proteinstrukturebene [14] und stellen eine neue Herangehensweise vor, um die beobachtete Komplexität unter Verwendung von Daten zur Proteinstrukturevolution und dem Konzept von ”evolutionären Isoformen” zu analysieren. Diese Analyse zeigt eine große Komplexität von alternativem Spleißing auf der Strukturebene, welche die Frage aufwirft, warum Proteine solche großen strukturellen Änderungen ertragen sollten. Wir werden dann zeigen, dass diese unerwartet hohe Toleranz von Proteinen gegen große Strukturänderungen in direktem Zusammenhang mit der evolutionären Geschichte der zugehörigen Faltungsklasse stehen könnte. Dies erlaubt es uns, Daten zur Strukturevolution für die Analyse von alternativem Spleißing, in Form von ”evolutionären Isoformen”, zu verwenden. Umgekehrt werden wir zeigen, dass Daten zu alternativem Spleißing helfen können existierende Hypothesen zur Evolution bestimmter Faltungsklassen zu bekräftigen und zusätzlich neue Hypothesen zur Beziehung zwischen Faltungsklassen und der Topologie des Strukturraumes aufzustellen.

Anschließend werden wir eine bestimmte Gruppe von Proteinen, nämlich Repeatproteine, analysieren und die durch Spleißing entstehende Variabilität in dieser Gruppe im Detail diskutieren. Wir können zeigen, dass Spleißing die einzelnen Motive der Repeatproteine in besonderem Maße betrifft und dass solche Ereignisse auf verschiedenen Ebenen maßgeblich zur strukturellen und funktionellen Diversität, beginnend bei der Kontrolle der Tranksription, ber die Mediation von Protein-Protein-Interaktionen bis hin zur Entstehung komplexer Gewebe in höheren Organismen, beitragen können [17].

Im letzten Teil (Part III) der Arbeit beschreiben wir Methoden und Tools, die wir zur genomweiten Detektion und Interpretation von alternativen Spleißereignissen entwickelt haben.

Wir diskutieren die ProSAS Datenbank [19] welche eine umfassende Resource für die genomweite, strukturbasierte Analyse von alternativem Spleißing, identifiziert in verschiedenen Datenbanken (Ensembl und Swissprot) und mittels verschiedener Methoden (z.B. Massenspektrometrie und Affymetrix Exonarrays), darstellt und als Rahmen für viele der in der Arbeit vorgestellten Analysen gedient hat. Wir werden dann kurz auf unsere Methode zur Analyse von Affymetrix Exon Array Daten, genannt PASS [89], eingehen.

Bisher erlauben die meisten verfügbaren Datensätze zu alternativem Spleißing nur die Existenz bestimmter Isoformen auf der mRNA-Ebene (über EST oder cDNA Sequenzierung) zu bestätigen. Alle weiteren Analysen, welche auf diesen Daten aufsetzen, müssen sich somit dem Problem des Abbaus von Stopcodon enthaltender mRNA (Nonsense-mediated decay) oder einer schnellen Degradation von unstabilen und nicht-faltenden Proteinen nach der Translation stellen, die ihre Ergebnisse und Erkenntnisse verfälschen könnten. Aus diesem Grund haben wir für die Analyse von Spleißereignissen auf der Strukturebene Daten aus einer weiteren Methode, nämlich Massenspektrometrie, hinzugezogen. Diese Daten von Mensch, Maus und Drosophila erlauben eine detaillierte sequenz- und strukturbasierte Analyse von Isoformen, welche auf der Proteinebene existieren und zeigen, dass sich unsere Ergebnisse zur strukturellen Komplexität von Isoformen [14] auch auf dieses Datenset erweitern lassen [20].

Chapter 1

Introduction

The first draft of the human genome was sequenced at a cost of $3 billion and took over 13 years to be completed. In contrast, the genome of James Watson, published in summer 2007, was completed within 2 months (at a cost of $1 million) and the Archon X Prize in Genomics has recently announced a $10 million cash award for the first group to sequence the genomes of 100 individuals in 10 days.

Advances in DNA sequencing are just one example for the explosion of biological data that became available from various experimental sources in the last 20 years. Starting from the availability of the complete sequences of hundreds of genomes from all kingdoms of life, the characterization of complete transcriptomes and proteomes via microarray analysis, deep sequencing and mass spectrometry to the large-scale experimental determination of protein structures. They all have contributed to an tremendous amount of data being available and therefore the growing need for their integrated computational analysis and professional management in terms of data storage, searchability and visualization.

Many of those novel data sources have significantly contributed to our understanding of cellular processes, evolution and disease. They have also revealed the large complexity of living organisms on a molecular basis. They might have even led to more questions than answers.

How do organisms evolve and change over time, how do they adapt to novel environmental requirements through mutation and selection, how do their genomes differ? What are the various roles and functions of small non-coding RNAs, genes, proteins, metabolites or functional modules in the cell, how do they interact and how can complexity arise from those rather simple, individual components? Can we understand organisms to a detail which allows to model the molecular basis of diseases, cellular processes and evolution? Such a *systems view* of organisms may lead to the development of novel strategies against diseases like cancer or to ”engineering” novel organisms which carry out specific functions like the optimized production of pharmaceuticals or bio fuel.

Given all those intriguing questions and the need to analyze and interpret data in a biologically meaningful way, bioinformatics and systems biology have become key players in life sciences. The diversity of the questions mentioned above and the resulting diversity of different fields in current bioinformatics research also makes clear that a thesis can only cover a very small proportion of this fast developing discipline.

Why genes in pieces?
(Walter Gilbert)

One focus of this thesis is the analysis of one intriguing mechanism contributing to functional diversity and complexity in higher organisms, namely alternative splicing. Splicing allows to assemble the exons of a gene in different ways during pre-mRNA splicing and therefore allows for generating various gene products or transcripts from one gene. The possibility to pack a large number of different gene products in one genomic locus, subject to dynamic regulation, is a very attractive idea as splicing may therefore account for the differences between different organisms despite their large similarity on the genome level.

The importance of gene products generated by alternative splicing has been shown in many cases. Nevertheless, one level in our understanding of this process is missing to date and will be addressed in this thesis. How does alternative splicing alter protein structures in order to generate isoforms which are able to carry out different functions in the cell?

The structure of a protein can reveal its function and its evolutionary history
(Philip E. Bourne)

It is the structure of a protein which finally accounts for the distinct functional properties like enzymatic reaction mechanisms or the specificity of protein-protein interactions and those can only be understood in detail on the structure level. Since the gap between known protein sequences and known protein structures is large and widening much effort has been put into methods for protein structure prediction from the sequence either in an *ab initio* fashion (mimicking the process of protein folding *in silico*) or using homology-based approaches. Such computational approaches to protein structure prediction have become more and more important for many biological and bioinformatics questions in the recent years and are evaluated in community-wide contests like the CASP or Livebench experiments.

Not only computational approaches to protein structure prediction have been improved significantly but also novel developments in experimental biology have led to a large increase in the number of known structures in the last years. In order to make use of the grown number of known protein structures for a further analysis, fast and automated methods are required to identify similarities and differences between protein structures. They also allow to address the question how proteins and protein structures are altered in the course of evolution to carry out different functions in the cell.

Outline of the thesis

In this thesis we address several intriguing questions to be asked in protein structure analysis and the computational interpretation of alternative splicing and propose novel methods in both, traditionally separated, fields of bioinformatics research. Surprisingly, some of the questions arising in the analysis of alternative splicing and the questions arising in the analysis of protein structure evolution turn out to be very similar. What are the operations used in the course of evolution to alter protein structures and how are structure evolution and the evolution of novel

functions related? Moreover, are those mechanisms similar to the operations used by alternative splicing to create functional diversity?
We therefore propose a joint analysis of alternative splicing and protein structure evolution which can be shown to lead to interesting insights into protein evolution and the origin of complex functions via alternative splicing.

Protein structure comparison is a crucial step towards a deeper understanding of the interplay of protein sequence and protein structure evolution, the principles of protein folding, topological features of the protein fold space as well as the relationships of protein structure and protein function. In **Part I** of this thesis we will therefore discuss several methods and analyses developed or co-developed by the author in the field of structural bioinformatics which may help to elucidate some of the intriguing existing and newly arising questions in the context of protein structures as well as their similarities and differences.

We will briefly introduce and discuss existing approaches in the field and some of the most important questions in current structural bioinformatics research in **Chapter 2**.

Our own work then includes the comprehensive and in-depth analysis and comparison of the two most important databases in the field of protein structure classification, namely SCOP and CATH (**Chapter 3**). We will use the orthogonal knowledge stored in both hierarchies to compile a novel benchmark set to evaluate automatic methods for protein structure classification or structure alignment and to extract interesting features on the topology of the protein structure space.

The major focus of Part I will then be on the development of novel methods to analyze similarities and differences observed in groups of proteins. We also account for the need for a fast classification and subsequent analysis of protein structures given the faster growing wealth of experimental data. In more detail we will introduce two methods for the analysis of structural relationships and the computation of protein structure alignments called PPM (in **Chapter 4**) and Vorolign (in **Chapter 5**). Both implement novel concepts to score and detect structural similarities. While PPM focuses on the computation of highly reliable, conserved cores of protein structure families in the presence of the natural variance observed in those groups (phenotypic plasticity), Vorolign describes a more sequence-based and contact driven similarity function between proteins which allows for a rapid and accurate classification of protein structures and the computation of accurate alignments in terms of structure and sequence quality. Vorolign finds its large-scale application in the AutoPSI database also discussed in **Section 5.3**. Furthermore we briefly discuss the potential application of the Vorolign scoring function for the identification of highly conserved patterns of functionally important residues in protein families (**Section 5.4**).

Part II of this thesis deals with the analysis of alternative splicing in the light of protein structure evolution. Splicing and its biological functions have been extensively studied in experimental and bioinformatics studies in the last years. Surprisingly, only little is known about how alternative splicing alters protein structures in order to generate protein variants (”isoforms”) with potentially different functions. In this part we will therefore describe several analyses and methods developed by the author in the context of a structure-based analysis of alternative splicing. We will show that an integrated analysis of alternative splicing, protein structures and protein structure evolution can help to understand this highly abundant mechanism in eukaryotic cells in more detail and leads to novel insights in both, traditionally separated, fields of research.

After a short introduction to alternative splicing and its functional roles in higher organisms in **Chapter 6** we will first of all provide a detailed analysis of the complexity of alternative splicing events on the protein structure level in **Chapter 7** introducing a novel approach to understand this complexity by using data on protein structure evolution and the concept of ”evolutionary isoforms”. This analysis reveals a large structural complexity of alternative splicing which requires for an explanation why proteins could be able to cope with such major rearrangements. We will then show that the tolerance of structures against major rearrangements can be linked to the evolutionary history of the corresponding protein fold and we can therefore use data on protein structure evolution to predict the outcome of certain splicing events (e.g. via evolutionary isoforms). Vice versa, we can use data on splicing to analyze protein structure evolution. In more detail we will show that we can confirm existing hypotheses on the evolution of specific fold classes via alternative splicing data and furthermore can propose novel hypotheses on fold evolution and the topology of the protein fold space utilizing known splicing events.

In **Chapter 8** we extend our analysis of splicing and protein structure evolution to a specific group of proteins, namely those made up from repetitive, structural motifs. We analyze the variability generated in this group through alternative splicing in very detail in a functional and evolutionary context. We will show that splicing preferentially makes use of those motifs to increase functional complexity in higher organisms in various functional categories starting from transcriptional control and protein-protein interactions to the organization of complex tissues.

Part III of the thesis will then deal with the detection and analysis of alternative splicing in genome-wide annotations and novel experimental sources like mass spectrometry or Affymetrix exon arrays. Both experimental methods have only recently become available for a genome-wide detection of proteins via small peptides (mass spectrometry) or via measuring the expression of every exon of a gene individually in contrast to standard expression arrays which use only one single sequence tag to identify the expression of specific genes. Moreover, they allow to test our hypotheses described in Part II to some extent and, together with other novel methods like Illumina’s GenomeAnalyzer, will significantly contribute to our understanding of the tissue-specific and disease-related expression of splice variants in the next years .

We will first of all introduce and discuss the different experimental methods and their ability to detect splice variants in a genome-wide fashion in **Chapter 9**.

We will then describe the ProSAS resource in **Chapter 10** which is a comprehensive database for the genome-wide analysis of alternative splicing events annotated in Swissprot and Ensembl as well as identified by high-throughput experiments, i.e. Affymetrix exon arrays and mass spectrometry, in the context of protein structures and other functional annotations. Experimental data sources and splicing events are integrated into this larger software system, including the visualization of the data in the ProSAS web application.

We will then give a brief overview on the PASS (Pairwise Alternative Splicing events via Scaling) method for analyzing Affymetrix Exon Array data in **Chapter 11**. Its main application is to detect and confirm the expression of tissue-specific splice variants annotated in Ensembl or Swissprot but PASS can potentially also be used to detect novel splice variants.

In **Chapter 12** we will analyze mass spectrometry data in the context of alternative splicing. Since most data on alternative splicing only provides evidence for the existence of certain isoforms on the mRNA level (via EST or cDNA confirmation) all analyses building on this data face the problem of potential nonsense-mediated decay of the final transcripts or a fast degradation of non-folding, unstable proteins after translation which could adulterate the conclusions made. We have analyzed all known isoforms observed in mass spectrometry datasets from Human, Mouse and Drosophila in some detail on a sequence and structure basis and find that our conclusions on the structural complexity of alternative splicing and the existence of non-trivial isoforms on the structure level (discussed in Chapter 7 and in Chapter 8) also hold in this set of isoforms confirmed on the protein level.

In the final **Chapter 13** we will discuss the methods and analyses carried out in this thesis and provide an outlook on promising extensions of our work in future research.

Part I

Protein Structure Comparison

Chapter 2

Introduction to Protein Structure Analysis

DNA and protein molecules are key players in the cell. The blue print of an organism is encoded in the DNA in the form of *genes* and most genes encode *proteins*. Via the processes of transcription and translation the information on the DNA is transcribed into messenger-RNA (mRNA) and then translated into a sequence of amino acids, a protein, by the ribosome. The primary structure, i.e. the sequence of a protein, then folds into a unique three-dimensional structure which determines the functional features of the protein like substrate specificity, stability or binding partners.

It has been shown by Anfinsen *"...that the three-dimensional structure of a native protein in its normal physiological milieu (...) is determined by the totality of inter atomic interactions and hence by the amino acid sequence..."* (Christian B. Anfinsen, Nobel Lecture, 11.12.1972). This finding remains one of the central paradigms in biology and bioinformatics and has inspired researchers in the last 35 years to study the principles that govern protein folding with the ultimate goal to fully predict the three-dimensional structure of a protein from the sequence. Indeed, there has been significant progress in the field for example in predicting secondary structure elements from the sequence [18] or complete protein structures using structural templates by comparative modeling techniques [102] also driven by community-wide competitions like the CASP experiments [108]. But nevertheless a reliable, ab initio prediction of the structure of a protein from the sequence is still not possible for most protein sequences.

Since the principles of protein structures and protein structure prediction approaches have been described and reviewed in many books and publications we would like to refer the reader to those articles for more details on proteins and their structures. An excellent introduction to protein structures, experimental protein structure determination by X-ray crystallography and NMR as well as state of the art protein structure prediction methods can be found in [26].

2.1 Goals and open questions in protein structure analysis

To date, about 50000 experimentally determined protein structures are deposited in the PDB [13] and the analysis of their similarities and differences provides essential insights into protein structure and protein sequence evolution. Similar to the content of protein sequence databases with the arrival of genome sequencing projects, the number of entries in the PDB has grown very fast in the recent years, driven by improvements in experimental methods as well as large-scale structural genomics projects [57]. The large number of protein structures being available today has led to novel, intriguing questions and problems that need to be addressed by bioinformatics methods.

How many distinct protein folds exist, how are they related and further, how have they evolved? Can computational methods detect similarities between protein structures with a quality comparable to human experts? What are common features of protein structure families, what are their similarities and differences on the sequence and the structure level? And finally, can we predict a protein's function and other functional features like interaction partners or the subcellular location from the structure?

2.2 Existing approaches to protein structure comparison

Soon after the first protein structures had been solved, it became clear that structures are much more conserved than protein sequences. While there are millions of sequences, there is only a relatively small number of distinct protein folds. Currently it is estimated that about 1000 topologically different fold classes exist in nature [28], although one has to note that our view on the fold space may be biased by limits and properties of the experimental methods used to solve structures, i.e. leading to our small knowledge on transmembrane proteins which are especially hard to crystallize.

2.2.1 SCOP and CATH - the standard of truth

The quest for structuring the protein fold space has led to the development of two (manually) curated databases, namely SCOP [6] and CATH [58], which both classify protein structures in a hierarchic manner. SCOP sorts protein domains into classes, folds, superfamilies and families while the four major levels of CATH are class, architecture, topology and homologous superfamily.

The SCOP database is mainly based on expert knowledge and on the first level of the hierarchy defines four major classes namely 'all α', 'all β', 'α/β' as well as '$\alpha + \beta$' roughly describing the content of secondary structure elements in the domain. According to the SCOP authors, proteins in a common fold have the same major secondary structures in the same arrangement with the same topological connections. In the same superfamily, proteins have low sequence identities but their structures and, in many cases, functional features suggest that a common evolutionary origin is probable while domains clustered in the same family are likely to have a common evolutionary origin based on sequence similarity or functional evidence.

Compared to SCOP, the building process of CATH contains more automatic steps and less human intervention. Analogous to SCOP, CATH starts at the class level defining three major classes of secondary structure content ('all α', 'all β' and 'α/β'). The second layer, called 'Architecture', clusters domains with common general features with respect to the overall protein-fold shape but does not take connectivity into account. The 'Topology' level is analogous to the

SCOP fold level and groups structures that have a similar number and arrangement of secondary structure elements with the same connectivity. The last (major) level, 'Homologous superfamily', clusters proteins with a high structural similarity and similar functions, which suggest that they may have evolved from a common ancestor.

SCOP and CATH both provide valuable and important gold standard datasets for the evaluation of protein structure prediction and classification methods ([31], [56], [104]) as well as further analysis of protein sequence and structure evolution ([129], [160]). Nevertheless, due to different approaches to classifying proteins, both 'gold standards' may classify the same chain differently. In Chapter 3 we will discuss the results of a detailed analysis of the differences and similarities of the two hierarchies which provides interesting insights into the properties of the two databases.

2.2.2 Automated protein structure alignment and comparison

Due to the arrival of structural genomics projects and faster structure determination pipelines the number of structures deposited in the PDB is growing fast. This leads to the fact that for example the most current version of the SCOP database (1.73, November 2007) does not contain classifications for more than 10000 protein structures stored in the PDB. Therefore, fast and accurate automatic structure classification and structure analysis methods that model expert knowledge have become more and more important in recent years. A crucial step in this task is the computation of a structural alignment.

The problem of computing structure alignments has been heavily investigated in computational biology for more than two decades and numerous approaches have been developed to address different aspects of the problem. The methods proposed differ in the representation of protein structures in the algorithm, the procedure to identify (structurally) equivalent residues, the approach to combine locally similar regions to an alignment, as well as in the treatment of protein structures, either rigid or flexible.

Many methods measure structural similarity based on a rigid body superposition of the input structures (given the alignment) and for that purpose several scores and optimization algorithms have been proposed. Among them are CE [143], optimizing the RMSD of two input structures, and TM-Align [187] optimizing the TM-Score [188].

Though being a valid approach for many applications, rigid superposition neglects the fact that proteins are dynamic and flexible objects that allow for (significant) structural variation even within families and much more so within superfamilies and folds. Such flexibilities may also be part of the unique function of a protein. Therefore, in recent years, also methods for flexible protein structure alignment have been published that account for large scale structural flexibilities and can align protein structures even in cases that show large conformational changes like movements around hinges. A well known method in this category is FATCAT [181].

While many approaches explicitly use the coordinates of the protein structure in the scoring and optimization process, other methods work on different representations of the input proteins like contact matrices [70] or graphs [66]. Those methods also allow for a certain amount of intrinsic flexibility of the protein structures as long as the contact networks and patterns are conserved.

All methods discussed so far address the problem of pairwise protein structure alignment. Today, also due to the growing number of protein structures in the PDB, multiple structure alignment methods are becoming more and more important for example in order to identify common structural cores of protein families. Methods for multiple structure alignment are e.g. MultiProt [142] and its extension STACCATO [141] or POSA [182].

The focus of the different methods varies greatly depending on the application but also due to the fact that it is not clear how exactly an ”optimal” solution to the structure comparison problem may be defined [182]. For example, for rigid body alignment algorithms, there is always a tradeoff between maximizing the number of residues that can be aligned below a distance threshold (so called number of equivalent residues, N_e) and minimizing the overall root mean square deviation ($RMSD$) of the superposition of the aligned parts of the structures. For alignment methods treating proteins as flexible objects, there must be a tradeoff between the flexibility allowed and the rigidity required to avoid an overestimation of the structural similarity.

For all the reasons described above protein structure alignment remains an important and active field in bioinformatics research where different methods are being developed for different applications and with different strengths and weaknesses.

In chapters 4 and 5 we will describe and discuss two novel methods, namely PPM and Vorolign, both addressing different aspects of the protein structure alignment and protein structure classification problem which introduce novel ideas to the field.

Chapter 3

A Systematic Comparison of SCOP and CATH

As already discussed, SCOP and CATH provide two independent, important 'standard of truth' datasets for protein structure analysis. Nevertheless, the they use to classify protein structures are not the same which may lead to differing classifications for the same protein.

In the following we will discuss the results of our detailed analysis of the differences and similarities of SCOP and CATH which is joint work of the author with Gergely Csaba [35].

Differences between SCOP and CATH are found with respect to the domain partition of the protein chain (a problem on its own [164]), as well as in the final classification of the protein domain into its corresponding structure family. Some differences and similarities between SCOP and CATH have already been evaluated ([38], [64]) and those analyses allowed for valuable insights into the problems and challenges of classifying protein structures. Nevertheless, since the latest study [38] the number of protein structures available in the PDB has more than doubled. This fact may have also increased the problem classifying all known structures in a consistent manner. Also, in contrast to previous studies, we will focus on the extraction of valuable, novel datasets based on the detailed comparison of the two hierarchies which are an invaluable source for e.g. machine learning methods applied to protein structure classification and prediction.

In more detail, we propose a new approach to compare SCOP and CATH on the different levels of the two hierarchies using a similarity measure for sets of domains. Based on a mapping of individual domains and on the similarity of two sets of domains, we identify for each set from one hierarchy the corresponding set(s) from the other hierarchy. This allows to map sets of domains on different levels of SCOP and CATH and to analyze the differences and similarities of the two hierarchies in detail.

SCOP and CATH are often used as 'standard of truth' datasets and inconsistencies and differences in the hierarchies unavoidably lead to problems in the training phase (since wrong or misleading concepts are learned) as well as in the testing / benchmarking phase. Proteins classified to be different by one hierarchy may indeed be similar according to the other classification leading to an overestimation of the errors made. To overcome those problems, we extract sets of pairs of protein domains from our SCOP-CATH mapping which are consistently classified in both hierarchies. Those pairs represent a novel and comprehensive benchmark (training) set which allows for a more consistent evaluation (and training) of protein structure comparison and protein structure prediction methods.

Finally, we utilize our mapping as orthogonal evidence in order to identify possible connections between different folds in one hierarchy which may be revealed via a connection of the two folds suggested by the respective other hierarchy. Such connections between different folds (which are supposed not to be evolutionary related i.e. due to the SCOP definition) provide interesting starting points to further analyze interfold similarities in the protein sequence-structure space, a problem which will also be addressed in the context of alternative splicing in Chapter 7.

3.1 Methods

In the following, we will handle the two hierarchies (SCOP and CATH) as labeled trees, where the leafs correspond to the domains classified by the corresponding hierarchy. Inner nodes represent sets of protein domains which are clustered together on a specific level of the hierarchy. For SCOP, inner nodes represent classes, folds, superfamiles or families. For CATH, inner nodes correspond to classes, architectures, topologies and homologous superfamilies. We denote the underlying sets of domains of the two hierarchies with D_1, D_2 and the hierarchy trees themselves as H_1, H_2. We further define $H_i = (V_i, E_i)$ where $D_i \subseteq V_i$ are the leafs of the tree. Since domain definitions in different hierarchies also may be different, we have to map the domains defined by SCOP and CATH ($D_1 \leftrightarrow D_2$) in a first step. In a second step we will then define and compute a mapping between inner nodes of the hierarchies.

3.1.1 Mapping of domain assignments

A protein domain is defined as a set of segments within one protein, where a segment is defined as a consecutive part of one chain of the protein. Please note that this definition also allows to define discontinuous domains and domains spanning different chains of a protein. In order to compare different domain assignments of the same protein we have to compare sets of segments. To do so, we use the sets of residue positions $RP(d)$ for all segments of domain d and define the similarity of segments via the intersection of their RP sets. Of course, such a mapping is not necessarily unique, i.e. it is possible that a domain in D_1 maps to more than one domain in D_2 or, more generally, that n domains in D_1 correspond to m domains in D_2. In such cases the definitions of the domains may be very different and we exclude domains from D_1, D_2 if their overlap o (see below) is smaller than a specified threshold. For two domains d_1 and d_2 from D_1 or D_2, respectively, we now define the overlap o of two domains as:

$$\mathrm{o}(d_1, d_2) = \frac{|\{\mathrm{RP}(d1)\} \cap \{\mathrm{RP}(d2)\}|}{|\{\mathrm{RP}(d1)\} \cup \{\mathrm{RP}(d2)\}|}$$

If we use a threshold $T_o > 0.5$ the mapping will of course be unique (but not necessarily complete).

3.1.2 Mapping inner nodes of the hierarchies

While the mapping of domains is more or less trivial (except for cases where the domain definitions differ to a large extent), it is not true for the mapping of inner nodes of the hierarchy. As already mentioned inner nodes represent sets of domains. The image of a set of domains in the one hierarchy is the set of domains in the other hierarchy where T_o exceeds a given threshold, i.e. for $S_1 \subseteq D_1$ (equivalently for $S_2 \subseteq D_2$) the image of S_1 is defined as follows:

$$img(S_1) = \{d_2 \in D_2 | \exists d_1 \in S_1 \text{ with } o(d_1, d_2) > T_o\}$$

Further, we define the sensitivity, specificity and the F-measure of a domain mapping of two sets $S_1 \subseteq img(D_2) \subseteq D_1$ and $S_2 \subseteq img(D_1) \subseteq D_2$ on the restricted hierarchies as:

$$\text{sensitivity}(S_1, S_2) = \frac{|S_1 \cap img(S_2)|}{|S_1|}$$

$$\text{specificity}(S_1, S_2) = \frac{|S_1 \cap img(S_2)|}{|S_1 \cup img(S_2)|}$$

$$\text{F-measure}(S_1, S_2) = \frac{2 * \text{sensitivity}(S_1, S_2) * \text{specificity}(S_1, S_2)}{\text{sensitivity}(S_1, S_2) + \text{specificity}(S_1, S_2)}$$

In order to map sets of domains, we search for all inner nodes S_1 from hierarchy H_1 and S_2 from hierarchy H_2 where F-measure(S_1,S_2) > 0, i.e. there needs to be at least one domain which occurs in both sets. The F-measure is especially useful as it accounts for a tradeoff between sensitivity and specificity which is necessary since, obviously, the most sensitive mapping will be always the root, the most specific one the direct parent nodes of two mappable domains.

Given the F-measure for every two nodes which have at least one mappable domain in common, we identify the nodes in H_2 which match best to a given node n_1 in H_1. From each path from the root to n_2 in H_2, only one node (the best one according to the F-measure) will be used in the mapping. Nevertheless, there may be different paths in H_2 containing nodes mapped to the query node from H_1. In those cases matches from different paths are sorted according to their F-measures.

Based on those definitions we calculate for each non-leaf node n_1 in H_1 a sorted (by their F-measures) set $MS(n_1)$ consisting of non-leaf nodes in H_2 where: $\forall a, b \in MS$ a is not descendant of b and b is not descendant of a.

3.2 Results

3.2.1 Datasets

For our analysis we use the most current version of SCOP (1.73, September 2007) as well as CATH version 3.1.0 (January 2007) which contains a similar number of proteins. The mapping containing the more recent CATH version 3.2.0 can be found on the supplementary website at http://www.bio.ifi.lmu.de/SCOPCath. The website and the benchmark datasets will be updated

	1	2	3	4	5	6
SCOP	49251	17162	1885	435	130	29
CATH	68270	11018	492	3	0	0

Table 3.1: Mapping of the domain definitions of the two hierarchies. An overlap threshold > 0.0 is used, i.e. all domains which share at least one residue are mapped onto each other. The SCOP row shows the number of CATH domains mapped onto a single SCOP domain, while the CATH row describes the number of SCOP domains mapped onto one domain defined in CATH. A single domain in SCOP may be mapped on up to 6 domains in CATH. Overall, about 20000 (out of 70000) SCOP domains map more than one CATH domain while 11500 out of 80000 CATH single domains map to more than one SCOP domain.

regularly when new versions of SCOP and CATH are released. SCOP 1.73 contains 34495 proteins deposited in the PDB (97178 domains) which are classified into 11 classes, 1283 folds, 2034 superfamilies and 3751 families. CATH comprises 30028 PDB proteins which are partitioned into 93885 domains and sorted into 4 classes, 40 architectures, 1084 topologies and 2091 homologous superfamilies. The union set of the proteins in the two classification schemes contains 36970 proteins. 27553 PDB proteins are classified in both hierarchies. Please note that throughout this article we regard the following levels of SCOP and CATH to be 'equivalent': SCOP family / superfamily → CATH homologous superfamily, SCOP fold → CATH topology, SCOP class → CATH class.

3.2.2 Detailed Comparison of SCOP and CATH

In the following we present the results of our analysis of similarities and differences between SCOP and CATH in more detail. We will first discuss the results of mapping the different domain definitions of SCOP and CATH onto each other, showing that there are (surprisingly) large differences between SCOP and CATH with respect to their domain definitions. We will then use the set of mappable domains (for which domain definitions largely agree), restrict the respective hierarchies to those domains and compute the mapping of inner nodes of the two restricted hierarchies. We then analyze this mapping of inner nodes in detail which turns out to be very complex indicating many inconsistencies between SCOP and CATH. The usefulness of the SCOP-CATH mapping is shown then by two applications.

Our analysis depends on whether we map SCOP to CATH or vice versa. For space reasons, we present in the following the mapping of SCOP → CATH. The results for mapping CATH → SCOP are available on http://www.bio.ifi.lmu.de/SCOPCath.

Domain mapping

In order to analyze the different domain definitions in SCOP and CATH, we regard a domain defined in one hierarchy to be fixed and count how often exactly one or more domains from the respective other classification are mapped onto it. A domain is mapped iff the overlap o, as defined in 3.1.1, is greater than 0.0, i.e. we map all domains which have at least one residue in common with the query domain. The results are shown in Table 3.1 and confirm results from previous studies [64] that SCOP tends to define larger domains which may be represented by several, smaller domains in CATH.

For our final mapping of domains we used a much more restrictive overlap threshold $T_o = 0.8$. This implies a bijective mapping of domains onto each other but leaves many domains unmapped. Including protein domains which overlap to only a small extent would lead to additional problems when comparing the two hierarchies, especially since domains are also classified according to their secondary structure elements and content. Therefore, including secondary structure elements in the domain definition of one hierarchy while not including them in the other one is likely to lead to differing classifications.

As shown in the following, differing domain assignments have a large impact on the resulting classification. Out of the 27553 proteins which are classified in both hierarchies, for only 19266 (about 70%) the domain definitions are highly similar ($T_o = 0.8$) leading to 56104 domains in the final dataset (increasing up to 66128 mappable domains with a $T_o = 0.5$). In SCOP, those domains are classified into 11 classes, 754 folds, 1258 superfamilies and 2228 families which means that on the other hand, for 538 folds, 776 superfamilies and 1523 families the domain definitions of SCOP and CATH differ to such a large extent that they can not be meaningfully mapped onto each other. According to CATH, the proteins belong to 4 classes, 38 architectures, 736 topologies and 1462 homologous superfamilies. Two architectures, 348 topologies and 629 superfamilies of CATH remain unmapped.

Those values show a surprisingly large number of domains in either of the two hierarchies which are defined in a very different manner in the respective other classification scheme according to their domain boundaries and result in the fact that for only 70% of the proteins the classifications can be compared. Moreover also only 70% of all SCOP families and CATH superfamilies are retained in the mapping due to largely differing domain assignments.

Mapping of Inner Nodes

Given the set of mappable domains as discussed aobve, we computed the mapping of inner nodes of the two hierarchies as described in the Methods section. The results are shown in Table 3.2. Using the F-measure (as defined in the Methods section) we are able to identify for every inner node of SCOP the corresponding, i.e. best fitting, node in the CATH. In such a mapping one would e.g. expect that SCOP superfamilies (and families) map best to the CATH 'homologous superfamily' level.

Surprisingly, when using a F-measure threshold of 0.0 (we map every query SCOP node onto the CATH node with maximal F-measure), the mapping of inner nodes and, therefore, the partitioning of the fold space according to SCOP and CATH appears to be more complicated than expected and many inconsistencies can actually be observed.

When we require a certain quality for a mapping, i.e. setting the F-measure threshold to 0.8, a large number of inner nodes do not find a partner in the other hierarchy. SCOP and CATH therefore define their sets of domains on every level of the hierarchies and for many cases very differently and a large number of unexpected mappings (all the cases except for the green cells

F>0	Unmapped	C	A	T	H
fold class	0	4	2	1	4
fold	0	0	5	504	236
superfamily	0	0	2	32	1224
family	0	0	1	9	2218
F>0.8	Unmapped	C	A	T	H
fold class	8	2	0	0	1
fold	125	0	4	439	177
superfamily	236	0	1	24	997
family	1055	0	1	6	1166

Table 3.2: Number of inner nodes from a hierarchy level in SCOP mapping best to a node from some level in CATH. Two different F-measure thresholds of 0.0 and 0.8 are shown. For example, 2218 SCOP families map best to a CATH topology node given a threshold of 0.0 dropping down to 1166 nodes for a F-measure threshold of 0.8

in Table 3.2) can be observed. For example 240 (178) homologous superfamilies in CATH can not be mapped to a corresponding SCOP superfamily or family for a F-measure threshold of 0 or 0.8, respectively.

The complete mapping and the observed differences between SCOP and CATH may be interactively and comprehensively explored on http://www.bio.ifi.lmu.de/SCOPCath.

Comparison of domain pairs

	consistent	inconsistent	folds	superfamilies
family	7.970.415	133.335	70	102
superfamily	8.208.965	713.181	121	159
fold	10.879.564	2.389.191	84	500
class	268.747.988	62.849.692	745	1258
other class	962.011.672	249.897.353	745	1258

Table 3.3: Shows the inconsistencies between SCOP and CATH with respect to the levels of the SCOP hierarchy. The second column displays the number of consistent pairs and the third column the number of inconsistent pairs. Columns four and five display the number of distinct folds and superfamilies which account for the inconsistencies observed.

In order to analyze the surprisingly large number of inconsistencies between SCOP and CATH in more detail, we tested all pairs of domains in the set of mappable domains for their consistency in the respective other hierarchy. For example, we test if a pair from the same SCOP superfamily is also classified to be in the same 'homologous superfamily' level in CATH. The results of this pairwise comparison of the two hierarchies are shown in Tables 3.3 and 3.4. This

	outer	class	fold	superfamily	family
outer	79.38%	8.31%	0.99%	0.40%	0.03%
class	18.16%	56.15%	2.55%	1.88%	0.87%
arch	2.42%	24.90%	2.80%	1.27%	0.09%
top	0.04%	10.50%	81.99%	4.44%	0.66%
hom	0.002%	0.14%	11.66%	92.01%	98.34%

Table 3.4: Displays the detailed mappings of domain pairs in percent from SCOP (columns) onto CATH (rows). Columns sum up to 100% and green table cells display consistent mappings. Please note that due to the very large number of pairs, even small percentage values correspond to many examples (see Table 3.3 for details as well as supplementary material).

analysis reveals a very large number of domain pairs which are not classified consistently in the two hierarchies. Even though on the family level, where the evolutionary relationship of the proteins should be clear, 98% of the pairs are consistently defined on the one hand, more than 130000 pairs classified into 70 different folds and 102 superfamilies are not classified in a consistent manner. More than 700000 pairwise errors are observed on the superfamily and more than two million errors on the fold level. Table 3.4 allows for a more detailed analysis of the mapping between the different levels of SCOP and CATH and the errors that occur. For example 0.866% (corresponding to 70.188 pairs) of the domain pairs from the same SCOP family are classified to be in different topologies (of the same CATH class) in CATH.

Fortunately, many errors are contributed by a relatively small number of 'superfolds' (Rossmann Folds, Immunoglobulins and some others). Those fold classes also build clusters of similar folds which are further discussed in section 3.3.2.

Nevertheless, a large number of inconsistencies can not be explained by these well known superfolds. All inconsistent pairs can be explored on http://www.bio.ifi.lmu.de/SCOPCath. An interesting example are d1bbxd_ and d1rhpa_ which are classified to belong to two different classes in SCOP (b.34.13.1 and d.9.1.1, respectively) and are indeed very different on the structure level, but belong to the same homology level according to CATH (2.40.50.40). A second example is the pair d1ku7a_ and d1j9ia_ (classified as a.6.1.5 and a.4.13.2). The two domains are indeed structurally similar (though they have a different number of helices). They are classified as different folds SCOP but belong to the same homology level in CATH (1.10.10.10).

All inconsistencies will lead to problems when benchmarking automatic structure classification methods. Also, they may lead to learning wrong concepts in the training phase of machine learning methods for protein structure classification as decision criteria are only learned with respect to one classification or correct criteria are ignored in the learning phase because of inconsistencies.

Extraction of a novel benchmark set

The pairwise comparison also allows us to extract sets of domain pairs which are consistently defined across the hierarchies and which may be used as novel benchmark sets to train and evaluate structure comparison methods. In particular, we extracted two sets of domain pairs:

- domains which are consistently defined as being similar in both hierarchies (in the following denoted as the SCOP-CATH set) corresponding to the consistent fold, superfamily and family pairs in Table 3.3.
- One set of non-similar, negative domain pairs, i.e. domains which are consistently classified to be in the same class but not in the same fold

Also, to avoid an overrepresentation of very similar domains in the dataset, we clustered the domains according to their sequence similarity. All domains with a pairwise sequence identity of more than 50% were clustered together. For each cluster we retained only one representative domain in the final benchmark set (SCOP-CATH50 set) which can be obtained at http://www.bio.ifi.lmu.de/SCOPCath. We also provide additional data, i.e. the details of the clustering process, which allows users to define their own benchmark sets using different sequence identity cutoffs in case that other sequence identity thresholds are appropriate for the specific application.

Redfern et al. also used a consistent set between SCOP and CATH in benchmarking their CATHEDRAL method [128]. Our approach is designed to contain all pairs of proteins which are consistently defined between the two databases. This is an important feature for benchmarking structure classification methods in very detail on a large set of different fold topologies. In contrast, the Redfern dataset, designed for a different purpose, focuses on consistently defined superfamilies whose members overlap to at least 80%. Extracting protein pairs from these consistent superfamilies would lead to a large number of pairs in the benchmark set (up to 20% of the proteins in a superfamily) which would be actually classified inconsistently between SCOP and CATH.

Our dataset can directly be employed for training and benchmarking novel methods developed in the field on different levels of the hierarchies and therefore different levels of structural similarity. In the following we show that this novel benchmark set allows for a much more consistent evaluation of structure comparison methods which is not biased by inconsistencies in the different gold standards.

3.3 Applications of the SCOP-CATH mapping

3.3.1 Benchmarking Structure-Comparison methods

For benchmarking purposes, and as an examplary structure comparison method, we used the TM-align method which computes a structural alignment optimizing the TM-Score [188]. TM-Score measures the similarity of two structures by an optimized rigid body superposition and a TM-score of above 0.4 has been described to indicate structural similarity ([186], [185]). TM-align has been chosen for this study since the TM-Score has already been used to discriminate between similar and non-similar proteins ([186], [185]) and should therefore allow for a good discrimination of similar and non-similar protein domains. Furthermore, the method is quite

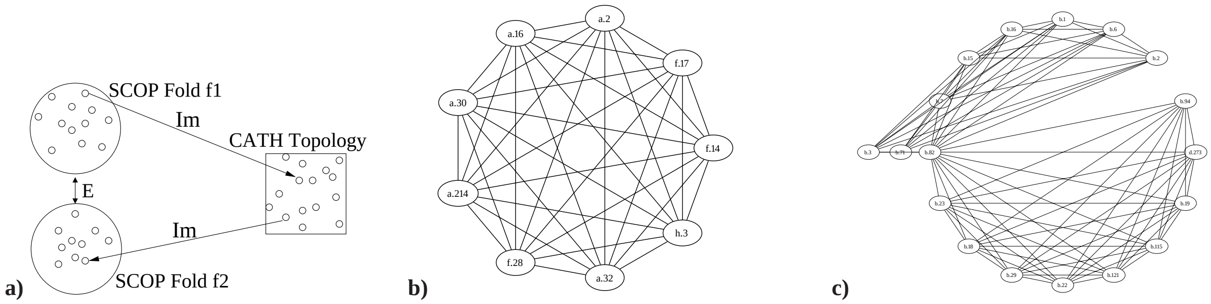


Figure 3.1: **a)** Shows the method of connecting different folds in i.e. SCOP via a link proposed by the mapping of SCOP and CATH. **c)**: Shows the interfold similarity of α-hairpin proteins in SCOP which are clustered in the same fold according to CATH (1.10.287). **c)**: Shows an more complicated fold graph clustering proteins of jelly-roll (2.60.40, Immunoglobulin-Like) and immunoglobulin topologies (2.60.120, Jelly-Roll) in a non-clique subgraph. All fold graphs may be interactively explored on http://www.bio.ifi.lmu.de/SCOPCath. (*Figure taken from [35]*)

fast allowing for the computation of the more than 5.000.000 structural alignments in reasonable time.

For our analysis, we compare the performance of TM-align on the complete benchmark set with the performance on the novel benchmark set proposed in this paper. The only difference between the two sets are the pairs being evaluated. While all pairs which are similar according to SCOP are evaluated in the original setting, our novel benchmark set contains only those pairs which are consistently defined to be similar or different in both SCOP and CATH. Therefore, while the domains contained in the sets are the same, the number of pairs being compared is much smaller in our novel benchmark set than in the original set (16% of the positive pairs have been removed).

In the following we will discuss the plots shown in Figure 3.2 which evaluate the performance of TM-align on the two benchmark sets in detail.

Plots **(a)** and **(b)** show the distribution of TM-Scores of domain pairs within the same class, fold, superfamily or family. The distributions of the scores are very similar between both sets indicating that the main properties of the benchmark sets are similar. There is no apparent bias in the benchmark set proposed here towards domains which are easier to classify and both sets appear to be equally difficult regarding their similarity relationships.

Plots **(c)** and **(d)** in row two as well as **(e)** and **(f)** in row three introduce a novel type of plot to benchmark the performance of structure comparison methods. The plots can be used for any structure comparison method to evaluate in detail the classification performance and in particular the errors made by a method. Especially, they allow to estimate the performance of a method given a template database where members of the family and superfamily are missing and analyze in detail the number of domains for which problems occur in a set of domains and also quantify the dimension of the problem. Plots **(c)** and **(d)** show the number of domains for which we observe problems according to the structural similarity detected by TM-Align. For every query domain, we show how many domains from a different fold have a higher similarity score

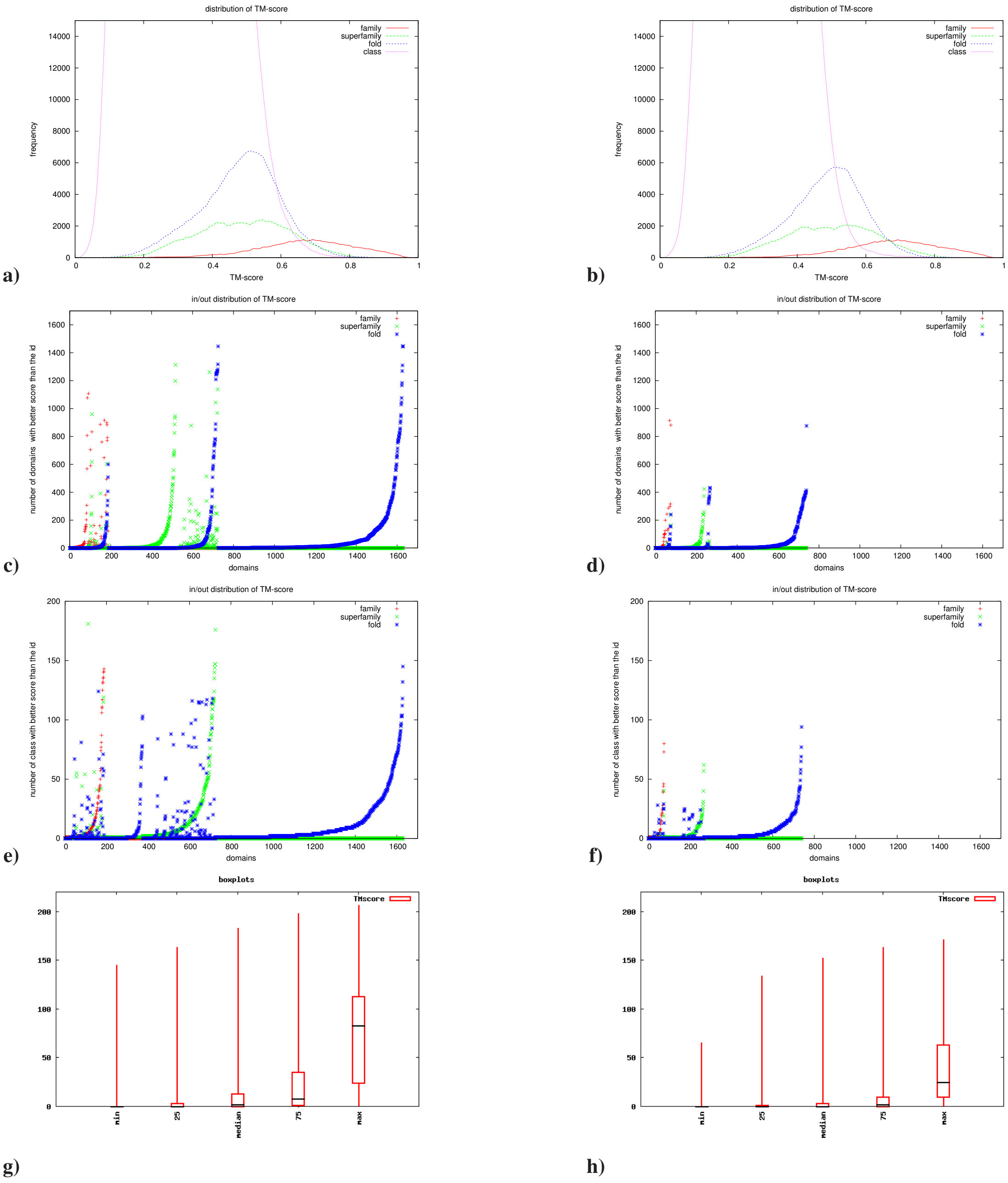


Figure 3.2: In detail evaluation of the performance of TM-align on the complete set of similarity relationships defined by SCOP (left column) and the performance on the novel benchmark set proposed in this study (right column). The plots are further discussed in section 3.3.1. (*Plots taken from [35]*)

than the highest scoring member of the domain's own family (red cross), superfamily (green x) or fold (blue star). On the x-axis we show all query domains for which we observe problems, while on the y-axis, the number of problematic cases for a query (i.e. the number of domains from different a different fold ranked higher than the own family/superfamily/fold) is plotted. For example if there are ten domains from a different fold scoring better than the most similar member from the domain's own family a red cross (at (x,10)) would be plotted. Similarly a blue dot is plotted if wrong proteins score better than a member of the querie's superfamily and a green x is plotted in the case of wrong domains scoring better than the own fold. Also, domains in columns which contain blue dots would not be assigned to their correct folds in the case of missing family and superfamily members since the best hit comes from a different fold.

Panels **(e)**-**(f)** in row three are similar to panels **(c)**-**(d)**, but instead of displaying the number of domains, they show the number of distinct folds (different from the domains own fold) which score better than the respective own family, superfamily or fold.

Comparing the plots that are computed based on the complete set of domain pairs (left column) with the plots computed on the benchmark set of consistent domain pairs (right column) we find that TM-Score / TM-align produces errors for only half of the domains tested and the dimension of the errors (i.e. the number of domains / folds which score better) also strongly decreases.

The only difference between the two sets tested is the removal of pairs which are inconsistently defined between SCOP and CATH. While we remove about 16% of the positive pairs (in SCOP) to obtain the consistent set, the number of errors observed is reduced by 53% (compared to 16% error reduction which would be expected when removing arbitrary pairs). Therefore, the removal of pairs which are inconsistently defined in SCOP and CATH allows to over-proportionally reduce the number of errors. Our conclusions are twofold: many errors reported for protein structure classification methods originate from pairs of domains which are similar to one another, but are classified differently by SCOP or CATH. A different set of errors results from pairs that are e.g. classified in the same family but not similar enough to be distinguished from random pairs by a structure-based comparison method. Using only pairs of domains consistently defined in SCOP and CATH allows to reduce the amount of errors significantly and to separate erroneous behaviour of a method (e.g. errors in the similarity model for protein structures implemented in a method) from problems arrising due to pairs of domains for which even gold standards and experts disagree in their classification.

This figure is completed and summarized by the plots **(g)** and **(h)**. To compute them, we sort the results obtained for every query domain according to their similarity scores. Then, we count for every member of the query fold, how many distinct other folds score better than the respective fold member (please note that every fold is counted only once even if multiple domains from a fold lead to errors). The boxplot in **(g)**-**(h)** shows the errors for five specific fold members: for the best and worst scoring fold members, as well as for the fold member placed at the 25%, median and 75% positions in the sorted list. As, unfortunately, correct fold members score quite differently, this allows to assess the overall performance of fold members by showing how often wrong members score better than the selected five fold member representatives. The boxplot now simply summarizes these numbers for all queries.

Thus the boxplots give a summarized overview of the observed errors. By comparing the two boxplots for the comprehensive and the consensus sets we again find a substantial reduction of errors in the consensus set. While the number of errors for the best scoring fold member is generally small, the errors for the low-scoring fold members quickly increase in both sets, but much more drastically in the comprehensive set as compared to the consensus set. For example, if we look at the fold members scored in the lower quarter (75%) of the fold members, we find 10 different random folds before a correct domain in the original dataset and only one fold in the novel benchmark set. This again indicates, that the number of errors as well as their quantity are significantly lower in the novel benchmark set compared to the original benchmark set.

Overall, the novel benchmark set proposed here is much more consistent than the original pairwise relationships defined by SCOP. It results in a much smaller number of errors (less than half the amount of the errors in the original set). Due to the largely reduced inconsistencies the set should also be well suited for training novel machine learning algorithms for protein structure classification, since it may allow for learning more consistent concepts from the input data.

Furthermore, we expect that the new benchmark set and also the new type of plots allow for a more instructive and objective evaluation of other structure comparison methods as well. The benchmark set has already been applied to measure the performance of PPM, Vorolign and TM-align as discussed in Section 5.2.2.

3.3.2 Inter-Fold Similarities revealed by Consistency Checks

As a second application, we have used our mapping of the two hierarchies in order to identify similarities of different folds / topologies defined in one hierarchy which are implied by mapping them onto the same fold / topology in the respective other classification scheme. Methods to detect possible interfold similarities have already been described for example by Friedberg et al. [52] and the CATH developers [65]. Here, we do not propose a novel approach to analyze such similarities from a structural point of view but utilize the orthogonal criteria and knowledge from two curated classification schemes to identify them. More specific, we search for folds f_1 defined in SCOP which map to a topology level in CATH t while this topology level in CATH also maps to a second fold f_2 in SCOP (see also Figure 3.1a).

The identification of such similarities provides interesting insights into the differences and similarities of fold classifications in SCOP and CATH and further allows to identify interesting links in the fold space. In order to propose a link we currently require the existence of at least five domains, which do not share a sequence identity of more than 50%, to support the link.

This analysis reveals a large number of singletons, i.e. folds / topologies with no link to another fold. 1137 folds in SCOP as well as 904 topologies in CATH turn out to be singletons. For relatively few folds / topologies similarities with other folds are identified which are interesting cases for further analyses in the context of protein structure and sequence evolution.

For SCOP, we identified 29 subgraphs, i.e. groups of folds which are connected via a link in CATH to another fold. 18 of the groups represent graphs of size 2, i.e pairs of folds while the other 11 subgraphs connect up to 39 different folds in SCOP. The largest graph contains a cluster of SCOP folds representing domains which are classified as Rossmann Fold Topology (3.40.50) in CATH but are splitted into 38 different folds in SCOP. Another large cluster comprises β-sandwich proteins with Greek-key topology which represent a cluster of 7 folds.

Two further interesting examples are shown in Figure 3.1. Figure 3.1b shows the interfold similarity of α-hairpin proteins in SCOP which are clustered in the same fold according to CATH (1.10.287). Part 3.1c shows a more complicated fold graph clustering proteins of jelly-roll (2.60.40, Immunoglobulin-Like) and immunoglobulin topologies (2.60.120, Jelly-Roll) in a subgraph which also shows that those graphs do not necessarily form a clique.

3.4 Conclusion

Here, we have carried out a detailed study of the similarities and differences between the two most prominent structure classification databases, SCOP and CATH, which have become gold standards in the field.

We find that there are essential differences between the two classification schemes due to their way of partitioning proteins into domains (which has already been described and discussed by [38], [64] and [69]). SCOP tends to partition domains into fewer but larger components than CATH. In total, only about 70% of the domain definitions for proteins classified in SCOP and CATH agree (overlap of at least 80%) and about one third of the families in SCOP and homologous superfamilies in CATH can not be mapped on domains of the respective other hierarchy.

For the remaining set of about 20000 proteins we then tested how well their classifications agreed with the classification in the respective other hierarchy. For this comparison we have used the F-measure to determine the similarity of two sets of domains on a specific level of two hierarchies. We find that both hierarchies show significant differences and often disagree also in their way to partitioning the protein structure space also in cases of nearly identical domain definitions.

Given those findings and our mapping of SCOP and CATH hierarchy nodes, we extract a novel benchmark set of protein domain pairs which are consistently defined across both hierarchies. This set should have enormous advantages for both, training and testing all kinds of prediction methods, especially machine learning approaches to protein structure classification since more consistent concepts may be learned in the training phase. We show that when benchmarking TM-align on the original dataset (with the similarity relationships defined by SCOP) as well as on the novel benchmark set leads to a largely improved performance. This is due to the fact that errors (proteins which are similar according to one hierarchy but separated into different classes in the other one) which occur due to inconsistencies are removed from the novel benchmark set. Therefore, the benchmark set proposed here provides interesting options to more confidentially measure the performance of protein structure comparison methods as the remaining errors are probably due to the method. We will further use this novel benchmark set to evaluate the performance of Vorolign and PPM in Chapter 5.

Finally, the mapping between SCOP and CATH provides interesting, orthogonal knowledge on the topology of the protein structure space which allows to identify non-trivial links between different folds in e.g. SCOP via their connection observed in CATH. There are some very interesting and large (up to more than 30 folds) sets which may be clustered together in SCOP according to CATH. Among them are some known clusters of folds like the Rossmann Fold Topology. But there are also several other clusters of folds which may be interesting starting points for a further analysis of their sequence-structure properties and may help to further understand the interplay of protein sequence and protein structure evolution, as different structure classifications reflect different viewpoints and criteria on structural and evolutionary similarity.

Chapter 4

Protein Structure Alignment considering Phenotypic Plasticity

We have already discussed the importance of structural alignments for protein structure analysis and comparison in Chapters 2 and 3. With PPM (phenotypic plasticity method) published in Bioinformatics in 2008 [36] we propose a novel method to address this problem. PPM is mainly the work of Gergely Csaba, who developed the method in the course of his Master thesis project in joint work with the author. Here we will only outline the basic ideas of the method (details are provided in [36]). The performance of PPM will be described and discussed in the the context of the Vorolign results in Chapter 5.

Sequence alignments are commonly computed using amino acid exchange matrices and gap penalties. Those parameters explicitly model the process of sequence evolution by mutations as well as insertions and deletions. It is difficult to propose an equivalent model for protein structure evolution which could be used for structure alignment. Therefore, to date, all structure alignment methods (discussed in Chapter 2) capture protein structure similarity not by the evolutionary cost of mutating (”morphing”) one sequence and structural conformation into the other but by a similarity of the three-dimensional objects.

The PPM method is based on the observation that, despite conformational changes and large scale flexibilities, protein structure families exhibit a high level of flexibility on a smaller scale a phenomenon we call *phenotypic plasticity* (see Figure 4.1 for an example). Phenotypic plasticity of a protein structure comprises the changes in the actual 3D-structure, which are to be expected for the given protein sequence (the ”genotype”) or within a given genotype population (i.e. a group of proteins related according to a family, superfamily or fold relationship). Further, phenotypic plasticity is grounded in evolutionary events that occur on the sequence level, namely mutations, insertions and deletions.

In contrast to existing approaches, PPM explicitly tries to model, score and optimize a pairwise structure alignment taking the changes which naturally occur in protein structure evolution into account. Under the hypothesis that two protein structures are similar, we assume that they

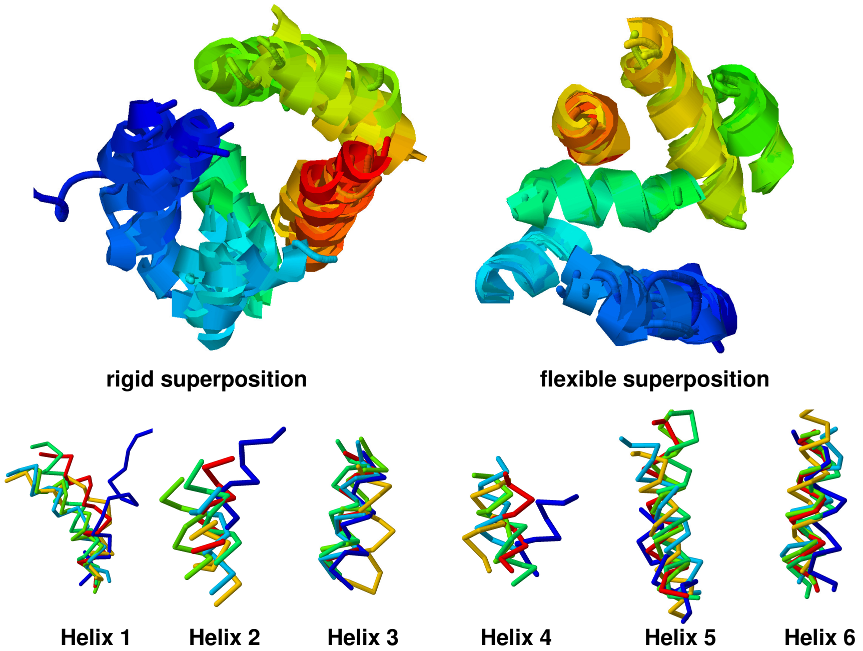


Figure 4.1: On the top left figure we show the overall multiple rigid superposition of members of the pheromone binding family (SCOP: a.39.2.1) implied by a manually curated alignment (guided by disulfide bridges and secondary structure elements). While the overall topology remains the same for all members of the family the exact position and orientation of the helices shows high plasticity (top left). As can be seen from a flexible multiple superposition (top-right) the helices are nevertheless very well conserved. Structural variations in the different helices are shown for each of the six helices separately in the lower part of the figure. (*Figure taken from [36]*)

share locally similar substructures (not necessarily restricted to secondary structure elements). For those, single amino acid exchanges (if they occurred) have not led to major structure rearrangements. Nevertheless, some point mutations as well as most insertions and deletions require certain topological rearrangements (movement of elements) on the structure level. In PPM we now try to measure the evolutionary, structural cost of mutating (or ”morphing”) one structure into the other one. During this process, we score the similarity of locally similar substructures as well as the conservation of their topological arrangement. This approach explicitly allows to map corresponding structural elements of the two structures onto each other and to observe their phenotypic plasticity in a population. This feature of PPM will be especially helpful to study the interplay of sequence and structure evolution as well as to define reliable core elements of protein structure families for further applications e.g. in protein structure prediction.

4.1 Outline of the method

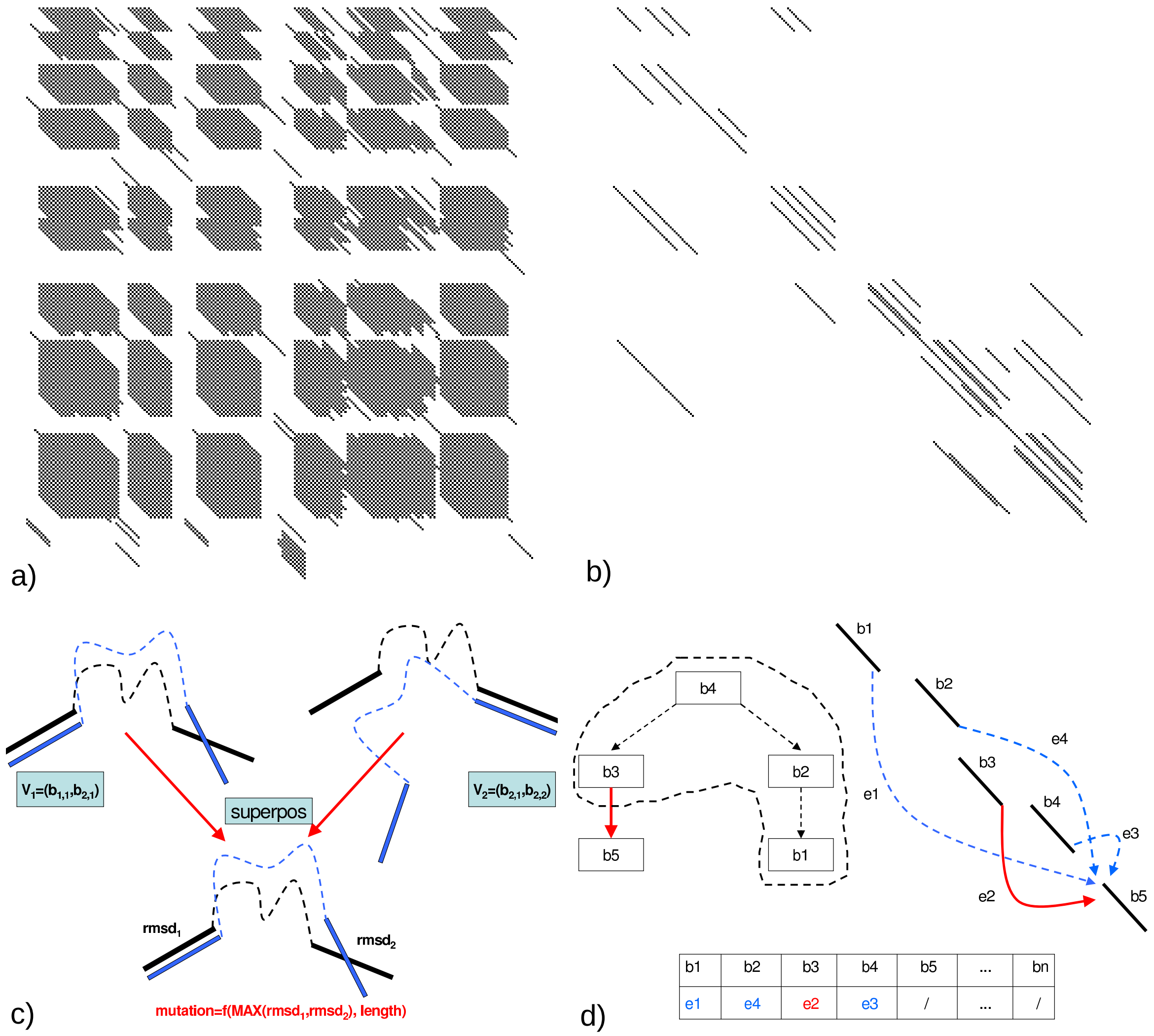

a)
b)
c)
d)
V$_1$=(b$_{1,1}$,b$_{2,1}$)
superpos
V$_2$=(b$_{2,1}$,b$_{2,2}$)
rmsd$_1$
rmsd$_2$
mutation=f(MAX(rmsd$_1$,rmsd$_2$), length)
b4
b3
b2
b5
b1
e1
e2
e3
e4
b1 b2 b3 b4 b5 ... bn
e1 e4 e2 e3 / ... /

Figure 4.2: Shows an overview on the single steps of PPM. **a)** In a first step, mappings of locally similar, ungapped blocks are identified in the two structures. **b)** The number of possible core block mappings is reduced avoiding shift errors. **c)** Visualizes the computation of the similarity of pairs of core block mappings. **d)** Shows the computation of the final alignment. The cost of adding block b_5 to the alignment consisting of blocks $b_1, ..., b_4$. The edges $e_1, ..., e_4$ are sorted with respect to their weights, i.e. e_1 represents the edge with the smallest, e_4 the edge with the highest cost. Depending on which edge is chosen in each step, adding a block becomes more or less expensive, allowing to adjust the degree of phenotypic plasticity in the alignment. We used level 3 throughout our experiments. (*Figure taken from [36]*)

In the following we will outline the PPM model of protein structure similarity based on phenotypic plasticity. The basic ideas of the method are only briefly described and details can be found in the corresponding publication [36]. The method consists of four major steps which are shown in Figure 4.2.

In the first step, we identify locally highly similar substructures, in the following called *core blocks*, of two protein structures being compared. Those substructures are required to be ungapped and need to have a minimal length of 6 residues. The two parts must further be rigidly superposable such that all pairwise residue distances are below a length dependent distance threshold, i.e. longer fragments are allowed to show a larger structural variance in their pairwise distances than shorter fragments. A pair of core blocks mapped onto each other is called a *core block mapping*. To reduce the complexity of the subsequent steps and to avoid shift errors in the core block mappings, especially in helices, we filter the mapping candidates in a second step using a global, superposition-based criterion.

Having identified possible mappings of core blocks we compute the topological similarity of two pairs of core block mappings as shown in Figure 4.2c by superposing the two pairs with respect to either of the core block mappings. Based on the maximal RMSD in this superposition of the respective other core block mapping we compute the topological similarity of the pair depending on the lengths of the blocks as well as their maximal RMSD.

In order to measure the similarity of the overall protein structure topologies we define a graph on the core block mappings. Each core block mapping represents one node in the graph. The graph is fully connected and called the *PPM graph*. The edges are weighted by the similarities of the pairs as determined in step three. An alignment of the two structures in the PPM graph is a subset of nodes which form a spanning tree in the graph. In order to identify the best alignment of the two structures, we have to define the score of such an alignment in the graph. We could, in principle, always connect a node to the alignment by the edge and the corresponding weight connecting it with the sequentially previous node. The problem of this procedure is that the overall topology of the protein is not taken into account since only the topological fit with the previous node would be scored in the alignment and the similarity of topologically different protein pairs could be overestimated. To overcome this problem we sort all edges connecting a node to the alignment by their weights and use the k-th edge to score the topological fit of the block to the alignment as shown in Figure 4.2d. Higher values of k enforce a larger rigidity in the alignment since changes in the topology are penalized stronger, while setting k to 1 would imply the largest flexibility and phenotypic plasticity in the alignment process. In our experiments k has been set to 3 which seems to represent a reasonable tradeoff between flexibility and rigidity.

In the final step the optimal path, i.e. alignment, through the graph is identified using the A* algorithm. If the scoring function being optimized can be shown to be admissible, i.e. always returns a upper bound of the true distance to the target, A* will identify the optimal solution in reasonable time. PPM needs on average one minute to compute an alignment.

4.2 Conclusion

With PPM we presented a novel idea and a new method to address the protein structure alignment and protein structure classification problem in the presence of phenotypic plasticity. In contrast to existing approaches PPM explicitly tries to model the evolutionary cost of the mutations, insertions and deletions that occurred on the structure level during the transformation of one structure to another as implied by changes on the sequence level. Thereby, it accounts for the natural variance observed in protein structure families, a phenomenon we call phenotypic plasticity. This plasticity can lead to problems and artifacts in the alignment process and those are effectively avoided by PPM since it only aligns highly similar local substructures in a first step and then optimizes the alignment of those core blocks using the A* algorithm. The process takes the overall topology of the structures into account and avoids overestimating the structural similarity due to a too large level of flexibility.

The performance of PPM is shown to be comparable or superior to other protein structure alignment methods as described in Chapter 5 even with the current, still rather heuristic, parameterization.

Future development of PPM will be into different directions. First, the method is completely modular according to which scoring functions are used to score the similarity of core blocks as well as their topologies and our final goal of finding an evolutionary measure for structural mutations similar to the idea behind accepted mutations like in the PAM [39] matrix is not reached yet. So far, the scoring functions are purely structure-based and due to our current lack of knowledge on the true costs of mutations and insertions/deletions on the structure level can only be optimized and parameterized empirically. It is an ongoing and iterative process to further refine those parameters to capture the real costs of mutations and insertions/deletions on the structure level, maybe even in a family or fold specific manner.

Second, we would like to extend the currently purely structure-based scoring functions used in the method to a combination of sequence and structure scores to account for both criteria and further improve on the quality of the protein alignments. High quality alignments with respect to sequence and structure form the basis for a detailed analysis of the interplay of protein sequence and structure evolution.

Third, conserved pairwise core blocks identified by PPM will further be used to identify conserved cores of protein structure families with interesting applications in protein structure prediction and sequence-structure alignment (e.g. threading). For this task, PPM needs to be extended in order to compute multiple structure alignments or to combine sets of pairwise alignments to a multiple alignment.

Finally, the model of structure evolution underlying PPM may be an interesting starting point towards a novel definition of protein structure similarity and protein structure flexibility within and between groups of structures.

Chapter 5

Vorolign - fast protein structure alignment using Voronoi contacts

With Vorolign, a joint work of the author with Jan E. Gewehr and Gergely Csaba, published in 2007 [15] we propose a fast method to flexibly align two or more protein structures.

The method is based on the assumption that the environments of two structurally equivalent residues are similar due to positive selection in order to ensure the structural integrity of the protein. We measure the similarity of the structural environments of two residues by their evolutionary relationship with respect to amino acid and secondary structure exchange scores and use this similarity function to align two protein structures using dynamic programming [111]. In contrast to other protein structure alignment methods that use mostly geometry-based similarity measurements we take the protein structure only implicitly into account, via the network of neighboring residues (3D contacts), allowing a certain degree of flexibility of the protein structures being compared.

Additionally, the sequence-based similarity scoring function reliably avoids common artifacts known from structural alignments where evolutionary equivalent residues are not aligned due to divergence of the structures (for an example see Figure 5.1).

To represent the biochemical environment of a residue, we use its nearest-neighbors residues as defined by the Voronoi tessellation [167] of the protein structure. Voronoi tessellation has been found to be useful for several tasks in structural bioinformatics including packing analysis [131], protein folding [53] and structure comparison ([75], [132]).

In our case, the major benefit of using the Voronoi tessellation, in contrast to distance-based contact definitions, is two-fold. First, we obtain a well defined set of nearest-neighbor contacts of a residue containing those amino acids that share a common face in the Voronoi diagram with the residue under consideration. Second, the contact set implicitly takes the geometry of the residue environment into account, since residues do not only have to be close in space but must also be direct neighbors without any other residues blocking the contact between them.

The performance of Vorolign is first evaluated for various applications of protein structure comparison including pairwise and multiple structure alignment as well as the automated assignment of a protein to its corresponding protein family. For the task of automatic family assignment, we will describe the results obtained on a set of almost 1000 difficult examples, derived

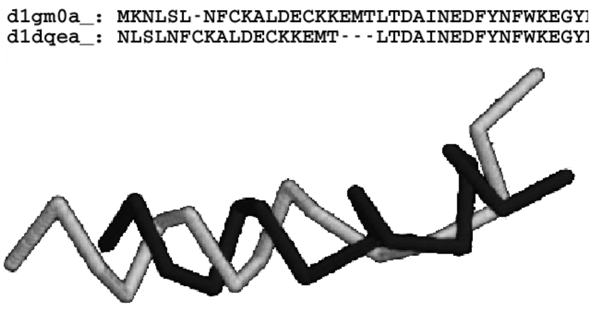

d1gm0a_: MKNLSL-NFCKALDECKKEMTLTDAINEDFYNFWKEGY
d1dqea_: NLSLNFCKALDECKKEMT---LTDAINEDFYNFWKEGY

Figure 5.1: Shows part of the superimposition induced by the pairwise structural alignment of d1gm0a_ and d1dqea_ from SCOP class a.39.2.1 as aligned by CE. Though having almost identical sequences, identical residues are de-aligned due to structural divergence. (*Figure taken from [15]*)

from the difference set between the ASTRAL [27] versions 1.67 and 1.65. Since the original publication of Vorolign in 2007 several additional tests have been performed on the much more comprehensive benchmark set of proteins consistently defined in SCOP and CATH and described in Chapter 3. We also re-evaluated the ability of Vorolign to compute accurate multiple structure alignments according to different criteria on a larger testset compared to the original publication.

Furthermore, we will describe two additional applications of the Vorolign method. The AutoPSI database discussed in section 5.3 provides a large-scale application of Vorolign to all known protein structures which are available in the PDB but not yet classified in SCOP. Moreover, we use the Vorolign idea and scoring function to identify conserved patterns of functional residues in protein families as described in section 5.4.

5.1 Methods

5.1.1 Voronoi tessellation of protein structures

Several methods to compute a Voronoi tessellation of a protein structure have been proposed, and are reviewed in [124], which are differently well suited for different applications. For Vorolign, the tessellation is only used to define contacts in protein structures and features like the exact volume of a cell depending of the atom type or other biochemically important features of the cells do not necessarily need to be as exact as possible. Therefore, the method used to define Voronoi contacts in Vorolign is as straightforward as possible and is exemplarily shown in 2D in Figure 5.2.

We use the C_β atoms (for Glycin the C_α atom) of the protein as the input set of points in the Euclidean space for the computation of the Voronoi decomposition. This decomposition partitions the space into convex polyhedra, called Voronoi cells. Each residue cell contains by definition all points that are closer to the corresponding C_β atom than to all other input nodes. All polyhedra, which are direct neighbors in space share a common face in the Voronoi decomposition corresponding to a contact of the two input points and, consequently, the two residues.

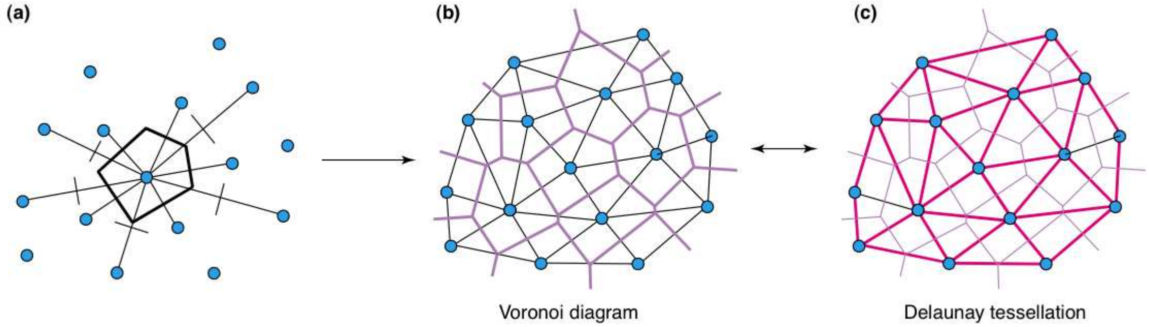


Figure 5.2: Construction of a two dimensional Voronoi tessellation (*Figure is taken from [124]*). **(a)** the Voronoi cell is built by constructing the planes bisecting the lines drawn from the centroid (in our case the amino acid C_β atom) to each of the other centroids in the set and selecting the innermost polyhedron formed by these planes **(b)** when repeated for all centroids in the set, the Voronoi tessellation is obtained. Single cells correspond to the areas surrounded by the thick purple lines. **(c)** The dual of the Voronoi tessellation (called the Delaunay tessellation or Delaunay graph) of the set of centroids (the thick pink lines) is obtained by drawing lines between all cell centroids which share a common face in the Voronoi tessellation.

The Voronoi decomposition of a set of input points can be computed efficiently via standard computational geometry algorithms [118]. We use the quickhull algorithm as implemented in the QHULL program [10] to obtain the Voronoi tessellation of a given protein structure.

There is a problem, though, with a straightforward application of the standard procedures to obtain the Voronoi decomposition of proteins. Residues on the surface of the protein will get unbounded, infinite Voronoi polyhedra and, which is even more unfavorable in our case, can share common faces with residues that are very distant in space. This effect can also be observed if the protein structure contains a larger cleft. Such residue pairs cannot be regarded as being nearest-neighbors in a native, aqueous environment and would lead to artifacts later on when scoring the evolutionary relationship of residue environments. In order to deal with this problem, we introduce explicit solvent atoms surrounding the protein using a method discussed by Zimmer et al. [189], which has been shown to be a fast and accurate method to approximate the water shell placed around the protein molecule. This method places solvent atoms in a regular three-dimensional grid with a distance between the grid nodes of d_{ll} Å within and around the protein structure. Then all solvent atoms that are closer than d_{pl} Å to any protein atom, are removed from the grid. We also remove all solvent molecules from the grid that are more than D_{pl} away from any protein site, with $D_{pl} = \sqrt{3 * d_{pl}^2}$. Following Zimmer et al. we set $d_{ll} = 3$ and $d_{pl} = 4$.

5.1.2 Properties of Voronoi cells

Our structural alignment routine employs the dynamic programming algorithm usually used to compare protein sequences [111] to align two protein structures. Standard sequence alignment routines use an amino acid exchange scoring matrix like PAM [39] to determine the similarity of two residues. In contrast, we calculate the evolutionary similarity of two residues based on features of their corresponding Voronoi cells.

A priori, one could think of several, possibly conserved, properties that could be used to calculate the similarity of two cells. Examples are geometrical features like volume, shape or the surface area, biochemical properties of the faces or the nearest-neighbors of a cell, i.e. residues whose cells share a common face with the cell under consideration.

In this study we compute the evolutionary conservation of two Voronoi cells, i.e. their similarity, by using their nearest-neighbor environments and therefore the conservation of Voronoi contacts. Those features implicitly take the structure of the protein into account since they are derived from a structure-based process, i.e. the Voronoi tessellation. Nevertheless, they are sufficiently general to allow a certain degree of flexibility in the two protein structures being compared (see also section 5.2.4). This is an important feature to detect similarities across more diverse protein structures and for the comparison of multi-domain protein alignments with domain movements.

5.1.3 Similarity of Voronoi cells

Given two proteins $X = x_1 x_2 ... x_p$ and $Y = y_1 y_2 ... y_q$ for which we want to calculate a structural alignment. The set of nearest-neighbors of a residue x_i, as defined by the Voronoi tessellation, is denoted as $N(x_i) = \{x_{i_1}, x_{i_2}, ..., x_{i_n}\}$. In order to measure the similarity of two residues x_i and y_j we will calculate the similarity of their corresponding nearest-neighbor sets $N(x_i)$ and $N(y_j)$.

Similarity of two nearest-neighbors As a first step, we need to define the similarity of two residues x_{i_k} and y_{j_l} in the nearest-neighbor sets $N(x_i)$ and $N(y_i)$. This similarity will be scored by the weighted sum of two scores as given in the equation below:

$$Sim(x_{i_k}, y_{j_l}) = \omega_1 * AA(x_{i_k}, y_{j_l}) + \omega_2 * SSE(x_{i_k}, y_{j_l}). \tag{5.1}$$

In the equation, $AA(x_{i_k}, y_{j_l})$ corresponds to an amino acid similarity score and $SSE(x_{i_k}, y_{j_l})$ scores the similarity of the corresponding secondary structure elements as defined by a secondary structure scoring matrix. The incorporation of the SSE-term avoids to bias the alignment towards equivalencing residues with conserved sequence environments. Preliminary tests showed that the combination of both scores leads to better results than achieved by any of the scores alone.

We also tested the incorporation of other features like the Euclidean or sequential distance of the two residues to their central residue, the Euclidean distance of x_{i_k} and y_{j_l} with respect to a reference frame [121] as well as PSI-BLAST profiles [3], into the similarity scoring function. This did not lead to a significant change in the alignment accuracy indicating that our two features are sufficient to characterize the evolutionary relationship of the nearest-neighbor environments for our needs.

Similarity of nearest-neighbor sets

Having defined the similarity of two nearest-neighbor residues, we estimate the similarity of the two nearest-neighbor sets $N(x_i)$ and $N(y_i)$. For this we need a matching of the residues in the $N(x_i)$ and $N(y_j)$ sets. Once we have found such a correspondence, the final score (similarity) of the two nearest-neighbor environments is simply given by the sum of the similarities of the residues matched onto each other minus a penalty score for each unmatched residue.

In principle such a matching of the two nearest-neighbor sets could be calculated by different methods to solve matching problems in bipartite graphs like the Hungarian algorithm [90]. But since the possible matchings of the neighbors onto each other may be constrained by their order in the protein sequence, the optimal solution can efficiently be computed using dynamic programming. The position of x_i (or y_j respectively) is taken into account by using two independent sets of nearest-neighbors, one containing all neighbors that are found left of the residue under consideration (x_i or y_i) in the protein sequence, the other set containing all residues found right of the residue in the sequence.

The similarity of two nearest-neighbor sets can be computed using dynamic programming with respect to the similarity function of two nearest-neighbor residues given in equation 5.1 and an additional penalty for unmatched residues p_u. This corresponds to calculating a global alignment with linear gap costs of the residues in the two nearest-neighbor sets. The entries $S(k,l)$ in the dynamic programming matrix S are filled using the following equation:

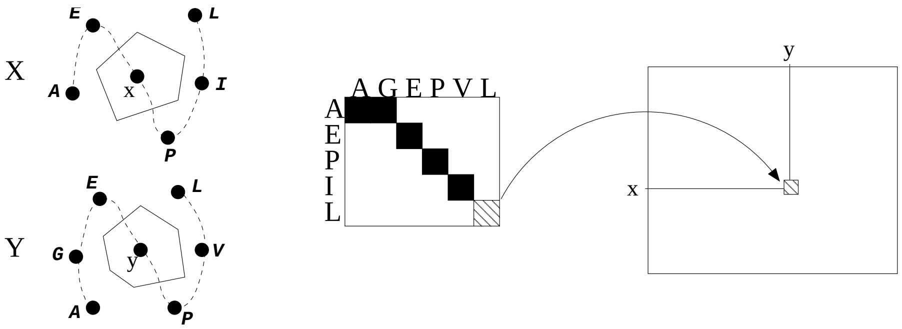
$$S(k,l) = max \begin{cases} S(k-1,l-1) + Sim(x_{i_k}, y_{j_l}) \\ S(k-1,l) - p_u \\ S(k,j-l) - p_u \end{cases} \tag{5.2}$$

The final score of that alignment, i.e. the entry in the last row and column of the matrix S, is used to score the similarity of the two environments of the residues x_i and y_i and is denoted as $Sim(N(x_i), N(y_i))$ in the following.

5.1.4 Pairwise alignment of protein structures

Having defined a function to measure the similarity of two residues x_i and y_j, we can align two protein structures using any dynamic programming algorithm and the similarity function $Sim(N(x_i), N(y_i))$. The complete method is summarized in Figure 5.3. The choice of the algorithm (global, free-shift, local) as well as the gap penalty model (linear or affine) depends on the application and all standard methods are available in the Vorolign program. Other features that influence the alignment quality are the scoring matrices chosen to measure the amino acid and secondary structure element similarity of two nearest-neighbor residues together with their corresponding weights ω_1 and ω_2 (see equation 5.1) as well as gap penalties for the low level (p_u) and the high level (gap open: G_O and gap extend: G_E) dynamic programming steps.

All parameters have been optimized for four different amino acid exchange matrices using a genetic algorithm [43] as discussed below.



Nearest−neighbour set after Voronoi tessellation

Simlarity of neighbour sets (low level matrix)

Similarity of residues x and y (high level matrix)

Figure 5.3: Shows the basic principle of the Vorolign method. First, the neighbor sets of each residue x and y in the two structures X and Y are defined by Voronoi Tessellation. The similarity of two neighbor sets is calculated using dynamic programming (low level matrix) and equation 5.2. The score of the low level matrix is used to fill the high level dynamic programming matrix. Please notice: to keep the figure simple, we did not split up the neighbor sets as discussed in section 5.1.3. (*Figure taken from [15]*)

The idea of using a low level dynamic programming step in order to fill a higher level dynamic programming matrix is also known as 'double dynamic programming' [158] and has already been used in the context of protein structure alignment by Taylor and Orengo ([157], [158]). Their way to calculate the similarity in the low level matrix differs substantially from ours. For a pair of residues (i, j), the two structures are centered and superimposed with respect to the adjacent residues. With respect to that superimposition all pairs of residues are examined and their similarity is measured using mainly geometric features. The score of an alignment of all residues of the two protein structures is used to fill the high level matrix entry for the pair (i, j). In comparison, Vorolign concentrates on the nearest-neighbor residues to compute the similarity score by using amino acid and secondary structure element conservation. This significantly reduces the costs of each similarity computation and also allows Vorolign to calculate flexible structural alignments (see also 5.2.4).

5.1.5 Multiple alignment of protein structures

Recently, the focus of structure comparison has shifted from pairwise to multiple protein structure alignments and several methods have been proposed (see for example [182], [142]). We adopted a standard approach of many multiple sequence alignment methods that first calculate all pairwise alignments of the input sequences, in our case structures, and combine the pairwise alignments following a guide tree to retrieve a multiple sequence alignment. We implemented an approach similar to T-Coffee [113] to generate multiple structure alignments from a set of pairwise structural alignments calculated by Vorolign.

The input of the T-Coffee method is a set of pairwise alignments, where the same protein pair can be included several times in the set to account for solutions generated by different alignment strategies. All residues aligned in that set of alignments form the so called "primary library".

All aligned residue pairs in the primary library are assigned a weight that represents the belief that the alignment of the two residues is correct. In the original publication [113], this weight is set to the sequence identity of the two sequences as given by the alignment. Since we want to measure not sequential but structural similarity, we use a structure based score, namely the TM-Score [188], to measure the quality of a pairwise alignment.

The next step of the algorithm is the ”library extension”. The purpose is to combine information contained in the primary library, such that any pair of residues reflects some of the information contained in the whole library. Therefore, a triplet approach is used. Each aligned pair is tested for consistency via all other alignments in the primary library.

For instance consider the residues x, y and z in the proteins X, Y and Z respectively as well as three pairwise alignments of the three proteins. We find x and y as well as x and z are aligned. Now, if y and z are also aligned in the third alignment there is an alignment of x and y through z. The weight of the pair x and y is increased by the minimum of the weights, i.e. TM-Scores, assigned to the alignment of X and Z or to the alignment of Y and Z. The extended library can now be used as a 'scoring matrix' for a progressive alignment strategy. The multiple alignment is constructed following a guide tree as produced by a neighbor joining strategy [137], with the similarity of two proteins given by their respective TM-Scores.

5.1.6 Fast scan for family members: VorolignScan

Protein structure alignment is often used for the automatic detection of structurally similar proteins in a database, given a structurally resolved, but still unclassified protein. A similar application would be the generation of all-against-all alignments of a set of proteins in the PDB in order to define protein families. For that application, not only the accuracy of the method in detecting proteins that belong to the same family, e.g. a SCOP class, but also the speed of the database scan is important. Since Vorolign is a very fast method for generating structural alignments, we tested the performance of the method in detecting the correct SCOP family for a target protein for several benchmark sets as discussed below.

For all database proteins, the Voronoi tessellation and the corresponding nearest-neighbor sets can be pre-calculated in order to speed up the scan. After the calculation of the Voronoi tessellation for the target protein, the protein can be aligned against the database. To avoid the calculation of Vorolign alignments between protein structures that do not show a significant similarity on the secondary structure level, e.g. an all-β protein against an all-α protein, we employ secondary structure element alignment (SSEA) [103] to pre-filter the template database. In the original setup, we use the 250 highest scoring hits (~5% of the template database) found by SSEA which are subsequently aligned to the target using the Vorolign method.

The results are then ranked with respect to their Vorolign alignment scores which are very useful to sort the results of one target protein against a database of structural templates. In order to obtain a target independent, normalized Vorolign score of two proteins, we divide the Vorolign alignment score by the mean length of two proteins being aligned. This average alignment score of the two Voronoi cell sequences can then be used to define a target-independent quality measure for Vorolign alignments. The distribution of normalized scores is shown in Figure 5.4b and e.g. a normalized score of larger than 15 is very likely to indicate a relationship of the two structures

Matrix	ω_1	ω_2	p_u	G_O	G_E
BC	0.3	2.52	1.23	-16.58	-6.15
P	2.03	2.42	1.78	-13.68	-4.83
T	0.58	0.71	5.04	-22.50	-6.50
N	0.26	0.96	4.45	-22.866	-8.91

Table 5.1: Optimized parameter settings for the three different amino acid exchange matrices as found by the genetic algorithm.

on either the superfamily or family level.

5.1.7 Automatic detection of domains in multi-domain structures

In the original publication, Vorolign has been evaluated on single domain proteins from the ASTRAL compendium. But for various applications like the AutoPSI database [16] or the Vorolign server version available on http://www.bio.ifi.lmu.de/Vorolign it is necessary to partition a query protein into potential domains before searching for similar structures against the database of single-domain templates. To allow for such queries we added an additional step to the Vorolign-Scan method. For every query protein being scanned against the database we search for possible domain partitions of the query on the structure level using the PDP program [2] which has been shown to reliably partition protein structures into domains similar to human experts. All parts of the query structure which represent single domains due to PDP are scanned against the database using the SSEA alignment filter. Additionally, we query the database with the complete structure to avoid a common problem of PDP which tends to partition single domains in SCOP into several smaller domains. In the current setup, the best 250 templates (which may only fit fragments, i.e. correspond to single domains of the query) ranked by the length-normalized SSEA filter scores are aligned with the query structure using Vorolign.

5.1.8 Parameter Optimization

Parameters used throughout the experiments have been optimized on a small set of 20 randomly chosen protein pairs. None of the training protein pairs has been included in the test cases discussed later on. We used a genetic algorithm [43] with the average TM-Score of the structural alignments as fitness function. Proteins were repeatedly aligned by Vorolign using the values of the population member until the genetic algorithm converged. As convergence criterion we used 5 generations without a change of the fittest population member. The optimization has been restarted 50 times.

Parameters were optimized for free-shift alignment with affine gap penalties (G_O and G_E) as well as four different structure-based amino acid exchange matrices: the Blake-Cohen matrix [22] (BC), a matrix by Prlic et al. [125] (P), by Nussinov et al. [8] (N) and the SM_THREADER matrix [44] (T). As secondary structure scoring matrix we used the matrix described by Walqvist et al. [168]. The optimization results are shown in Table 5.1.

5.2 Results

5.2.1 Family recognition on a testset of 979 query proteins

Method	Family	Superfamily	Fold	Wrong
PPM	**88.3%**	**94.5%**	97.5%	2.5%
Vorolign, T	86.4%	92.4%	**97.7%**	**2.3%**
Vorolign, N	83.7%	90.4%	96.1%	3.9%
Vorolign, P	82.9%	88.9%	91.4%	8.6%
Vorolign, BC	81.6%	87.6%	89.9%	10.1%
TM-align	83.8%	92.6%	95.9%	4.1%
CE_1	81.8%	88.4%	90.6%	9.4%
CE_2	84.6%	91.9%	94.1%	5.9%
C_β, T	82.9%	89.5%	94.4%	5.6%
PPA	80.8%	87.5%	91.9%	8.1%
Free-shift alignment	65.8%	68.0%	69.6%	30.4%
SSEA	60.8%	68.9%	75.6%	24.4%
BLAST	48.9%	52.5%	52.8%	47.2%

Table 5.2: Results of the family recognition scan and the corresponding average pairwise alignment quality of the methods. Methods are evaluated taking the best hit with respect to the method specific score into account. For Vorolign, PPM, TM-align, C_β, SSEA, free-shift alignment and PPA those are the respective alignment scores, for BLAST we use the best E-Value and for CE the best z-score. If CE returns more than one alignment for a target-template pair the z-Score of the longest alignment (CE_1) or the best z-Score of a pair (CE_2) is used.

In order to test the ability of Vorolign to detect the correct protein family in a database of representative protein structures, we originally used 979 proteins from the ASTRAL compendium that are contained in version 1.67 (February 2005) but not in the previous version 1.65 (December 2003). From the 10034 proteins contained in the difference set of version 1.67 and 1.65 (excluding genetic domains), we removed all proteins for which the sequence identity using standard sequence alignment methods against the ASTRAL95 (version 1.65) was found to be above 30%, with more than 30 identical residues. The final set of 979 proteins therefore comprises the non-trivial cases and allows us to asses the 'blind test' performance of our and other methods in predicting the SCOP family for a given target.

Those proteins were scanned against the ASTRAL25 database (version 1.65) which includes 4358 proteins using the setup described in section 5.1.6. All four amino acid exchange matrices from section 5.1.8 are evaluated.

The performance of all methods will be measured as the fraction of proteins that can be assigned to their correct family, superfamily or fold with respect to the SCOP classification. For Vorolign, only the best template due to the Vorolign alignment score is taken into account for the evaluation. The results can be found in Table 5.2.

The performance of Vorolign is compared against several other methods. SSEA, free-shift sequence alignment with affine gap costs (Pam250 matrix, $G_O = -12$, $G_E = -1$), BLAST (standard options, against ASTRAL25 database) [3], Profile-Profile alignment (PPA) as used in [166] as well as CE, which is faster than for example Dali and performed best in a recent study of protein fold comparison servers [114]. Additionally to the results in the original Vorolign publication we now evaluated the performance of PPM and TM-align on this set.

All methods (except for BLAST) were given the same 250 best hits from the SSEA pre-filter and again only the best template with respect to the method specific quality score is taken into account.

To show the advantage of using Voronoi instead of C_β distance-based contacts, we test the performance of the method using neighbor-sets defined by C_β contacts (threshold 6.5 Å) together with the SM_THREADER matrix. Parameters have been optimized as discussed in 5.1.8 and are set to $G_O = -20.07, G_E = -8.76, \omega_1 = 0.71, \omega_2 = 1.47, p_u = 4.76$.

When comparing the performance of the four different amino acid exchange matrices in combination with the Vorolign method, we find the different matrices are differently well suited to score the evolutionary similarity of residue environments. The SM_THREADER matrix performs best with respect to all quality measurements applied. This finding is interesting, since this matrix is not computed with respect to evolutionary criteria but expresses the contribution of a residue to the total energy of the protein. This could be an interesting starting point for extending Vorolign in the future, using potentials instead of amino acid exchange scores to score the similarity of two residue environments in combination with the Vorolign method.

A comparison of the results of the SM_THREADER matrix in combination with nearest-neighbor sets defined by Voronoi or C_β contacts finds Voronoi contacts to perform better than C_β-based contacts according to recognition rates as well as alignment quality. This difference cannot be expected to be large since the nearest-neighbor sets will of course overlap, but the advantage of the Voronoi-based contact definition gained in comparison to C_β contacts justifies the small overhead of computing the Voronoi tessellation in order to retrieve the nearest-neighbor sets.

The results also show that standard sequence alignment methods, i.e. BLAST and Smith-Waterman, are not able to make accurate family, superfamily or even fold assignments for our test set and demonstrate that the set contains the more challenging cases. The structural quality of the alignments is also poor. For example, simple sequence-based alignment algorithms reach an average TM-Score of 0.49 on the set.

Methods applied in the field of fold recognition, namely SSEA and PPA, gain higher recognition rates. The results of SSEA justify its application as a pre-filter to speed up computationally more expensive methods like Vorolign, CE and PPA. Profile-Profile alignment performs surprisingly well with respect to the family, superfamily and fold recognition rates. In more than 90% of the cases the method is able to identify the correct fold of the target. Nevertheless, the structural quality of the alignments is not comparable structure-based alignment programs with an average subset size of 66.2% and a TM-Score value of 0.58.

Of course, a fair comparison of the abilities of Vorolign can only be made other structural alignment methods. With respect to the recognition rates the structure-based alignment programs outperform all other methods discussed above.

Interestingly, Vorolign in combination with the T matrix is more accurate in predicting the correct family, superfamily or fold of a target than CE and TM-align making only 2.3% wrong assignments in comparison to 5.9% wrong assignments made by CE and 4.1% wrong assignments made by TM-align. In this updated evaluation, PPM performs even better than Vorolign on the family and superfamily and comparable to Vorolign on the fold level. Both programs are more accurate in automatic family assignment than the other structural alignment methods tested in this scenario.

The very good performance of Vorolign compared to other structure alignment methods and especially TM-Align and CE in recognizing the correct family of a target protein seems to be due to its more sequence-based scoring function. While structurally similar proteins can often not be distinguished on a pure structural basis, the more sequence-based similarity score of Vorolign seems to be an advantage, taking not only structural but also sequential knowledge into account. Also, the scoring function seems to better capture the intention behind the SCOP database taking also sequential and functional criteria into account when classifying protein structures. Nevertheless, the example of PPM shows that with a meaningful similarity measure which allows for a certain amount of flexibility in the protein structures (i.e. phenotypic plasticity) it is also possible to achieve very good and even better results on this set. So flexibility as well as similarities detectable on the sequence but not easily on the structure level seem to account many of the errors made by standard structure alignment methods like CE and TM-Align.

5.2.2 Detailed Evaluation on the SCOP-CATH set

Since the publication of Vorolign in 2007 we have evaluated the performance of Vorolign, PPM and TM-Align (the three top-scoring methods on the Vorolign set) on the larger and also more recent test set which has been introduced in Chapter 3.

To benchmark the methods, every domain is used as query and compared against all domains in the set which belong to the same family, superfamily, fold (positive examples) and the same class but not the same fold (negative examples). Templates from other classes are not taken into account to limit the number of pairs to be evaluated. In total, we compute more than 3.6 million pairwise structure comparisons (the set can be obtained from http://www.bio.ifi.lmu.de/PPM).

When evaluating the same scenario as above, namely the family, superfamily and fold recognition rate of the methods according to the first hit (shown in Table 5.3) the methods turn out to perform very similar. Vorolign slightly outperforms PPM and TM-Align on family and superfamily level while PPM performs better than the other two methods on fold level. The somewhat different results of the two benchmarks shown in Tables 5.2 and 5.3 on family and superfamily as compared to fold level may be due to a smaller sequence identity allowed for pairs in the Vorolign set (<30%) compared to the larger set (< 50%) where the advantages of the more sequence-based Vorolign scoring function become evident.

In contrast to Table 5.3, Table 5.4 evaluates the performance of the methods in a different and more difficult scenario which has not been benchmarked for the small set of 979 domains. Here, the performance is evaluated in a situation where all members of the true family or superfamily of the target have been removed from the benchmark set to test a methods ability to detect sim-

Method	Family	Superfamily	Fold
PPM	78.6% (3817)	91.4% (4442)	**98.0%** (4763)
Vorolign	**79.9%** (3881)	**91.9%** (4464)	97.5% (4739)
TM-align	78.5% (3815)	91.4% (4429)	97.4% (4735)

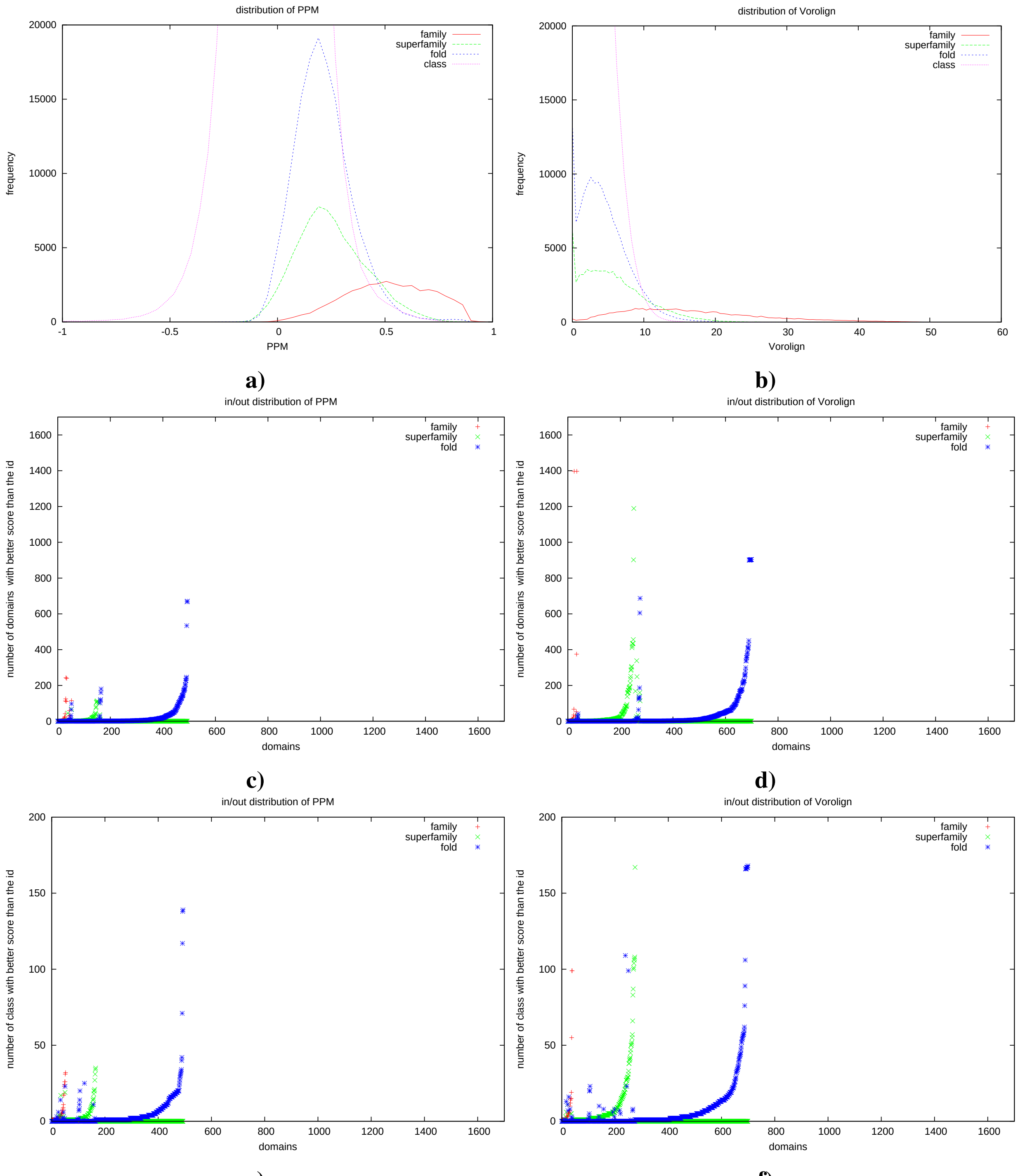
Table 5.3: Evaluation of the family, superfamily and fold recognition rate of PPM, TM-Align and Vorolign on a large set of almost 5000 SCOP domains. The absolute number of correct predictions accounting for the rates are shown in brackets.

Method	Family	Superfamily	Fold
PPM	1.07% (53)	**3.62%** (122)	**13.61%** (357)
Vorolign	**0.83%** (41)	6.29% (212)	18.68% (490)
TM-align	1.56% (77)	6.02% (203)	19.79% (519)

Table 5.4: Evaluation of the errors made by a method in the presence of the correct family (Family column), the presence of the correct superfamily, but not the correct family (Superfamily column) as well as in the absence of the correct family and superfamily but in the presence of the correct fold (Fold column). An error is defined as a domain from a different fold scoring better than the query's own family, superfamily or fold members, respectively.

ilarities in the absence of clear sequence similarity. This setup becomes for example important for folds which contain only one family and need to be extended by a new family resulting in a situation which could be described as ”protein structure threading” or ”fold recognition” (while knowing both structures).

Therefore, all hits from the query domain's own fold are regarded as correct hits while members of the same class but from a different fold are regarded to be errors. The table shows the number of query proteins where templates from wrong folds score better than templates from the own family, superfamily or fold, respectively. The error rates obtained allow for a detailed analysis of the performance of a method in the absence of the true family or superfamily since in those cases wrong folds would be predicted. In this more difficult test scenario (compared to pure family recognition, where only the first hit is taken into account) PPM improves on both, Vorolign and TM-Align, on the superfamily and fold level, clearly reducing the number of errors. On the family level, Vorolign performs better and again, the sequence-based scoring function may account for the advantage. The specific test scenario is also visualized in Figure 5.4 using the plots introduced in Chapter 3, section 3.3.1. First of all plots a) and b) display the distribution of scores obtained by PPM and Vorolign for different levels of similarity, i.e. proteins in the same family (red line), same superfamily (green line), same fold (blue line) and the same class (magenta line). One can see that for example above a threshold of 15, the normalized Vorolign score indicates a superfamily or family relationship of the two proteins with a very high probability. Plots c)-f) show the number of domains (c and d) and distinct folds (e and f), respectively, which would score better than a hit from the same family, superfamily or fold in the absence of the true family or true superfamily. Those cases would therefore lead to very wrong classifications. As already discussed in the context of Table 5.4 in this test scenario PPM performs better than



a) b)

c) d)

e) f)

Figure 5.4: Detailed evaluation of the performance of PPM (left column) and Vorolign (right column) on the novel benchmark set proposed in Chapter 3. The different types of plots have been introduced in section 3.3.1 of Chapter 3 and are discussed in the text.

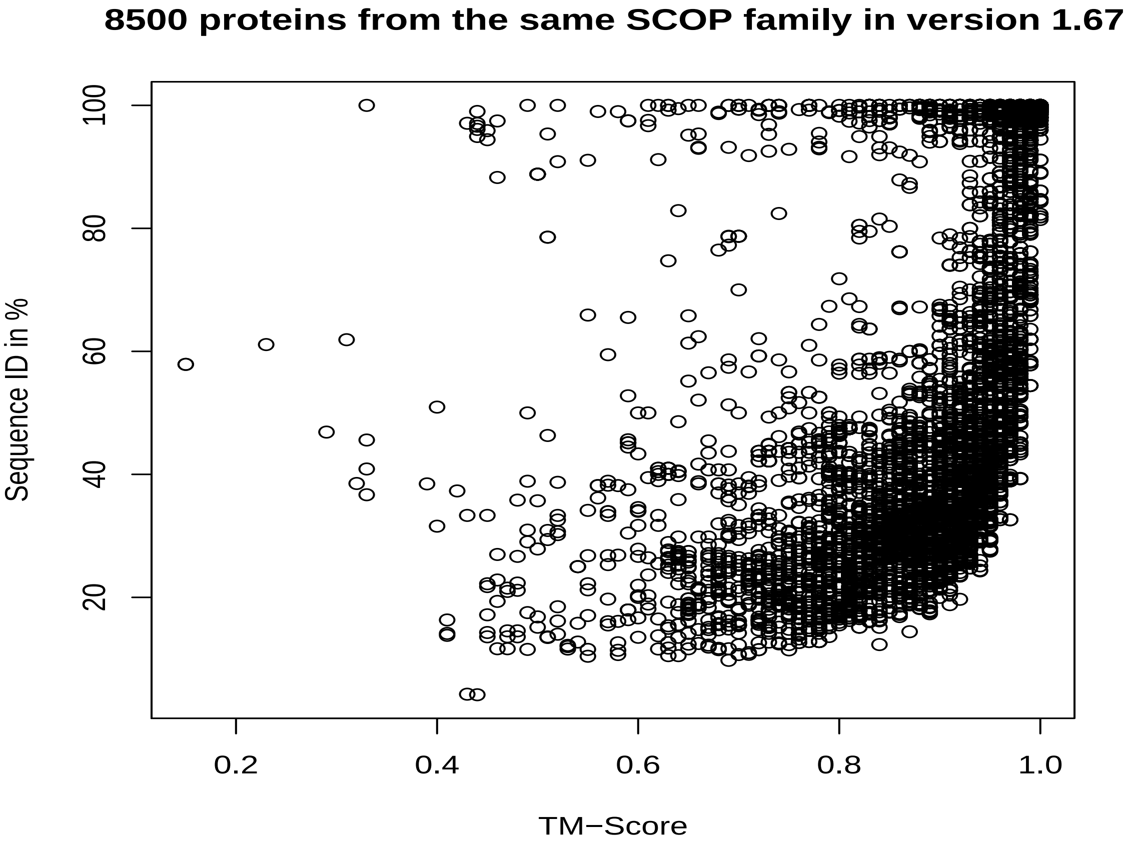


Figure 5.5: Displays the sequence identity of 8500 protein pairs plotted against the structural quality of the Vorolign alignment measured by the TM-Score. Alignments with a high structural quality can be computed across the complete range of sequence identity. Protein pairs with a very high sequence identity and low TM-scores are often due to structural flexibilities whose structural similarity can be detected by Vorolign but is not captured by the TM-score.

Vorolign according to the number of errors made as well as their dimensions, i.e. the number of wrong proteins or folds which score better than the own fold, superfamily or family.

5.2.3 Alignment quality

The structural quality of Vorolign alignments is only slightly worse than the quality of structure based methods. While Vorolign aligns about 76% of the residues with a high structural accuracy and achieves an average TM-score of 0.74, CE equivalently aligns 78% of the residues and reaches an average TM-score of 0.78. This result is not very surprising since Vorolign does not attempt to optimize the superimposition of the two structures at any point in the algorithm.

One could assume that Vorolign mainly detects sequence similarity in the scanning process and that the structural quality of the resulting alignments heavily depends on the sequence similarity of the two input proteins. In order to evaluate this we have scanned the complete set of 10034 proteins in the ASTRAL difference set described above against the template database. This results in more than 8500 proteins for which Vorolign identifies the correct family in the scan. For those protein pairs we have plotted their sequence identity (as implied by the Vorolign alignment) against their pairwise TM-Score shown in Figure 5.5. As it can be seen in the Figure, Vorolign is able to compute high quality structure alignments as measured by the TM-Score across the complete range of sequence identity observed for the family pairs. The majority of the alignments has a TM-Score higher than 0.75 which indicates a high structural similarity implied by the Vorolign alignment. Also, the short-comings of TM-Score as a rigid-body similarity measure may be seen as several protein pairs with very high sequence similarities of e.g. above 90% are structurally non-similar due to structural flexibilities or domain movements.

5.2.4 Specific examples for multiple alignment of protein structures

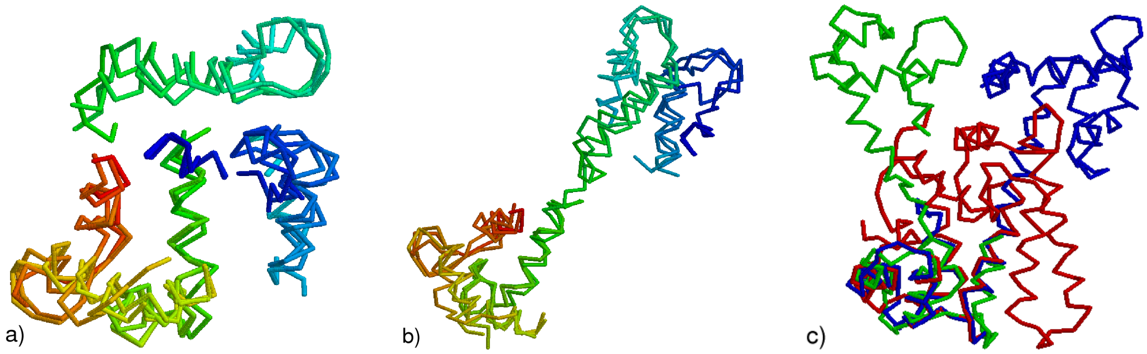


Figure 5.6: Figure **a** shows the flexible superimposition induced by the multiple Vorolign alignment of 1ncx and 1jfjA onto 2sas (closed conformation), while figure **b** shows the flexible superimposition of 1jfjA and 2sas onto 1ncx (fully open conformation). One can see that all four EF-Hands can be superimposed onto each other even in the different conformations. Figure **c** displays the superimposition of 1jfjA, 2sas and 1ncx as calculated by Multiprot. A rigid superimposition is not possible leading to the very small common core found by Multiprot (all figures were generated with Rasmol). (*Figure taken from [15]*)

To evaluate the quality of multiple alignments computed by Vorolign in the original publication we applied the multiple alignment routine described in section 5.1.5 to three structural families from the literature [182] and compared the performance of the method with the performance of POSA and MultiProt. The examples contain the relatively easy case of 15 proteins from the Globin family, 10 proteins of the NAD(P)-binding Rossmann fold, a highly divergent family, as well as three Calmodulin-like proteins that contain conformational flexibilities to show that our method is able to deal with such a case. Alignments from MultiProt are obtained from the standalone program, while POSA alignments are retrieved directly from the webserver. All results are evaluated using the same protocol applied to the Vorolign results in order to guarantee a fair comparison.

Globin family

Globins are an extensively studied example of evolution at the molecular level (see [94] for a recent review) and the family is considered to be a relatively easy case for multiple structure alignment. Vorolign has been used to align a set of 15 globin structures described in Ochagavia et al. [115] and identifies a common core of 72 residues with an average pairwise RMSD of 1.46 Å. In contrast, MultiProt finds a common core, including all 15 input structures, of 75 aligned positions with a pairwise average RMSD of 1.95 Å. POSA identifies 67 positions with an RMSD of 1.12 Å. As expected, different programs lead to different structural alignments of various length and RMSD values. Nevertheless, the superimpositions of all programs are very similar. Therefore, the performance of all three programs tested here can be regarded as being comparable.

NAD(P)-binding Rossmann fold

The NAD(P)-binding Rossmann fold (SCOP fold c.2) consists of a three-layer sandwich structure with a parallel β-sheet of 6 strands packed between two helices and the sheet order 321456. We align one representative of each of the 10 families in SCOP 1.65 (d1heta2, d1ek6a_, d1obfo1, d2naca1, d1kyqa1, d2cmd_1, d1np3a2, d1bgva1, d1id1a_, d1oi7a1). Vorolign identifies a common core of 44 residues with an average RMSD of 1.82 Å, the POSA alignment represents a core of 37 residues with an average RMSD of 1.15 and Multiprot identifies 33 residues with an average RMSD of 1.39. Vorolign finds 3 β-strands (S1, S4 and S5) and 2 helices (H1, H5) to be relatively conserved among all 10 input structures which represent only a small part of the overall structure of the proteins. The example displays the capability of Vorolign to compute good multiple alignments for a structurally highly diverse set of proteins with a quality similar to other structure-based algorithms.

Calmodulin-like proteins

Calmodulin consists of two domains, each domain having a pair of so called EF-hands which are helix-loop-helix Ca^{2+}-binding motives. Ca^{2+} induced domain movements lead to an open or closed conformation of the protein. Different conformations are directly linked to the proteins function. Therefore, proteins from that family are good examples to test the ability of a protein structure alignment method to flexibly align protein chains.

We use three proteins from this family to demonstrate the ability of Vorolign to find good structural alignments even for proteins that show significant conformational changes. Proteins aligned are troponin C from Chicken (1ncx, open conformation), sarcoplasmic calcium-binding protein from *Branchiostoma lanceolatum* (2sas, closed conformation) as well as EHCABP from *Entamoeba histolytica* (1jfjA, partly open conformation).

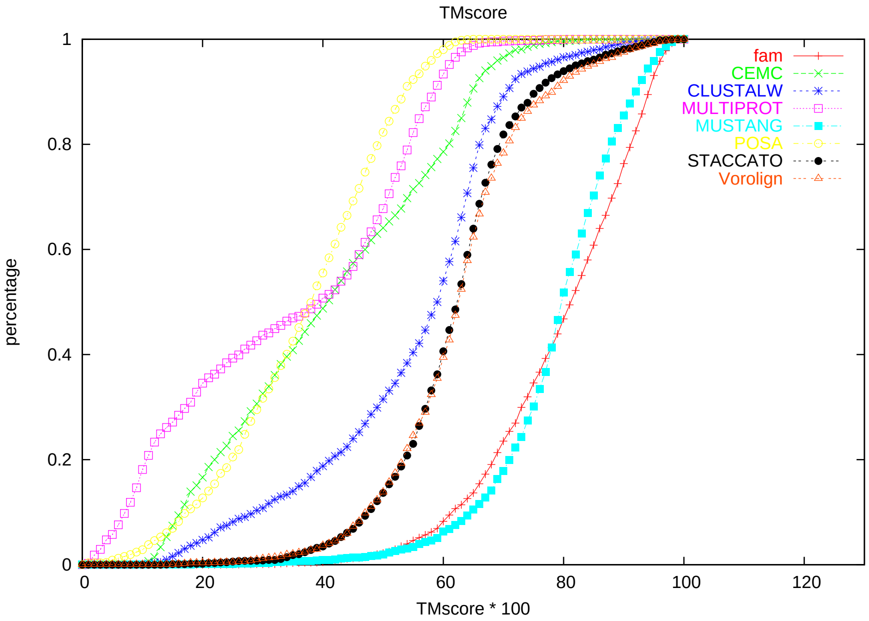


Figure 5.7: Shows the quality of the pairwise alignments induced by their corresponding multiple alignments of the family. The structural quality of the alignments is measured by the TM-Score.

Standard, rigid-body multiple structure alignment routines obtain only small, structurally conserved regions. Multiprot returns an alignment containing 55 positions with an RMSD of 1.58 Å corresponding to one ”hand” of the calmodulin protein. This result is not surprising since a complete, rigid body superimposition of the molecules is not possible (see also Figure 5.6c).

In comparison, the POSA method returns a full length alignment of the three structures containing 131 positions with an RMSD of 2.58 Å in a flexible superimposition. The Vorolign method is able to align 124 positions with an RMSD of 2.77 Å. Although the common core is found to be smaller than the core found by POSA the Vorolign alignment spans the whole length of all three proteins, and a flexible superimposition of the three structures as (see Figures 5.6a and b) induced by the Vorolign alignment shows that the three proteins can indeed be structurally aligned if flexibilities are incorporated. This exemplifies Vorolign’s ability to align protein structures that contain larger conformational changes.

5.2.5 Large scale evaluation of multiple alignment quality

In addition to the examples discussed above and in the original Vorolign publication, Gergely Csaba has evaluated the quality of multiple protein sequence and structure alignments on a large scale as part of his Bachelor thesis which was co-supervised by the author. In order to evaluate the quality of multiple alignments we used two different criteria. First, the structural quality of the pairwise alignments induced by the multiple alignment were evaluated to test the ability of a method to preserve a high pairwise structural alignment quality in the process of computing the multiple alignment. Second, we tested the ability of the programs to correctly align biologically important residues (such as functional sites or sequence patterns) in the multiple alignment. The ability to align functionally relevant residues of a protein family with a high accuracy is especially important for many applications of multiple (structure) alignments like the definition of sequence family signatures [74] or the reliable training of family hidden Markov models [153].

Seven multiple (structure) alignment methods, namely Vorolign, Staccato [141], its predecessor Multiprot [142], Mustang [85], CE-MC [62], POSA [182] and CLUSTALW [159] have been evaluated with respect to those two criteria. The benchmark set consisted of all SCOP families in version 1.67 which contained at least three members after a filtering step removing too similar proteins. In more detail only one representant for every single-linkage cluster of proteins with TM-scores larger than 0.8 or sequence identities of more than 80% were retained in the dataset. The final set contains 861 SCOP families.

Induced pairwise alignment quality

Figure 5.7 shows the fraction of induced pairwise alignments which have a TM-score above x percent in a cumulative way. As shown the structural quality of the pairwise alignments induced by the multiple family alignments varies greatly. Mustang is able to reach almost the maximally possible pairwise alignment quality in terms of the optimal pairwise TM-Score computed by TM-Align. Staccato and Vorolign come second and both perform superior to the other methods. This is an interesting result since both methods are more sequence driven than POSA, CE-MC and Multiprot. Staccato can clearly improve on the Multiprot superpositions it is based on, POSA and CE-MC seem to align too few residues to gain high overall TM-Scores. The structural quality of the aligned parts is nevertheless good.

Conservation of important residues

In order to evaluate the ability of the multiple alignment methods to align functionally important residues we evaluated two different types of feature categories. The plots should be read in the following way: On the X-axis both figures show the percentage of the features in the dataset for which at least a certain percentage of conservation in the multiple alignment is reached (shown on the Y-axis). The conservation is computed as follows. For example if a family has 5 members and member 1 has an annotated class one feature at position p, and 3 other members have the same amino acid aligned to this position. The fourth member has a different amino acid at position p. Then the conservation rate is 3/4=0.75. The ”sequence maximum” and ”TM-superpos maximum” curves represent upper bounds for the maximally possible conservation rate with respect to sequence and structure by either aligning the two sequences optimizing feature conservation or trying to guide the superposition of the two structures via the residues corresponding to the features.

Features of class one are supposed to be highly conserved and important for the function and / or the structural integrity of a family. They are represented by functionally important residues annotated in the Catalytic Site Atlas (CSA) [123] as well as cysteins taking part in cystein bridges

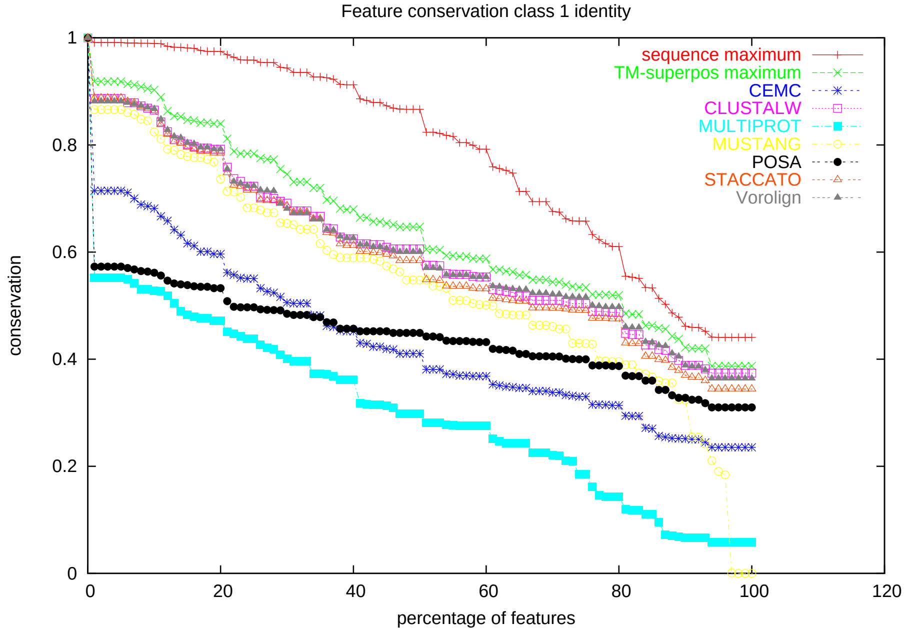


Figure 5.8: This figure shows the conservation of CSA functional residues and residues taking part in cystein bridges in the multiple alignments analyzed.

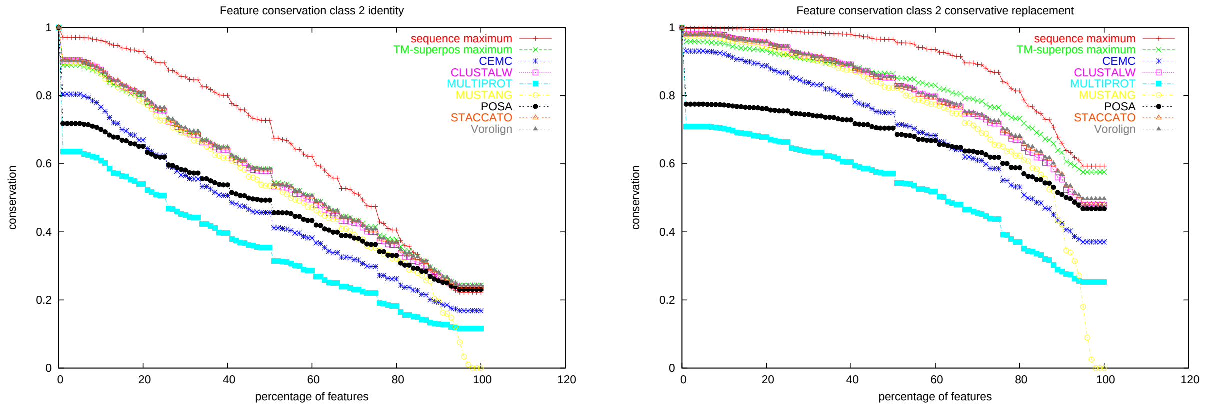


Figure 5.9: This figure shows the conservation of PROSITE pattern residues as well as Swissprot annotated functional residues in the multiple alignments analyzed. The left hand plot expects the same residue as the pattern residue at a position p, while the right hand plot allows for all residues with a positive Dayhoff matrix score.

which are often important for the structural stability of a protein. The results of the conservation of CSA residues and cystein bridges are shown in Figure 5.8.

Features of class two comprise residues included in PROSITE [74] patterns as well as residues annotated as functional residues in the Swissprot [175] database. Those features are also supposed to be conserved in multiple alignments. Nevertheless, we test not only the exact conservation of the corresponding residue in the multiple alignment but we also allow for mutations with a positive PAM matrix score at the corresponding alignment column. The results are shown in Figure 5.9.

As it can be seen from this detailed analysis, Vorolign and Staccato perform best out of the seven methods tested and outperform the other structure alignment methods regarding feature conservation. The fact that they take the protein sequence into account during the alignment process allows for a higher conservation of important residues in the alignments. This finding is also confirmed by the comparable performance of CLUSTAL W. Surprisingly, Mustang does not align the functionally important residues at all for almost 10% of the features (as indicated by the sharp drop of the conservation rate around 90% on the X-axis).

Together with the results of the structural quality of the induced pairwise alignments, Vorolign (and Staccato) seem to provide a reasonable tradeoff between a good structural quality of the alignments as well as a high quality with respect to aligning functionally and structurally relevant residues. The combination of those two features makes both methods very interesting for application in further analyses of common features protein structure families with respect to functionally and structurally important residues.

5.3 The AutoPSI database

With the AutoPSI (Automated Protein Structure Identification) database published in 2008 [16] we implemented a large-scale application of Vorolign to predict the SCOP families of all proteins unclassified in the latest SCOP release. While the SCOP classification pipeline consists of several manual steps and therefore is quite time consuming, structural genomics projects and improved structure determination pipelines have led to a significant increase of known protein structures published each year. Taken together, this leads to the fact that thousands of protein structures are not classified in the most current version of the SCOP database and therefore are not available for methods and analyses relying on SCOP as an important resource.

In AutoPSI we have joined the predictive power of Vorolign with the recently developed AutoSCOP method [56] in order to provide SCOP classifications for thousands of unclassified protein structures and millions of unclassified protein sequences in UniProt [175]. In the following we will focus on the Vorolign component of the database in order to analyze the capabilities of Vorolign in predicting SCOP classifications at a large scale in some more detail.

5.3.1 The current database content

The most current version of the database contains SCOP classification predictions for all protein structures available in the PDB in September 2008. SCOP classifications were obtained from the SCOP parsable files of the most recent SCOP version 1.73 (November 2007). In total, the current PDB version contains 51317 distinct protein structures made up from 126186 protein chains. Currently SCOP provides predictions for 34426 PDB entries consisting of 75746 chains.

Therefore, only 67% of the entries and 60% of all known chains are classified in SCOP showing the need for an automated and up to date supplement of the SCOP releases.

We ran the VorolignScan pipeline for all 16000 entries missing in the current SCOP release and predicted the SCOP family of the targets based on 7828 templates of SCOP release 1.71. As described in section 5.1.6, potential domain borders were predicted by PDP and the best 250 templates that pass the SSEA filter are aligned against the target. In total, the process took 2 days on a cluster of 20 CPUs which again displays Vorolign's capability for a very fast scan for similar protein chains in a database of potential templates.

We then analyzed the VorolignScan results for predictions which are reliable either in terms of their high structural similarity to the template (corresponding to a TM-score of greater than 0.6) or according to a very good Vorolign score. To obtain a length independent score we normalized the Vorolign alignment score by the average length of the two proteins being aligned. The threshold has been set to 15 which in our experiments proved to be a highly reliable indicator for either being a correct superfamily or family hit (see also Figure 5.4b). Using the Vorolign score as additional criterion for the quality of the assignment allows to predict a SCOP classification also in cases where large flexibilities exist in the protein structures being compared leading to bad superposition-based scores (TM-Score and RMSD).

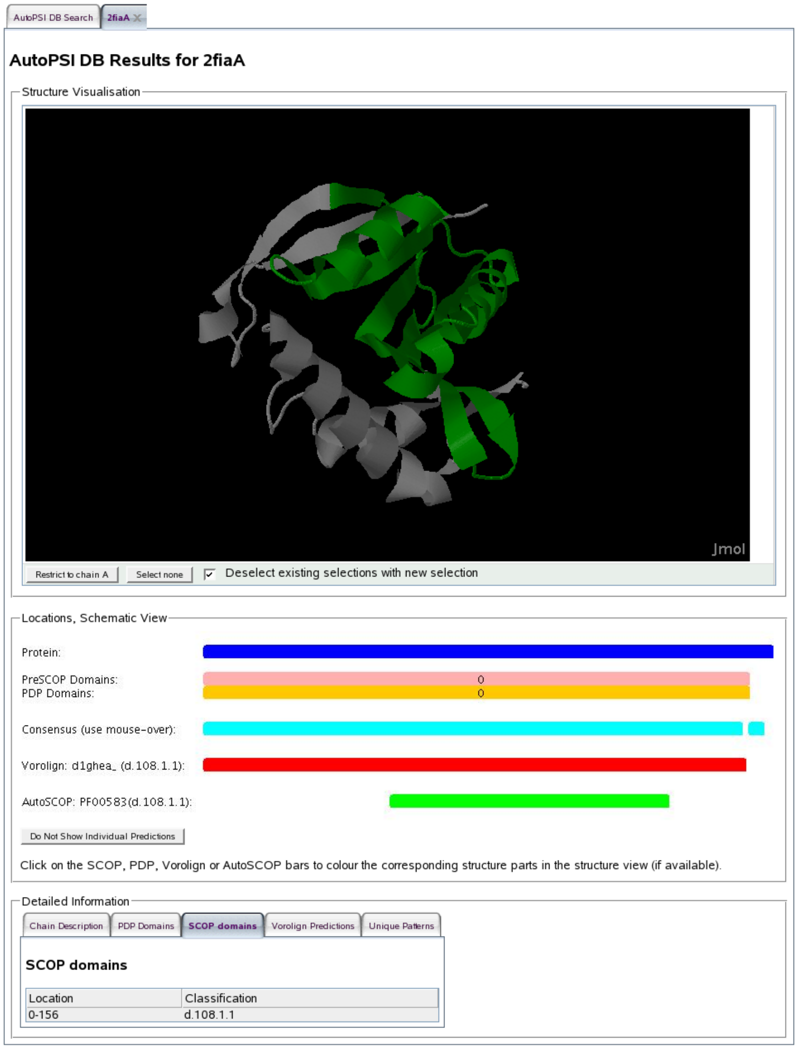
In total, Vorolign can identify highly reliable (sometimes only partial) templates for 13673 PDB entries and 30953 chains. Therefore, using Vorolign, we can assign reliable SCOP classifications to 93% of all known PDB entries and almost 85% of all known protein chains and close the gap between SCOP releases. Assignments can also be made in cases where larger flexibilities exist. For example, filtering for proteins with a TM-score smaller than 0.4 but having a Vorolign score greater than 15 and also sharing a sequence identity of more than 50% we identify 221 PDB entries and 322 protein chains from 90 different SCOP families which are have highly flexible protein structures (as indicated by the rather small TM-scores) but are clearly related on the sequence level. The similarity of template and target can still be detected by the sequence-driven scoring function of Vorolign but is hard to detect on the structure level.

In contrast, Vorolign can also identify structural similarities (indicated by TM-scores larger than 0.6) in cases where the sequence identity is below 20%. In total, we find 588 distinct protein entries where sequence similarity is very low, but the structural quality of the superposition as well as the normalized Vorolign scores indicate a clear structural and likely evolutionary relationship.

5.3.2 Database access

The data stored in the AutoPSI database is accessible in two different ways. Firstly, we provide a flat file distribution of the database which contains the individual predictions made for PDB proteins as well as UniProt proteins. Secondly, the data is accessible by a intuitive web interface available at http://www.bio.ifi.lmu.de/AutoPSIDB which is implemented using the Zkoss (http://www.zkoss.org) framework.

The database can be searched in multiple ways and for every entry, a detail view (one example can be seen in Figure 5.10) can be requested. This page allows to view detailed information about SCOP classifications annotated to the proteins and shows the protein structure (if avail-



AutoPSI DB Results for 2fiaA

☑ Deselect existing selections with new selection

Click on the SCOP, PDP, Vorolign or AutoSCOP bars to colour the corresponding structure parts in the structure view (if available).

SCOP domains

Location	Classification
0-156	d.108.1.1

Figure 5.10: Shows the detail view of the AutoPSI database

able) using Jmol (http://www.jmol.org). The image below the structure visualization shows the protein, its domains and consensus predictions. By clicking on the button ”Show Individual Predictions” a user can get a larger image which contains detailed information about matching InterPro patterns and Vorolign predictions as well as their locations on the sequence (see Figure 5.10). The locations of all patterns and Vorolign predictions can be visualized on the structure by clicking on the corresponding pattern in the image. Below the overview image and the structure visualization, a box containing tabs for the sequence itself and the outputs of the different predictions mechanisms and annotations found for it provides further information as well as links to external databases.

The AutoPSI database will be extended by additional methods for automated structure classification in the future. Currently we plan to include the PPM method in the pipeline in order to allow for consensus predictions of 3 distinct methods to further enhance the prediction capability of the database and to join the different strengths of the different approaches to protein structure classification.

5.4 Voronoi contact patterns around functionally important sites

With the arrival of structural genomics projects [171] and the faster growing number of protein structures available in the PDB not only the automatic classification of protein structures into families but also the computational characterization of the biological functions of those proteins has become an important issue. For many protein structures solved in those projects the biological function, their enzymatic sites and ligands or their potential binding partners are unknown. Since this is a reversal of the classical structural biology approach to determine a protein structure in order to understand how the protein performs its specific function on the molecular level, it requires for novel strategies and methods to predict a proteins function given its structure.

Protein function prediction has a long history and many methods to characterize protein function based on similarities detected on the sequence level (homology) have been developed. The most prominent one may be PSI-BLAST [3] which allows to search a query sequence against a large protein database in reasonable time. But despite the usefulness of BLAST searches and homology-based function inference the problem remains that the exchange of only one amino acid in the active center of an enzyme or at the surface of the structure being responsible for binding other proteins may change the function of the protein. Such subtle differences can hardly be identified and taken care of in PSI-BLAST and other sequence-based searches.

To partly overcome this problem on the sequence level, databases providing patterns or ”fingerprints” of highly conserved residues involved in or near functionally important sites have been developed. Among them are PROSITE [74] providing a set of regular expressions and the PRINTS database [7] collecting position specific scoring matrices of functionally important sites. Of course, all other member databases of InterPro [109] collecting e.g. hidden Markov models of protein families may help in order to assign a potential biological role to a protein sequence of unknown function.

In difference to sequence-based methods, structure-based approaches to function prediction can rely on a larger set of additional features known for the protein and not available on the pure sequence level. Among them are the detection of surface clefts, structural similarity not detectable on the sequence level via structural alignments or residue conservation and correlated mutations of surface residues. All those methods are, among others, integrated in the ProFunc server [93].

Among the most powerful methods for function assignments based on the structure are methods relying on sets of structurally conserved residues in the active site of the protein, so-called 3D templates or 3D motifs. Such motifs can either be annotated manually, similar to the example of PROSITE on the sequence level, like the CSA (Catalytic Site Atlas) database [123] or can be derived automatically from a set of structures with the same function ([122], [86], [110]).

5.4.1 Outline of the method

Matthias Siebert has evaluated the usefulness of the Vorolign scoring function to identify functionally important, conserved 3D patterns in protein families in his Diploma thesis under the supervision of the author. Similar to other 3D template based methods we aim at the automatic identification of highly conserved 3D residue patterns in protein structures which are known to have the same function. In a first step, we identify (functionally) important *seed residues*, e.g. using PROSITE patterns or by identifying highly conserved and spatial close residues in multiple alignments of the family. For the identification of larger patterns, we assume that not only the functional residues themselves but also surrounding residues (called *supporting shell residues* in the following) may have important functions e.g. to coordinate functionally important residues or exhibit previously undiscovered functions and therefore may be highly conserved in the family as well. Therefore, in a second step, we aim at the identification of highly conserved, consistent residue networks in the supporting shells of functional sites across all members of the family. Given a Vorolign pattern for a family we can then search a complete database of protein structures for matches of the pattern in a third step and therefore identify proteins with potentially similar functions. In the following we will briefly outline and describe the single steps of the method.

In the first step, we need to identify residues of potential interest which are used as starting points for the further exploration of the contact network conservation. The most simple approach to define such residues is to make use of PROSITE patterns describing functionally important residues and active sites like the histidine residue of the catalytic triad in trypsin-like serine proteases (Pattern PS000134: [LIVM]-[ST]-A-[STAG]-H-C). In this first attempt to test our ideas we made mainly use of such active site PROSITE patterns in order to define seed residues (see also Figure 5.12). However, an other approach which does not require prior knowledge could be to explore all sites (e.g. all pairs or triplets of residues) which are highly conserved in a multiple Vorolign alignment of the family and which are close in space on the surface of the protein as potential seeds.

The second step aims at the identification of 3D patterns containing highly conserved residues in the supporting shell of residues around the seeds. The idea is supported by the finding that interactions between catalytic and non-catalytic residues may play functional roles in catalysis [63], e.g. in controlling tautomerization of the histidine imidazole ring or in the stabilization of charged residues. Starting from the seed residues defined in step one we extent the potential pattern to all direct neighbors of the seed residues in the Voronoi tessellation of the structure (supporting shell) and to all direct neighbors of the supporting shell residues, i.e. the supporting-supporting shell. All those residues are potential candidates for being contained in the final Voronoi pattern (see also Figure 5.13).

Given now two protein structures together with their seed and shell residues, the task is to identify the conserved residue network in both structures. Given the mapping of the seed residues (which is given by the PROSITE pattern or the multiple alignment) we first aim at the mapping of supporting shell residues of the seed residues. In our method we make use of the Vorolign scoring function in order to compute the similarity of residues in the cell neighborhood and then we extract the final residue mapping from the Vorolign low level matrix (see also Figure 5.3). Having mapped the shell residues of the seeds we carry out the same procedure for mapped shell residues and their nearest neighbors according to the Voronoi tessellation. Finally, this results in a set of residues placed around the seed residues which are mapped onto the respective residues in the other structure. Those can then be used to define the Voronoi pattern using different features for the edges and vertices in the pattern like secondary structure and amino acid conservation of the vertices or geometrical constraints (distance, face area...) of the edges.

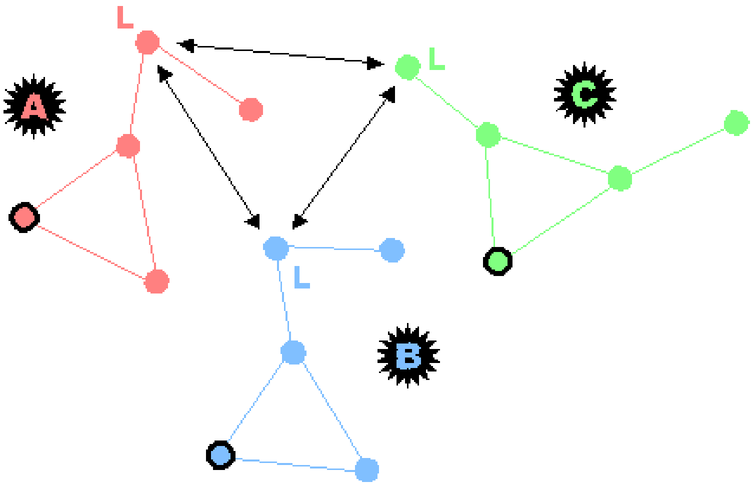


Figure 5.11: Consistency check example considering three patterns from proteins A (red), B (blue) and C (green). Seed residues are indicated by black circles and the patterns are depicted in a graph-like representation. The (consistent) mapping of the Leucine residues is shown by black arrows. (*Figure taken from the diploma thesis of Matthias Siebert*)

In the case of more than two structures in the family, we need to identify the contact networks being conserved across all members of the family, given a set of all pairwise Voronoi contact patterns of pairs of proteins in the set. The need combine those sets of patterns into one consensus pattern is a similar problem like the computation of a multiple alignment given a set of pairwise alignments. The idea of our mapping procedure is similar to the T-Coffee method for multiple alignments. Our goal is to identify all residues which are consistently mapped between two proteins A and B given a third structure C (a simple mapping case is also shown in Figure 5.11), and finally to identify all such residues which are consistent and conserved among all members of the family. The final consensus Voronoi pattern (one example is shown in Figure 5.12) is defined by the pair of patterns from two structures A and B which implies the largest consistency across all other patterns and structures in the group and therefore the multiple pattern alignment which includes the most alignment columns without any gaps.

Given a mapping of all patterns in the set, we can extract different features to form the consensus pattern. Those can in principle include all geometric, structural or biochemical features. In the diploma thesis of Matthias Siebert several such feature representations have been tested. Surprisingly the features also used by Vorolign, namely the combination of secondary structure and amino acid features also turns out to be the best performing one and is therefore applied in the case study described below.

The final question now is how to search with a pattern or a set of patterns against a database of e.g. newly resolved protein structures. The task is to match a pattern, which corresponds to a Delaunay graph, to the Delaunay graph of the protein structure to be searched which corresponds to the subgraph isomorphism problem. Despite the fact that the problem is known to be NP-complete, we use a brute-force enumeration procedure, more precisely, an exhaustive depth-first tree-search algorithm. This approach is computationally feasible since the patterns and their vertex and edge features allow for pruning the search space dramatically.

5.4.2 Results: Case study of the Trypsin-like serine protease family

In the following we will show the individual steps of the method and the result for a specific protein family, namely Trypsin-like serine proteases which are well characterized group of enzymes. They are covered by two PROSITE patterns (TRYPSIN_HIS (PS00134) and TRYPSIN_SER (PS00135)). In SCOP eukaryotic proteases correspond to family b.47.1.2 which (in SCOP 1.67) consists of 53 structures, 52 of which are matched by at least one pattern, while only 43 of them are matched by both patterns. Those 43 structures are used to construct the final consensus pattern.

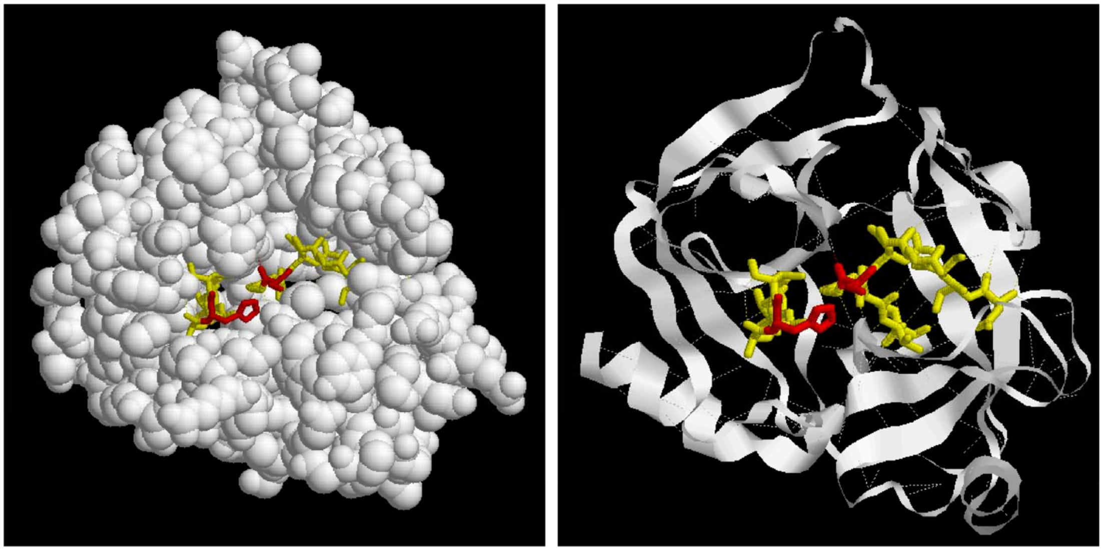


Figure 5.12: PROSITE pattern matches in trypsin (PDB: 1a0j, chain A). The residues matched by the pattern are shown in yellow, seed residues, corresponding to the histidine and serine residues of the catalytic triad are shown in red. (*Figure taken from the diploma thesis of Matthias Siebert*)

In the first step, PROSITE patterns (TRYPSIN_HIS and TRYPSIN_SER) are searched in protein sequences and are mapped onto the corresponding protein structures. The two catalytic residues defined in the PROSITE patterns shown in red in Figure 5.12 are used as seed residues of the Voronoi pattern to be constructed.

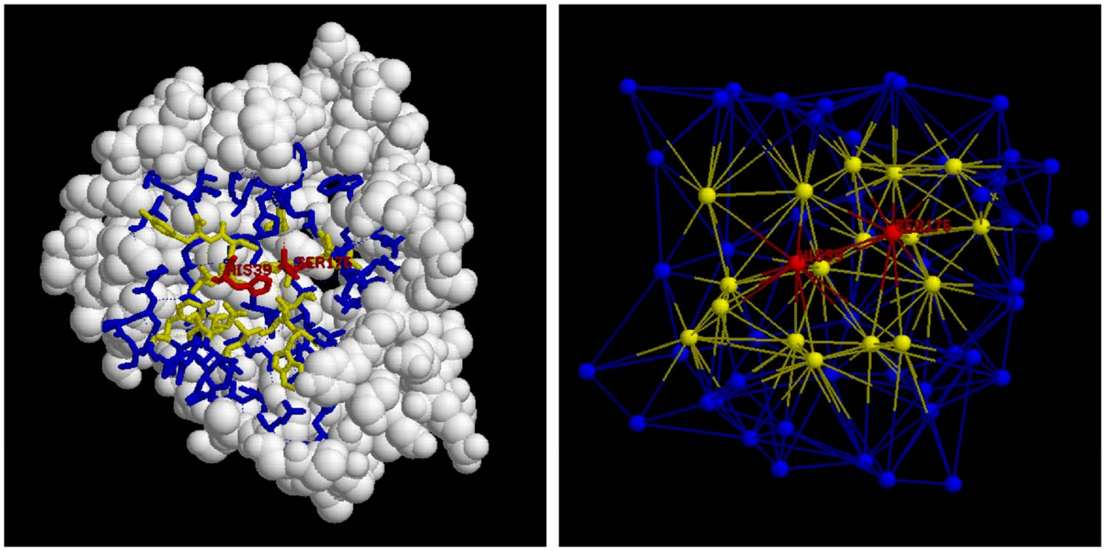


Figure 5.13: Initial Voronoi pattern in trypsin (PDB: 1a0j, chain A). Seed residues are shown in red, direct neighbors of the seeds are shown in yellow and neighbors of the yellow residues are shown in blue. On the right hand side, the Delaunay graph of the corresponding Voronoi tessellation around the active site is shown. (*Figure taken from the diploma thesis of Matthias Siebert*)

Starting from the seed residues, the initial Voronoi pattern is constructed containing all residues being direct neighbors of the seed residues (supporting shell residues) or direct neighbors of seed neighbors (supporting-supporting shell residues). The result of this process is shown in Figure 5.13.

Based on all those initial Voronoi patterns for the members of a family, we construct a consensus pattern as described above. The consensus pattern resulting from this procedure for eukaryotic Trypsin-like serine proteases is shown in Figure 5.14 and consists of 8 highly conserved residues among all 43 proteins used to construct the pattern.

In order to evaluate the stability of the pattern construction process, we performed a leave-one out crossvalidation by constructing the consensus pattern from n-1 (i.e. 42) structures and then testing if the resulting pattern matches the structure being left out. The crossvalidation for this family pattern yields a accuracy of 95% indicating that the patterns generated by our algorithm are quite stable and robust against smaller changes in the training set in the specific case of Trypsin-like serine proteases. One has to note that the pattern represents the consensus of many structures in this case, while in most families much fewer family members are available which often lead to less robust patterns.

We then searched the complete SCOP database (release 1.67) for matches of the 3D Voronoi pattern and identify 54 hits of the pattern in all protein structures. Those contain the 43 matches the pattern was build from, 4 additional members of family b.47.1.2 where only one of the two PROSITE patterns matches (indicating that the regular expressions of PROSITE are too spe-

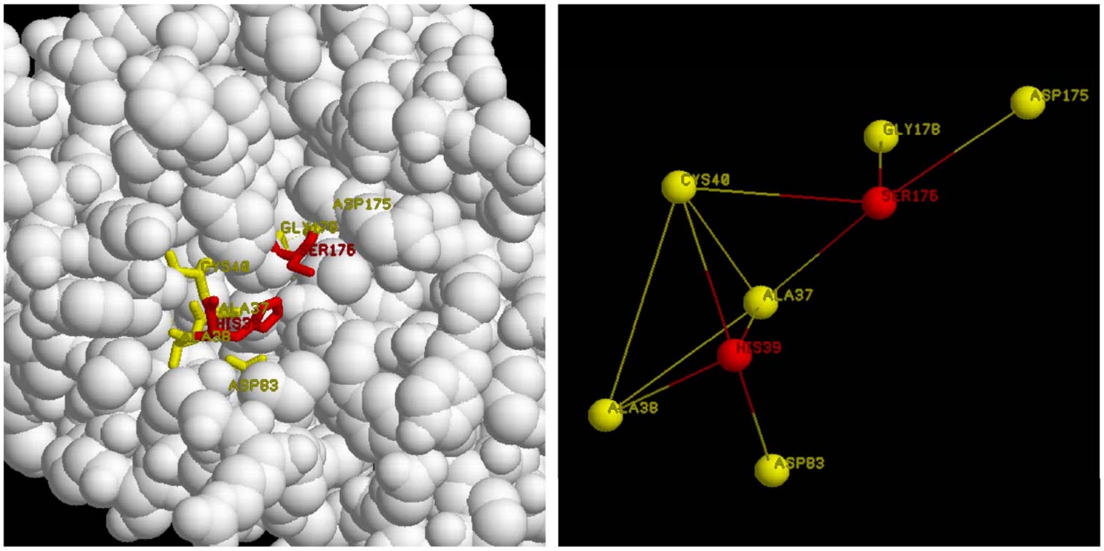


Figure 5.14: PROSITE-based consensus Voronoi pattern for SCOP family b.47.1.2 containing Trypsin-like serine proteases. The pattern is visualized in Trypsin (PDB: 1aj0, chain A). On the left, the 8 residues constituting the pattern are colored and shown in the protein structure. On the right, the consensus pattern is shown as Delaunay graph which represents the catalytic unit of this family conserved among 43 members of the family. Please note that the pattern also contains the third residue of the catalytic triad (aspertate) which is not represented by a single PROSITE pattern (*Figure taken from the diploma thesis of Matthias Siebert*)

cific to detect the other catalytic residue) and moreover, 6 additional protein structures which are classified into SCOP class b.47.1.1 corresponding to prokaryotic Trypsin-like serine proteases. Therefore, the pattern appears to be highly specific and can even detect members from the prokaryotic domain, indicating the high evolutionary conservation of not only the catalytic triad but also the surrounding, coordinating residues in the evolution of the family.

We then tested the ability of the pattern to functionally classify newly resolved protein structures. In order to do so, we scan for pattern matches in the difference set of proteins being classified in SCOP 1.69 but which are not yet contained in SCOP 1.67. In this set, which contains about 5000 structures, 30 structures are detected which match the Voronoi pattern for b.47.1.2. Those comprise 28 of the 34 structures (82%) classified into SCOP family b.47.1.2, while two hits correspond to false positive hits and lower the specificity of our method.

5.4.3 Discussion

In the thesis of Matthias Siebert we have tested the usability of the Vorolign scoring function for a different task, namely the identification of highly conserved subnetworks of protein structures which have potential applications in the prediction of protein function or protein family membership. Based on preliminary tests using Trypsin-like serine proteases we could show that protein contact networks may be highly conserved around functionally important residues and the identification of such conserved subnetworks can lead to very specific and sensitive contact patterns which can, in turn, be used for protein function prediction. Also, an advantage of the approach is the fact that the method, in contrast to other approaches, does not require for any (structural) alignment or superposition of the active site and therefore, just like Vorolign, allows for an intrinsic flexibility of the protein.

But despite those promising results several questions and problems remain. The method highly depends on specific seed residues which, at the moment, are only identified by PROSITE pattern matches. This needs to be extended towards an automatic identification of conserved, spatial close residues in protein families which then allow to process more and maybe less well studied examples and might also allow for the identification of previously undiscovered contact patterns of functional importance or other important core networks of protein structures like early folding units.

Furthermore, in the current stage, the method performance is strongly affected by the number of structures available in the pattern generation process where too few structures lead to large patterns containing many residues which are not very general leading to a disappointing robustness in the leave-one out crossvalidation and / or the performance in detecting novel structures of the same family. To this end we need to improve the consensus detection methodology and the usage of features in the final consensus pattern in the future.

Also, more experiments are required to evaluate strengths and weaknesses of the approach in comparison to other prediction methods in the field and on a larger set of families.

5.5 Conclusions

With Vorolign and PPM we presented two novel methods to address the protein structure alignment problem. While PPM introduces a novel approach to measure the evolutionary relationship between protein structures, Vorolign is based on the idea to measure the similarity of protein structures using an alternative representation, i.e. Voronoi cells, which implicitly captures structural features of the proteins being compared. The similarity of two sets of nearest-neighbor residues (which is extracted from the information in the Voronoi cell representation of the protein) is then scored using amino acid and secondary structure exchange matrices. This similarity criterion is then optimized by double dynamic programming.

PPM has its specific strength in the detection of conserved structural cores of protein families in the presence of natural structural variation (phenotypic plasticity). Also, as shown in our benchmark results, the method performs very well in predicting the SCOP classifications for newly resolved protein structures.

The major benefit of Vorolign is its speed (on average 5 alignments per second, excluding time for Voronoi tessellation) that allows for a rapid although very accurate comparison of a protein structure against a large database of potential structure templates in order to classify protein structures with a high accuracy. Those abilities of Vorolign have led to the incorporation of the method in the ProKSI server (http://www.procksi.net) [11] developed at the University of Nottingham and the development of the AutoPSI database as discussed above.

Both methods proposed in the thesis perform superior or comparable to other standard algorithms for protein structure comparison like CE, TM-align and STACCATO in terms of their speed (especially CE takes up to a minute to compute an alignment) as well as their classification and alignment accuracy.

Vorolign's ability to produce accurate pairwise and multiple structure alignments, even of highly flexible protein structures, is an important feature to detect similarities across the natural variance of protein structures. The detailed analysis of multiple alignments produced by Vorolign shows that Vorolign performs comparable to standard multiple structure alignment algorithms in terms of structural but also sequence-based quality criteria. Especially the good conservation of functionally and structurally important residues in the multiple alignments represents an important and interesting feature of the method.

Moreover, with the AutoPSI database and our approach to detect conserved residue patterns using the Vorolign scoring function we have described two applications of the Vorolign method. While the AutoPSI database is already a useful resource for researchers that rely on very up-to-date structure classification databases, e.g. for protein structure prediction and others, the power of Voronoi patterns, despite some promising initial results, needs to be evaluated in detail in future research.

Flexible, pairwise or multiple alignments computed by Vorolign or PPM for families, superfamilies or folds can therefore be useful to define common cores of proteins which are an important step towards a deeper understanding of protein structure and sequence evolution with applications in protein structure prediction or the analysis of alternative splicing in the context of protein structures which will be discussed and addressed in detail in the following chapters.

Part II

Alternative Splicing in the Context of Protein Structure

Chapter 6

Introduction to the Analysis of Alternative Splicing

Many eukaryotic genes are separated into coding parts, so called exons, and non-coding parts, so called introns. In order to obtain the mature mRNA which is translated into protein by the ribosome, the cell needs to remove the intronic parts from the pre-mRNA. This process is called *splicing*. For many genes in higher eukaryotes, the exons of a gene are assembled in different ways during pre-mRNA splicing such that different mRNAs and, thus, protein *isoforms* are produced from the same gene. This process is called *alternative splicing*.

Alternative splicing appears to be the rule rather than the exception in eukaryotic cells and is thought to be one of the major sources for functional diversity in the proteomes of multicellular organisms. Based on very recent next-generation sequencing data it is estimated that more than 90% of all human multi-exon genes are alternatively spliced [169] which drastically increases the number of proteins in the human proteome and, together with a time and tissue-specific regulation increases the functional complexity of an organism.

Spliced proteins are involved in many biological processes such as apoptosis [139] or the control of synaptic function [98] and play important roles in human diseases like cancer where the influence of alternatively spliced genes on transcription factors or signaling pathways has been described [163]. The effects of alternative splicing on the function of a single protein range from changes in substrate or interaction partner specificity to the regulation of DNA binding properties [148].

In order to change the function of a protein by alternative splicing, its structure may be changed accordingly. In this chapter we will mainly focus on our methods developed for a structure-aided analysis and interpretation of alternative splicing events. But first we will briefly review the fundamentals of alternative splicing as well as the results of recent analyses of the roles of alternatively spliced proteins in cellular processes and disease.

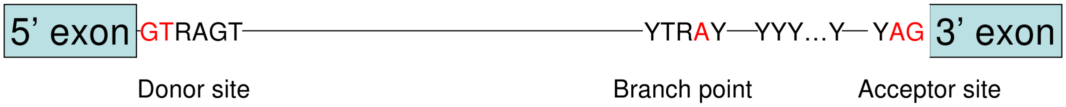


Figure 6.1: Important splicing signals on the intron mRNA sequence recognized by the snRNPs. Please note that Y stands for a C or a T and R stands for A or G. The GT and AG nucleotides at the exon-intron borders, as well as the central adenosine in the branch point sequence are highly conserved and, in the case of the central adenosine, are of catalytic importance.

6.1 Fundamentals of alternative splicing

6.1.1 The splicing process in the cell

While there are known examples of self-splicing introns, in higher organisms splicing of the pre-mRNA is usually carried out by the spliceosome a molecular complex is almost as large as the ribosome. Similar to the ribosome it consists of several RNA and protein molecules which are combined to so called snRNPs (small nuclear ribonucleo-protein particles). Five snRNPs, termed U1, U2, U4, U5 and U6, are especially important for the splicing process. The snRNPs detect certain signals on the mRNA sequence (see Figure 6.1) via specific base pairings of the RNA parts of the snRNPs with the pre-mRNA and are responsible for catalyzing the splicing process.

According to House et al. [71] the spliceosome assembly takes the following steps. First, the splicing signals on the mRNA sequence are detected by U1 which binds to the splicing donor site and a protein heterodimer, U2AF, binding to the acceptor site. Those form the so-called *E-complex* of the spliceosome. ATP-dependent addition of U2 which binds to the branch point leads to the formation of the *A-complex*. Then a larger, tri-snRNP complex consisting of U4, U5 and U6 enters the spliceosome to form the *B-complex* consisting of all snRNPs. U6 replaces U1 and binds to the donor site, U1 and U4 are released from the spliceosome, the *C-complex* is formed. The mRNA is cleaved at the donor site and the 5’ intron end is attached to the adenosine at the branch point forming a lariat structure. Then the mRNA at the acceptor site is cleaved and the 3’ exon start is ligated to the 5’ exon end. Finally, the intron is released.

As already mentioned, the recognition of the intron start, its end and the branch point requires certain conserved splicing signals on the mRNA level. The locations and consensus sequences of the three most important signals, donor and acceptor site as well as the branch point, are shown in Figure 6.1. The branch point is located about 40 nucleotides upstream of the acceptor site. Between branch point and acceptor site a pyrimidine rich region, the so called polypyrimidine tract, can be found which is also important for the acceptor detection by U2AF.

6.1.2 Different types of alternative splicing events

All genes containing introns and exons undergo splicing in order to remove intronic, non-coding parts from the pre-mRNA and to form the mature mRNA transcript. Additionally to the standard splicing process which links exon n with exon $n + 1$, many transcripts are formed by alternative

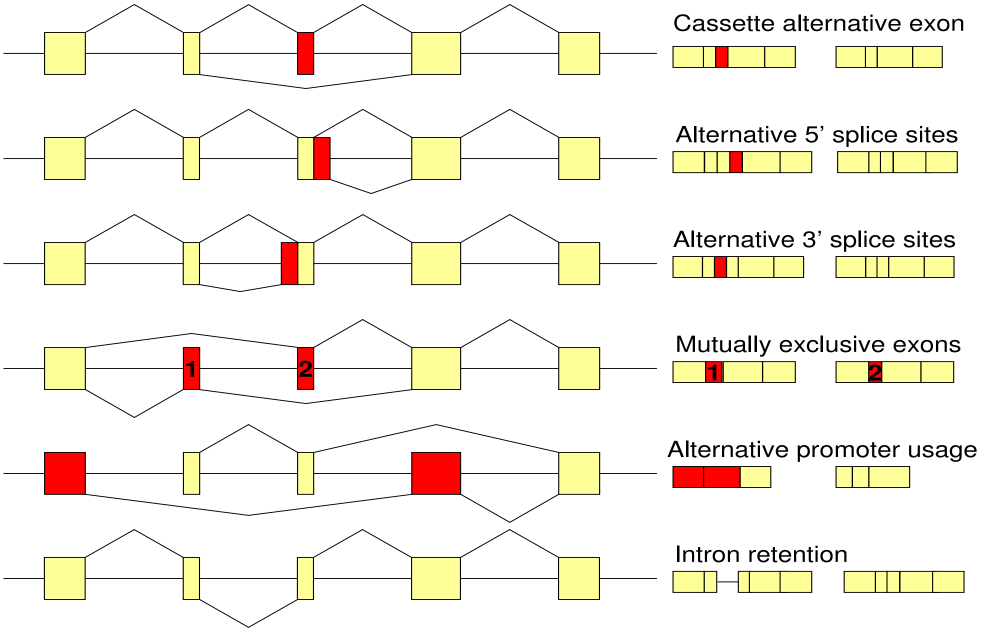

Cassette alternative exon
Alternative 5' splice sites
Alternative 3' splice sites
Mutually exclusive exons
1
2
1
2
Alternative promoter usage
Intron retention

Figure 6.2: Different types of splicing events observed in eukaryotic transcripts accounting for the generation of functionally distinct transcripts according to [23]. Alternative exon usage is shown in dark gray, constitutive exons in light gray. While the splicing pattern is shown on the left side of the figure, the resulting transcripts are shown on the right.

splicing, i.e. not all exons of the gene are present in the final transcript. Exons which are present in all transcripts of a gene are said to be *constitutive exons* while exons which may or may not be present in a transcript are called *alternative exons*. As shown in Figure 6.2 alternative splicing events may lead to complex patterns of exon usage. At least one-third of the known splicing events are contributed by cassette-type alternative exons which may be inserted or be omitted in a transcript. Such exons may e.g. account for the introduction of an additional domain in the isoform leading to a membrane-bound or free protein product [179]. Also, alternative 3' and 5' splice sites are frequently used which often only introduce subtle changes in the protein sequence, e.g. alternative NAGNAG splice sites [68], accounting for about one-quarter of the events. Other types of splice events described in Figure 6.2 or combinations of those types can lead to highly complex and very different transcripts from one genomic locus whose function may differ significantly.

6.1.3 Regulation of alternative splicing

As we will discuss below in detail, the functions of different isoforms may be very different and even antagonistic to one another. While one isoform may have apoptotic function leading to cell death, the other isoform lacking one or more exons may lead to cell survival [173]. Therefore it is clear that similar to the expression of a gene in general, the generation of specific isoforms by alternative splicing must be highly regulated in the cell.

Surprisingly, the core splicing signals described above are only poorly conserved in higher organisms and do not contain sufficient information for the splicing machinery to distinguish between correct and cryptic splice sites. Additionally, due to the length of many introns which are much longer than the flanking exons, false positive, random splice signals are much more abundant than correct splice sites leading to the question how the production of the correct transcript required at a specific time in a specific tissue is controlled. As shown by several studies (e.g. reviewed in [23], [71] and [172]) additional *cis*-acting sequence elements in exons and introns are important for splice site recognition and to distinguish correct from wrong pairs of splicing donors and acceptors.

Those sequence elements can either promote recruitment of the spliceosome and therefore the inclusion of the corresponding exon (splicing enhancers) or disrupt the assembly of the splicing machinery and cause exon skipping (splicing silencers). Today the use of most exons is believed to be under the combinatorial control of multiple regulatory RNA elements as well as the inherent strength or weakness of its splice sites. Some of those splicing enhancers and silencers act through their specific RNA secondary structure but the majority acts primarily as platforms for binding non-snRNP regulatory proteins. Exonic splicing enhancers often bind a family of proteins known as ”SR proteins”, which contain an RNA binding domain and a region rich in Arg-Ser dipeptides. By contrast, exonic silencers typically function to repress exon inclusion by recruiting members of the hnRNP family of proteins, a structurally diverse set of RNA binding proteins.

Overall, the splicing pattern of an mRNA transcript seems to be determined by the interplay of several enhancers and silencers which are found along the nascent transcript and directly or indirectly regulate the spliceosome assembly. The combination of those sites in a pre-mRNA transcript is often referred to as the ”splicing code” [23]. Bioinformatics studies (e.g. [165]) have already led to the discovery of potentially novel binding motifs in intronic and exonic regions. Data on known alternative splicing events, genomic data from different organisms and conserved motifs in those genomes as well as the integration of those data sources with protein network data will further help elucidate the interplay of genomic elements and their corresponding proteins in regulating (alternative) splicing in the post-genomic area and will therefore help to decipher the splicing code.

6.2 Biological functions of alternative splicing events

Alternatively spliced isoforms are involved in many biological processes and for hundreds of splicing events differing functions of the isoforms have been characterized by experimental studies. Functional alterations range from changes in the ligand-binding specificity in FGFR2 [183] and the subcellular localization, e.g. membrane-bound vs. free in the cytoplasm in genes of the tumor necrosis factor receptor family [179], to changes in the domain composition of the resulting protein [87]. Several examples are also reviewed in [148].

Besides the analysis of the function of specific isoforms, several large-scale studies have analyzed the effects of alternative splicing on a genomic scale for example focusing on the impact of splicing on interaction domains [130] or the distribution of alternatively spliced transcripts in certain categories of the gene ontology (GO) [112]. The study by Resch et al. [130] has for example led to the interesting insight that genes encoding proteins of the Kruppel family of transcription factors are often alternatively spliced and that splicing tends to remove the protein-protein interaction domain instead of the DNA-binding domain. Spliced isoforms therefore are still able to bind to DNA but exert a positive rather than a negative effect on gene expression due to the loss of the repression domain.

The outcome of alternative splicing may therefore be a functional protein which folds into a stable three-dimensional structure and carries out a specific function in the cell. This function may also be the down-regulation of the active gene product by the production of non-functional isoforms lacking i.e. the active site or protein-protein interaction domains required for the function.

A different mechanism which regulates the level of the active gene product of a gene in the cell is nonsense-mediated decay (NMD) induced as a consequence of alternative splicing events. NMD is an RNA surveillance function which recognizes mRNAs containing premature stop codons, located more than 50 nucleotides upstream of the site of removal of an intron, and targets those transcripts for destruction rather than translation into protein [96]. NMD seems to be an efficient mechanism to avoid the translation of transcripts which have been mis-spliced but to date it is unclear how many of the transcripts observed in public databases are prone to NMD as other studies report evidence against the coupling of alternative splicing with NMD to control gene expression [120].

Experimental and computational studies in the last years in combination with the wealth of information available through sequencing complete genomes have sharpened our understanding of the different roles of alternative splicing for generating functional diversity in eukaryotic genomes. Many open questions regarding the time and tissue specific regulation of alternative splicing, the evolution of splicing events in different organisms, the functional effects of splicing on a single protein and the protein network level as well as the prediction and detection of splicing events in experimental and genomic data remain.

Additionally, besides those open questions, most experimental and computational studies on the effects of alternative splicing on the function of different isoforms so far have neglected the effects of alternative splicing on protein structures. In the following we will discuss our tools developed for a structure-based analysis of alternative splicing events on a genomic scale and the results obtained from our analyses on the structural effects of alternative splicing.

6.3 Alternative splicing and disease

Alternative splicing allows for the development of a highly dynamic proteome in higher organisms on the one hand. On the other hand, it requires for a very tight regulation of the expression of specific splice variants in tissues or developmental stages. Expressing the wrong isoform at the wrong time or expressing splice variants which simply contain the wrong combination of exons can lead to non-functional proteins or proteins which protect a cell from apoptosis instead of inducing cell death [173].

Indeed, we are just starting to understand the complex mechanisms of splicing regulation, isoform function and the connection to human disease but it is obvious that *cis*- or *trans*-acting mutations which disrupt the regulatory splicing code or affect the splicing machinery or associated proteins themselves have various roles in human diseases. Currently it is estimated that up to 50% of all disease-causing mutations in human affect alternative splicing [170]. Such mutations may have various effects on splicing and can lead to an increased rate of exon skipping (gain-of-splicing-function mutations) or exon inclusion (Loss-of-splicing-function mutations) which finally disrupt the required mRNA transcript and are reviewed e.g. by Wang et al. [170]. Well known and characterized examples of mis-splicing can lead to an abnormal form of *cystic fibrosis* via skipping exon number 9 of the *CFTR* gene. Also changes in the rations between two isoforms can have deleterious effects as known for example for the *MAPT* gene encoding the *tau* protein where shifting the balance of the two major isoforms seems to be associated with fronto-temporal dementia with parkinsonism (FTDP-17) [55].

Due to the importance of spliced transcripts on disease and even the response of patients to certain drugs have led to attempts to influence alternative splicing and the transcription of specific isoforms in the cell. The recent years have shown that antisense oligonucleotides [156] or RNAi [55] may lead to the development of powerful therapies in clinical applications of diseases [55]. But even if such treatments work *in vitro* and *in vivo* we need to understand in detail what the regulation, function and structure of those alternative isoforms is like in order to understand the wanted and also unwanted effects of their expression and the treatment in the cell and in the organism.

Chapter 7

Alternative Splicing and Protein Structure Evolution

The consequences of alternative splicing events on the protein structure level are largely unknown. To analyze different aspects of this problem we have carried out a comprehensive structural analysis of alternative splicing isoforms annotated in Swissprot published in Nucleic Acids Research in 2008 [14]. The analysis has been a joint work with Gergely Csaba.

To date, the structures of less than 10 isoforms are available in the PDB and known structural implications of splicing events on some of those proteins have been reviewed in [148]. Those examples include a protein called Piccolo and the surprisingly drastic rearrangement of its C2-domain altered by a short insert of nine residues [54]. Despite those few examples only little is known from experimental data about how and to what extend alternative splicing alters protein structures. Due to this lack of knowledge from biological data, several recent studies ([134], [161], [171]) have mapped alternative splicing events onto predicted protein structures and have analyzed features of the regions being affected. While they could link some structural properties like protein disorder [134] to a group of splicing events, the effects on many isoforms appear to be non-trivial. Based on this surprising complexity of alternative splicing on the proteome level, the most recent study by Tress et al. [161] even comes to the converse conclusion that ”it seems unlikely that the spectrum of conventional enzymatic or structural functions can be substantially extended through alternative splicing”.

Here we present the results from comprehensively mapping all splice variants annotated in Swissprot [175] onto known protein structures from the PDB leading to almost 500 isoforms annotated to more than 350 Swissprot entries whose structures can be modeled with a very high accuracy (see Materials and Methods section). While about half of the events fall into variable regions of protein structures or affect complete domains of multi-domain proteins, the effects on the other half appear to be non-trivial since they affect structured and well conserved regions of the corresponding protein family. The large number of such non-trivial events, which are also found to be conserved among different species, can be explained in two ways: Firstly, non-trivial splicing events are non-functional on the mRNA or protein level leading to nonsense-mediated mRNA decay or unstructured proteins which are degraded after translation. This would indeed allow only few exons of an organism to be alternatively spliced, clearly questioning the importance of splicing due to its complexity on the proteome level. Secondly, they may represent evidence that non-trivial splice events can produce functional isoforms where the absence of highly conserved parts of the structure might even allow for new structural and new functional properties of the isoform.

7.1 Hypotheses and major concepts

Here we provide evidence for the second hypothesis. Having mapped a large set of splicing events onto protein structures we first explore the natural variation of the corresponding protein structure family, namely the ”evolutionary isoforms” of the respective protein, which allow us to explain about 50% of the splicing events. We will show that the analysis of known ”evolutionary isoforms” can be a helpful tool to predict the outcome of splicing events even in cases which appear to be non-trivial at the first glance. The other half of the isoforms defines the set of non-trivial splicing events which can not be explained by the observed variation in the respective protein family and which we examined in more detail.

Based on evolutionary considerations of known fold changing events [59] we group non-trivial events into eight different categories comprising different effects to be expected on the structure level. We then show that an extensive search of the biological literature provides clear evidence of stable protein products originating from such isoforms as well as evidence for a well defined functional role of those proteins in the cell. The existence of such isoforms can sharpen our understanding of a protein’s tolerance against major structural changes and additionally largely increases the importance of alternative splicing for generating functional and structural diversity. We will therefore review the function as well as the structural complexity of some of those isoforms in detail.

Based on those findings, we try to explain the tolerance of such isoforms against the splicing events and will show that it can be linked to the evolutionary history of the corresponding fold. In more detail we can use alternative splicing data to add additional evidence to specific, existing hypotheses on the evolution of specific fold classes. Moreover, splicing events which can not be explained by the variability observed in one fold may be explained by members from different folds. Structures of splice isoforms may therefore be located in different regions of the protein fold space indicating previously undiscovered and novel links between different protein topologies which can be explored by alternative splicing events. We find evidence that such links between different folds in the sequence-structure space may indeed exist, and, for the first time, suggest a simple and common genetic mechanism, namely alternative splicing, for nature to explore them in vivo.

7.2 Methods

In the following the methods and data sources used for our study are described in detail. The methods are summarized in Figure 7.1.

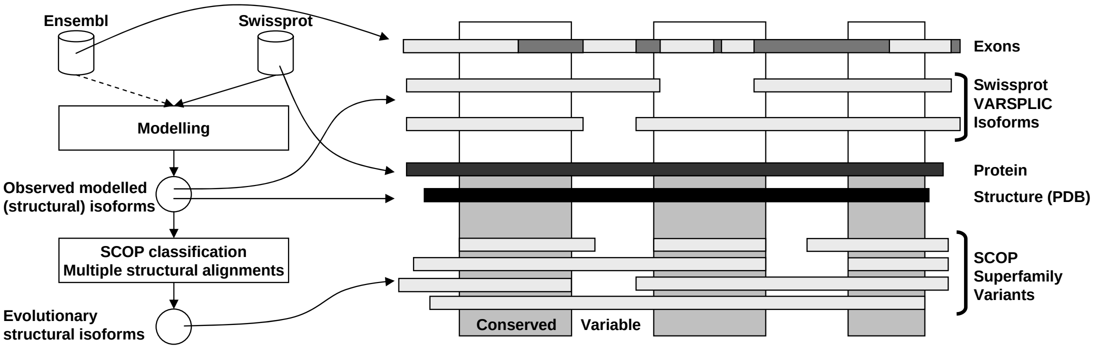

Ensembl
Swissprot
Modelling
Observed modelled (structural) isoforms
SCOP classification
Multiple structural alignments
Evolutionary structural isoforms
Exons
Swissprot VARSPLIC Isoforms
Protein
Structure (PDB)
SCOP Superfamily Variants
Conserved
Variable

Figure 7.1: The figure shows the main modeling steps of our approach and the resulting gene products. Validated isoforms annotated in Swissprot are conservatively mapped onto 3D structures from the PDB resulting in structural models for the Swissprot isoforms. The SCOP protein classification and high quality multiple structural alignments with variability defines 'Conserved' and 'Variable' regions and thereby evolutionary structural isoforms observed in superfamilies. This allows assessing whether the validated isoforms affect only variable regions or structural cores of proteins. (*Figure taken from [14]*)

7.2.1 Alternative splicing and literature data

The data for alternatively spliced proteins used in this work was obtained from the Swissprot protein database (September 2006), which annotates splicing events for 9135 out of 231434 protein entries. Those 9135 entries harbor 20845 alternative splicing events where 56.6% of the events are deletion events and the other 43.3% of the events represent replacements. In 22.2% of the replacements the original sequence is shorter than the replacement sequence (insertions), in 27.7% the replacement sequence is shorter than the original sequence (deletions) while in 50.1% of the cases the original sequence and the replacement sequence are of the same length. Literature assignments to different isoforms are also provided by Swissprot. We have examined them manually for evidence for the experimental proof of a stable protein product as well as experimental validation of its function.

7.2.2 Protein structure assignment

To obtain protein structure data we ran BLAST [3] against all proteins in the PDB (August 2006) for all Swissprot proteins with annotated splicing events. For each alternatively spliced protein we then used free-shift alignment [111] (PAM250 matrix, gap open: 12, gap extend: 1) to compute full length sequence-structure alignments of the alternatively spliced Swissprot entries with their respective homologs identified by BLAST. From all alternatively spliced Swissprot proteins only those whose structure could be modeled with a very high sequence identity of at least 60% between template and target and whose sequence is covered to at least 75% by protein structure are used for further analysis.

7.2.3 Assignment of protein structures to families

The protein structures used to model the Swissprot proteins have been assigned to their corresponding SCOP families as defined in SCOP version 1.71 (December 2006). All Swissprot entries that are modeled with structures not yet classified in the SCOP version 1.71 were assigned to their respective protein families using Vorolign [15]. The final dataset contains 367 Swissprot proteins with 488 annotated isoforms which are classified into 166 different families, 134 superfamilies and 119 folds with respect to the SCOP hierarchy.

7.2.4 Multiple structure alignments and evolutionary ”isoforms”

Multiple structure alignments were computed from multiple structure superpositions with STACCATO [141] which has been shown to compute accurate alignments with respect to both, sequence and structure. In order to guarantee enough variability within the set of evolutionary related protein structures, we use proteins from the same SCOP superfamily. On this level of the SCOP hierarchy, protein structures exhibit enough structural variance to allow the definition of conserved and variable regions without overestimating structure conservation due to too similar proteins. Each set must contain at least three members and their structural similarity is measured by the TM-Score [188]. Proteins in a set have to be similar enough (indicated by a pairwise TM-Score of larger than 0.4) while still showing structural variability (TM-Score smaller than 0.8). Given a multiple structure alignment, a conserved region of a SCOP superfamily is defined as a block of at least 10 residues which are conserved among all members of the superfamily. Each protein in a block may contain two gaps at maximum to account for some small variability within blocks. All regions outside of the conserved blocks are defined as variable regions. The proteins in the set define what we call evolutionary ”isoforms” which display insertions, deletions and substitutions and define the set of evolutionary events that are likely to be tolerable for a protein structure.

7.2.5 Alternative splicing and alternative structural models

In order to suggest alternative structures for non-trivial alternative splicing isoforms we applied the splicing event onto the structure (e.g. removed the structural parts belonging to a skipped exon). We then searched for reliable structural superpositions of the resulting structure model against all known folds (according to the SCOP classification). Such a search resulted in a number of structurally similar proteins from SCOP folds different than the proteins own fold. The soundness of such superpositions was measured by different criteria. First, as argued by Zhang and Skolnick [188] a TM-Score larger than 0.4 is a clear evidence for a structural similarity (criterion 1). Since we believe this criterion is not strict enough to claim such remote structural similarities we also used more stringent criteria. Therefore, we filtered the superpositions to those superposing at least 80% of the spliced structure and 60% of its secondary structure elements and additionally examined the resulting superpositions manually for the conservation of core secondary structure elements with the correct connectivity and topology (criterion 2).

	Coil	α	β(p)	β(i)	$\alpha\beta$(p)	$\alpha\beta$(i)	Repeat	Large	Total
Region	14	50	29	25	49	35	21	37	260
Isoform confirmed	4	15	2	9	6	3	1	3	43
Function described	2	7	1	5	5	3	1	2	26

Table 7.1: This table displays the distribution of non-trivial isoforms in the eight categories defined based on evolutionary considerations with respect to different features. (p) and (i) indicate the position of the corresponding β-strands either at internal or peripheral positions of the sheet. The *Region* row displays the number isoforms which affect conserved regions of the corresponding superfamily. The *Isoform confirmed* row displays isoforms which have been confirmed in the literature on the protein level, while the *Function described* row references isoforms in the different categories which have been described in the literature to perform a well-defined function.

7.3 Results

7.3.1 Structural complexity of functional isoforms

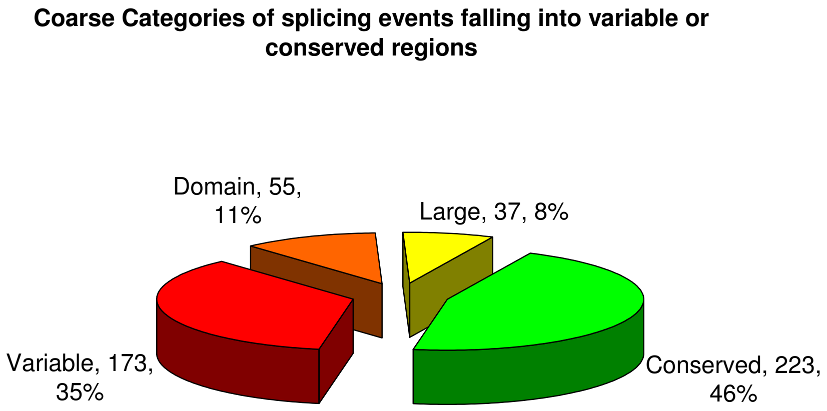


Figure 7.2: Distribution of 488 splicing events in the four major categories. 35% of the events fall into variable regions of the corresponding superfamily while 11% affect complete domains of multi-domain proteins. 8% of the isoforms affect larger regions (more than 50% of the structure) while 46% affect conserved regions of their corresponding superfamily which are present in all superfamily members. (*Figure taken from [14]*)

Our study is based on 367 Swissprot proteins and 488 additional splicing isoforms which can be modeled on the structure level with a very high accuracy. As shown in Figure 7.2 about 50% of the events fall into variable, often terminal, regions of the corresponding protein superfamily (”evolutionary isoforms”) or affect complete domains. The other half (260 events) of the events are harder to explain since they affect regions conserved in all superfamily members including core secondary structure elements as well as highly conserved residues.

Based on evolutionary considerations about possibly fold changing events proposed by Grishin [59] we defined eight categories describing different types of non-trivial splicing events. Due to their hydrogen bonding patterns β-strands being located at the edge of a larger β-sheet are known to be more variable than internal β-strands. Therefore, two categories describe events affecting peripheral or internal β-strands. Similarly, in proteins belonging to α-β fold classes, often $\alpha\beta$-secondary structure motifs tend to be affected and, accordingly, we defined two classes comprising peripheral and internal $\alpha\beta$-motifs. Additional categories contain conserved coil regions, conserved helices, large-scale events (affecting more than 50% of the structure) as well as events affecting repetitive protein structure families whose repeat number is known to vary in evolution. Table 7.1 shows the distribution of the splicing events in the different categories.

The biological literature annotated to the isoforms in Swissprot provides a valuable source of information. While Swissprot annotates most reference proteins only if they have been verified experimentally, this must not be the case for their corresponding isoforms which may originate from large scale EST or cDNA experiments. Indeed, most isoforms have only been experimentally verified on the mRNA level, while the protein products of the isoform were not investigated in the corresponding study. In even fewer cases, studies try to characterize the novel function of the identified isoforms. Nevertheless, out of the 260 non-trivial isoforms, 43 (17%) have been experimentally validated on the protein level and for 26 isoforms (10%) the function of the spliced variant has been described. Surprisingly, and in contrast to previous findings [161], we find literature evidence for functionally important and well characterized isoforms in all of our eight categories (see Table 7.1) indicating that even large scale events may lead to functional and interesting protein products. A complete list of all literature references and isoforms is given in the supplementary material published in correspondence with [14]. In the following we will shortly review the function of some interesting isoforms from different categories.

Alternative splicing of a terminal region of a protein Many splicing events in our dataset change amino- or carboxy-terminal parts of a protein structure which are found to be more variable and differ significantly among the members of a protein family. One of the examples, where the splicing event can be explained by the variability observed in the corresponding protein family is found in the human protein p38α (Q16539), a member of the mitogen-activated kinase (MAPK) family. Those proteins are integral parts of several signal transduction pathways and known to play important roles e.g. in the stress response of the cell. For p38α several splice variants are annotated in public databases and among the well studied splice variants are two proteins known as Mxi2 (Q16539-3) and Exip (Q16539-4) which both differ in large parts of their carboxy-terminal ends compared to p38α. Also, a similar splicing event is annotated for a homologous protein in mouse (P47811-2). The splicing event annotated for Exip [150] removes 46 residues from the protein structure, in addition 52 residues differ in their sequence due to a frameshift introduced by the event (see Figure 7.3a). It results in the loss of a well conserved interaction domain used to interact with upstream kinases and downstream substrates. This leads to the fact that the protein is not targeted by MKK6 anymore. Expression of the isoform in the cell leads to an earlier onset of apoptosis and seems to target signal transduction pathways which are different from those targeted by p38α [150]. In the example of Exip, the splicing event in-

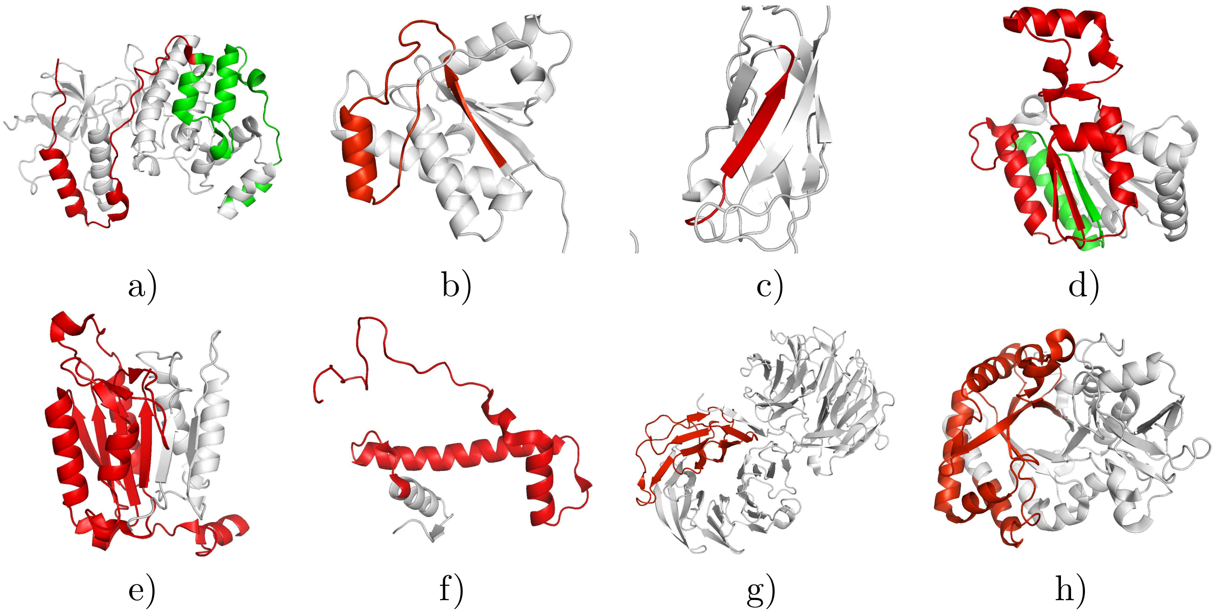


Figure 7.3: Visualization of alternative splicing events on the structure level. Substitutions are colored in green while deletions are colored in red. All figures have been created using PyMOL (http://www.pymol.org). a) shows the removal of the carboxy-terminal part from MK14_HUMAN (Q16539-4, pdb: 1zzlA), b) the removal of one external strand and helix motif in PPAC_HUMAN (P24666-3, pdb: 5pnt), c) the removal of an internal strand in TF65-HUMAN (Q04206-3, pdb 1nfi), d) the removal of a large part of the protein from TIP30_HUMAN (Q9BUP3-2, pdb: 2bkaA), e) the removal of several strands in CASP9_HUMAN (P55211-2, pdb: 1nw9B), f) the removal of a large part of H4_RAT (P62804-2, pdb: 2f8nF) leading to the osteogenic growth peptide, g) the removal of β-propeller blades from WDR1_CAEEL (Q11176-2: pdb: 1pevA) as well as h) the removal of one half of a TIM-barrel structure in CHIA_HUMAN (Q9BZP6-3, pdb: 1vf8A). A comprehensive database of alternative splicing events mapped onto protein structures can be found at http://www.bio.ifi.lmu.de/ProSAS/NARSupplement.html. (*Figure taken from [14]*)

deed targets a more variable part of the protein family (SCOP superfamily d.144.1). Structures, which lack the carboxy-terminal part are known to fold into stable conformations.

Alternative splicing at the edges of β-sheets Alternative splicing events that involve β-strands naturally lead to the disruption of important hydrogen bonds in the protein structure. Nevertheless, such events are typical in the evolution of globular proteins if they affect peripheral β-strands of a larger β-sheet [59]. Therefore, these events might be tolerable by a protein structure, even if they affect conserved regions of the protein's family. One such event is the removal of one peripheral $\alpha\beta$-motif from LMPTP (P24666), a tyrosine phosphatase. The protein is known to be expressed in three different isoforms, all of which differ in a 38 amino acid long part corresponding to one $\alpha\beta$-motif. The β-strand represents a peripheral strand of a β-sheet consisting of 4 strands in total. While the original sequence is replaced by another 38 residues in isoform 2, the corresponding part is removed in isoform 3 (LMPTP-C, P24666-3) (see also Figure 7.3b). Detailed analysis of LMPTP-C [154] shows that the protein is lacking phosphatase activity and can also not be phosphorylated by Lck kinase indicating that the active center has been tackled by the splicing event. When being co-expressed with isoform 2, LMPTP-C is shown to act as an antagonist to its native variant. The proposed mechanism [154] is that LMPTP-C competitively associates with the cellular substrates or regulators of its native counterparts and thereby blocks dephosphorylation of their targets. LMPTP-C represents an example how a splicing event changes a protein's function by removing the active center of the protein. While this goes hand in hand with a loss of its native function, the isoform is still able to mimic features of the native structure which allows it act as antagonist of LMPTP.

Alternative splicing of internal strands of β-sheets As shown above, the deletion of peripheral $\alpha\beta$-motif can result in a functional protein revealing an interesting mechanism for the regulation of enzyme activity. The deletion of internal strands from a β-sheet or a β-barrel appears to be more problematic since this results in the loss of hydrogen bonds on both sides of the strand and requires the formation of several new ones to retain the native-like structure of the protein. Nevertheless, there are known examples for strand deletion events that occurred in structure evolution as discussed in [59]. The p65 (Q04206) subunit of the NF-κB transcriptional activator has one splice variant (Q04206-3) which exhibits such a removal event [135] as shown in Figure 7.3c, where nine residues, corresponding to one internal β-strand, are removed. Again, for a homologous protein in mouse (Q04207-2) the same splicing event is annotated. While the splice variant lost its capability to bind to p50, the second subunit of the NF-κB complex, it can form weak heterodimers with the native isoform of p65. Those heterodimers are found to be greatly reduced in their ability to bind DNA. This finding allows for two possible conclusions. Either, co-expression of the isoform and the native protein negatively regulates the NF-κB function, again revealing a pattern where the inactivation of a protein feature may act as a antagonist for the native protein. Or, in case that the isoform is still able to bind IκB (the inhibitor of the NF-κB complex), it may act as a regulatory ”sink” binding excess IκB and allowing p65 or the p65/p50 complex to enter the nucleus [135].

Alternative splicing may affect large and conserved regions of the protein structure In the following, we give evidence for the fact that alternative splicing events may affect large and conserved regions of a protein structure and still can result in an isoform with unique functional features. CC3 (Q9BUP3) is known to be a metastatis suppressor inducing apoptosis in human cells which is not expressed in highly metastatic lines of ”small cell lung carcinoma”. A variant, called TC3 (Q9BUP3-2) [173], undergoes an alternative splicing event which removes 107 residues from its carboxy-terminal end and further replaces 21 residues at the new terminus which do not share any sequence similarity with the original sequence. As shown in Figure 7.3d the splicing event affects more than 50% of the protein structure. It removes two peripheral β-strands from a β-sheet consisting of seven strands as well as several additional helices and strands which are not involved in the formation of the core $\alpha\beta\alpha$-fold. Strikingly, TC3 has, in contrast to its native variant CC3, an anti-apoptotic function which seems to be located in its unique C-terminal part. Even though the protein lost several conserved elements of its fold it seems to be able to fold into a stable, functional isoform (see also Figure 7.5, rightmost column). It is shown to be short-lived due to a degradation signal located in the new carboxy-terminal end of the protein [173] which possibly represents another physiological feature. A second example which exhibits a similar splicing event that removes an even larger part from the structure is an isoform of caspase 9 (P55211-2) (see Figure 7.3e). The isoform, named caspase 9b, is again shown to function as an endogenous apoptosis inhibitory molecule [146]. The isoforms of CC3 and caspase 9 reveal a surprising tolerance of $\alpha\beta\alpha$-fold proteins to large scale aberration events. This tolerance might originate in the evolutionary history of proteins of this fold class as they might have evolved by successively adding $\alpha\beta$-motifs to the edges of the core sheets. This might result in the fact that they can also be removed from the structure without loosing capability to fold into a stable conformation.

Alternative splicing events resulting in small peptides Another, rather drastic, example for a large scale splicing event is observed for histone H4 in Rat [9] and seems to occur in other mammals as well. The isoform is likely a result of alternative translational initiation via an alternative AUG site which leads to the production of a small peptide of less than 20 amino acids (see also Figure 7.3f). This peptide, called Osteogenic Growth Peptide (OGP), was initially isolated from bone marrow and later on was shown to be transcribed from the H4 gene locus (or one of its copies) [9]. In vivo, OGP increases bone mass and stimulates blood and bone marrow cellularity and therefore plays an important role in the control of cell proliferation. This example shows that even very large splicing events leading to small peptides may play important and functional roles in the cell.

7.3.2 Interpretation of splicing events using evolutionary isoforms

In the previous examples we have mainly focused on the analysis of non-trivial isoforms and their functional implications. Most of the isoforms where variable parts of the superfamily are affected represent trivial alterations often at the C- or N-terminal ends of a structure. But there are also cases which appear to be non-trivial at the first glance but whose effects can be understood in the light of protein structure evolution. One interesting example, namely the splicing of one half of the MHC antigen recognition domain of the Hemochromatosis protein (HFE_HUMAN) annotated for isoform 2 in Swissprot, is shown in Figure 7.4. At the first glance and without additional knowledge this removal event appears to be deleterious for the domain since one half of the central β-sheet (shown in red in Figure 7.4a) is removed by the splicing event. Strikingly, members of the same family (shown in Figure 7.4b) contain only one half of the domain and the part is consequently annotated as variable in the superfamily alignment. As shown in Figure 7.4c proteins which contain only one half of the domain form dimers in order to resemble the complete antigen recognition domain. This feature can also be proposed for the respective isoform and can explain its tollerance against the annotated splicing event.

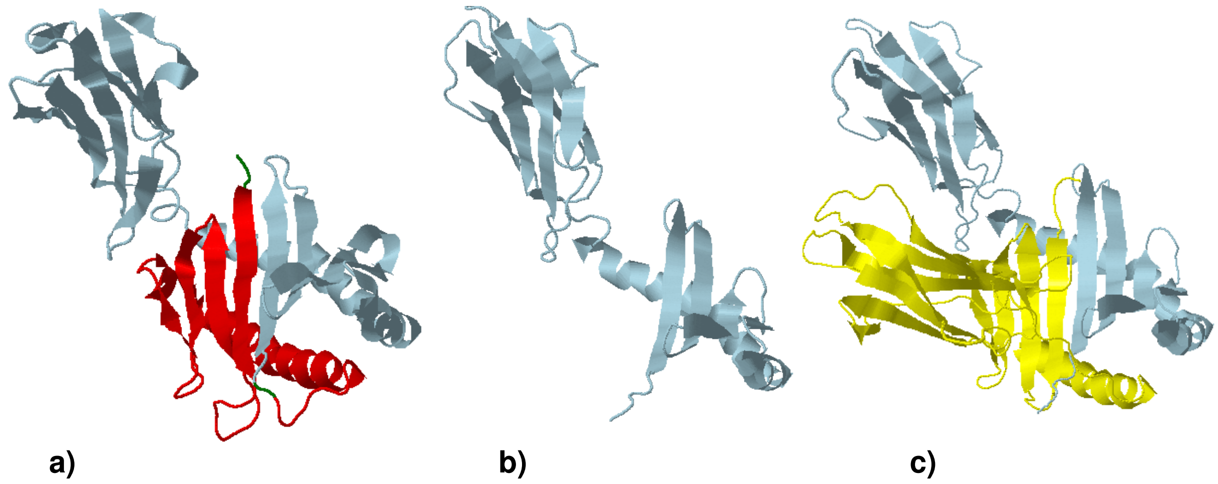


Figure 7.4: Understanding the effects of splicing events via the analysis of evolutionary isoforms. a) Splicing event annotated for the MHC antigen recognition domain of Hemochromatosis protein (HFE_HUMAN, Isoform 2, PDB: 1de4G, SCOP: d.19.1.1). b) evolutionary isoform from the same family (PDB: 1aqdH, SCOP: d.19.1.1) which is equal to the expected structure of the splicing isoform c) which is known to form dimers of two chains (G and H, shown in blue and yellow) to resemble the complete antigen recognition domain.

7.3.3 Supporting hypotheses on fold evolution via alternative splicing data

So far we have discussed the large complexity of alternatively spliced isoforms on the protein structure level and the usefulness of knowledge on protein structure evolution, i.e. ”evolutionary isoforms”, in characterizing those events. In the following we will show that data on alternative splicing can also be used to analyze protein structure evolution. As a first application of splicing data in this context we will show the correlation between existing hypotheses on the evolution two well known protein families, namely TIM-Barrels and β-propellers and alternative splicing data.

The origin of TIM-Barrels from Half-Barrels For a number of protein structures and protein structure families it is well known that they resulted from ancient gene duplication and/or fusion events. Such duplication events are not always obvious from sequence data since the two subdomains have possibly already evolved to an extent where sequence similarity blurs to random. A well studied and recurrent motif in protein structures is the $(\alpha/\beta)8$-barrel family (”TIM-barrel”, SCOP fold c.1). Proteins of this family adopt a large variety of different functions and based on sequence and structure analysis it has been proposed [92] for some members of that family that they originated from a gene duplication and fusion event of two ancestral half-barrel proteins. Those ancient half-barrels probably formed a homodimer consisting of two identical half-barrels [106]. Our analysis now provides additional support for this hypothesis since it reveals two splicing isoforms (Q9BZP6-3 from CHIA HUMAN, and P27934-2 from AMY3E ORYSA) where one half of the barrel is removed by a large-scale removal event (see Figure 7.3h). The isoform of the human chitinase gene (Q9BZP6-3) has been described by Saito et al. [136] to be specifically expressed in lung though experimental validation of the existence of the stable protein product is lacking. In comparison to its native isoform the protein lacks a secretory signal sequence leading to the conclusion that it might be present in the cytoplasm instead of being secreted. It also lacks the amino-terminal active site essential for chitinase activity. So far, we have no experimental validation for a functional gene product and a stable protein resulting from those splicing events. Nevertheless, based on the proposed evolutionary mechanism of fusing two half-barrels by an ancient gene duplication and fusion event, the splicing isoforms possibly form a (homo-)dimer to reassemble the complete barrel. The possibility to express proteins of the TIM-Barrel family as half-barrels might offer an increased functional variability by combining half-barrels containing different functional sites in heterodimeric complexes.

Common evolutionary origin of different β-propeller folds Splicing events observed for β-propellers (one example is shown in Figure 7.3g) are especially also interesting to elucidate the evolutionary history of those proteins. Propellers consisting of different numbers of blades have been sorted into different folds by the authors of the SCOP and CATH which displays the low level of sequence identity between different families comprised of different numbers of repeats as well as their diverse functions. Nevertheless, it has only recently been shown by Chaudhuri, Söding and Lupas [30] that despite low sequence identities there are indications for a common evolutionary origin of all β-propeller folds.

Splicing data further contributes evidence for this hypothesis. Splicing frequently generates transcripts with different numbers of blades (see Chapter 8 for a detailed analysis of splicing events in this family) and different β-propeller folds seem to be explored from one genomic locus via alternative splicing. This data supports the common evolutionary origin proposed by Chaudhuri et al. and displays the usefulness of splicing data in the analysis of protein structure evolution.

7.3.4 Alternative splicing: a mechanism to explore the protein fold space?

So far we have shown that the changes imposed on protein structures by alternative splicing can often not be explained by the variation within the own fold (corresponding to non-trivial isoforms) and we suggested a connection between splicing events and protein structure evolution. In the following we will describe our approach to identify novel connections between different folds in the protein fold space which lead to testable hypotheses on the structure of non-trivial isoforms.

For 225 non-trivial isoforms (excluding repetitive and conserved coil cases as these splice events will presumably not result in a different fold) we searched for similar structures as described in the Materials and Methods section 7.2. Applying the TM-Score criterion (TM-Score > 0.4) [188] alone we find for 139 (66%) isoform structures resulting from splicing events a similar structure from a different fold. Applying the more stringent criterion (secondary structure and isoform coverage) results in 49 isoforms (47 of which have a TM-Score larger than 0.4). For these, we superposed the spliced structure with the target fold and visually inspected the super-

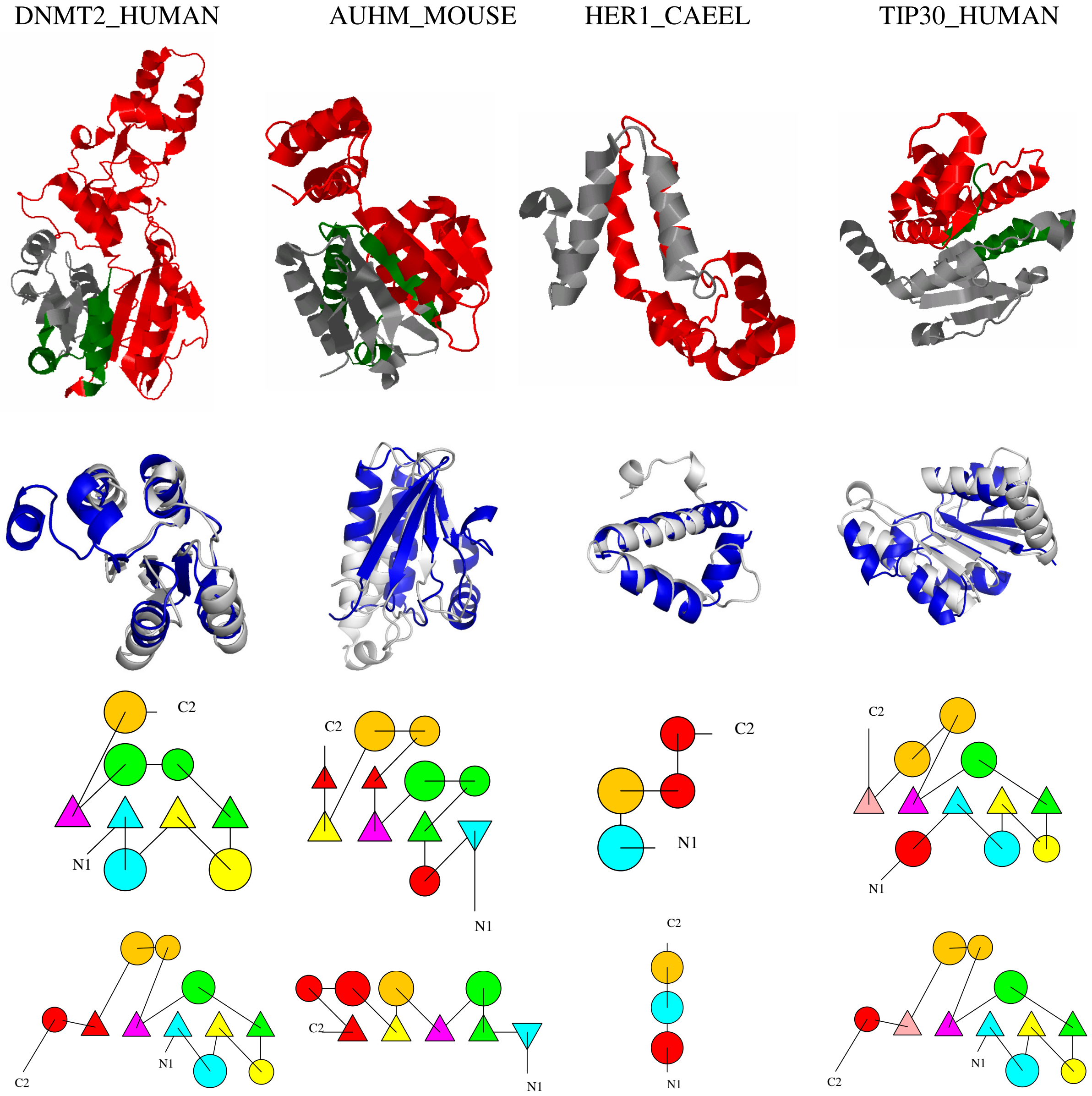


Figure 7.5: This figure shows four examples for possibly fold changing splicing events by non-trivial splicing events (i.e. those which cannot be accommodated in the native structure) and by superposing the spliced structure to a different SCOP fold. The examples can also be explored interactively following this link http://www.bio.ifi.lmu.de/ProSAS/NARSupplement.html. Each column represents one example which are discussed in the text. All structures have been visualized using Jmol (http://www.jmol.org) and PyMol (http://www.pymol.org). (*Figure taken from [14]*)

positions for conservation of core secondary structure elements as well as their connectivity, i.e. the topology of the core elements. We observe 21 (10%) highly confident superpositions, i.e. models for the spliced structures having a fold different from the one of the non-spliced protein. Thus, these different folds are probable structural models for the isoform, which could explain the drastic changes caused by the splicing event (four examples are shown and briefly discussed in Figure 7.5). In the first row of this figure the splicing event is visualized on the native protein structure of the Swissprot protein. In the second row the superposition of the spliced protein with the corresponding protein from a different fold is shown. Rows three and four display TOPS [105] diagrams of the spliced protein (row three) and the protein belonging to the different fold (row four). Corresponding secondary structure elements are colored the same, elements missing in the other protein are colored in red. Sometimes corresponding helices are split up which frequently results from breaks in the DSSP assignments. From left to right the following examples are shown:

Column 1: DNMT2_HUMAN (O14717-6, Astral: d1g55a_, SCOP: c.66.1.26). The spliced protein superposes very well (TM-Score: 0.68) with d1gsoa2 (SCOP: c.30.1.1). Topologically the proteins are very similar, except for a very short strand (length 2) - helix (length 4) motif at the C-terminal end of d1gsoa2.
Column 2: AUHM_MOUSE (Q9JLZ3-2, Astral: d1hzda_, SCOP: c.14.1.3) which superposes well (TM-Score: 0.56) with d1vc1a_ (SCOP: c.13.2.1). Topologically the proteins are similar, except for two small strands (both of length 2) and one short helix (length 3) missing in d1vc1a_. Additionally, the C-Terminal part of d1vc1a_ has an additional, short helix-strand motif.
Column 3: HER1_CAEEL (P34704-2, Astral: d1szha_, SCOP: a.226.1.1) superposed with d1ni8a_ (SCOP: a.155.1.1, TM-Score 0.49). Only a small fragment (helix-turn-helix-motif) is left over by the splicing event. The two TOPS diagrams are similar with the two main helices being preserved while short helical parts are missing in either of the two proteins. Interestingly, d1ni8a_ is described to contribute to DNA binding after dimerization (35) which might also be the way how the isoform resulting from the HER1 splicing event is stabilized.
Column 4: TIP30_HUMAN (Q9BUP3-2, PDB: 2bka, SCOP: c.2.1.2) which again superposes well with d1gsoa2 (see also DNMT2_HUMAN) from SCOP fold c.30.1.1 (TM-Score: 0.54). Topologically the two proteins are very similar according to TOPS except for two helices missing at the C- and N-Terminal ends. The function of the isoform is discussed in the text (isoform TC3).

Of course, proteins resulting from splicing events might not be able to fold at all into a stable structure and often this will be the case. In other cases the structure might be stable but will form a novel fold (so far not solved and deposited in the PDB). Such drastic cases also will not be detectable by our approach as we explicitly search for similar protein structures assuming that the structure of the isoform remains unchanged except for the part being spliced.

In rare cases, the modified structure might be similar to a known fold different from the native one. In the latter case we would observe links between different folds by defined genetic changes (alternative splicing) transforming one stable 3D structure into a different stable 3D structure. Despite many attempts and research on structure classifications and structural descriptions and features, which led to the well known structural resources such as SCOP and CATH, reliable and traceable links between fold classes are very rare. This is even more the case for evolutionary explanations of the observed similarities and events. Here we do not only observe a considerable number of such transformation events but also provide a simple genetic mechanism explaining them as all the events correspond to known observed transcripts.

7.4 Discussion

Our study reveals a large number of functionally important, alternatively spliced proteins that harbor non-trivial splicing events and hints to a high degree of plasticity and a large tolerance of protein structures and folds against major rearrangements. The possibility to express the antagonist of a protein as an isoform of the native variant represents an intuitive mechanism to increase the functional complexity of an organism by alternative splicing and has been discussed by several studies before. The structural explanation for this mechanism may be grounded in the removal of highly conserved parts, which are essential for the function of the native variant. If the isoform is still able to fold into a native-like structure, which often seems to be the case, it can mimic native structural features and e.g. interact with native interaction partners without processing them further. Thus, alternative splicing immediately provides a mechanism for turning an activator into an effective inhibitor via a simple, possibly regulated, genetic mechanism.

The sequence-structure protein space tries to link different folds by appropriate similarities and differences but examples for fold transitions are rare and typically difficult to explain biologically. Our study provides examples for such links and explains them with a simple and common genetic mechanism. Thus, alternative splicing may be a new approach to chart the protein space and gain insights into mechanism of protein structure evolution. Future work will be on exploring the fold space as well as the changes that occur within and between folds in the context of alternative splicing in more detail. Therefore, structural analysis of alternative splicing events may help to identify common paths of protein fold evolution similar to the events discussed by Grishin [59] and to describe events that may be tolerated within protein families. This knowledge may have interesting applications in protein design and protein structure analysis.

Without experimental proof we can currently only speculate about the structures of isoforms resulting from non-trivial splicing events. Several facts indicate that at least some of those isoforms could have a well defined structure. They perform a well defined function in the cell and are able to mimic features of their native counterparts. Additionally, the identification of fold transitions exemplifies that they could adopt structures from different folds. Nevertheless they might also be unstructured or fold into yet unknown conformations not detectable by our approach. Therefore, this study will provide interesting starting points for experimentalists trying to gain a deeper understanding of the non-trivial alterations of protein structure produced by alternative splicing and will lead to new insights into protein structure stability and the principles of protein fold evolution.

We also expect recently established experimental techniques like exon-level micro arrays to contribute significantly to our understanding of the functional and, in the context of this study more important, the structural effects of alternative splicing. Novel NMR-techniques [48] will contribute to validate or falsify the importance of alternative splicing for the functional diversity of complex organisms. Another source of protein-level confirmations of splicing isoforms can be provided by mass spectrometry data [1] measuring complete proteomes. In Chapter 12 we will therefore carry out a comprehensive analysis of mass spectrometry confirmed isoforms identified in Human, Mouse and Drosophila.

As in principle alternative splicing is a mechanism to produce a combinatorial number of transcripts, even a small percentage of stable structures implies a very large number of new protein variants. Thus, we believe that evolution makes use of alternative splicing to produce structural and functional diversity and this diversity is due to the large structural plasticity of proteins.

Chapter 8

Alternative Splicing and Protein Repeats

One of the driving forces in the evolution of novel protein functions is (segmental) gene duplication. While gene duplication can lead to two distinct genes which may be regulated differently and adopt different functions, the duplication and recombination within a gene leads to proteins which contain so-called tandem repeats of the same element or domain. It has been shown that at least 14% of all proteins contain repetitive elements and, especially in multi-cellular organisms, the number of genes containing repeats appears to be increased compared to organisms with smaller genomes and proteomes [101]. Repeat evolution can lead to highly complex patterns of repetitive motifs [21].

Proteins containing repeats have many interesting functions which often serve the special needs of multi-cellular organisms. They are involved in the control of gene expression, the formation of protein-protein interactions and multi-protein complex assemblies as well as cell signaling and cell adhesion. Therefore, it has been suggested that the larger number of repetitive elements is a consequence of an increased proteome complexity resulting in a fast development of novel protein functions under the constraint of longer generation times [101].

We have already introduced alternative splicing as an additional mechanism contributing to functional diversity [148] in higher eukaryotes in previous chapters. Further, in the last chapter (Chapter 7) we have proposed a connection between structural and functional protein diversity generated in the course of protein (structure) evolution and the diversity that originates from alternative splicing. Here we test this hypothesis for proteins containing repetitive elements [17].

Repetitive proteins are especially attractive targets for evolution and alternative splicing with respect to generating structural and functional diversity for several reasons. First, their specific function often origins from the simple and modular combination of repetitive motifs leading to proteins with different DNA or protein binding properties. Second, the structural effects of a repeat duplication or deletion are relatively easy to explain. Protein structures containing repeats are highly tolerant against changes in the repeat number which is also supported by many known protein structures of repetitive protein families in the PDB which show a large variation in the number of repeats.

A simple evolutionary model (Figure 8.1a) of increasing the number of repeats in a protein by duplication in combination with the possibility to alter the repeat number via alternative splicing after duplication events has the potential to produce a combinatorial number of functionally different protein isoforms as shown in Figure 8.1b. The specific expression of an isoform can be altered by small variations and mutations in the regulatory signals controlling the splice pattern of the gene [172] allowing for a fast evolution of novel functions. Therefore, repeat expansion in cooperation with alternative splicing may have provided higher organisms an evolutionary advantage in driving functional diversity of their proteomes.

In the following we will first discuss the content of repetitive motifs in the Swissprot database and the human genome identified by PFAM patterns. We will then analyze known alternative splicing events annotated in Ensembl and in Swissprot in order to find any preferences for splicing events to affect repetitive motifs. Finally, we will discuss functional, structural and evolutionary implications of alternative splicing for several highly abundant repeat classes which are found to be affected by splicing events significantly more often than expected in our analysis.

8.1 Methods

We analyzed splicing events observed in two, largely independent datasets. The first set corresponds to splicing events annotated for human in the Ensembl database [73]. The second dataset corresponds to all splicing events from various organisms annotated in the Swissprot database [175].

8.1.1 Genomic data and annotated splice variants

Genomic data for human has been obtained from the Ensembl database (Release 48, December 2007) which annotates alternative splicing events observed in EST data and was then processed by our ProSAS database [19] pipeline. The dataset contains 27.316 human genes with 54.640 annotated transcripts. In the following we will refer to this dataset as ”human dataset”.

For each gene annotated with more than one transcript we defined the reference transcript as the protein annotated as major isoform in Swissprot. If no Swissprot annotation but a high quality structure prediction is available we use the transcript which best fits the structure annotation in terms of coverage and sequence identity. All splicing events for a gene were then defined and analyzed with respect to this reference isoform.

8.1.2 Splicing events annotated in Swissprot

Additionally to the human dataset, we also analyzed splicing events obtained from the Swissprot resource (Release 54, December 2007). This set will be called ”Swissprot dataset” in the following. Due to the fact that Swissprot is manually curated, splicing events are annotated in a more conservative way than in Ensembl. 14.086 Swissprot proteins from various organisms have annotated VARSPLIC information on alternative isoforms. In total, the Swissprot dataset contains 22.724 splicing isoforms (on average 1.6 additional splice isoforms per entry).

8.1.3 Repeat assignment and data collection

Following other studies ([21], [130]) on protein repeats we searched for repetitive protein elements using patterns from the PFAM [10] database. Matches of those patterns against the human genome and Swissprot resulting from InterProScan [126] were obtained from the SIMAP database [127].

A pattern is said to match a specific position of a transcript if the E-value is smaller than 0.01. We define patterns as ”repetitive” if it occurs on average more than two times in proteins containing the pattern. The average value of two is chosen to distinguish truly repeated patterns from simple domain duplications. Repetitive patterns may overlap to 10% at maximum in order to be defined as distinct pattern matches. A pattern is said to be significantly affected by an alternative splicing event if more than 75% of the residues matched by the pattern are affected. Please note that we only analyze events leading to the removal of a part of the reference transcript contributed by e.g. exon skipping or alternative 5’ or 3’ splice sites.

For every pattern we analyze how many entries (reference transcripts or Swissprot entries) contain the pattern and how often a pattern is found to be affected by a splicing event annotated in the data. In order to avoid any over representation of repetitive patterns in the data a splicing event affecting a pattern is only counted once per Swissprot entry or Ensembl gene, respectively, regardless how often the pattern occurs in the entry and regardless of how many (repetitive) patterns of the same type are affected by a splicing event. If more than one isoform exists patterns are also counted only once regardless how many different isoforms suggest a splicing event affecting the corresponding pattern.

8.1.4 Assessing the statistical significance of the results

Given how often a pattern matches an entry and how often it is alternatively spliced we can compute the significance of the ratio between observed occurrences and splicing events, given the background probability of any pattern of the corresponding database being spliced. In order to address this question the use of Logarithm of Odds score (LOD), known from genetic linkage studies, has been proposed by Resch et al. [130]. Given s observations that a pattern is spliced out of n total occurrences of a pattern as well as the background probability b that any pattern from the corresponding pattern database is spliced the LOD value can be computed as the log ratio of two probabilities with $L(s, n, p)$ as derived from the binomial distribution.

$$LOD(s, n, b) = log_{10}\left(\frac{L(s,n,s/n)}{L(s,n,b)}\right) \text{ with } L(s, n, p) = \frac{n!}{s!(n-s)!}p^s(1-p)^{n-s}$$

8.2 Results

8.2.1 Frequency of repeats in Swissprot and Ensembl

In the following we define a repetitive motif as a PFAM pattern which occurs on average more than twice among all proteins containing the motif. We then analyzed the content of repeat-containing proteins annotated with alternative splicing events in the human genome in Ensembl

Pattern	Repeats	Proteins	Max. rep.	Avg. rep.	Name
Human					
PF00096	3493	390	33	8,96	Zinc finger, C2H2 type
PF00400	643	150	11	4,29	WD domain, G-beta repeat
PF00023	623	148	26	4,21	Ankyrin repeat
PF00041	432	111	24	3,89	Fibronectin type III domain
PF00036	202	99	5	2,04	EF hand
Swissprot					
PF00096	2034	301	29	6,76	Zinc finger, C2H2 type
PF00041	1140	209	132	5,45	Fibronectin type III domain
PF00400	761	189	11	4,03	WD domain, G-beta repeat
PF00023	834	172	26	4,85	Ankyrin repeat
PF07679	966	170	167	5,68	Immunoglobulin I-set domain

Table 8.1: The five, most common repeat patterns matching in 10722 Human genes and 14086 Swissprot entries at an e-value smaller than 0.01.

[73] (human dataset) as well as in Swissprot [175] (Swissprot dataset) which covers alternative splicing events in a wide variety of organisms.

Among the 10.722 genes in the human dataset annotated with more than one transcript, 2.209 (20.6%) contain multiple, non-overlapping copies of a PFAM pattern type. From the 14.086 proteins in the Swissprot dataset 2.901 (20.5%) contain repeats. Interestingly, the content of repetitive domains in human (including those which do not have annotated splicing events) is smaller with 15.1% which already hints to a small preference for alternative splicing to occur in genes containing repeats.

As shown in Table 8.1 the most frequent repetitive motif in both datasets are C_2H_2-Zinc finger domains (PF00096) which are found in 390 genes and 3.493 copies in the human genome. They are followed by ankyrin repeats, β-propellers and fibronectin type III domains. Out of 2.638 different PFAM patterns matching Swissprot entries 164 are found more than twice on average per entry (6.2%), while 146 patterns out of 2.507 (5.8%) represent repetitive motifs in the human genome.

Among the non-repetitive patterns frequently found in the human dataset are protein kinases (PF00069) matching 289 genes and 7 transmembrane receptors of the rhodopsin family (PF00001), occurring in 199 copies. In the Swissprot dataset PF00069 represents the most abundant pattern matching 440 entries followed by PH domain proteins (PF00169) involved in intracellular signaling and cytoskeleton formation (found in 247 entries).

8.2.2 Alternative splicing of proteins containing repeats

As described in Section 8.1, we used alternative splicing events annotated in the Swissprot and human dataset to test whether splicing affects repetitive elements more often than expected.

Out of 24.537 patterns matching human genes with splicing annotation, 8.805 are alternatively spliced in at least one isoform. Counting every pattern type only once per gene / entry (to avoid an over representation of repeats) we observed 14.498 pattern matches out of which 4.382 are alternatively spliced in the human dataset and 18.229 matches in the Swissprot dataset of which 3.938 appear to be spliced. This corresponds to a background probability for a pattern to be spliced of 30.2% in the human dataset and 21.6% in the Swissprot dataset. The smaller percentage of spliced patterns observed in Swissprot may be due to the more conservative, manual

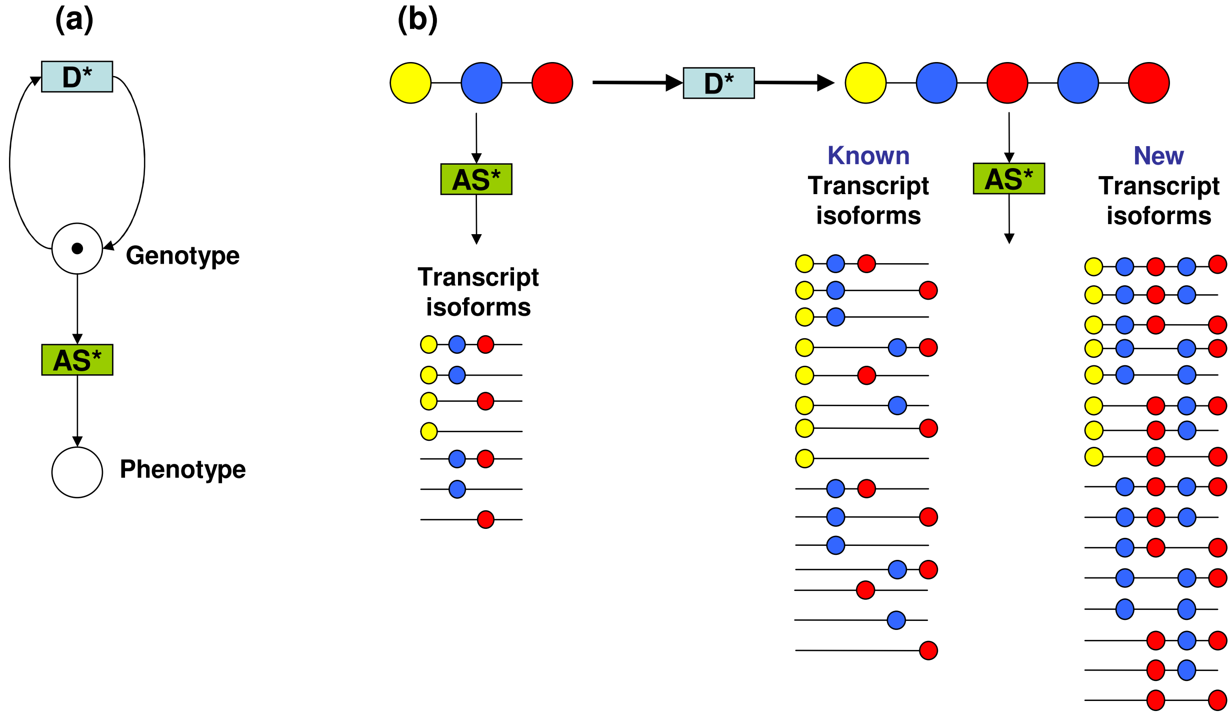


Figure 8.1: Figure 1(a) shows our simplified model of evolution to generate complex and new phenotypes via segmental duplication (D) and alternative splicing (AS). The model assumes an operating splicing mechanism (AS), which allows to splice individual exons or not in a regulated way. A piece of DNA can be duplicated via duplication events D. A sequence of duplications (D*) can already give rise to quite complicated genotypes and a variety of sequences and transcripts. The action of alternative splicing (AS*) on such a (duplicated) genotype allows for a combinatorial number of transcripts to express a complex phenotype (collection of proteins). Regulation of AS* can give rise to a variety of phenotypes from the same genotype depending on the conditions. Selection and mutation complicate the picture, but allow to modify the duplicated units independently and to change the motifs regulating the splicing thereby creating and manifesting certain phenotypes. Figure 1(b) shows an example how the combination of D* and AS* gives rise to a huge variety of new isoforms and at the same time allows to reproduce the old ones (before D*), thereby reducing the potentially deleterious effects of duplication and allowing for potentially new variants. (*Figure taken from [17]*)

annotation of splicing events. Based on those background probabilities we can now compute the statistical significance of the observed frequencies for a pattern to be spliced using the LOD score (see Section 8.1). This score allows us to identify patterns which are spliced more often than expected under the null-hypothesis that splicing events affect all pattern classes similarly, i.e. with the given background probability.

The results for the top 15 alternatively spliced PFAM patterns are shown in Tables 8.2 and 8.3 (for the human and Swissprot dataset respectively). All top 15 patterns are significantly more often spliced than expected with a LOD score greater than 3 corresponding to a p-value smaller than 0.001. In total, 33 patterns in the human dataset and 34 patterns in the Swissprot dataset are significantly more often spliced (p-value < 0.01 or LOD > 2) than expected. Ten patterns are found to be significant in both sets of top 15 hits.

For splicing events annotated in the human dataset there is a strong preference to affect repeat regions. 75% of the top 15 patterns (Table 8.2) are found to be repetitive (according to our definition). The pattern class with the best LOD score is the C_2H_2-Zinc finger pattern which appears to be affected by splicing events in 218 out of 390 genes. Zinc fingers are followed by splicing events affecting sushi domain repeats, spectrin repeats and ankyrin repeats which harbour annotated splicing events in 76 out of 148 genes matching the pattern. PFAM pattern PF00400, corresponding to β-propeller proteins, also appears to be significantly more often spliced than expected with a LOD score of 3.28.

Splicing events observed in the Swissprot dataset (Table 8.3) show a similar picture although

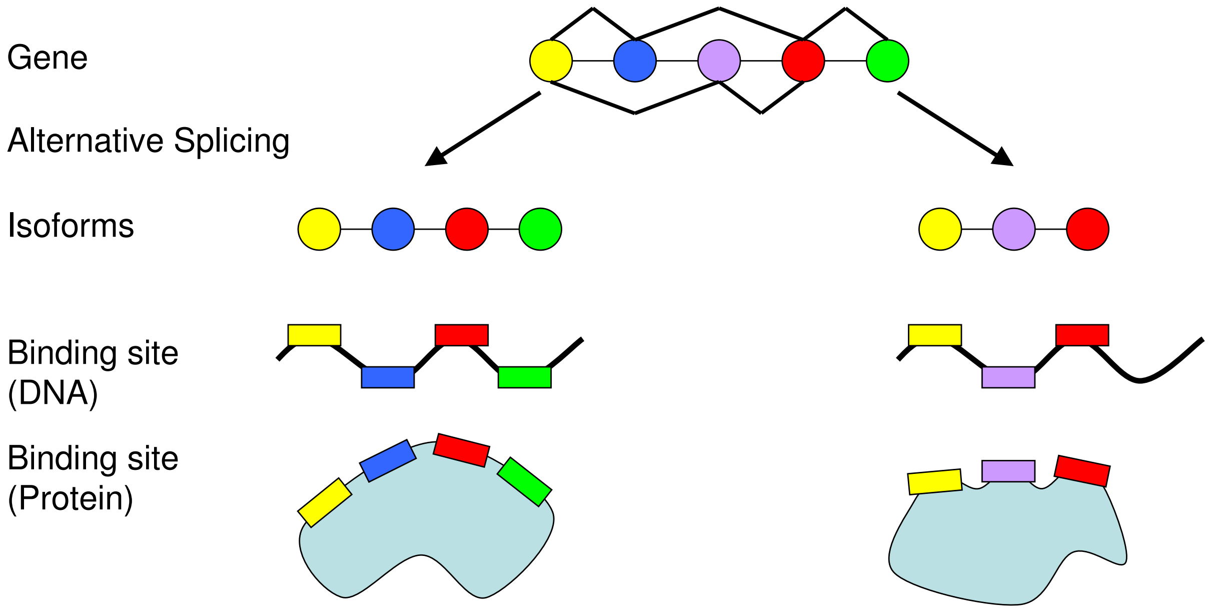


Figure 8.2: Shows the proposed mechanism of generating functional complexity of repetitive proteins via alternative splicing. Colored circles correspond to repetitive motifs with different binding properties (e.g. recognition patterns). Colored rectangles correspond to the recognized motifs on the DNA or protein level, for e.g. Zinc-finger or ankyrin domains. (*Figure taken from [17]*)

Pattern	Name	spliced	observed	% spliced	LOD
PF00096	Zinc finger, C2H2 type	218	390	0,56	23,94
PF00084	Sushi domain (SCR repeat)	32	43	0,74	7,73
PF00435	Spectrin repeat	20	24	0,83	6,32
PF00023	Ankyrin repeat	76	148	0,51	6,22
PF06758	Repeat of unknown function	9	9	1	4,68
PF00307	Calponin homology (CH) domain	35	60	0,58	4,4
PF01391	Collagen triple helix repeat	36	64	0,56	4,04
PF02210	Laminin G domain	21	31	0,68	4,01
PF00681	Plectin repeat	10	11	0,91	3,9
PF01193	RNA polymerase Rpb3/Rpb11	7	7	1	3,64
PF02383	SacI homology domain	7	7	1	3,64
PF07728	ATPase family (AAA)	7	7	1	3,64
PF03143	Elongation factor Tu	9	10	0,9	3,42
PF00400	WD domain, G-beta repeat	68	150	0,45	3,28
PF00041	Fibronectin type III domain	53	111	0,48	3,24

Table 8.2: The top 15 alternatively spliced patterns in human which were significantly more often removed by a splicing event than expected ($p < 0.001$). 10 of the 15 patterns (75%) usually occur in repeats of more than 2 copies per gene.

repetitive patterns are less dominant. Out of the first 15 patterns 40% represent repeats. Nevertheless, the top 4 patterns with the highest LOD values correspond to Zinc fingers, β-propellers, fibronectin domains and ankyrin repeats. Among the significantly spliced but non-repetitive patterns is the KRAB box domain (PF01352) which is often part of C_2H_2-Zinc-finger proteins and is involved in protein-protein interactions [162]. The function of splicing events affecting KRAB box domains has been discussed by Resch et al. [130].

It should be noted that, as shown in Table 8.4, not all splicing events alter the number of repeats in a protein but events which remove the complete, repetitive domain can also be observed. The implications of such events are likely to be similar to other events changing the domain content of a protein. As shown, between 56-76% of the annotated events in Swissprot and 50-73% in the human dataset change the number of repeats (depending on the pattern type) and have the potential to lead to novel, functional isoforms through repeat variation.

In the following, we will discuss the structure and function of four repetitive patterns which are significantly more often affected by splicing events than expected in both datasets and will describe possible mechanisms which allow the generation of a large variety of different functions from the recombination of repeats. Recombination of those elements has played an important role in the regulation of DNA transcription, the mediation of protein-protein interactions, DNA repair and the organization of complex tissues in evolution. As we will show below, this variety may be further extended through alternative splicing. The functional mechanism is illustrated in Figures 8.1 and 8.2.

Pattern	Name	spliced	observed	% spliced	LOD
PF00096	Zinc finger, C2H2 type	125	290	0,43	14,53
PF00400	WD domain, G-beta repeat	79	187	0,42	8,68
PF00041	Fibronectin type III domain	80	193	0,41	8,32
PF00023	Ankyrin repeat	68	168	0,4	6,58
PF00168	C2 domain	51	114	0,45	6,56
PF00084	Sushi domain (SCR repeat)	32	61	0,52	6,03
PF00622	SPRY domain	35	73	0,48	5,36
PF00696	Amino acid kinase family	12	15	0,8	5,04
PF08686	PLAC (protease and lacunin) domain	8	9	0,89	4,07
PF03256	Anaphase promoting complex 10	6	6	1	3,99
PF04760	Translation initiation factor IF-2	6	6	1	3,99
PF02985	HEAT repeat	21	41	0,51	3,75
PF01352	KRAB box	35	85	0,41	3,57
PF00418	Tau and MAP protein repeat	8	10	0,8	3,36
PF00551	Formyl transferase	5	5	1	3,33

Table 8.3: The top 15 alternatively spliced patterns in Swissprot which were significantly more often removed by a splicing event than expected ($p < 0.001$). 6 out of the 15 patterns (40%) are found on average twice per entry. Repeats are marked in bold font. Additionally, the HEAT repeat family is known to be repetitive but patterns occur on average with only 1.73 times per entry which may be due to highly diverse sequence patterns and, corresponding, rather large E-values observed for a match.

8.2.3 Alternative splicing of C_2H_2-Zinc-finger motifs

C_2H_2-Zinc-finger motifs are among the most common DNA binding motifs found in eukaryotic transcription factors. As shown in Table 8.1, 390 genes match the Zinc-finger pattern, which occurs on average 9 times per protein and, as discussed above, this family appears to be significantly more often affected by splicing events than expected (see Tables 8.2 and 8.3).

Zinc-finger domains consist of approximately 30 amino acids which forms a $\beta\beta\alpha$-motif. Amino acids on the surface of the α-helix make specific contacts with bases in the major groove and, therefore, allow the motifs to specifically bind certain DNA sequences [174].

The fact that splicing events in Zinc-fingers are so frequent hints to an important role of alternative splicing in generating Zinc-fingers with different DNA sequence recognition patterns. The importance of multiple Zinc finger isoforms resulting from the CF2 gene in different developmental stages of Drosophila have been shown by Hsu et al. [72]. The effect of one of the splicing events for CF2_DROME annotated in Swissprot, removing one Zinc-finger motif is shown in Figure 8.3a.

The high significance of splicing events affecting Zinc-finger motifs observed in our study is strong support for the hypothesis that modulations of the Zinc-finger repeat number allows to control the spectrum of regulatory processes that are served by a single transcriptional regulator gene [72].

Pattern	SP %removed	SP %changed	HG %removed	HG %changed
PF00096 (Zinc finger)	24.5 (28)	75.5 (86)	49.2 (105)	50.8 (107)
PF00023 (Ankyrin)	43.6 (31)	56.4 (40)	45.0 (34)	55.0 (46)
PF00400 (WD40 Repeat)	24.0 (18)	76.0 (57)	31.3 (21)	68.7 (46)
PF00041 (Fibronectin)	29.2 (20)	70.8 (47)	26.1 (11)	73.9 (31)

Table 8.4: Displays the percentage (and absolute number) of entries / genes containing at least two repetitive pattern matches. It compares isoforms for which all copies of the corresponding pattern are removed (%removed) and with isoforms where the number of repeats is changed (%changed) in Swissprot (SP) and the human genome (HG)

8.2.4 Sculpting protein binding domains - Ankyrin repeats

As discussed above, repeats of DNA binding motifs are frequently used to recognize specific DNA sequences and the large number of splicing events observed for those motifs hints to a dominant role of splicing in generating isoforms with different DNA binding properties. Now, we will focus on splicing events observed for ankyrin repeats (an example is shown in Figure 8.3b) which act as a molecular architecture for protein recognition [107].

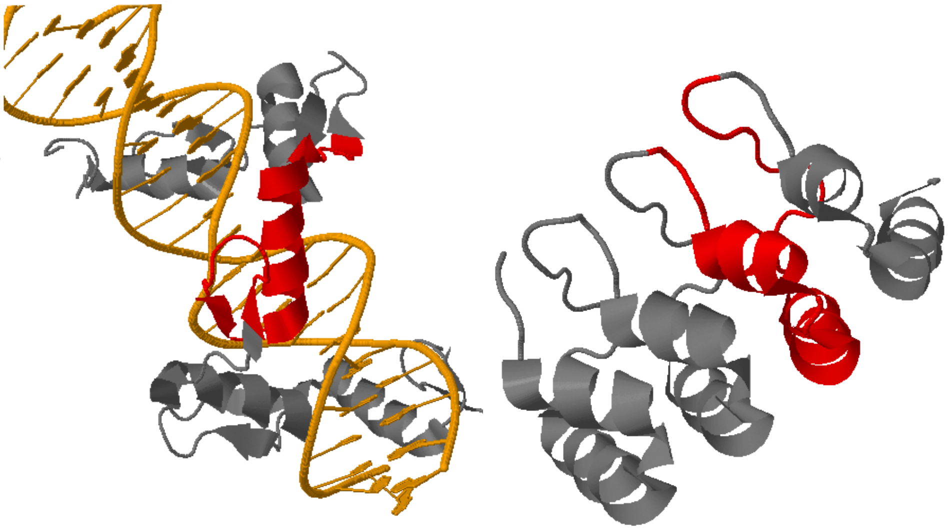
For many ankyrin proteins with varying numbers of repeats the binding partners are known. They play important roles in diverse processes like transcription factor regulation, cell-cell adhesion, cell cycle regulation and signaling [107].

Ankyrins are made up from several copies of a usually 33 residue long motif which folds into two anti-parallel α-helices followed by a longer loop region. The loop points outwards forming an almost 90° angle that sometimes folds into a β-hairpin structure. Stacking of those repeats leads to an elongated structure with a large, solvent accessible surface which is responsible for binding the targets of the protein. Unlike other protein-binding domains like e.g. SH2 or SH3 which recognize specific amino-acid sequences of the target, ankyrins recognize their specific binding partners through variations in adaptive surface residues dispersed over the whole target molecule [107]. Although different ankyrin repeat proteins are quite similar in sequence the binding of targets seems to be highly specific [97].

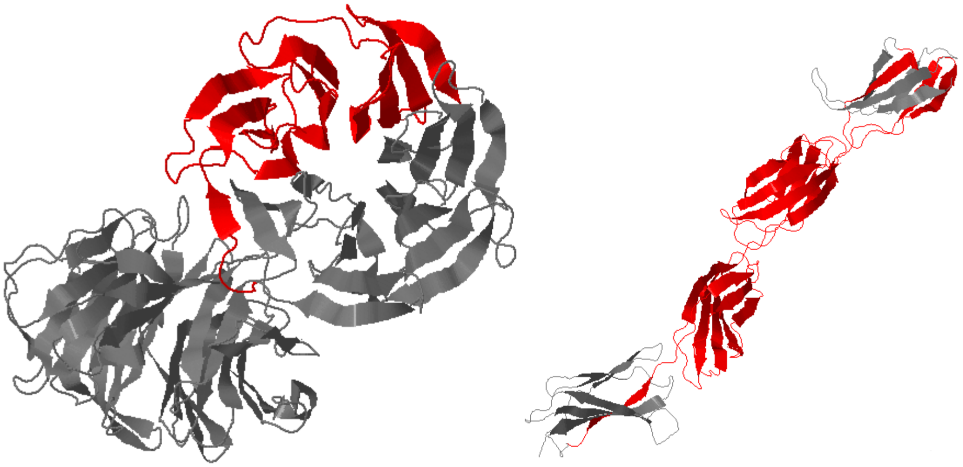
The fact that alternative splicing affects ankyrin repeat regions in a highly significant way (as observed in both datasets) may point to a use of ankyrin repeat variation for generating variable protein-protein interaction domains. In order to generate this variation, single repeat modules may be combined via alternative splicing to recognize specific targets (see also Figure 8.2).

We searched the biological literature assigned to splicing events annotated in Swissprot (68 entries) for ankyrin containing proteins for evidence of differing binding partners of the isoforms. For only very few cases the isoform has been characterized on the protein level and for even fewer cases potential binding partners different from the native one have been identified.

One example, where it is known that the native binding partner is not bound anymore but the transcript is specifically expressed in a tissue (pancreas) is the p12 isoform of CD2A1_HUMAN. CD2A1 belongs to the very well studied INK4 family of proteins which inhibit cyclin-dependent kinase 4 (CDK4) and 6 (CDK6) and modulate cell progression through G1-S transition. Proteins of the family consist of four or five ankyrin repeats. The p12 isoform described in [133] consists of exon 1 contributing one 1.5 ankyrin repeats and a novel C-terminus encoded by an intronic sequence. It is unclear if this sequence contributes another ankyrin motif or at least completes the missing half of the repeat such that the minimum number of two ankyrin repeats necessary for folding is obtained. The isoform does not interact with the native targets CDK4 and CDK6 anymore but is suggested to play a different role in growth suppression in the cell [133].



a) DNA binding b) Protein – Protein interactions



c) Protein – Protein interactions d) Tissue organization

Figure 8.3: Shows several interesting splicing events annotated in Swissprot on the structure level. Red parts are removed by the corresponding event. a) shows the removal of one Zinc-finger motif from CF2_DROME b) the removal of one ankyrin repeat in ANKR2_HUMAN c) the removal of three blades of a 7 blade β-propeller in WDR1_HUMAN leaving 4 blades (which may fold into a propeller structure) and c) two partially and two completely removed fibronectin type III domains annotated for FINC_HUMAN. (*Figure taken from [17]*)

More support for the modular usage of ankyrin repeats comes from in vitro experiments where it has been shown by Amstutz et al. [4] that it is indeed possible to create ankyrin proteins binding specific targets through recombination. In their study, novel intracellular kinase inhibitors have been designed using combinatorial libraries of ankyrin repeats. A similar mechanism may act in nature through alternative splicing.

The recombination of ankyrins can be used for the generation of specific protein-binding domains similar to the variation of Zinc-finger motifs for DNA recognition. Therefore, splicing may be a simple mechanism to control binding partner specificity in complex proteomes and interactomes.

8.2.5 Alternative splicing of β-propellers

While multiple copies of C2H2-Zinc-finger and ankyrin motifs do not fold into globular domains, β-propeller represent intermediates between solenoid domains (e.g. ankyrin repeats) formed by the simple repetition of super-secondary structure elements and globular proteins. Those domains are called toroids. Toroids, which also comprise TIM-Barrels and β-trefoils, are formed by relatively simple elements but fold into closed rather than open structures [30]. Propellers form a closed structure consisting of 4-8 repetitive elements, called blades, each of which usually consists of four β-strands forming one β-sheet. Those proteins carry out a wide variety of enzymatic and non-enzymatic functions including ligand binding and the mediation of protein-protein interactions [5]. We have already discussed the evolution of β-propeller folds and the connection to alternative splicing data in section 7.3.2.

An interesting example for an isoform of a β-propeller protein with a different number of repeats is D1 from DDB2_HUMAN described in [77]. The native variant contains 5 WD-repeats (according to PFAM) and is involved in DNA repair. D1 which is highly expressed in brain and heart contains only 2 WD-repeats due to a larger removal event and is shown to act as dominant negative inhibitor for DNA repair mediated by DDB1 and DDB2 [77]. Therefore, the splicing event seems to lead to a functional isoform despite the fact that only 2 motifs are left in the protein. This isoform is also interesting from an evolutionary point of view since an oligomeric structure of a β-propeller consisting of 3 two-bladed propeller motifs is known (PDB: 2bt9) [30].

8.2.6 Importance of splicing for complex tissue organizations

Fibronectin is a large, multi-domain protein which is involved in many different functions including cell adhesion, migration, differentiation and proliferation. The most abundant domain found in human fibronectin (FINC_HUMAN), as well as in various other proteins, is the fibronectin type III-domain which is an about 90 residues long all-β domain. It is repeated 16 times in FINC_HUMAN. As shown in Tables 8.2 and 8.3 splicing events affecting those domains are significantly more often observed than expected, especially in splicing events annotated in Swissprot.

Similar to other extracellular proteins, like the Sushi domain repeat, Fibronectin is organized as an extended molecule like ”beads on a string”. Specific domains are e.g. responsible for making contact with collagen, heparin and cell surface receptors of the extracellular matrix [12].

FINC_HUMAN (a part of the molecule is shown in Figure 1d) harbors several, well characterized splicing events which are known to be highly regulated during development, repair and disease and generate functionally different isoforms appropriate for growth and stability [47].
As shown in [12], splicing events do not only change the domain composition of the molecule but also change the interfaces between neighboring domains. Therefore, not only the presence or absence but also the order of the domains may influence their binding properties further increasing the complexity generated by recombination. The high frequency of splicing events observed for fibronectin type III domains hints to a very prominent role of those repetitive elements in several processes involved in the formation of complex tissues in multi-cellular organisms.

8.3 Discussion

We have carried out a large scale study on alternative splicing events observed in the human genome and in Swissprot affecting proteins containing repeats. We found strong evidence for a widely used combination of repeat expansion and alternative splicing to generate and drive functional and structural diversity in higher organisms.

Protein repeats are known to play major roles in all modes of protein-based regulation: the regulation of transcription by transcription factors (e.g. C_2H_2-Zinc-fingers), the mediation of protein-protein interactions (e.g. ankyrins, β-propeller), and the organization of complex tissues (e.g. fibronectin type III, collagen, sushi domains). The structural and functional properties of protein repeats make them attractive targets of evolutionary processes to generate structural and functional diversity. Repeat expansion has been found to be a successful means in the evolutionary history of higher organisms. The combination with alternative splicing appears to further increase and modulate the variety of repeat structures in proteins to introduce novel functions.

Especially C_2H_2-Zinc-fingers and Ankyrin repeats seem to be targeted by splicing events which may lead to novel isoforms regulating a different set of genes or protein-protein interaction domains with different binding partners. The function of many repeat proteins originates from the combination of repeat modules and their specific DNA or protein binding features.

The impact of protein-protein interaction domains on organism and network complexity has only recently been discussed [177]. The increase of repetitive elements in combination with alternative splicing, as proposed in this study and summarized in Figure 8.1, may have provided higher organisms an evolutionary advantage in evolving novel functions in the light of highly complex genome and proteome organization and therefore may have significantly contributed to organism complexity.

Despite several well characterized examples which support our hypothesis, the functional role of many splicing isoforms remains to be explored. Tissue and time specific expression detected by exon-level micro arrays and mass spectroscopy measuring complete proteomes as well as novel techniques to detect protein-protein and protein-DNA interactions will not only help to understand the function of those isoforms generated by alternative splicing but might also provide insights into the generation of complex organisms due to the interplay of alternative splicing and DNA duplication in the course of evolution.

From a structural point of view, increases or decreases in the repeat number can be expected to be easily tolerable by the protein structure as they have been a successful strategy to evolve novel functions. For many repeat classes crystal structures with varying repeat numbers are known. Nevertheless, the common evolutionary history of proteins with different numbers of repeats is not always obvious as discussed by the example of β-propeller repeats. Here, alternative splicing provides additional evidence for the hypothesis that those proteins may indeed be related as splicing often creates isoforms with different numbers of repeats. Data on alternative splicing therefore is a novel and important tool to study protein fold evolution.

Overall, the analysis of this specific group of protein structures made up from repetitive structural motifs further supports our hypothesis on a coupling of functional and structural diversity generated in the course of evolution and by alternative splicing as described and discussed in Chapter 7.

Part III

Genome-wide Detection and Analysis of Alternative Splicing

Chapter 9

Detection and prediction of Alternative Splicing Events

As discussed above, alternative splicing is a highly abundant process in eukaryotic cells but also needs to be tightly regulated in a time and tissue-specific manner. In this part we will describe the methods and tools developed in the course of the thesis to aid a structure- and feature-based, genome-wide analysis of alternative splicing and to identify alternative splicing events in experimental datasets of only recently developed experimental techniques. In this Chapter we will briefly review some of the most important experimental approaches for the identification of alternatively spliced isoforms in high-throughput methods and will also describe some bioinformatics approaches to predict alternatively spliced exons *ab initio*. An overview on the different experimental approaches to identify splice isoforms is shown in Figure 9.1.

EST and cDNA-based annotation of alternative transcripts

Most of our knowledge on alternative transcripts expressed in eukaryotic cells comes from experiments collecting and sequencing mRNA transcripts. Those transcripts may either be sequenced completely (called cDNA transcripts) or they are only partly sequenced resulting in Expressed Sequence Tags (ESTs). Both types of data are available at a large scale in several databases ([24], [76]) and may be used in combination with completely sequenced genomes to identify expressed (alternative) transcripts in the cell.

While cDNA sequences of complete genes represent the highest quality data available for mRNA transcripts and usually can easily be aligned to a genomic locus and directly lead to the sequence of the expressed mRNA transcript (see also the H-DBAS database [155]), the relatively short ESTs only give an imperfect view on possibly present transcripts in a cell. Splicing events predicted from ESTs tend to be biased towards the 3' and 5' ends of the transcripts and in general there is not enough data to infer frequency and tissue/time specificity of the splicing events from this data [23]. Also, EST data usually comes from a small set of different cell types, often from tumor libraries, where the accuracy and functional importance of the observed splicing events may be questionable. Nevertheless, most of the data on known splicing events comes from the analysis of ESTs using bioinformatics methods (reviewed in [25]) and have led to the

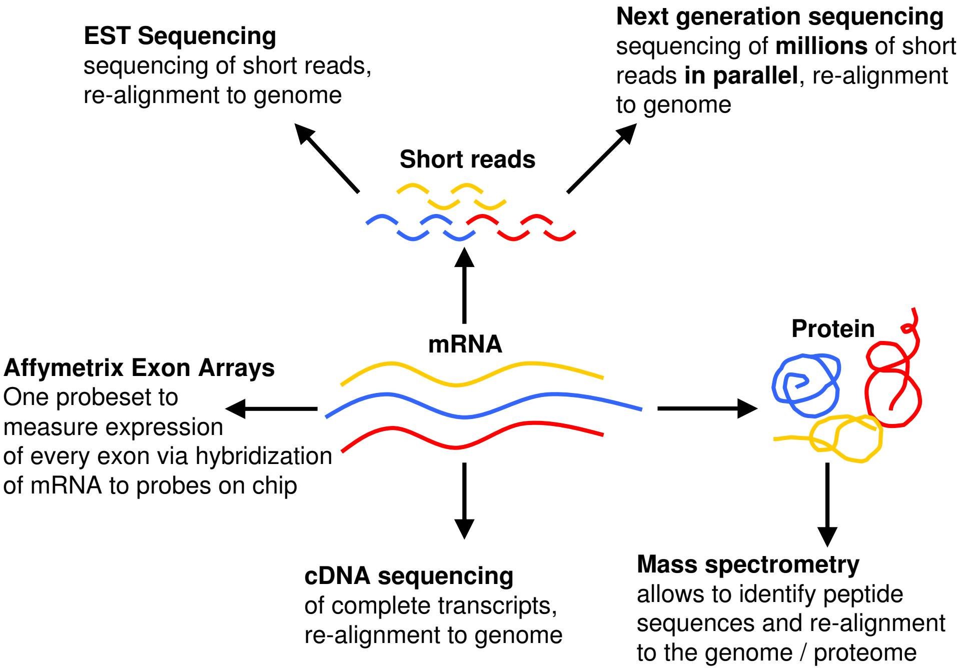

EST Sequencing
sequencing of short reads, re-alignment to genome
Next generation sequencing
sequencing of millions of short reads in parallel, re-alignment to genome
Short reads
mRNA
Protein
Affymetrix Exon Arrays
One probeset to measure expression of every exon via hybridization of mRNA to probes on chip
cDNA sequencing
of complete transcripts, re-alignment to genome
Mass spectrometry
allows to identify peptide sequences and re-alignment to the genome / proteome

Figure 9.1: Summarizes experimental approaches to identify alternatively spliced isoforms from experimental data sources. While next-generation sequencing, EST sequencing rely on sequencing short reads and the subsequent re-alignment onto the reference genome, cDNA Sequencing sequences complete mRNA transcripts and therefore is a very reliable method to determine the nature of splice variants. Affymetrix exon arrays potentially measure all splice variants in a sample at a time with the drawback of problems in the analysis of isoform mixtures and others. Finally, mass spectrometry overcomes the problem of mRNA-level identified isoforms and experimentally identifies the existence of short peptides (similar to EST sequencing) on the proteome level which are then re-aligned to the reference proteome and may confirm specific isoforms.

development of several databases like the ASAP database [83] or the ASD resource [147].

Array-based detection of alternative splicing

Especially to overcome the problem of analyzing time and tissue specific splicing events using EST data and to allow for a genome-wide screen for alternative isoforms, several platforms for analyzing the presence of alternative transcripts using microarrays have been developed. Among those, mainly two different approaches can be distinguished.

Several platforms contain thousands of oligonucleotide probes that are specific for individual splice-junction sequences formed by the inclusion or skipping of an exon. Using splice-junction arrays, specific splicing events may be detected with a high accuracy in a time and tissue specific manner. Nevertheless, their major disadvantage is the fact that only splice variants can be detected which have been known / expected to exist at the design time of the chip.

The second chip platform for detecting splicing events on a genomic scale is the Exon Chip technology available from Affymetrix. Those chips measure the exons of a gene individually potentially allowing for the analysis of the presence or absence of every exon of a gene in any experimental condition. Currently the chips are available for Human, Mouse and Rat and we will shortly discuss the a method to analyze such data as well as the data content of those chips in the Chapter on ProSAS (Chapter 10) and in the context of the PASS method (Chapter 11).

Next-generation sequencing methods

Recent development of sequencing methods (termed *next generation sequencing*) by Illumina's GenomeAnalyzer, Roche's 454 and the ABI SOLiD technology, allow to sequence gigabases of short reads of about 40-400 nucleotides in length (depending on the platform used) at minimal time and cost. Using those techniques, one can sequence all mRNA transcripts present in the cell under a certain experimental condition or in a certain developmental stage and therefore allow for the detection and possibly also the quantification of all transcripts present in the cell. Splice junctions and alternative transcripts can then be identified by aligning those short reads back to the genome and transcript frequencies can be (in theory) determined by simply counting the number of identical reads observed in the dataset which is described by Illumina to be proportional to the transcript abundance.

Next generation sequencing has already been applied to the analysis of gene activity and alternative splicing in humans ([151] and [169]). The authors find a larger number of splice junctions (4096 out of 94214 confirmed junctions) which are novel and may correspond to new, often minor, splice variants of more than 3000 genes. Overall, those experiments indicate that up to 90% of all human multi-exon genes may be alternatively spliced (more than the 74% previously reported based on EST data [79]) and show that those methods may become highly interesting experimental approaches for the future investigation of alternative splicing.

Detecting splice variants in mass spectrometry data

Since only recently, large-scale mass spectrometry experiments which have the capability to measure the composition of complete proteomes allow for the detection of transcripts on the proteome level similar to EST based approaches on the mRNA level. Tandem mass spectrometry allows to measure the presence of virtually all proteins present in a cell or tissue at a specific time via the identification of small peptides. The major advantage of splice variants identified in mass spectrometry is the fact that those isoforms are identified on the proteome instead of the transcriptome level. Therefore, they are likely not to be prone to nonsense mediated decay or protein degradation.

Usually the spectra measured in a run are matched against a database of known proteins and known splice variants (e.g. UniProt [175]) and the theoretical peptides contained in the database matching the observed spectra best are predicted to be present. A large number of database search engines and post-processing tools are available ([33], [82]). Those tools allow for a reliable detection of known proteins and splice variants via the identification of unique peptides as discussed in Chapter 12.

Two problems remain. Rare splice variants which are observed only in few spectra are likely to be ignored by the standard pipelines. Second, the pipelines are not able to detect novel splice variants which are not yet annotated in public databases. To overcome this problem two approaches have been suggested. While some publications suggest to use extended peptide databases (e.g. all exon junctions observed in EST data [45]) other approaches predict the sequences of the peptides present in the data from the scratch corresponding to de novo peptide identification [51]. We will present our own approach to the problem in Section 12.4. But despite those difficulties mass spectrometry will become a major source for splice isoforms identified on the proteome level in the future.

Bioinformatics approaches to predict splicing events

Besides the experimental tools described above for the detection of splicing events, several bioinformatics approaches to predict constitutive and alternatively spliced exons have been published. All methods use certain features that allow to distinguish between constitutive and alternative exons. Those features include the conservation of the exons and their flanking regions in other organisms [145], the fact that the length of most alternative exons is divisible by three to avoid frame-shifts [100], the exon and intron length and nucleotide composition or other features like PFAM domains [67]. The computational tools for the prediction range from support vector machines [144] to hidden Markov models [116]. To date, those methods can predict constitutive and alternative exons to a certain extent. Nevertheless, our knowledge on the regulation of alternative splicing and the role of the various splice signals in exonic and intronic regions is still too limited to reliably predict the expression of specific transcripts possibly even in a tissue or time-specific manner.

Chapter 10

The ProSAS - Database

Due to the importance of alternative splicing and the large amount of data being available, several databases ([40], [50], [83], [95], [147], [155]) dedicated to the analysis of alternative splicing have been published in the last years. The main purpose of most databases ([95], [83], [147], [155]) is the collection of alternative transcripts on the mRNA level based on EST and cDNA data and the development of specific methods like partial order alignment to determine potential splice variants [178]. Additionally they often annotate features to allow a functional characterization of the different isoforms based on InterPro [109] patterns, tissue specificity and literature references describing specific isoforms.

Nevertheless, one dimension in the analysis of the different isoforms, namely the protein structure, is missing in all those databases to date. We describe and discuss the importance of analyzing and interpreting alternative splicing events in the context of protein structures in Chapters 7, 8, 11 and 12. In order to allow for an easy and straightforward, structure-aided analysis of alternative splicing, we have build the ProSAS (Protein Structure and Alternative Splicing) resource and database. ProSAS unifies protein structure and alternative splicing data for several mammalian genomes, Swissprot as well as isoforms identified in mass spectrometry runs and provides tools for a detailed analysis of those splicing events on the protein structure level. All data stored in the ProSAS database can be accessed via the interactive web interface implemented with help of a recent Java-based technology, namely the Zkoss framework.

ProSAS has been used as basis for all studies on alternative splicing discussed above. In the following we will therefore focus on its technical features and properties. First of all we will describe the ProSAS pipeline used to annotate and analyze splicing events. We will then discuss the current database content and finally, we will briefly describe the main features of the ProSAS web interface.

10.1 The ProSAS pipeline

The ProSAS database consists of several parts of data and integrates them into one large resource. Figure 10.1 shows an schematic overview on the basic data sources and their relationships which will now be described in some more detail.

10.1.1 Genome and alternative splicing data

Genomic datasets available in ProSAS are based on the Ensembl database [73]. Currently, data for Human, Mouse, Rat and Chimpanzee is available, but additional genomes can easily be integrated into the database. Alternative splicing events for each gene are based on the alternative transcripts annotated in Ensembl. Additionally, all Swissprot entries with annotated splicing events (VARSPLIC annotation) are stored in the ProSAS database. In cases where Ensembl transcripts correspond to a Swissprot entries, both sources are interlinked which allows for an integrated analysis of splicing events annotated in different databases.

10.1.2 Protein structure prediction

For all transcripts annotated in the database, we searched for homologous structures in the PDB using the SIMAP database and the SIMPAT webservice. All homologs in the PDB identified by SIMAP with an E-value smaller than $10e^{-5}$ are stored in our database. Pairwise sequence-structure alignments have been computed using standard free-shift alignment with the PAM250 matrix and a gap open penalty of -12 and a gap extend penalty of -1. For each part of the transcript one protein structure has been selected as template and in the case of multiple, overlapping structures, the template with the best E-value has been chosen for a region of the transcript. For further analysis we usually use only transcripts which can be modeled with highly confident structural models with a sequence identity of e.g. more than 40% between transcript and template and which are covered by one template to a large extent (e.g. more than 75%). Other thresholds may be chosen depending on the application.

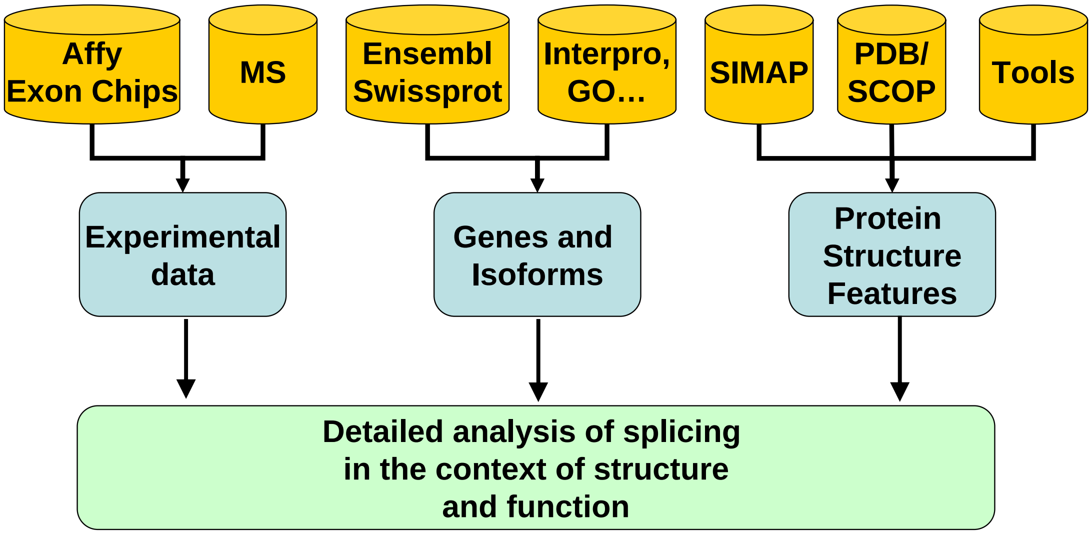


Figure 10.1: Shows a schematic overview on the various data sources and their relationships in the ProSAS database (The MS resource corresponds to mass spectrometry data, e.g. available in PeptideAtlas [41]).

10.1.3 Characterization of splicing events

In order to describe the effects of different splicing events onto the structure and domain composition of a protein we automatically assigned the following features to every exon and every splicing event observed in the data. The features are motivated by the properties used to characterize splicing events as described by various studies in the field ([14], [130], [134], [171]).

Sequence-Patterns To characterize splicing events on the sequence level and also in cases where no reliable protein structure prediction is available we use patterns of the Interpro [109] member databases like PFAM [49] or PROSITE [74]. Those patterns, which can be identified using InterProScan [126], allow for a sequence-based analysis of splicing events in terms of the domain organization of the spliced isoform (e.g. by PFAM domains), changes in the active site of a protein (e.g. via PROSITE active site patterns) or changes in the localization of the isoform e.g. membrane-bound or free in the cytoplasm depending on the presence or absence of e.g. transmembrane-helix signals. Matches of those patterns to all proteins from human, mouse, rat, chimp and Swissprot have been obtained from SIMAP database. A splicing event is said to affect a pattern if at least 75% of the residues involved in the pattern / domain are changed or removed by the splicing event.

Structure-based domain assignments In order to identify changes in the domain structure of a transcript on the structure level we use two types of features. Structural domains in template structures are identified using the SCOP database. Further, globular parts of a protein chain are detected using the PDP [2] program. A splicing event is said to affect a complete domain if at least 75% of residues forming the domain are affected by the event.

Secondary structure based features Secondary structures are annotated to the template structures using DSSP [80]. Besides the secondary structure assignments, DSSP is also used to define the solvent accessibility surface of residues as well as the location of beta-strands in larger beta-sheets. Those strands can either be ”peripheral” which means that they are located on the edges of a sheet (with hydrogen bonds only on one side) while strands with hydrogen bonds to other strands on both sides are classified as ”internal”. Additionally, larger regions which do not contain secondary structure elements are defined as unstructured regions.

Definition of core elements of a structure family Protein families as defined by most recent SCOP version (1.73) partially display the evolutionary history of a group of proteins, i.e. the set of insertions, deletions and replacements that have been tolerated by the structure without changing the overall fold of the protein. Therefore, family members provide a valuable source of information to judge the effects of alternative splicing events as already described in section 7.

For each template a multiple structure alignment of the corresponding SCOP superfamily is computed using STACCATO [141]. On the superfamily level of the SCOP hierarchy proteins exhibit enough structural variance to define conserved and variable regions without overestimating the structure conservation due to too similar proteins. Each group must contain at least 3 members and the maximally allowed sequence identity between two members is 60%.

Given a multiple structure alignment a conserved region is defined as a block of at least 10 residues which are conserved among all members of the superfamily. Each member in the block may contain 2 gaps at maximum to account for some small variability within conserved blocks. All regions outside of the conserved blocks are defined as variable regions. The proteins in the set define what we call evolutionary isoforms which display changes in the structures which are likely to be tolerable for the protein structure if they occur by alternative splicing.

An alternative definition of an integral part of a protein structure has been given by Wang et al. [171]. Here, core elements are defined as secondary structure elements of at least length four (for alpha-helices) and length three for beta-strands which are in contact with at least one other core secondary structure element.

10.1.4 Affymetrix data mapping

Affymetrix provides a new platform (”exon arrays”) to measure the expression of human, mouse and rat genes on the exon level by measuring each exon individually on the chip. To aid the analysis of data obtained from such expression experiments, we mapped probesets measured on the different mammalian exon array chips onto exons of human, mouse and rat, respectively. A probeset was mapped onto an Ensembl-defined exon if the chromosomal position of the Affymetrix gene annotated in the NetAffx file overlapped at least 75% with the chromosomal position of the human gene and the position of the probeset overlapped to at least 75% with the chromosomal position of the human exon.

Additionally we provide information on exons which are orthologous to one another in different organisms. This allows for a mapping of exons and consequently Affymetrix probesets across organisms and for an inter-organism comparison of alternative splicing events.

We make use of this mapping provided by ProSAS in Chapter 11 when describing our PASS method for analyzing Affymetrix exon array data.

10.1.5 Further data sources

Other data sources linked to genes and transcripts available in ProSAS include orthologous genes and exons between human, mouse, rat and chimpanzee as well as gene descriptions and GO annotations which are obtained from Ensembl using the BioMart [81] service.

10.2 The ProSAS database content

The current release of the ProSAS database contains data from four genomes, namely Homo sapiens (Human), Mus musculus (Mouse), Rattus norvegicus (Rat) and Pan troglodytes (Chimp), as well as data on alternatively spliced entries stored in the Swissprot database. Genomic data has been obtained from Ensembl (Release 48, December 2007), while Swissprot data corresponds to release 54 (also downloaded in December 2007).

	Genes	Transcripts	Exons	Avg. exon length	Probesets
Human	27316	54640	283857	158 (52 AA)	348258 (221228)
Mouse	24423	32041	218322	167 (55 AA)	271806 (196312)
Rat	22959	33350	229091	161 (53 AA)	233260 (188890)
Chimp	21624	35130	232883	152 (50 AA)	0 (0)
All	96322	155161	964153	159 (53 AA)	853324 (606430)
	Entries	**Isoforms**	**Human**	**Mouse**	**Rat**
Swissprot	14086	22724	11489 (6095)	6083 (3620)	1709 (1085)

Table 10.1: Genomic database content of the ProSAS database in June 2008. The Affymetrix probesets column displays the number of probeset mapped onto exons of the corresponding genome as well as the number of distinct exons covered by at least one probeset (in brackets). Swissprot entries and their isoforms are shown in the last row. The distribution of isoforms in the three major organisms contributing isoforms in Swissprot (in the Human, Mouse and Rat columns) is also shown. The values represent the number of isoforms as well as the number of distinct Swissprot entries (in brackets) of the corresponding organism

	modeled genes	40/60% sid	40/60% sid, 75% cov	90% sid, 90% cov
Human	17363	10345 / 7162	3917 / 2945	1462
Mouse	13082	7499 / 5186	3012 / 2257	862
Rat	10620	6185 / 4329	2497 / 1880	681
Chimp	8482	5344 / 3844	2129 / 1682	929
Swissprot	10445	6543 / 6543	1500 / 1116	437

Table 10.2: Displays the number of genes which are modeled by any structure and any quality criteria (modeled genes column) as well as all genes which are modeled with a specified quality according to the sequence identity (sid) and coverage (cov).

The basic database content of ProSAS is shown in Table 10.1. Human harbors most known splicing events in Ensembl as well as in Swissprot. In Ensembl, 54640 transcripts are annotated to all known human genes while in Swissprot, about 50% of all known isoforms are annotated for human entries. This is likely to be caused by the larger set of known Human EST and cDNA sequences as well as the larger interest in characterizing human splice isoforms in Swissprot and not by a systematically larger presence of alternatively spliced products in the human proteome.

The unique feature of the ProSAS database is to allow a structure-based analysis of splicing events. Therefore, we have modeled the structure of all proteins in the database using the methods described above. Table 10.2 displays the number of Ensembl genes and Swissprot entries which can be modeled according to different quality criteria. Models covering more than 75% of a protein with a sequence identity of more than 40% can be said to be in the save modeling zone and Above 60% sequence identity they can be said to be high quality structural models. As shown, 10-15% of the entries in the database can be fully modelled (coverage $>$ 75%) with a good quality. Between 24% (chimp) and 46% (Swissprot) of all entries are covered at least partly by a good model with more than 40% sequence identity between template and target.

	Any InterPro	PFAM	SMART	PROSITE	Superfamily
Human	26556	18809 (3393)	9608 (589)	9043 (942)	16710 (943)
Mouse	19963	14906 (3239)	6888 (575)	7293 (902)	13236 (917)
Rat	15256	12353 (3055)	5356 (564)	5992 (864)	10730 (882)
Chimp	13110	9897 (2868)	4885 (563)	4634 (850)	8596 (861)
Swissprot	14044	11619 (2615)	6519 (556)	5255 (735)	10202 (790)

Table 10.3: InterPro pattern coverage of the genes and entries annotated in Swissprot and Ensembl at an e-value of smaller than $10e^{-5}$ for different, important member databases. In brackets, the number of distinct patterns is shown.

Additionally to protein structures, features identified on the sequence level provide important and additional knowledge and we annotated InterPro member database patterns to all genes and entries. The content of ProSAS according to those features is shown in Table 10.3.

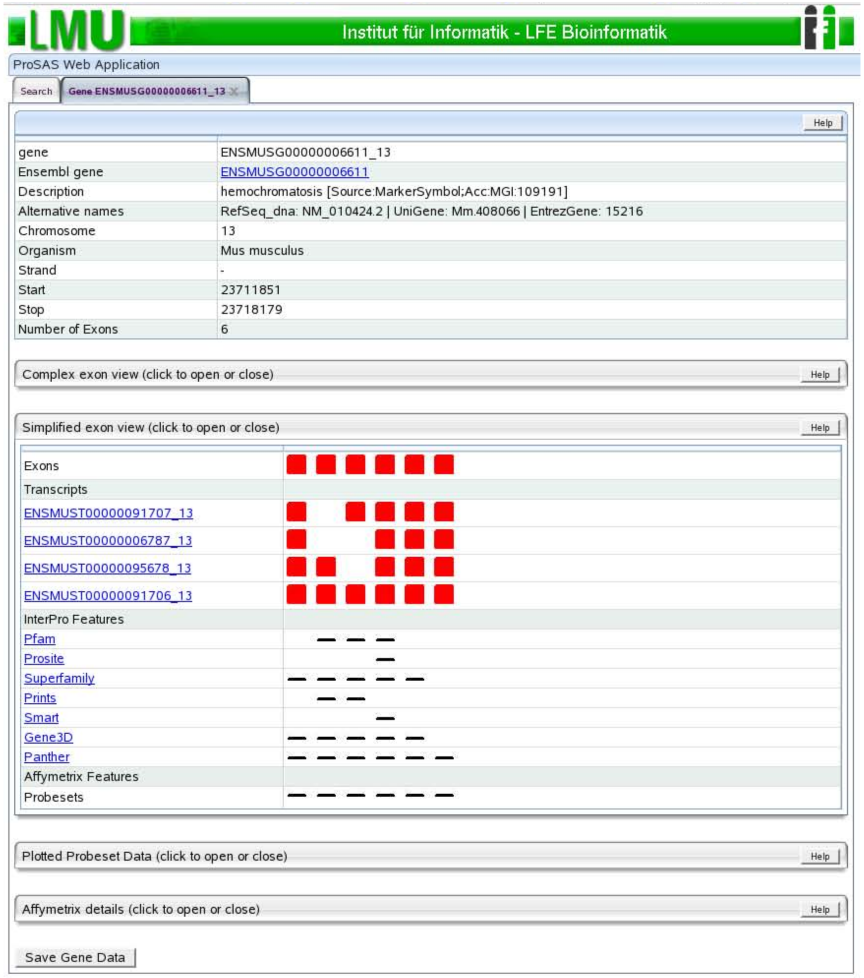
In the current mapping of Affymetrix probesets onto the exons of Human, Mouse and Rat we can cover 22290 (82%) of the Human genes, 22613 (93%) of the Mouse genes and 21429 (93%) of the Rat genes by probesets. On the exon level 221228 (78%) of the Human exons, 196312 (90%) of the Mouse exons and 188890 (92%) of all Rat exons are measured by at least one probeset on the corresponding Affymetrix Exon Array. Additionally, for many exons and their corresponding probesets (e.g. 70% in Human) we have information on orthologous exons which are highly similar according to their sequences and according to their structure and following ”orthologous” probesets stored in the database such that inter-organism comparisons of exon array data become possible.

10.3 The ProSAS database web interface

The data stored in ProSAS is made available for download (MySQL table dumps) or is accessible through the web interface of the ProSAS database which allows users to access and browse the data in order to explore the relationship between protein structure and alternative splicing. The interface is implemented using the Zkoss framework (http://www.zkoss.org) which allows to implement dynamic web interfaces and pages using Java. It consists of several components, where the most important ones are now shortly described.

Database searches The database can be searched by different identifiers in the search dialog. Among them are Ensembl gene and transcript ids, Ensembl gene descriptions (fulltext keyword search), Uniprot Names, InterPro member database patterns names as well as Swissprot / Uniprot names. The search may be limited to genes and Swissprot entries that are modeled by protein structure according to certain structure quality criteria, namely sequence identity and structural coverage of the transcript or entry.

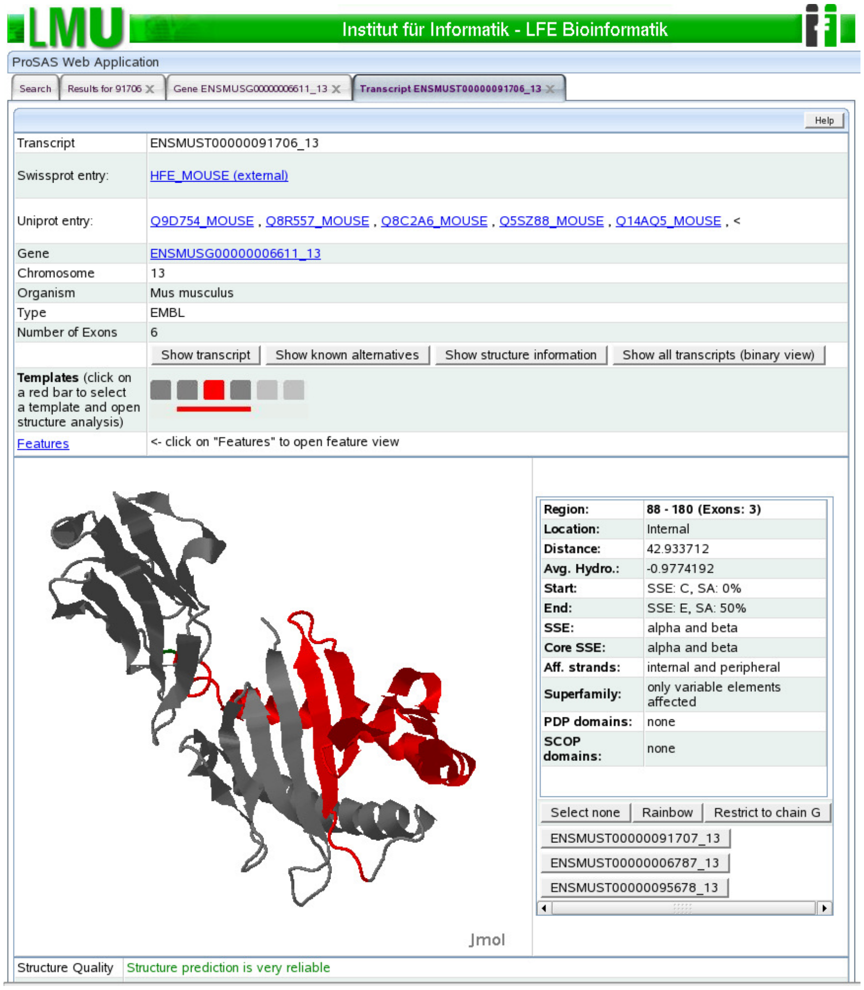
Gene report This view (see Figure 10.2) provides access to gene specific information. All exons of the gene as well as the exon composition of the corresponding transcripts of the gene


LMU
Institut für Informatik - LFE Bioinformatik
ProSAS Web Application
Search
Gene ENSMUSG00000006611_13
Help

gene	ENSMUSG00000006611_13
Ensembl gene	ENSMUSG00000006611
Description	hemochromatosis [Source:MarkerSymbol;Acc:MGI:109191]
Alternative names	RefSeq_dna: NM_010424.2 | UniGene: Mm.408066 | EntrezGene: 15216
Chromosome	13
Organism	Mus musculus
Strand	-
Start	23711851
Stop	23718179
Number of Exons	6

Figure 10.2: Details for gene ENSMUSG00000006611_13 from mouse showing all exons of the gene as well as different transcripts annotated for the gene in Ensembl. Matches of InterPro patterns and Affymetrix probesets onto exons can also be accessed

are visualized as linear arrays in two subsections displaying either the exon sizes relative to their true size or in a simplified way as equally sized blocks (shown in Figure 10.2) telling if exons are present or absent in the transcript which allows for a faster overview on annotated splicing events. Exon information (sequence, positions and phase) can be viewed by clicking on the specific exon. Links to Ensembl and Swissprot (if annotated for the gene) are provided to obtain additional information. InterPro patterns are linked to their corresponding source databases to obtain specific information about the pattern. These patterns allow to judge splicing events in a functional context by the absence or presence of patterns in different transcripts. Expression values for all exons of the gene according to the Affymetrix exon array technology can also be analyzed as shown in Figure 11.1 and splicing events observed on the chip can be explored. Following the transcript link leads to the transcript view panel.



Transcript	ENSMUST00000091706_13
Swissprot entry:	HFE_MOUSE (external)
Uniprot entry:	Q9D754_MOUSE , Q8R557_MOUSE , Q8C2A6_MOUSE , Q5SZ88_MOUSE , Q14AQ5_MOUSE , <
Gene	ENSMUSG00000006611_13
Chromosome	13
Organism	Mus musculus
Type	EMBL
Number of Exons	6
Templates (click on a red bar to select a template and open structure analysis)	
Features	<- click on "Features" to open feature view

Region:	**88 - 180 (Exons: 3)**
Location:	Internal
Distance:	42.933712
Avg. Hydro.:	-0.9774192
Start:	SSE: C, SA: 0%
End:	SSE: E, SA: 50%
SSE:	alpha and beta
Core SSE:	alpha and beta
Aff. strands:	internal and peripheral
Superfamily:	only variable elements affected
PDP domains:	none
SCOP domains:	none

Structure Quality	Structure prediction is very reliable

Figure 10.3: Transcript details view for transcript ENSMUST00000091706_13 from gene ENSMUSG00000006611_13. The structure of the transcript is visualized with Jmol (http://www.jmol.org/) and the difference with respect to transcript ENSMUST00000091707_13, namely the deletion of a larger N-terminal part is visualized on the structure. The alternatively spliced region is characterized with respect to different features such as solvent accessibility, secondary structure content and the conservation of the removed part in the corresponding SCOP superfamily.

Transcript report and structure analysis The transcript view (see Figure 10.3), as well as the Swissprot entry report which is designed in a very similar way, allows access to transcript specific data like the protein and DNA sequence and colors the exons of the transcript on the sequences. Other transcripts of the gene can be visualized with respect to the current transcript and deletions, replacements and insertions that occur due to the splicing event are color-coded on the transcript sequence.

The analysis of alternative splicing events in the context of protein structures is also available in this view. Users can choose a protein structure that matches the transcript. This structure is then visualized using Jmol and the correspondence of the gene structure (i.e. its exons) and the protein structure can be explored interactively. All exons of the transcript can be colored individually in the structure. That way, exon positions and substructures that belong to one or more exons on the structure level can easily be identified. Such an analysis provides useful insights into the correspondence of exons to small structural motifs or the location of exon boundaries on the structure level. An exon or a group of exons can additionally be structurally classified in terms of many features like secondary structure content, solvent accessibility, structural contacts, evolutionary knowledge on the variation observed in the corresponding SCOP superfamily as well as domain classifications as defined by SCOP or PDP.

Despite analyzing each exon individually, known splicing events, i.e. other known transcripts of the gene can also be colored on the structure level relative to the current transcript. This conveniently visualizes deletions, insertions and replacements on the structure level observed in different transcripts.

Those tools and features allow users to judge the importance of a spliced exon for the stability of a protein structure and will therefore provide an interesting additional dimension in the analysis of splicing events and isoforms.

While some events appear to be non-critical for the structure since complete domains, globular parts or unstructured regions are removed or replaced, others that remove large or well structured parts of the protein are much harder to explain and might hint to non-sense mediated mRNA decay or non-trivial effects in the isoform structure as described in 7. The web interface therefore allows to identify the cases where additional experimental validation of the protein product might be necessary.

10.4 Conclusion

With the ProSAS database we provide a unified framework for a structure-based analysis of alternative splicing events. The application and usefulness of ProSAS is twofold:

The web application is an interesting tool for biologist and bioinformaticians to analyze specific splice variants and others who want to gain a deeper understanding of the connection between splicing events and protein structures. The interface provides easy access to thousands of known, spliced proteins in Swissprot and Ensembl with highly confident structure annotations, visualizes splicing events on the template structures, provides a feature-based analysis of the splicing events observed and allows for the functional interpretation of the alternative transcripts linking to InterPro member databases, Swissprot and other sources.

The data stored in the underlying ProSAS database has been an important resource for the analyses described in Chapters 7, 8, 11 and 12. Since the publication in December 2007 [19] several additional features have been added to the database itself (including the chimpanzee genome, data on introns, orthologous exons and others). Also, we have implemented several new features for the web interface (e.g. the integration of the PASS method and the visualization of mass spectrometry confirmed isoforms discussed in the following). The functionality and content of ProSAS will further be extended in the future.

Chapter 11

PASS - Identifying splicing events in Affymetrix Exon Arrays

Alternative splicing is rarely detected by standard microarray platforms as they provide only a few probes per gene. In contrast, Affymetrix exon arrays are targeted at a more comprehensive detection of alternative splicing, providing in general one or more probesets for one exon (see [37] for a review on exon microarray designs and splicing detection methods). The quantitative detection of splice forms known at design time is performed by focused arrays using specific probe arrangements, e.g. exon junction arrays [120].

With the PASS (Pairwise Alternative Splicing with Scaling) method [89] Robert Küffner proposed a novel approach to detect alternative splicing events in Affymetrix exon array data and described a benchmark system to reliably evaluate the performance of such methods based on splicing annotation in Ensembl. PASS has mainly been developed by Robert Küffner in collaboration with the author and is therefore only briefly described here.

PASS is fully integrated into the ProSAS database and the ProSAS web application (see also Figure 11.1). For the public exon array datasets provided by Affymetrix for Human, Mouse and Rat, we visualize the data for all human, mouse and rat genes before and after applying PASS and display exons being absent in one of the tissues, i.e. potential splicing events identified by PASS on the exon and the protein structure level. The PASS method, its performance and its integration in ProSAS is discussed by some selected examples in the following.

11.1 Outline of the method

PASS aims at the de-novo prediction of AS events from Affymetrix exon arrays. Existing approaches (e.g. the MIDAS method proposed by Affymetrix) usually compare relative changes in the signal levels of one probeset between pairs of samples in order to detect potentially spliced exons, i.e. an exon which is present in one sample, but absent in the other.

In contrast, with PASS we are able to compare the changes in expression levels of different probesets via a direct comparison of pairs of probesets. In order to do so, we first need to account for probeset-specific response effects that cause each probeset to respond differently to the observed target abundance. Those effects are accounted for by a multiplicative scaling factor. A second source of errors is introduced by background noise and cross-hybridization. We account for those errors by a second, additive scaling factor.

Therefore, probesets of a gene are linearly scaled using additive and multiplicative scaling constants which are estimated from the distribution of sample group means across all samples. Among several strategies tested to identify the scaling constants fitting the scaling constants by range fitting (Min-Max scaling) leads to the best results in our test scenario. To scale the probeset signals, the smallest group mean is used as additive scaling factor a and the difference between the largest mean and a is assigned as multiplicative factor m. After scaling, the distribution of both probesets will cover the same range with respect to the minimal and maximal group means. No assumptions are made on the type of the probeset signal distribution.

Subsequent to scaling, we can detect pairwise alternative splicing events, i.e. pairs of exons where one probeset is present whereas the other one is absent. Such interesting pairs of probesets can be identified using a simple one-way ANOVA and the log ratio of the two scaled probeset signals.

We first compute the ratio of two scaled probeset values s_1 and s_2 within the same sample. Pseudocounts are added to reduce the bias from dividing by low (unreliable) signal levels.

$$x_{ij} = log \frac{s_1 + pseudocount}{s_2 + pseudocount}.$$

The sample mean within I replicates $\overline{x}_j$ and the mean across all samples (and J sample groups, i.e. tissues) $\overline{x}$ are computed as:

$$\overline{x}_j = \frac{1}{I} \sum_{i=1}^{I} x_{ij} \qquad \overline{x} = \frac{1}{J} \sum_{j=1}^{J} \overline{x}_j$$

We can then compute the variance *within* the I replicates of a sample group V_W and the variance *between* the J sample groups V_B as:

$$V_W = \frac{1}{J} \sum_{j=1}^{J} [\frac{1}{I} \sum_{i=1}^{I} (x_{ij} - \overline{x}_j)^2] \qquad V_B = \frac{1}{J} \sum_{j=1}^{J} (\overline{x}_j - \overline{x})^2$$

Finally, the F-value is computed as $F = \frac{V_B}{V_W}$ where large F-values thus correspond to a high probability for a pairwise alternative splicing event according our model, i.e. the variance between the sample groups V_B differs significantly from the signals values observed within the replicates V_W. Using this approach detected splicing events between pairs of probesets are indicated across all sample groups. In cases where events observed in a specific group are of interest (e.g. tissue specific absence or presence of an exon) we extract such information from pairs of probesets which are conspicuous according to ANOVA and exhibit a signal ratio of greater than a predefined threshold (e.g. 1.5) in the given sample group.

Pairwise events	Individual events	Affected Probesets	Genes	Precision
1000	694	273	9	98.2%
10000	3780	945	108	96.2%
100000	33472	10995	1339	70.2%
1000000	255902	70440	7482	46.2%

Table 11.1: Prediction accuracy of the first 1000 to 1000000 predicted splicing events by PASS with respect to Ensembl as standard of truth. Events are sorted with respect to their ANOVA F-values and it should be noted that the same event can be detected in different tissue pairs (Pairwise events vs. Individual events). The background probability of a spliced exon in Ensembl is 19.3% indicating that even one million splice events are predicted with an accuracy well above the random level.

11.2 Results and Discussion

We have evaluated the performance of the PASS method using known splice variants annotated in Ensembl on a dataset of 11 human tissues, with 3 technical replicates each, provided by Affymetrix. As shown in Table 11.1, PASS is able to predict the ”splicability” of specific exons with high accuracy, which means that exons predicted to be alternatively spliced by PASS are also often annotated to be spliced in at least one known transcript in Ensembl. The first 20000 splicing events can be predicted at an estimated accuracy of larger than 90% and even the first million of events affecting almost 7500 genes are predicted with an accuracy of more than 46% and therefore well above the background probability of 19.3% corresponding to the percentage of spliced exons in the Ensembl dataset.

For several cases PASS predicts splicing events which are known to occur in the corresponding tissue. Two of those examples are shown in Figure 11.1 which also displays the analysis capabilities of the PASS method in correspondence with the ProSAS web application.

The upper panel shows a gene known as UAP1_HUMAN (ENSG00000117143 on chromosome 1) which has three different splice variants (transcripts) annotated in Ensembl. Two splice variants are especially interesting and well characterized. Transcript ENST00000367926, in which exon number 8 is skipped, is also known as AGX1. The protein is specifically found in testis, is known to form homodimers and binds GalNAc-1-p with a very high affinity. In contrast, ENST00000367925 (also known as AGX2) contains all 10 exons of the gene, binds GclNac-1-p with a high affinity and does not form homodimers but is monomeric. On the protein structure level (shown in the upper right corner and usually part of the transcript panel shown in Figure 10.3) the removal event of exon number 8 corresponds to a part on the surface of the protein which is not completely crystalized and (not shown in the Figure) may play an important role in the formation of the protein-protein interaction interface. As shown in the PASS panel exon number 8 (corresponding to the light magenta line) behaves very similar to all other probesets in all tissues except for samples 28-30 which represent testes samples. Therefore, PASS correctly identifies a functionally important splice variant in testes.

The lower panel shows a gene known as FYN_HUMAN (ENSG00000010810 on chromosome 6). Transcripts of this gene play important roles in the control of cell growth and the regulation of intracellular calcium levels. The gene has several annotated splice variants in Ensembl and two of them are formed by the alternative, mutually exclusive usage of exons 6 and 7. Isoform 1 where exon 6 used and 7 is skipped is known to be brain specific and plays an important role in axon growth and neurite extension in developing and mature human brain. This isoform (exon 7 absent, exon 6 present) is also predicted by PASS to be expressed in samples 4-6 which correspond to human brain samples. On the protein structure level the mutually exclusive splicing event leads to exchanging a domain linker region (shown in blue on the structure level).

Even though the PASS method performs accurate in terms of detecting exons known to be spliced (see Table 11.1) and can reliably predict tissue specific splice forms (two examples have been discussed above) several open questions remain which are hard to address within the Affymetrix exon array platform. Especially mixtures of isoforms which are likely to exist in many tissues (e.g. isoforms of *tau* in brain as discussed above in section 6.3) are a problem as they can for example lead to the fact that an exon / probeset is not found to be significantly different across the tissues even though it is alternatively spliced. Also, identifying the correct isoform, i.e. the exons the final transcript is composed of, from a set of pairwise events predicted by PASS in the presence of such mixtures is a hard problem if not even impossible at all. Finally, microarray platforms will probably be not sensitive enough to detect minor form transcripts, i.e. transcripts accounting for substantially less than 50% of the mRNA of a given gene.

Some of those tasks may be addressed by future extensions of the PASS method, others may be intrinsic problems of the experimental approach itself which can not be accounted for by any method analyzing the data. Compared to other approaches to analyze Affymetrix exon array data PASS offers an interesting alternative with its unique features which allow for the detection of a specific set of splicing events with a very high accuracy. In combination with the ProSAS database PASS provides interesting and orthogonal knowledge on the nature and tissue-specific expression of specific splice isoforms as shown above.

Expression data analysis will change dramatically in the next years. The current methods have to face recent developments in sequencing technologies ("next-generation-sequencing") which are likely to replace standard chip-based technologies in the near future. Those sequencing approaches, e.g. Illumina's GenomeAnalyzer method and others, allow to sequence the complete mRNA of a sample via millions of short reads, at a cost which will soon be comparable to the cost of today's chip experiments. The technology promises a much easier detection of splice variants (and also mixtures of them) via sequencing all exon-exon junctions present in the transcripts as well as a quantitative analysis of all mRNA transcripts in the cell. Those data require for different analysis strategies and will surely substantially change our current understanding of alternative splicing.

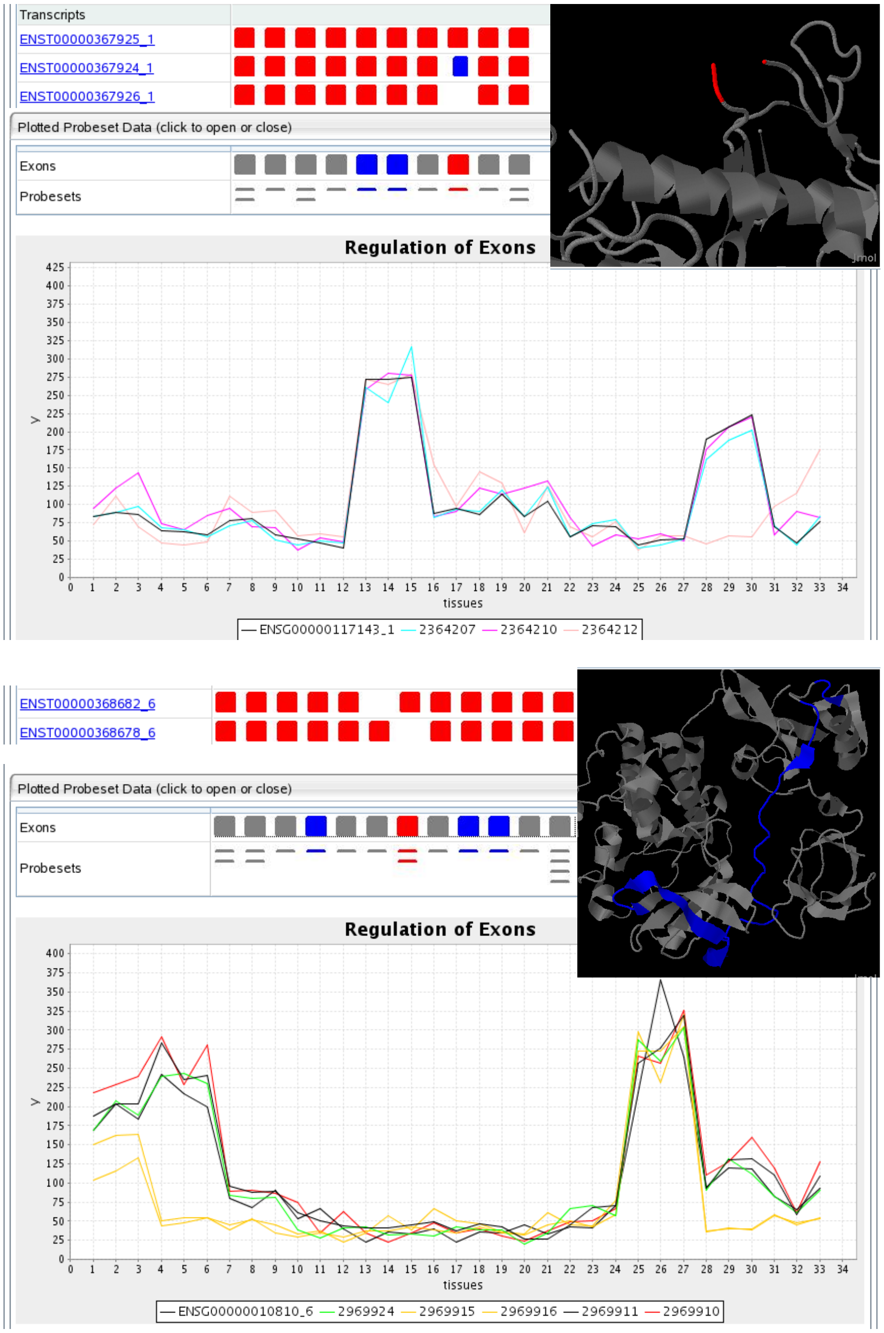


Figure 11.1: Shows two tissue-specific splice variants detected by PASS. The upper panel shows a testes specific splice variant (probeset of exon number 8 is missing in samples 28-30 corresponding to the three testes specific replicates) while the lower panel displays the identification of a brain-specific splice variant (missing exon 7) corresponding to samples 4-6. Both examples are discussed in the text.

Chapter 12

Alternative splicing and proteome complexity in mass spectrometry data

We have already mentioned that on the mRNA level, splicing seems to be the rule rather than the exception. But although we could show in Chapter 7 that non-trivial isoforms can have well defined functions and the tolerance of a structure against major rearrangements may be linked to the evolutionary history of the corresponding protein fold it remains unclear to which extent structurally non-trivial isoforms are present on the proteome level in the cell.

This is due to a major problem in the analysis of alternative splicing isoforms which is their detection by experimental methods and the following annotation in public databases. To date, most of our knowledge on the nature of splicing isoforms stems from the analysis of EST or cDNA libraries or more recently from next-generation sequencing experiments [151]. Therefore most alternative transcripts have only been observed on the mRNA level. Even in the manually annotated Swissprot database only few isoforms are confirmed on the protein level and for even less the function of the novel protein product is known and annotated.

Given now the large complexity of splicing observed on the protein structure level we do not know to what extent the large number of splicing isoforms observed on the mRNA level lead to stable protein products. Because of nonsense mediated decay or degradation right after their translation they may not contribute many novel function to the proteome.

Recent developments in mass spectrometry [32] (MS) allow to analyze alternative splicing on a large scale on the proteome level as they have led to the possibility to determine the (quantitative) presence of all proteins in complete organisms or tissues ([29], [42], [117], [184]). Here we analyze alternative splicing isoforms found in Drosophila melanogaster [29], Mus musculus [184] and Homo sapiens [42] in three large-scale datasets of peptides identified in mass spectrometry experiments and available in PeptideAtlas [41]. Those highly confident peptide identifications allow for the first time for a large scale analysis of the presence and complexity of alternative splicing events on the proteome level. Therefore, this data has the potential to sharpen our understanding on the mechanisms to generate functional and structural diversity via alternative splicing on the proteome level.

12.1 Methods

12.1.1 Identification of isoforms in precompiled MS peptide datasets

We have downloaded three sets of peptides identified by mass spectrometry experiments from PeptideAtlas [41]. Those contain the peptide atlas build for the Human proteome [42] (April 2007), the Mouse plasma proteome [184] (May 2007) as well as the Drosophila proteome [29] (March 2007). All contain only highly confident peptide identifications which have a PeptideProphet [82] score greater than 0.9.

The peptides are matched against all proteins and splice variants known for the corresponding organism in the Uniprot [175] database. In order to identify known splice variants of human genes, the Uniprot datasets have been processed in the following way:

We downloaded all non-fragmented proteins for the corresponding organism from the UniProt database. From those entries all duplicates or entries being equal to known VARSPLIC splice isoforms were removed (keeping the reviewed entries and known VARSPLIC isoforms were possible). Additionally, all protein entries which were not annotated to a UniGene or Ensembl gene were removed from the dataset.

In a second step, we identified known splice variants in UniProt. Proteins annotated in the Swissprot complement of Uniprot often have manually annotated splice variants (VARSPLIC annotation) and all those are used as isoforms in our study. Second, we filter the entries for proteins which are found to belong to the same gene (as identified by the same UniGene or Ensembl identifier) and therefore are likely to represent splicing isoforms. One isoform is used as reference for the gene and all other entries in UniProt are marked as splice variants. As reference protein we choose the protein of the gene which is marked as being ”reviewed” in Uniprot (those usually correspond to Swissprot entries). If no reviewed entry is available, we require at least one protein to have annotated evidence on the protein level. In the case of several entries known on the protein level, we use the longest isoform as reference. If no entry of the gene has such evidence, the entries remain in the dataset (to avoid an overestimation of unique peptides) but we do not treat them as isoforms.

We then match all peptides in the corresponding dataset against all protein sequences (reference and isoform sequences) of an organism. Only perfect, un-gapped peptide matches are taken into account. We then identify all peptides which match uniquely to a known splice variant of a reference protein, i.e. which do not match to any other database protein and also not to the reference protein. All isoforms which are confirmed by at least one unique peptide are used for further analyses. This results in a conservative estimate on the presence of splicing isoforms in the dataset.

12.1.2 Expected versus observed unique isoform peptides

In order to determine if unique isoform peptides are underrepresented in MS data we computed the expected number of such peptides from an artificial tryptic digest (cleavage rules followed Swissprot's PeptideMass tool) of all proteins of the respective organism in Uniprot. The artificial peptides were then matched against all Uniprot sequences of the corresponding organism using the same approach as described above for the MS confirmed peptides. We therefore identified peptides matching uniquely to reference or alternative isoforms. Only proteins which had annotated splicing isoforms were taken into account to measure the fraction of peptides matching uniquely to the isoform or to the reference in order to avoid a bias introduced by proteins which have no splicing annotation and therefore only consist of reference peptides.

12.1.3 Sequence-based analysis of alternative splicing

In order to analyze the effects of alternative splicing events on the sequence level we used pattern information available from the InterPro database [109] like PFAM domains, transmembrane domains as identified by TMHMM or TargetP / SignalP peptides. Matches of those patterns to proteins with confirmed splicing isoforms have been obtained from the SIMAP [127] database. A pattern is said to be significantly affected by a splicing event if at least 75% of the residues involved in the pattern / domain are changed or removed.

12.1.4 Protein structure prediction pipeline

The structure prediction pipeline and features assignment protocols are discussed in the context of the ProSAS database in Chapter 10. All data on protein structures and structural features has been obtained from ProSAS for isoforms identified in the peptide datasets.

12.2 Results

In the following we will describe and discuss the results of our comprehensive, structure-aided analysis of alternative splicing isoforms in Human, Mouse and Rat which are confirmed by highly confident MS peptide identifications (MS-confirmed isoforms). The basic principle of the method is shown in Figure 12.1. All peptides contained in those datasets are identified with PeptideProphet scores greater than 0.9 and are often identified by several spectra and in different samples.

We will first of all discuss the presence of alternative transcripts in MS data which are identified by a very conservative approach (see Section 12.1) and correlate the observed number of isoforms with what is theoretically expected. We will then analyze some interesting functional effects of alternative splicing events observed in MS data on the sequence and protein domain level, which reveals some interesting and obviously very common mechanisms of splicing to generate functional diversity. We will finally describe the expected effects of splicing events on the protein structure level in order to characterize the structural complexity of splicing isoforms present on the proteome level.

All isoforms observed in MS data can be interactively analyzed in a structure-based context in a complement of the ProSAS database (see Chapter 10). The database complement can be accessed at http://www.bio.ifi.lmu.de/ProSAS/MSSupplement.html.

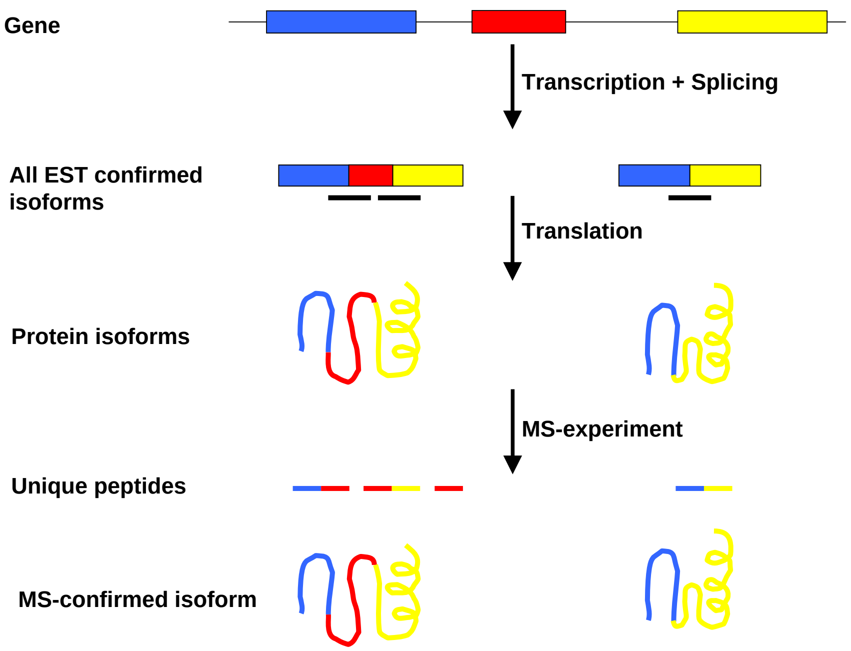


Figure 12.1: EST libraries can confirm isoforms by unique EST sequences (black bars) which match i.e. to specific exon boundaries or unique exons. Most isoforms stored in public databases are therefore confirmed on the mRNA level. Those isoforms may be translated into stable proteins or deleted by nonsense-mediated decay or protein degradation right after translation. Stable protein products can then be detected by mass spectrometry, in analogy to EST sequences via unique peptides matching exon boundaries or other unique parts of the isoforms. Those isoforms comprise the set of MS-confirmed isoforms and are likely to exist on the proteome level.

12.2.1 Presence of isoforms in mass spectrometry data

In order to collect all MS-confirmed isoforms, peptides identified by MS experiments in Human [42], Mouse [184] and Drosophila [29] were mapped onto all UniProt proteins, including major transcripts (reference proteins) and known isoforms, of the corresponding organism in a first step.

Reference entries and isoforms in UniProt were defined as described in Section 12.1 and the content of our final datasets is shown in Table 12.1. Most known splice variants are annotated for Human (18930), followed by Mouse (12142) and Drosophila (3116) which may display the larger number of EST / cDNA data available for Human and Mouse and not necessarily a larger presence of alternatively spliced transcripts in Mouse and Human compared to Drosophila.

We then matched the peptides in the three datasets against all proteins of the respective organism. Only exact, un-gapped hits were allowed and we could match 95% of the Human peptides, 91% of the Mouse peptides and 98% of the Drosophila peptides onto at least one protein. The detailed results are shown in Table 12.2.

Organism	Genes	Uniprot entries	VARSPLIC Isoforms	UNIPROT isoforms	Peptides
Human	21597	35773	7122	11808	84087
Mouse	20820	34311	3110	9032	13779
Drosophila	12258	20727	286	2830	74541

Table 12.1: Statistics on the UniProt datasets used for the three organisms. The number of VARSPLIC isoforms displays isoforms annotated in Swissprot while UNIPROT isoforms correspond to isoforms identified in UniProt by the procedure described in Section 12.1. The peptides column displays the number of peptides available in the mass spectrometry datasets of the corresponding organisms in PeptideAtlas.

Organism	Matching peptides	Unique peptides	Unique for isoform	Isoforms confirmed Varsplic / Uniprot
Human	81653	37858	619	216 / 266
Mouse	12565	6838	119	29 / 57
Drosophila	73546	42491	651	74 / 156

Table 12.2: Displays the statistics of mapping the mass spectrometry peptides onto all UniProt entries. The ”Unique for isoform” column displays the number of peptides in the dataset which match uniquely to a known isoform from Varsplic or UniProt. The number of distinct isoforms from Varsplic and Uniprot is shown in the ”Isoform confirmed” column.

Isoforms confirmed in MS experiments

MS-data can confirm 482 known isoforms in Human, 86 isoforms in Mouse and 230 isoforms in Drosophila by unique peptides matching no other protein in the dataset. 34% of the isoforms in Mouse, 31% in Drosophila and 45% in Human correspond to isoforms annotated in Swissprot (VARSPLIC), while the other isoforms are annotated in UniProt to belong to the same gene but have not been annotated manually in Swissprot. Nevertheless, the number of confirmed VARSPLIC isoforms is consistently higher than expected from their corresponding background probabilities where 25% from all Mouse isoforms, 9% of the Drosophila isoforms and 37% of Human isoforms are annotated in Swissprot. This may be due to their higher quality and the fact that they also correspond to more frequent and more stable isoforms.

When analyzing the 1350 events annotated for the MS-confirmed isoforms, we find that 23% correspond to removal events while 77% of the events replace a part of the reference transcript by a different sequence. In 30% of the cases the replacement sequences are shorter than the original sequences (corresponding to deletions), 42% of the replacement sequences are longer than the original (corresponding to insertion events) while 28% are of the same size as the original sequence. It is noteworthy that in the complete VARSPLIC dataset the fraction of removal events is 58% while only 42% of the events correspond to replacements. This bias towards replacement events is likely due to the fact that deletion events can only be confirmed by a peptide matching the new exon boundary arising from the splicing event. In contrast, splicing events leading to replaced sequences allow potentially more different peptides to be observed especially for longer

Dataset	Peptides	Unique	Unique (Reference)	Unique (Isoform)	Fraction
Human (art.)	505544	100157	65608	34549	34.5%
Human (obs.)	52085	8290	7671	619	7.5%
Mouse (art.)	422220	81299	58998	22301	27.4%
Mouse (obs.)	6895	1168	1049	119	10.1%
Drosophila (art.)	220298	22975	14856	8119	35.3%
Drosophila (obs.)	34425	3370	2719	651	19.3%

Table 12.3: Displays the expected (art. dataset) and observed (obs. dataset) number of peptides unique for reference proteins and alternatively spliced isoforms. Splice isoforms are consistently underrepresented in MS data across all three organisms, although the proportions vary.

parts being replaced.

Underrepresentation of alternative splicing isoforms in MS data

The number of isoforms confirmed by MS data and the number of peptides confirming isoforms in a unique manner appear to be rather small at the first glance. In order to find out if and to what extent splicing isoforms are underrepresented in MS data we compared the observed number of unique peptides confirming isoforms with the theoretically expected number of such peptides (see Section 12.1 for details). As shown in Table 12.3 artificial peptides uniquely confirming the reference or major transcript are roughly two times more frequent than peptides confirming an isoform as shown in all three artificial datasets. This can be explained by the fact that many splicing events correspond to deletions and therefore lead to fewer peptides which could uniquely confirm the isoform (in the case of cassette exon removal, one of the most frequent events, exactly one peptide can confirm the new exon border). But despite this bias observed in the artificial data, splice isoforms are still less frequently found in MS data than expected. The fraction varies from 34.5% expected compared to 7.5% observed unique isoform peptides in Human to 35.3% expected versus 19.3% observed isoform peptides in Drosophila. Interestingly the proportions are negatively correlated with the number of known isoforms which may point to the fact that many isoforms annotated for Human correspond to rare events (but contributing potentially unique peptides) while the relatively few isoforms annotated for Drosophila correspond to more frequent ones. Also, a certain bias towards the identification of the reference transcript in MS data can be expected as it represents the fact that major isoforms should be more frequent than splice isoforms. The bias may additionally be enhanced by the current peptide identification pipelines which aim at the identification of highly significant peptides rather than the identification of rare splice isoforms.

12.2.2 Changing signal peptides, localization and domain composition via alternative splicing

In the following we will describe some interesting functional implications of alternative splicing events observed in our MS dataset confirms the soundness of the identified splicing products.

We will describe some very interesting functional mechanisms of changes in localization signals, transmembrane helices and the domain composition of multi-domain proteins which are discussed in the biological literature.

Terminal modifications of protein domains

A large fraction of isoforms (81%) have modifications at the terminal ends of the reference protein which may contribute to changes in localization or degradation signals. Using SignalP and TargetP annotations [46] from InterProScan [126], we find 117 SignalP and 160 TargetP peptide matches. 25% and 31%, respectively, of those are found to be altered by alternative splicing potentially accounting for changes in the cellular localization or faster degradation.

One example is the Human Deoxyuridine Triphosphate Nucleotidohydrolase gene (P33316) which is important for DNA replication. Two distinct isoforms DUT-N and DUT-M originating from the same gene are located in the nucleus or targeted to the Mitochondria respectively [91] depending on the unique N-terminal signal which can be altered by alternative splicing. The existence of both isoforms is confirmed by unique mass spectrometry peptides in our dataset. Another example is the BACH (Human Brain Acyl-CoA Hydrolase) protein (O00154) in Human, where different isoforms (including the isoform 3 confirmed by mass spectrometry data) [180] are shown to be located in the cytoplasm (and potentially targeted to the nucleus) or transported to the Mitochondria.

Known isoforms with alterations of signal peptides in our dataset contain isoforms important for brain development and neuronal activity. While the Vol-L and Vol-S isoforms of ITA3 (O44386) from Drosophila [60] take part in Integrin mediated short-term memory learning, the isoforms of the Gamma-aminobutyric acid receptor subunit beta-3 (P28472) are involved in the synaptic inhibition of neuronal activity [84]. The detailed functions of those signal peptide alterations are unknown in both cases.

Alternative splicing of transmembrane regions

Similar to changes in the cellular localization of a protein via alternative splicing of target or signal peptides, the removal or the inclusion of trans-membrane domains can lead to membrane-bound or secreted / cytoplasmic isoforms. According to TMHMM [88] annotations 79 proteins in our dataset contain transmembrane domains from which 33 (41%) appear to be alternatively spliced. Among the isoforms discovered in our dataset are several interesting cases also described in the biological literature.

ADAM23 (O75077), a metalloprotease and disintegrin protein which plays an important role in brain development gives rise to three isoforms which differ in their C-terminal transmembrane domain. While the native isoform and isoform 2 have different transmembrane domains, isoform 3 lacks this domain and is found to be secreted. Isoform 3 (confirmed in MS-data) can then compete with the native, membrane-bound variant for binding Integrin. Different ADAM23 isoforms may therefore alter cell-cell or cell-matrix interaction during brain development [152]. Another interesting case is TNR25 (Q93038) in Human where several isoforms are produced. TNR25 is involved in receptor signaling inducing apoptosis. Different isoforms are found to be integrated into the membrane or secreted from the cell. The existence of the secreted isoform number 8 can be specifically confirmed by MS data. The secreted forms seem to bind extracellular ligands and block receptor signaling before the signal is transmitted into the cell. Therefore they are suggested to set the balance between TNR25 mediated apoptosis and cell survival [140]. Both cases represent examples for an interesting functional principle of splicing resulting from secreted or membrane-bound isoforms. While membrane bound isoforms are responsible for the interaction of the cell with its environment, secreted isoforms are able to block extra-cellular signals before they reach the cell by acting as regulatory signal sinks.

Changing domain composition by alternative splicing events

In order to analyze the frequency of splicing events altering the domain composition of multi-domain proteins on a sequence basis we used PFAM patterns. Out of the 2002 patterns annotated for the 733 reference proteins, 370 (18%) are found to be altered significantly (more than 75% of the residues of the domain affected) by alternative splicing events. Unfortunately the dataset is too small to identify patterns being spliced significantly more often than expected e.g. via the LOD analysis used in Chapter 8.

Nevertheless, two interesting examples for changing the function of a protein via changing its domain composition are the Human proteins ZN396 (Q96N95) and ZSCA1 (Q8NBB4). Both consist of a SCAN-box domain as well as three C_2H_2-Zinc finger motifs and members of this family often act as transcriptional repressors. In both cases isoforms are confirmed by MS displaying the deletion of all three zinc-finger domains leaving only the SCAN domain. Isolated SCAN domains have been shown to act as selective oligomerization domains which may modulate the formation of functional SCAN domain zinc-finger transcription factors [138] and the isoform of ZN396 has been shown to e.g. hetero-associate with the native variant [176]. The fact that two of those events are confirmed in MS-data may point to a large importance of this mechanism to regulate the transcriptional landscape of the human genome via changes in the domain composition of multi-domain transcription factors.

12.2.3 Structural analysis of isoforms identified in mass spectrometry data

Many biological properties of proteins can only be understood and analyzed in the context of protein structures. Splicing can alter existing, stable protein domains via insertion, deletion or replacement of one or more exons and it was shown recently (e.g. Chapter 7) that the expected effects of splicing on the structure level are often non-trivial as integral parts of a structure may be changed or removed. This unexpectedly large complexity has led to the question if splicing can contribute to functional diversity or if the outcome of most splicing events is an unstable, non-functional protein [161]. Our dataset of MS-confirmed isoforms allows for the first time for an in depth and large-scale analysis of splicing products which exist on the protein level and are likely not to be prone to nonsense mediated decay or degradation. In order to characterize the alterations on the protein structure level to be found in MS-confirmed isoforms we predicted the protein structures of all isoforms in our dataset and structurally analyzed the regions being alternatively spliced as described in Section 12.1.

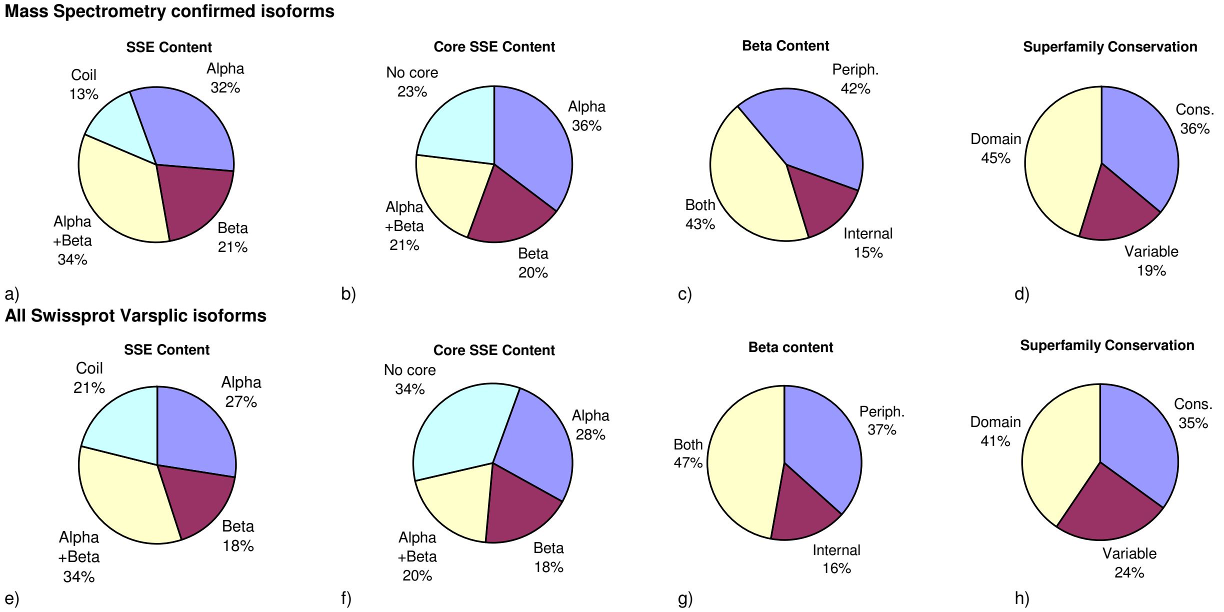


Figure 12.2: Displays the complexity of alternative splicing events observed in Mass spectrometry data on the protein structure level (a-d) in terms of affected secondary structure elements (SSE content), core secondary structure elements (Core SSE content) and beta strands (Beta content). Additionally, we show how many isoforms change complete domains or conserved and variable parts of the corresponding SCOP superfamily (Superfamily conservation). Charts e-h display the same analysis on the complete set of all Swissprot annotated isoforms which shows a very similar level of structural complexity.

Structural complexity of MS-confirmed isoforms:

In total, we are able to model 182 splicing events from 116 protein isoforms (15% of all MS confirmed isoforms) on the protein structure level, meaning that in those cases the spliced region is covered with structure and not the complete PDB domain is affected by the splicing event (which is the case for 61 isoforms). 78 of the modelled events correspond to removal events (or replacement events by a different, shorter sequence), 82 events represent replacements by different sequences of the same length while 22 events insert novel parts.

The results of our feature-based analysis on the complexity of splicing in MS-confirmed isoforms are shown in Figure 12.2. Parts a-c show that most spliced regions correspond to well-structured parts of the proteins in terms of their secondary structure content. Indeed, the fraction of non-core secondary structure elements (as defined in [171]) is a bit larger than the percentage of coil regions indicating that about 10% of the secondary structure elements being spliced have no contact to other, integral elements of the protein structure.

Peripheral beta strands (located on the edges of larger beta sheets) appear to be more frequently spliced than internal strands of a beta sheet. This is to be expected from a structural point of view as only hydrogen bonds on one face of the sheet are altered by such a splicing event. Also, we frequently observe the removal of several internal and peripheral strands in a block from the edge of a larger beta sheet (corresponding to the ”both” slice in Figure 12.2c). Despite the fact that those events affect larger parts of the structure they may have similar effects like the removal of a peripheral strand as only hydrogen bonds on one side of the remaining strands are altered.

While secondary structure based features give only an incomplete picture on the structural complexity of a splicing event, we have introduced the use of evolutionary knowledge about the corresponding protein family in the analysis of alternative splicing (see Chapter 7. The annotation of evolutionary highly conserved and variable regions of a protein family allows characterizing the importance of a region for the viability and stability of a structure. Figure 1d displays the percentage of splicing events removing highly conserved or variable parts or of the corresponding SCOP superfamily or complete domains of the template structure. As shown, 36% of the events confirmed by MS data remove conserved elements, while only 19% of the events fall into variable parts of the corresponding SCOP superfamily. The other 45% correspond to removals of complete protein domains from multi-domain proteins (including the 61 isoforms where complete PDB structures are altered / removed).

Comparison with structural complexity of all VARSPLIC isoforms:

In order to compare the complexity of MS-confirmed splicing events with the structural complexity observed in all Swissprot annotated VARSPLIC isoforms, including those only confirmed by EST and cDNA data, we carried out the same analysis for all structurally modeled Swissprot proteins with VARSPLIC annotations. The results of the analysis of 4866 annotated and structurally modeled splicing events are shown in Figure 12.2e-h). As it can be seen, the structural complexity of those splicing events turns out to be very similar to the complexity observed in MS data with respect to all criteria applied. Especially the fraction of events affecting conserved and variable superfamily regions as well as complete domains turns out to be the same. Therefore, nonsense mediated decay and protein degradation do not seem to preferentially target those isoforms which appear to be complex on the structure level pointing to the existence of a large number of non-trivial, structurally complex isoforms on the proteome level.

Functionally interesting examples of structurally covered events:

Figure 2 shows nine examples for the observed structural complexity of isoforms in MS data.

Several events appear to be easy to explain as rather unstructured regions are being altered by splicing events. An interesting case is ATNA_DROME (P13607, Figure 12.3a) where a linker region between two helices is changed by a splicing event. This loop region of ATNA, which catalyzes the hydrolysis of ATP coupled with the exchange of sodium and potassium ions across the plasma membrane, is known to be of functional importance and changing this part via alternative splicing is suggested to lead to changes in pump kinetics, ion selectivity or the regulatory pattern of the corresponding channel [119]. The removal of several domains in MASP2 (Q91WP0) in Mouse (one part is shown in Figure 12.3b), a serum protease that plays an important role in the activation of the complement system, leads an isoform which binds different binding partners compared to the native variant [149] and the removal of several HAT repeats, which are

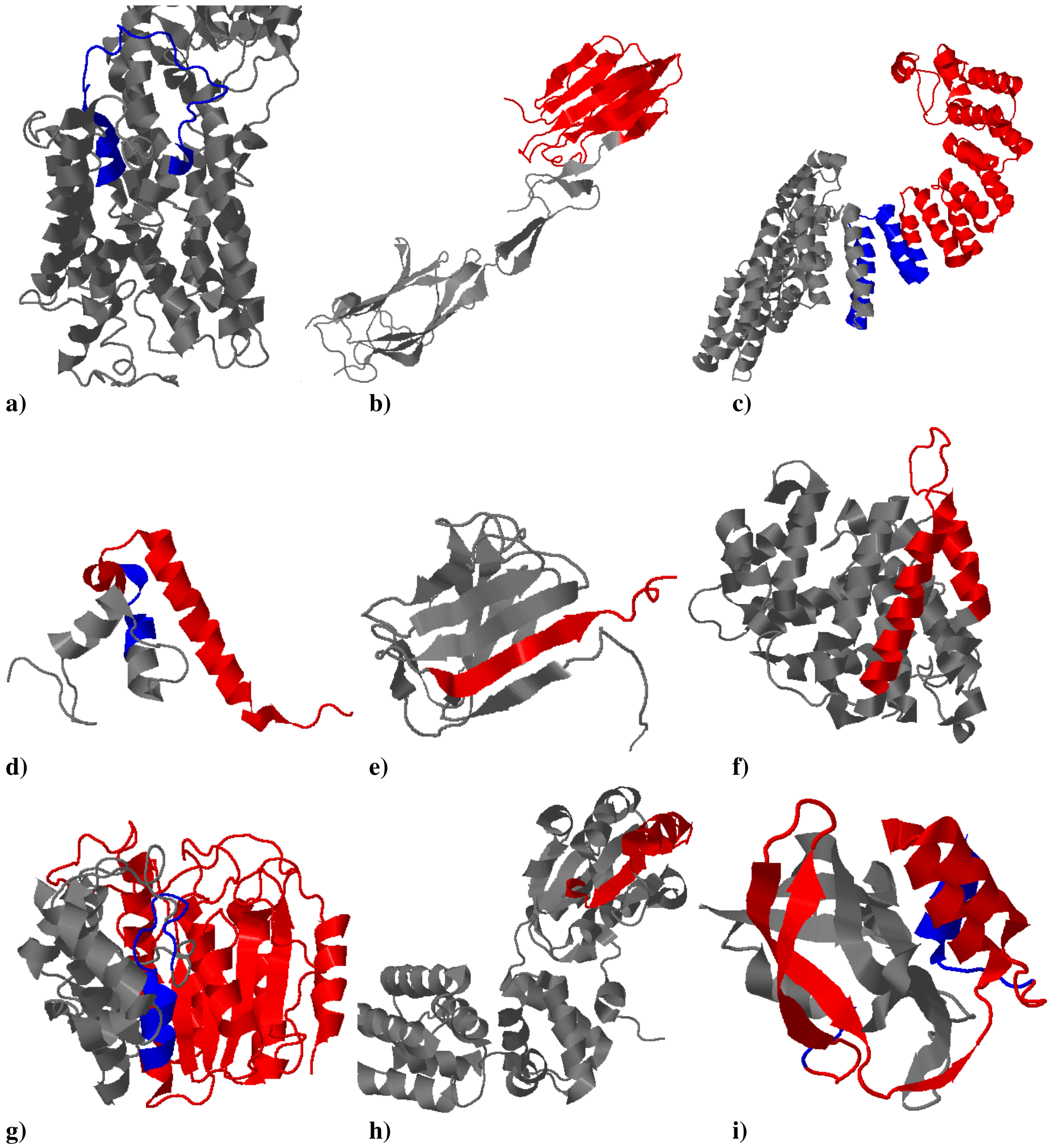


Figure 12.3: Red parts are being removed in the isoform structure, blue regions are replaced by different sequences. a) the replacement of a loop region in ATNA_DROME b) removal of a CUB domain and several others (not shown on the structure) from MASP2_MOUSE c) removal of several HAT repeats in SUF_DROME d) shows part of the structure of DFFA_HUMAN and the alteration of the C-terminal part of the molecule e) deletion of an internal β-strand in LMNA_HUMAN (only a part of the structure is known) f) deletion of one helix-turn-helix motif in PTPA_HUMAN g) deletion of a large part of the central β-sheet in HDAC8_HUMAN h) deletion of one internal α-β motif in RFC2_HUMAN and i) complex deletion of α helices and β-strands in CG024_HUMAN.

well-known protein-protein interaction motifs in SUF_DROME (P25991, Figure 12.3c) displays potential alterations in binding partners through domain removal events.

While those events can be expected to be tolerable for the corresponding protein structure, the tolerance against larger removal events of highly conserved elements is not easy to explain. Nevertheless such events are also often observed in MS confirmed isoforms. Figures 12.3d-i display several categories of such deletion events, Figure 12.3d shows the deletion of a C-terminal helical motif from DFFA_HUMAN (O00273, DNA fragmentation factor alpha) which triggers DNA fragmentation during apoptosis and leads to an isoform known as DFF35. The isoform can bind its native target DFF40 more strongly than the native isoform DFF45 and may play an important role in putting the death machinery of the cell under strict control [61]. The removal of an internal beta-strand from LMNA_HUMAN (P02545) shown in Figure 12.3e is proposed to result in a different localization and potentially a different function [99]. Further non-trivial splice variants whose function is unclear but for which tissue specific expression has been annotated include isoform 3 of HDAC8_HUMAN (Q9BY41, Figure 12.3g) removing a large part of the central beta sheet, isoform 4 from PTPA_HUMAN (Q15257, Figure 12.3f), a regulatory subunit of serine/threonine protein phospatase 2A, where one conserved helix-turn-helix motif is removed from the structure and whose tissue specific expression has been shown [78]. Figures 12.3h and 12.3i display the complex removal of a beta-hairpin and several helical parts from CG024_HUMAN (O75223) and the removal of one alpha-beta motif from RFC2_HUMAN (P35250). In both cases the function is not described in the biological literature.

12.3 Discussion

In this chapter we described our analysis of three datasets of highly confident MS peptides. We have identified a set of 796 splicing isoforms uniquely confirmed by those short fragments. Those peptides correspond to very conservative confirmations of the existence of specific isoforms on the protein level and allow for the first time to comprehensively study alternative splicing on the proteome instead of the transcriptome level. We have implemented a complement to the ProSAS database which allows the large scale, structure aided analysis of protein isoforms confirmed by MS data and can be accessed at http://www.bio.ifi.lmu.de/ProSAS/MSSupplement.html. ProSAS will be further extended by more isoforms confirmed in mass spectrometry experiments in the future.

In our dataset we can confirm a large number known effects of alternative splicing on the localization of proteins and their domain composition. Those include changes in targeting signals as well as the well documented examples of membrane-bound or secreted isoforms playing important roles in brain development and apoptosis. MS peptides can also confirm an interesting set of splicing events altering the domain composition of SCAN-Box containing Zinc-Finger proteins where splicing events remove all Zinc-Finger motifs from those proteins leaving only the SCAN-box domain. The function of those single SCAN-box domains has been shown to have important effects on transcriptional regulation which differ significantly from the native isoform of the gene.

We also carried out a comprehensive, structure-based analysis of the splicing events observed in our dataset. This analysis reveals a very large complexity of splicing observed in MS data on the protein structure level which is comparable to the complexity observed in all Swissprot annotated splicing events. Therefore there seems to be no preference for structurally complex isoforms to the prone to nonsense-mediated decay or protein degradation compared to other splicing events annotated in EST or cDNA data. This is further support for the hypothesis that protein structures are highly tolerant even against major rearrangements and the effects of splicing on the protein structure level are much more complex than previously thought 7.

Our study also shows that, to date, we know only little about the structure and function of most isoforms annotated in public databases (in our case in UniProt). Novel experimental techniques like mass spectrometry or exon-level micro-arrays will further help to elucidate the time and tissue specific expression patterns of splicing isoforms and therefore will add another dimension to our understanding of their functional importance. Nevertheless, in order to fully understand and characterize the complexity of proteomes of higher organisms and the contribution of alternative splicing to this complexity we need an integrated and concerted analysis of novel experimental data in the context of protein function, protein structures and protein-protein interaction networks.

12.4 Outlook: Identification of novel splice variants in mass spectrometry data

The analysis carried out in this chapter so far focused on the analysis and identification of already known splice variants (annotated in Uniprot or Swissprot) in precompiled high-confidence peptide datasets obtained from PeptideAtlas. In order to identify such high-confidence peptides, most mass spectrometry pipelines rely on reference databases of known protein sequences (like UniProt or Ensembl) to assign spectra to peptides. The protein sequences in the database are digested *in silico* and the theoretical mass spectra resulting from this *in silico* digest are then matched against the spectra observed in the experiment. The theoretical spectrum matching best to an observed one is used to assign the sequence of the peptide. Pipelines based on this protocol are therefore limited to the detection of known splice variants and proteins provided with the database of known sequences. They are not able to identify peptides resulting from any unknown splice variant or a novel genomic locus in the assignment process and such spectra are lost. One has to note that splicing variation is only one of several factors like phosphorylation and other post-translational modification which account for the usually quite large number of spectra which can not be reliably assigned to a spectrum.

One possible solution to this problem is to assign peptide sequences to spectra based on a *de novo* prediction of the peptide sequence from the spectrum. Such approaches have been suggested and successfully used for example by the group of Pavel Pevzner [51]. Here, we take a different approach to address this problem which has the advantage, that it can easily be combined with existing pipelines. We suggest to enumerate all theoretically possible splice junctions based on current gene structure annotations and search against such a database. More specifically, we built a database of all pairs of exons e_i and e_j (where $i < j$) of a gene comprising all potential exon junctions and translate the sequence resulting from this combination in all three open reading frames. If the combination translates to a protein sequence which does not contain any stop codons (except for the last codon) the combination is, in principle, interesting. We then apply tryptic digest rules (as described in the http://www.expasy.ch PeptideCutter tool documentation) to the protein sequence of the potential exon combination and disregard all peptides which result in a fragment of a length smaller than 5 around the potentially new splice junction.

Results

In the following we will briefly outline the results of using the proposed database for the identification of novel splice junctions on a large-scale mass spectrometry dataset of mouse plasma [184]. The raw data has, again, been obtained from the PeptideAtlas data repository. In order to process the data, we built a pipeline containing the free X!Tandem [34] program to search a set of spectra against a specified database and PeptideProphet [82] to additionally estimate the significance of the peptides identified in a dataset.

We first of all generated the database of potential exon junctions for Mus musculus using the protocol described above. Over all, 18881 genes consisting of 205.732 exons represent multi-exon genes in which 183.508 exon junctions are observed (i.e. found in at least one known Ensembl transcript). To generate our database we tested 5.731.788 theoretical exon junctions out of which 1.415.400 (24.7%) lead to at least one open reading frame which is not interrupted by a stop codon. Requiring a peptide of length 5 or larger crossing the new exon border which results from an *in silico* tryptic digest leads to 557.719 valid peptides. 26.168 of them are non-unique and are therefore present only once in the database, 73.929 represent peptides resulting from known exon junctions, while 457622 peptides represent new, previously undiscovered exon-exon junctions.

As a first case study, we downloaded the mouse plasma PeptideAtlas data [184] consisting of 449 raw mass spectrometry files. We then matched the spectra against a database containing all known mouse transcripts as well as all potential exon junctions using the X!Tandem software. We then required a X!Tandem E-value of smaller than 0.001 (which was also used by Zhang et al. [184]) in order to identify confident peptide assignments. Additionally, we use PeptideProphet which allows to assign a probability to every peptide that the peptide belongs to the fraction of true positives in the dataset. Usually, a probability of 0.9 is used as a threshold [184]. In our case, this corresponds to less than 1% false positives to be expected in the final peptide set as estimated by PeptideProphet.

With this setup and cutoff, we identify 1853 distinct genes by a total of 11600 unique peptides. 119 of those correspond to unknown ones which are identified only in our database of exon junctions. Among those, 32 peptides are identified in multiple runs (i.e. multiple raw mass spectrometry data files) with E-values smaller than 0.001. Using PeptideProphet to assign confidence scores to the peptides, 792 novel splice-junction peptides reach a PeptideProphet score of larger than 0.9.

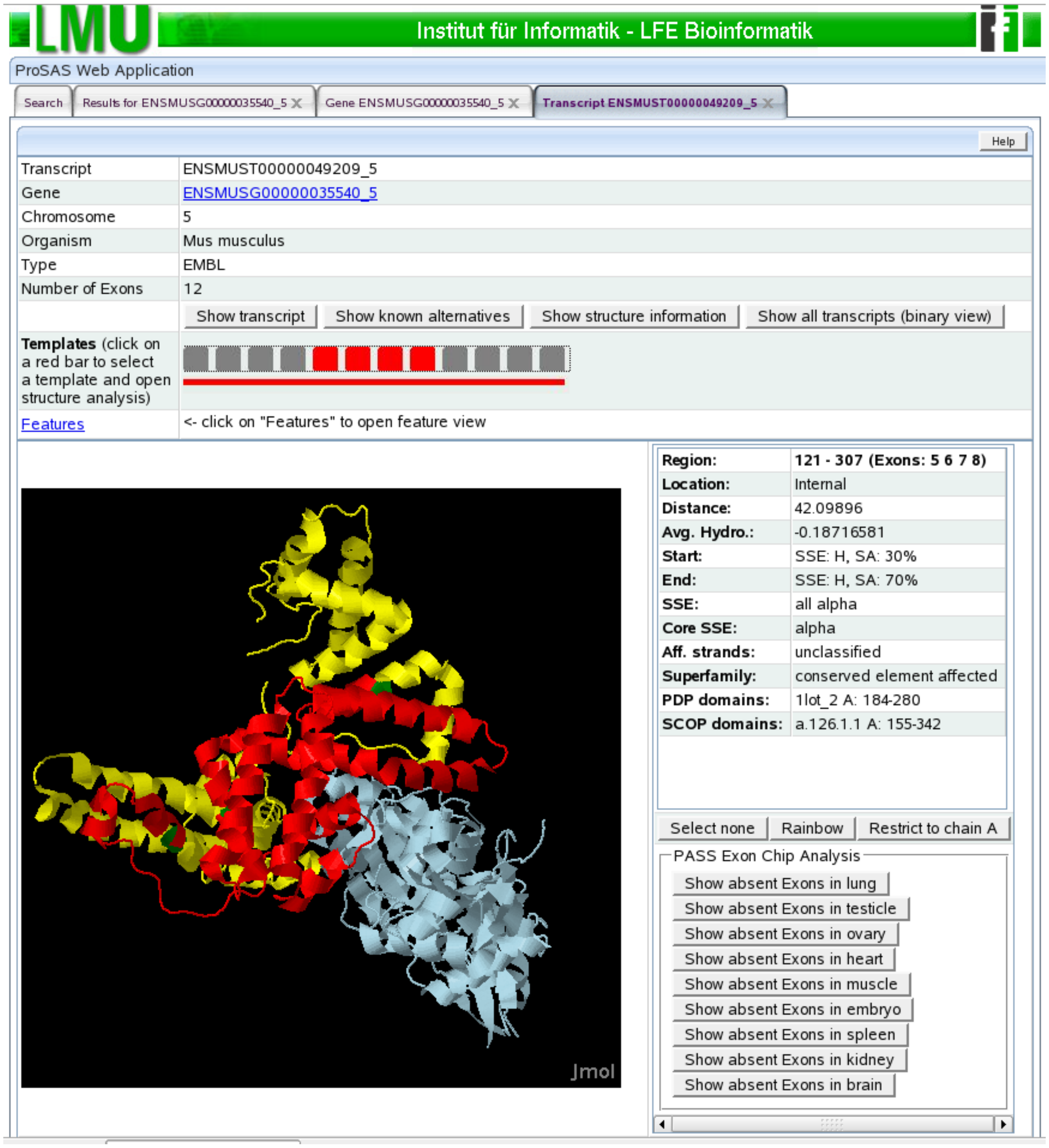
One interesting example which is identified 360 times (with an E-value as small as 0.00021 and a PeptideProphet score up to 0.93) is a peptide with the sequence *GTPVGPAAAGGHAPQLANALL-EKEV* in the dataset. It corresponds to a novel exon junction of gene ENSMUSG00000000738 on chromosome 8 between the first and the last exon. The gene is annotated as a homolog of the human spastic paraplegia 7 (SPG7) gene which encodes for a mitochondrial metalloprotease. The protein has 17 exons, and consists (according to PFAM patterns) of three domains (a ATPase domain, an extracellular FtsH domain as well as a domain of the peptidase family M41). The novel protein which is implied by this peptide consists only of two exons, the first and the last one, and has a length of 116 residues. The splicing event does not introduce a frame shift, the sequence of the potentially novel protein is shown below (the residues belonging to the newly identified peptide are shown in upper case).

```
Exon 1:  maaallllrglrpgpeprprrlwgllsgrgpglssgagarr
         pyaarGTPVGPAAAGGHAPQ
Exon 17: LANALLEKEVinyediealigppphgpkkmiapqkwidaek
         erqasgeeeapap
```

According to BLAST (against the NR) the resulting protein has no homology to any known protein in public databases which makes a validation of this novel gene product in terms of its potential function impossible. Therefore, it is also hard to tell if this protein represents a valid gene product at all. The fact that it is detected so frequently and the reverse sequence is also not known in the NR (such that a misassignment of the spectra is unlikely) might point to a valid protein product. This product either origins from a novel splicing event of this gene or points at the existence of a novel gene leading to a protein which contains this peptide sequence.
A second example, which hints to a novel, unknown splice variant and may have a well defined function in mouse plasma, is represented by a peptide with the sequence *KGFADQYTFELSR* (E-Value 0.00033, PeptideProphet score: 1.0, 5 matches above 0.9). The peptide matches a novel, in-frame exon junction of gene ENSMUSG00000035540 on chromosome 5. This gene is homologous to the human Vitamin-D binding protein which is present in plasma and on the surfaces of several cell types. The protein has various functions. In plasma, it carries the Vitamin D sterols and prevents polymerization of actin by binding its monomers. The splicing event which is predicted by the peptide corresponds to the removal of 4 out of 12 exons in the middle of the protein as shown in Figure 12.4. On the protein structure level, the removal event affects a domain-like part which provides the interface for binding monomeric actin (shown in blue in Figure 12.4). Therefore, the spliced isoform is likely not to be able to bind actin and may carry out a different, yet unknown, function in plasma.

While our pipeline and database provides a new approach to the identification of novel isoforms and of novel genes, several problems remain and need to be addressed in future research.

How can we further validate the existence of novel gene products, given their identification in mass spectrometry data, how can we get a clue on their function? Moreover, why have those proteins never been observed before such that sometimes even no homolog is available in the NR


LMU
Institut für Informatik - LFE Bioinformatik
ProSAS Web Application
Search
Results for ENSMUSG00000035540_5
Gene ENSMUSG00000035540_5
Transcript ENSMUST00000049209_5
Help
Show transcript
Show known alternatives
Show structure information
Show all transcripts (binary view)
Jmol
Select none
Rainbow
Restrict to chain A
PASS Exon Chip Analysis
Show absent Exons in lung
Show absent Exons in testicle
Show absent Exons in ovary
Show absent Exons in heart
Show absent Exons in muscle
Show absent Exons in embryo
Show absent Exons in spleen
Show absent Exons in kidney
Show absent Exons in brain

Transcript	ENSMUST00000049209_5
Gene	ENSMUSG00000035540_5
Chromosome	5
Organism	Mus musculus
Type	EMBL
Number of Exons	12
Templates (click on a red bar to select a template and open structure analysis)	
Features	<- click on "Features" to open feature view

Region:	**121 - 307 (Exons: 5 6 7 8)**
Location:	Internal
Distance:	42.09896
Avg. Hydro.:	-0.18716581
Start:	SSE: H, SA: 30%
End:	SSE: H, SA: 70%
SSE:	all alpha
Core SSE:	alpha
Aff. strands:	unclassified
Superfamily:	conserved element affected
PDP domains:	1lot_2 A: 184-280
SCOP domains:	a.126.1.1 A: 155-342

Figure 12.4: Shows a novel splice junction identified for a human Vitamin-D binding protein homolog in mouse visualized in ProSAS. The novel junction corresponds to the removal event of 4 out of 12 exons (shown in red on the structure level). Actin, bound to the protein is shown in blue, the part of the protein left over after the splice event is shown in yellow.

(see the first example discussed above)? Furthermore, what are good quality score thresholds which avoid the identification of false positives as far as possible?

Overall, a large number of spectra can not be interpreted using todays standard pipelines and novel search methods are required to identify post-translational modifications, phosphorylation and novel splice variants in this data.

Part IV

Conclusions and Outlook

Chapter 13

Conclusion and Outlook

During our crossing, Einstein explained his theory to me every day,
and by the time we arrived I was fully convinced he understood it.
(Chaim Weizmann)

The increase in biological data from various sources has led to the development of numerous methods in the fields of bioinformatics and systems biology to manage, analyze and interpret biological data as well as to predict biological features of small molecules, DNA, RNA, proteins, pathways and even of complete biological systems. This thesis has focused on the analysis and combination of two traditionally separated fields, namely the analysis of protein structures and protein structure evolution on the one hand and the detection of alternative splicing events in high-throughput experiments and their structure-aided interpretation on the other hand.

In **Part I** of the thesis we have described several methods developed and analyses carried out in the field of structural bioinformatics. Similar to other fields of experimental biology, data in structural biology, i.e. crystal or NMR structures of proteins, are generated faster and at lower cost than some years ago. The growing number of proteins available in the PDB has therefore led to the need for their automatic and accurate interpretation in the context of already characterized protein structures. To this end, especially methods which allow to rapidly search for similar protein structures in databases of already classified ones (e.g. in the SCOP database) and compute accurate structure-based alignments of protein families have become important tools in structural biology. They allow to group protein structures in an hierarchic manner which describes their evolutionary, structural and often their functional relationship. Such groups can therefore be used to study the correspondence of protein sequence and protein structure evolution, characterize evolutionary relationships between protein folds and to assign functional features to still uncharacterized protein structures (which become more and more available via structural genomics projects). We have developed several methods and analyses for a fast characterization of structural and functional features of newly resolved protein structures.

SCOP and CATH have become gold standard datasets for protein structure classification in the last years. Since both attempts to structure the protein fold space are based around different ideas and methods the same protein domain may be classified differently in the two hierarchies. Therefore, structural relationships between pairs of proteins may be defined differently according to the criteria used by the respective databases. In **Chapter 3** we have discussed the results of our detailed study of the similarities and differences of SCOP and CATH. We find a surprisingly large number of protein domains for which the two databases disagree in their classification and describe effects of including such domains for which the classification is unclear in datasets used for training and benchmarking structure classification methods. Furthermore, we extract valuable knowledge from the orthogonal information stored on both datasets in order to identify interfold similarities between protein structures classified into different folds.

An important step in structure analysis is the computation of structural similarities using structural alignments. We have developed two novel methods for protein structure comparison namely PPM and Vorolign which both introduce novel concepts to measure protein structure similarity based on different criteria.

With PPM, described in **Chapter 4**, Gergely Csaba (in collaboration with the author) addressed common problems in many current structure comparison algorithms and proposed a method which computes structure similarity by estimating the evolutionary distance by their structural mutation cost. PPM proves to be a very accurate method to detect structural similarities also in very difficult benchmark scenarios and is furthermore able to compute highly consistent cores of protein structure families in the presence of phenotypic plasticity avoiding shift errors and other common problems in standard structural alignment tools.

While the current scoring function of the PPM method is based on structural criteria (similar to most other existing structure alignment methods), the similarity function of our Vorolign method, described in **Chapter 5**, is much more sequence-driven and allows for a very fast and accurate scan for structurally and sequentially similar proteins which is a useful feature to support automatic structure classification attempts. We showed that pairwise and multiple structure-aided alignments computed by Vorolign are accurate in terms of their biological meaningfulness (according to criteria like the conservation of functionally important residues in the alignment).

Also, Vorolign's ability to align structures accurately solely based on the conservation of the Voronoi neighborhood may have interesting applications in protein structure prediction as the method is very similar to threading approaches with the exception that both structures (instead of one structure) are known. Vorolign shows that, in principle, threading methods which score the environment of a residue in the template structure given the current alignment can lead to reasonable alignments and very high recognition rates. From this point of view, Vorolign corresponds to performing threading, knowing the native residues neighborhoods in the structures. Nevertheless, our attempts to integrate the Vorolign scoring function into threading methods and to optimize the score given only one known structure has not yet been successful but may be an interesting direction for future research.

The good performance of the method in detecting structural similarity based on a sequence-driven scoring function points to a certain evolutionary conservation of residue neighborhoods in protein structures even if the sequence identity is near the random level. Such similarities are captured by the method when aligning two structures as the similarity of important residues, i.e. those which turn out to be structurally equivalent in the alignment, is expected to be larger than the conservation of any pair of cells representing non-equivalent residues.

Moreover, we have described two applications and extensions of the Vorolign method and its scoring function in the sections on the AutoPSI database (**Section 5.3**) and the identification of functionally important residues via contact patterns defined using the Vorolign scoring function (**Section 5.4**).

The AutoPSI database provides access to all currently available protein structures and their either predicted or known classifications with respect to the SCOP database and allows to close the widening gap between the known protein structure space and its available classification in gold standard datasets. AutoPSI has applications in all structural bioinformatics methods relying on template databases covering the known protein structure space as complete as possible.

With our method to identify conserved, functionally important patterns in protein structure families we have shown that Vorolign, in principle, may provide a useful tool in this application of structural bioinformatics as well. But despite promising initial results, the methodology needs to be improved and extended in the future to identify highly confident structure patterns describing functionally important and highly conserved sites in a wide range of protein families.

In **Part II** we applied our methods and knowledge developed in the analysis of protein structures to a different field of current bioinformatics research, namely the analysis and interpretation of alternative splicing events. In more detail, we analyzed data on alternative splicing in the context of protein structures and protein structure evolution which led to interesting and novel insights in both fields.

In **Chapter 7** we first introduced the concept of ”evolutionary isoforms” in the analysis of splicing events in order to distinguish structurally tolerable from structurally non-trivial events. Second, we provided evidence that proteins are much more tolerant against major rearrangements due to alternative splicing by analyzing non-trivial splice isoforms which carry out specific and well defined functions in the cell. Finally, we could show that there may be a connection between events observed in alternative splicing data and events occurring in protein structure evolution. In this analysis, alternative splicing data could add additional evidence to existing hypotheses on protein structure evolution and this data could be used to detect novel connections between different folds in the protein fold space. Such links point to the ability of splice isoform structures to adopt two different folds in the protein structure space.

In **Chapter 8** we further examined the connection between functional diversity generated by alternative splicing and functional diversity arising in the course of evolution proposed in Chapter 7 and focused on alternative splicing events annotated for a specific class of proteins, namely proteins containing repetitive elements. Such proteins play important roles in various processes like the regulation of transcription, the mediation of protein-protein interactions or tissue development especially in higher organisms. Our data suggests, that the extension of protein repeats in the evolution of complex organisms in combination with an additional increase in repeat protein variability may have significantly contributed to the development of complex transcriptomes, proteomes and tissue organizations and provides additional evidence for the coupling of protein evolution and alternative splicing.

Today, most data on the nature of alternative isoforms like their tissue and time specific expression is collected in large-scale, high-throughput experiments. In the final **Part III** of this thesis we describe our tools and pipelines developed for the analysis of diverse data sources. All those experimental methods have the potential to provide additional evidence for the hypotheses proposed in Part II and we have discussed the results of analyzing alternative splicing with the help of those methods and with respect to those hypotheses in some detail.

In **Chapter 10** we described our ProSAS database which is a comprehensive resource to study and analyze alternative splicing in the context of protein structures and other functional features of protein sequences. ProSAS has been extensively used as a data source and visualization platform for several analyses carried out in this thesis. More specifically, we have used data on known splice variants annotated in Ensembl and Swissprot and stored in ProSAS for our analyses on alternative splicing and protein structure evolution (Chapter 7) as well as our study on alternative splicing and protein repeats (Chapter 8). Furthermore, we have used ProSAS for the validation and development of the PASS method (Chapter 11) and for the analysis of mass spectrometry data in the context of alternative splicing (Chapter 12). Both experimental data sources are integrated into the ProSAS system and the ProSAS web application.

In order to analyze data generated by the Affymetrix exon array platform, Robert Küffner (in collaboration with the author) has introduced the PASS method discussed in **Chapter 11**. Despite the intrinsic problems of the platform in terms of the detection of isoform mixtures and the correct identification of specific isoforms in the presence of such mixtures, PASS turns out to be a useful tool for the experimental characterization and identification of splicing events in a time and tissue-specific manner in a genome-wide, large-scale fashion.

Most of our knowledge on alternative splicing stems from experiments acting on the transcriptome level. In contrast, mass spectrometry experiments allow for the detection of isoforms existing on the proteome level and we have described our analysis of such data sources in **Chapter 12**. We characterized the structural and functional complexity of splice isoforms existing on the proteome level using large-scale proteomics datasets and the splice variants uniquely identified by those experiments. This data supports the large tolerance of protein structures against rearrangements proposed in Chapter 7 and shows interesting patterns of generating functional diversity in the proteomes of higher organisms through alternative splicing. Furthermore, we discussed our approach to identify novel splice variants in mass spectrometry data using a database of potential exon junctions to extend existing pipelines.

Directions for future research

The ideas, hypotheses and methods proposed in this thesis may be extended in various directions in the future.

For the structure alignment methods described in Part I, namely Vorolign and PPM, several interesting questions remain. Especially the combination of the PPM algorithm with the Vorolign scoring function seems to be an attractive direction for future research in order to combine the powerful optimization procedure and model of protein structure evolution proposed with PPM with the well-perfoming features of the Vorolign similarity measurement. Such a hybrid method, combining the strength of both ideas will hopefully lead to even more accurate alignments and an even more accurate structure similarity measure.

Also, the integration of the PPM method into the AutoPSI database will help to further improve on the consensus prediction of automatic SCOP assignments and may also lead to interesting insights on the strength and weaknesses of the single methods.

Using both methods in combination to define highly conserved cores of protein structures may be another application useful for protein structure prediction approaches which can be extended by a large set of background knowledge on protein families and important core elements in order to improve sequence to structure alignment and template selection methods.

We have also already discussed the application of Vorolign to identify conserved residue patterns in protein families with interesting applications in protein function prediction and the identification of other important residue contact networks in protein families like early folding units.

For our hypotheses on alternative splicing and protein structure evolution especially the experimental confirmation of non-trivial isoforms via protein structure determination and the confirmation of the fold changing events proposed in this thesis have the potential to contribute to our understanding of protein structure flexibility and protein structure evolution in the future. Therefore, we hope that our ideas are further confirmed by novel, experimental protein structures of splice isoforms as our growing notion of the involvement of splicing isoforms in disease and the importance of isoforms for many important processes in higher organisms will lead to a growing need of a functional and also structured analysis of isoforms.

Splicing data may also be helpful to elucidate the specific evolution and of relationships within and between large protein families like Rossmann folds and others similar to the examples of TIM-Barrels and β-propellers discussed above. Such a specific evaluation of specific families together with potentially fold changing events detected by our approach may contribute to our understanding of the topology of the protein fold space.

Our pipelines for analyzing alternative splicing events in different sources of experimental data should be extended to address novel data potentially of use for the analysis of splicing generated by novel experimental techniques. Those contain next-generation sequencing methods which surely will lead to the detection of many known and also unknown isoforms in a genome-wide manner but require for novel methods and algorithms to analyze, process and manage those millions of short reads.

Finally, splicing needs to be tightly regulated in all cells in order to avoid the expression of non-functional, or even worse, harmful isoforms. We are currently investigating potential regulatory mechanisms of alternative splicing in the context of short RNAs, especially miRNA, which may help to keep the splicing machinery and its products under a strict control.

Overall, the mechanisms of time and tissue specific regulation of alternative splicing, the detailed functional mechanisms of (non-trivial) isoforms and the connection of functional and structural diversity generated in the course of evolution and generated by alternative splicing in our opinion correspond to some of the most interesting and most exiting questions in current bioinformatics research which have important applications in disease therapy and our understanding of organism complexity.

Bibliography

[1] J. Adachi, C. Kumar, Y. Zhang, and M. Mann. In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol Cell Proteomics*, 6(7):1257–73, 2007.

[2] N. Alexandrov and I. Shindyalov. Pdp: protein domain parser. *Bioinformatics*, 19(3): 429–30, 2003.

[3] S F Altschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

[4] P. Amstutz, H. K. Binz, P. Parizek, M. T. Stumpp, A. Kohl, M. G. Grutter, P. Forrer, and A. Pluckthun. Intracellular kinase inhibitors selected from combinatorial libraries of designed ankyrin repeat proteins. *J Biol Chem*, 280(26):24715–22, 2005.

[5] M. A. Andrade, C. Perez-Iratxeta, and C. P. Ponting. Protein repeats: structures, functions, and evolution. *J Struct Biol*, 134(2-3):117–31, 2001.

[6] A. Andreeva, D. Howorth, J.M. Chandonia, S.E. Brenner, T.J. Hubbard, C. Chothia, and A.G. Murzin. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, 36:D419–425, Jan 2008.

[7] T.K. Attwood, P. Bradley, D.R. Flower, A. Gaulton, N. Maudling, A.L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, and C. Zygouri. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, 31:400–402, Jan 2003.

[8] E Azarya-Sprinzak, D Naor, H J Wolfson, and R Nussinov. Interchanges of spatially neighbouring residues in structurally conserved environments. *Protein Eng*, 10(10):1109–1122, Oct 1997.

[9] I. Bab, E. Smith, H. Gavish, M. Attar-Namdar, M. Chorev, Y.C. Chen, A. Muhlrad, M.J. Birnbaum, G. Stein, and B. Frenkel. Biosynthesis of osteogenic growth peptide via alternative translational initiation at AUG85 of histone H4 mRNA. *J. Biol. Chem.*, 274: 14474–14481, May 1999.

[10] C. Bradford Barber, David P. Dobkin, and Hannu T. Huhdanpaa. The Quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, December 1996. URL `http://www.thesa.com/software/qhull/qhull-96.ps`.

[11] D. Barthel, J.D. Hirst, J. Blazewicz, E.K. Burke, and N. Krasnogor. ProCKSI: a decision support system for Protein (structure) Comparison, Knowledge, Similarity and Information. *BMC Bioinformatics*, 8:416, 2007.

[12] S. Bencharit, C. B. Cui, A. Siddiqui, E. L. Howard-Williams, J. Sondek, K. Zuobi-Hasona, and I. Aukhil. Structural insights into fibronectin type iii domain-mediated signaling. *J Mol Biol*, 367(2):303–9, 2007.

[13] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[14] F. Birzele, G. Csaba, and R. Zimmer. Alternative splicing and protein structure evolution. *Nucleic Acids Res.*, 36:550–558, Feb 2008.

[15] F. Birzele, J.E. Gewehr, G. Csaba, and R. Zimmer. Vorolign–fast structural alignment using Voronoi contacts. *Bioinformatics*, 23:e205–211, Jan 2007.

[16] F. Birzele, J.E. Gewehr, and R. Zimmer. AutoPSI: a database for automatic structural classification of protein sequences and structures. *Nucleic Acids Res.*, 36:398–401, Jan 2008.

[17] F. Birzele, E. Hoffmann, G. Csaba, and R. Zimmer. Alternative splicing and protein repeats. *submitted.*

[18] F. Birzele and S. Kramer. A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics*, 22:2628–2634, Nov 2006.

[19] F. Birzele, R. Kueffner, F. Meier, F. Oefinger, C. Potthast, and R. Zimmer. ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.*, 36:D63–68, Jan 2008.

[20] F. Birzele and R. Zimmer. Alternative splicing and proteome complexity in mass spectrometry data. *submitted.*

[21] A. K. Bjorklund, D. Ekman, and A. Elofsson. Expansion of protein domain repeats. *PLoS Comput Biol*, 2(8):e114, 2006.

[22] J D Blake and F E Cohen. Pairwise sequence alignment below the twilight zone. *J Mol Biol*, 307(2):721–735, Mar 2001.

[23] B.J. Blencowe. Alternative splicing: new insights from global analyses. *Cell*, 126:37–47, Jul 2006.

[24] M.S. Boguski, T.M. Lowe, and C.M. Tolstoshev. dbEST–database for expressed sequence tags. *Nat. Genet.*, 4:332–333, Aug 1993.

[25] P. Bonizzoni, R. Rizzi, and G. Pesole. Computational methods for alternative splicing prediction. *Brief Funct Genomic Proteomic*, 5:46–51, Mar 2006.

[26] P. H. Bourne and H. Weissig. *Structural Bioinformatics*. Wiley, 2003. ISBN 0-471-20200-2.

[27] S E Brenner, P Koehl, and M Levitt. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–256, Jan 2000.

[28] S.E. Brenner, C. Chothia, and T.J. Hubbard. Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.*, 7:369–376, Jun 1997.

[29] E. Brunner, C.H. Ahrens, S. Mohanty, H. Baetschmann, S. Loevenich, F. Potthast, E.W. Deutsch, C. Panse, U. de Lichtenberg, O. Rinner, H. Lee, P.G. Pedrioli, J. Malmstrom, K. Koehler, S. Schrimpf, J. Krijgsveld, F. Kregenow, A.J. Heck, E. Hafen, R. Schlapbach, and R. Aebersold. A high-quality catalog of the Drosophila melanogaster proteome. *Nat. Biotechnol.*, 25:576–583, May 2007.

[30] I. Chaudhuri, J. Soding, and A. N. Lupas. Evolution of the beta-propeller fold. *Proteins*, 71(2):795–803, 2008.

[31] K. Chen and L. Kurgan. PFRES: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics*, 23:2843–2850, Nov 2007.

[32] J. Cox and M. Mann. Is proteomics the new genomics? *Cell*, 130:395–398, Aug 2007.

[33] R. Craig and R.C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20:1466–1467, Jun 2004.

[34] Robertson Craig and Ronald C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

[35] G. Csaba, F. Birzele, and R. Zimmer. Systematic Comparison of SCOP and CATH: A new Gold Standard for Protein Structure Analysis. *submitted*.

[36] G. Csaba, F. Birzele, and R. Zimmer. Protein structure alignment considering phenotypic plasticity. *Bioinformatics*, 24:98–104, Aug 2008.

[37] M. Cuperlovic-Culf, N. Belacel, A.S. Culf, and R.J. Ouellette. Data analysis of alternative splicing microarrays. *Drug Discov. Today*, 11:983–990, Nov 2006.

[38] R. Day, D.A. Beck, R.S. Armen, and V. Daggett. A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci.*, 12:2150–2160, Oct 2003.

[39] M O Dayhoff, R Schwartz, and B C Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5:345–352, 1978.

[40] P. de la Grange, M. Dutertre, N. Martin, and D. Auboeuf. FAST DB: a website resource for the study of the expression regulation of human gene products. *Nucleic Acids Res.*, 33: 4276–4284, 2005.

[41] F. Desiere, E.W. Deutsch, N.L. King, A.I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S.N. Loevenich, and R. Aebersold. The PeptideAtlas project. *Nucleic Acids Res.*, 34: D655–658, Jan 2006.

[42] F. Desiere, E.W. Deutsch, A.I. Nesvizhskii, P. Mallick, N.L. King, J.K. Eng, A. Aderem, R. Boyle, E. Brunner, S. Donohoe, N. Fausto, E. Hafen, L. Hood, M.G. Katze, K.A. Kennedy, F. Kregenow, H. Lee, B. Lin, D. Martin, J.A. Ranish, D.J. Rawlings, L.E. Samelson, Y. Shiio, J.D. Watts, B. Wollscheid, M.E. Wright, W. Yan, L. Yang, E.C. Yi, H. Zhang, and R. Aebersold. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.*, 6:R9, 2005.

[43] Aleksandra B Djurisic, Jovan M Elazar, and A D Rakic. Genetic algorithms for continuous optimization problems - a concept of parameter-space size adjustment. *Journal of Physics A: Mathematical and General*, 30(22):7849–7861, 1997. URL `http://stacks.iop.org/0305-4470/30/7849`.

[44] Z Dosztanyi and A E Torda. Amino acid similarity matrices based on force fields. *Bioinformatics*, 17(8):686–699, Aug 2001.

[45] N.J. Edwards. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol. Syst. Biol.*, 3:102, 2007.

[46] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*, 2:953–971, 2007.

[47] C. ffrench Constant. Alternative splicing of fibronectin–many different proteins but few different functions. *Exp Cell Res*, 221(2):261–71, 1995.

[48] F. V. Filipp and M. Sattler. Conformational plasticity of the lipid transfer protein scp2. *Biochemistry*, 46(27):7980–7991, 2007.

[49] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. The pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–8, 2008.

[50] S. Foissac and M. Sammeth. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.*, 35:W297–299, Jul 2007.

[51] A.M. Frank, M.M. Savitski, M.L. Nielsen, R.A. Zubarev, and P.A. Pevzner. De novo peptide sequencing and identification with precision mass spectrometry. *J. Proteome Res.*, 6:114–123, Jan 2007.

[52] I. Friedberg and A. Godzik. Fragnostic: walking through protein structure space. *Nucleic Acids Res.*, 33:W249–251, Jul 2005.

[53] H H Gan, A Tropsha, and T Schlick. Lattice protein folding with two and four-body statistical potentials. *Proteins*, 43(2):161–174, May 2001.

[54] J. Garcia, S. H. Gerber, S. Sugita, T. C. Sudhof, and J. Rizo. A conformational switch in the piccolo c2a domain regulated by alternative splicing. *Nat Struct Mol Biol*, 11(1): 45–53, 2004.

[55] M.A. Garcia-Blanco, A.P. Baraniak, and E.L. Lasda. Alternative splicing in disease and therapy. *Nat. Biotechnol.*, 22:535–546, May 2004.

[56] J.E. Gewehr, V. Hintermair, and R. Zimmer. AutoSCOP: automated prediction of SCOP classifications using unique pattern-class mappings. *Bioinformatics*, 23:1203–1210, May 2007.

[57] M. Grabowski, A. Joachimiak, Z. Otwinowski, and W. Minor. Structural genomics: keeping up with expanding knowledge of the protein universe. *Curr. Opin. Struct. Biol.*, 17: 347–353, Jun 2007.

[58] Lesley H Greene, Tony E Lewis, Sarah Addou, Alison Cuff, Tim Dallman, Mark Dibley, Oliver Redfern, Frances Pearl, Rekha Nambudiry, Adam Reid, Ian Sillitoe, Corin Yeats, Janet M Thornton, and Christine A Orengo. The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic acids research*, 35(Database issue):D291–7, 2007.

[59] N. V. Grishin. Fold change in evolution of protein structures. *J Struct Biol*, 134(2-3): 167–85, 2001.

[60] M.S. Grotewiel, C.D. Beck, K.H. Wu, X.R. Zhu, and R.L. Davis. Integrin-mediated short-term memory in Drosophila. *Nature*, 391:455–460, Jan 1998.

[61] J. Gu, R.P. Dong, C. Zhang, D.F. McLaughlin, M.X. Wu, and S.F. Schlossman. Functional interaction of DFF35 and DFF45 with caspase-activated DNA fragmentation nuclease DFF40. *J. Biol. Chem.*, 274:20759–20762, Jul 1999.

[62] C. Guda, S. Lu, E.D. Scheeff, P.E. Bourne, and I.N. Shindyalov. CE-MC: a multiple protein structure alignment server. *Nucleic Acids Res.*, 32:W100–103, Jul 2004.

[63] A. Gutteridge and J.M. Thornton. Understanding nature's catalytic toolkit. *Trends Biochem. Sci.*, 30:622–629, Nov 2005.

[64] C. Hadley and D.T. Jones. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, 7:1099–1112, Sep 1999.

[65] A. Harrison, F. Pearl, R. Mott, J. Thornton, and C. Orengo. Quantifying the similarities within fold space. *J. Mol. Biol.*, 323:909–926, Nov 2002.

[66] A. Harrison, F. Pearl, I. Sillitoe, T. Slidel, R. Mott, J. Thornton, and C. Orengo. Recognizing the fold of a protein structure. *Bioinformatics*, 19:1748–1759, Sep 2003.

[67] M. Hiller, K. Huse, M. Platzer, and R. Backofen. Non-EST based prediction of exon skipping and intron retention events using Pfam information. *Nucleic Acids Res.*, 33: 5611–5621, 2005.

[68] M. Hiller, K. Huse, K. Szafranski, N. Jahn, J. Hampe, S. Schreiber, R. Backofen, and M. Platzer. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat. Genet.*, 36:1255–1257, Dec 2004.

[69] T.A. Holland, S. Veretnik, I.N. Shindyalov, and P.E. Bourne. Partitioning protein structures into domains: why is it so difficult? *J. Mol. Biol.*, 361:562–590, Aug 2006.

[70] L Holm and C Sander. Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–138, Sep 1993.

[71] A.E. House and K.W. Lynch. Regulation of alternative splicing: more than just the ABCs. *J. Biol. Chem.*, 283:1217–1221, Jan 2008.

[72] T. Hsu, J. A. Gogos, S. A. Kirsh, and F. C. Kafatos. Multiple zinc finger forms resulting from developmentally regulated alternative splicing of a transcription factor gene. *Science*, 257(5078):1946–50, 1992.

[73] T. J. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res*, 35(Database issue), 2007.

[74] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, B.A. Cuche, E. de Castro, C. Lachaize, P.S. Langendijk-Genevaux, and C.J. Sigrist. The 20 years of PROSITE. *Nucleic Acids Res.*, 36:D245–249, Jan 2008.

[75] Valentin A Ilyin, Alexej Abyzov, and Chesley M Leslin. Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Sci*, 13(7):1865–1874, Jul 2004.

[76] T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, K.O. Koyanagi, R.A. Barrero, T. Tamura, Y. Yamaguchi-Kabata, M. Tanino, K. Yura, S. Miyazaki, K. Ikeo, K. Homma, A. Kasprzyk, T. Nishikawa, M. Hirakawa, J. Thierry-Mieg, D. Thierry-Mieg, J. Ashurst, L. Jia, M. Nakao, M.A. Thomas, N. Mulder, Y. Karavidopoulou, L. Jin, S. Kim, T. Yasuda, B. Lenhard, E. Eveno, Y. Suzuki, C. Yamasaki, J. Takeda, C. Gough, P. Hilton, Y. Fujii, H. Sakai, S. Tanaka, C. Amid, M. Bellgard, M.d.e. F. Bonaldo, H. Bono, S.K. Bromberg, A.J. Brookes, E. Bruford, P. Carninci, C. Chelala, C. Couillault, S.J. de Souza,

M.A. Debily, M.D. Devignes, I. Dubchak, T. Endo, A. Estreicher, E. Eyras, K. Fukami-Kobayashi, G.R. Gopinath, E. Graudens, Y. Hahn, M. Han, Z.G. Han, K. Hanada, H. Hanaoka, E. Harada, K. Hashimoto, U. Hinz, M. Hirai, T. Hishiki, I. Hopkinson, S. Imbeaud, H. Inoko, A. Kanapin, Y. Kaneko, T. Kasukawa, J. Kelso, P. Kersey, R. Kikuno, K. Kimura, B. Korn, V. Kuryshev, I. Makalowska, T. Makino, S. Mano, R. Mariage-Samson, J. Mashima, H. Matsuda, H.W. Mewes, S. Minoshima, K. Nagai, H. Nagasaki, N. Nagata, R. Nigam, O. Ogasawara, O. Ohara, M. Ohtsubo, N. Okada, T. Okido, S. Oota, M. Ota, T. Ota, T. Otsuki, D. Piatier-Tonneau, A. Poustka, S.X. Ren, N. Saitou, K. Sakai, S. Sakamoto, R. Sakate, I. Schupp, F. Servant, S. Sherry, R. Shiba, N. Shimizu, M. Shimoyama, A.J. Simpson, B. Soares, C. Steward, M. Suwa, M. Suzuki, A. Takahashi, G. Tamiya, H. Tanaka, T. Taylor, J.D. Terwilliger, P. Unneberg, V. Veeramachaneni, S. Watanabe, L. Wilming, N. Yasuda, H.S. Yoo, M. Stodolsky, W. Makalowski, M. Go, K. Nakai, T. Takagi, M. Kanehisa, Y. Sakaki, J. Quackenbush, Y. Okazaki, Y. Hayashizaki, W. Hide, R. Chakraborty, K. Nishikawa, H. Sugawara, Y. Tateno, Z. Chen, M. Oishi, P. Tonellato, R. Apweiler, K. Okubo, L. Wagner, S. Wiemann, R.L. Strausberg, T. Isogai, C. Auffray, N. Nomura, T. Gojobori, and S. Sugano. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, 2:e162, Jun 2004.

[77] T. Inoki, S. Yamagami, Y. Inoki, T. Tsuru, T. Hamamoto, Y. Kagawa, T. Mori, and H. Endo. Human ddb2 splicing variants are dominant negative inhibitors of uv-damaged dna repair. *Biochem Biophys Res Commun*, 314(4):1036–43, 2004.

[78] V. Janssens, C. van Hoof, E. Martens, I. de Baere, W. Merlevede, and J. Goris. Identification and characterization of alternative splice products encoded by the human phosphotyrosyl phosphatase activator gene. *Eur. J. Biochem.*, 267:4406–4413, Jul 2000.

[79] J. M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, and D. D. Shoemaker. Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *Science*, 302(5653):2141–4, 2003.

[80] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, Dec 1983.

[81] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. Ensmart: a generic system for fast and flexible access to biological data. *Genome Res*, 14(1):160–9, 2004.

[82] A. Keller, A.I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, 74:5383–5392, Oct 2002.

[83] N. Kim, A.V. Alekseyenko, M. Roy, and C. Lee. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res.*, 35:D93–98, Jan 2007.

[84] E.F. Kirkness and C.M. Fraser. A strong promoter element is located between alternative exons of a gene encoding the human gamma-aminobutyric acid-type A receptor beta 3 subunit (GABRB3). *J. Biol. Chem.*, 268:4420–4428, Feb 1993.

[85] A.S. Konagurthu, J.C. Whisstock, P.J. Stuckey, and A.M. Lesk. MUSTANG: a multiple structural alignment algorithm. *Proteins*, 64:559–574, Aug 2006.

[86] D.M. Kristensen, R.M. Ward, A.M. Lisewski, S. Erdin, B.Y. Chen, V.Y. Fofanov, M. Kimmel, L.E. Kavraki, and O. Lichtarge. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics*, 9:17, 2008.

[87] E.V. Kriventseva, I. Koch, R. Apweiler, M. Vingron, P. Bork, M.S. Gelfand, and S. Sunyaev. Increase of functional diversity by alternative splicing. *Trends Genet.*, 19:124–128, Mar 2003.

[88] A. Krogh, B. Larsson, G. von Heijne, and E.L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, 305:567–580, Jan 2001.

[89] R. Kueffner, F. Birzele, and R. Zimmer. PASS: Reliable detection of alternative splicing with microarrays. *submitted*.

[90] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Res. Logist. Quart.*, 2:83–97, 1955.

[91] R.D. Ladner, D.E. McNulty, S.A. Carr, G.D. Roberts, and S.J. Caradonna. Characterization of distinct nuclear and mitochondrial forms of human deoxyuridine triphosphate nucleotidohydrolase. *J. Biol. Chem.*, 271:7745–7751, Mar 1996.

[92] D. Lang, R. Thoma, M. Henn-Sax, R. Sterner, and M. Wilmanns. Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science*, 289 (5484):1546–50, 2000.

[93] R.A. Laskowski, J.D. Watson, and J.M. Thornton. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, 33:89–93, Jul 2005.

[94] Juliette T J Lecomte, David A Vuletich, and Arthur M Lesk. Structural divergence and distant relationships in proteins: evolution of the globins. *Curr Opin Struct Biol*, 15(3): 290–301, Jun 2005.

[95] Y. Lee, Y. Lee, B. Kim, Y. Shin, S. Nam, P. Kim, N. Kim, W.H. Chung, J. Kim, and S. Lee. ECgene: an alternative splicing database update. *Nucleic Acids Res.*, 35:99–103, Jan 2007.

[96] B.P. Lewis, R.E. Green, and S.E. Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U.S.A.*, 100:189–192, Jan 2003.

[97] J. Li, A. Mahajan, and M. D. Tsai. Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry*, 45(51):15168–78, 2006.

[98] D. Lipscombe. Neuronal proteins custom designed by alternative splicing. *Curr Opin Neurobiol*, 15(3):358–63, 2005.

[99] B.M. Machiels, A.H. Zorenc, J.M. Endert, H.J. Kuijpers, G.J. van Eys, F.C. Ramaekers, and J.L. Broers. An alternative splicing product of the lamin A/C gene lacks exon 10. *J. Biol. Chem.*, 271:9249–9253, Apr 1996.

[100] A. Magen and G. Ast. The importance of being divisible by three in alternative splicing. *Nucleic Acids Res.*, 33:5574–5582, 2005.

[101] E. M. Marcotte, M. Pellegrini, T. O. Yeates, and D. Eisenberg. A census of protein repeats. *J Mol Biol*, 293(1):151–60, 1999.

[102] M.A. Mart-Renom, A.C. Stuart, A. Fiser, R. Snchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29: 291–325, 2000.

[103] L J McGuffin, K Bryson, and D T Jones. What are the baselines for protein fold recognition? *Bioinformatics*, 17(1):63–72, Jan 2001.

[104] I. Melvin, E. Ie, R. Kuang, J. Weston, W.N. Stafford, and C. Leslie. SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. *BMC Bioinformatics*, 8 Suppl 4:S2, 2007.

[105] I. Michalopoulos, G. M. Torrance, D. R. Gilbert, and D. R. Westhead. Tops: an enhanced database of protein structural topology. *Nucleic Acids Res*, 32(Database issue):D251–4, 2004.

[106] E. W. Miles and D. R. Davies. Protein evolution. on the ancestry of barrels. *Science*, 289 (5484):1490, 2000.

[107] L. K. Mosavi, T. J. Cammett, D. C. Desrosiers, and Z. Y. Peng. The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci*, 13(6):1435–48, 2004.

[108] J. Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.*, 15:285–289, Jun 2005.

[109] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A. N. Nikolskaya, S. Orchard, M. Pagni, C. P. Ponting, E. Quevillon, J. Selengut, C. J. Sigrist, V. Silventoinen, D. J. Studholme, R. Vaughan, and C. H. Wu. Interpro, progress and status in 2005. *Nucleic Acids Res*, 33(Database issue):D201–5, 2005.

[110] J.C. Nebel, P. Herzyk, and D.R. Gilbert. Automatic generation of 3D motifs for classification of protein binding sites. *BMC Bioinformatics*, 8:321, 2007.

[111] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, Mar 1970.

[112] A.D. Neverov, I.I. Artamonova, R.N. Nurtdinov, D. Frishman, M.S. Gelfand, and A.A. Mironov. Alternative splicing and protein function. *BMC Bioinformatics*, 6:266, 2005.

[113] C Notredame, D G Higgins, and J Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217, Sep 2000.

[114] Marian Novotny, Dennis Madsen, and Gerard J Kleywegt. Evaluation of protein fold comparison servers. *Proteins*, 54(2):260–270, Feb 2004. Evaluation Studies.

[115] Maria Elena Ochagavia and Shoshana Wodak. Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins. *Proteins*, 55(2): 436–454, May 2004.

[116] U. Ohler, N. Shomron, and C.B. Burge. Recognition of unknown conserved alternatively spliced exons. *PLoS Comput. Biol.*, 1:113–122, Jul 2005.

[117] G.S. Omenn, D.J. States, M. Adamski, T.W. Blackwell, R. Menon, H. Hermjakob, R. Apweiler, B.B. Haab, R.J. Simpson, J.S. Eddes, E.A. Kapp, R.L. Moritz, D.W. Chan, A.J. Rai, A. Admon, R. Aebersold, J. Eng, W.S. Hancock, S.A. Hefta, H. Meyer, Y.K. Paik, J.S. Yoo, P. Ping, J. Pounds, J. Adkins, X. Qian, R. Wang, V. Wasinger, C.Y. Wu, X. Zhao, R. Zeng, A. Archakov, A. Tsugita, I. Beer, A. Pandey, M. Pisano, P. Andrews, H. Tammen, D.W. Speicher, and S.M. Hanash. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics*, 5:3226–3245, Aug 2005.

[118] Joseph O'Rourke. *Computational Geometry in C*. Cambridge University Press, 1993. ISBN 0-521-44034-3.

[119] M.J. Palladino, J.E. Bower, R. Kreber, and B. Ganetzky. Neural dysfunction and neurodegeneration in Drosophila Na+/K+ ATPase alpha subunit mutants. *J. Neurosci.*, 23: 1276–1286, Feb 2003.

[120] Q. Pan, A.L. Saltzman, Y.K. Kim, C. Misquitta, O. Shai, L.E. Maquat, B.J. Frey, and B.J. Blencowe. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.*, 20:153–158, Jan 2006.

[121] X Pennec and N Ayache. *An* $O(n^2)$ *algorithm for 3D substructure matching of proteins.* Plenum Publishing, 1994. 25–40 pp.

[122] B.J. Polacco and P.C. Babbitt. Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*, 22:723–730, Mar 2006.

[123] C.T. Porter, G.J. Bartlett, and J.M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, 32:D129–133, Jan 2004.

[124] A. Poupon. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr. Opin. Struct. Biol.*, 14:233–241, Apr 2004.

[125] A Prlic, F S Domingues, and M J Sippl. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng*, 13(8):545–550, Aug 2000.

[126] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. Interproscan: protein domains identifier. *Nucleic Acids Res*, 33(Web Server issue):W116–20, 2005.

[127] T. Rattei, P. Tischler, R. Arnold, F. Hamberger, J. Krebs, J. Krumsiek, B. Wachinger, V. Stmpflen, and W. Mewes. SIMAP structuring the network of protein similarities. *Nucleic Acids Res*, Nov 2007.

[128] O.C. Redfern, A. Harrison, T. Dallman, F.M. Pearl, and C.A. Orengo. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput. Biol.*, 3:e232, Nov 2007.

[129] G.A. Reeves, T.J. Dallman, O.C. Redfern, A. Akpor, and C.A. Orengo. Structural diversity of domain superfamilies in the CATH database. *J. Mol. Biol.*, 360:725–741, Jul 2006.

[130] A. Resch, Y. Xing, B. Modrek, M. Gorlick, R. Riley, and C. Lee. Assessing the impact of alternative splicing on domain interactions in the human proteome. *J. Proteome Res.*, 3: 76–83, 2004.

[131] F M Richards. The interpretation of protein structures: total volume, group volume distributions and packing density. *J Mol Biol*, 82(1):1–14, Jan 1974.

[132] Jeffrey Roach, Shantanu Sharma, Maryna Kapustina, and Charles W Jr Carter. Structure alignment via Delaunay tetrahedralization. *Proteins*, 60(1):66–81, Jul 2005.

[133] K. D. Robertson and P. A. Jones. Tissue-specific alternative splicing in the human ink4a/arf cell cycle regulatory locus. *Oncogene*, 18(26):3810–20, 1999.

[134] P. R. Romero, S. Zaidi, Y. Y. Fang, V. N. Uversky, P. Radivojac, C. J. Oldfield, M. S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic, and A. K. Dunker. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A*, 103(22):8390–5, 2006.

[135] S. M. Ruben, R. Narayanan, J. F. Klement, C. H. Chen, and C. A. Rosen. Functional characterization of the nf-kappa b p65 transcriptional activator and an alternatively spliced derivative. *Mol Cell Biol*, 12(2):444–54, 1992.

[136] A. Saito, K. Ozaki, T. Fujiwara, Y. Nakamura, and A. Tanigami. Isolation and mapping of a human lung-specific gene, tsa1902, encoding a novel chitinase family member. *Gene*, 239(2):325–31, 1999.

[137] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, Jul 1987.

[138] C. Schumacher, H. Wang, C. Honer, W. Ding, J. Koehn, Q. Lawrence, C.M. Coulis, L.L. Wang, D. Ballinger, B.R. Bowen, and S. Wagner. The SCAN domain mediates selective oligomerization. *J. Biol. Chem.*, 275:17173–17179, Jun 2000.

[139] C. Schwerk and K. Schulze-Osthoff. Regulation of apoptosis by alternative pre-mrna splicing. *Mol Cell*, 19(1):1–13, 2005.

[140] G.R. Screaton, X.N. Xu, A.L. Olsen, A.E. Cowper, R. Tan, A.J. McMichael, and J.I. Bell. LARD: a new lymphoid-specific death domain containing receptor regulated by alternative pre-mRNA splicing. *Proc. Natl. Acad. Sci. U.S.A.*, 94:4615–4619, Apr 1997.

[141] M. Shatsky, R. Nussinov, and H. J. Wolfson. Optimization of multiple-sequence alignment based on multiple-structure alignment. *Proteins*, 62(1):209–17, 2006.

[142] Maxim Shatsky, Ruth Nussinov, and Haim J Wolfson. A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1):143–156, Jul 2004.

[143] I N Shindyalov and P E Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11(9):739–747, Sep 1998.

[144] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rtsch. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 8 Suppl 10:S7, 2007.

[145] R. Sorek and G. Ast. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, 13:1631–1637, Jul 2003.

[146] S. M. Srinivasula, M. Ahmad, Y. Guo, Y. Zhan, Y. Lazebnik, T. Fernandes-Alnemri, and E. S. Alnemri. Identification of an endogenous dominant-negative short isoform of caspase-9 that can regulate apoptosis. *Cancer Res*, 59(5):999–1002, 1999.

[147] S. Stamm, J.J. Riethoven, V. Le Texier, C. Gopalakrishnan, V. Kumanduri, Y. Tang, N.L. Barbosa-Morais, and T.A. Thanaraj. ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, 34:46–55, Jan 2006.

[148] J. Stetefeld and M. A. Ruegg. Structural and functional diversity generated by alternative mrna splicing. *Trends Biochem Sci*, 30(9):515–21, 2005.

[149] C.M. Stover, S. Thiel, N.J. Lynch, and W.J. Schwaeble. The rat and mouse homologues of MASP-2 and MAp19, components of the lectin activation pathway of complement. *J. Immunol.*, 163:6848–6859, Dec 1999.

[150] T. Sudo, Y. Yagasaki, H. Hama, N. Watanabe, and H. Osada. Exip, a new alternative splicing variant of p38 alpha, can induce an earlier onset of apoptosis in hela cells. *Biochem Biophys Res Commun*, 291(4):838–43, 2002.

[151] M. Sultan, M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321:956–960, Aug 2008.

[152] Y.P. Sun, K.J. Deng, F. Wang, J. Zhang, X. Huang, S. Qiao, and S. Zhao. Two novel isoforms of Adam23 expressed in the developmental process of mouse and human brains. *Gene*, 325:171–178, Jan 2004.

[153] J. Sding, A. Biegert, and A.N. Lupas. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, 33:W244–248, Jul 2005.

[154] P. Tailor, J. Gilman, S. Williams, and T. Mustelin. A novel isoform of the low molecular weight phosphotyrosine phosphatase, lmptp-c, arising from alternative mrna splicing. *Eur J Biochem*, 262(2):277–82, 1999.

[155] J. Takeda, Y. Suzuki, M. Nakao, T. Kuroda, S. Sugano, T. Gojobori, and T. Imanishi. H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res.*, 35:D104–109, Jan 2007.

[156] J.K. Taylor, Q.Q. Zhang, J.R. Wyatt, and N.M. Dean. Induction of endogenous Bcl-xS through the control of Bcl-x pre-mRNA splicing by antisense oligonucleotides. *Nat. Biotechnol.*, 17:1097–1100, Nov 1999.

[157] W R Taylor. Protein structure comparison using iterated double dynamic programming. *Protein Sci*, 8(3):654–665, Mar 1999.

[158] W R Taylor and C A Orengo. Protein structure alignment. *J Mol Biol*, 208(1):1–22, Jul 1989.

[159] J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, Nov 1994.

[160] A.E. Todd, C.A. Orengo, and J.M. Thornton. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, 307:1113–1143, Apr 2001.

[161] M. L. Tress, P. L. Martelli, A. Frankish, G. A. Reeves, J. J. Wesselink, C. Yeats, P. L. Olason, M. Albrecht, H. Hegyi, A. Giorgetti, D. Raimondo, J. Lagarde, R. A. Laskowski, G. Lopez, M. I. Sadowski, J. D. Watson, P. Fariselli, I. Rossi, A. Nagy, W. Kai, Z. Storling, M. Orsini, Y. Assenov, H. Blankenburg, C. Huthmacher, F. Ramirez, A. Schlicker, F. Denoeud, P. Jones, S. Kerrien, S. Orchard, S. E. Antonarakis, A. Reymond, E. Birney, S. Brunak, R. Casadio, R. Guigo, J. Harrow, H. Hermjakob, D. T. Jones, T. Lengauer, C. A. Orengo, L. Patthy, J. M. Thornton, A. Tramontano, and A. Valencia. The implications of alternative splicing in the encode protein complement. *Proc Natl Acad Sci U S A*, 104(13):5495–500, 2007.

[162] R. Urrutia. Krab-containing zinc-finger repressor proteins. *Genome Biol*, 4(10):231, 2003.

[163] J. P. Venables. Aberrant and alternative splicing in cancer. *Cancer Res*, 64(21):7647–54, 2004.

[164] S. Veretnik, P.E. Bourne, N.N. Alexandrov, and I.N. Shindyalov. Toward consistent assignment of structural domains in proteins. *J. Mol. Biol.*, 339:647–678, Jun 2004.

[165] R.B. Voelker and J.A. Berglund. A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res.*, 17:1023–1033, Jul 2007.

[166] Niklas von Ohsen, Ingolf Sommer, Ralf Zimmer, and Thomas Lengauer. Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics*, 20(14):2228–2235, Sep 2004. Evaluation Studies.

[167] G. Voronoi. Nouvelles applications des parametres continus a la theorie des formes quadratiques. *J. f. d. Reine und Angewandte Mathematik*, 134:198–287, 1908.

[168] A Wallqvist, Y Fukunishi, L R Murphy, A Fadel, and R M Levy. Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics*, 16(11):988–1002, Nov 2000.

[169] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–476, Nov 2008.

[170] G.S. Wang and T.A. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, 8:749–761, Oct 2007.

[171] P. Wang, B. Yan, J. T. Guo, C. Hicks, and Y. Xu. Structural genomics analysis of alternative splicing and application to isoform structure modeling. *Proc Natl Acad Sci U S A*, 102(52):18920–5, 2005.

[172] Z. Wang and C.B. Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14:802–813, May 2008.

[173] S. Whitman, X. Wang, R. Shalaby, and E. Shtivelman. Alternatively spliced products cc3 and tc3 have opposing effects on apoptosis. 20(2):583–93, Jan 2000.

[174] S. A. Wolfe, L. Nekludova, and C. O. Pabo. Dna recognition by cys2his2 zinc finger proteins. *Annu Rev Biophys Biomol Struct*, 29:183–212, 2000.

[175] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue):D187–91, 2006.

[176] Y. Wu, L. Yu, G. Bi, K. Luo, G. Zhou, and S. Zhao. Identification and characterization of two novel human SCAN domain-containing zinc finger genes ZNF396 and ZNF397. *Gene*, 310:193–201, May 2003.

[177] K. Xia, Z. Fu, L. Hou, and J.D. Han. Impacts of protein-protein interaction domains on organism and network complexity. *Genome Res.*, 18:1500–1508, Sep 2008.

[178] Y. Xing, A. Resch, and C. Lee. The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.*, 14:426–441, Mar 2004.

[179] Y. Xing, Q. Xu, and C. Lee. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Lett.*, 555: 572–578, Dec 2003.

[180] J. Yamada, Y. Kuramochi, M. Takagi, T. Watanabe, and T. Suga. Human brain acyl-CoA hydrolase isoforms encoded by a single gene. *Biochem. Biophys. Res. Commun.*, 299: 49–56, Nov 2002.

[181] Yuzhen Ye and Adam Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19 Suppl 2:II246–II255, Oct 2003.

[182] Yuzhen Ye and Adam Godzik. Multiple flexible structure alignment using partial order graphs. *Bioinformatics*, 21(10):2362–2369, May 2005. Evaluation Studies.

[183] B.K. Yeh, M. Igarashi, A.V. Eliseenkova, A.N. Plotnikov, I. Sher, D. Ron, S.A. Aaronson, and M. Mohammadi. Structural basis by which alternative splicing confers specificity in fibroblast growth factor receptors. *Proc. Natl. Acad. Sci. U.S.A.*, 100:2266–2271, Mar 2003.

[184] Q. Zhang, R. Menon, E.W. Deutsch, S.J. Pitteri, V.M. Faca, H. Wang, L.F. Newcomb, R.A. Depinho, N. Bardeesy, D. Dinulescu, K.E. Hung, R. Kucherlapati, T. Jacks, K. Politi, R. Aebersold, G.S. Omenn, D.J. States, and S.M. Hanash. A mouse plasma peptide atlas as a resource for disease proteomics. *Genome Biol.*, 9:R93, 2008.

[185] Y. Zhang, I.A. Hubner, A.K. Arakaki, E. Shakhnovich, and J. Skolnick. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. U.S.A.*, 103:2605–2610, Feb 2006.

[186] Y. Zhang and J. Skolnick. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. U.S.A.*, 102:1029–1034, Jan 2005.

[187] Y. Zhang and J. Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, 33:2302–2309, 2005.

[188] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, Dec 2004.

[189] R Zimmer, M Wohler, and R Thiele. New scoring schemes for protein fold recognition based on Voronoi contacts. *Bioinformatics*, 14(3):295–308, 1998.

Acknowledgements

Diese Arbeit wäre ohne die Hilfe von vielen Menschen, die mich in diesen Jahren begleitet haben, nicht möglich und nur halb so spannend gewesen. Besonders bedanken möchte ich mich bei Prof. Dr. Ralf Zimmer für seine Unterstützung, seine Ideen, sein offenes Ohr zu jeder Zeit und dafür dass er mir die Freiheit gegeben hat meine eigenen Ideen umzusetzten und am Lehrstuhl weiterzuentwickeln. Danken möchte ich auch meinen Kollegen, insbesondere Gergely Csaba, Dr. Jan Erik Gewehr, Prof. Dr. Stefan Kramer, Dr. Robert Küffner, Franziska Schneider und Frank Steiner für die hervorragende und inspirierende Zusammenarbeit bei gemeinsamen Artikeln, Projekten, in der Lehre und am Lehrstuhl.
Dank schulde ich Prof. Dr. Dimtrij Frishman für sein Interesse an dieser Arbeit und seine Bereitschaft diese zu begutachten, sowie Prof. Dr. Hans-Peter Kriegel und Prof. Dr. Volker Heun, dafür dass sie sich bereit erklärt haben als Dritt- bzw. Viertprüfer der Prüfungskommision anzugehören.

Großer Dank gebührt natürlich meinen Eltern, meinen beiden Geschwistern und meinen Freunden auf deren Unterstützung und Freundschaft ich mich zu jeder Zeit verlassen kann und konnte.

Ganz besonders möchte ich mich bei meiner Frau Eva bedanken, die nun seit fast 12 Jahren an meiner Seite ist und mir in allen Lebenslagen mit all ihrer Kraft hilft. Auch wenn sie immer noch nicht so genau weiß was Bioinformatik ist und was ich den ganzen Tag mache, enthält diese Arbeit einen großen Teil von ihr. Danke für so Vieles und Alles!

Curriculum Vitae

Fabian Birzele wurde am 10. Juli 1979 in Würzburg, Deutschland, geboren. Er besuchte die Grundschule St. Walburg und das Gabrieli-Gymnasium in Eichstätt wo er 1999 mit dem Abitur abschloss. Er leistete seinen Zivildienst von August 1999 bis Juli 2000 im Altenheim St. Elisabeth, wo er mit der Betreuung und Pflege alter Menschen betraut war. Im Oktober 2000 begann er das Studium der Bioinformatik an der Ludwig-Maximilians-Universität München welches er im April 2005 mit dem Diplom mit Auszeichnung (1.0) abschloss. Im Juni 2005 trat er eine Position als wissenschaftlicher Mitarbeiter am Lehrstuhl für praktische Informatik und Bioinformatik der Ludwig-Maximilians-Universität München unter der Leitung von Prof. Dr. Ralf Zimmer an.

Publications

Gergely Csaba, **Fabian Birzele** and Ralf Zimmer, Protein Structure Alignment considering Phenotypic Plasticity, Bioinformatics, 2008, 24:i98-i104

Fabian Birzele, Gergely Csaba and Ralf Zimmer, Alternative Splicing and Protein Structure Evolution, Nucleic Acids Research, 36:2, 2008, 550-558

Fabian Birzele, Robert Küffner, Franziska Meier, Florian Oefinger, Christian Potthast, Ralf Zimmer. ProSAS: A database for analyzing alternative splicing in the context of protein structure. Nucleic Acids Research, 36, 2008, D63-D68

Fabian Birzele, Jan E. Gewehr, Ralf Zimmer. AutoPSI: A database for automatic structural classification of protein sequences and structures. Nucleic Acids Research, 36, 2008, D398-D401

Fabian Birzele, Jan E. Gewehr, Gergely Csaba and Ralf Zimmer. Vorolign - Fast Structural Alignment using Voronoi Contacts, Bioinformatics, Vol. 23, No. 2, 2007, e205-e211

Fabian Birzele and Stefan Kramer. A New Representation for Protein Secondary Structure Prediction based on Frequent Patterns, Bioinformatics, Vol. 22, No. 24, 2006, 2628-2634.

Fabian Birzele, Jan E. Gewehr, and Ralf Zimmer. QUASAR - Scoring and Ranking of Sequence-Structure Alignments, Bioinformatics, Vol. 21, No. 24, 2005, 4425-4426.

Facius A, Englbrecht C, **Birzele F**, Groscurth A, Benjamin S, Wanka S, Mewes W., PRIME: a graphical interface for integrating genomic/proteomic databases, Proteomics. 2005 Jan;5(1):76-80.

Submitted Manuscripts

Fabian Birzele, Eva Hoffmann, Gergely Csaba and Ralf Zimmer, Alternative Splicing and Protein Repeats, submitted

Fabian Birzele and Ralf Zimmer, Alternative Splicing and Proteome Complexity in Mass Spectrometry Data, submitted

Robert Küffner, **Fabian Birzele** and Ralf Zimmer, PASS: Reliable detection of alternative splicing on microarrays, submitted

Gergely Csaba, **Fabian Birzele** and Ralf Zimmer, Systematic Comparison of SCOP and CATH: A new Gold Standard for Protein Structure Analysis, submitted

Refereed Talks

Fabian Birzele, Alternative Splicing and Protein Structure Evolution, Highlights Track Talk, ISMB 2008, Toronto

Fabian Birzele, Alternative Splicing and Protein Structure Evolution, Highlights Track Talk, GCB 2008, Dresden

Awards

Best student paper award (1st prize), Vorolign - Fast structural alignment using Voronoi Contacts, ECCB 2006, Eilat