# Functional Sites in Structure and Sequence

## - Protein Active Sites and miRNA Target Recognition -

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

**Alexander Stark**

aus

Bietigheim

2004

Berichterstatter: Prof. Dr. Schomburg

Prof. Dr. Waffenschmidt

Tag der mündlichen Prüfung: 14. Juni 2004

# Table of Contents

# List of Figures

## List of Tables

# Zusammenfassung

Die Anzahl bekannter Proteinstrukturen wächst exponentiell und sogenannte „Structural Genomics“ Projekte haben es sich zum Ziel gesetzt, die Strukturen aller Proteine aufzuklären, um dadurch deren Funktionen zu bestimmen. Ich habe in meiner Dissertation eine Methode zum Vergleich lokaler struktureller Muster, wie zum Beispiel katalytischer Zentren von Enzymen, entwickelt. Die Methode berechnet die statistische Signifikanz von Suchergebnissen und erlaubt dadurch die Unterscheidung von bedeutsamen und zufällig auftretenden Ähnlichkeiten. Sie stellt eine wichtige Ergänzung zu Methoden dar, die Proteine anhand deren Gesamtstruktur vergleichen („Structural Alignment“), da signifikante Suchergebnisse sowohl Funktionen bestätigen können, die aufgrund ähnlicher Gesamtstrukturen vermutet werden, als auch funktionelle Ähnlichkeiten in Proteinen unterschiedlicher Gesamtstruktur vorhersagen oder erklären können. Im Internet ist eine einfach zu bedienende Benutzeroberfläche für die Funktionsuntersuchung von Proteinstrukturen verfügbar (http://pints.embl.de).

Im zweiten Teil meiner Dissertation präsentiere ich eine systematische computerbasierte Suche nach *Drosophila* Genen, die von microRNAs (miRNAs) reguliert werden („Targets“). miRNAs sind kurze RNA Moleküle, die in Tieren die Translation ihrer Targets blockieren, indem sie an komplementäre Stellen in deren 3’ untranslatierten Bereichen binden. Methoden zur Vorhersage von miRNA Targets wurden dringend benötigt, da Targets für nur drei der insgesamt 700 bekannten miRNAs beschrieben waren. Sechs meiner Target-Vorhersagen wurden experimentell bestätigt und viele weitere sind mit hoher Wahrscheinlichkeit ebenfalls zutreffend, so dass die Ergebnisse eine wertvolle Hilfe zur Erforschung von miRNAs darstellen. Die Studie erweitert die bislang bekannten Funktionen von miRNAs um die Kontrolle ganzer Signaltransduktions- und Stoffwechselwege sowie ihre Beteiligung an der Entwicklung des Nervensystems. Weiterhin zeigte sich, dass eine miRNA oft mehrere Targets kontrolliert, umgekehrt aber auch ein Gen von mehreren miRNAs reguliert werden kann.

# Abstract

The number of protein three-dimensional structures is increasing steeply, and structural genomics projects aim to solve the structures for all proteins as a means to understanding function. In the first part of my thesis, I developed a method for the comparison of local structural patterns (e.g. enzyme active sites) that provides a reliable statistical measure to discern meaningful matches from noise. The method is complementary to structural alignment as it is able to confirm functional similarities suggested by an overall similar structure but also detects functional similarities between different folds. An easy-to-use interface is available on the Internet for functional annotation of protein structures (http://pints.embl.de).

In the second part of my thesis, I present a computational screen for microRNA (miRNA) targets in *Drosophila*. miRNAs are short RNAs that inhibit translation of target messenger RNAs in animals by binding to complementary sites in their 3' untranslated regions. Target predictions were urgently needed as targets were known for only three of the more than 700 miRNAs. Of my predictions, six were validated experimentally and others are likely to be functional, making the results a useful resource for miRNA research. The screen extended miRNA function to pathway control, nervous system development and regulation of metabolism, and revealed that one miRNA typically regulates several targets but also that one gene is likely to be targeted by several miRNAs.

# 1 Introduction

During the last decade, the number of known biological sequences has increased dramatically and complete genome sequences of many prokaryotes and some eukaryotes including human are now available. The key problem of today's biology is to make sense of these sequences and to understand their function and functional interplay. Traditionally, genes and their functions were identified in laborious and time-consuming experiments prior to the knowledge of the genes' sequence. Bioinformatics tries to infer function directly by means of comparisons, as high sequence or structure similarity between two proteins is often indicative of a common function. However, some similarities required for detecting specific functions can also be quite subtle and may comprise only a few residues.

For my thesis, I worked on two types of subtle similarities. Both are difficult to detect but can be sufficient to infer function. In the first part of the thesis, I describe a new method for functional annotation of protein structures by the comparison of active sites. In the second, I present a screen for short sequence motifs indicative of microRNA target genes in *Drosophila*. I then, summarize four other projects to which I contributed. In the introductory section that follows, I review methods typically used to compare protein sequences and structures and discuss the problem of statistical significance. I then introduce microRNAs as a novel class of genes and discuss the importance of microRNA target prediction.

## *1.1 Protein Structure, Function and Evolution*

Proteins are central to all biological processes, including metabolism, immune response, signal transduction and gene expression, and defective proteins have been implicated in many human diseases. It is thus crucial to know the functions of the growing number of proteins identified from genome sequencing. Many proteins (especially large eukaryotic proteins) are modular: they consist of several *domains* that are able to fold independently into stable structures. Domains also frequently contribute distinct molecular functions to the overall protein. Sequence and structural similarity (see next sections) have shown that domains are often re-used by nature: a domain may occur in different contexts and combinations and prediction of function must thus take the overall domain structure into account (e.g. Apic *et al.*, 2001; Chervitz *et al.*, 1998; Copley *et al.*, 2002a; Copley *et al.*, 2002b; Gerstein & Levitt, 1997; Koonin *et al.*, 2002). Evolutionary and structural classifications of proteins are often

domain-centric (e.g. Pfam and SCOP; see below) and here, I will use the terms protein and domain interchangeably.

Proteins evolve via point mutations or insertion/deletion events in their corresponding genes, or less frequently through duplications of partial or whole genes. This process of accumulating changes that ultimately results in new functions or species is called *divergent evolution*. All proteins that diverged from a common ancestor (i.e. have a shared evolutionary history) are *homologous* irrespective of current similarity or function. However, our ability to reliably assign homology to two proteins is dependent on the remaining similarity that is preserved from the ancestor. Over short evolutionary distances, this is possible through sequence comparison (see Chapter 1.2). Proteins then typically have the same structure and a similarity in function and are thus often clustered into (homologous) *families*. However, over long distances eventually too many mutations accumulate and these methods fail as the similarity between homologous proteins becomes indistinguishable from the similarity between two random sequences.

Protein three-dimensional (3D) structures can be similar in the absence of detectable sequence similarity and structure is thus often said to be more conserved than sequence. Proteins structures are typically compared or classified according to their *folds*, the spatial arrangements of secondary structure elements and their connectivity (e.g. Blundell & Johnson, 1993; Murzin *et al.*, 1995b). Structural alignment methods (see Chapter 1.4.1) are often able to detect remote homologies not evident from the sequence. However, not all proteins that adopt a similar fold are homologous. For instance, despite the large number of protein sequences sampled during evolution, it has been estimated that all naturally occurring proteins belong to only a few thousand folds (Blundell & Johnson, 1993; Chothia, 1992; Koonin *et al.*, 2002; Orengo *et al.*, 1994a) suggesting that independent convergence is prevalent. A more theoretical approach enumerated all possible sequences for an average sized domain (150 residues) and predicted that about $10^{26}$ different sequences with less than 20% sequence identity might adopt one common fold (Branden & Tooze, 1999). Proteins with the same fold that are not thought to share a common ancestor but rather believed to have evolved independently and converged to a stable structure are often referred to as *analogs* (Fitch, 1970). Analogy is a more general term in biology, used to describe *convergent evolution*: classical examples of analogy are fly and bird wings that arose independently to serve the same function. Interestingly, the population of different folds varies greatly with

only a few folds accounting for more than half of all protein structures. One of these so-called *superfolds* (Orengo *et al.*, 1994b) is the $(\beta\alpha)_8$-(TIM)-barrel adopted by many metabolic enzymes and first named after triosephosphate isomerase (TIM, see for example Farber & Petsko, 1990). Remote homology has been discovered between many of these enzyme families but for others independent evolution seems more likely (Copley & Bork, 2000; Nagano *et al.*, 2002).

Protein function is largely determined by structure and some highly specialized structures have consequently been developed. Some proteins such as collagen or keratin form elongated fibres to withstand mechanical stress; others are globular and suitable for transport (e.g. albumin or haemoglobin). The immunoglobulin (Ig) fold is common to many proteins in the immune system and is well suited as a general template for recognition and binding of a variety of molecules (e.g. Bork *et al.*, 1994). Specialized structures are also apparent on the level of domains or motifs: domains that mediate protein-protein interactions are important components of many signal transduction proteins (e.g. SH2 and SH3; see Pawson & Gish, 1992) and structural motifs such as helix-loop-helix (HLH) motifs or zinc-fingers are often common to otherwise different proteins to mediate protein-DNA interactions (e.g. Nelson, 1995).

In many cases, only a small part of the protein – the *active site* – is directly involved in a specific function and the rest serves other functions (e.g. regulation, interactions, etc.) or acts only as a scaffold. For example, the active site of the serine proteases is a characteristic arrangement of a serine, histidine and an aspartate residue (catalytic triad; e.g. Dodson & Wlodawer, 1998). Other proteins often rely on cofactors such as metal ions or small molecules (also called coenzymes) to carry out their function. Some cofactors are only transiently associated with the enzyme and function as cosubstrates (e.g. ATP, NAD) whereas prosthetic groups such as heme or FAD are permanently bound to the enzyme. Catalytic residues or cofactor binding sites can be common to unrelated proteins and comparisons of active sites can thus give clues about function independent of overall sequence or structural similarity (see Chapter 1.4.2.2 and Denessiouk *et al.*, 1998; Kobayashi & Go, 1997a/b; Nobeli *et al*., 2001).

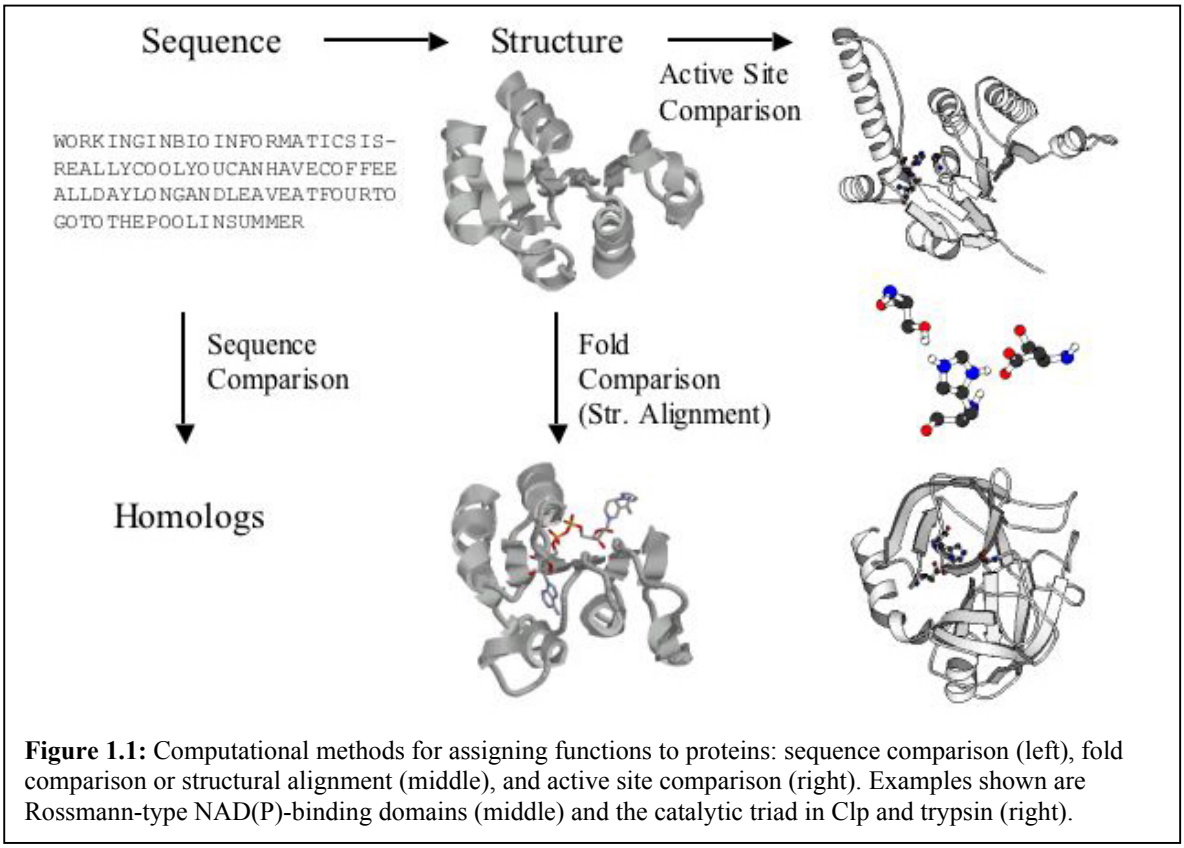



**Figure 1.1:** Computational methods for assigning functions to proteins: sequence comparison (left), fold comparison or structural alignment (middle), and active site comparison (right). Examples shown are Rossmann-type NAD(P)-binding domains (middle) and the catalytic triad in Clp and trypsin (right).

Computational methods to assign functions to proteins compare sequences (Chapter 1.2), overall structures (folds, 1.4.1) or structural details such as active site residues (1.4.2.2, see Figure 1.1 for an overview). A sufficiently non-random similarity that covers the whole length of a protein can be used to establish homology and often to infer function. In contrast, functional similarities can be common to unrelated proteins and can thus reside in completely different sequence or structural contexts. The residues involved are typically close in sequence or structure and comprise only a very small fraction of the overall protein. Examples include independently evolved enzyme active sites or short linear motifs that are recognized by cellular receptors. If this type of similarity is reliably detected, it can be used directly for functional annotation.

## *1.2 Protein Sequence Comparison*

### 1.2.1 Pairwise Sequence Alignments and Sequence Searches

Protein sequence searches are generally performed as pairwise comparisons of a query sequence to each sequence in a database. For this, each pair of sequences is aligned so that a maximal number of identical or similar residues match, while mismatches or gaps are penalized. All algorithms to perform this rely on schemes to score amino acid

substitutions/changes between the sequences (e.g. Durbin *et al.*, 1998). The simplest scheme considers only identical residues and gives a percent identity value for a pair of sequences. Others try to weight the differences by considering the physico-chemical properties of the amino-acids (for a review see Vogt *et al.*, 1995). However, empirical *substitution matrices* that are derived from observed substitution frequencies in multiple sequence alignments of homologous sequences are now used almost exclusively (PAM (Dayhoff *et al.*, 1978) or BLOSUM (Henikoff & Henikoff, 1992)).

Exact alignment algorithms based on dynamic programming (Needleman-Wunsch (Needleman & Wunsch, 1970) or Smith-Waterman algorithm (Smith & Waterman, 1981)) are guaranteed to find the optimal alignment according to a particular scoring scheme but are often too slow for comparisons involving large databases. Therefore, heuristic algorithms have been developed that allow much faster searches but sacrifice some sensitivity, i.e. they may miss distantly related sequences or lead to errors in ranking of remote matches.

Fasta (Pearson & Lipman, 1988) and BLAST (Altschul *et al.*, 1990) are the most widely used programs for sequence similarity searches. Both search first for identical short stretches in both sequences to align (also called words or k-tuples) and then extend and join these words into an alignment. Whereas Fasta searches for all possible words of a given length, BLAST is based on the observation that alignments for true matches likely contain short stretches of very high similarity and thus considers only the most significant of these. The speed of the programs increases and sensitivity decreases from the Smith-Waterman algorithm to Fasta and BLAST (Pearson, 1995). Both programs allow searches with protein or nucleic acid queries against databases containing either sequence type and associate matches with measures for statistical significance that allow the user to compare results from different searches and distinguish true from random matches (see Chapter 1.5).

### 1.2.2 Multiple Sequence Alignments

When a database search reveals several matches, a logical next step is to create a *multiple sequence alignment (MSA)*. MSAs are powerful tools when looking for residues that are especially important for structural or functional reasons. Whereas they may not be revealed in a pairwise alignment (e.g. when the two sequences are too divergent or too similar), important residues usually stand out in an MSA containing several related sequences.

Three widely used programs for multiple sequence alignments are PileUp (Feng & Doolittle, 1987), ClustalW (Thompson *et al.*, 1994), and T-Coffee (Notredame *et al.*, 2000). These programs use a progressive alignment algorithm that first calculates all pairwise sequence alignments and similarities and builds a guidance tree (dendrogram). It then starts joining the two most similar sequences and extends the alignment by adding single or multiple sequences with decreasing similarity according to the dendrogram until all sequences are aligned.

### 1.2.3 Sequence Profiles and Hidden Markov Models

In a multiple sequence alignment of homologous proteins, some positions are typically more conserved or show higher preferences for some type of amino acid (e.g. hydrophobic) than others. This is due to structural requirements (e.g. disulphide bonds, hydrophobic core) or functional features (e.g. catalytic residues) specific to that family. *Sequence profiles* (also called *weight matrices* or *position specific score matrices* (*PSSM*) (Henikoff & Henikoff, 1994)) capture these features by specifying for each position the frequency that each amino acid appears (Gribskov *et al.*, 1987). A profile not only captures the different amino acid preferences at different positions but also highlights positions of particular importance or weights down others that are not conserved. Profiles are thus more sensitive than the traditional methods used to search databases that assign equal weight to each position along the sequence (e.g. Brenner *et al.*, 1998; Park *et al.*, 1998).

Position-Specific Iterated BLAST (PSI-BLAST, Altschul *et al.*, 1997) makes use of sequence profiles to allow more sensitive sequence searches. It first compares a query sequence to a database and builds a profile from the matches. It then iteratively searches the database with the profile and adds all new matches to the profile until no new sequence is detected. By successively incorporating more sequences, the method efficiently detects remote homologs, hence the general usage of PSI-BLAST in genome annotation and structure prediction.

Although sequence profiles capture some conservation features, they are inadequate to represent all the information in an MSA of a protein family. One must still rely on (arbitrary) gap penalties as in pairwise sequence alignment, or combine multiple ungapped blocks described below for BLOCKS. Hidden Markov models (HMMs) provide a full probabilistic model for all sequences in a sequence family (e.g. Baldi *et al.*, 1994; Eddy, 1998; Krogh *et*

*al.*, 1994). They consist of a repetitive structure of different *states*, typically *match*, *insertion* and *deletion*. For each position in the alignment, the probabilities for the transition between the states (e.g. match→insertion or match→match) and for the value of the matches (i.e. the amino acid preferences) are different. Given a collection of sequences that need not be aligned, all probabilities are adjusted so that sequences similar to those in the training set score best. This also means that a random walk through the states of the model considering the different probabilities most likely leads to a sequence from the training set (this is why it is said that HMMs *emit* sequences and *match* states are sometimes called *emission* states). HMMs thus provide a complete statistical framework for sequence searches and alignments including a consistent treatment of gaps (insertions) and deletions. If enough sequences are known to train an HMM for a given family, it allows the most sensitive sequence searches possible and provides reliable significance scores to all matches. The most widely used HMM software packages are HMMer (http://hmmer.wustl.edu, Eddy, 1998), and SAM (Karplus *et al.*, 1998).

Several pattern databases store conserved features from multiple sequence alignments and derived sequence profiles. The evolutionary information of individual protein families can be used to infer family membership for new sequences. The most widely used include PROSITE (Hulo *et al.*, 2004) that uses patterns (regular expressions) and sequence profiles characteristic for a protein family or domain, the BLOCKS database (Henikoff & Henikoff, 1991) that stores ungapped multiple alignments (blocks) that correspond to the most conserved regions of documented protein families, and PRINTS-S (Attwood *et al.*, 2003) that uses the most conserved regions of multiple sequence alignments to build signatures (fingerprints) diagnostic for family membership. The databases and analysis tools Pfam (Bateman *et al.*, 2002) and SMART (Letunic *et al.*, 2004; Schultz *et al.*, 1998) make use of the great value of MSAs and their representation as HMMs. Both are essentially collections of carefully constructed MSAs for different protein domain families. Sequences can be searched against HMMs built for each family to detect domain recurrences, help classifying the protein, or aid in functional annotation. The metasite InterPro combines many resources of this type and provides a consensus view that overcomes the specific weaknesses of the individual databases (Apweiler *et al.*, 2001).

## 1.2.4 Limits of Sequence Comparison

### 1.2.4.1 Thresholds for Inference of Homology

In 1986, Cyrus Chothia and Arthur Lesk reported the first systematic comparison of structures from different protein families and showed that the extent of structural changes is directly related to the extent of sequence changes. Specifically, the overall structural divergence measured by the root mean square deviation (RMSD) of the superimposed backbone Cα atoms increased exponentially with decreasing residue identity (Chothia & Lesk, 1986), a trend that was later confirmed on a much larger scale and for different measures of sequence similarity (Russell & Barton, 1994; Russell *et al.*, 1997; Wilson *et al.*, 2000).

However, it became clear that shorter alignments require a higher degree of similarity for structural significance. A systematic comparison of protein sequence and structure determined the sequence identity required to infer structural similarity dependent on the length of the alignment, allowing the authors to quantify the notion that "two protein sequences are sufficiently similar to be considered homologous" (homology cutoff; Sander & Schneider, 1991). This cut off is at 25% sequence identity for long alignments whereas for alignments shorter than 70-80 residues, the structural significance of a given sequence identity drops sharply and at 10 or fewer residues, even 100% sequence identity is not sufficient to infer homology and/or structural information. Indeed, identical pentapeptides have been found to adopt completely different structures (Kabsch & Sander, 1984). Sequences with similarities below this threshold are in the so-called *twilight zone*: they are not necessarily unrelated but their homology remains uncertain. Modern methods for sequence comparison take the length of the sequences into account and provide reliability scores for the likelihood that matches are meaningful (see Chapter 1.5).

### 1.2.4.2 Thresholds for Reliable Inference of Function

Divergent evolution implies that the descendents of a given ancestral protein (i.e. homologs by definition) have adapted to perform different functions. This is clear for remote homologs where 10% were found to have completely different functions (Hegyi & Gerstein, 1999; Russell *et al.*, 1998). However, depending on the number of changes necessary to achieve this, proteins can still exhibit remarkable sequence similarities despite functional differences. Functional similarities cannot be measured with a simple metric but instead rely on

classifications schemes. The Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) assigns a four-digit EC (Enzyme Commission) number to all enzymes to classify them according to the nature of chemical reactions they catalyze. The first digit denotes the class of reaction (i.e. oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases) and the remaining levels specify the reaction more precisely (i.e. substrates, etc. – but the exact meaning of each level depends on the primary number, Webb *et al.*, 1992). The EC classification system is used by all major databases specializing in protein sequences (e.g. SwissProt; Boeckmann *et al.*, 2003), enzymes (e.g. BRENDA; Schomburg *et al.*, 2004, IntEnz; Fleischmann *et al.*, 2004), or complete metabolic pathways (e.g. KEGG; Kanehisa *et al.*, 2004).

In the last few years, several groups used this system to examine the relationship between sequence similarity and similarity in function and to establish thresholds for reliable inference of function (Rost, 2002; Tian & Skolnick, 2003; Todd *et al.*, 2001; Wilson *et al.*, 2000). The thresholds reported in these publications vary greatly: Todd *et al.* (2001) and Wilson *et al.* (2000) found that 40% sequence identity was sufficient to transfer precise function reliably or 30% sequence identity to transfer the first 3 digits of the EC number with 90% accuracy, whereas Rost reported that less than 30% of the sequence pairs above 50% identity have entirely identical EC numbers and that even BLAST E-values below $10^{-50}$ did not allow transfer of enzyme function without errors (Rost, 2002). Indeed, even the most cautious thresholds do not account for outliers, such as the extreme example of melamine deaminase and atrazine chlorohydrolase, which have different overall function despite 98% sequence identity (Seffernick *et al.*, 2001). More general examples include the so-called non-catalytic enzymes that are similar to active enzymes but have lost their catalytic function (e.g. crystallins in vertebrate lenses that are homologous to glutathione S-transferases and other enzymes (Tomarev & Zinovieva, 1988; Wistow & Piatigorsky, 1987), carboxypeptidase II that has homologous catalytic and non-catalytic domains (Aloy *et al.*, 2001a), the similarity between the signalling factor sonic hedgehog and the transthyretin domain of carboxypeptidase (Gomis-Ruth *et al.*, 1999), significant similarity between α-lactalbumin and lysozyme, or more recently *Mycobacterium tuberculosis* MPT51 as a founding member of a new family of non-catalytic α/β hydrolases (Wilson *et al.*, 2004; for a review see Murzin, 1993a; Murzin, 1998; Todd *et al.*, 2002).

### 1.2.5 Sequential Motifs

Some important functional parts of proteins such as post-translational modification sites (Yaffe *et al.*, 2001), targeting signals for specific cellular compartments (Emanuelsson *et al.*, 2000), and protein interaction or cleavage sites (Nielsen *et al.*, 1997) consist of only a few residues. They often occur in exposed loops or unstructured regions outside globular domains and can be common to unrelated proteins. Whilst consensus sequences (*motifs*) have been experimentally determined for many such sites (e.g. phosphorylation sites), others show only preferences for certain residue types such as the positively charged residues found in nuclear targeting signals (Cokol *et al.*, 2000).

Predictions based on sequence searches with motifs are usually not specific and give a high number of false positive matches (i.e. they tend to *overpredict*) because short sequences often match by chance (the same problem exists for miRNA target prediction, see below and Chapter 3.5). Protein kinase C for example recognizes the tripeptide SVK, but searches for that motif match about every tenth protein and the vast majority of the matches are not functional. The largest resource for linear motifs is the ELM (eukaryotic linear motif) server (Puntervoll *et al.*, 2003), followed by PROSITE (Hulo *et al.*, 2004) and Scansite (Yaffe *et al.*, 2001). ELM tries to reduce overprediction using different filters such as a cellular compartment filter, a globular domain filter or a taxonomy filter that remove sequence matches unlikely to be functional. Scansite is specific for motifs involved in signalling pathways (e.g. kinase recognition motifs, protein interaction motifs) and scores matches with a sequence profile around the putative sites to enhance the specificity.

## *1.3 Structure Prediction*

The structure of a protein is ultimately determined by its sequence, and scientists have long tried to predict protein structure from sequence. Resulting models can provide a more detailed picture of protein function but can also reveal similarities not apparent from the sequence. I will review three types of methods that are applicable depending on the similarity of the protein of interest to proteins of known structure.

### 1.3.1 Homology Modelling

Homology modelling (or comparative modelling) is based on the observation that homologous proteins have a common fold. It uses experimentally defined protein structures as

*templates* to predict the conformation of other proteins with similar sequence (*targets*). The first modelling studies were carried out in the late 1960s and early 1970s for α-lactalbumin (Browne *et al.*, 1969) and the α-lytic protease (McLachlan & Shotton, 1971).

Today, obtaining 3D structures by homology modelling can be achieved with an accuracy approaching that of a low-resolution X-ray structure or a medium resolution NMR structure for protein sequences that have at least 35-40% sequence identity to a known structure (Sali, 1998; Sanchez *et al.*, 2000). This vastly increases the number of proteins for which reasonable structural information can be inferred and facilitates the design of experiments and functional annotation (Vitkup *et al.*, 2001).

All current methods are based on four steps. The first is to identify the templates, i.e. proteins with known structure that exhibit significant sequence similarity to the target sequence. Then target and template sequences are aligned and the most suitable template is chosen. The third step is to build the structural model for the target based on the alignment to the template and the template structure. Finally, the model is evaluated by several criteria and the alignment and model building is improved until a satisfactory model is obtained (for a review, see Marti-Renom *et al.*, 2000). At present, the main problems in homology modelling are template selection and alignment, modelling insertions and deletions (i.e. regions without template structure), and predicting the side-chain packing (Tramontano & Morea, 2003).

### 1.3.2 Threading or Fold Recognition

There are many examples where proteins share a similar 3D structure despite having no apparent sequence similarity. Threading programs make use of this observation by testing how well a sequence fits a particular fold. After "threading" the sequence through a collection of 3D template folds, the programs evaluate the fit by energetic or statistical potentials (e.g. Sippl, 1995). The original methods created the sequence-to-structure alignment purely by optimizing these potentials using time-consuming algorithms (Godzik *et al.*, 1992; Jones *et al.*, 1992). Newer approaches like GenTHREADER (Jones, 1999a) or 3D-PSSM (Kelley *et al.*, 2000) align the target sequences to profiles derived from sequence or structure alignments. The statistical potentials are only used to score the different fits, often in combination with other measures such as solvent exposure or secondary structure. In the past three years, several consensus prediction methods or meta-predictors were developed to

combine predictions from a variety of available methods and are often more accurate than any individual server (Bujnicki *et al.*, 2001). Pcons (Lundstrom *et al.*, 2001) ranks predictions generated by a set of servers with a scoring function that takes the confidence of the servers' predictions into account. 3D-SHOTGUN (Fischer, 2003) reassembles high-scoring fragments taken from different predictions into a new model that can be closer to the native structure than any of the original models. Another approach (3D-Jury) assumes that the most abundant model (i.e. a structure that is predicted by many methods) is closer to the native structure than any single model and thus re-ranks the models according to their abundance (Ginalski *et al.*, 2003).

Threading is able to identify similarities that are not found by conventional sequence comparison. Although the more sensitive methods like PSI-BLAST or HMMer are often able to identify remote similarities and bridge the gap between comparative modelling and threading, threading methods (e.g. Kelley *et al.* 2000) can often successfully detect structural similarity for proteins where PSI-BLAST and HMMer fail.

### 1.3.3 Ab initio Methods

When homology modelling and threading methods fail, hints about protein structure can come from *ab initio* or template-free methods. One class of these methods predicts the secondary structure, i.e. whether local segments adopt an α-helix, β-sheet or a coiled structure. Knowing something about the secondary structure of a protein is often seen as a necessary step towards determining the full structure. It has indeed been used for fold prediction (e.g. Koretke *et al.*, 1999; Russell *et al.*, 1996; Sheridan *et al.*, 1985) and fragment-based methods described below). The first generation of methods such as those by Chou & Fasman (1974) or Garnier *et al.* (1978) was developed during the 1970s. They averaged the empirical propensities of residues to adopt one of the three secondary structure states over segments with a typical length of 11-21 residues but seldom achieved accuracies better than 60%. This was greatly improved by incorporating evolutionary information derived from MSAs with artificial neural networks into methods like PhD (Rost & Sander, 1993) or PSI-PRED (Jones, 1999b), which could reach accuracies above 70%. New consensus methods (meta-predictors) such as Jpred (Cuff *et al.*, 1998) combine several methods and are currently the most accurate. However, the improvements in performance tend to plateau at around 80% accuracy (Aloy *et al.*, 2003b),

which is expected given the conservation of secondary structure in homologous proteins (Kabsch & Sander, 1983b; Rost *et al.*, 1994; Russell & Barton, 1993).

During the last few years, methods have been developed that show remarkable success in predicting overall structures for novel folds (Aloy *et al.*, 2003b; Lesk *et al.*, 2001). They are based on the assumption that all possible structures that can be adopted by small fragments (e.g. 3 and 9 residues) are sufficiently sampled in known folds (Bystroff & Baker, 1998; Jones, 1997). For structure prediction, the sequence of interest is first compared to sequence profiles created from structurally similar fragments (fragment library) and high-scoring fragments are then assembled considering secondary structure prediction, hydrophobic burial and steric clashes. Recently, one of these methods (Rosetta, Bonneau *et al.*, 2002) was used to predict structures for major protein families with no structural information. Interestingly, some predictions showed structural similarity to known folds with similar functions that was not found by either sequence comparisons or fold recognition. This establishes *a priori* structure predictions as a new means for the annotation of function.

### 1.3.4 CASP

Since 1994, the performance of protein structure prediction methods has been assessed every two years in CASP (Critical Assessment of Structure Prediction). In this blind test, target sequences are released for structure prediction prior to the availability of the experimentally solved protein structure (Moult *et al.*, 1995). This forces predictors to exercise caution when making claims of success and CASP has thus played a major role in charting the progress of the field. In autumn 2002, our group assessed the predictions in the CASP5 *new fold* (formerly *ab initio*) category and presented the results in the subsequent meeting at Asilomar in California (USA). I will summarize our assessment in Chapter 3.6.3.

## *1.4 Assigning Function from Protein Structure*

The 3D structure of a protein provides a much more detailed view of its properties and function than the sequence alone. Residues can for example form spatial clusters that are not seen in the sequence but might have certain characteristics indicative of function such as positively charged surface crevices for binding DNA (e.g. Boggon *et al.*, 1999). Furthermore, catalytic mechanisms and active site residues have been determined by careful examination and comparison of structures for many enzymes. Protein structures are also much more

conserved than sequences so that remote homology can often only be detected by structure comparison.

Recent improvements in structural biology have greatly increased the number of protein structures in the public Protein Data Bank (PDB, Berman *et al.*, 2000). Today, the PDB holds more than 24,000 structures and is growing exponentially. In addition, several structural genomics projects aim to systematically solve the structures for all proteins as a means to understanding function (Burley, 2000; Burley & Bonanno, 2002; Hurley *et al.*, 2002; Kim *et al.*, 2003; Vitkup *et al.*, 2001; Zhang & Kim, 2003). Because of the increasing availability of structures and the advantages above, methods to annotate function through structure are now of growing importance. These methods belong to two classes: structural alignment or fold comparison (next section), and structural pattern matching (Chapter 1.4.2.2).

### 1.4.1 Structural Alignment

As mentioned above, remote homologs and analogs can adopt very similar structures indicative of specific functions even in the absence of detectable sequence similarity. Methods have thus been developed to compare or align protein structures independent of their sequence. They use the pairwise distances between Cα atoms (e.g. Dali (Holm & Sander, 1993; Holm & Sander, 1995), STAMP (Russell & Barton, 1992), SSAP (Orengo & Taylor, 1996)), the geometry of secondary structure elements (e.g. VAST (Gibrat *et al.*, 1996)) or a combination of different information (CE (Shindyalov & Bourne, 1998), GRATH (Harrison *et al.*, 2002)) to find proteins with a similar fold, i.e. common spatial arrangements of secondary structure elements in the same order along the sequence.

Such similarities can identify ancient evolutionary relationships that are not always apparent when only sequences are known, but that are often associated with a similarity in function. Indeed, the location of active sites, binding surfaces or substrate type is often conserved and their function can be easily tested by further experiments. Highly populated protein folds (*superfolds*) often have a common location for substrate binding sites between remote homologs or even analogs. The best-known examples are the TIM-barrels, that are known to bind substrates at the C-terminal end of the barrel-forming β-strands (Russell *et al.*, 1998).

Based on manual inspection and automated structure comparison, several structure classification systems are available, in particular SCOP (Structural Classification of Proteins), CATH (Class Architecture Topology Homology) and FSSP (Families of Structurally Similar Proteins).

SCOP (Murzin *et al.*, 1995a) organizes protein domains into a hierarchy consisting of class, fold, superfamily, family and species. Proteins are first broadly grouped into classes by their secondary structure content (i.e. all $\alpha$, all $\beta$, $\alpha/\beta$, $\alpha+\beta$) or unique features (i.e. small proteins, coiled coil proteins). Proteins within one fold share a common core as determined by manual inspection of the number, arrangement and connectivity of secondary structure elements. This structural similarity might have evolved convergently and different folds might thus represent analogous folds, although distant evolutionary links may exist. Proteins that are grouped in the same superfamily often have no detectable sequence similarity but show some evidence of a common ancestor, based on high structural similarity, conservation of unusual structural features or functions, or significant sequence identity after structural superimposition. Close homologs with high structural similarity and detectable sequence similarity are grouped into one family.

CATH (Orengo *et al.*, 1997) uses a semi-automated method to classify proteins into a hierarchy consisting of the levels class, architecture, topology and homology. Whereas class, topology and homology are comparable to class, fold and family in SCOP, the manually annotated architecture level is unique to CATH. It encapsulates broad features of the protein shape such as the orientation of secondary structure elements independent of connectivity or direction.

FSSP (Holm & Sander, 1996) presents the results of pairwise structural comparisons rather than a structural classification. Proteins with greater than 25% sequence identity are first grouped together and representatives of all groups are compared to one another using the structural alignment program Dali. The user can retrieve all significant matches of these comparisons and browse a fold tree that is computed from the results. FSSP also assigns a six-character fold index at different similarity cutoffs that do not correspond to the hierarchy levels of the other databases.

A systematic comparison of these protein structure classifications revealed that approximately two-thirds of the proteins in each database are common to all three databases and that the classifications agree in the majority of cases. No database was found to be distinctly superior and it was suggested that all three should be used in combination. The great strength of SCOP is the careful manual assignment of evolutionary relationships - even in the absence of sequence similarity - with drawbacks being update frequency and limited coverage. In contrast, FSSP is updated continuously and covers all structures in the PDB but the data are left to the users' own assessment (Hadley & Jones, 1999). We mainly used SCOP during our work as we its manual curation most reliable and useful for our purposes.

### 1.4.2 Active Site Identification and Comparison

Sequence similarity does not always imply a common function (see above) and structural alignment-based search methods do not always provide functional clues. This is clear if a protein adopts a new fold (i.e. does not resemble any known structure), but problems can also arise when proteins adopt very common folds like TIM-barrels, ferredoxins or Ig-like structures that perform many different functions, (e.g. Orengo *et al.*, 1994a). Here, functional inferences are difficult to make, since structural alignments can show an equal degree of similarity between functionally similar and dissimilar proteins. Two types of methods thus concentrate on the functional parts of proteins and aim to detect the active sites by means of sequence conservation or local structural similarities.

#### 1.4.2.1 Identification of Active Sites using Conservation

Functionally important residues are typically more conserved than the overall protein sequence because they are under evolutionary pressure to maintain their functional integrity. Furthermore, one expects a higher degree of conservation of the catalytically active residues directly involved in the reaction mechanism whereas the substrate binding residues are only conserved among close homologs but then altered to allow for different substrate specificities.

The Evolutionary Trace (ET) method (Lichtarge *et al.*, 1996) is based on these observations and "traces" conserved residues within an MSA by following an evolutionary tree. When residue conservation across different numbers of tree branches is required, universally conserved catalytic site residues (high selectivity) or intermediately conserved binding sites

(high sensitivity) can be detected. Projection of the selected residues onto the protein structure helps to manually select solvent accessible clusters of conserved residues that are potentially important for function. Many other groups published work on the identification of active sites using evolutionary conservation similar to the original ET method (Aloy *et al.*, 2001b; Armon *et al.*, 2001; Landgraf *et al.*, 2001; Lichtarge & Sowa, 2002; Madabushi *et al.*, 2002; Oliveira *et al.*, 2003; Ota *et al.*, 2003). The main differences are in automation and large-scale benchmarking (Aloy *et al.*, 2001b), more careful construction of evolutionary trees and consideration of physicochemical properties of amino acids (Armon *et al.*, 2001) and the assessment of significance for spatial clusters of conserved residues (Madabushi *et al.*, 2002). Recently, various characteristics of catalytic residues were carefully examined (Bartlett *et al.*, 2002) and used to predict catalytic residues (Gutteridge *et al.*, 2003).

#### 1.4.2.2 Comparison of Active Site 3D Patterns

Although the methods described above are sometimes able to detect active site residues by their evolutionary conservation, they do not provide further functional annotation or allow comparisons between sites, but merely highlight centres of conservation. One class of methods makes use of the growing number of known protein structures and tries to obtain functional clues by directly comparing functional sites, which can be common to proteins with different folds. Residues within these spatial *patterns* are not necessarily adjacent in the protein sequence and can occur in any order. A classic example of this phenomenon is the trypsin-like catalytic triad, which nature has reinvented more than ten times (Dodson & Wlodawer, 1998), although several other examples have been reported (e.g. Denessiouk *et al.*, 1998; Endicott & Nurse, 1995; Russell, 1998). Methods to detect the functional similarities must thus be independent of the overall sequence or fold similarity and the sequence order of the residues. This prevents the use of alignments and instead requires a view of protein structures as collections of unconnected points or atoms in space. Most of the available methods use search algorithms adapted from computer vision (geometric hashing) or mathematical graph theory. Geometric hashing requires the transformation of the coordinates to many internal reference frames and performs searches by directly comparing the coordinate values. Graph theory algorithms use the distances between atoms or residues that are independent of the absolute coordinates, i.e. the orientation of the structures. The structures are represented as graphs: the atoms or residues correspond to nodes and the distances between them to edges. The search algorithms aim to detect common patterns, i.e. sub-graphs

or cliques and are often named accordingly, although all essentially perform recursive depth-first searches (see Chapter 2.3.3). They were first successfully applied to small molecules in pharmacophoric pattern matching (see Willett (1987) for a review).

Once equivalent patterns are found, a score to assess their similarity (i.e. the quality of the match) is calculated. Most methods use the RMSD, which accurately scores the (geometrical) difference between two sets of coordinates and is often used in structure comparison. The definition and properties of RMSD are critical to assess its statistical significance and are discussed below (Chapter 3.1).

Artymiuk *et al.* (1994) developed a method (ASSAM) to search protein structures for the recurrence of user-defined side-chain patterns using a subgraph isomorphism algorithm. Residues are represented by one, two or three pseudo-atoms and patterns are characterised by the distances between these atoms. Matches are required to consist of the same type of residues with equal inter-atom distances although a distance tolerance for near-exact matches is allowed. The method successfully detected recurrences of the serine protease catalytic triad, the two-arginine active site from staphylococcal nuclease, and a zinc-binding site. However, depending on the distance tolerance, different numbers of matches are reported, that are also not associated with a similarity score and require visual inspection for ranking or separation from noise. In addition, inter-atom distances do not reflect the chirality of biological structures and discrimination of true matches from mirror images is not possible.

Fischer and co-workers used a geometric hashing algorithm (originally developed by Nussinov & Wolfson, 1991) to compare spatial patterns of Cα atoms (Fischer *et al.*, 1994). Matching Cα atoms are optimally superimposed and the RMSD is reported. In addition, an empirical similarity score is calculated from the number of equivalent Cα atoms normalized to the overall size of the two proteins. The authors detected similarities between trypsin and subtilisins (i.e. across folds) and extended their work to the comparison of protein surfaces (Lin *et al.*, 1994; Norel *et al.*, 1994). The method requires intensive pre-calculations and lacks specificity as information about the residue type or the orientation of the residue within a pattern is lost when only Cα atoms are considered. This is reflected in the large number of Cα atoms common to all matches (more than about 100 atoms or 50% of the query).

Wallace *et al.* (1996) derived a 3D template for the catalytic sites of serine proteases and demonstrated its use for finding novel examples. Subsequently, the authors introduced a geometric hashing algorithm (TESS) for the construction of templates and databases searches and created a template database PROCAT (Wallace *et al.*, 1997). Matches were evaluated by the RMSD, and the number of matches at different RMSD cutoffs was compared. A two-residue active site template from lysozyme, for example, had low specificity as many false positive matches were reported. The authors found that the number of matches for a given RMSD generally depended on the number of atoms or residues in the template but did not provide a measure for the statistical significance (i.e. the meaning, see Chapter 1.5) of the matches.

Fetrow and Skolnick (1998) also defined structural templates for active sites. Their templates (fuzzy functional forms or FFFs) contain information about the residue type, Cα atom distances, residue conformations (e.g. *cis* versus *trans*-proline) and the sequence context. They found matches to FFFs from glutaredoxin/thioredoxin and ribonuclease active sites in experimentally-determined and predicted protein structures. However, the different nature of information used made the manual template definition difficult and often inapplicable and prevented a score that would allow ranking of the matches.

Russell (1998) uses a recursive depth-first search to find residues in similar spatial arrangements as assessed by similar pairwise distances between Cα, Cβ and functional atoms of the residues. Matches are evaluated by a weighted RMSD that renders the contribution of each amino acid independent of the number of atoms used for fitting. For each pattern size (i.e. number of residues or atoms), the RMSD distribution of random patterns is calculated to assess the statistical significance of the matches in the form of a P-value (see Chapter 1.5). The recurrence of several known side-chain patterns was analyzed and an all-against-all search with conserved hydrophilic residues found several significant similarities in proteins from different folds.

SPASM and RIGOR (Kleywegt, 1999) are essentially based on the same algorithm and were designed to find occurrences of a small pattern in proteins or to scan a protein against a collection of small patterns, respectively. The programs allow similar amino acids to substitute for each other and purely geometrical searches considering only Cα atoms are

possible. Matches are evaluated by the RMSD as in (Russell, 1998), however no measure of statistical significance is given.

During the course of my work, several new methods were developed that search for recurrences of predefined patterns or templates in protein structures. One method is restricted to three-residue patterns, which allows the representation of inter-atomic distances as vectors that can be efficiently searched using multidimensional index-trees (Hamelryck, 2003). Barker and Thornton (2003) use a tree-based backtracking algorithm similar to other methods (e.g. Kleywegt, 1999; Russell, 1998) to search for patterns of different sizes. Both methods evaluate matches by RMSD and provide measures of statistical significance that are estimated by fitting empirical background distributions (similar to Russell, 1998). Jambon and co-workers represent protein structures by stereochemical groups independent of amino acids. Recurrences of at least three of these groups are scored by the RMSD and the differences in local atom density (Jambon *et al.*, 2003). Another method searches for cavities with similar arrangements of functional groups to detect ligand binding sites and scores the overall surface overlap of matching cavities (Schmitt *et al.*, 2002). However, neither method assesses the significance of the matches.

To detect novel active sites, Wangikar *et al.* (2003) searched for local patterns shared between members of one family or superfamily, which merely combines sequence conservation with the requirement for spatial proximity and shape similarity. All matches that satisfied distance and RMSD requirements were reported but were not associated with a measure of significance. This work was extended to find patterns characteristic for groups of proteins and to classify proteins according to these spatial fingerprints (Tendulkar *et al.*, 2003).

## *1.5 Statistics for Sequence and Structure Comparison*

If two entities are compared, similarity is expressed using a score with a value and an entity. For everyday life comparisons, we immediately understand the meaning of this score and judge whether a similarity is significant. However, having only the score is insufficient if the nature of the compared entities is not known and the same score can have very different meanings. A price difference of one Euro for example is high when considering a packet of cigarettes but is negligible for a computer or a car.

The meaning of scores for protein sequence or structure comparisons is also highly dependent on the proteins under consideration. Whether a particular sequence identity (or indeed any measure of sequence similarity) is meaningful greatly depends on sequence length (see above and Sander & Schneider, 1991). A similar effect is seen when comparing protein structures using RMSD, which measures the difference between two sets of atoms. Two proteins with an RMSD of 2 Angstroms (Å) over 150 Cα atoms are normally homologous, while the same value observed between two Asp-His-Ser patterns can easily occur by chance (e.g. Russell, 1998). The number of matches to active site templates are dependent on the number and type of atoms in the template (see above and Wallace *et al.*, 1997).

*Statistical significance* directly addresses this problem by assessing whether a particular similarity is different from a random similarity between unrelated proteins or whether it could have arisen by chance. This is especially important for searches in large databases that typically produce many matches and require separation of biologically relevant matches from noise (i.e. random matches, Vingron, 2001). As scores cannot be compared for different searches and do not allow a reliable assessment of significance, specific measures for statistical significance have been developed. Z-scores, P- and E-values are the most commonly used in bioinformatics.

The Z-score is the number of standard deviations σ above the mean $\bar{x}$ of a distribution:

$$Z = \frac{x - \bar{x}}{\sigma(x)}$$

To assess the significance of search results, Z-scores are used to normalize observed similarities to the average of a background distribution (i.e. consisting of random matches) and are thus a measure of non-randomness. An average random similarity would score Z=0, and better similarities have positive Z-scores. For normal distributions, Z-scores are directly related to the number of expected random matches. For example, 50% of random matches are better than Z=0, 5% are better than Z=2, and 0.3% achieve Z=3 or better.

However, most scores in sequence or structure comparison are not distributed normally: Because they are calculated from optimal rather than random pairwise alignments or superimpositions, their distribution is enriched in good (i.e. extreme) scores and is thus called

an *extreme value distribution* (EVD) (Gumbel, 1958). The EVD can be used to calculate the *expectation* value (E-value), i.e. the number of expected random matches with equally good or better scores. Highly significant matches have E-values close to 0, whereas matches with high E-values are insignificant.

For an alignment of two sequences with a similarity S, the E-value depends on the lengths *m* and *n* and on two parameters *K* and λ (Karlin & Altschul, 1990):

$$E = Kmne^{-\lambda S}$$

Many groups have worked on adjusting these parameters for searches involving biological sequences so that E-values can be calculated *a priori* without the need for fitting empirical background distributions (Altschul & Gish, 1996; Karlin & Altschul, 1990; Mott *et al.*, 1990; Pearson, 1998; Waterman & Vingron, 1994). However, it is important to note that they are not always valid and that E-values for matches with uncommon features (e.g. low complexity regions consisting of only a few types of amino acids) are typically overestimated.

A P-value is the probability [0 – 1] that at least one equally good or better score occurs by chance (e.g. Karlin & Altschul, 1990; Mott *et al.*, 1990). Highly significant matches will not occur by chance and thus have very small P-values (close to 0) whereas P-values close to 1 correspond to insignificant matches. Because the number of random matches with scores equal or better than a particular score follows a Poisson distribution, P-values can be calculated from E-values as:

$$P = 1 - e^{-E}$$

Many search programs report E-values as they allow an easier comparison of insignificant similarities: for example, E-values of 5 and 10 correspond to P-values of 0.993 and 0.99995. However, for E < 0.01, P- and E-values are nearly identical. For database searches, statistical significance not only depends on the query but also on the size of the database that is searched: in huge databases, the number of individual comparisons (trials) is sufficiently high for any pattern to be found by chance. This is equivalent to lotteries: there are lottery millionaires each week although the success rate for an individual participant is close to nothing. As the chance of getting random matches increase with that database size (i.e. the

number of entries in the database), search programs also appropriately adjust the E- and P-values.

Today, measures of statistical significance are an integral part of all commonly used sequence or structure alignment methods. For example, BLAST, Fasta and HMMer report E-values, Vast reports P-values and Dali uses Z-scores for all matches. I used all three measures for protein active site comparison (Chapters 3.1 – 3.4) and miRNA target prediction (3.5).

The statistical significance of RMSD has been considered previously when comparing continuous protein backbone segments (or even entire structures), either during structural alignment (e.g. Levitt & Gerstein, 1998) or assessment of prediction quality (e.g. Cohen & Sternberg, 1980; Reva *et al.*, 1998). During these studies, significance was estimated by comparison with various background distributions derived from real or artificial proteins. Other methods to evaluate overall structural similarity often lack statistical evaluation and results are therefore difficult to compare or often even interpret (for an overview see Cristobal *et al.*, 2001). However, when I started my thesis, only one method provided empirically derived P-values to assess the significance for structural comparison of active sites (Russell, 1998), and the methods developed since also rely on fitting empirical background distributions (Barker & Thornton, 2003; Hamelryck, 2003). The relationship between pattern size, amino acid composition and the statistical parameters remained unknown and P- or E-values could not be calculated *a priori*. In addition, the statistics reported by Russell (1998) sometimes overestimated significance.

## *1.6 Scope: PINTS–Patterns in Non-homologous Tertiary Structures*

Although active site comparison has the potential to directly detect functional similarities and aid in functional annotation of protein structures, structural biologists seldom use the methods described above. Indeed, whereas new protein structures are routinely compared to others by structural alignment (often with Dali), results from active site comparison are almost never mentioned. This may be due to three main reasons: the lack of a measure for statistical significance that would allow inexperienced users to interpret search results, the non-existence of a large up-to-date database of functional patterns and the absence of an easy-to-use Internet service such as Dali, to perform comparisons. Only PROCAT, the method by Fischer *et al.* (1994), and SPASM are available on the Internet. PROCAT allows searches against a small

number of templates, SPASM only detects recurrences of user-defined patterns in protein structures (Madsen & Kleywegt, 2002), and Fischer's method lacks specificity as described above.

For my thesis, I developed a method (PINTS; Patterns in Non-homologous Tertiary Structures) to compare local structural patterns typical of active sites. I used PINTS to derive a statistical model for such similarities, that allows the significance to be estimated for any local similarity with a particular RMSD *a priori,* without requiring a fit to background data (Chapter 3.1). I then assessed the potential of active site comparison for functional annotation of proteins on a large number of structures solved by structural genomics projects (3.2) and during a detailed case study for an archaeal fructose-bisphosphate aldolase (3.3). Finally, I built a user-friendly Internet server for PINTS that includes several pattern databases (3.4).

## *1.7 MicroRNAs: A Novel Class of Genes*

A second major part of my thesis was the prediction of microRNA (miRNA) function. miRNAs are a class of 21-22 nucleotide non-protein-coding RNAs. They are excised from longer precursor transcripts that fold locally into 70-100 nucleotide-long hairpin-like structures. They are found in all higher eukaryotes and are thought to play major regulatory roles in post-transcriptional gene regulation. Although the first miRNA was identified more than ten years ago, the general abundance and importance of miRNAs has been discovered only during the last three years. They represent a novel class of genes and add a new level of regulatory complexity to gene expression.

### 1.7.1 miRNAs regulate post-transcriptional gene expression

The first miRNAs (*lin-4* and *let-7*) were identified in the nematode *Caenorhabditis elegans* by their mutant phenotypes in 1993 and 2000, respectively (Lee *et al.*, 1993; Reinhart *et al.*, 2000; Wightman *et al.*, 1993). Because of their temporally regulated expression, they were originally called *small temporal RNAs* or *stRNAs*. Genetic interactions suggested that both miRNAs negatively regulate protein-coding genes (*target genes*) involved in developmental timing. Inspection of the target messenger RNA (mRNA) sequences revealed sites complementary to the miRNAs within the 3' untranslated region (UTR). Reporter gene assays showed that these sites were sufficient to infer miRNA-dependent regulation, supporting a

direct mechanism that acts through a miRNA-mRNA duplex. However, at that time, it was not clear whether miRNAs were peculiarities of *C. elegans* development or if they existed beyond nematodes.

### 1.7.2 General Importance of miRNAs

The picture changed dramatically when the Ruvkun lab reported that *let-7* was found in a wide range of animals (e.g. *C. elegans*, *Drosophila*, and human) and showed that its temporal regulation and the complementary sites in known targets were conserved (Pasquinelli *et al.*, 2000). Since then, hundreds of plant and animal miRNAs have been isolated and sequenced in systematic large-scale studies (Lagos-Quintana *et al.*, 2001; Lau *et al.*, 2001; Lee & Ambros, 2001; Reinhart *et al.*, 2002). Combining this data with computational cross-genome comparison predicts 100-120 miRNA genes in *C. elegans* and *Drosophila* and about 250 in mouse and human, and miRNAs are thought to comprise about 1% of all genes in each of these species (Ambros *et al.*, 2003; Grad *et al.*, 2003; Lai *et al.*, 2003; Lim *et al.*, 2003a; Lim *et al.*, 2003b). Their abundance and the evolutionary sequence conservation of many miRNAs suggest that they have ancient and important biological functions.

### 1.7.3 miRNA Biogenesis and Function

Post-transcriptional regulation of gene expression by RNA also occurs in RNA interference (RNAi) that is known for most eukaryotes. RNAi and translational inhibition by miRNAs are both mediated by short RNAs of similar length and both pathways are now known to share core components (see Figure 1.2 for an overview). In RNAi, double stranded RNA (dsRNA) causes a rapid and sequence-specific depletion of the corresponding mRNA by endonucleolytic cleavage. According to the current model, the RNAse III endonuclease *Dicer* cuts the long dsRNA into pieces of about 22 nucleotides. These *siRNA* (*small interfering RNA*) duplexes have a 2-3 nucleotide 3' overhang characteristic of RNAse III cleavage products and can mediate RNAi when introduced into mammalian cells (Elbashir *et al.*, 2001). The two strands are separated and the siRNAs are incorporated into the RNA-induced silencing complex (RISC). RISC then specifically recognizes target mRNAs that are complementary to the siRNA template and cleaves them between residue 10 and 11 from the 5' end of the siRNA. RISC has been purified from insect and mammalian cells, but most of its components and their functions are still unknown. One core component, apart from the

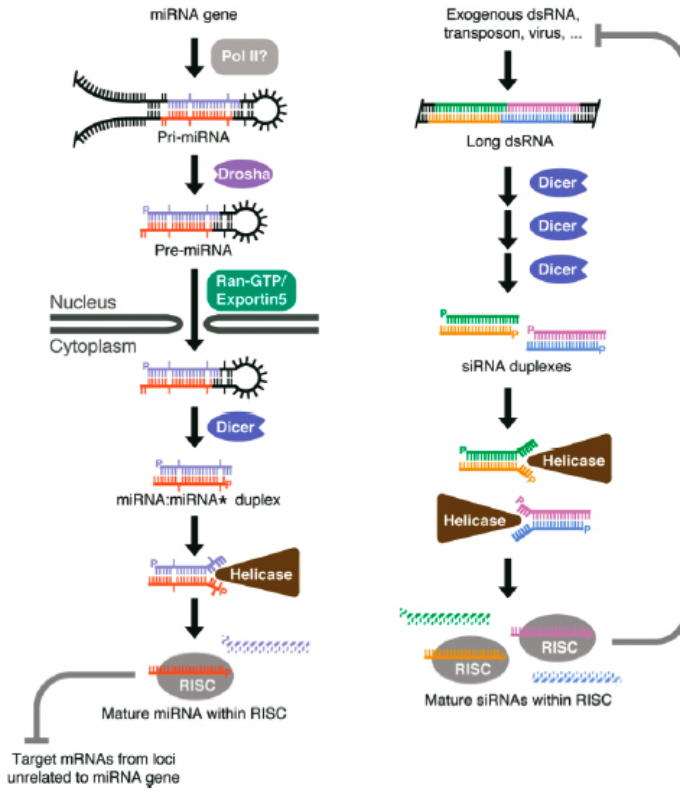



**Figure 1.2:** Biogenesis of miRNAs (left) and siRNAs (right) after (Bartel, 2004). See text for details.

uncharacterized RNAse III responsible for target cleavage, is made up of the highly basic Argonaute proteins that can bind single and double stranded RNA by a conserved PAZ-domain and might be involved in RNA incorporation and/or target recognition (Lingel *et al.*, 2003; Song *et al.*, 2003; Yan *et al.*, 2003).

As siRNAs always target the gene they are derived from, RNAi probably evolved as a potent defense mechanism rather than for general regulation of gene expression. Its involvement in viral defense and silencing of endogenous parasitic elements such as transposons is supported by several findings: Certain viruses for example express viral suppressors of gene silencing that specifically target the RNAi machinery and mutations in these genes can be rescued by inhibiting the RNAi pathway (Kasschau *et al.*, 2003). Some *C. elegans* strains have increased mutation rates caused by defects in the RNAi machinery that lead to unusually high transposon activity (mutator-strains; Ketting *et al.*, 1999; Tabara *et al.*, 1999). Some recent evidence establishes a link between siRNAs in RNAi and transcriptional silencing or chromatin maintenance (Hall *et al.*, 2002; Wassenegger *et al.*, 1994).

In contrast to siRNAs, miRNAs are genes themselves and not derived from transcripts of other genes. miRNAs act in *trans* on sequences different from their origin and are thought to regulate or fine-tune gene expression. Nevertheless, the biogenesis that is common to all miRNAs has some similarities to the RNAi pathway such as the involvement of *Dicer* and RISC.

According to the current model, miRNAs are transcribed as long primary transcripts (pre-miRNAs) that can contain several miRNAs. The pre-miRNAs fold locally into a 60-70

nucleotide stemloop or hairpin structure that contains the miRNA in one arm. In the nucleus, the RNAse III *Drosha* liberates this stemloop precursor (pre-miRNA) with a staggered cut that defines one end of the miRNA (Lee *et al.*, 2003). The pre-miRNAs are then exported from the nucleus by Ran-GTP and the export receptor Exportin-5 (Lund *et al.*, 2004; Yi *et al.*, 2003). In the cytosol, *Dicer* cuts an RNA:RNA duplex from the pre-miRNA that consists of the miRNA paired to the opposing RNA fragment in the stemloop. These fragments, that are called miRNA*, occur with much lower frequencies than the miRNA itself, which is preferentially incorporated into RISC. This asymmetry was recently explained by the relative stability of the 5' ends of the two opposing RNAs: typically, the strand with the less stable (i.e. more loose) 5' end is preferably incorporated into RISC and thus found as the miRNA (Khvorova *et al.*, 2003; Schwarz *et al.*, 2003).

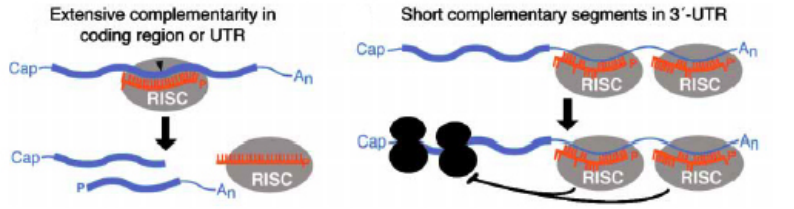



**Figure 1.3:** Mechanism of mRNA-cleavage in RNAi and miRNA mediated translational repression. After (Bartel, 2004).

Similar to RNAi, the miRNA in RISC is used as a template to recognize complementary sites in the target mRNA. However, depending on the degree of miRNA target complementarity, two different modes of miRNA-directed target inhibition have been demonstrated (see Figure 1.3): Target RNAs containing sequences with perfect or near-perfect complements of the miRNA are cleaved by RISC similar to RNAi (Hutvagner & Zamore, 2002; Martinez *et al.*, 2002). Endogenous plant miRNAs have been shown to regulate target RNAs by RNAi involving perfect or near-perfect target site complementarity (Kasschau *et al.*, 2003; Llave *et al.*, 2002; Martinez *et al.*, 2002; Palatnik *et al.*, 2003; Tang *et al.*, 2003; Xie *et al.*, 2003). In contrast, all animal miRNAs tested until now pair imperfectly with their targets and systematic analysis has confirmed the absence of targets with perfect or near-perfect sequence complementarity for all *C. elegans* miRNAs (Ambros *et al.*, 2003). The mismatches, bulges and loops are thought to prevent cleavage of the target but instead inhibit translation, leading to reduced protein levels without affecting the mRNA of the target protein (Brennecke *et al.*, 2003; Doench *et al.*, 2003; Lee *et al.*, 1993; Reinhart *et al.*, 2000; Zeng *et al.*, 2002). Interestingly, the loading of target mRNA with ribosomes does not change during translational inhibition (Olsen & Ambros, 1999) and many miRNAs have recently been found to be associated with polysomes, suggesting a mechanism where ribosomes are stalled on the mRNA (Kim *et al.*, 2004).

The two different outcomes seem to be independent of the miRNA itself as the same small RNA can cause degradation of its target mRNA or block its translation solely depending on the degree of miRNA target sequence complementary (Doench *et al.*, 2003; Hutvagner & Zamore, 2002). It is currently speculated that different types of RISCs may be involved in RNAi, miRNA mediated target cleavage or translational inhibition. For example there are several different Argonaute proteins that exhibit a tendency towards siRNA (*dAgo1*) or miRNA (*dAgo2*) substrates, respectively (Caudy *et al.*, 2002). In addition, some of the unknown core components or only transiently associated proteins are likely to differ between the RISC complexes in different mechanisms.

### 1.7.4 Assigning Function to miRNAs

Although more than 700 miRNAs have been deposited in central databases (Griffiths-Jones *et al.*, 2003) and their abundance and conservation suggest highly important functions, the assignment of these and the identification of target genes lag far behind. For some miRNAs, expression profiles suggest an involvement in organ or tissue development. Mouse *miR-290* and *miR-295* are for example expressed in embryonic stem cells but not differentiated cells, whereas *miR-1* is preferentially expressed in the mammalian heart and *miR-122* in the liver (Lagos-Quintana *et al.*, 2002). For other miRNAs, their genomic organization indicates interesting regulatory connections and networks. About one quarter of human miRNAs are located in introns of protein-coding genes leading to co-expression of protein and miRNA, potentially allowing antagonistic regulation of genes or pathways (e.g. *miR-7* is located in an intron of *hnRNP K*; Aravin *et al.*, 2003). Others are clustered in the genome and probably co-transcribed as multi-cistronic transcripts, suggesting broad effects on gene expression.

However, prior to this thesis, specific functions were known only for four animal miRNAs that are required for correct timing of developmental events (*lin-4* and *let-7*), to regulate apoptosis and cell proliferation (*bantam*), or to prevent cell-death and take part in fat metabolism (*miR-14*) (Brennecke *et al.*, 2003; Lee *et al.*, 1993; Moss *et al.*, 1997; Reinhart *et al.*, 2000; Wightman *et al.*, 1993; Xu *et al.*, 2003). Direct targets were experimentally validated for only three of them: the *lin-4* targets *lin-14* (Wightman *et al.*, 1993) and *lin-28* (Moss *et al.*, 1997), the *let-7* targets *lin-41* (Reinhart *et al.*, 2000) and *lin-57/hbl-1* (Abrahante *et al.*, 2003; Lin *et al.*, 2003) and the *bantam* target *hid* (see below and Brennecke *et al.*, 2003).

### 1.7.4.1 Genetic Approaches

Traditional genetic *loss-of-function* screens were critical to determine the functions of *lin-4* and *let-7* and subsequently their target genes. They remained however restricted to these examples, probably because the small miRNAs are difficult to target for mutagenesis or because a clear *loss-of-function* phenotype is prevented by functional redundancy that is expected for some groups of miRNAs with similar sequences. Overexpression of miRNAs (or the corresponding genomic region) in *gain-of-function* screens does not suffer from these limitations and was successfully used to identify *bantam*. However, visual assessment of mutants might still miss certain phenotypes that are less severe or fall outside the range of interest. In addition, even when a specific phenotype is observed, the target gene itself often remains elusive: Although *Drosophila miR-14* has a clear anti-apoptotic effect, no direct target gene has yet been identified (Xu *et al.*, 2003).

### 1.7.4.2 Computational Approaches

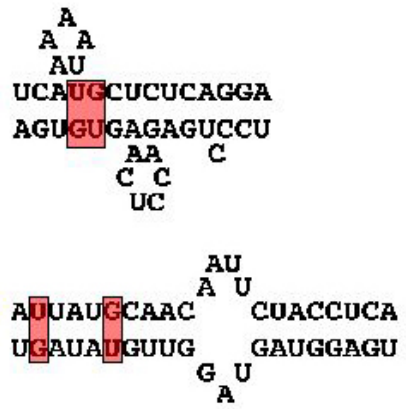


**Figure 1.4:** miRNA target complexes for lin-4/lin-14 (top) and let-7/lin-14 (bottom) as proposed by (Banerjee & Slack, 2002). miRNAs are the bottom strands and G:U base paris are highlighted.

The increasing number of known miRNA sequences and large databases of genome and transcript sequences suggests more direct approaches for target discovery, e.g. experimental tests of candidates resulting from computational screens (Ambros, 2001). Given that miRNAs interact with their target through sequence complementarity, the prediction of putative target genes from miRNA sequence alone seems feasible and very promising. Simple sequence searches indeed revealed perfectly complementary sites in putative target genes for plant miRNAs (Llave *et al.*, 2002; Rhoades *et al.*, 2002). However, for the known miRNAs in animals no target sites with perfect or near-perfect sequence complementarity could be found (Ambros *et al.*, 2003; Rhoades *et al.*, 2002). The RNA:RNA duplexes for the known targets are discontinuous and contain mismatches, gaps and G:U base pairs at different positions. Even allowing for G:U base pairs, the longest contiguous alignments in these examples range from 8-10 nucleotides. Such limited information content makes it difficult to identify targets within whole genome or transcriptome databases, since

standard alignment methods produce many false positives with such short variable sequences. Furthermore, the small number of validated examples makes the development and benchmarking of a generally applicable computational method problematic at present.

### 1.7.5 Scope: A Screen for miRNA Targets in *Drosophila*

For my thesis, I developed a method to screen for miRNA targets in *Drosophila* that combines a lenient sequence search with an RNA secondary structure prediction algorithm. It identifies all of the previously known miRNA targets and successfully predicts new targets, some of which were validated experimentally. The *bantam* target *hid* was identified with a preliminary version of the screen described below (Chapter 3.6.2, Brennecke *et al.*, 2003).

# 2 Materials and Methods

## 2.1 General Equipment

For all studies, I used personal computers with Intel Pentium processors and the SuSe Linux operating system. I typically automated all sequence or structural comparisons and data collection with scripts in the Perl programming language.

## 2.2 Programs

For sequence comparisons, I used BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997) or HMMer (Eddy, 1998) as indicated. Multiple sequence alignments were created with ClustalW (Thompson *et al.*, 1994) and usually edited manually with Jalview (Clamp *et al.*, 2004). We used STAMP (Russell & Barton, 1992) for all structural alignments or searches for similar folds in local databases. PINTS (described below and in Stark & Russell, 2003; Stark *et al.*, 2003b) was used for all comparisons involving active sites or other local residue patterns. For visual inspection protein structures or matching patterns were displayed with RasMol (Sayle & Milner-White, 1995). Images were created using Molscript (Kraulis, 1991) and Raster3D (Merritt & Murphy, 1994). For colourful displays of multiple sequence alignments, I used Alscript (Barton, 1993).

## 2.3 The PINTS Program

For the structural comparison of biological molecules, I developed the program PINTS (Patterns of Non-homologous Tertiary Structures). It finds all possible patterns of residues (or atoms or points defined by other criteria) common to two sets of coordinates. For protein structure comparison, I used PINTS to compare residue patterns (e.g. active sites) to proteins, proteins to collections (databases) of such patterns, or two protein structures against each other. Comparisons can take single atoms such as Cαs or several atoms per residue into account. PINTS uses the common PDB format without the need for pre-computation and can also read residue accessibility to restrict searches to the protein surface. It is implemented in the C programming language to achieve high search performance. Below, I discuss the basic data-type used (*points*), the input format, the search algorithm, the RMSD score, E-value and the output formats.

### 2.3.1 Representing Structure Data: Points

Functionally important patterns such as protein active sites can occur in any sequence order, which prevents the use of alignment methods and requires treating the structure as *independent* (or *unconnected*) *points* in space (see Introduction). Typically, the chemical nature and physicochemical properties of the atoms or residues are important for function. In protein active sites for example, different types of amino acids are not functionally equivalent and cannot freely replace each other. PINTS thus represents all macromolecules and patterns internally with points consisting of 3D coordinates and a *type* (i.e. an integer value) that summarizes their properties. All points with the same type are regarded as equivalent and are allowed to substitute for one another. The core of the program is thus independent of the concepts of atom or amino-acid residue and can handle any kind of structural data (e.g. grid points with physicochemical properties, atoms or functional units of drug-like molecules, etc.).

### 2.3.2 Data Input

**A**

```
A CA 0
C CA 0
D CA 0
E CA 0
F CA 0
G CA 0
H CA 0
I CA 0
K CA 0
L CA 0
M CA 0
N CA 0
P CA 0
Q CA 0
R CA 0
S CA 0
T CA 0
V CA 0
W CA 0
Y CA 0
```

**B**

```
A CA  0 50
C CA  1  0
D CA  2  0
E CA  3  0
F CA  4 50
G CA  5 50
H CA  6  0
I CA  7 50
K CA  8  0
L CA  9 50
M CA 10 50
N CA 11  0
P CA 12 50
Q CA 13  0
R CA 14  0
S CA 15  0
T CA 16  0
V CA 17 50
W CA 18  0
Y CA 19  0
```

**C**

```
E (OE1,CD,OE2|F) 0
D (OD1,CG,OD2|F) 0
S (OG,CB)        1
Y (CD1,OH,CD2|F) 2
T (OG1,CB,CG2)   3
R (NH1,NE,NH2|F) 4
K (CD,NZ)        5
C (SG,CB)        6
F (CE1,CG,CE2|F) 7
W (CD1,CE3,CZ2)  8
H (CG,ND1,NE2)   9
N (OD1,ND2,CB)  10
Q (OE1,NE2,CG)  10
G CA            11
```

**D**

```
* N 0
* C 1
* O 2
```

**Table 2.1**: Examples of *Definition files*. A. All amino acids have type 0 (purely geometrical matching of Cα-atoms). B. All amino acids are distinguished (types 0 – 19), and non-polar residues are required to be at least 50% solvent accessible. C. Different side-chain atoms are considered for each amino acid. Alternative matching is allowed for carboxyl-, guanidino-, and aromatic ring atoms (Syntax: (OE1,CD,OE2|F) means 'allow flipping about CD'. (OE1,CD,OE2;OE2,CD,OE21) is equivalent). D. Atoms (N, C, O) are considered irrespective of the residue or ligand they belong to.

Structural data is read and translated into points by input modules. The module currently implemented reads coordinate files in PDB text format and atom or residue accessibility (i.e. surface exposure) from NACCESS (Lee & Richards, 1971) or DSSP (Kabsch & Sander, 1983a) files. PINTS can load and compare all types of residues or atoms (ATOM or HETATM section of PDB -files).

*Definition files* (see Table 2.1) specify how the coordinates in the coordinate file should be translated into points and which residues or atoms should be considered. The files consist of three

mandatory and two optional columns separated by spaces (blanks). The first column specifies the residue name (corresponding to columns 18 – 20 of the PDB file). Standard amino acids can be specified in three- or one-letter code (e.g. ALA or A for alanine, etc.) and a wildcard '*' can be used. The second column defines the individual atom(s) for each residue that should be considered (PDB file columns 13 – 16). When several atoms per residue are specified, the average coordinate is used during the search but all atoms can be used for the RMSD calculation (see below). The PDB format requires a unique labelling of all atoms for each residue, even in cases where the atoms are equivalent (e.g. in carboxyl- and guanidino-groups or the ring-carbons in phenylalanine and tyrosine) or cannot be distinguished easily in the electron density (e.g. amides); PINTS allows alternative matching for these atoms when specified (see legend to Table 2.1). Column three specifies the type of the points. Types are integer numbers and all points with identical numbers can be superimposed. Columns four and five specify the minimal relative or absolute solvent exposure of a residue to be considered according to DSSP or NACCESS, that can be used to compare solvent-accessible surface patterns only.

### 2.3.3 The Search Algorithm

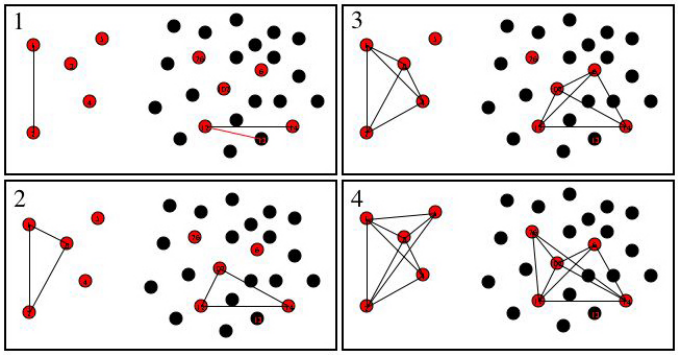



**Figure 2.1:** The PINTS search algorithm. A pattern (left) is found within a larger structure (right) by extending smaller patterns (steps 1-4). New points are only added when the distances to other points are similar (black lines) and not otherwise (e.g. red line in (1)).

The search must detect all *matches*, i.e. patterns common to two structures. Matches can either be *complete* (i.e. one structure matches completely within the other) or *partial* when only a part of one structure is similar to a part of the second. To achieve this, essentially all possible combinations of points similar to the two sets are generated recursively. As a measure of similarity, the distances between the points in set (1) and those between points in set (2) are compared and required to differ by no more than a given tolerance cutoff.

The algorithm (*recursive depth-first search*) is based on the fact that all matches of a given size (i.e. consisting of N points) contain smaller matches with N-1 points and can thus be

found as extensions of those (see Figure 2.1). It starts by equivalencing one point from set (1) with each point in set (2), given that their types match. Then, the match is recursively extended by checking, for each of the remaining (i.e. unmatched) points in set (1), whether points in set (2) can be found with similar distances to all previously found points. Comparing the distances prevents the need for superimpositions at each step but cannot distinguish mirror images. A chirality-check is thus performed once four points are found to discard matches with wrong handedness. The sensitivity of the search can be altered by changing the distance tolerance cutoff (-*mt* option).

### 2.3.4 The RMSD Score and Statistical Significance

Once similar patterns are found, their similarity or quality is measured with the RMSD after optimal superimposition by least-squares fitting (McLachlan, 1979), which also provides the transformations needed for optimal superimposition. The RMSD can be calculated on the points (i.e. the average coordinates when several atoms are specified) or on all atoms of a given residue. The user can choose whether all atoms or all residues should be given equal weight (-*rf* or -*wf* option).

As the RMSD value that implies a meaningful similarity varies between searches, PINTS provides E-values (see Chapter 3.1 and Stark *et al.*, 2003b) similar to those used in sequence searches that assess the probability that the obtained matches occurred just by chance without further functional implications (e.g. Altschul & Gish, 1996; Karlin & Altschul, 1990). It is this feature that sets PINTS most apart from previous methods (Artymiuk *et al.*, 1994; Fetrow & Skolnick, 1998; Fischer *et al.*, 1994; Kleywegt, 1999; Russell, 1998; Wallace *et al.*, 1997; Wallace *et al.*, 1996) or servers (Madsen & Kleywegt, 2002) that perform such searches.

### 2.3.5 Output formats

Matches are reported in three different output formats (-*of* option) that are readable (*short*, 0), can serve as input of programs from the STAMP package (*stamp*, 1), or can be easily parsed as one line per match is printed (*line*, 3). In addition, PINTS can directly print the superimposed coordinates in PDB format (*pdb*, 2) for visual inspection of the matches.

### 2.3.6 Other Parameters or Command Line Options

The user can change the behaviour and many parameters of PINTS by command-line options. An overview of all parameters is given in Table 2.2.

| **Modes** | | **Default** |
|---|---|---|
| -r | Residue-based search | + |
| -a | Atom-based search | - |
| **Input** | | |
| -d <file> | Domain-file (STAMP-package) | |
| -db <file> | PINTS-Database | |
| -ex <file> | Exclude Database entries | |
| -c <file> | Coordinate File (PDB- or NACCESS-format) | |
| -def <file> | Definition File (PINTS) | |
| **Output** | | |
| -o <file> | Redirect output to file | - |
| -of <string> | Output format (0,1,2,3) | 0 |
| -gfa <float> | Combine matches with x shared atoms | - |
| -gfp <float> | Combine matches with x % shared atoms | - |
| -rb <integer> | Report only x best matches | - |
| -rbe <integer> | Report only x best matches per DB entry | - |
| -vp | Report Permutations | - |
| -vs | Report Subpatterns | - |
| **Search** | | |
| -ma <integer> | Minimum number of points required for match | 5 / 10 |
| -mt <float> | Maximum distance tolerance allowed during search | 3 / 1 |
| -ms <float> | Maximum pattern diameter | 15 |
| -md <integer> | Maximum search depth | 20 |
| -mr <float> | Maximum pattern RMSD | 3 / 1 |
| -me <float> | Maximum E-value | 10 |
| -mes <float> | Maximum standard E-value (E700) | - |
| -mh <integer> | Maximum number of matches during search | - |
| -all [<float>] | Pattern has to be found entirely or to x % | - |
| **Special** | | |
| -wf | Weighted RMSD (equal weight for all residues) | - |
| -rf | RMSD calculation on multiple atoms per residue | - |
| -rs [<int>] | Minimum relative surface accessibility (x%) | - |
| -as [<int>] | Minimum absolute surface accessibility | - |
| -v | Verbose | - |
| -ha | Load hetero-atoms | - |
| -pf <float> | Factorial Parameter for Statistics | - |
| -ef <float> | Exponential Parameter for Statistics | - |
| -lr <integer> | Long-range filter (sequential separation of x residues required) | - |
| -linear | Requires same sequential order | - |
| -aaa | All-against-all comparison | - |
| -help | Prints Help | |

**Table 2.2**: Command line options for PINTS and default values (+/- for on/off).

## *2.4 PINTS Databases*

### 2.4.1 PINTS Databases of Functional Patterns

For functional annotation of protein structures, a database of functionally relevant side-chain patterns would be ideal (Thornton *et al.*, 2000; Wallace *et al.*, 1997), though no sufficiently complete database is currently available. We thus decided to automatically collect residue patterns likely to be of functional importance. Although manual curation is more accurate, automation allows us to cover all structures and to update our databases constantly. The

*ligand binding sites* database, contains residues that have at least one atom within 3.0 Å of a HETATM entry (excluding waters), and the *SITE annotations* database those defined by structural biologists to form a functional site (SITE entries in PDB files). We update the databases on a weekly basis with every PDB release. Currently, the *ligand binding sites* and *SITE annotations* database contain 15200 and 7500 patterns respectively (21.03.2004). I noticed, that many patterns include residues not directly related to or required for function. The databases can thus be seen as reducing the search space to potentially interesting parts of proteins and we allow for partial matches where only part of the database entry is found in the query. Recently, the Thornton group at the European Bioinformatics Institute (EBI) in Hinxton published a large set of manually annotated catalytic sites (Bartlett *et al.*, 2002; Porter *et al.*, 2004) and made the data available to be searched with PINTS.

### 2.4.2 Non-Redundant Databases of Protein Structures

For some proteins, the PDB contains many structures that are nearly identical (e.g. lysozyme, myoglobin). For searches, it is crucial to avoid such redundancy and consider only representatives for groups of similar structures. I thus created databases of entire protein structures for different levels of redundancy by collecting one representative of each SCOP (Murzin *et al.*, 1995a) fold, superfamily, family or protein but also used a representative set of protein structures suggested by PDB-select (Hobohm & Sander, 1994b).

## *2.5 The Development of a Statistical Model*

### 2.5.1 Background database

To avoid any bias of the parameters, background databases (BDs) must be non-redundant, i.e. consist of database entries that are unrelated and do not share functional or structural similarities. I thus used one member of each of the 723 folds in SCOP version 1.55 as a BD for all searches to determine the statistical parameters.

### 2.5.2 Parameter Determination

To determine the parameters A, B, C, and D in our statistical model, I compared random patterns of two to eight residues to the BD and fit the number of matches with RMSD≤RM to the function $AR_M^B$ (least squares). Each data point was calculated as the average of 9

independent searches with three patterns from each of three structures that were not in the BD (1a6m, 4rhv, 5p21) For the independence model, I represented residues by their Cα atoms and for the dependence model with one, two or three atoms (e.g. Cα, Cγ1 and Cγ2) as indicated.

### 2.5.3 Cumulative Distribution of P-values

I randomly split the BD into 10 test databases and searched 6 of the patterns used above against these databases and calculated the P-value of the best match with a corrected parameter $A_{1/10}={}^{1}/_{10}A$. The cumulative distribution of these is plotted and compared to a linear function with the slope of 1.

### 2.5.4 Search with the Trypsin Catalytic Triad

I searched the BD and all structures in the PDB with the catalytic residues from trypsin (1mct: His 57 Cβ, Nδ1, Nε2; Asp 102 Oδ1 Oδ2 Cγ; Ser 195 Cβ Oγ) and derived P-values derived by fitting the cumulative distribution of of matches against the BD.

### 2.5.5 Comparing Proteins to Pattern Databases

I compared the unliganded structure of *Trypanosoma cruzi* PEPCK (1ii2) to the *ligand binding sites* database and the structure of LuxS (1j98) to the *SITE annotations* database using the default settings of PINTS that consider multiple side-chain atoms per residue (Table 2.1, Column C), disregard hydrophobic residues and require the pattern diameter to be within 12Å.

### 2.5.6 Over-represented Patterns

I collected one representative structure from each fold in SCOP (classes 1–8, version 1.61) and compared all 706 structures to each other with PINTS requiring an E-value ≤ 10 and restricted the pattern diameter to 4 (5, 6, 7, or 8) Å depending on whether the pattern contained 2 (3, 4, 5, or 6) residues. I ignored sequential matches by requiring at least two residues per pattern to be at least 5 residues apart. I clustered all matches to identical patterns into groups by recursive single-linkage clustering and kept only those groups with at least 10 patterns. To determine the abundance of each pattern without the constraints required above, I searched the same database for recurrences of each pattern and calculated the number of matches for E ≤ 10, 1, or 0.1 and the average, median, and maximum number for all groups.

For the ranking of the patterns, I required the highest similarity in side-chain arrangements (E ≤ 0.1), but also inspected the lists for the other (less stringent) cutoffs.

## *2.6 Analysis of Structural Genomics Proteins*

We considered 254 structures labeled as structural genomics with release dates up to October 2003. The PDB accessions were:

2mjp, 1ufh, 1udk, 1ucr, 1uan, 1qwk, 1qu9, 1q8c, 1q53, 1q2y, 1pug, 1pt8, 1pt7, 1pt5, 1pqy, 1pm3, 1pgv, 1pav, 1p9v, 1p8c, 1p5f, 1p1m, 1p1l, 1oz9, 1oyz, 1oy1, 1ovq, 1otk, 1oru, 1oq1, 1ooj, 1ooe, 1on0, 1o5n, 1o5j, 1o5h, 1o54, 1o51, 1o50, 1o4w, 1o4t, 1o3u, 1o22, 1o1z, 1o1y, 1o13, 1o0u, 1o0i, 1nza, 1nyn, 1ny4, 1ny1, 1nxz, 1nxu, 1nxj, 1nxi, 1nx8, 1nx4, 1nwb, 1nvo, 1ns5, 1nri, 1nr9, 1nr3, 1nqn, 1nqm, 1nqk, 1npy, 1npd, 1nog, 1nnw, 1nn4, 1nkv, 1njk, 1njh, 1nij, 1nig, 1ni9, 1ni7, 1ng6, 1nf2, 1ne8, 1ne2, 1nc7, 1nc5, 1n91, 1n81, 1n6z, 1n1q, 1mzh, 1mzg, 1mw7, 1mtp, 1mog, 1mo0, 1ml8, 1mk4, 1mjh, 1m98, 1m94, 1m68, 1m65, 1m3s, 1m33, 1m25, 1m1s, 1m0s, 1ly7, 1lxn, 1lxj, 1lv3, 1lur, 1lql, 1lpl, 1lkn, 1ljo, 1lj7, 1lel, 1ldq, 1ldo, 1lcz, 1lcw, 1lcv, 1l7y, 1l7b, 1l7a, 1l6r, 1l5x, 1l1s, 1l0b, 1kyt, 1kyh, 1kuu, 1kut, 1ktn, 1ks2, 1kr4, 1kq4, 1kq3, 1kon, 1kkg, 1kk9, 1kjn, 1k8v, 1k8f, 1k7k, 1k7j, 1k77, 1k4n, 1k3r, 1k2e, 1k26, 1jzt, 1jyh, 1jyg, 1jx7, 1jw3, 1jsx, 1jsb, 1jru, 1jrm, 1jrk, 1jri, 1jov, 1jop, 1jog, 1jo0, 1jn1, 1je3, 1jdq, 1jcu, 1jbm, 1jbi, 1jay, 1jax, 1jal, 1j9l, 1j9k, 1j9j, 1j8c, 1j8b, 1j7h, 1j7d, 1j74, 1j6r, 1j6p, 1j6o, 1j5x, 1j5u, 1j5p, 1iyg, 1ixl, 1iw5, 1ivz, 1iv0, 1iuy, 1iur, 1iul, 1iuk, 1iuj, 1in0, 1ilv, 1ilo, 1ij8, 1iio, 1ihn, 1ie0, 1i9h, 1i8f, 1i81, 1i6n, 1i60, 1i36, 1i17, 1hy2, 1hxz, 1hxl, 1hu7, 1hu6, 1hu5, 1htw, 1hru, 1hqq, 1h2h, 1gtd, 1gh9, 1g9x, 1g6y, 1g6w, 1g6p, 1g2r, 1g04, 1fux, 1fl9, 1f89, 1f3o, 1exc, 1ex2, 1ew4, 1eo1, 1eiw, 1ehx, 1dus, 1dm9, 1dm5, 1di7, 1di6, 1dcj, 1dbx, 1dbu, 1d1r, 1ct5, 1b78, 1apa.

I compared all proteins to the sequences in the using BLAST. Proteins that matched any other sequence in the database with an E-value ≤ $10^{-10}$ were not considered further. Although this degree of sequence similarity is not always associated with a similarity in function (Rost, 2002; Tian & Skolnick, 2003; Todd *et al.*, 2001), our threshold ensures that the analysis excludes all cases where functional similarities are obvious from sequence comparison and is thus reliable in assessing the added value of structure comparison. Altering this threshold does not greatly affect the overall findings. I then used these structures to search for similarities in the *ligand-binding site* and *SITE annotations* databases of side-chain patterns using PINTS. For comparison, I also compared the structure to PDB representatives in the FSSP database using the Dali server (Holm & Sander, 1993) with default options.

### 2.6.1 Structural similarity thresholds

PINTS usually detects binding site similarities for chemically similar ligands with E-values between $10^{-4}$ - $10^{-2}$ whereas negative matches generally have E≥0.1 (see Chapter 3.1). We thus expect reliable functional clues to come from matches with E≤$10^{-3}$. However, I also

manually inspected the best matches for each structure. The accuracy of inferring a functional relationship for a given Dali Z-score is case-specific. For example, TIM-barrels can have different functions at comparatively high Z-score values (e.g. Z=18; Lorentzen *et al.*, 2003), while Rossman-type NAD-binding domains are reliably detected with values as low as Z=6. A general threshold above which function can be reliably assigned based on fold comparison is not possible (Liisa Holm, personal communication). However, earlier observations showed that fewer than 10% of functionally unrelated structures have values above 10 (Holm & Sander, 1997). We thus decided to use this as the threshold for our study. I report and discuss only the best matches for PINTS or Dali and only consider one representative for groups of structures sharing 90% sequence identity.

## *2.7 Analysis of Archaeal FBPA IA*

### 2.7.1 Structural Alignments

I compared the Tt FBPA monomer to representatives of all families with a TIM-barrel fold (according to SCOP) using STAMP, which also reports the number of structurally equivalent residues. I then calculated the percent sequence identity for structurally equivalent residues and the probability (P-value) that the percent identity arose independently (i.e. by chance, Murzin, 1993b).

### 2.7.2 Comparison of FBPA Active-Site

I compared the active site of FBPA to all proteins classified as TIM-barrels by SCOP (SCOP fold c.1.) and the FBPA structure against all TIM-barrel active-site patterns in the PINTS databases.

## *2.8 miRNA Target Prediction*

### 2.8.1 Accession numbers

**miRNAs:**

lin-4 NR_000799; let-7 NR_000938; bantam AJ550546, Rfam MI0000387; miR-2a RF00047, AJ421757; miR-4 AJ421762; miR-7 AJ421767; miR-9 AJ421769; miR-11 AJ421771 ; miR-13a AJ421773; miR-14 AJ421776; miR-277 RFAM MI0000360.

**Target genes:**

lin-14 NM_077516; lin-28 NM_059880; lin-41 NM_060087; lin-57 NM_076575; hid, NM_079412; reaper NM_079414; grim NM_079413; sickle AF460844; dpld NM_080033; m4 NM_079786; HLHm3 NM_079785; Tom NM_079349; drice NM_079827; hairy NM_079253; *D. simulans* hairy AY055843; *T. castaneum* hairy AJ457831; Lyra NM_080079; CG5599 NM_132772; CG1673 NM_132656; CG8199 NM_141648; CG1140 NM_167928; scu NM_078672; CG15093 NM_166306; CG17896 NM_130489.

### 2.8.2 Conserved 3' UTR-database

*Drosophila melanogaster* 3' untranslated regions (UTRs) were obtained from the Berkeley Drosophila Genome Project (BDGP, www.fruitfly.org/annot/release3.html) and those of >50nt were selected. Duplicate UTRs from different splice variants of the same transcript were removed. For each of the resulting 10196 non-redundant 3' UTRs, I mapped the last 50 amino acids of the corresponding ORF to the *D. pseudoobscura* genome sequence with TBLASTN (Altschul *et al.*, 1990) ($E \leq 10^{-5}$; http://hgsc.bcm.tmc.edu/projects/drosophila). I selected UTR matches that included the last 10 residues and had a sequence identity ≥80% or $E \leq 10^{-10}$ and compared these UTRs to the 3000 nucleotides downstream of the putative *D. pseudoobscura* ortholog with BLASTN (word-size 7; E≤10000, assuming a database the size of the whole *D. pseudoobscura* genome). Non-conserved nucleotides or those outside the matched regions were replaced by Ns in the *D. melanogaster* 3' UTR database to produce the conserved 3' UTR database.

The *D. pseudoobscura* genome has not been fully assembled. This means that some *D. pseudoobscura* genes are located close enough to the end of a contig that the UTR sequences may be missed. 386 *D. melanogaster* genes mapped to the *D. pseudoobscura* genome less than 1 Kb from a contig end; 189 mapped less than 500nt from a contig end. UTR conservation may be underestimated for these genes. For 3564 genes I did not detect a suitable ortholog using this protocol. 571 of these are known genes, the others are predicted genes about which little is known. For the 4662 *D. melanogaster* genes lacking annotated UTRs I assumed 3' UTRs of 2 Kb after the stop codon and built a separate database of predicted UTRs. The search for *Anopheles* orthologs was done using TBLASTN for the last 50 amino acids of each *D. melanogaster* ORF. Due to the more extensive sequence divergence, a lower cutoff threshold was allowed (E≤0.05) if the last exon of the predicted ORF mapped to the same location (+/- 1 Kb) in the annotated genome as the orthologous gene

(Zdobnov *et al.*, 2002). If not, the cutoff was $E \leq 10^{-5}$ as for *D. pseudoobscura*. The second more stringent step of comparing the last 10 amino acids was omitted.

### 2.8.3 miRNA-Screen

HMMer (Eddy, 1998) profiles were constructed for each of two alignments per miRNA containing copies of the reverse complement of the first (5') 8 nucleotides of the miRNA. The first alignment contained 5 copies of the exact complement, the second had an additional 5 copies with C replaced by T and A replaced by G to allow for G:U mismatches. I searched the conserved 3'UTR database with both profiles and a lenient domain bit score threshold (domT≥3) and combined the results. Sequence matches were extended to miRNA length+5nt, the hairpin loop and miRNA sequence were added and the sequence was evaluated using Mfold (Mathews *et al.*, 1999; Zuker *et al.*, 1999). Mfold uses dynamic programming to predict RNA secondary structure by free energy minimization. It includes experimentally determined thermodynamic parameters and knowledge about available RNA structures to account for sequence dependencies revealed in some RNA motifs (Mathews *et al.*, 1999). For *Anopheles*, predicted UTRs were searched for the presence of residues 2-7 of the predicted target site. The target sequences were extended and evaluated using Mfold. Only the best site in the *Anopheles* UTR was reported.

### 2.8.4 Statistics

For each miRNA, we calculated the mean and standard deviation of a background distribution, i.e. the Mfold free energy ΔG of 10,000 randomly selected sequences from the conserved UTR database with lengths of miRNA+5nt. For each prediction I calculated the Z-score as the number of standard deviations above the mean (see Introduction). To compute E-values, I fit an exponential function to the cumulative background distributions and extrapolated it to give a value for any observed energy and database size. E-values are not restricted to normal distributions and can scale with database size, so different searches can be compared.

# 3 Results and Discussion

The comparison of local spatial patterns like active sites in protein structures is complementary to sequence or fold comparison and can be used to annotate protein function. Prior to this thesis however, this type of comparison lacked a model to assess the statistical significance of matches. Specifically, although it had been recognized that the meaning of a specific RMSD value depended on the size and amino acid composition of the matches, this relation had not been investigated and remained unknown (see Introduction). I developed the program PINTS (Patterns In Non-homologous Tertiary Structures) to search for recurrences of residue or atom patterns in protein structures (see Materials and Methods). In the following chapter, I present a statistical model for the significance of local patterns in protein structure that I developed for my thesis. I used PINTS and the statistics to compare protein structures solved by structural genomics projects to databases of functionally relevant patterns to assess the use of active site comparisons on a large sample (Chapter 3.2). I also performed a detailed comparison of an archaeal fructose-bisphosphate aldolase using PINTS and the structural alignment program STAMP (3.3). Finally, I describe the development and current use of a server that allows for PINTS searches via the Internet (3.4). The following publications resulted from the results presented below.

**A. Stark**, S. Sunyaev, R.B. Russell; A Model for Statistical Significance of Local Similarities in Structure. *J. Mol. Biol*., 326, 1307-1316, 2003.

**A. Stark**, A. Shkumatov, R.B. Russell; Finding Functional Sites in Structural Genomics Proteins. *Structure*, submitted, 2004.

E. Lorentzen, E. Pohl, P. Zwart, **A. Stark**, R.B. Russell, T. Knura, R. Hensel, B. Sievers; Crystal structure of an Archaeal Class I Aldolase and the Evolution of $(\beta\alpha)_8$ Barrel Proteins. *J. Biol. Chem.*, 278(47), 47253-47260, 2003.

**A. Stark**, R.B. Russell; Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.*, 31(13), 3341-3344, 2003.

## 3.1 Statistical Model for Local Structural Patterns

Most methods to compare protein active sites – including PINTS – report RMSD values that accurately score the quality in terms of geometrical similarity. However, as outlined in the Introduction, the value that implies a meaningful similarity is highly dependent on the number and type of atoms being compared. To avoid ambiguities, or the choice of an arbitrary RMSD cutoff for any particular pattern, PINTS provides P- and E-values as a measure of statistical significance. For this, I derived a rigorous model for the behaviour of RMSD. Following previous work on statistical models for sequence comparison (Altschul & Gish, 1996; Karlin & Altschul, 1990; Mott *et al.*, 1990; Pearson, 1998), I use P-values to derive and present the model below, that can be easily converted to E-values (see Introduction).

### 3.1.1 Rationale for a Statistical Model of RMSD

For database searches, statistical significance is generally assessed by an *extreme value distribution* (EVD). This allows the calculation of a significance P-value from an *expectation function (EF)* that predicts the number of matches with an equally good or better score found in a database (i.e. *cumulative distribution* (CD) of scores):

$$P(x) = 1 - e^{-EF(x)}$$

*P(x)* is the probability of finding a score equal or better than *x* by chance, thus scores with high P-values are not meaningful.

For any CD there are just three models for the asymptotic behaviour of the EVD. If the CD decreases quickly with good scores, the EVD is a double exponent widely used in sequence comparison. However if it has a slowly decreasing tail or a finite terminal (i.e. bound by a lower value such as zero for RMSD), the EVDs are exponents of power functions that differ only in the sign of the exponent (Aldous, 1989; Gumbel, 1958; Kendall *et al.*, 1977). The choice of the correct model is critical for accurate statistics and must precede parameter estimation by fitting or calculation.

I performed searches of *query* patterns against a database using PINTS that optimally superimposed the matches according to the method of McLachlan (1979) and used the

associated RMSDs for the calculations below. For all searches I used lenient distance constraints that did not affect the range of the RMSD distribution considered.

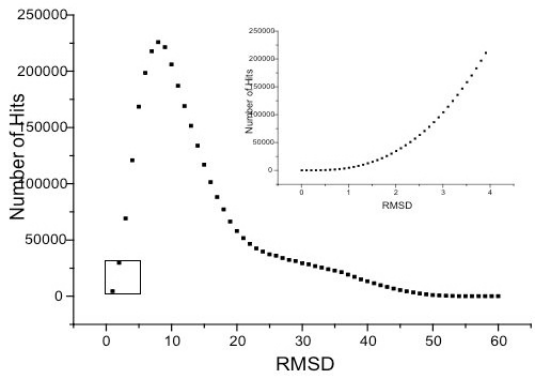



**Figure 3.1:** Example distribution of RMSDs (number of matches versus RMSD) for a typical query (3 Cα atoms from PDB entry 1a6m). The inset shows a magnification for RMSD≤4Å.

Figure 3.1 shows the background distribution of RMSDs for a typical local structural pattern. We considered only the increasing region of the curve, as the form of the distribution depends only on the tail of the CD for good scores (i.e. low values) and we believe that the decrease of the curve for high RMSDs is due to limitations in protein size. A closer view of this region (Figure 3.1, inset) shows a slow increase (slope approaches 0) for small RMSDs as is typical for power but not exponential functions.

When compared to a database, a query can be considered to have all ($N$) atoms in *ideal* positions, with deviations of other patterns scored by RMSD. RMSD$^2$ is the average squared distance between equivalent atoms and RMSD itself can be seen as an approximation of the average distance error (indeed if all deviations are equal, it is exactly this):

$$RMSD = \left[ \frac{1}{N} \sum_{N} (\Delta x^2 + \Delta y^2 + \Delta z^2) \right]^{\frac{1}{2}}$$

A perfect match thus has an RMSD=0 and increasing RMSDs correspond to an increasing dissimilarity. Restricting this to a maximum (RMSD≤$R_M$) requires atoms to be in a sphere around the ideal positions with an *allowed volume* proportional to $R_M^3$. Calculating RMSD involves finding the best superimposition of equivalenced sets of atoms by translation and rotation (McLachlan, 1979), which reduces these constraints on the positions for the first three atoms in a pattern. The first atom can be moved into the ideal position without any volume restriction. The second can be placed anywhere within a shell defined by two spheres and the third can lie in a ring-like volume as shown in Figure 3.2A.

### 3.1.2 Model assuming independence of atoms

We first develop an *independence model* where we consider only one atom per residue and assume they are independent and randomly distributed in space. We expect the probability of a residue from the database to match one from the query to increase with the allowed volume (above), and to be proportional to the database size ($D$) and residue abundance ($\phi$). Thus for a query with $N$ residues:

$$EF(R_M) = AR_M{}^B \cong D\phi_1 \prod_{i=2}^{N} \phi_i \rho V_i$$

$$with \quad A \cong D\Phi\rho^N \quad and \quad B = \begin{cases} 1 & for\ N = 2 \\ 3N-6 & for\ N \geq 3 \end{cases} \qquad (1)$$

where $V_i$ is the *allowed volume* for the $i^{th}$ residue (see Figure 3.2A), $\Phi$ is the product of all residue abundances and $\rho$ is a constant.

The simple power function $EF=AR_M^B$ is monotonously increasing as expected for a cumulative distribution and correctly assigns a probability of zero to perfect matches (RMSD=0). We expect $A$ to be correlated with residue abundance, to increase linearly with database size and to decrease exponentially with the number of residues in the query pattern. We expect $B$ to increase linearly with the query size and to be independent of database size or residue abundance. The linear behaviour for $N$=2 is expected since for two atoms RMSD merely describes a deviation from an ideal distance.

I searched a background database with random patterns of between 2 and 8 Cα atoms, and found the function above to fit the observed CDs accurately (Figure 3.2B). Plots of CDs in logarithmic scale (Figure 3.2C) show logarithmic behaviour typical of power but not exponential functions, which would be linear. The curves cross because high $R_M$s resemble random choices of $N$ atoms from the database, which creates an increasing number of permutations and combinations. The power function $AR_M^B$ naturally accounts for this behaviour as larger exponents always overtake smaller.

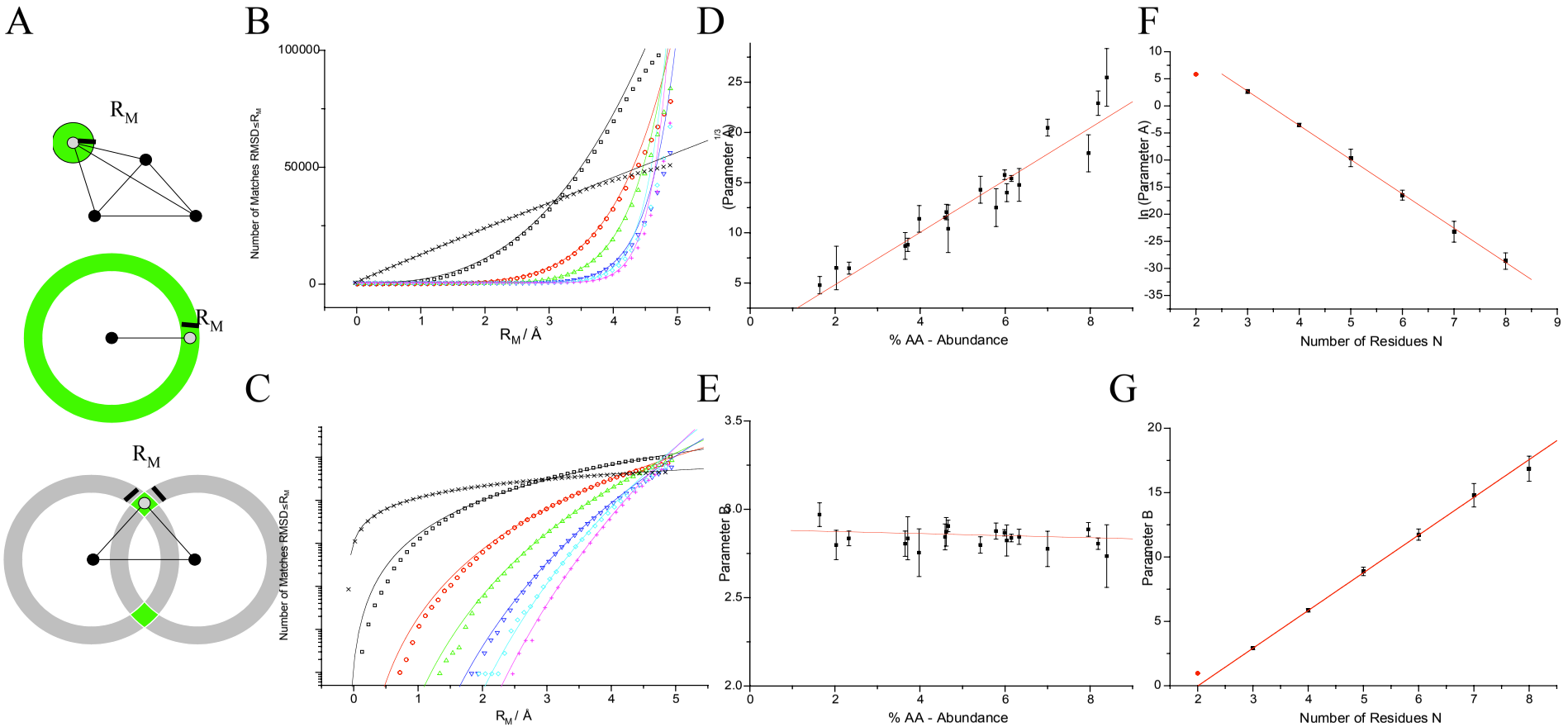



**Figure 3.2:** (A) Schematic two-dimensional representation of *allowed volumes* (green) for placing atoms (grey) of a pattern within an RMSD limit ($R_M$). RMSD≤$R_M$ restricts atoms on average to an allowed volume described by a sphere of radius $R_M$ around the ideal positions ($V_{4+}=4/3\pi R_M^3 \cong R_M^3$, left). The first atom is not restricted, but the second (middle) must be placed within a shell defined by two spheres around the first in a volume $V_2=4/3\pi(2R_M^3+6d_{intra}^2R_M)\approx 8\pi d_{intra}^2 R_M \cong R_M$ (for $R_M << d_{intra}$). The third atom (right) can lie in a ring-like volume $V_3=\pi d'_{intra}(2R_M)^2=4\pi d'_{intra}R_M^2 \cong R_M^2$. *(B)* RMSD distributions are typical of power and not exponential functions. I compared random patterns of 2-8 residues (1a6m) to the background database (BD) and fit the number of matches with RMSD≤$R_M$ (CD) to the function $AR_M^B$ (least squares). Linear (B) and logarithmic plots (C) are shown with fits represented by lines and observed values (every 10th) as symbols (from 2 - 8 residues as x, □, O, Δ, ∇, ♦, +). (D,E) Parameters A and B as a function of residue abundance For each residue, I randomly selected 3 patterns consisting of 3 residues from each of 3 structures (1a6m, 4rhv, 5p21) and searched the BD. The CD for each search was fit to the function $AR_M^B$. The cubic root of A (D) increases approximately linearly with the residue abundance ϕ (linear regression: A=2.6ϕ-0.64) whereas B (E) is independent of it (linear regression B=-0.006ϕ+2.9). (F) Parameters A and B as a function of pattern size. I randomly selected two patterns (Cα atoms) for each N between 2 – 8 residues from each of 3 structures (1a6m, 4rhv, 5p21), searched the BD and fit the CD for each search to the function $AR_M^B$. For each pattern, A was divided by residue abundances to remove the dependency shown in Fig. D. The averages of log(A) and B are plotted against N. Linear regression shows the exponential or linear dependence of A (F) or B (G) on N, respectively (excluding N=2 shown in red, see text).

I used searches with random queries of three residues of one type to explore the relationship between $A$ and $B$ and residue abundance. I fit the $EF(R_M)$ function and calculated $A$ and $B$. As expected (Eq. 1), the cubic root of $A$ increases linearly with abundance (Figure 3.2C) and $B$ shows no dependency (Figure 3.2D). I also explored the dependence of $A$ and $B$ on $N$ (Eq. 1) with random queries of various sizes ($N$=2-8). Here, I divided calculated values for $A$ by residue abundance to remove the dependency above. log($A$) versus $N$ shows a decreasing linear behaviour for $N \geq 3$ as expected (Figure 3.2F). For $N=2$, the values are below the extrapolated line due to the volume allowances in $A$. Linear regression gives a function to calculate $A$ for any pattern:

$$A = \begin{cases} a_0 \Phi a_2 & for\, N = 2 \\ a_0 \Phi a_3^N & for\, N \geq 3 \end{cases} \qquad (2)$$

with $a_0$=2.678x10$^9$, $a_2$=1.277x10$^{-7}$ and $a_3$=1.790x10$^{-3}$. The above also confirms our predictions (Eq. 1) about the behaviour of $B$: it increases linearly with $N$ for $N \geq 3$ (Figure 3.2G), with observed slope and intercept (2.93, 5.88) close to expected values (3, 6). For the special case of $N$=2, $B$ is 0.97±0.01 (i.e. close to 1).

The power function $EF(R_M)=AR_M^B$ models the behaviour of RMSD for searches with simple queries against a database. In the independence model, $A$ and $B$ for any query can either be obtained by fitting the function $EF(R_M)=AR_M^B$ to the CD or estimated with the equation above given only $N$ and residue abundance.

### 3.1.3 Accounting for dependency of covalently linked atoms

For protein functional sites the correct relative orientation of residues rather than their simple presence is crucial for activity. Moreover, representing a residue by only one atom (i.e. simple presence) is not sufficient to separate true matches from background (e.g. Russell (1998); Wallace *et al.* (1997) and below). One can account for this by considering multiple atoms per residue when calculating the RMSD. However, as these atoms are linked by covalent bonds they violate the assumption of random and independent atom distribution in the model above. We thus modified our geometrical arguments to account for the effect of atoms that depend on one another (*dependence model*).

The position of a second atom for a residue is constrained to the surface of sphere with a radius equal to its distance ($d$) from the first (Figure 3.3A). If $R_M \leq 2d$ then the position is restricted to a cap on this sphere. For each second atom we use the ratio of the two areas, $R_M^2/(4d_i^2)$, to correct $EF(R_M)$. Similar arguments can be applied to a third atom per residue. This is constrained to a circle around an axis formed by the first and second atoms, with small $R_M$s restricting this to an arc (Figure 3.3A), roughly proportional to $R_M$. We correct for both of these effects by adding two terms for each query residue with 2 or 3 atoms: $c_2 R_M^2$ and $c_3 R_M^3$, respectively. These corrections predict an abrupt transition around $R_M=2d$ where the restriction for the second atom no longer applies and the model assuming independence holds. Differences in RMSD when considering more than 3 atoms per residue are due to conformational differences (i.e. rotamers). In practice, 3 atoms is sufficient to define residue orientations, though our model is normally conservative when more than 3 are used.

We now have a modified EF:

$$EF(R_M) = AR_M^{\,B}\left[c_2 R_M^{\,2}\right]^S\left[c_3 R_M^{\,3}\right]^T = A' R_M^{\,B'} \qquad (3)$$

where $S$ and $T$ are the numbers of query residues where 2 and 3 atoms are used, respectively; $A'=Ac_2^S c_3^T$, and $B'=B+2S+3T$.

To test this modified model, I searched with random patterns with $N$=2-4 gradually increasing the number of residues with second or third atoms and fit the above function. The effect of dependent atoms is clear in the CDs for $N$=2 with 2 or 3 atoms per residue (Figure 3.3B): there are fewer matches with low $R_M$ (i.e. flattened curves). At about $R_M$=2Å, the CDs show the predicted transition to the independence model.

The effect of dependent points is also evident in the behaviour of the variables. Figure 3.3C shows the effect of an increasing number of dependent atoms on log($A'$), which the model predicts to decrease linearly. The initially flat curves for $N$=2 are due to the transition from 2-3 residues discussed above for the independence model. The curves for $N$=4 are just above those for $N$=3, which reflects a lack of data for small $R_M$. From these curves we determined $c_2$=0.196±0.026 and $c_3$=0.094±0.024. $c_2$ corresponds to an atom distance $d$=2.3Å which is the range of intra-residue distances observed between Cα and Cγ/Oγ atoms.

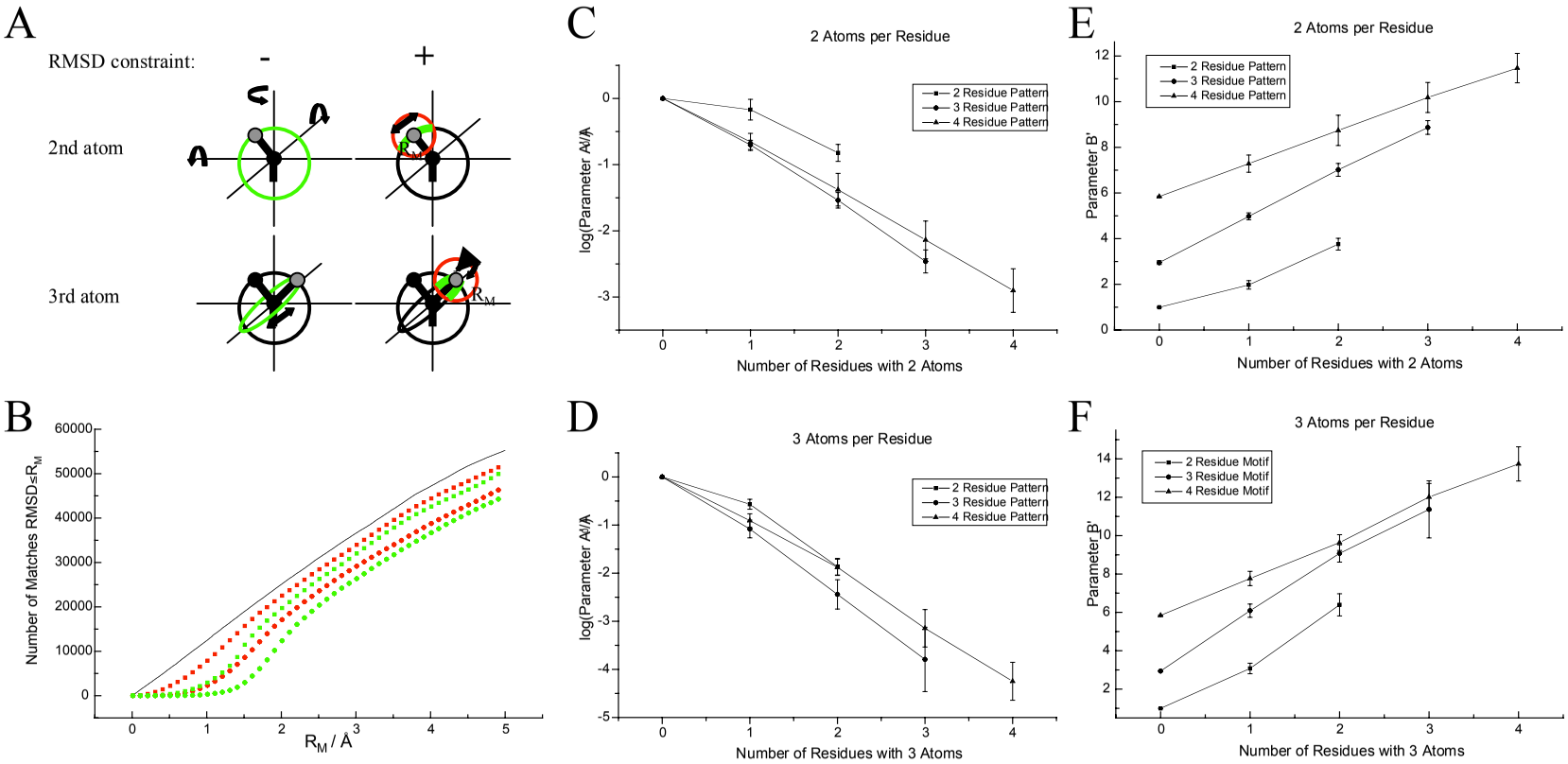



**Figure 3.3:** (A) Schematic two-dimensional representation of *allowed areas* (green) for placing *dependent atoms* (grey). Without RMSD constraints, a second atom per residue could be anywhere on the surface of a sphere around the first atom with a radius defined by the chemical bond (top left). RMSD≤$R_M$ additionally restricts the position of the second atom to a sphere (red) around the ideal position (top right). The allowed area is the intersection, i.e. a sphere cap with an area of $\pi R_M^2$. Once two atoms are fixed, the position of a third atom is restricted by the residue conformation to a circle around the axis formed by the first two atoms (bottom left). The intersection of this circle and the sphere of RMSD≤$R_M$ is an arc of the circle with a length roughly proportional to $R_M$. (B) Effect of dependent atoms on the CD for N=2. 3 patterns (Cα atoms) for each N between 2 – 4 residues were randomly selected from each of the structures (1a6m, 4rhv and 5p21) and searched against the BD using a second (e.g. Cα and Cγ) and third (e.g. Cα, Cγ1 and Cγ2) atom for an increasing number of residues in the pattern (from none to all). Representing one (□) or both (O) residues for N=2 by 2 (red) or 3 (green) atoms leads to a flattening of the curve due to fewer matches for small RMSDs. A around $R_M$=2Å, the resulting power-law like curves show the predicted transition to the independence model (for N=2: linear). (C,D) A' as a function of the number of dependent atoms For all searches (as in *B*), I fit $A'R_M^{B'}$ to the leftmost part of the CDs, divided A' by A (independence model) and plotted log(A'/A) against the number of residues with 2 or 3 atoms used during the searches. log(A'/A) decreases linearly with the number of oriented residues if 2 (C) or 3 (D) atoms per residue are used. (E,F) B' as a function of the number of dependent atoms Considering the same fits used in (C,D). B' increases linearly with a slope of 2 or 3 if 2 (E) or 3 (F) atoms per residue are used respectively.

Figure 3.3D shows the effect of an increasing number of dependent atoms on *B'*, which the model predicts will increase by 2 or 3 for each residue in the query containing 2 or 3 atoms, respectively. The slopes of the curves are indeed close to 2 and 3, though the 4-residue queries are hampered by a lack of data for small $R_M$, and we see an initial effect similar to that above for $N$=2.

### 3.1.4 Final P-value for Local Structural Pattern Comparison

We can now calculate the P-value for any RMSD observed for a query pattern:

$$P(RMSD \le R_M) = 1 - e^{-EF(R_M)}$$

$$EF(R_M) = \begin{cases} a_0 \Phi a_2 R_M{}^{0.97} \left[ c_2 R_M{}^2 \right]^{S-1} \left[ c_3 R_M{}^3 \right]^{T-1} & for \quad N = 2 \\ a_0 \Phi a_3{}^N R_M{}^{2.93N-5.88} \left[ c_2 R_M{}^2 \right]^S \left[ c_3 R_M{}^3 \right]^T & for \quad N \ge 3 \end{cases} \qquad (4)$$

where $N$ is the number of residues, $S$ and $T$ the numbers of query residues where 2 and 3 atoms are used in fitting, $\Phi$ is the product of all abundances and $a_0$, $a_2$, $a_3$, $c_2$ and $c_3$ are empirically determined constants (see above). The corrections in square brackets apply only if their values are <1, corresponding to $R_M < d_{intra}$. We emphasise that a key step in the derivation of this function is the demonstration above that the EVD for RMSD is the exponent of a power and not an exponential function.

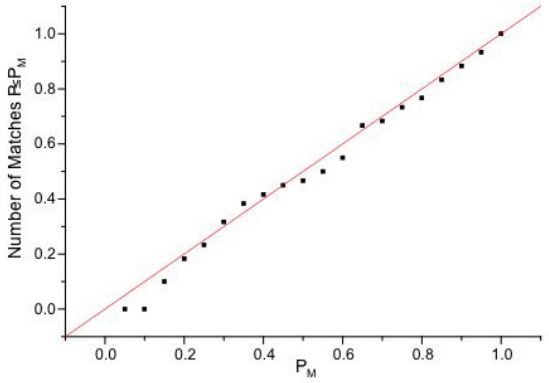



**Figure 3.4:** P-value plot. The cumulative distribution of P-values for the best match ($P_M$) of each of 60 searches is close to a straight line with the slope of 1 (goodness of fit $\chi^2$=0.28426).

P(x) normalizes a distribution of scores (x), a property that can be used to test a model. Specifically, if a model to predict EF(x) is valid then a cumulative histogram of P(x) for the best scores in a series of searches should give a straight line with a slope of 1. P(x) indeed shows this behaviour, providing confidence in our model for the behaviour of RMSD (Figure 3.4).

### 3.1.5 Comparing Patterns to Databases of Proteins

To demonstrate how the above formula can be applied to detecting recurrences of a known functional site, I compared the trypsin catalytic triad to all structures in the PDB (Figure 3.5). Triads from homologous proteases have RMSD≤0.6Å (associated with P≤0.009) with the exception of distorted sites owing to bound inhibitors, and the distribution of triads from

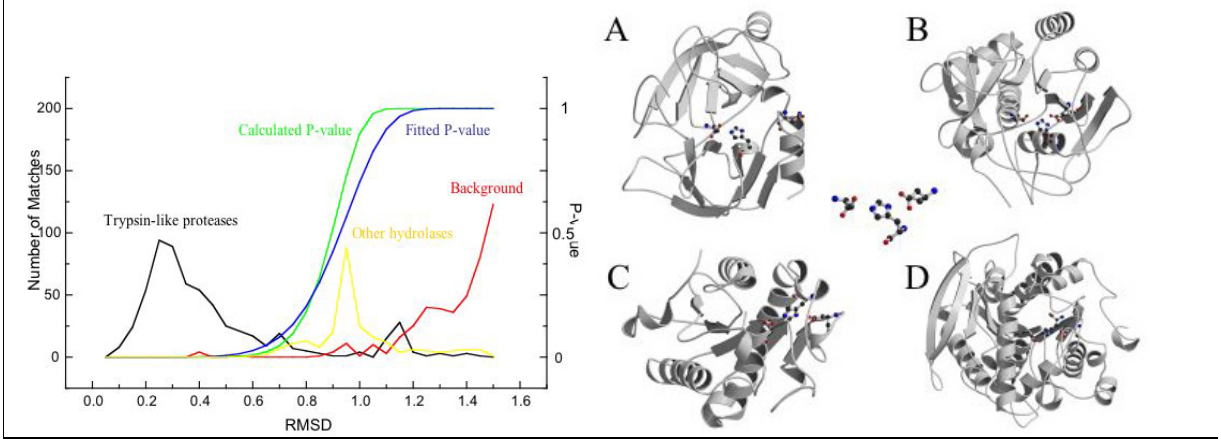



**Figure 3.5:** Comparing patterns to databases of proteins: search with a trypsin catalytic triad. Left: P-values derived by fitting the CD of background matches (blue) and from parameters calculated using the statistical model (green) are compared against the distribution of matches in the PDB classified as trypsin-like proteases (black), other hydrolases (yellow) and other proteins (background, red). Right: Top scoring matches are to trypsin (1mct, A, E=0), the subtilisin-like fold (E=0.4, 1cse, subtilisin), the α/β-hydrolases (E=0.8, 1jkm, carboxylesterase) and the flaxodoxin-like cutinases (E=0.8, 1cex, cutinase).

different folds peaks at around 0.9Å (P=0.0009-0.9). Plots for calculated and observed (i.e. fitted) P-values superimpose very well and show a sharp transition from small values to 1 at an RMSD of 0.9Å where the first false matches appear. Top scoring matches are to the subtilisin-like fold, the α/β-hydrolases, and the flaxodoxin-like cutinases (Figure 3.5).

During these searches, I also detected a previously undescribed Ser-His-Glu triad in the yeast proteasome α-subunit (1g0u (Groll *et al.*, 2000); Ser144, His147, Glu159; P=0.03 when compared to thrombin or cutinase). The location of these residues on the surface of the structure near to the pore, and their conservation in many homologs, suggests a possible catalytic function.

I was also able to detect all patterns studied or reported by other methods (Artymiuk *et al.*, 1994; Fetrow & Skolnick, 1998; Fischer *et al.*, 1994; Kleywegt, 1999; Russell, 1998; Wallace *et al.*, 1997; Wallace *et al.*, 1996). For nearly all patterns true matches had significant P-values (e.g. ribonuclease A and T, thermolysin, Zn-fingers, heme binding sites, cellobiohydrolase). However, for three examples matches were insignificant according to our

formula and were indeed found among background matches. These were active sites of just two residues from lysozyme (Wallace *et al.*, 1997) and staphylococcal nuclease (Artymiuk *et al.*, 1994) and a putative three-residue disulphide oxidoreductase site in a phosphatase (Fetrow *et al.*, 1999). For the lysosome and nuclease examples the insignificance was expected as other methods were also unable to discern them from noise, and to our knowledge the oxidoreductase site has not yet been verified.

### 3.1.6 Comparing Proteins to Databases of Patterns

I also tested the formula in a reverse situation: that is to compare an entire protein structure to databases of functional patterns (see Materials and Methods). As our database entries include residues not directly related to function, I allowed for partial matches where only part of the database entry is found in the query. To correct *EF* to account for this, it is multiplied by the total number of possible partial matches of the same size (i.e. consisting of any residues) that are contained in the database entry (correction for multiple testing).

I compared the phosphoenolpyruvate carboxykinase (PEPCK) structure (without bound ligands) to the *ligand-binding sites* database. PEPCK contains a P-loop but adopts a structure that is otherwise quite different from other P-loop containing nucleotidyl transferases (Marquez *et al.*, 2002; Russell *et al.*, 2002). Indeed, if one ignores homologs, the most similar protein according to FSSP (Holm & Sander, 1996) is hydrogenase-2-maturation protease (1cfz, Z=4.0), which performs a very different function. In contrast, our search identified nucleotide-binding sites with high significance in proteins of different folds that are structurally dissimilar from PEPCK (DALI Z<3). As expected, the best matches were to sites from P-loop proteins (top rank, elongation factor Tu, 1eft, $P<10^{-6}$) where the nucleotide binds in a very similar orientation (Figure 3.6). These were followed by matches containing residues from the ATP binding site of GroEL-like proteins (best, GroEL, 1der, $P=4x10^{-3}$) and the FMN binding site in flavodoxin (6nul, $P=6x10^{-2}$). Here the residues common to the matches make similar contacts to phosphates attached to nucleotides in otherwise different conformations (Figure 3.6). The first negative match (P=0.88) comprises 5 residues from a large heme-binding site.

We also compared the bacterial quorum-sensing protein LuxS to the database and found highly significant matches to Zn-binding active sites across folds and no other site among the

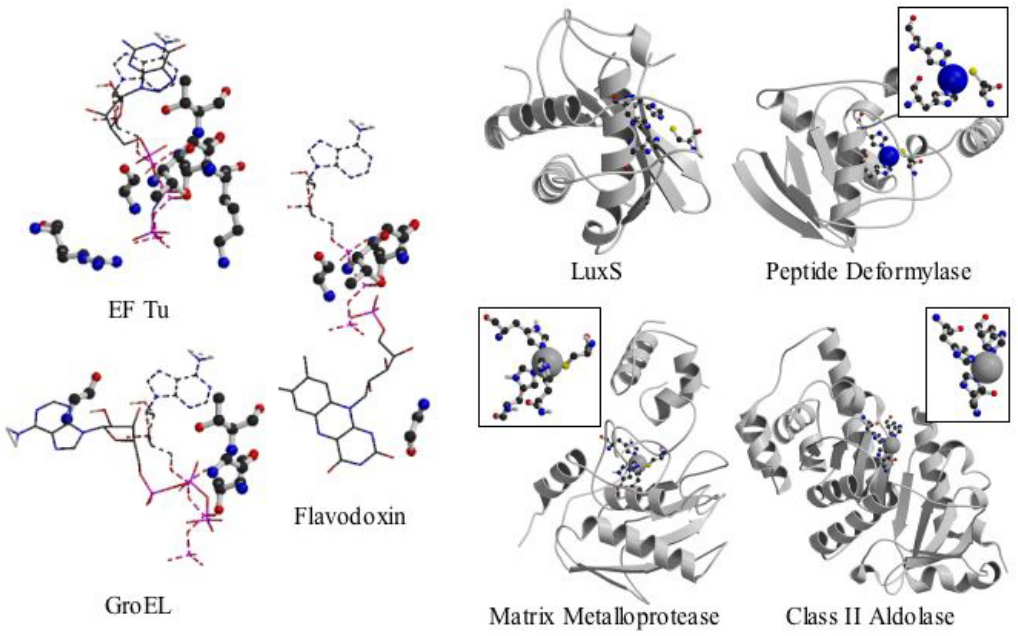



**Figure 3.6:** Comparing Proteins to Databases of Patterns. Representatives from the top scoring folds are shown; Residues in ball-and-stick are matches found during the search. Left: Matches found after comparing the unliganded structure of *Trypanosoma cruzi* PEPCK (1ii2) to the ligand binding sites database. Figures showing matches to elongation factor EF Tu ($E=1x10^{-6}$, 1eft), GroEL ($4x10^{-3}$, 1der), and flavodoxin ($E=6x10^{-2}$, 6nul). The location of ATP within *E. Coli* PEPCK (1ayl; superimposed on 1ii2) is shown as a broken line, and the location of bound ligands in the matches are shown as continuous lines. Right: Matches found after comparing LuxS (1j98) to the *SITE annotations* database are: peptide deformylase ($E=2x10^{-6}$, 1bs6), the zincin-like matrix-metalloprotease stromelysin-1 ($E=8x10^{-5}$, 1slm) and the class II aldolase rhamnulose-1-phosphate aldolase (E=0.2, 1gt7). Matches to sites within proteins of the LuxS/MPP-like metallohydrolase (i.e. proteins similar to LuxS), cysteine rich, dioxygenase and classical C2H2,C2HC zinc finger folds are not shown for clarity.

top 100 matches ($P \leq 0.97$). Representatives of the different folds were peptide deformylase, the zincin-like matrix-metalloprotease stromelysin-1 and the class II aldolase rhamnulose-1-phosphate aldolase, all suggestive of a hydrolytic activity (see Figure 3.6). Indeed, LuxS was found to cleave S-ribosylhomocysteine to homocysteine and the autoinducer molecule 5-dihydroxy-2,3-pentadione (Schauder *et al.*, 2001). A structure with the putative substrate bound and noticeable clustering of conserved residues suggests that we identified the correct active site in the absence of overall sequence or structural similarity to known structures (Hilgers & Ludwig, 2001; Ruzheinikov *et al.*, 2001).

### 3.1.7 Comparing Entire Protein Structures

It is possible to apply our formula when comparing entire protein structures, with no prior definition of active or functional sites. A pairwise comparison of two proteins that share a biochemical or cellular function (e.g. catalytic activity, specific binding characteristics, etc.)

can suggest the molecular basis for the common feature. For example, when cytochrome c6 is compared to the non-homologous multiheme cytochrome c7 three similarities were detected. All involve the CxxCH heme attachment site, of which one is present in cytochrome c6 and three in c7. One of these matches also identified an additional serine common to both proteins (Figure 3.7).

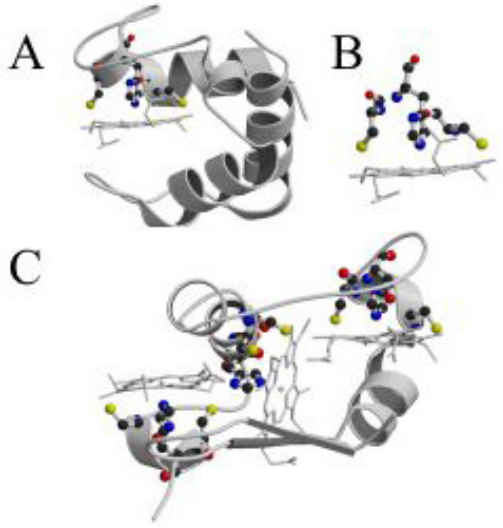



**Figure 3.7:** Pairwise comparison of entire proteins: we compared cytochrome c6 (1c75) and the multiheme cytochrome c7 (1hh5). Three similarities were detected, of which one is present in cytochrome c6 (residues 32,35,36) and three in c7 (1: 26,29,30, $E=1x10^{-6}$; 2: 49,52,53, $E=0.18$; 3: 62,65,66, $E=3.6x10^{-4}$). The match to site 2 also identified an additional serine (residue 44 in c6 and 55 in c7) common to both proteins.

Generally however, the greatly increased search space has a critical effect on the statistics: searching more amino acids increases the number of random matches, and can render true matches insignificant i.e. bury them in noise. For example, a protein versus pattern search comparing trypsin to a database of functional patterns, or a pattern versus protein database search comparing only the catalytic triad (His57, Asp102, Ser195) to a database of whole structures identifies true functional similarities to be significant. However a pairwise comparison between trypsin and subtilisin detects the similarity, but does not find it to be significant owing to the large number of background matches with equivalent RMSD and size introduced by the comparison of two whole proteins (of 223 and 275 amino acids respectively). This is not a limitation of the method, but a fact of life when searching for similarities within large databases (see Jones & Swindells, 2002) for a similar discussion about sequence searches).

### 3.1.8 Deviations from predicted E-Values

Searches with a salt bridge and disulphide bond show curious differences between the observed and calculated distributions (Figure 3.8) due to their unusually high frequency in proteins with different folds. Here our model has a clear advantage over an empirical fit to the observed distribution as it does not down-weight the importance of these structural features. For example, the observed distribution would suggest that an RMSD of 0.18Å between two

Arg-Glu salt bridges is insignificant (P>0.99), though our model gives P=0.0014. This difference is even more pronounced for a disulphide bond: $P_{fit}$=1 versus $P_{model}$=0.001 for RMSD=0.08Å. We can thus assess significance for highly abundant, but still functionally or structurally important patterns.

To investigate the possible correlation between deviation from our model and contribution to protein stability, I specifically searched for residue patterns overrepresented in unrelated proteins (see Materials and Methods). For two residue patterns, we expected results similar to the widely used pair potentials (e.g. Bahar & Jernigan, 1997; Hendlich *et al.*, 1990; Miyazawa & Jernigan, 1996; Sippl, 1990). However, while pair potentials are derived from merely measuring residue-residue distances in proteins, we required a precise relative orientation of the residues and thus expected to gain a more precise understanding. In addition, the search also found higher order patterns (i.e. with more than two residues) to be overrepresented suggesting, cooperative effects not captured by pair potentials (e.g. Carter *et al.*, 2001; Kannan & Vishveshwara, 1999).

For two-residue patterns, I found an equal number of polar or charged and hydrophobic interactions in the top 20 in agreement with previous studies of stabilization centres (Dosztanyi *et al.*, 1997) or residue pairs (Bahar & Jernigan, 1997). The most abundant two-residue patterns, that are about 2000-fold more frequent than expected, were arginine-aspartate/glutamate salt bridges and disulphide bonds (see Figure 3.8). I thus found pairs of residues known to be beneficial for protein stability: a disulphide bond for example contributes 17 kJ/mol and buried salt bridges 12 – 20 kJ/mol (Fersht, 1999). An interacting leucine pair that is part of a larger hydrophobic cluster was 1500-fold overrepresented followed by aspartate-serine/threonine hydrogen bonds (800-fold). This changed dramatically to predominantly hydrophobic interactions when I relaxed the requirement for residue orientation. Polar interactions seem to occur with well-defined residue orientations, whereas hydrophobic contacts appear more flexible, a trend also seen by (Bahar & Jernigan, 1997).

When I inspected top ranking interactions for the individual amino acids, I found that charged residues (including histidine) are mainly involved in salt bridges, cysteine forms disulphide bonds, and hydrophobic residues preferably cluster together. The polar non-charged serine, threonine, asparagine and glutamine most frequently form hydrogen bonds to aspartate and glutamate. Proline often binds to aromatic residues, which generally make hydrophobic

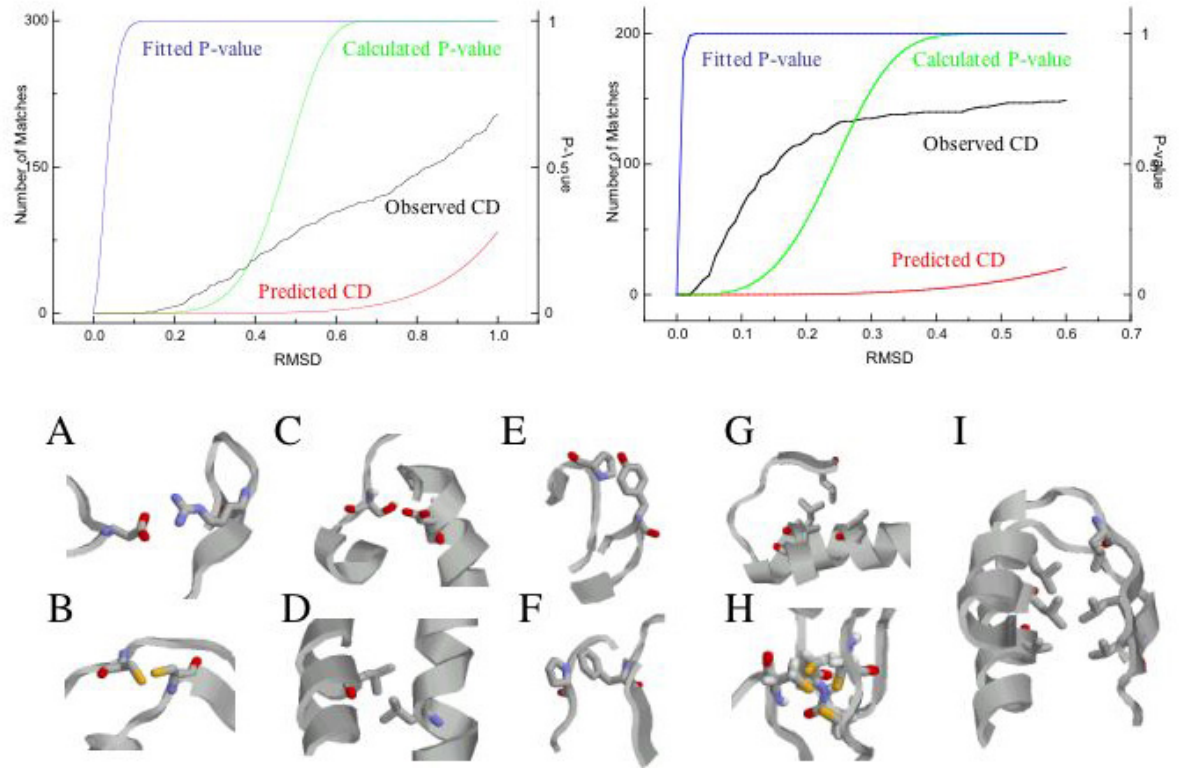



**Figure 3.8:** Deviations from predicted E-values. Top: Comparison of observed and calculated P-values and cumulative distributions (CDs) for searches with an Arg-Glu salt bridge (left, 1xel, Arg176, Glu157) and a disulphide bond (3ckb, Cys9, Cys86) against the background database. Bottom: Residue patterns found to be overrepresented (representatives): Asp-Arg salt bridge (A: 2350x overrepresented, 1ckq, residues 135,187), disulphide bond (B: 1900x, 9wga, 17,35), Asp-Ser hydrogen bond (C: 860x, 1pys, 247,226), Leu-Leu pair (D: 1560x, 2end, 24,67). Top-scoring partner of Pro are Tyr (E: 240x, 1gof, 88,99) and Phe (F: 140x, 1kek, 448,474). A 4 leucine cluster (G: 1870x, 1e8c, 5,8,21,36), a 4 residue disulphide cross (H: 1afp, 1460x, 26,28,49,51), and a 6 residue hydrophobic cluster (I: 1210x, 2bnh, leucines 47,53,56,75,85, valine 71) are examples of larger patterns found to be overrepresented.

contacts. This is particularly interesting as aromatic residues have been shown to be critical for proline-binding in SH3- and WW-domains, or other proteins (e.g. Bjorkegren *et al.*, 1993; Carl *et al.*, 1999; de Beer *et al.*, 2000; Lim & Richards, 1994; Macias *et al.*, 1996; Wang *et al.*, 2003; Yu *et al.*, 1992). It remains to be seen whether some interactions that seem to be beneficial in the protein interior are also important for protein-protein or protein-ligand binding, despite generally different pair potentials for these cases (see for example Aloy & Russell, 2002).

The most frequent among the patterns consisting of 3-6 residues are hydrophobic clusters with the exception of a four residue disulphide cross (Harrison & Sternberg, 1996) and a six cysteine pattern common to different folds such as the knottin-, metallothionein-, or disintegrin-fold (e.g. Mas *et al.*, 1998) shown in Figure 3.8. The prevalence of hydrophobic clusters was observed in previous studies examining spatial clustering (Kannan &

Vishveshwara, 1999) or nearest-neighbours relations (Carter *et al.*, 2001) and emphasizes their importance for protein stability and folding.

### 3.1.9 Concluding Remarks Regarding the Statistical Model

We have presented a formula to calculate the significance of any local protein structural similarity, and have shown that it can discern meaningful similarities from noise. There are likely many undetected similarities between protein structures, related to protein function, stability or transport. Reliable statistics are pivotal if patterns consisting of as few as two residues from disparate parts of the polypeptide chain are to be distinguished from noise.

For all searches, I inspected the ranking of matches and found that our P-values not only distinguish true positives from noise, but also permit the comparison of results involving patterns of different sizes. The RMSD required for a given significance (e.g. $P=10^{-4}$) varies with the pattern as is clearly seen using the examples above: whereas Arg-Glu salt bridges (2 residues, 6 atoms) have to be within 0.02Å, catalytic triads (3 residues, 8 atoms) can deviate up to 0.4Å and the PEPCK ATP-binding site (7 residues, 16 atoms) up to 1.5Å according to our model. RMSD alone places insignificant matches with few residues or atoms above those that are larger and significant. For the comparison of a large set of protein structures determined by structural genomics projects (e.g. Teichmann, *et al.*, 2001; see Chapter 3.2) to pattern databases, I found that P-values for true matches typically have $P<10^{-5}$, those with similar chemical groups $P\approx10^{-4}$ - $10^{-2}$ and negatives $P\geq0.1$. Our statistical formula is thus able to discern significant similarities from noise when entire protein structures are searched against databases of functional patterns, even when matches are only parts of larger structures.

We confirm that active site descriptors using only the positions or distances of Cα or Cβ atoms are generally not sufficient to reach an appropriate level of specificity while retaining sensitivity (Russell, 1998; Wallace *et al.*, 1997). For example, when searching with Cαs from the trypsin catalytic triad, negative matches have RMSDs as low as 0.5Å. This value is comparable to those seen between close trypsin homologs, and true triads from subtilisins have values as high as 2.0Å (data not shown). Previous approaches recognised this limitation and used constraints on sequence and secondary structure context (Fetrow & Skolnick, 1998) or Cα atom geometry around the active site (Fischer *et al.*, 1994) to reduce the number of false matches (i.e. increase specificity). This approach can fail if only catalytic residues are common and might thus restrict methods to proteins with similar folds. Other methods

considered residue orientation (Artymiuk *et al.*, 1994; Kleywegt, 1999; Russell, 1998) or concentrated on functionally important atoms (Wallace *et al.*, 1997; Wallace *et al.*, 1996). Our formula permits searches to be made without requiring manual definitions of queries (Fetrow & Skolnick, 1998; Wallace *et al.*, 1997), and provides a general basis to separate true matches from background without the need to define query-specific RMSD or distance thresholds.

To compare our statistic to previous work on the significance of RMSD, I extrapolated our results to larger numbers of residues, considering only Cαs. In this situation our model is conservative: for example, to give a significance of $P=10^{-5}$ when comparing two 70 residue proteins one previous study required an RMSD=6 (Reva *et al.*, 1998), whereas our model needed RMSD=4. We suspect that this discrepancy is due to differences in background models. The covalently linked amino acids and the tendency of proteins to form regular secondary or super-secondary structures create additional dependencies. We did not wish to consider these in a model for similarities involving a few residues close in space, but not necessarily adjacent in sequence.

More generally, robust methods for structure comparison are key to the success of structure-based functional annotation required for structural genomics (e.g. Burley, 2000). Identification of common local structural patterns is highly complementary to fold comparison (e.g. DALI, Holm & Sander, 1993): it can both confirm functional similarities suggested by a common fold and identify instances of convergent evolution where common local patterns are found in different folds. Like the sequence comparison methods used to annotate genomes, these methods can be applied automatically to thousands of new structures and provide initial functional predictions without human intervention.

## *3.2 Assigning Function to Protein Structures – Examples*

We wanted to get a general picture of the applicability of PINTS for functional assignment of novel proteins based on an analysis of a large test set. I thus probed for functional site similarities in the 254 currently available structures from structural genomics projects and compared this to overall fold similarities reported by Dali. We could confirm functional similarities suggested by an overall similar fold but also found examples of functional similarities despite no sequence or overall fold similarity demonstrating the complementarity of this approach to those based on structural alignment.

### 3.2.1 Overall Performance of Functional Site Comparison

We considered 254 structures labelled as structural genomics with release dates up to October 2003 but excluded all cases where functional similarities were obvious from sequence comparison (see Materials and Methods). This filter left 157 of the original 254 structures for further analysis. I compared these structures to representative structures in the FSSP database with Dali and to databases of functionally relevant patterns with PINTS applying thresholds usually associated with a similarity in function (i.e. Dali: $Z \geq 10$; PINTS: $E \leq 10^{-3}$; see Materials and Methods). Dali finds matches with $Z \geq 10$ for 61 (39%) and PINTS reports matches with $E \leq 10^{-3}$ for 29 (18%). For 17 (11%) both methods find significant matches, 44 (28%) were only found by Dali and 12 (8%) only by PINTS. The proportions are similar when structures labeled as "unknown function" are used instead (Dali: 41%, PINTS: 21%; Overlap: 12%; Dali-only 29%, PINTS-only 8%).

There are several reasons why similarities are found by Dali and not by PINTS. For example, active sites can sometimes be distorted by binding to other molecules and cannot be detected with statistical significance. This effect is most pronounced for similarities involving a small number of residues. For example, our best match for Tm1158 (1o1y) is to three residues from the active site of a glutamine amidotransferase domain (1a9x). Although the E-value of 0.035 is above the threshold used here, the match is from the same family as the best Dali match (1qdl; Z=20.4). Other missed similarities include those lacking common small-ligand binding sites, such as scaffolding proteins (e.g. 1oyz/1b3u; Dali Z=15), or DNA/RNA binding proteins (1jyh/1d5y, Z=14; 1ljo/1d3b, Z=12). Some Dali similarities are to other proteins that are also of unknown function, where no functional pattern is present in any database (e.g. 1o13/1p90; Z=11.5) or involve fold matches without a similarity in function (e.g. helical bundles (1n1q/1bcf, Dali Z=18) or a periplasmic divalent cation tolerance protein with fold similarity to anthranilate isomerase (1p1l/2pii; Z=10).

The 12 structures matched only by PINTS were mostly novel folds where a functional similarity was found between proteins with different overall folds. Of these 5 were metal binding sites, 2 were ligand binding sites, 3 were anion binding sites and 2 were short linear motifs with similar conformations probably due to their secondary structure context but lacking an apparent functional role.

Using a large number of structural genomics targets without sequence similarity to known structures, we can find functional centres within an overall similar fold for 11% and detect functional similarities across folds that cannot be detected by structural alignment methods for an additional 6% of all structures. Specific examples of how functional site similarity can aid structure-based annotation of function are discussed in the sections that follow.

### 3.2.2 Confirmation of Superfamily, or Resolution of Ambiguity

Overall sequence or fold similarity does not always reveal the correct function. For example, the archaeal fructose-1,6 bisphosphate aldolase shows the highest fold similarity to a triosephosphate isomerase (1hg3, Dali Z=17.7) high above the FBPAs from eukaryotes (Dali Z=7.4 for 1fbp) (see below and Lorentzen *et al.*, 2003). Functional site comparison methods have already shown some promise in resolving these situations (e.g. TIM-barrels, Lorentzen *et al.*, 2003), or α/β hydrolases (Wilson *et al.*, 2004; Sanishvili *et al.*, 2003), see Babbitt, 2003 for a general discussion).

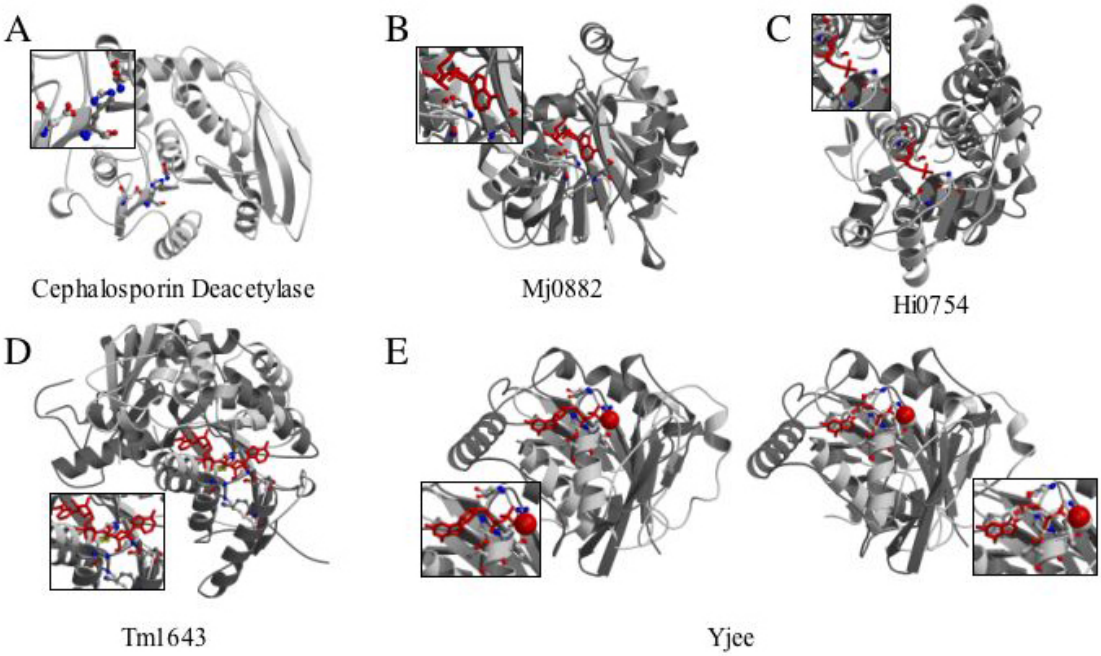



**Figure 3.9:** Functional site conservation within superfamily or fold. Structural genomics proteins (query) are shown in light grey and the match in dark grey cartoons. Matched residues are shown in ball-and-stick, with the ligand of the database structure in red (magnified in insets). (A) Similarity between cephalosporin deacetylase (1l7a) and the catalytic triad of prolyl-oligopeptidase (1h2x; $E=1x10^{-8}$). (B) Residues Gly63, Asp84 and Asp113 from the hypothetical protein (HP) Mj0882 (1dus) matched to the S-adenosylmethionine binding site from isoflavone o-methyltransferase (1fpx; $E=3x10^{-5}$; Dali Z=13.2). Here Dali's first match (1nv8; Z=18.2) ranks 2nd in PINTS ($E=9x10^{-3}$). (C) Residues Thr78, Ser79, Ser147, Thr150 from the HP Hi0754 (1nri) match to the glucosamine 6-phosphate binding site in the isomerase domain of glucosamine 6-phosphate synthase (1moq; $E=6x10^{-9}$; Dali Z=12.6). Dali's best match (1jeo, Z=14.2) belongs to the same superfamily (c.80.1). (D) Residues Gly7, Gly9, Gly12, Asp28, Lys32 and Cys55 from HP Tm1643 (1j5p) match the NAD binding site in Lactate dehydrogenase (2ldb, $E=3x10^{-4}$; Z=9.3). (E) Residues Gly43 and 45-48 from Yjee (1fl9, unliganded, left (Teplyakov *et al.*, 2002)) match the Ran GDP binding site (1a2k, $E=2x10^{-5}$). Superposition of the ADP-bound form of Yjee (1htw, right) showing the similar position of the nucleotide.

There are several structures for which we could support functional similarity also inferred by Dali in addition to highlighting the functional centre. These include the similarities between cephalosporin c deacetylase and $\alpha/\beta$ hydrolases (PINTS $E=1x10^{-8}$; Dali Z=20; Figure 3.9A), between Mj0882 and methyltransferases ($E=3x10^{-5}$; Z=13.2; B), between Hi0754 and glucosamine 6-phosphate synthase ($E=6x10^{-9}$; Z=14.2; C) or between Tm1643 and lactate dehydrogenase ($E=3x10^{-4}$; Dali Z=9.3; D).

For Yjee (Teplyakov *et al.*, 2002) (Figure 3.9E), the best Dali match is marginal (RecA; Z=6.3), not readily allowing any functional conclusions. However, the functional site found here is highly significant, involving 5 residues from the GDP binding sites of Ran ($E=2x10^{-5}$) or other P-loop nucleotide hydrolases from the same superfamily. The subsequently determined ADP-bound form of Yjee (deposited six months later) shows that the two nucleotides superimpose perfectly (Figure 3.9E, right).

### 3.2.3 Sites Found by Similarities Between Different Folds

Functional sites found across different folds are both intriguing and useful: they can suggest aspects of convergent evolution or can suggest functional details for proteins adopting folds not seen before. Those detected here fall into broad classes that we discuss below.

**Metal or phosphate binding sites**

Nature frequently reinvents similar metal-binding sites (Russell, 1998), and unsurprisingly several similarities observed across folds involve metals. For example, Tm1083 has a highly significant similarity with the calcium-binding site of staphylococcal nuclease ($E=5x10^{-5}$), despite an obvious difference in fold (Figure 3.10A). Aq_1354 contains a site similar to the Zn containing active site of carbonic anhydrase ($E=8x10^{-3}$) and other Zn-binding sites (Figure 3.10B). Although no metal is present in the structure, the conservation of the histidine residues suggests that the site is real, Zn being absent from the structure owing to EDTA in the purification protocol (Oganesyan *et al.*, 2003). The Dali server in contrast, reported only a marginal match to glucuronidase (1mqp; Z=3.3) that did not allow reliable functional inferences.

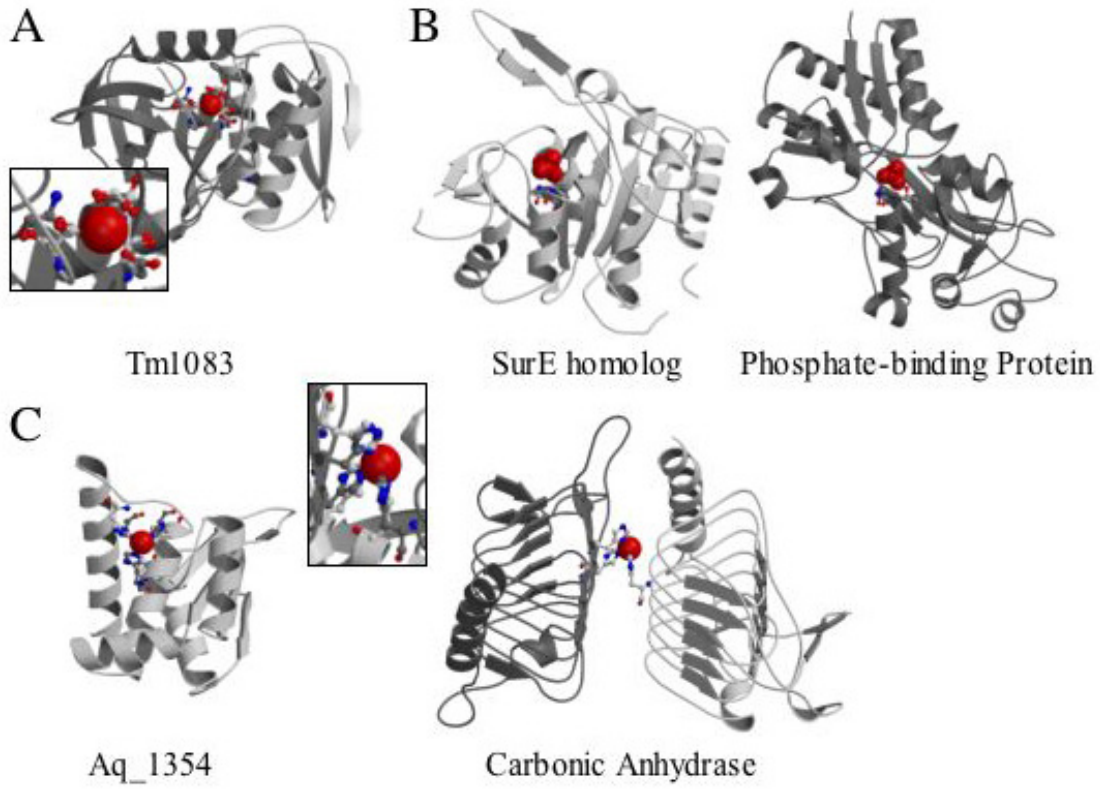



**Figure 3.10:** Functional similarities between different folds (details are as in Fig. 3.9). (A) Residues Asp16, Asp129 and Thr130 from the hypothetical protein (HP) Tm1083 (1j5u) match to a Ca-binding site in staphylococcal nuclease (1sty; E=$5x10^{-5}$). B) His115, His125 and His119 of HP Aq_1354 (1oz9) match to the zinc-binding active site of carbonic anhydrase (1thj; E=$9x10^{-3}$). (C) Ser106, Gly107 and Thr108 of the survival protein Ea (SurEa) homolog (1l5x) match to the phosphate-binding site of a phosphate-binding periplasmic protein (1a40; E=$2x10^{-4}$).

Phosphate site similarities also arise convergently. For example, the survival protein Ea (SurEa) has a site similar to the active site of a phosphate-binding periplasmic protein (E=$2x10^{-4}$; Figure 3.10C). Although SurEa is not liganded itself, a homolog (SurE; 1j9l) contains a vanadate ion ($VO_4^{3-}$) at the corresponding site. Conserved residues lining this surface lead to the protein being identified as a putative phosphatase site with a preferred specificity towards purine nucleotides (Mura *et al.*, 2003).

**Active Site in Ribose-5-Phosphate Isomerase**

The alternate Ribose-5-Phosphate Isomerase (RpiB/AlsB) catalyses the conversion of ribose 5-phosphate to ribulose-5-phosphate. Structure comparison confirmed the similarity to Rossman-fold proteins but did not reveal any insights into the reaction, though the authors used their knowledge about preferred binding sites, residue conservation and surface curvature (i.e. surface cavity) to locate a putative active site pocket (Zhang *et al.*, 2003). The best match for this protein in our study is the similarity between three residues that line one

side of this pocket and the substrate bound active site of phosphoglycerate mutase ($E=3x10^{-2}$; Figure 3.11). Although the similarity does not comprise the full active site pocket, it is useful for determining the location and specificity for phosphorylated ligands.

**A predicted Phosphotyrosine-binding site in an archaeal protein**

Structure comparison revealed that Mth1187 adopted a ferrodoxin-like fold, and the authors speculated that it might be a protein-protein interaction module (Tao *et al.*, 2003). They noted however that the residues lining the binding site of a sulfate ion showed enhanced conservation indicative of a functional site or a binding site for an unknown ligand. We found a highly significant similarity to the phosphotyrosine-binding sites in SH2 domains (1fyr; $E=3x10^{-4}$; Figure 3.11). The similarity includes residues contacting the sulfate ion in addition to others that contact the tyrosine ring. Indeed, a reverse search of the Mth1187 binding site against all phosphate/sulphate binding sites or against a representative set of complete structures (Hobohm & Sander, 1994a) finds no other significant similarities. Phosphotyrosine is thus an excellent candidate for the natural ligand. This is especially interesting, as tyrosine specific protein kinases and phosphatases have only recently been recognized to play important roles in prokaryotic and archaeal organisms and key proteins might still be unknown (Bakal & Davies, 2000; Kennelly, 2002; Kennelly, 2003; Shi *et al.*, 1998).

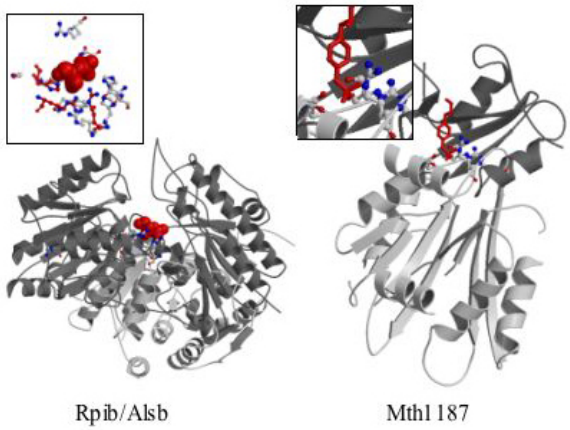



**Figure 3.11:** New functional convergences. (Left) Residues His10, Arg133 and Arg40 of Alternate Ribose-5-Phosphate Isomerase Rpib/Alsb (1nn4) match to the active site of phosphoglycerate mutase (1o98 and 1o99; 2-phosphoglycerate bound; $E=3x10^{-2}$). The inset shows the residues that line the putative active site pocket with the matching residues in red. (Right) Residues Ser17, Ser19, Arg1081, Arg1086 and Ser1094 of Mth1187 (1lxn) match to the phosphotyrosine-binding site of an SH2-domain (1fyr; $E=3x10^{-4}$).

## 3.2.4 Discussion

We have tested the applicability of functional site comparison on a large dataset of new proteins with unknown function. Many structures show similarities between functional residues to those solved previously, which can lead to functional hypotheses being tested further. The examples show a variety of situations, ranging from confirmation of a similarity

inferred by overall structural similarity to detecting a convergently evolved mode of ligand binding.

Overall, the results demonstrate how searches for similar functional sites complement those for similar folds. A combined strategy where both types of searches are used for structure-based functional annotation can help overcome problems inherent to each when applied separately. Even when structural alignment searches reveal fold similarities, active site comparison can highlight the presence (Sanishvili *et al.*, 2003) or absence (Wilson *et al.*, 2004) of an active site, and can sometimes resolve functional ambiguities (Lorentzen *et al.*, 2003). It can also help to identify "migrating" catalytically equivalent residues that are located on different parts of homologous structures (e.g. Todd *et al.*, 2001). Moreover, newly determined active sites can be sought in previously existing structures regardless of any similarity in overall fold.

The complementarity can also work in reverse: a similarity in fold as revealed by structural alignment can boost confidence in a marginally significant functional site match. This is particularly relevant for matches involving only a few residues that require too narrow geometrical constraints (i.e. small RMSD) to be distinguishable from noise (see Chapter 3.1) or those involving residues distorted by bound ligands. Functional site matches involving proteins of the same fold can be more believable even when the matches themselves are marginal.

Both approaches will benefit from the increasing number of functionally annotated protein structures. There are also recent efforts to catalogue active sites in structures based on studies of their function (Bartlett *et al.*, 2002; Porter *et al.*, 2004). These will increase the coverage, sensitivity and specificity of our searches or methods similar to ours. Investigating both types of similarities discussed here, while the number of structures and known functional sites grows, will also complete the picture of how nature evolves or reinvents proteins to perform different functions with a diverse array of ligands.

## *3.3 Structural Analysis of the Archaeal Class I FBP-Aldolase*

Fructose-1,6-bisphosphate aldolase (FBPA, EC4.1.2.13) catalyzes the reversible cleavage of fructose-1,6-bisphosphate into glyceraldehyde 3-phosphate and dihydroxyacetone phosphate and is central to the reversible Embden-Meyerhof-Parnas pathway (i.e glycolysis and glyconeogenesis) and the Calvin cycle. Two classes of FBPAs with different catalytic mechanisms are known: The eukaryotic class I enzymes use an active site lysine to form a Schiff-base with the substrate, whereas the bacterial class II FPBAs stabilize the intermediate with divalent metal ions (Lebherz & Rutter, 1969). Both adopt the common TIM-barrel fold and might share a common ancestor despite insignificant overall sequence similarity (Copley & Bork, 2000; Nagano *et al.*, 2002). Archaeal organisms appear to rely solely on a recently identified third class of *archaeal FBPAs* (*FBPA IA* or *aFBPA*). This family was identified as a divergent group, comprising of members from almost all sequenced archaeal organisms, some eubacteria but no eukaryotes (Gamblin *et al.*, 1990; Siebers *et al.*, 2001). Although they share the catalytic mechanism with the classical eFBPAs, they do not show any significant overall sequence similarity to this or other TIM-barrel superfamilies (Siebers *et al.*, 2001).

Esben Lorentzen from Ehmke Pohl's (EMBL-Hamburg) solved the structure for archaeal FBPA from *Thermoproteus tenax* (Tt) to 1.9Å resolution and I performed a detailed structural analysis with other TIM-barrel enzymes that allowed us to establish evolutionary links between the archaeal and the classical FBPAs, and the triosephosphate isomerases (TIMs) (Lorentzen *et al.*, 2003).

### 3.3.1 Overall Structure of FBPA IA

The Tt-FBPA forms a homo-decamer where each monomer adopts a TIM-barrel fold (Figure 3.12). Due to the wide range of reactions catalyzed and the low sequence similarity often observed between them, it is believed that the simple barrel architecture might have arisen multiple times during evolution. SCOP for example currently distinguishes 26 superfamilies, which may represent individual convergences although it has been suggested that many of them might share a common ancestor (Copley & Bork, 2000; Farber & Petsko, 1990; Nagano *et al.*, 2002).

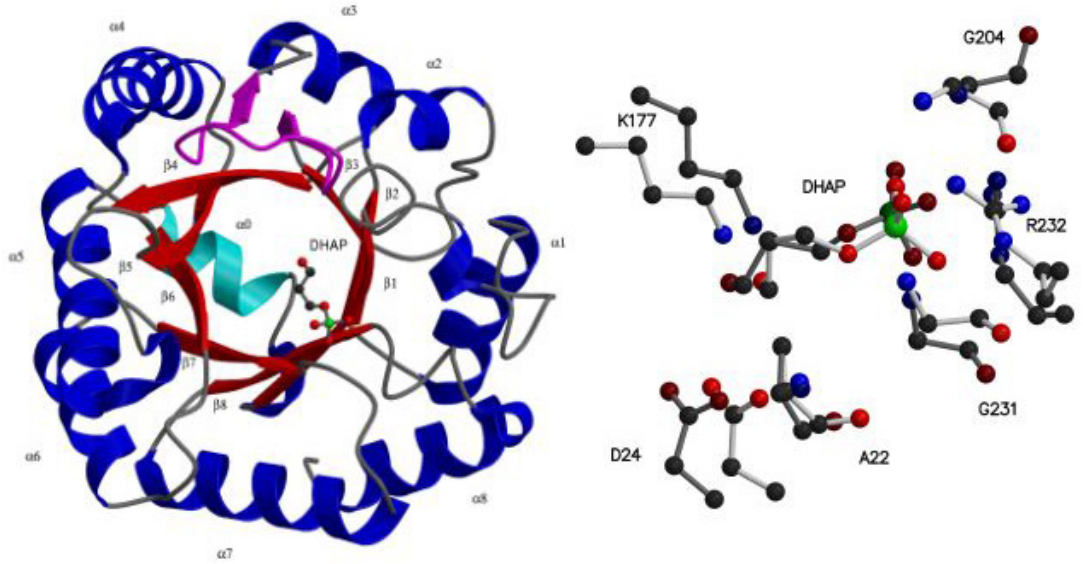



**Figure 3.12:** Structure of the archaeal FBPA from *Thermoproteus tenax* (Tt). (Left) The Tt-FBPA momomer adopts a TIM-barrel fold (helices shown in blue, strands in red). The substrate binding site (bound DHAP) and the unusual structural features (N-terminal helix α0, cyan; small β-sheet, purple; for details see text) are highlighted. (Right) Active sites of FBPA IA from Tt (light grey) and classical FBPA I from rabbit muscle (dark grey) after the active site residues, found to be equivalent by PINTS ($E=3x10^{-4}$), have been superimposed. Residues are labeled according to the Tt-FBPA sequence.

### 3.3.2 Comparison to the classical FBPA I

The classical FBPA I and aFBPA adopt the same fold and catalyze identical reactions but share no overall sequence identity. It was thus of particular interest to see if a detailed structural comparison could unravel their evolutionary relationship. I first performed a structural alignment of the Tt-FBPA subunit and the subunit of a human FBPA I (Dalby *et al.*, 1999) using STAMP (Russell & Barton, 1992). The structures superimpose with an RMSD of 1.9 Å over 250 residues and the sequence identity for the 154 residues that occupy equivalent structural positions is 13% (Figure 3.13). This value is not sufficient to infer homology directly (P=1; see Material and Methods and Murzin, 1993b). However, two unusual structural features are observed in both types of FBPAs (Figure 3.12). The first is the presence of an additional N-terminal helix, which precedes the first β-strand of the barrel. As seen for the classical FBPA I (Gamblin *et al.*, 1990) and for the KDPG aldolase (Mavridis *et al.*, 1982), this helix runs across the N-terminal part of the barrel and closes it. The second structurally equivalent feature is the insertion of a small two-stranded anti-parallel β-sheet between strand β3 and helix α3. This loop is involved in one of the dimer interfaces of the tetrameric classical FBPA I (Gamblin *et al.*, 1990). In Tt-FBPA, however, the loop is turned

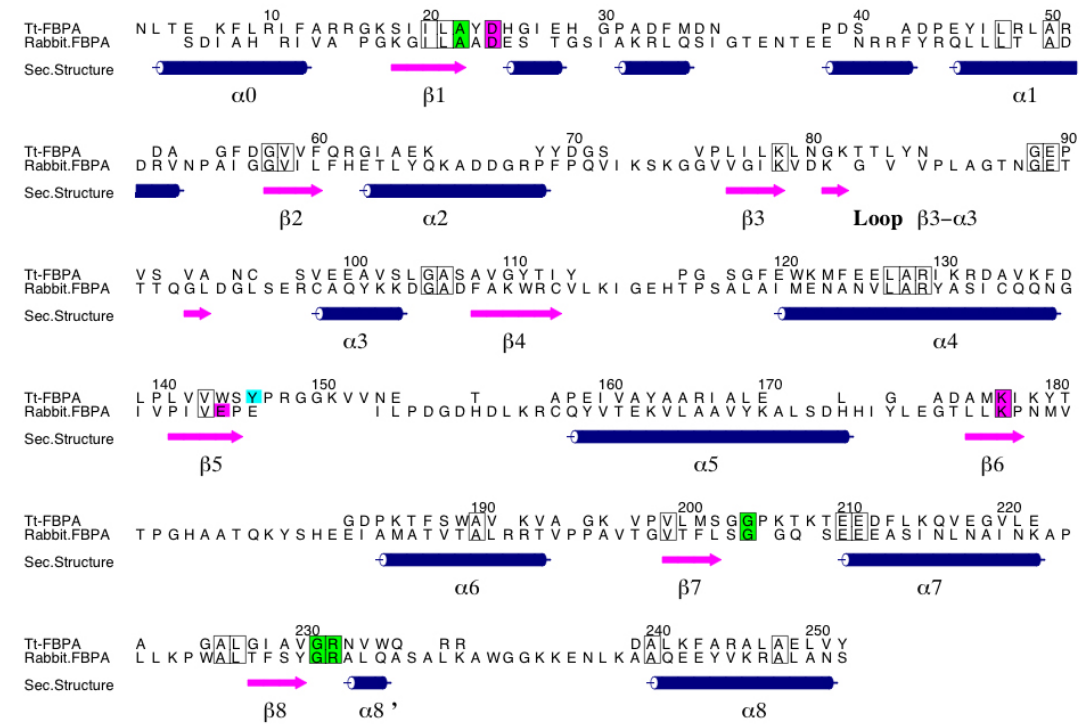



**Figure 3.13:** Structure-based sequence alignment of Tt-FBPA and rabbit muscle FBPA. The numbering and the secondary structure assignment are shown according to Tt-FBPA. Residues identical to both sequences are surrounded by boxes, conserved catalytic residues are colored in magenta, substrate binding residues in green, and the proposed catalytic proton donor Tyr-146 in cyan.

and moved about 10 Å from the position seen in the classical FBPA I. Different positions allow the same loop to be involved in dimer interactions in the classical FBPA I and in pentamer formation in Tt-FPBA. Such unusual features are the key to assigning ancient relationships to structures as in SCOP.

The strongest evidence for a common evolutionary origin between the archaeal type and the classical FBPA I is based on the comparison of the two active sites. PINTS detected six residues in a similar spatial arrangement in both aldolases with an RMSD of 1.1 Å and an E-value of $3x10^{-4}$ (Figure 3.12). The residues form equivalent structural substrate binding sites as seen in structures with covalently bound DHAP (Choi *et al.*, 2001; Lorentzen *et al.*, 2003) and furthermore occupy the same positions in the protein sequence (Figure 3.13). In contrast, no significant similarities between aFBPA and active site patterns from other TIM-barrel enzymes were found as expected based on the known function. The similarity extends beyond the core of catalytically active residues required for function but is not a general feature of all TIM-barrel enzymes. Such similarities are very unlikely to arise independently and we thus conclude that the classical and the archaeal FBPA I share a common ancestor.

Despite highly conserved active site residues, one important difference is found between FBPA I and aFBPA. A glutamic acid (Glu187 in rabbit FBPA corresponding to Trp144 in Tt-FBPA, see Figure 3.13), positioned at the end of strand β5 in all classical FBPA I, acts as an acid in the catalytic mechanism by donating a proton in the dehydration of the carbinolamine to form the imine (Choi *et al.*, 2001). The carboxyl is located 3.9 Å from the Schiff-base-forming carbon of the substrate (2.5-3.0 Å from the proton-accepting hydroxyl of the substrate). Of the 27 genes identified to date as belonging to the FBPA IA family (Siebers *et al.*, 2001), only two encode a glutamic acid at this position. The other 25 genes code for a hydrophobic residue, which cannot participate in proton donation. In the structure of the liganded Tt-FBPA, Tyr146 is positioned with a distance of 3.7 Å from its hydroxyl to the Schiff-base-forming carbon of the substrate and is the only plausible candidate for a proton donor. This residue is a tyrosine in 20 of the archaeal FBPA sequences and we propose it to be the catalytic proton donor in these proteins. As the proton donor differs with respect to type of amino acid as well as sequence position between FBPA I and FBPA IA and within the FBPA IA family it is reasonable to assume that it is of recent evolutionary origin. Other important catalytic and substrate binding residues seem to be of more ancient origin. The active site lysine (Lys177 in Tt-FBPA) is conserved in all FBPA Is as well as IAs (FBPA I/IA) and is therefore likely to have been present in an ancestor protein common to all FBPA I/IAs. The phosphate moiety of the substrate is tightly bound by the main-chain nitrogens of the two glycines in loops β7-α7 and β8-α8' (Gly204 and Gly231) and by the side-chain of an arginine (Arg232). As these three residues are conserved in all FBPA I/IAs identified to this date, they probably represent an ancestral phosphate-binding site. In addition, Asp24, which acts as a base in the catalytic mechanism (Choi *et al.*, 2001; Wilmanns *et al.*, 1991), is conserved in all FBPA I/IAs and might be of ancestral origin. Ala22 is found to make hydrophobic contacts to the carbon atoms of the substrate, but is not conserved in all FBPA IAs.

In conclusion, we have identified and described similar active sites for the archaeal type and the classical FBPA I. This common site contains many of the catalytic and substrate binding residues conserved with respect to structure as well as sequence. It is thus very likely that the classical and the archaeal type FBPA I share a common evolutionary origin.

### 3.3.3 TIM and the aldolase superfamilies are homologous

Triosephosphate isomerases (TIMs) and aldolases are usually grouped into different superfamilies (e.g. by SCOP), but the presence of a common phosphate binding site and recent results from stepping-stone or transitive PSI-BLAST approaches provided some indication for divergent evolution (Banner *et al.*, 1975; Copley & Bork, 2000; Nagano *et al.*, 2002).

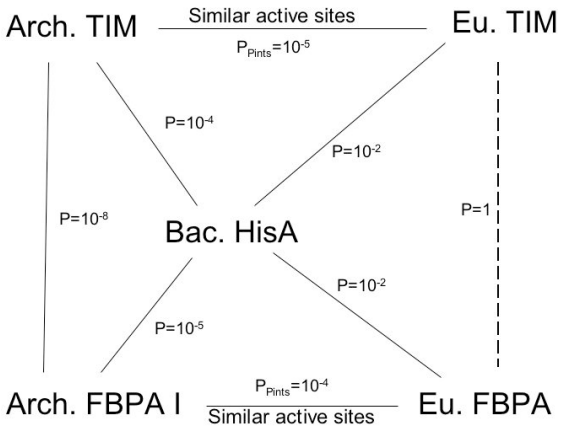



**Figure 3.14:** Summary of the evolutionary links between different families of TIM barrel proteins discussed (Arch., archaeal; Bac., bacterial; Eu., eukaryotic; broken line: link in eukaryotes not significant on its own).

When we compared human FBPA Is and TIMs, we did not find convincing structural evidence for a common ancestor: Only 10 out of the 126 residues (8%) found to be structurally equivalent are identical between human muscle FBPA (Gamblin *et al.*, 1990) and TIM (Mande *et al.*, 1994), which is not sufficient to infer homology reliably (P=1). However, as it has been suggested previously that hyperthermophilic archaea have slower evolutionary rates than bacteria or eukaryotes (Kollman & Doolittle, 2000; Pace, 1991; Woese, 1987), we wondered whether we could establish a evolutionary link between the two superfamilies using the archaeal structures as a bridge.

Indeed, we found that the structure of the Tt-FBPA monomer is, from a structural perspective, closer to archaeal TIMs (from *P. woesei* (Walden *et al.*, 2001) and *T. tenax* (Tt-TIM; Lorentzen, submitted)) than to any other structure in the PDB: the structures of Tt-TIM and Tt-FBPA superimpose with an RMSD of 1.7 Å and 30 of the 149 structurally equivalent residues are identical (20%). Except for the common phosphate-binding site, no catalytic or substrate binding residues are shared by the two proteins. Most are hydrophobic residues in the cores of the proteins and involved in similar hydrophobic contacts (e.g. Leu127, Val143, Ala168 and Leu171; numbering according to the Tt-FBPA sequence), or other structurally important residues such as salt bridges (e.g. Arg129 that contacts Glu126 in Tt-FBPA or Asp102 in Tt-TIM). Although this similarity is not detected by sequence comparisons but is only found after structure-based alignment, it provides strong evidence for a common origin

between the aldolase and TIM superfamilies as the probability to observe it between unrelated proteins is small ($P=3.7x10^{-8}$, Murzin, 1993b). In addition, Tt-TIM also shares the extended loop with a small anti-parallel β-sheet between β3 and α3 that is also seen in the classical and the archaeal FBPA I (see above).

The possibly slower evolutionary rate is also indicated by the fact that both Tt-TIM and Tt-FBPA are more reminiscent of the HisA protein from eubacteria (Lang *et al.*, 2000) than the eukaryotic proteins. Although still controversial, the HisA protein has been suggested to be reminiscent of a putative common ancestor for all TIM-barrel enzymes (Copley & Bork, 2000). After structural alignment, HisA shows a highly significant sequence identity to both Tt-FBPA (22 out of 123 structurally equivalent residues (18%); $P=1.4x10^{-5}$) and Tt-TIM (23/152 residues (15%); $P=1.2x10^{-4}$). In contrast, the similarities for the human enzymes are much less pronounced (11% corresponding to $P=1x10^{-2}$; Figure 3.14). An indication for the common evolutionary origin of HisA and Tt-FBPA is the identical positions of essential catalytic residues at the end of β-strand 1 and 5 in both enzymes.

Figure 3.14 summarizes the evolutionary links we were able to establish between TIMs and FBPAs using the significant structural similarity between the archaeal enzymes, which is not apparent when comparing sequences or structures of the eukaryotic enzymes. We can also link the archaeal enzymes from both superfamilies to HisA that has been placed near the root of the evolutionary tree of TIM-barrel proteins. The greater degree of similarity found in archae furthermore supports the hypothesis that these organisms undergo fewer evolutionary changes and that the last common ancestor might have been a hyperthermophilic or thermophilic organism (Pace, 1991).

## 3.4 The PINTS Server

One of the reasons why active site comparisons are far less frequently used by structural biologists than structural alignment (e.g. Dali) might be the lack of an easy-to-use Internet service. As I wanted PINTS to be most useful, I implemented a web interface that is accessible on the Internet (http://pints.embl.de) (Stark & Russell, 2003). The server offers several databases of complete protein structures or patterns and contains detailed information about its uses (*Info*, *Help* and *FAQs* (*Frequently Asked Questions*) pages). We currently allow for three types of searches described in the following:

### 3.4.1 Searches

We distinguish between three types of searches as a single, all-encompassing, all-against-all search would have a critical effect on the statistics (see Chapter 3.1). In addition, patterns in our databases are more directly related to an individual molecular function (e.g. catalytic activity or ligand-binding) than the overall protein.

**Protein versus pattern database:** For a new protein structure (e.g. from structural genomics projects), hints about function or the location of a functional site can come from searches against databases of patterns likely to be of functional importance (see Chapter 3.2).

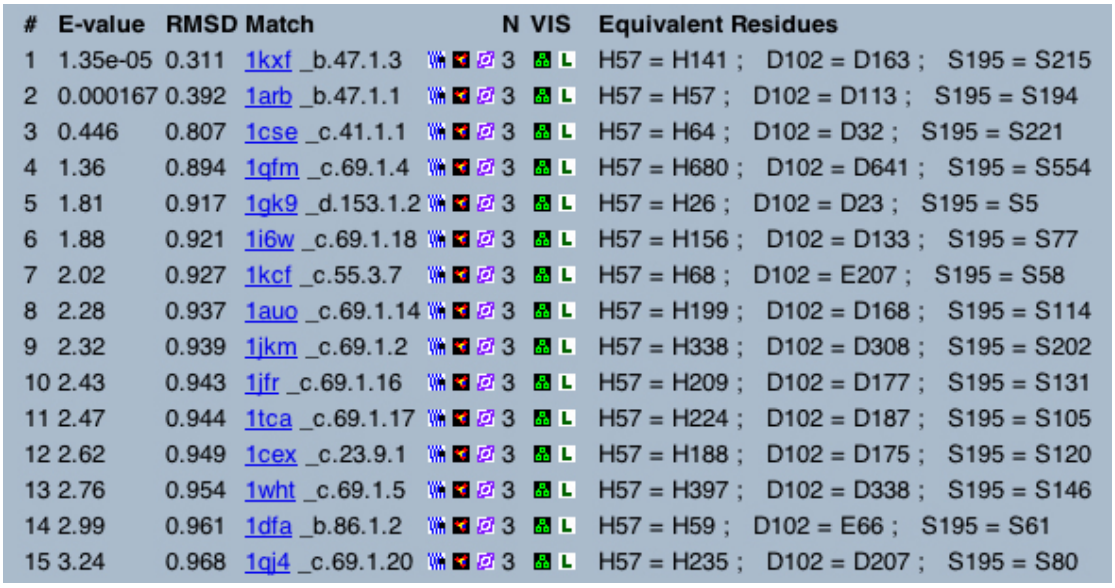
**Pattern versus protein database**: The recurrence of a known functional or an interesting new pattern in other structures can suggest common properties. We therefore allow patterns of up to 100 residues to be compared to protein databases (i.e. containing complete structures) at different levels of redundancy (see Materials and Methods) or the pattern databases above.

**Pairwise comparison**: We also allow a pairwise comparison of two structures. This can suggest the molecular basis for the common biochemical or cellular function (e.g. catalytic activity, specific binding characteristics, see Chapter 3.1.7).

For all searches, the user can either upload files in PDB text format or specify the four-letter PDB identifiers for publicly available structures. The query can be restricted to specific residues (i.e. a pattern) within the submitted or specified PDB file by an easy syntax.

### 3.4.2 Output

For all searches, matches up to a user-defined E-value maximum and that contain at least three residues are reported (Figure 3.15). We allow for partial matches to be detected, which is particularly important if an active site is not fully understood or when the similarity may not cover the whole of a pre-defined site. Automatically or manually annotated patterns (such as the *ligand binding sites* or the *SITE annotations* database entries in PINTS) often contain additional residues that are not absolutely required for function.



| # | E-value | RMSD | Match | N | VIS | Equivalent Residues |
|---|---|---|---|---|---|---|
| 1 | 1.35e-05 | 0.311 | 1kxf _b.47.1.3 | 3 | L | H57 = H141 ; D102 = D163 ; S195 = S215 |
| 2 | 0.000167 | 0.392 | 1arb _b.47.1.1 | 3 | L | H57 = H57 ; D102 = D113 ; S195 = S194 |
| 3 | 0.446 | 0.807 | 1cse _c.41.1.1 | 3 | L | H57 = H64 ; D102 = D32 ; S195 = S221 |
| 4 | 1.36 | 0.894 | 1qfm _c.69.1.4 | 3 | L | H57 = H680 ; D102 = D641 ; S195 = S554 |
| 5 | 1.81 | 0.917 | 1gk9 _d.153.1.2 | 3 | L | H57 = H26 ; D102 = D23 ; S195 = S5 |
| 6 | 1.88 | 0.921 | 1i6w _c.69.1.18 | 3 | L | H57 = H156 ; D102 = D133 ; S195 = S77 |
| 7 | 2.02 | 0.927 | 1kcf _c.55.3.7 | 3 | L | H57 = H68 ; D102 = E207 ; S195 = S58 |
| 8 | 2.28 | 0.937 | 1auo _c.69.1.14 | 3 | L | H57 = H199 ; D102 = D168 ; S195 = S114 |
| 9 | 2.32 | 0.939 | 1jkm _c.69.1.2 | 3 | L | H57 = H338 ; D102 = D308 ; S195 = S202 |
| 10 | 2.43 | 0.943 | 1jfr _c.69.1.16 | 3 | L | H57 = H209 ; D102 = D177 ; S195 = S131 |
| 11 | 2.47 | 0.944 | 1tca _c.69.1.17 | 3 | L | H57 = H224 ; D102 = D187 ; S195 = S105 |
| 12 | 2.62 | 0.949 | 1cex _c.23.9.1 | 3 | L | H57 = H188 ; D102 = D175 ; S195 = S120 |
| 13 | 2.76 | 0.954 | 1wht _c.69.1.5 | 3 | L | H57 = H397 ; D102 = D338 ; S195 = S146 |
| 14 | 2.99 | 0.961 | 1dfa _b.86.1.2 | 3 | L | H57 = H59 ; D102 = E66 ; S195 = S61 |
| 15 | 3.24 | 0.968 | 1qj4 _c.69.1.20 | 3 | L | H57 = H235 ; D102 = D207 ; S195 = S80 |

**Figure 3.15:** Example of a PINTS Server results page. Shown are the top 15 matches (and their SCOP classification) for a search with the trypsin catalytic triad (1mct, 57, 102, 195) against one representative of each SCOP family. The three buttons are links to NCBI-Entrez, SCOP and PDBsum and the two buttons in the VIS column link to the superimposed coordinates as explained in the text.

Matches are ranked by their statistical significance and the equivalent residues and associated RMSDs are provided, as are cross-references to useful Internet resources: SCOP (Murzin *et al.*, 1995a), NCBI-Entrez (http://www.ncbi.nlm.nih.gov/Entrez/), and PDBsum (Laskowski, 2001; Laskowski *et al.*, 1997; Luscombe *et al.*, 1998). For visual inspection with RasMol (Sayle & Milner-White, 1995), we provide superimposed coordinates for both the matched patterns alone (i.e. the equivalent residues) and within the whole protein context. For an example of the PINTS results page, see Figure 3.15. Search results are kept for 8 days and can be retrieved by an identifier, E-mail or IP address as preferred.

### 3.4.3 Settings

We exclude hydrophobic aliphatic residues (Ala,Ile,Leu,Met,Val) from the search as proteins often contain hydrophobic centres that are structurally very similar (see Chapter 3.1.8) and would lead to false positive matches comparable to low complexity regions in sequence comparison. We regard only the amide (Asn,Gln) and acidic (Asp,Glu) residues to be similar enough to be equivalenced generally (see the corresponding PINTS definition file (Table 2.1, Column 3). The search parameters are currently restricted to standard settings that we know would be applicable to a wide variety of different submissions (maximum pattern diameter 15 Å, distance tolerance during the depth-first search 3 Å, minimum and maximum number of residues per pattern 3 and 100, respectively).

As PINTS was specifically designed to find spatial patterns in non-homologous proteins, its algorithm suffers heavily if two very similar proteins are compared. Especially in database searches, most of the search time would be spent on database entries with high sequence similarity and the output would be swamped with many patterns from these. The PINTS server therefore removes all database entries with high sequence similarity to the query (i.e. detected by BLAST with high confidence $E \leq 1x10^{-10}$) from the search and reports them separately.

### 3.4.4 PINTS-Weekly

In addition to the service above for individual structures, we now constantly monitor the PDB for new functional similarities. With each weekly update of the PDB, we compare all new structures to our pattern databases. Updates to the PDB often contain structures that are either slight variants (e.g. different bound small molecules, mutants, etc.) or close homologs of proteins already present in the database. We thus classify the structures into two categories (in addition to *All Structures*) to facilitate browsing or finding formerly unknown similarities. *Structural Genomics* are structures that are labeled as "Structural Genomics" or "Unknown Function" by the authors of the PDB file. *Unique Sequences* are those that do not match any previously known structure using BLAST (E-value $\leq 10^{-20}$, sequence identity $\geq 80\%$, and length difference of $\leq 90\%$ or $\leq 50$ residues. This service is available at http://www.russell.embl.de/pints-weekly (Stark, *et al.*, submitted).

### 3.4.5 Access Statistics

We announced the PINTS server on the PDB mailing list (pdb-l) on January, 29th 2003 (http://www.rcsb.org/pdb/lists/pdb-l/200301/000412.html). Since then, our site has been accessed nearly 50,000 times by more 3300 different non-EMBL users identified by their IP-addresses. There were on average 130 unique searches per month or nearly 2000 altogether (Figure 3.16).

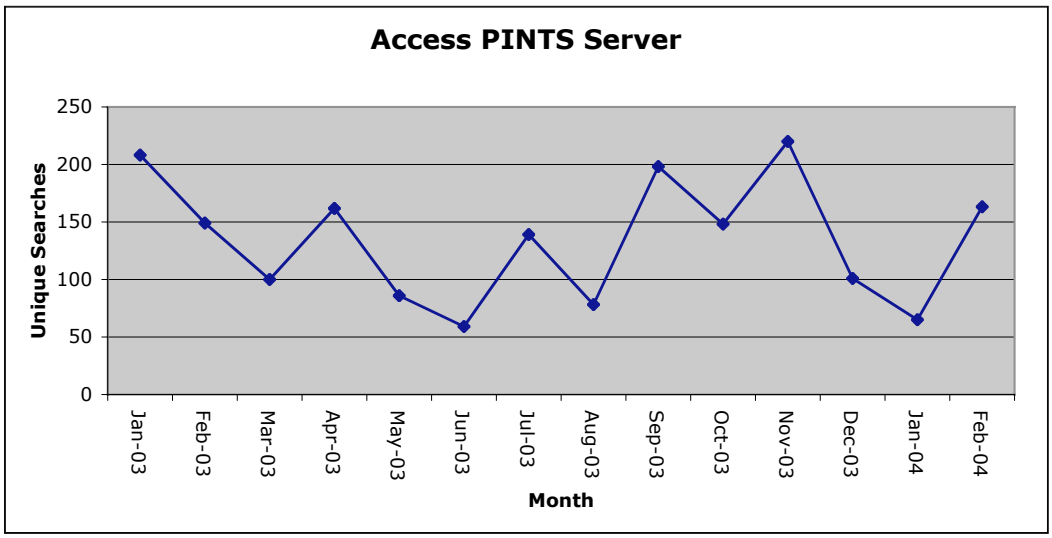



**Figure 3.16:** Access statistics for the PINTS-Server. Shown are unique searches from non-EMBL users between January 2003 and February 2004. The announcement on the PDB-mailing list (01/2003) and the publication (06/2003, Stark, *et al.*, 2003) are seen as maxima, but the server usage was highest from September-November 2003.

## *3.5 miRNA Target Prediction*

Although miRNAs are thought to play major regulatory roles in all higher organisms, not much is known about their functions or the target genes they regulate. The limited sequence complementarity between miRNAs and their targets means that target prediction is difficult and requires careful statistical evaluation (see Introduction). In close collaboration with Julius Brennecke in the group of Stephen M. Cohen (EMBL-Heidelberg), I developed a computational screen for miRNA targets in *Drosophila*. I examined properties of valid miRNA targets such as their structures or conservation that I then used to predict targets for miRNAs. I showed that the method identifies all known miRNA targets and predicted new targets that Julius Brennecke validated experimentally. An earlier version of the method was used to predict the *bantam* target *hid* (see Chapter 3.6.3 and Brennecke *et al.*, 2003). The screen presented in the following section has been published:

**A. Stark**, J. Brennecke, R.B. Russell, S. M Cohen; Identification of *Drosophila* miRNA Targets. *PLoS Biology*, 1(3), E60, 2003.

### 3.5.1 Conserved 3' UTR Database

For each of the validated miRNA/target pairs, functional target sites are located in the 3' untranslated region (UTR) of the mRNA and are conserved in the 3' UTRs of the homologous genes from related species (Abrahante *et al.*, 2003; Brennecke *et al.*, 2003; Lee *et al.*, 1993; Lin *et al.*, 2003; Moss *et al.*, 1997; Reinhart *et al.*, 2000; Wightman *et al.*, 1993). I used pairwise comparison of the 3' UTRs of orthologous genes in related genomes to identify conserved 3' UTR sequences. Figure 3.17A shows the resulting pattern of 3' UTR conservation for the known targets in worms and flies. The vast majority of miRNA target sites (red bars) are located in blocks of conserved sequence (white blocks). Figure 3.17B shows cross-genome conservation of these miRNA target sites. A striking pattern of uninterrupted conservation emerges at the end of the target sequences that pair with the 5' end of the miRNAs.

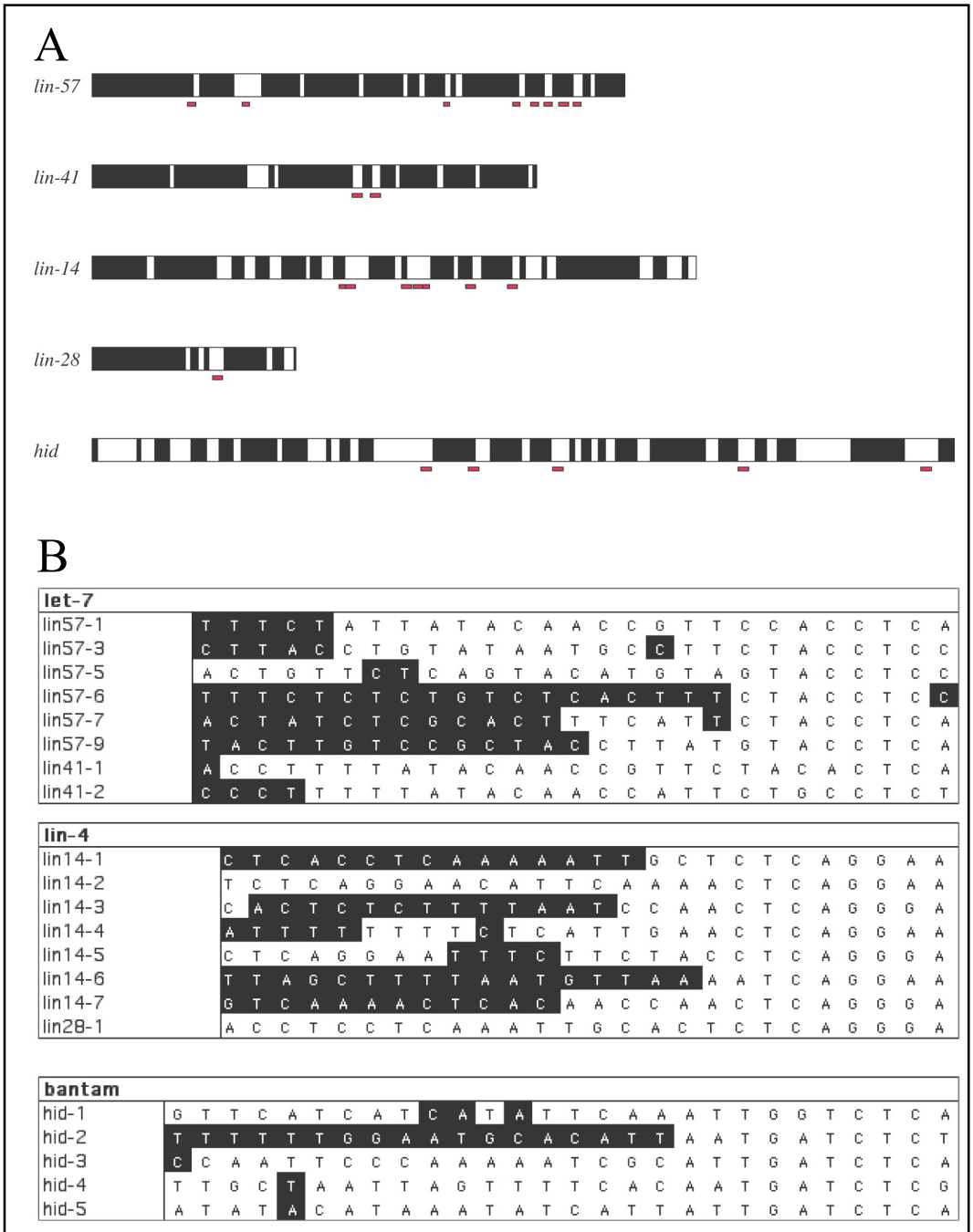



**Figure 3.17**: Features of known miRNA targets. (A) miRNA target sites (red bars) are generally in conserved regions of 3' UTRs of known miRNA targets (white background). Comparison was done according to conditions used to construct the 3' UTR database (see Materials and Methods) between *D. melanogaster* and *D. pseudoobscura* (for *hid*), or between *C. elegans* and *C. briggsae* (otherwise). Most UTRs contain multiple predicted target sites and function for individual sites not been tested generally. (B) Sequence conservation within the predicted miRNA binding sites that are conserved (shown is miRNA length plus 5 nts). All residues that pair with positions 2-8 of the miRNA are identical in the conserved sites in both genomes compared (white background).

To permit genome-wide searches for targets of *Drosophila* miRNAs, a conserved 3' UTR database was prepared by comparison of *D. melanogaster* and *D. pseudoobscura* 3' UTRs. As very few 3' UTRs are defined by cDNA sequence data in *D. pseudoobscura*, I used genomic sequence following the last exon of the *D. pseudoobscura* gene as the orthologous UTR (see Materials and Methods). Last exons were reliably detected in *D. pseudoobscura* for about two-thirds of *D. melanogaster* genes. On average 22% of the *D. melanogaster* 3' UTR sequence is conserved in the predicted *D. pseudoobscura* 3' UTR. Much of this reflects isolated blocks of very high conservation interspersed among less conserved sequence. Use of conserved 3' UTRs reduces the expected number of sequence matches that would occur at random by 4 to 5 fold in relation to full-length 3' UTRs, and several fold further compared to the full transcriptome. We considered using the *Anopheles gambiae* genome to extend the cross species comparison. Although genome annotation identifies orthologs for two-thirds of *D. melanogaster* genes (Zdobnov *et al.*, 2002) I was unable to identify the last *D. melanogaster* exon for approximately half of these and therefore chose not to require conservation in *Anopheles*, but to use it as an additional level of validation for predicted *Drosophila* targets where possible.

### 3.5.2 Screening strategy

We have adopted a two step approach to target identification that combines a sensitive sequence database search with an RNA folding algorithm to evaluate the quality of the RNA duplex formed between the miRNA and its predicted targets. We examined the known target sites for *lin-4*, *let-7* and *bantam* for common features. All of these sites showed better complementarity to the 5' end of the miRNA, with no obvious common features elsewhere (Figure 3.18 A and B). There were few sequence mismatches or G:U base pairs in the alignment of the first eight residues at the 5' end of the miRNA. I used HMMer (Eddy, 1998) to search for sequences complementary to the first eight residues of the miRNA, allowing for G:U mismatches. Where possible the corresponding sites were also identified in the *D. pseudoobscura* 3' UTR and the sites from both genomes were considered, since the regions outside of the sequence match can vary between the two organisms, leading to difference in subsequent steps (see below).

The identified sequences were extended to the length of the miRNA plus five residues to allow for bulges and were evaluated for their ability to form energetically favorable RNA-

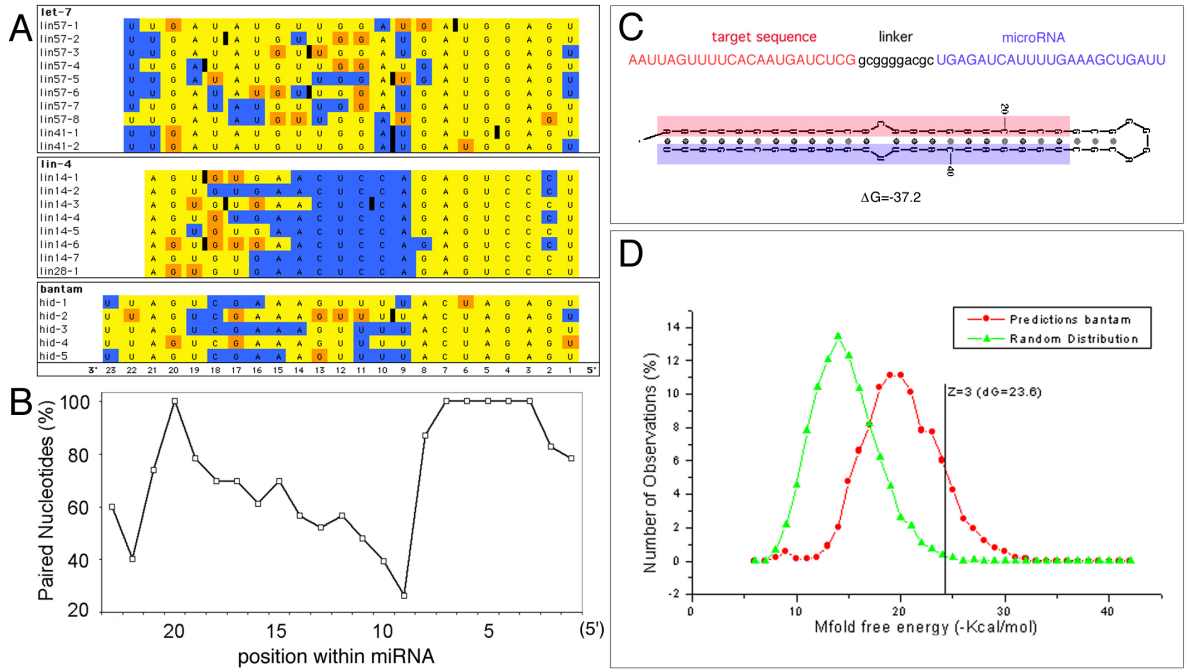



**Figure 3.18:** miRNA target prediction strategy. (A) *let-7*, *lin-4* and *bantam* miRNA sequences showing the pattern of base pairing to their known targets (yellow: conventional base pair; orange: G:U base pair; blue: mismatch, black bars: insertions in the target; all sequences at the length of the miRNA). (B) Quantitation of the data from (A): the 5' ends of the miRNA are always well paired to the target, suggesting searches for complementarity to the first 8 residues of the miRNA. (C) Graphical representation of the Mfold output for *bantam* and a target site from the 3' UTR of its target *hid*. To use Mfold it is necessary to join the predicted target site (red) and the miRNA (blue) into a single sequence using a hairpin-forming linker sequence. In this example the target sequence and the miRNA are the same length, so the additional 5 nt in the tail of the predicted target sequence are not shown. (D) Plot of the Mfold free energy distribution for 10,000 random sequences (green) and for predicted targets of the bantam miRNA (red). X-axis: DG calculated for each site by Mfold.

RNA duplexes with the miRNA using Mfold, which combines knowledge of known RNA structures with thermodynamic parameters, such as those involved in base-pairing to evaluate the free energy of folding (ΔG; Mathews *et al.*, 1999; Zuker *et al.*, 1999). Mfold requires a single linear sequence as input, so each predicted target was linked to the miRNA using a standard hairpin-forming linker sequence (GCGGGGACGC). An example of the Mfold output is shown in Figure 3.18C for the top scoring *bantam* miRNA target site that we had previously identified in the 3' UTR of *hid* (Brennecke *et al.*, 2003).

The Mfold free energy of folding (ΔG) was determined for each predicted target, which allows predicted sites to be ranked according to ΔG. However, ΔG depends on miRNA length and GC content, so it is not possible to distinguish systematically real targets from random matches using the raw ΔG score, or to compare different miRNAs. Instead, we calculated Z-scores as a measure of non-randomness, with an average random site scoring Z=0. Figure

3.18D shows the distribution of folding energies for predicted targets of the *bantam* miRNA compared to 10,000 randomly selected target sequences.

Most of the previously validated targets have more than one predicted miRNA-binding site in their 3' UTRs. The use of Z-scores allows us to add the scores of several sites within one UTR by selecting only those scores that are different from background matches. This is not possible with ΔG alone because even average random matches have favorable energy values (Figure 3.18D) and the sum of several average random matches in a UTR could score better than a single true site. We selected Z=3 as a cutoff value as folding energies of more than three standard deviations above the mean (Z≥3) are expected to occur for only 0.3% of random matches. Use of a higher Z-score increases the likelihood that predictions are correct, but also increases the risk of missing out contributions from real sites of lower folding energy. For example, only three of the five conserved *bantam* sites previously identified in the *hid* 3' UTR score Z≥3 (with the best site at Z=7.4). We evaluated our predictions by the best single site in the 3' UTR ($Z_{max}$) and by the sum of sites with $Z \geq 3$ ($Z_{UTR}$).

### 3.5.3 Tests with previously validated targets

Table 3.1 summarizes the performance of the method in predicting the known targets of the *C. elegans* miRNAs *lin-4*, *let-7* and the *Drosophila* miRNA *bantam*. The *Drosophila hid* gene was ranked first of all predicted *bantam* targets sorted by the single best site ($Z_{max}$) or by the sum of sites ($Z_{UTR}$). All of the known targets of *lin-4* and *let-7* were found when their 3' UTRs were added to the *Drosophila* 3' UTR database. Like *hid*, the *let-7* target *lin-57* ranked near the top of the list by both measures: *lin-57* ranked first by $Z_{UTR}$ due to several predicted sites with Z≥3 and its best single site ranked in position 2 (Z=6.8). *C. elegans lin-14* was predicted to contain multiple *lin-4* sites (Lee *et al.*, 1993; Wightman *et al.*, 1993). Three of these scored Z≥3. *lin-14* was ranked first when the list of predicted *lin-4* targets was sorted for $Z_{UTR}$ although the best single site in lin-14 placed it in position 20 (Z=4.3). The *lin-4* target *lin-28* and the *let-7* target *lin-41* ranked highly when the lists were sorted for the best single site, but ranked lower when multiple sites were combined because they had few high-scoring sites. The *Drosophila* homolog of *lin-41*, *dpld*, also ranked high among *let-7* targets (Z=5.6, see below). We compared our results with previous target predictions from the literature that have not been experimentally validated (Table 3.1). Our screen supports some of them (e.g. *let-7* regulating *lin-14*), but we consider others unlikely because they rank very low on their

| miRNA/target pair | ΔG | $Z_{Max}$ | Rank $Z_{Max}$ | # Z≥3 | Rank $Z_{UTR}$ | References |
|---|---|---|---|---|---|---|
| ***Confirmed Pairs*** | | | | | | |
| *lin-4/lin-14* | -29.9 | 4.3 | 20 | 3 | 1 | (Wightman *et al.*, 1993) |
| *lin-4/lin-28* | -30.9 | 4.6 | 8 | 1 | 15 | (Moss *et al.*, 1997) |
| *let-7/lin-41* | -32.3 | 6.4 | 3 | 2 | 20 | (Reinhart *et al.*, 2000) |
| *let-7/lin-57 (hbl-1)* | -33.4 | 6.8 | 2 | 14 | 1 | (Abrahante *et al.*, 2003; Lin *et al.*, 2003) |
| *bantam/hid* | -37.4 | 7.4 | 1 | 3 | 1 | (Brennecke *et al.*, 2003) |
| ***Predicted Pairs*** | | | | | | |
| *lin-4/lin-41* | -28.9 | 4.0 | 32 | 1 | 36 | (Slack *et al.*, 2000) |
| *lin-4/lin-57* | -21.6 | 1.7 | 361 | 0 | - | (Abrahante *et al.*, 2003; Lin *et al.*, 2003) |
| *let-7/lin-14* | -35.1 | 7.2 | 1 | 13 | 2 | (Reinhart *et al.*, 2000) |
| *let-7/lin-28* | -20.6 | 2.8 | 861 | 0 | - | (Moss & Tang, 2003) |
| *miR-13a/hb* | - | - | - | 0 | - | (Abrahante *et al.*, 2003) |
| *miR-4/hb* | - | - | - | 0 | - | (Abrahante *et al.*, 2003) |
| *miR-3/hb* | - | - | - | 0 | - | (Abrahante *et al.*, 2003) |
| *miR-11/HLHm8* | -29.4 | 4.7 | 27 | 1 | 46 (pred. UTR) | (Lai, 2002) |
| *miR-4/m4* | -21.5 | 2.1 | 272 | 0 | - | (Lai, 2002) |
| *miR-7/HLHm3* | -37.3 | 7.0 | 2 | 1 | 16 | (Lai, 2002) |
| *miR-7/Tom* | -34.5 | 6.1 | 5 | 2 | 1 | (Lai, 2002) |
| *miR-14/Drice* | - | - | - | 0 | (not conserved) | (Xu *et al.*, 2003) |

**Table 3.1:** Assessment of predictions for known and predicted miRNA Targets (see text for details). Confirmed pairs are experimentally validated, predicted pairs are predicted in the cited reference. The predicted pairs *let-7/lin-14*, *lin-4/lin-41*, *miR-11/HLHm8*, and the two *miR-7* targets seem plausible whereas *miR-14/Drice* and predictions for *hunchback* (*hb*) seem unlikely according to our method.

lists or have no sites of Z≥3 (e.g. *let-7* and *lin-28* or *miR-4* and *m4*. The predicted *miR-14* target *Drice* (Xu *et al.*, 2003) is unlikely to be valid because the site is not conserved in the predicted *Drice* 3' UTR from *D. pseudoobscura*.

This analysis shows that all known targets were detected and ranked among the top scoring predictions in genome-wide searches. This suggests that other valid targets should rank among the small number of best predictions that can be tested experimentally. Of particular interest were three miRNAs for which we predicted clusters of functionally related targets: *miR-7*, the *miR-2* family and *miR-277* (Table 3.2 and Table 3.3). Clustering of top-scoring sites in a group of related genes is likely to be significant when it arises from an unbiased genome-wide analysis. *miR-7* and *miR-2* were selected for target validation.

### 3.5.4 *miR-7* regulates *Notch* targets

Among the top ten predictions for *miR-7*, we found *Enhancer of split* (*E(spl)*) and *Bearded* (*Brd*) complex genes (Figure 3.19A). *HLHm3* encodes a basic-helix-loop-helix (bHLH) transcriptional repressor; *Tom* and *m4* encode *Brd* family proteins. The bHLH repressor *hairy*

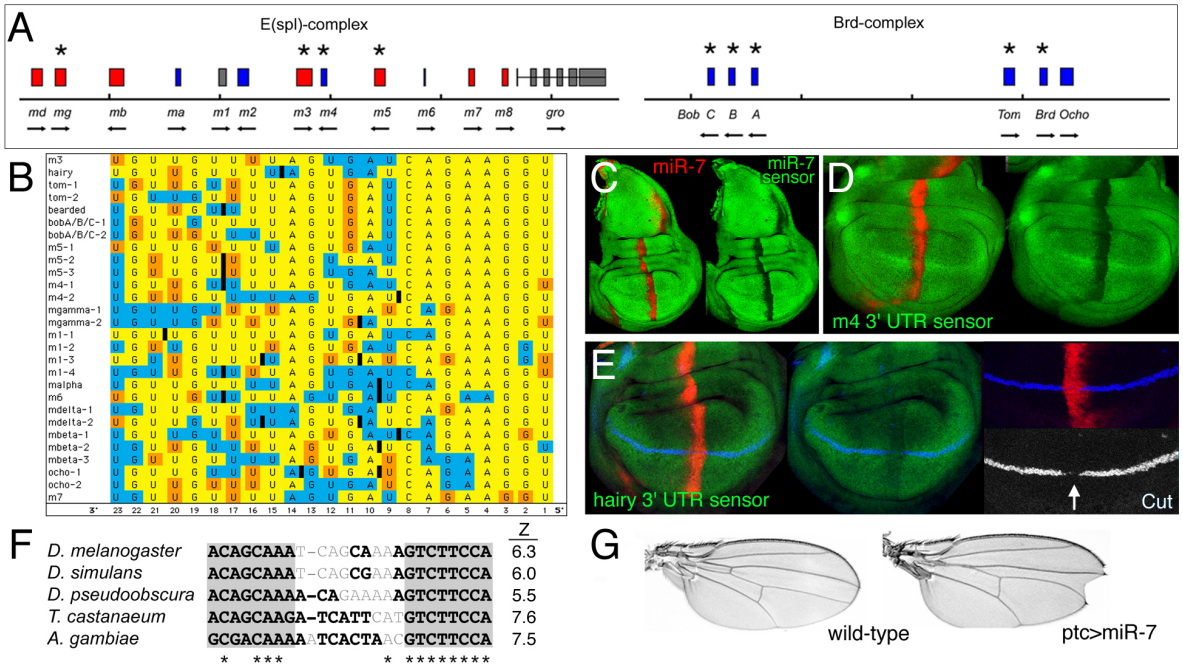



**Figure 3.19:** Experimental validation of *miR-7* targets. (A) Schematic representation of the *E(spl)* and *Brd* gene complexes highlighting bHLH-type transcriptional repressors (red) and Brd-type proteins (blue). Black asterisks indicate likely valid *miR-7* targets with no mismatch in the first 8 residue. (B) *miR-7* miRNA sequence showing the pattern of base pairing with target sites in *E(spl)* and *Brd* complex genes ranked according to folding energy (details as in Figure 3.18). (C-G) Experimental validation (J. Brennecke, see Stark *et al.*, 2003a for details) (C) Expression of the *miR-7* sensor transgene (both panels, green) is lost when *miR-7* is present (left, red). Right panel shows the *miR-7* sensor alone. (D, E) Expression of the *m4* 3' UTR and hairy 3' UTR sensor transgenes (green) are down-regulated by *miR-7* (red). Cut protein (top: blue, botton: white) was down-regulated in *miR-7* expressing cells. (F) Alignment of *miR-7* target sites in the 3' UTRs of *hairy* from several species (asterisks: identical nucleotides (nts); black type: nts paired to the miRNA; grey shading: miRNA binding site conserved in all species. (G) Cuticle preparations of a wild type adult wing and a wing expressing *miR-7* under *ptc-Gal4* control in the region between veins 3 and 4. Note the notching of the wing, and the reduction of the region between veins 3 and 4 leading to partial fusion proximally. The size of the posterior compartment was increased apparently to compensate for reduction of the vein 3-4 region.

was also among the top ten. These sites were conserved in the orthologous genes from *Anopheles*, when those could be identified (e.g. for *m4*, *hairy*, *HLHm3*) or even found in two additional insects: *Drosophila simulans* and the flour beetle *Tribolium castanaeum* (*hairy*).

This prompted us to examine all the genes in *E(spl)* and *Brd* complexes for *miR-7* sites. We found possible target sites in many of them. Alignment of these sites showed a pattern of 5' end conservation quite similar to that for validated targets, with no mismatches and few G:U base pairs for about half of these genes (Figure 3.19B). Julius Brennecke assessed the validity of some of these predictions experimentally (see Stark *et al.*, 2003a). Expression of *miR-7* in transgenic flies caused a clear downregulation of an EGFP (enhanced green fluorescent protein) reporter bearing either the *m4*, *hairy*, or *HlHm3* 3'UTR suggesting that all three are valid targets. In addition, we observed phenotypes reminiscent of *Notch* mutant flies such as

notching of the wing margin or a reduced spacing of veins 3 and 4 (Baonza & Garcia-Bellido, 2000; de Celis & Garcia-Bellido, 1994; Diaz-Benjumea & Cohen, 1995; Micchelli *et al.*, 1997; Rulifson & Blair, 1995) and noticed reduced levels of *Cut* protein, whose expression is dependent on bHLH transcription factors and *Brd*-like genes of the *E(spl)* complex (Ligoxygakis *et al.*, 1999). *miR-7* expression could provide a means to simultaneously down-regulate these and other proteins, that might otherwise function redundantly to mediate *Notch* activity in the wing margin. Taken together, these findings support the prediction that the *miR-7* miRNA regulates expression of bHLH and *Brd*-like proteins encoded by *hairy* and the *E(spl)* and *Brd* complex genes and implicates *miR-7* as a possible regulator of *Notch* target gene expression. A more detailed analysis of the physiological function of *miR-7* will require isolation of lack-of-function mutations in the *miR-7* gene.

Lai has reported complementarity between some miRNAs and sequence elements known as K boxes, Brd boxes and GY boxes in the 3' UTRs of *E(spl)* and *Brd* complex genes (Lai, 2002). K boxes and Brd boxes have been implicated in post-transcriptional regulation, though no function was assigned to the *miR-7* complementary GY boxes (Lai *et al.*, 1998; Lai & Posakony, 1997). The presence of GY boxes in several *E(spl)* and *Brd* complex genes as well as in *hairy* and *extramachrochatae* (*emc*) has been reported (Lai & Posakony, 1998). Based on the presence of these boxes, Lai predicted *miR-7* target sites in *HLHm3* and in *Tom* (Lai & Posakony, 1998). We extend these predictions to a much larger gene family and provide experimental validation for some of them, indicating that GY boxes participate in the regulation of *Notch* target genes.

| miR-7 | ΔG | Z(max) | # Z ≥ 3 | Z(UTR) | Gene | Alignment | Flags | Ag Z≥3 |
|---|---|---|---|---|---|---|---|---|
| 1 | -38.7 | 7.47 | 1 | 7.82 | CG14989-RB | ACAGCAGAAUCACGC-AGGG-CUUCCA<br>UGUUGUUUUAGUG--AUC--AGAAGGU<br>***+***+******---**---****** | - | + |
| 2 | -37.3 | 7.03 | 1 | 7.40 | **HLHm3** | GCAACAAGAUCCGUU----GUCUUCCA<br>UGUUGUUUUAG----UGAUCAGAAGGU<br>+******+***--------******** | - | NF |
| 3 | -35.3 | 6.39 | 1 | 6.81 | CG17657-RA | ACAACGGUUAAG---CGCUGCGUCUUCCA<br>UGUUGUU-----UUAGUGAU-CAGAAGGU<br>*****++--------*+**+-******** | - | NF |
| 4 | -35.0 | 6.29 | 1 | 6.72 | **hairy** | ACAGCAAAUCAG--CAAA--AGUCUUCCA<br>UGUUGUUU----UAGU--GAUCAGAAGGU<br>***+****------**----********* | - | + |
| 5 | -34.5 | 6.13 | 2 | 11.78 | Tom | -UAGCC-GAAUCAUU-GUCUUCCA<br>UGUUG-UUUUAGUGAUCAGAAGGU<br>-+*+*--+*****+*-******** | - | + |
| 6 | -33.9 | 5.94 | 1 | 6.39 | hep | GCAGCAACAGUCGC-AGUUUUUCA<br>UGUUGUU-UUAGUGAUCAGAAGGU<br>+**+***-*+**+*-***+**+** | 5'cons<br>CDS+ | NF |
| 7 | -33.8 | 5.91 | 1 | 6.36 | CG8944-RA | ACGACAAGAUCAAGCGCUACGUCUG-CCA<br>UGUUGUUUUAG----UGAU-CAGA-AGGU<br>**+****+***----+***-****--*** | 5'helix<br>CDS+ | NF |
| 8 | -33.1 | 5.68 | 1 | 6.15 | CG10540-RA | -CGACAAAGCG--GCCCAAUAGUCUUCCA<br>UGUUGUUUU--AGUG----AUCAGAAGGU<br>-*+*****+----+*----********** | - | - |
| 9 | -31.8 | 5.27 | 1 | 5.76 | CG10444-RA | GCGACC-AAAA-CAG--AGUCUUCCA<br>UGUUG-UUUU-AGU-GAUCAGAAGGU<br>+*+**--***--**---********* | - | NF |
| 10 | -31.5 | 5.17 | 1 | 9.29 | **m4** | -CAGCUUU--AAUCAAC---GUCUUCCG<br>UGUUG---UUUUAGU--GAUCAGAAGGU<br>-**+*-----*****-----*******+ | - | + |

| miR-2a | ΔG | Z(max) | # Z ≥ 3 | Z(UTR) | Gene | Alignment | Flags | Ag Z≥3 |
|---|---|---|---|---|---|---|---|---|
| 1 | -39.0 | 6.78 | 2 | 11.85 | CG1969-RB | ----------GCUGGCUGGC-GGUG<br>CGAGUAGUUUCGACCGAC--ACUAU<br>----------**********-*+*+ | 5'cons<br>5'helix<br>mispairing | NF |
| 2 | -38.6 | 6.66 | 1 | 6.66 | CG4269-RA | GCUCCUG--CAU-GGAUUGGCUGUGAUA<br>CGAG---UAGU-UUC-GACCGACACUAU<br>****------**--+*-+********** | - | - |
| 3 | -38.0 | 6.49 | 1 | 6.49 | **reaper** | -CUCAUCAAAGCGA---UUGUGAUA<br>CGAGUAGUUUCG--ACCGACACUAU<br>-**********-----+******* | - | NF |
| 4 | -34.3 | 5.42 | 1 | 5.42 | GLaz | GCUUUGAU----GAGC--GCUGUGAUA<br>CGAG----UAGUUUCGACCGACACUAU<br>***+---------+***--********** | mispairing | - |
| 5 | -33.5 | 5.19 | 1 | 5.19 | BG:DS05899.3 | GUUCAUCCCUU--GGCGUUG-GGCUGUGU-UA<br>CGAGUAG----UUUCG----ACCGACAC-UAU<br>*+*****------+**-----********--** | 5'helix | NF |
| 6 | -33.2 | 5.1 | 1 | 5.1 | Scr | GCUCGGUG-GGAGUG-GGUG-GUGGUG<br>CGAGU-A-GUUUCG-ACCG-ACACUAU<br>****+-*--++**+--**+--***+*+ | - | - |
| 7 | -33.0 | 5.04 | 1 | 5.04 | hbs | ---CAUGCGGC-GCUCGAAGGCUGUGAUA<br>CGAGUA-GUU-UCGA----CCGACACUAU<br>---***-*++--********---********** | - | - |
| 8 | -32.8 | 4.99 | 1 | 4.99 | amon | GUUCAA-UAAAAGUGCUGGCUGUG---<br>CGAGU-AGUUU---CGACCGACACUAU<br>*+*****+***---*********--- | 5'helix | - |
| 9 | -32.5 | 4.9 | 1 | 4.9 | **grim** | GCUCAAUCAAAGCGCA---UUGUGAU-<br>CGAGU-AGUUUCG---ACCGACACUAU<br>*****-*******------+******- | - | NF |
| 10 | -32.1 | 4.78 | 2 | 8.01 | CG7187-RA | GCUUUGAU----GAGC--GCUGUGGUG<br>CGAG----UAGUUUCGACCGACACUAU<br>***+---------+***--*******+*+ | 5'cons<br>mispairing | NF |

**Table 3.2:** Top Ten Predictions for *miR-7* and *miR-2a* ΔG, $Z_{max}$ and $Z_{UTR}$ are explained in the text. Alignment: top: target site; middle: miRNA; bottom: *, conventional base pair; +, G:U base pair, -, mismatche or gap. Flags: The "5' conservation" flag identifies sites that differ in the two genomes at any residue complementary to positions 2 to 7 of the miRNA (5'cons). The "5' helix" flag identifies sites that do not have at least 6 contiguous base pairs in positions 1-8. The "CDS+" flag indicates that the predicted site overlaps coding sequence on the same strand; "CDS-" indicates that the overlap is on the opposite strand. In some cases Mfold structures include base pairs that are not between the miRNA and its target. "Mispairing" flags sites with artificially high folding energies. Ag (Z≥3): Anopheles genes that cannot be reliably identified by our criteria are indicated "NF". For the cases where the orthologous Anopheles gene was found, the presence of a target site with Z≥3 is indicated (+, otherwise -). Heavy outlining indicates those loci that would pass stringent filtering of the lists using the flags and lack of a conserved target in an Anopheles ortholog.

### 3.5.5 *miR-2* regulates pro-apoptotic genes

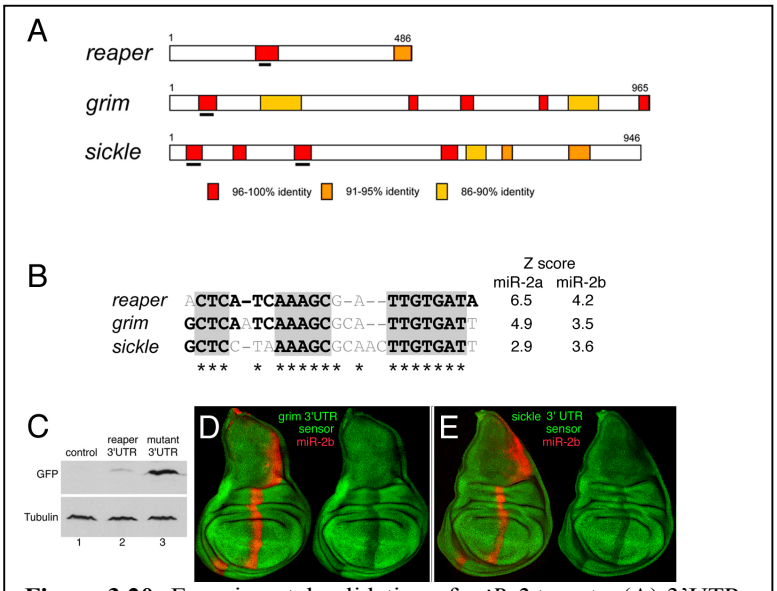



**Figure 3.20:** Experimental validation of *miR-2* targets. (A) 3'UTRs of *reaper*, *grim* and *sickle* highlighting blocks of high sequence similarity between *D. melanogaster* and *D. pseudoobscura* (underlined: pedicted *miR-2/13* target sites). (B) Alignment of predicted *miR-2/13* target sites in the *reaper*, *grim* and *sickle* 3' UTRs with associated Z-scores. (C-E) Experimental validation (J. Brennecke, for details see Stark *et al.*, 2003a) (C) Immunoblot probed for GFP and Tubulin of S2 cells transfected to express an EGFP-reaper 3' UTR construct (lane 2) or a comparable construct from which the *miR-2a* site has been deleted (lane 3; lane 1: control with empty vector). (D, E) Expression of the *grim* and *sickle* 3' UTR sensor transgenes (green) were down-regulated by *miR-2b* (red).

The pro-apoptotic genes *reaper* and *grim* were among the top predictions for *miR-2a* and *miR-2b* (Table 3.2). *reaper*, *grim* and the third pro-apoptotic gene *sickle* are clustered in the genome and show blocks of high conservation in their 3' UTRs, which include the *miR-2* family target sites (Figure 3.20A). The putative miRNA-binding site is very similar between the three genes suggesting an identical RNA-complex structure (Figure 3.20B). The corresponding sites are highly similar in *D. pseudoobscura*, but the orthologous genes cannot be identified in *Anopheles*. We could again show a miRNA dependent downregulation of a 3'UTR EGFP-reporter by *miR-2* in transgenic flies (*grim*, *sickle*) or *Drosophila* Schneider S2 cells (*reaper*, Figure 3.20). The *miR-13* family is similar in sequence to the *miR-2* family. Experimental validation will be needed to determine if *reaper*, *grim* and *sickle* are also regulated by *miR-13*. Identification of *reaper*, *grim* and *sickle* as targets suggests that *miR-2* family miRNAs may be involved in control of apoptosis.

### 3.5.6 Statistical Evaluation of Target Predictions

Although a number of the top-ranking sites identified in our screen have been experimentally validated, we wanted to assess the likelihood that sites with equivalent scores can be found by chance. To do so I calculated E-values for the *bantam* miRNA based on the tail of the cumulative distribution of ΔG values for 10,000 random matches. An E-value predicts the number of background matches with a similar or better score (E-values scale with database

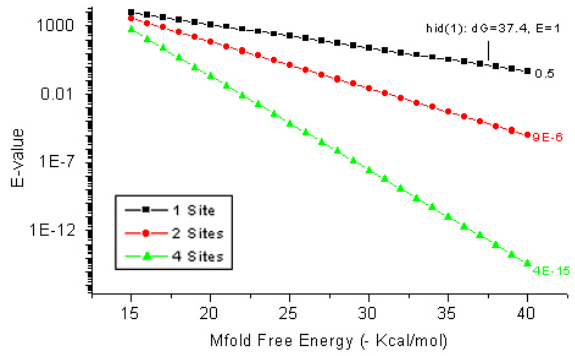



**Figure 3.21**: Statistical evaluation of predicted targets. Plot of E-values as a function of free energy of folding predicted by Mfold. E-value plots are shown for one, two, and four sites of equal energy. The position of the best *bantam* site in *hid* is shown for reference.

size and are applicable to any distribution profile; see Introduction). The results are presented on a logarithmic scale for UTRs containing 1, 2 or 4 sites of a given ΔG value (Figure 3.21). The best single *bantam* site in the *hid* UTR had an E-value of 1.3. This means that background matches reach RNA-duplex energies similar to the best sites, even in the smaller conserved 3' UTR database. Indeed, target sites predicted using shuffled *bantam* miRNA sequences give folding energy distributions very similar to the native sequence (not shown). Although single sites are not statistically significant, the presence of multiple sites within a single UTR can greatly increase the significance of the prediction. Combining the three *bantam* sites (Z>3) predicted in the *hid* 3' UTR gives an E-value of $1.8 \times 10^{-5}$. Some single sites are sufficient to mediate regulation by a miRNA, however, we emphasize that the lack of statistical significance for even the best single site means that they require experimental validation.

### 3.5.7 Additional validation by cross genome comparison

One means to improve the significance of the predictions would be to require conservation in a third genome. The two *Drosophila* species are separated by an estimated 30 million years. The mosquito *Anopheles gambiae* is separated from *Drosophila* by 250 million years. Orthologous mosquito genes have been defined for approximately two-thirds of *Drosophila* genes, however, systematic comparison showed great differences in length between orthologous gene pairs (Zdobnov *et al.*, 2002). Indeed we were able to identify orthologous last exons with confidence for only half of these pairs, or one-third of *D. melanogaster* genes. We have therefore chosen to use conservation in *Anopheles* to provide a more stringent evaluation of target site conservation, instead of requiring it generally. The presence of a conserved site with a high Z-score across all three genomes increases the confidence that the site is functional. To illustrate the utility and limitations of this we examine the top 100 predictions for *miR-7* and *miR-2*. The *Anopheles* orthologs were identified for 52 of the top 100 predicted *miR-7* targets. Eleven of these had conserved target sites (Z≥3), including four

of the top ten predictions, *hairy*, *Tom*, *m4* and *CG14989* (Table 3.2). For *miR-2a* 40 of the top 100 predictions had a detectable ortholog in *Anopheles*. Five of these sites were conserved in *Anopheles* (Z≥3), and none of these were among the top ten predictions. Conservation in *Anopheles* can thus be used to find sites with a higher probability of being valid, but increases the risk of missing real targets. It is only useful in cases where the orthologous UTR region can be identified, which is not the case for the validated *miR-2a* targets *grim*, *reaper* and *sickle*.

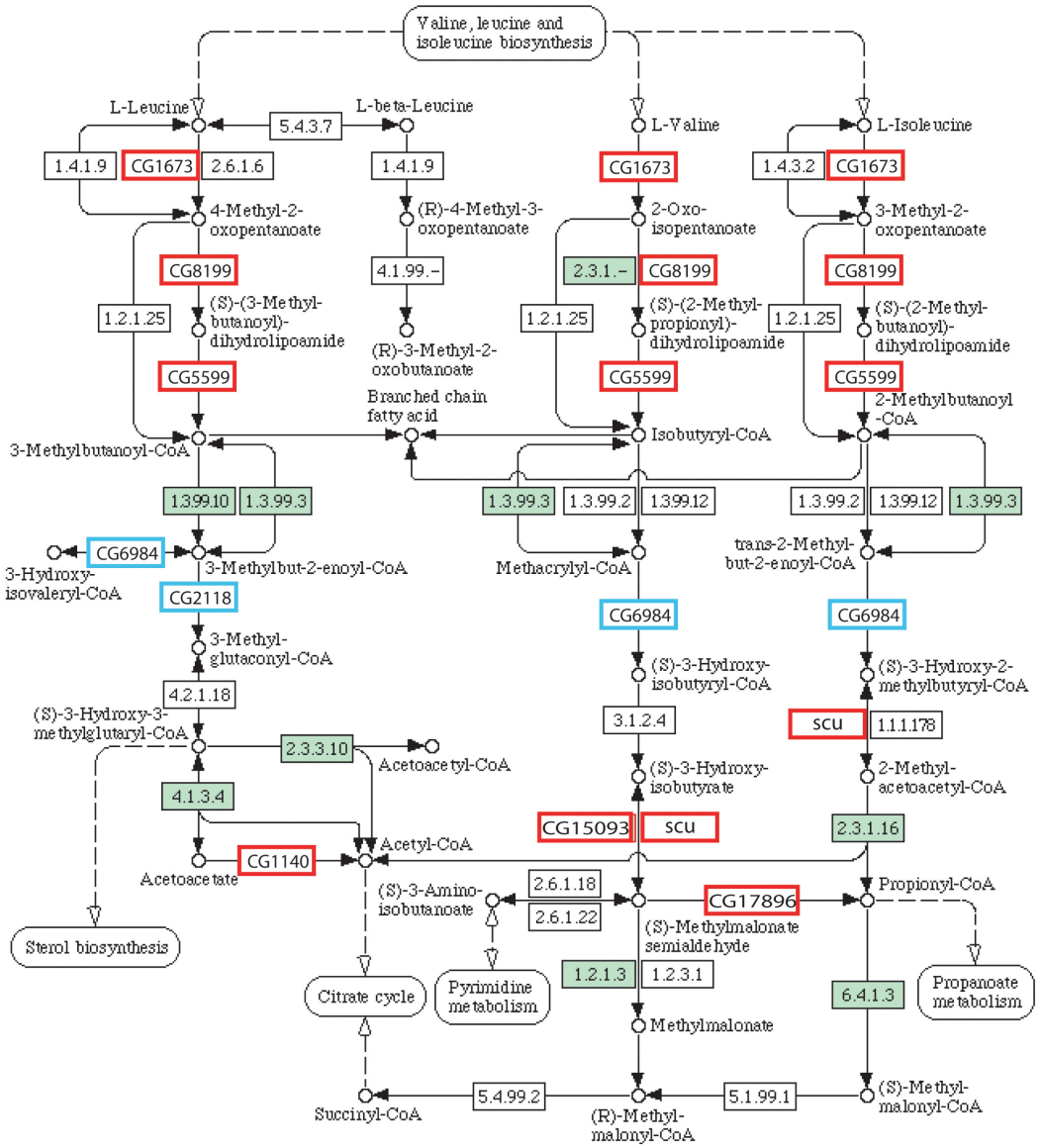



**Figure 3.22:** Valine, Leucine, and Isoleucine Catabolic Pathway (adopted from KEGG). Predicted *miR-277* targets conserved in Anopheles are boxed (red: Z>3, blue: Z>2), other enzymes that are known in *Drosophila* are shaded in green.

### 3.5.8 *miR-277* is a putative metabolic switch

Table 3.3 shows predicted *miR-277* targets that are conserved (Z≥3) in *Anopheles*. Seven of the top eleven are enzymes involved in branched chain amino acid degradation and two additional enzymes were identified at more relaxed stringency (Z≥2), along with a number of unrelated loci (Figure 3.22). This striking clustering of functionally-related enzymes suggests that *miR-277* regulates the pathway for valine, leucine and isoleucine degradation by down-regulating many of its enzymes and thus acts as a metabolic switch. The degradation of these essential amino acids is presumably regulated under conditions of starvation or excess dietary intake. *miR-277* expression has so far only been detected in adult flies (Aravin *et al.*, 2003; Lai *et al.*, 2003) suggesting a role in regulating metabolic responses to environmental conditions. Interestingly, the human homolog of *CG8199* is mutated in maple syrup urine disease. It remains to be determined if these enzymes are regulated by miRNAs in vertebrates.

| Rank | Gene | Function | Enzyme | # Z≥3 | $Z_{UTR}$ |
|---|---|---|---|---|---|
| 1 | CG31651 | Protein GalNAc transferase | EC:2.4.1.41 | 2 | 7.31 |
| 2 | CG5599 | Val Leu Ile degradation | EC:2.3.1- | 2 | 6.53 |
| 3 | CG1673 | Val Leu Ile degradation | EC:2.6.1.42 | 1 | 5.75 |
| 4 | fz | Cell polarity | | 1 | 4.89 |
| 5 | CG8199 | Val Leu Ile degradation | EC:1.2.4.4 | 1 | 4.53 |
| 6 | CG18549 | - | | 1 | 4.23 |
| 7 | CG1140 | Val Leu Ile degradation | EC:2.8.3.5 | 1 | 3.9 |
| 8 | scu | Val Leu Ile degradation | EC:1.1.1.35 | 1 | 3.81 |
| 9 | CG15093 | Val Leu Ile degradation | EC:1.1.1.31 | 1 | 3.79 |
| 10 | CG7740 | Membrane protein | | 1 | 3.64 |
| 11 | CG17896 | Val Leu Ile degradation | EC:1.2.1.27 | 1 | 3.61 |

**Table 3.3**: Top *miR-277* targets include many enzymes from the valine, leucine, isoleucine degradation pathway (required: Z≥3 *D. melanogaster, D. pseudoobscura, Anopheles*).

### 3.5.9 Features shared by validated targets

Comparison of the five previously validated targets and the six new targets validated here revealed three features shared by all sites. (1) Cross-genome comparison showed perfect sequence identity in the target site residues that base pair with residues 2-8 of the miRNA (Figure 3.17). This was also true for the newly validated target sites (data not shown). (2) The pattern of base pairing between the miRNAs and their targets shown in Figure 3.18A suggested that a continuous helix of at least six of the first eight base pairs might be required (allowing for G:U base pairs). This was also true for the newly validated target sites (Figures

3.19B and 3.20B). (3) Many transcripts in the *D. melanogaster* genome overlap other transcripts on the same strand or on the opposite strand of the DNA. There are many examples of alternate splicing that produces alternate 3' UTRs, so that one UTR variant may include coding exons from another variant. In such cases the basis for the sequence conservation between genomes is unclear. None of the validated sites from *Drosophila* overlaps coding sequence on either strand (this feature was not examined for the *C. elegans* sites).

Target sites that do not share these features are indicated in Table 3.2. These features can be used to increase the stringency of the screen, by discounting sites that differ from validated targets. For *miR-7* this would eliminate two of the top ten predictions so that the validated targets would constitute three of the remaining top eight predictions. For *miR-2a* this would eliminate four of the top ten predictions, so that the validated targets *reaper* and *grim* would rank in positions two and six. We have chosen not to implement the flags as filters to exclude predictions because they are based on a relatively small set of eleven validated targets. Although, we note that all nine predicted *miR-277* targets would pass such a filter. When more targets are validated we will learn if these features have a general predictive value.

### 3.5.10 Discussion

One of the major limitations in studying animal miRNA function is the difficulty in identifying their targets. Our screening strategy has proven to be useful for predicting new miRNA targets. Three new targets have been experimentally validated for *miR-7* and for *miR-2*, bringing the total number of validated targets of animal miRNAs to eleven. In addition we predict a number of miRNA/target pairs or target families that seem likely to be valid, but require experimental validation. Our study depended on the high quality annotation of the *D. melanogaster* genome and the availability of the *D. pseudoobscura* genome sequence. Where possible we have extended the analysis to include evaluation of predicted sites in the *Anopheles gambiae* genome. More complete annotations of the fly and mosquito genomes, aided by cDNA sequencing projects, will increase the number of genes for which orthologous UTR sequences can be defined. This will permit more sensitive and more extensive cross genome comparison.

We have made a number of assumptions based on the inspection of known targets in designing the screen and it remains to be determined if all of them will prove to be generally

applicable. For example, as all previously known animal miRNA target sites are located in 3' UTRs of protein coding genes, we restricted our screen to these sequences (as were all similar approaches, see below). However, we also find sites with similar scores in coding sequences, in 5' UTRs and indeed through DNA sequence in general. It remains to be determined whether miRNAs can act to control translation via these sites such as RNAi can act via sites in the coding region (e.g. Kasschau *et al.*, 2003; Llave *et al.*, 2002; Tang *et al.*, 2003). The use of whole transcriptome or whole genome databases would greatly increase the search space, which might bury valid targets among background matches. Based on the observation that miRNA target complementarity was best for the first eight residues of the miRNA in the previously known examples, we searched for complementarity only to these residues. However, we noticed that miRNA-mRNA duplexes with preferential pairing in the 3' end or middle regions had Mfold energies similar to our predictions (not shown) but it is currently unknown whether such sites function with a given miRNA, or more importantly, whether some miRNAs might indeed favor them. On the other hand, the unknown requirements for functional pairing also mean that some of our predictions might not be functional and we expect improvements in specificity to come from a more precise understanding of the structural requirements for miRNA/target pairing.

In designing the screening strategy we considered the balance between sensitivity and specificity. We chose a search strategy that was based on the known examples, but generalized to allow detection of similar targets. By doing so, we risk missing fewer valid targets at the expense of including more false positives, as indicated by our statistical analysis (Figure 3.21). This enabled us to detect all known miRNA targets and to predict clusters of functionally related targets, not detected by other methods with more stringent requirements (see below). To help distinguish false positives from potentially valid targets we identify features shared by valid targets and, where possible, test predictions for conservation in a third, more distantly related, genome. Both positive and negative results in tests of new predictions will provide a better understanding of how miRNAs bind their targets, perhaps highlighting positions that are particularly critical, which will permit high sensitivity and specificity for future target prediction methods.

Complete tables of target site predictions are available on the web at www.miRNA.embl.de. These tables report Z-scores and sequences for the *D. melanogaster*, *D. pseudoobscura* target sites and where possible for the *Anopheles* target site. The tables contain flags to identify sites

that share the features described above. We recommend using these flags to filter the lists, but note that this may exclude valid targets. We recommend making use of the *Anopheles* data to discount predictions where the orthologous gene is identified and the site is absent or has a low Z-score (Z<2).

The presence of a conserved site in all three genomes increases the confidence that a predicted site is valid, as in the case of the *miR-7* sites in *hairy* and *Tom*. Also, *dpld*, the *Drosophila* homolog of the *let-7* target *lin-41*, ranks second among *Drosophila let-7* targets when conservation in *Anopheles* was required. A number of other target predictions that meet these requirements look quite promising. We have high confidence that the cluster of enzymes in the branched-chain amino-acid degradation pathway will prove to be valid *miR-277* targets. Another promising candidate is the predicted *miR-9a* target *Lyra*. *Lyra* contains two predicted *miR-9a* sites. The best *Lyra* site ranks first among all predicted *miR-9a* targets that are conserved in *Anopheles*. Intriguingly, mutations affecting the *Lyra* 3' UTR lead to a dominant phenotype and to increased *Lyra* protein levels, an observation that strongly suggests that *Lyra* is subject to translational regulation. *miR-9* is an excellent candidate to mediate this regulation. Many other miRNA/target pairs are identified with sites of a similar quality to those mentioned here (examples include four conserved sites for *miR-309* in *Ets65a* at Z≥2). Interestingly, we found that many genes are predicted to be regulated by different miRNAs. An outstanding example is *nerfin-1*, which has binding sites in its 3' UTR for *bantam*, *miR-9b*, *miR-5*, *miR-279*, and *miR-286*, many of which are conserved in *Anopheles*. Regulation of *nerfin-1* by miRNAs is also expected from the observation that nerfin-1 mRNA is ubiquitously present during the embryonic development of the central nervous system, whereas the protein is only expressed in neuronal precursor cells (A. Kuzin et al., unpublished). This shows, that miRNAs are probably involved in nervous system development and that miRNAs seem to form complex regulatory networks as one miRNA can regulate several targets, but one gene can also be regulated by several miRNAs.

Although it is more difficult to distinguish functional sites from false positives in the cases where only two genomes are compared, we have made use of clustering of related genes to identify real targets. *reaper grim* and *sickle* have been validated as miR-2 targets. We note that the *Netrin* receptor *unc-5* and *Netrin-A* rank second and fourth among predicted *miR-288* targets. We observe an abundance of transcription factors among the predicted targets of *miR-9*, *miR-279* and *miR-286* for which orthologous UTRs were not identified in *Anopheles*.

Our statistical analysis shows that the very best single predicted target sites are not statistically significant, even though we have used a reduced database consisting of conserved 3' UTR sequences. This means that prediction of any single target site cannot be taken as evidence for regulation of a transcript by a miRNA without experimental validation. Sites that are not statistically significant alone can be significant when combined. For example, although none of the *bantam* sites are significant individually, their combined scores are highly significant and supported by experimental validation. 3' UTRs with multiple predicted target sites are likely to be valid targets for regulation by the miRNA, particularly if their best single sites also rank high in the lists of predicted targets. Despite the advantages conferred by multiple sites, single miRNA target sites can mediate regulation in vivo. The *C. elegans lin-4* miRNA appears to regulate its target *lin-28* through a single site (Moss *et al.*, 1997). We have presented evidence that *miR-2* family miRNAs can regulate expression of transgenes containing the 3' UTRs of *reaper* and *grim*, which have one predicted target site, as well as the *sickle* 3' UTR, which has two predicted sites. Similarly, *miR-7* can regulate expression of transgenes containing the *HLHm3*, *m4* and *hairy* 3' UTRs, which have one predicted target site. Further work will be needed to gain insight into what makes some single sites functional and others not. One possibility is that a single site for one miRNA might function in conjunction with independent target sites for other miRNAs in the same UTR. Indeed, a survey of our lists of target predictions indicates that many 3' UTRs are predicted to contain binding sites for more than one miRNA.

Recently, other studies on miRNA target prediction in *Drosophila* and mammals have been reported (Enright *et al.*, 2003; Lewis *et al.*, 2003; Rajewsky & Socci, 2004). Enright *et al.* (2003) use a sequence alignment algorithm to find sites complementary to the miRNA in 3' UTRs from *D. melanogaster* and *D. pseudoobscura*. The algorithm allows for G:U pairing, rewards complementarity in the 5' region of the miRNA and applies empirical filters similar to our flags. For conserved sites, the miRNA:target duplex energy is evaluated with an RNA secondary structure program. The authors used shuffled sequences to estimate that their screen has a false positive rate of about 35% that however improves when multiple sites per target UTR are required (see our statistical evaluation). They found that genes involved in transcriptional and translational control, cell adhesion, enzyme regulation and apoptosis were overrepresented among the predicted targets and discussed implications for body axis specification, ecdysone signalling and development. However, as none of the putative targets

have been validated experimentally, predictions have to be treated with caution. Rajewsky and Socci (2004) directly hybridised *Drosophila* miRNAs to 31 developmental genes involved in body patterning and found (unsurprisingly) that the genes are probably targeted by some of the 74 known miRNAs. Lewis and co-workers predicted miRNA targets that are conserved across three mammalian species (human, mouse, and rat) using a method that combines sequence comparison and evaluation of RNA-binding energy similar to ours (Lewis *et al.*, 2003). Because of more stringent criteria during the search, they cannot detect all known targets but achieve a statistical signal indicating that two-thirds of their predictions should be correct. Indeed, experimental tests in cell culture showed that 11 out of 15 predicted sites had an influence on the expression levels of a reporter gene. Although the authors found many genes involved in transcription and regulation of transcription among their predictions, their prevalence was not as pronounced as in plants where nearly all known miRNAs target transcription factors. This indicated that animal miRNAs act in a broad diversity of biological processes.

All methods are very similar to the one described here and all together, they show that miRNA targets can be predicted computationally, which will be of great help for miRNA research. Ongoing work to refine the structural requirements (e.g. Doench, *et al.*, unpublished; Brennecke, *et al.*, unpublished) will improve the sensitivity and specificity of these methods and lead to a better understanding of miRNA function.

## *3.6 Other Projects*

For my thesis, I also worked on other projects in collaboration with people from our group or other groups at the EMBL-Heidelberg. As these are not directly related to my main projects and the collaborators contributed more to the overall results, I will only summarize the main findings and refer to the resulting publications for further information. The publications corresponding to the following chapters are:

R. Metivier, **A. Stark**, G. Flouriot, M.R. Huebner, H. Brand, G. Penot, D. Manu, S. Denger, G. Reid, M. Kos, R.B. Russell, O. Kah, F. Pakdel, F. Gannon; A Dynamic Structural Model for Estrogen Receptor-alpha Activation by Ligands, Emphasizing the Role of Interactions between Distant A and E Domains. *Molecular Cell*, 10, 1019-1032, 2002.

J. Brennecke, D.R. Hipfner, **A. Stark**, R.B. Russell, S.M. Cohen; *Bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the pro-apoptotic gene *hid* in *Drosophila*. *Cell*, 113, 25-36 , 2003.

P. Aloy, **A. Stark**, C. Hadley, R.B. Russell; Predictions without templates: New folds, secondary structure and contacts in CASP5. *PROTEINS: Struct. Funct. Genet*., 53, 436-456, 2003 (CASP5 special issue).

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

### 3.6.1 Estrogen receptor: a case study

Estrogen receptors (ERα and ERβ) are cellular receptors for the steroid hormone estradiol that plays a pivotal role in the female reproductive tract physiology and is important for the homeostasis of other tissues (Feigelson & Henderson, 1996; Nilsson *et al.*, 2001). They belong to the superfamily of nuclear receptors and are composed of six modular domains, A to F. ERs bind to DNA at specific *estrogen response elements* with the C domain (DNA binding domain) and two activation functions (AFs) are located in the B-domain and the C-terminal D – F domains; the E domain is the ligand-binding domain (LBD). In contrast, a specific role for the N-terminal A domain was unknown. Some evidence for its function came from a short ERα variant expressed in fish but also in some human cell types that lacked the A domain and showed transcriptional activity in the absence of ligand (Flouriot *et al.*, 2000; Pakdel *et al.*, 2000). Previous work demonstrated that the A domain interacted with the LBD

in the absence of ligand suggesting a possible mechanism for a function in silencing ligand-independent ERα activities (Metivier *et al.*, 2000).

I used the available sequence and structural data to detect possible determinants for the interaction of the A domain and the LBD that were then tested experimentally. We showed that the interaction depended on charged residues and on a hydrophobic helix that were conserved in multiple sequence alignments of A domains from different species. Inspection of the LBD structure suggested that the A domain might compete with the ERα C-terminal helix and corepressors for the same binding site. R. Metivier (F. Gannon's group) experimentally validated this model and showed that mutations in the identified interaction sites had consistent effects on the A domain binding and on ERα activity. We could thus propose a model that integrates different domain functions and provides insight into the dynamic structure of the full-length ERα when bound to different ligands (Metivier *et al.*, 2002).

### 3.6.2 The miRNA Oncogene *Bantam*

The *Drosophila bantam* locus was identified in a gain-of-function screen for genes that affect tissue growth. Artificial overexpression of the genomic region leads to tissue overgrowth by an increase in cell number. Conversely, flies lacking about 21 kb in this region grew poorly and died as early pupae (Hipfner *et al.*, 2002). The *bantam* region does not contain any predicted protein coding genes, but we found a highly significant similarity (30/31 nucleotides) to the *Anopheles gambiae* genome. An alignment of the two genomic sequences showed a block of about 90 nucleotides with considerable similarity. Both sequences were predicted to form stable hairpin structures characteristic for miRNA-precursors (see Introduction). Experiments showed that the mature miRNA and the precursor are indeed present in flies and that a short sequence containing the predicted precursor could rescue *bantam* mutants and reproduce the overexpression phenotype. J. Brennecke (Stephen Cohen's group) also showed that the spatial expression of *bantam* correlates with cell proliferation and that *bantam* can rescue cells from undergoing apoptosis. Using an earlier version of the target prediction method described in this thesis, I predicted the pro-apoptotic gene *hid* to be a potential *bantam* target and J. Brennecke showed experimentally that this is true. *bantam* was the first miRNA found outside nematodes and extended miRNA function from developmental timing to regulation of cell proliferation and apoptosis. As it is able to stimulate proliferation and also prevents apoptosis, it is the first miRNA oncogene to be discovered.

### 3.6.3 CASP5 – Assigning Structures to Sequences

In Autumn 2002, our group assessed the CASP5 (Critical Assessment of Structure Prediction) predictions in the *new fold* category, i.e. for proteins that adopt a structure not observed before. We considered a total of 4840 models from 165 groups for five targets with new folds and eight lying on the borderline to fold recognition or threading (see Introduction). Our detailed visual and numerical inspection showed that the quality of the best predictions was very good: for nearly every target at least one group predicted a structure close to the correct one. The group of David Baker that used a method based on fragment assembly (Bradley *et al*., 2003; Chivian *et al*., 2003, see Introduction), proved to be best overall, but we also saw high quality and consistency from others, suggesting that the community is moving toward general procedures to predict accurate structures for proteins showing no resemblance to anything seen before. We did not find secondary structure predictions significantly improved since CASP4 (Lesk *et al.*, 2001) and the good performance of most groups suggests that the limits of accuracy for *a priori* secondary structure prediction may be reached (Kabsch & Sander, 1983b; Rost *et al.*, 1994; Russell & Barton, 1993).

### 3.6.4 The Relationship Between Sequence and Interaction Divergence

Protein interactions and complexes have recently been the subject of great interest. Methods like the yeast-two-hybrid system or affinity purifications have identified many interactions and complexes (von Mering *et al.*, 2002). The 3D structure of complexes can provide a deeper understanding of the function and has become the focus of much of experimental structural biology. However, there is currently a gap in knowledge between complexes of known structure and those known from other experimental methods. Modeling interactions based on a homologous complex structure assumes that the components will interact in the same way (e.g. Aloy & Russell, 2002). Several studies have explored the nature of different protein-interaction types (Chakrabarti & Janin, 2002; Jones & Thornton, 1997; Ofran & Rost, 2003), how domain orientations can vary within specific superfamilies (e.g. Bashton & Chothia, 2002), or the possibility of transferring interface information to homologous (Aloy & Russell, 2002; Lu *et al.*, 2002). However, no study of the general relationship between similarity in interaction and sequence has been performed.

We used pairs of interacting domains of known 3D structure to explore the relationship between sequence divergence and similarity in interaction (Aloy *et al.*, 2003a). We compared the relative orientations of different instances of the same interacting domain pair with a newly devised measure, interaction RMSD (iRMSD). iRMSD is a purely geometric difference between domain orientations related to the RMSD used in this thesis and accounts for both translational differences (i.e. different location of the centre of masses) and domain rotations over a wide range of differences in domain orientation.

We observed that interactions tend to be generally similar when sequence identity is above 30 – 40%, which is just above the traditional twilight zone for the relationship between sequence and structural similarity (Chothia & Lesk, 1986). Domains from the same family or superfamily are more likely to interact similarly than domains from different superfamilies in the same fold. Here, exceptions were found to mainly due to crystal contacts that are probably not biologically meaningful or domains from the immune system such as immunoglobulins. We identified the few instances where unrelated proteins that share a common fold interact similarly such as trypsin and trypsin inhibitors from different superfamilies or similarities between homo- or pseudohomo-dimers. We also saw that interactions between domains in separate proteins compared to intramolecular interactions of similar domains (i.e. in one protein after fusion events) are rarely similar. One should thus exercise caution when inferring a domain-domain interaction between separate proteins based on a similar pair of domains in a single polypeptide (Apic *et al.*, 2001; Enright & Ouzounis, 2001).

# 4 References


Abrahante, J. E., Daul, A. L., Li, M., Volk, M. L., Tennessen, J. M., Miller, E. A., and Rougvie, A. E. (2003). The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. Dev Cell *4*, 625-637.

Aldous, D. J. (1989). Probability approximations via the Poisson clumping heuristic (New York, Springer-Verlag).

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

Aloy, P., Companys, V., Vendrell, J., Aviles, F. X., Fricker, L. D., Coll, M., and Gomis-Ruth, F. X. (2001a). The crystal structure of the inhibitor-complexed carboxypeptidase D domain II and the modeling of regulatory carboxypeptidases. J Biol Chem *276*, 16177-16184.

Aloy, P., Querol, E., Aviles, F. X., and Sternberg, M. J. (2001b). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol *311*, 395-408.

Aloy, P., and Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci U S A *99*, 5896-5901.

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

Altschul, S. F., and Gish, W. (1996). Local alignment statistics. Methods Enzymol *266*, 460-480.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

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

Ambros, V. (2001). microRNAs: tiny regulators with great potential. Cell *107*, 823-826.

Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T., and Jewell, D. (2003). MicroRNAs and other tiny endogenous RNAs in C. elegans. Curr Biol *13*, 807-818.

Apic, G., Gough, J., and Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol *310*, 311-325.

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D.*, et al.* (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res *29*, 37-40.

Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small RNA profile during Drosophila melanogaster development. Dev Cell *5*, 337-350.

Armon, A., Graur, D., and Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol *307*, 447-463.

Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W., and Willett, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol *243*, 327-344.

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P.*, et al.* (2003). PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res *31*, 400-402.

Babbitt, P. C. (2003). Definitions of enzyme function for the structural genomics era. Curr Opin Chem Biol *7*, 230-237.

Bahar, I., and Jernigan, R. L. (1997). Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol *266*, 195-214.

Bakal, C. J., and Davies, J. E. (2000). No longer an exclusive club: eukaryotic signalling domains in bacteria. Trends Cell Biol *10*, 32-38.

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci U S A *91*, 1059-1063.

Banerjee, D., and Slack, F. (2002). Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. Bioessays *24*, 119-129.

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E.*, et al.* (1975). Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 angstrom resolution using amino acid sequence data. Nature *255*, 609-614.

Baonza, A., and Garcia-Bellido, A. (2000). Notch signaling directly controls cell proliferation in the Drosophila wing disc. Proc Natl Acad Sci U S A *97*, 2609-2614.

Barker, J. A., and Thornton, J. M. (2003). An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics *19*, 1644-1649.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bartlett, G. J., Porter, C. T., Borkakoti, N., and Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. J Mol Biol *324*, 105-121.

Barton, G. J. (1993). ALSCRIPT: a tool to format multiple sequence alignments. Protein Eng *6*, 37-40.

Bashton, M., and Chothia, C. (2002). The geometry of domain combination in proteins. J Mol Biol *315*, 927-939.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002). The Pfam protein families database. Nucleic Acids Res *30*, 276-280.

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

Bjorkegren, C., Rozycki, M., Schutt, C. E., Lindberg, U., and Karlsson, R. (1993). Mutagenesis of human profilin locates its poly(L-proline)-binding site to a hydrophobic patch of aromatic amino acids. FEBS Lett *333*, 123-126.

Blundell, T. L., and Johnson, M. S. (1993). Catching a common fold. Protein Sci *2*, 877-883.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I.*, et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res *31*, 365-370.

Boggon, T. J., Shan, W. S., Santagata, S., Myers, S. C., and Shapiro, L. (1999). Implication of tubby proteins as transcription factors by structure-based functional analysis. Science *286*, 2119-2125.

Bonneau, R., Strauss, C. E., Rohl, C. A., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., and Baker, D. (2002). De novo prediction of three-dimensional structures for major protein families. J Mol Biol *322*, 65-78.

Bork, P., Holm, L., and Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol *242*, 309-320.

Bradley, P., Chivian, D., Meiler, J., Misura, K. M., Rohl, C. A., Schief, W. R., Wedemeyer, W. J., Schueler-Furman, O., Murphy, P., Schonbrun, J.*, et al.* (2003). Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins *53 Suppl 6*, 457-468.

Branden, C.-I., and Tooze, J. (1999). Introduction to protein structure, 2nd edn (New York, Garland Pub.).

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B., and Cohen, S. M. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell *113*, 25-36.

Brenner, S. E., Chothia, C., and Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A *95*, 6073-6078.

Browne, W. J., North, A. C., Phillips, D. C., Brew, K., Vanaman, T. C., and Hill, R. L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol *42*, 65-86.

Bujnicki, J. M., Elofsson, A., Fischer, D., and Rychlewski, L. (2001). LiveBench-2: large-scale automated evaluation of protein structure prediction servers. Proteins *Suppl 5*, 184-191.

Burley, S. K. (2000). An overview of structural genomics. Nat Struct Biol *7 Suppl*, 932-934.

Burley, S. K., and Bonanno, J. B. (2002). Structural genomics of proteins from conserved biochemical pathways and processes. Curr Opin Struct Biol *12*, 383-391.

Bystroff, C., and Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol *281*, 565-577.

Carl, U. D., Pollmann, M., Orr, E., Gertlere, F. B., Chakraborty, T., and Wehland, J. (1999). Aromatic and basic residues within the EVH1 domain of VASP specify its interaction with proline-rich ligands. Curr Biol *9*, 715-718.

Carter, C. W., Jr., LeFebvre, B. C., Cammer, S. A., Tropsha, A., and Edgell, M. H. (2001). Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol *311*, 625-638.

Caudy, A. A., Myers, M., Hannon, G. J., and Hammond, S. M. (2002). Fragile X-related protein and VIG associate with the RNA interference machinery. Genes Dev *16*, 2491-2496.

Chakrabarti, P., and Janin, J. (2002). Dissecting protein-protein recognition sites. Proteins *47*, 334-343.

Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., Smith, T.*, et al.* (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science *282*, 2022-2028.

Chivian, D., Kim, D. E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C. E., Bonneau, R., Rohl, C. A., and Baker, D. (2003). Automated prediction of CASP-5 structures using the Robetta server. Proteins *53 Suppl 6*, 524-533.

Choi, K. H., Shi, J., Hopkins, C. E., Tolan, D. R., and Allen, K. N. (2001). Snapshots of catalysis: the structure of fructose-1,6-(bis)phosphate aldolase covalently bound to the substrate dihydroxyacetone phosphate. Biochemistry *40*, 13868-13875.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. Nature *357*, 543-544.

Chothia, C., and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. Embo J *5*, 823-826.

Chou, P. Y., and Fasman, G. D. (1974). Prediction of protein conformation. Biochemistry *13*, 222-245.

Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004). The Jalview Java alignment editor. Bioinformatics *20*, 426-427.

Cohen, F. E., and Sternberg, M. J. (1980). On the prediction of protein structure: The significance of the root- mean-square deviation. J Mol Biol *138*, 321-333.

Cokol, M., Nair, R., and Rost, B. (2000). Finding nuclear localization signals. EMBO Rep *1*, 411-415.

Copley, R. R., and Bork, P. (2000). Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. J Mol Biol *303*, 627-641.

Copley, R. R., Doerks, T., Letunic, I., and Bork, P. (2002a). Protein domain analysis in the era of complete genomes. FEBS Lett *513*, 129-134.

Copley, R. R., Letunic, I., and Bork, P. (2002b). Genome and protein evolution in eukaryotes. Curr Opin Chem Biol *6*, 39-45.

Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L., and Elofsson, A. (2001). A study of quality measures for protein threading models. BMC Bioinformatics *2*, 5.

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

Dalby, A., Dauter, Z., and Littlechild, J. A. (1999). Crystal structure of human muscle aldolase complexed with fructose 1,6-bisphosphate: mechanistic implications. Protein Sci *8*, 291-297.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In Atlas of protein sequence and structure, M. O. Dayhoff, and R. V. Eck, eds. (Silver Spring, Md., National Biomedical Research Foundation), pp. 345-352.

de Beer, T., Hoofnagle, A. N., Enmon, J. L., Bowers, R. C., Yamabhai, M., Kay, B. K., and Overduin, M. (2000). Molecular mechanism of NPF recognition by EH domains. Nat Struct Biol *7*, 1018-1022.

de Celis, J. F., and Garcia-Bellido, A. (1994). Roles of the Notch gene in Drosophila wing morphogenesis. Mech Dev *46*, 109-122.

Denessiouk, K. A., Lehtonen, J. V., Korpela, T., and Johnson, M. S. (1998). Two "unrelated" families of ATP-dependent enzymes share extensive structural similarities about their cofactor binding sites. Protein Sci *7*, 1136-1146.

Diaz-Benjumea, F. J., and Cohen, S. M. (1995). Serrate signals through Notch to establish a Wingless-dependent organizer at the dorsal/ventral compartment boundary of the Drosophila wing. Development *121*, 4215-4225.

Dodson, G., and Wlodawer, A. (1998). Catalytic triads and their relatives. Trends Biochem Sci *23*, 347-352.

Doench, J. G., Petersen, C. P., and Sharp, P. A. (2003). siRNAs can function as miRNAs. Genes Dev *17*, 438-442.

Dosztanyi, Z., Fiser, A., and Simon, I. (1997). Stabilization centers in proteins: identification, characterization and predictions. J Mol Biol *272*, 597-612.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). Biological sequence analysis (Cambridge, UK, Cambridge University Press).

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494-498.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol *300*, 1005-1016.

Endicott, J. A., and Nurse, P. (1995). The cell cycle and suc1: from structure to function? Structure *3*, 321-325.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). MicroRNA targets in Drosophila. Genome Biol *5*, R1.

Enright, A. J., and Ouzounis, C. A. (2001). Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol *2*, RESEARCH0034.

Farber, G. K., and Petsko, G. A. (1990). The evolution of alpha/beta barrel enzymes. Trends Biochem Sci *15*, 228-234.

Feigelson, H. S., and Henderson, B. E. (1996). Estrogens and breast cancer. Carcinogenesis *17*, 2279-2284.

Feng, D. F., and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol *25*, 351-360.

Fetrow, J. S., Siew, N., and Skolnick, J. (1999). Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase- 1 subfamily. Faseb J *13*, 1866-1874.

Fetrow, J. S., and Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol *281*, 949-968.

Fischer, D. (2003). 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins *51*, 434-441.

Fischer, D., Wolfson, H., Lin, S. L., and Nussinov, R. (1994). Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. Protein Sci *3*, 769-778.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. Syst Zool *19*, 99-113.

Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K. B., Bairoch, A., Schomburg, D., Tipton, K. F., and Apweiler, R. (2004). IntEnz, the integrated relational enzyme database. Nucleic Acids Res *32 Database issue*, D434-437.

Flouriot, G., Brand, H., Denger, S., Metivier, R., Kos, M., Reid, G., Sonntag-Buck, V., and Gannon, F. (2000). Identification of a new isoform of the human estrogen receptor-alpha (hER-alpha) that is encoded by distinct transcripts and that is able to repress hER-alpha activation function 1. Embo J *19*, 4688-4700.

Gamblin, S. J., Cooper, B., Millar, J. R., Davies, G. J., Littlechild, J. A., and Watson, H. C. (1990). The crystal structure of human muscle aldolase at 3.0 A resolution. FEBS Lett *262*, 282-286.

Garnier, J., Osguthorpe, D. J., and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol *120*, 97-120.

Gerstein, M., and Levitt, M. (1997). A structural census of the current population of protein sequences. Proc Natl Acad Sci U S A *94*, 11911-11916.

Gibrat, J. F., Madej, T., and Bryant, S. H. (1996). Surprising similarities in structure comparison. Curr Opin Struct Biol *6*, 377-385.

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

Godzik, A., Kolinski, A., and Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. J Mol Biol *227*, 227-238.

Gomis-Ruth, F. X., Companys, V., Qian, Y., Fricker, L. D., Vendrell, J., Aviles, F. X., and Coll, M. (1999). Crystal structure of avian carboxypeptidase D domain II: a prototype for the regulatory metallocarboxypeptidase subfamily. Embo J *18*, 5817-5826.

Grad, Y., Aach, J., Hayes, G. D., Reinhart, B. J., Church, G. M., Ruvkun, G., and Kim, J. (2003). Computational and experimental identification of C. elegans microRNAs. Mol Cell *11*, 1253-1263.

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

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). Rfam: an RNA family database. Nucleic Acids Res *31*, 439-441.

Groll, M., Bajorek, M., Kohler, A., Moroder, L., Rubin, D. M., Huber, R., Glickman, M. H., and Finley, D. (2000). A gated channel into the proteasome core particle. Nat Struct Biol *7*, 1062-1067.

Gumbel, E. J. (1958). Statistics of extremes (New York,, Columbia University Press).

Gutteridge, A., Bartlett, G. J., and Thornton, J. M. (2003). Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol *330*, 719-734.

Hadley, C., and Jones, D. T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure Fold Des *7*, 1099-1112.

Hall, I. M., Shankaranarayana, G. D., Noma, K., Ayoub, N., Cohen, A., and Grewal, S. I. (2002). Establishment and maintenance of a heterochromatin domain. Science *297*, 2232-2237.

Hamelryck, T. (2003). Efficient identification of side-chain patterns using a multidimensional index tree. Proteins *51*, 96-108.

Harrison, A., Pearl, F., Mott, R., Thornton, J., and Orengo, C. (2002). Quantifying the similarities within fold space. J Mol Biol *323*, 909-926.

Harrison, P. M., and Sternberg, M. J. (1996). The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. J Mol Biol *264*, 603-623.

Hegyi, H., and Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol *288*, 147-164.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol *216*, 167-180.

Henikoff, S., and Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. Nucleic Acids Res *19*, 6565-6572.

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A *89*, 10915-10919.

Henikoff, S., and Henikoff, J. G. (1994). Position-based sequence weights. J Mol Biol *243*, 574-578.

Hilgers, M. T., and Ludwig, M. L. (2001). Crystal structure of the quorum-sensing protein LuxS reveals a catalytic metal site. Proc Natl Acad Sci U S A *98*, 11169-11174.

Hipfner, D. R., Weigmann, K., and Cohen, S. M. (2002). The bantam gene regulates Drosophila growth. Genetics *161*, 1527-1537.

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

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

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

Holm, L., and Sander, C. (1995). Dali: a network tool for protein structure comparison. Trends Biochem Sci *20*, 478-480.

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

Holm, L., and Sander, C. (1997). Decision support system for the evolutionary classification of protein structures. Proc Int Conf Intell Syst Mol Biol *5*, 140-146.

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

Hurley, J. H., Anderson, D. E., Beach, B., Canagarajah, B., Ho, Y. S., Jones, E., Miller, G., Misra, S., Pearson, M., Saidi, L*., et al.* (2002). Structural genomics and signaling domains. Trends Biochem Sci *27*, 48-53.

Hutvagner, G., and Zamore, P. D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. Science *297*, 2056-2060.

Jambon, M., Imberty, A., Deleage, G., and Geourjon, C. (2003). A new bioinformatic approach to detect common 3D sites in protein structures. Proteins *52*, 137-145.

Jones, D. T. (1997). Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins *Suppl 1*, 185-191.

Jones, D. T. (1999a). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol *287*, 797-815.

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

Jones, D. T., and Swindells, M. B. (2002). Getting the most from PSI-BLAST. Trends Biochem Sci *27*, 161-164.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. Nature *358*, 86-89.

Jones, S., and Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. J Mol Biol *272*, 121-132.

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

Kabsch, W., and Sander, C. (1983b). How good are predictions of protein secondary structure? FEBS Lett *155*, 179-182.

Kabsch, W., and Sander, C. (1984). On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proc Natl Acad Sci U S A *81*, 1075-1078.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res *32 Database issue*, D277-280.

Kannan, N., and Vishveshwara, S. (1999). Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol *292*, 441-464.

Karlin, S., and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A *87*, 2264-2268.

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

Kasschau, K. D., Xie, Z., Allen, E., Llave, C., Chapman, E. J., Krizan, K. A., and Carrington, J. C. (2003). P1/HC-Pro, a viral suppressor of RNA silencing, interferes with Arabidopsis development and miRNA unction. Dev Cell *4*, 205-217.

Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol *299*, 499-520.

Kendall, M. G., Stuart, A., and Ord, J. K. (1977). The advanced theory of statistics, 4th edn (London, C. Griffin).

Kennelly, P. J. (2002). Protein kinases and protein phosphatases in prokaryotes: a genomic perspective. FEMS Microbiol Lett *206*, 1-8.

Kennelly, P. J. (2003). Archaeal protein kinases and protein phosphatases: insights from genomics and biochemistry. Biochem J *370*, 373-389.

Ketting, R. F., Haverkamp, T. H., van Luenen, H. G., and Plasterk, R. H. (1999). Mut-7 of C. elegans, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. Cell *99*, 133-141.

Khvorova, A., Reynolds, A., and Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209-216.

Kim, J., Krichevsky, A., Grad, Y., Hayes, G. D., Kosik, K. S., Church, G. M., and Ruvkun, G. (2004). Identification of many microRNAs that copurify with polyribosomes in mammalian neurons. Proc Natl Acad Sci U S A *101*, 360-365.

Kim, S. H., Shin, D. H., Choi, I. G., Schulze-Gahmen, U., Chen, S., and Kim, R. (2003). Structure-based functional inference in structural genomics. J Struct Funct Genomics *4*, 129-135.

Kleywegt, G. J. (1999). Recognition of spatial motifs in protein structures. J Mol Biol *285*, 1887-1897.

Kobayashi, N., and Go, N. (1997a). ATP binding proteins with different folds share a common ATP-binding structural motif. Nat Struct Biol *4*, 6-7.

Kobayashi, N., and Go, N. (1997b). A method to search for similar protein local structures at ligand binding sites and its application to adenine recognition. Eur Biophys J *26*, 135-144.

Kollman, J. M., and Doolittle, R. F. (2000). Determining the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs. J Mol Evol *51*, 173-181.

Koonin, E. V., Wolf, Y. I., and Karev, G. P. (2002). The structure of the protein universe and genome evolution. Nature *420*, 218-223.

Koretke, K. K., Russell, R. B., Copley, R. R., and Lupas, A. N. (1999). Fold recognition using sequence and secondary structure information. Proteins *Suppl 3*, 141-148.

Kraulis, P. J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J Appl Cryst *24*, 946-950.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol *235*, 1501-1531.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853-858.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. Curr Biol *12*, 735-739.

Lai, E. C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet *30*, 363-364.

Lai, E. C., Burks, C., and Posakony, J. W. (1998). The K box, a conserved 3' UTR sequence motif, negatively regulates accumulation of enhancer of split complex transcripts. Development *125*, 4077-4088.

Lai, E. C., and Posakony, J. W. (1997). The Bearded box, a novel 3' UTR sequence motif, mediates negative post-transcriptional regulation of Bearded and Enhancer of split Complex gene expression. Development *124*, 4847-4856.

Lai, E. C., and Posakony, J. W. (1998). Regulation of Drosophila neurogenesis by RNA:RNA duplexes? Cell *93*, 1103-1104.

Lai, E. C., Tomancak, P., Williams, R. W., and Rubin, G. M. (2003). Computational identification of Drosophila microRNA genes. Genome Biol *4*, R42.

Landgraf, R., Xenarios, I., and Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol *307*, 1487-1502.

Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., and Wilmanns, M. (2000). Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. Science *289*, 1546-1550.

Laskowski, R. A. (2001). PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res *29*, 221-222.

Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L., and Thornton, J. M. (1997). PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci *22*, 488-490.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science *294*, 858-862.

Lebherz, H. G., and Rutter, W. J. (1969). Distribution of fructose diphosphate aldolase variants in biological systems. Biochemistry *8*, 109-121.

Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. J Mol Biol *55*, 379-400.

Lee, R. C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science *294*, 862-864.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415-419.

Lesk, A. M., Lo Conte, L., and Hubbard, T. J. (2001). Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins *Suppl 5*, 98-118.

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

Levitt, M., and Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci U S A *95*, 5913-5920.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787-798.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol *257*, 342-358.

Lichtarge, O., and Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol *12*, 21-27.

Ligoxygakis, P., Bray, S. J., Apidianakis, Y., and Delidakis, C. (1999). Ectopic expression of individual E(spl) genes has differential effects on different cell fate decisions and underscores the biphasic requirement for notch activity in wing margin establishment in Drosophila. Development *126*, 2205-2214.

Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., and Bartel, D. P. (2003a). Vertebrate microRNA genes. Science *299*, 1540.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B., and Bartel, D. P. (2003b). The microRNAs of Caenorhabditis elegans. Genes Dev *17*, 991-1008.

Lim, W. A., and Richards, F. M. (1994). Critical residues in an SH3 domain from Sem-5 suggest a mechanism for proline-rich peptide recognition. Nat Struct Biol *1*, 221-225.

Lin, S. L., Nussinov, R., Fischer, D., and Wolfson, H. J. (1994). Molecular surface representations by sparse critical points. Proteins *18*, 94-101.

Lin, S. Y., Johnson, S. M., Abraham, M., Vella, M. C., Pasquinelli, A., Gamberi, C., Gottlieb, E., and Slack, F. J. (2003). The C elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. Dev Cell *4*, 639-650.

Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2003). Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain. Nature *426*, 465-469.

Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science *297*, 2053-2056.

Lorentzen, E., Pohl, E., Zwart, P., Stark, A., Russell, R. B., Knura, T., Hensel, R., and Siebers, B. (2003). Crystal structure of an archaeal class I aldolase and the evolution of (betaalpha)8 barrel proteins. J Biol Chem *278*, 47253-47260.

Lu, L., Lu, H., and Skolnick, J. (2002). MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins *49*, 350-364.

Lund, E., Guttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science *303*, 95-98.

Lundstrom, J., Rychlewski, L., Bujnicki, J., and Elofsson, A. (2001). Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci *10*, 2354-2362.

Luscombe, N. M., Laskowski, R. A., Westhead, D. R., Milburn, D., Jones, S., Karmirantzou, M., and Thornton, J. M. (1998). New tools and resources for analysing protein structures and their interactions. Acta Crystallogr D Biol Crystallogr *54*, 1132-1138.

Macias, M. J., Hyvonen, M., Baraldi, E., Schultz, J., Sudol, M., Saraste, M., and Oschkinat, H. (1996). Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. Nature *382*, 646-649.

Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E., and Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol *316*, 139-154.

Madsen, D., and Kleywegt, G. J. (2002). Interactive motif and fold recognition in protein structures. J Appl Cryst *35*, 137-139.

Mande, S. C., Mainfroid, V., Kalk, K. H., Goraj, K., Martial, J. A., and Hol, W. G. (1994). Crystal structure of recombinant human triosephosphate isomerase at 2.8 A resolution. Triosephosphate isomerase-related human genetic disorders and comparison with the trypanosomal enzyme. Protein Sci *3*, 810-821.

Marquez, J. A., Hasenbein, S., Koch, B., Fieulaine, S., Nessler, S., Russell, R. B., Hengstenberg, W., and Scheffzek, K. (2002). Structure of the full-length HPr kinase/phosphatase from Staphylococcus xylosus at 1.95 A resolution: Mimicking the product/substrate of the phospho transfer reactions. Proc Natl Acad Sci U S A *99*, 3458-3463.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct *29*, 291-325.

Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell *110*, 563-574.

Mas, J. M., Aloy, P., Marti-Renom, M. A., Oliva, B., Blanco-Aparicio, C., Molina, M. A., de Llorens, R., Querol, E., and Aviles, F. X. (1998). Protein similarities beyond disulphide bridge topology. J Mol Biol *284*, 541-548.

Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol *288*, 911-940.

Mavridis, I. M., Hatada, M. H., Tulinsky, A., and Lebioda, L. (1982). Structure of 2-keto-3-deoxy-6-phosphogluconate aldolase at 2 . 8 A resolution. J Mol Biol *162*, 419-444.

McLachlan, A. D. (1979). Gene duplications in the structural evolution of chymotrypsin. J Mol Biol *128*, 49-79.

McLachlan, A. D., and Shotton, D. M. (1971). Structural similarities between alpha-lytic protease of Myxobacter 495 and elastase. Nat New Biol *229*, 202-205.

Merritt, E. A., and Murphy, M. E. P. (1994). Raster3D Version 2.0. A program for photorealistic molecular graphics. Acta Crystallogr D *50*, 869-873.

Metivier, R., Petit, F. G., Valotaire, Y., and Pakdel, F. (2000). Function of N-terminal transactivation domain of the estrogen receptor requires a potential alpha-helical structure and is negatively regulated by the A domain. Mol Endocrinol *14*, 1849-1871.

Metivier, R., Stark, A., Flouriot, G., Hubner, M. R., Brand, H., Penot, G., Manu, D., Denger, S., Reid, G., Kos, M.*, et al.* (2002). A dynamic structural model for estrogen receptor-alpha activation by ligands, emphasizing the role of interactions between distant A and E domains. Mol Cell *10*, 1019-1032.

Micchelli, C. A., Rulifson, E. J., and Blair, S. S. (1997). The function and regulation of cut expression on the wing margin of Drosophila: Notch, Wingless and a dominant negative role for Delta and Serrate. Development *124*, 1485-1495.

Miyazawa, S., and Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol *256*, 623-644.

Moss, E. G., Lee, R. C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. Cell *88*, 637-646.

Moss, E. G., and Tang, L. (2003). Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites. Dev Biol *258*, 432-442.

Mott, R. F., Kirkwood, T. B., and Curnow, R. N. (1990). Tests for the statistical significance of protein sequence similarities in data-bank searches. Protein Eng *4*, 149-154.

Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. Proteins *23*, ii-v.

Mura, C., Katz, J. E., Clarke, S. G., and Eisenberg, D. (2003). Structure and function of an archaeal homolog of survival protein E (SurEalpha): an acid phosphatase with purine nucleotide specificity. J Mol Biol *326*, 1559-1575.

Murzin, A. G. (1993a). Can homologous proteins evolve different enzymatic activities? Trends Biochem Sci *18*, 403-405.

Murzin, A. G. (1993b). Sweet-tasting protein monellin is related to the cystatin family of thiol proteinase inhibitors. J Mol Biol *230*, 689-694.

Murzin, A. G. (1998). How far divergent evolution goes in proteins. Curr Opin Struct Biol *8*, 380-387.

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

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

Nagano, N., Orengo, C. A., and Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol *321*, 741-765.

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

Nelson, H. C. (1995). Structure and function of DNA-binding proteins. Curr Opin Genet Dev *5*, 180-189.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng *10*, 1-6.

Nilsson, S., Makela, S., Treuter, E., Tujague, M., Thomsen, J., Andersson, G., Enmark, E., Pettersson, K., Warner, M., and Gustafsson, J. A. (2001). Mechanisms of estrogen action. Physiol Rev *81*, 1535-1565.

Nobeli, I., Laskowski, R. A., Valdar, W. S., and Thornton, J. M. (2001). On the molecular discrimination between adenine and guanine by proteins. Nucleic Acids Res *29*, 4294-4309.

Norel, R., Fischer, D., Wolfson, H. J., and Nussinov, R. (1994). Molecular surface recognition by a computer vision-based technique. Protein Eng *7*, 39-46.

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

Nussinov, R., and Wolfson, H. J. (1991). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci U S A *88*, 10495-10499.

Ofran, Y., and Rost, B. (2003). Analysing six types of protein-protein interfaces. J Mol Biol *325*, 377-387.

Oganesyan, V., Busso, D., Brandsen, J., Chen, S., Jancarik, J., Kim, R., and Kim, S. H. (2003). Structure of the hypothetical protein AQ_1354 from Aquifex aeolicus. Acta Crystallogr D Biol Crystallogr *59*, 1219-1223.

Oliveira, L., Paiva, P. B., Paiva, A. C., and Vriend, G. (2003). Identification of functionally conserved residues with the use of entropy-variability plots. Proteins *52*, 544-552.

Olsen, P. H., and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev Biol *216*, 671-680.

Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994a). Protein superfamilies and domain superfolds. Nature *372*, 631-634.

Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994b). Protein superfamilies and domain superfolds. Nature *372*, 631-634.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. Structure *5*, 1093-1108.

Orengo, C. A., and Taylor, W. R. (1996). SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol *266*, 617-635.

Ota, M., Kinoshita, K., and Nishikawa, K. (2003). Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J Mol Biol *327*, 1053-1064.

Pace, N. R. (1991). Origin of life--facing up to the physical setting. Cell *65*, 531-533.

Pakdel, F., Metivier, R., Flouriot, G., and Valotaire, Y. (2000). Two estrogen receptor (ER) isoforms with different estrogen dependencies are generated from the trout ER gene. Endocrinology *141*, 571-580.

Palatnik, J. F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J. C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. Nature *425*, 257-263.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol *284*, 1201-1210.

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Muller, P*., et al.* (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature *408*, 86-89.

Pawson, T., and Gish, G. D. (1992). SH2 and SH3 domains: from structure to function. Cell *71*, 359-362.

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. Protein Sci *4*, 1145-1160.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. J Mol Biol *276*, 71-84.

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A *85*, 2444-2448.

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

Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D. M., Ausiello, G., Brannetti, B., Costantini, A*., et al.* (2003). ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res *31*, 3625-3630.

Rajewsky, N., and Socci, N. D. (2004). Computational identification of microRNA targets. Developmental Biology *267*, 529-535.

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901-906.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). MicroRNAs in plants. Genes Dev *16*, 1616-1626.

Reva, B. A., Finkelstein, A. V., and Skolnick, J. (1998). What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? Fold Des *3*, 141-147.

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. Cell *110*, 513-520.

Rost, B. (2002). Enzyme function less conserved than anticipated. J Mol Biol *318*, 595-608.

Rost, B., and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol *232*, 584-599.

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

Rulifson, E. J., and Blair, S. S. (1995). Notch regulates wingless expression and is not required for reception of the paracrine wingless signal during wing margin neurogenesis in Drosophila. Development *121*, 2813-2824.

Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol *279*, 1211-1227.

Russell, R. B., and Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins *14*, 309-323.

Russell, R. B., and Barton, G. J. (1993). The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J Mol Biol *234*, 951-957.

Russell, R. B., and Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. J Mol Biol *244*, 332-350.

Russell, R. B., Copley, R. R., and Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. J Mol Biol *259*, 349-365.

Russell, R. B., Marquez, J. A., Hengstenberg, W., and Scheffzek, K. (2002). Evolutionary relationship between the bacterial HPr kinase and the ubiquitous PEP-carboxykinase: expanding the P-loop nucleotidyl transferase superfamily. FEBS Lett *517*, 1-6.

Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A., and Sternberg, M. J. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol *269*, 423-439.

Russell, R. B., Sasieni, P. D., and Sternberg, M. J. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol *282*, 903-918.

Ruzheinikov, S. N., Das, S. K., Sedelnikova, S. E., Hartley, A., Foster, S. J., Horsburgh, M. J., Cox, A. G., McCleod, C. W., Mekhalfia, A., Blackburn, G. M.*, et al.* (2001). The 1.2 A structure of a novel quorum-sensing protein, Bacillus subtilis LuxS. J Mol Biol *313*, 111-122.

Sali, A. (1998). 100,000 protein structures for the biologist. Nat Struct Biol *5*, 1029-1032.

Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M. A., Madhusudhan, M. S., Mirkovic, N., and Sali, A. (2000). Protein structure modeling for structural genomics. Nat Struct Biol *7 Suppl*, 986-990.

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

Sanishvili, R., Yakunin, A. F., Laskowski, R. A., Skarina, T., Evdokimova, E., Doherty-Kirby, A., Lajoie, G. A., Thornton, J. M., Arrowsmith, C. H., Savchenko, A.*, et al.* (2003). Integrating structure, bioinformatics, and enzymology to discover function: BioH, a new carboxylesterase from Escherichia coli. J Biol Chem *278*, 26039-26045.

Sayle, R. A., and Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. Trends Biochem Sci *20*, 374.

Schauder, S., Shokat, K., Surette, M. G., and Bassler, B. L. (2001). The LuxS family of bacterial autoinducers: biosynthesis of a novel quorum-sensing signal molecule. Mol Microbiol *41*, 463-476.

Schmitt, S., Kuhn, D., and Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol *323*, 387-406.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004). BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res *32 Database issue*, D431-433.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A *95*, 5857-5864.

Schwarz, D. S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P. D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199-208.

Seffernick, J. L., de Souza, M. L., Sadowsky, M. J., and Wackett, L. P. (2001). Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. J Bacteriol *183*, 2405-2410.

Sheridan, R. P., Dixon, J. S., Venkataraghavan, R., Kuntz, I. D., and Scott, K. P. (1985). Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. Biopolymers *24*, 1995-2023.

Shi, L., Potts, M., and Kennelly, P. J. (1998). The serine, threonine, and/or tyrosine-specific protein kinases and protein phosphatases of prokaryotic organisms: a family portrait. FEMS Microbiol Rev *22*, 229-253.

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

Siebers, B., Brinkmann, H., Dorr, C., Tjaden, B., Lilie, H., van der Oost, J., and Verhees, C. H. (2001). Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase. J Biol Chem *276*, 28710-28718.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol *213*, 859-883.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. Curr Opin Struct Biol *5*, 229-235.

Slack, F. J., Basson, M., Liu, Z., Ambros, V., Horvitz, H. R., and Ruvkun, G. (2000). The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. Mol Cell *5*, 659-669.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.

Song, J. J., Liu, J., Tolia, N. H., Schneiderman, J., Smith, S. K., Martienssen, R. A., Hannon, G. J., and Joshua-Tor, L. (2003). The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. Nat Struct Biol *10*, 1026-1032.

Stark, A., Brennecke, J., Russell, R. B., and Cohen, S. M. (2003a). Identification of Drosophila MicroRNA Targets. PLoS Biol *1*, E60.

Stark, A., and Russell, R. B. (2003). Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. Nucleic Acids Res *31*, 3341-3344.

Stark, A., Sunyaev, S., and Russell, R. B. (2003b). A model for statistical significance of local similarities in structure. J Mol Biol *326*, 1307-1316.

Tabara, H., Sarkissian, M., Kelly, W. G., Fleenor, J., Grishok, A., Timmons, L., Fire, A., and Mello, C. C. (1999). The rde-1 gene, RNA interference, and transposon silencing in C. elegans. Cell *99*, 123-132.

Tang, G., Reinhart, B. J., Bartel, D. P., and Zamore, P. D. (2003). A biochemical framework for RNA silencing in plants. Genes Dev *17*, 49-63.

Tao, X., Khayat, R., Christendat, D., Savchenko, A., Xu, X., Goldsmith-Fischman, S., Honig, B., Edwards, A., Arrowsmith, C. H., and Tong, L. (2003). Crystal structures of MTH1187 and its yeast ortholog YBL001c. Proteins *52*, 478-480.

Teichmann, S. A., Murzin, A. G., and Chothia, C. (2001). Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struct Biol *11*, 354-363.

Tendulkar, A. V., Wangikar, P. P., Sohoni, M. A., Samant, V. V., and Mone, C. Y. (2003). Parameterization and classification of the protein universe via geometric techniques. J Mol Biol *334*, 157-172.

Teplyakov, A., Obmolova, G., Tordova, M., Thanki, N., Bonander, N., Eisenstein, E., Howard, A. J., and Gilliland, G. L. (2002). Crystal structure of the YjeE protein from Haemophilus influenzae: a putative Atpase involved in cell wall synthesis. Proteins *48*, 220-226.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). 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.

Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., and Orengo, C. A. (2000). From structure to function: approaches and limitations. Nat Struct Biol *7 Suppl*, 991-994.

Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol *333*, 863-882.

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol *307*, 1113-1143.

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2002). Sequence and structural differences between enzyme and nonenzyme homologs. Structure (Camb) *10*, 1435-1451.

Tomarev, S. I., and Zinovieva, R. D. (1988). Squid major lens polypeptides are homologous to glutathione S-transferases subunits. Nature *336*, 86-88.

Tramontano, A., and Morea, V. (2003). Assessment of homology-based predictions in CASP5. Proteins *53 Suppl 6*, 352-368.

Vingron, M. (2001). Bioinformatics needs to adopt statistical thinking. Bioinformatics *17*, 389-390.

Vitkup, D., Melamud, E., Moult, J., and Sander, C. (2001). Completeness in structural genomics. Nat Struct Biol *8*, 559-566.

Vogt, G., Etzold, T., and Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J Mol Biol *249*, 816-831.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. Nature *417*, 399-403.

Walden, H., Bell, G. S., Russell, R. J., Siebers, B., Hensel, R., and Taylor, G. L. (2001). Tiny TIM: a small, tetrameric, hyperthermostable triosephosphate isomerase. J Mol Biol *306*, 745-757.

Wallace, A. C., Borkakoti, N., and Thornton, J. M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci *6*, 2308-2323.

Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci *5*, 1001-1013.

Wang, W. K., Tereshko, V., Boccuni, P., MacGrogan, D., Nimer, S. D., and Patel, D. J. (2003). Malignant brain tumor repeats: a three-leaved propeller architecture with ligand/peptide binding pockets. Structure (Camb) *11*, 775-789.

Wangikar, P. P., Tendulkar, A. V., Ramya, S., Mali, D. N., and Sarawagi, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. J Mol Biol *326*, 955-978.

Wassenegger, M., Heimes, S., Riedel, L., and Sanger, H. L. (1994). RNA-directed de novo methylation of genomic sequences in plants. Cell *76*, 567-576.

Waterman, M. S., and Vingron, M. (1994). Rapid and accurate estimates of statistical significance for sequence data base searches. Proc Natl Acad Sci U S A *91*, 4625-4628.

Webb, E. C., and International Union of Biochemistry and Molecular Biology. Nomenclature Committee. (1992). Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes (San Diego ; London, Academic Press).

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell *75*, 855-862.

Willett, P. (1987). A review of chemical structure retrieval systems. J Chemometrics *1*, 139-155.

Wilmanns, M., Hyde, C. C., Davies, D. R., Kirschner, K., and Jansonius, J. N. (1991). Structural conservation in parallel beta/alpha-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. Biochemistry *30*, 9161-9169.

Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol *297*, 233-249.

Wilson, R. A., Maughan, W. N., Kremer, L., Besra, G. S., and Futterer, K. (2004). The Structure of Mycobacterium tuberculosis MPT51 (FbpC1) Defines a New Family of Non-catalytic alpha/beta Hydrolases. J Mol Biol *335*, 519-530.

Wistow, G., and Piatigorsky, J. (1987). Recruitment of enzymes as lens structural proteins. Science *236*, 1554-1556.

Woese, C. R. (1987). Bacterial evolution. Microbiol Rev *51*, 221-271.

Xie, Z., Kasschau, K. D., and Carrington, J. C. (2003). Negative feedback regulation of Dicer-Like1 in Arabidopsis by microRNA-guided mRNA degradation. Curr Biol *13*, 784-789.

Xu, P., Vernooy, S. Y., Guo, M., and Hay, B. A. (2003). The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. Curr Biol *13*, 790-795.

Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S., and Cantley, L. C. (2001). A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat Biotechnol *19*, 348-353.

Yan, K. S., Yan, S., Farooq, A., Han, A., Zeng, L., and Zhou, M. M. (2003). Structure and conserved RNA binding of the PAZ domain. Nature *426*, 468-474.

Yi, R., Qin, Y., Macara, I. G., and Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev *17*, 3011-3016.

Yu, H., Rosen, M. K., Shin, T. B., Seidel-Dugan, C., Brugge, J. S., and Schreiber, S. L. (1992). Solution structure of the SH3 domain of Src and identification of its ligand-binding site. Science *258*, 1665-1668.

Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R. R., Christophides, G. K., Thomasova, D., Holt, R. A., Subramanian, G. M., *et al.* (2002). Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. Science *298*, 149-159.

Zeng, Y., Wagner, E. J., and Cullen, B. R. (2002). Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol Cell *9*, 1327-1333.

Zhang, C., and Kim, S. H. (2003). Overview of structural genomics: from structure to function. Curr Opin Chem Biol *7*, 28-32.

Zhang, R. G., Andersson, C. E., Skarina, T., Evdokimova, E., Edwards, A. M., Joachimiak, A., Savchenko, A., and Mowbray, S. L. (2003). The 2.2 A resolution structure of RpiB/AlsB from Escherichia coli illustrates a new approach to the ribose-5-phosphate isomerase reaction. J Mol Biol *332*, 1083-1094.

Zuker, M., Mathews, D. H., and Turner, D. H. (1999). Algorithms and thermodynamics for RNA secondary structure prediction. In A practical guide in RNA biochemistry and biotechnology, J. Barciszewski, and B. F. C. Clark, eds. (Kluwer Academic Publishers), pp. 11-43.

# 5 Ausführliche Zusammenfassung

Die Anzahl bekannter Proteinstrukturen wächst exponentiell und sogenannte „Structural Genomics“ Projekte haben es sich zum Ziel gesetzt, die Strukturen aller Proteine aufzuklären, um dadurch deren Funktionen zu bestimmen. Dies wird oft durch Vergleiche der Gesamtstrukturen von Proteinen erreicht, da ähnliche Strukturen oft auf Homologie der Proteine und auf funktionelle Gemeinsamkeiten schließen lassen. Hinweise auf bestimmte Funktionen lassen sich aber auch durch Vergleiche lokaler struktureller Muster – zum Beispiel katalytischer Zentren – bekommen, die in Proteinen mit unterschiedlichen Gesamtstrukturen auftreten können. Ich habe in meiner Dissertation eine Methode zum Vergleich dieser Muster entwickelt und ein Modell zur Berechnung der statistischen Signifikanz von Suchergebnissen hergeleitet, das die Unterscheidung von bedeutsamen und zufällig auftretenden Ähnlichkeiten erlaubt. Im Internet ist eine einfach zu bedienende Benutzeroberfläche meiner Methode für die Funktionsuntersuchung von Proteinstrukturen verfügbar (http://pints.embl.de). Dieser Server erlaubt dem Nutzer Strukturen von Proteinen unbekannter Funktion mit Datenbanken funktionell wichtiger Muster zu vergleichen, nach dem Auftreten eines bestimmten Musters in verschiedenen Proteinen zu suchen aber auch alle gemeinsamen Muster zweier Proteine zu bestimmen. Um eine Vorstellung der praktischen Anwendbarkeit dieser Methode zu bekommen, habe ich über 250 Proteinstrukturen untersucht, die von den oben genannten „Structural Genomics“ Projekten aufgeklärt wurden. Signifikante Suchergebnisse haben dabei sowohl Funktionen bestätigt, die aufgrund ähnlicher Gesamtstrukturen vermutet wurden, als auch funktionelle Ähnlichkeiten in Proteinen unterschiedlicher Gesamtstruktur vorhergesagt. Die Methode stellt daher eine wichtige Ergänzung zu dem oben genannten Vergleich von Gesamtstrukturen dar. Durch eine detaillierte Analyse der Struktur der Fruktose-1,6-Bisphosophat Aldolase von Archaea, konnte ich zusätzlich die Homologie einiger Enzymfamilien zeigen.

Im zweiten Teil meiner Dissertation präsentiere ich eine systematische computerbasierte Suche nach *Drosophila* Genen, die von microRNAs (miRNAs) reguliert werden („Targets“). miRNAs sind kurze RNA Moleküle mit einer Länge von ungefähr 22 Nukleotiden, die in Tieren die Translation ihrer Targets blockieren, indem sie an komplementäre Stellen in deren 3’ untranslatierten Bereichen binden. Methoden zur Vorhersage von miRNA Targets wurden dringend benötigt, da Targets für nur drei der insgesamt 700 bekannten miRNAs beschrieben

waren. Meine Suche basiert auf einer systematischen Untersuchung der bekannten miRNA-Target Komplexe, und kombiniert Sequenzvergleiche mit Methoden zur Vorhersage von RNA-Strukturen. Zusätzlich wurde angenommen, dass biologisch wichtige miRNA-Target Interaktionen in den eng verwandten Spezies *Drosophila melanogaster* und *Drosophila pseudoobscura* identisch und miRNA Bindestellen daher konserviert sind. Dadurch konnte ich alle bekannten Targets in einer Datenbank aller Drosophila Gene zuverlässig detektieren und Target-Vorhersagen treffen, von denen sechs experimentell bestätigt wurden und viele andere mit hoher Wahrscheinlichkeit ebenfalls zutreffend sind. Da eine statistische Analyse jedoch zeigte, dass das vorhandenen Signal oft zu gering für zuverlässige Aussagen war, wurde als unabhängige Bestätigung *Anopheles gambiae* systematisch auf das Vorhandensein entsprechender Bindestellen untersucht. Spezifische Funktionen konnten den miRNAs *miR-7* (Kontrolle von *Notch*-aktivierten Genen), *miR-2* (Inhibierung von Apoptose) und *miR-277* (Regulation des Abbaus einiger essentieller Aminosäuren) zugeschrieben werden. miRNAs scheinen zudem generell Transkriptionsfaktoren zu regulieren und an der Entwicklung des Nervensystems beteiligt zu sein. Sowohl die Ergebnisse der Untersuchung von miRNA-Target Komplexen und die getroffenen Target Vorhersagen stellen eine wertvolle Hilfe zur Erforschung von miRNAs dar.

# 6 Acknowledgements

I would like to thank everybody that contributed to this thesis over the past 3 years. In particular, I want to thank…

Rob Russell, first of all, for having me in his group, lots of interesting discussions and ideas, and his generous support.

The members of my thesis advisory committee, Carsten Schultz, Peer Bork and especially my external supervisor and "Doktorvater" Prof. Dr. Dietmar Schomburg for helpful discussions and advice. Special thanks also to Prof. Dr. Sabine Waffenschmidt for kindly agreeing to referee this thesis.

Shamil Sunyaev, Raphael Metivier, Frank Gannon, Esben Lorentzen, Ehmke Pohl, Julius Brennecke, and Steve Cohen for interesting and fruitful collaborations.

Alexander Shkumatov and Zeynep Arziman for spending their time working with me.

The Russell Group, Patrick Aloy, Caroline Hadley, Victor Neduva, Matthieu Pichaud and all the visitors for a nice lab atmosphere, helpful discussions and collaborations, and the lab excursions.

Marc Hemberger for always helping me out with computer problems.

Caroline, Lena, Peter, and Rob for critically reading the thesis.

All my friends and family – especially Lena – for making the last years so enjoyable.

# 7 Erklärung

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Professor Dr. Dietmar Schomburg betreut worden.

Heidelberg, den 16. März 2004 Alexander Stark

**Teilpublikationen:**

- R. Metivier, **A. Stark**, G. Flouriot, M.R. Huebner, H. Brand, G. Penot, D. Manu, S. Denger, G. Reid, M. Kos, R.B. Russell, O. Kah, F. Pakdel, F. Gannon; A Dynamic Structural Model for Estrogen Receptor-alpha Activation by Ligands, Emphasizing the Role of Interactions between Distant A and E Domains. *Molecular Cell*, 10, 1019-1032, 2002.
- **A. Stark**, S. Sunyaev, R.B. Russell; A Model for Statistical Significance of Local Similarities in Structure. *J. Mol. Biol*., 326, 1307-1316, 2003.
- **A. Stark**, R.B. Russell; Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.*, 31(13), 3341-3344, 2003.
- J. Brennecke, D.R. Hipfner, **A. Stark**, R.B. Russell, S.M. Cohen; Bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the pro-apoptotic gene hid in Drosophila. *Cell*, 113, 25-36 , 2003.
- E. Lorentzen, E. Pohl, P. Zwart, **A. Stark**, R.B. Russell, T. Knura, R. Hensel, B. Sievers; Crystal structure of an Archaeal Class I Aldolase and the Evolution of $(\beta\alpha)_8$ Barrel Proteins. *J. Biol. Chem.*, 278(47), 47253-47260, 2003.
- **A. Stark**, J. Brennecke, R.B. Russell, S. M Cohen; Identification of Drosophila miRNA Targets. PLoS Biology, 1(3), E60, 2003.
- P. Aloy, H. Ceulemans, **A. Stark**, R.B. Russell; The relationship between sequence and inteaction divergence in proteins. *J. Mol. Biol*., 332, 989-998, 2003.
- P. Aloy, **A. Stark**, C. Hadley, R.B. Russell; Predictions without templates: New folds, secondary structure and contacts in CASP5. *PROTEINS: Struct. Funct. Genet*., 53, 436-456, 2003 (CASP5 special issue).
- **A. Stark**, A. Shkumatov, R.B. Russell; Finding Functional Sites in Structural Genomics Proteins. *Structure*, submitted, 2004.

# 8 Lebenslauf

Alexander Stark
Klingenteichstrasse 6c
69117 Heidelberg

| | |
|---|---|
| Geburtsdatum | 19. September 1974 |
| Geburtsort | Bietigheim (jetzt: Bietigheim-Bissingen) |
| Familienstand | ledig |
| Schulausbildung | 1981 – 1985 Grundschule Besigheim<br>1985 – 1994 Gymnasium Besigheim |
| Schulabschluß | Allgemeine Hochschulreife (16. Juni 1994) |
| Zivildienst | August 1994 – November 1995<br>Schule am Favoritepark, Ludwigsburg |
| Studium | Wintersemester 1995 – Sommersemester 1998<br>Universität Tübingen<br>Diplomstudiengang Biochemie<br><br>Wintersemester 1998 – Sommersemester 1999<br>University of North Carolina, Chapel Hill, USA<br>Biochemistry and Biophysics<br><br>Wintersemester 1999 – Wintersemester 2000<br>Universität Tübingen, Biochemie |
| Diplom | Diplom-Biochemiker (Universität Tübingen, 29. November 2000) |
| Praktikum | Dezember 2000 – Mai 2001<br>Friedrich-Miescher Labor der Max-Planck Gesellschaft, Tübingen<br>Betreuer: Dr. Christoph Schuster |
| Doktorarbeit | Juni 2001 – Heute<br>European Molecular Biology Laboratory (EMBL) Heidelberg<br>Betreuer: Prof. Dr. Schomburg (Universität zu Köln)<br>Dr. Robert B. Russell |