1 Overview

1.1 Aim

The focus of the research in this Thesis is to generate novel biological knowledge through the transfer of information between related protein sequences. Currently there are around 1 million unique protein sequences available in public databases. Around a third of these proteins do not belong to any recognised and characterised family, and the majority contain regions that have not been described. Within these regions remains a huge amount of important biological information – and clustering them into sequence families allows both the synthesis of information from each family member and global analyses of family characteristics. The work carried out in this thesis aims to identify novel families of high interest, to refine known families and to correctly establish the homology borders within the member proteins. Statistical methods are used to identify potential new families in a high throughput manner, which are then manually investigated. Functional predictions are provided through the use of sequence analysis software and through the analysis of associated literature.

1.2 Background

As more protein structures have been solved, using X-ray crystallography and NMR, several trends and constraints of protein structure have become apparent. Of these, the most striking observation was that proteins are usually made up from several independently folding units, with the overall function of the protein being a composite of these substructures' functions. Furthermore, these substructures have been found to

have shuffled during evolution to create novel proteins with new emergent functions. The discrete and modular nature of these elements has led to them being termed domains; this also makes understanding protein domains a powerful way of understanding proteins.

It is also of note that there are already over 1 million proteins in public sequence databases, whereas it is estimated that there are between 1000 and 5,000 folds – a fold being the three dimensional structure a protein assumes in its native state – that exist in nature, with about 50% of proteins belonging to one of 800 folds (reviewed in Grant, Lee *et al*., 2004; first estimated by Chothia, 1992). Therefore grouping these sequences into fold families and subfamilies makes the data much more manageable. Solving structures is expensive, time-consuming and labour intensive at best, and at worst is currently impossible – particularly with the extremely biologically interesting cell membrane-associated proteins. So while three dimensional structural analyses are highly informative, comparative methods of protein sequence and structure analysis are essential.

Certain observations from sequence analyses have led to the development of powerful tools for protein comparison and structure determination. First and foremost is that protein amino acid sequences divide up into discrete units, which can be found in differing contexts. Mapping these to the corresponding structures has shown that a "sequence domain" almost certainly maps directly to a "structural domain". There are of course exceptions and qualifiers – for instance β-propellers are typically made up from between 6-8 sequence repeats and form a fold made up from 6-8 "blades"

(Murzin, 1992) - though of note haemopexin forms a four-bladed propeller (Gomis-Ruth, Gohlke *et al.*, 1996). All the blades are required to form the propeller, and hence all should be included as part of a single fold, but at a sequence level it would be seen as a series of homologous, and possibly gapped, repeats.

The second key observation is that if two protein sequences are shown to be evolutionarily related (homologous) then they will have the same tertiary structure – though again there can be exceptions (Grishin, 2001). There are now several powerful statistical tools for determining the likelihood that two sequences are related, some of which are described in chapter 1.6.

It is a common maxim in structural biology that function is encapsulated within the structure. If, through sequence analysis, we are able to demonstrate that set of sequences or sub-sequences are homologous, then we can transfer functional information associated with these regions. These two observations imply that if we can describe a family of related protein sequences and we know the physical structure of one of the proteins, then we can describe the function of all of them. This is because we should be able to construct comparative models based on the known structure and identify changes to the biochemistry of the protein. Developing comparative analysis technologies is currently the main approach in protein analysis as the cost of sequencing the gene is several orders of magnitude less than solving the structure of the protein, and *ab initio* structural prediction methods are still prone to significant inaccuracies (Aloy, Stark *et al.*, 2003).

In general the volume of publicly available protein sequence data has been expanding rapidly, driven by the current wave of genome sequencing projects, as indicated by the growth of the sequence repository, UniProt (see Figure 1.1). In turn sequence analysis has become part of the standard repertoire of biological research methods, and is now carried out on desktop computers by lab bench researchers and *en masse* on supercomputers by trained informaticians. A subfield of protein sequence analysis is domain hunting – the identification of novel protein domains from sequence data.

The concept of protein domains became apparent soon after the first structures were solved, and by the mid-1970s they were being considered in both sequence and structural terms (e.g. Wetlaufer, 1973; Edelman and Gall, 1969; Rossman and Liljas, 1974), with the first defined domain being the Ig domain (Edelman and Gall, 1969). The principle that they could be considered as mobile genetic units was put forward by Rossman and Liljas (1974) after analysing the similarity of nucleotide-binding domains in different structures.

Led by researchers like Eugene Koonin, Peer Bork, Chris Ponting & Kay Hoffman (to name a few) the *de novo* identification of domains has become a field in its own right. Approaches range from the purely automated (e.g. ProDom, see chapter 1.6.3; Servant, Bru *et al*., 2002) to manually intensive (e.g. the BRCT domain; Bork, Hofmann *et al*., 1997), and encompass combinatorial approaches (e.g. Ponting, Mott *et al*., 2001). Several databases now collect and curate descriptions of these domains (see chapter 1.6.3), and provide tools for identification of known domains in new sequences. These use a variety of statistical methods and design philosophies. Others

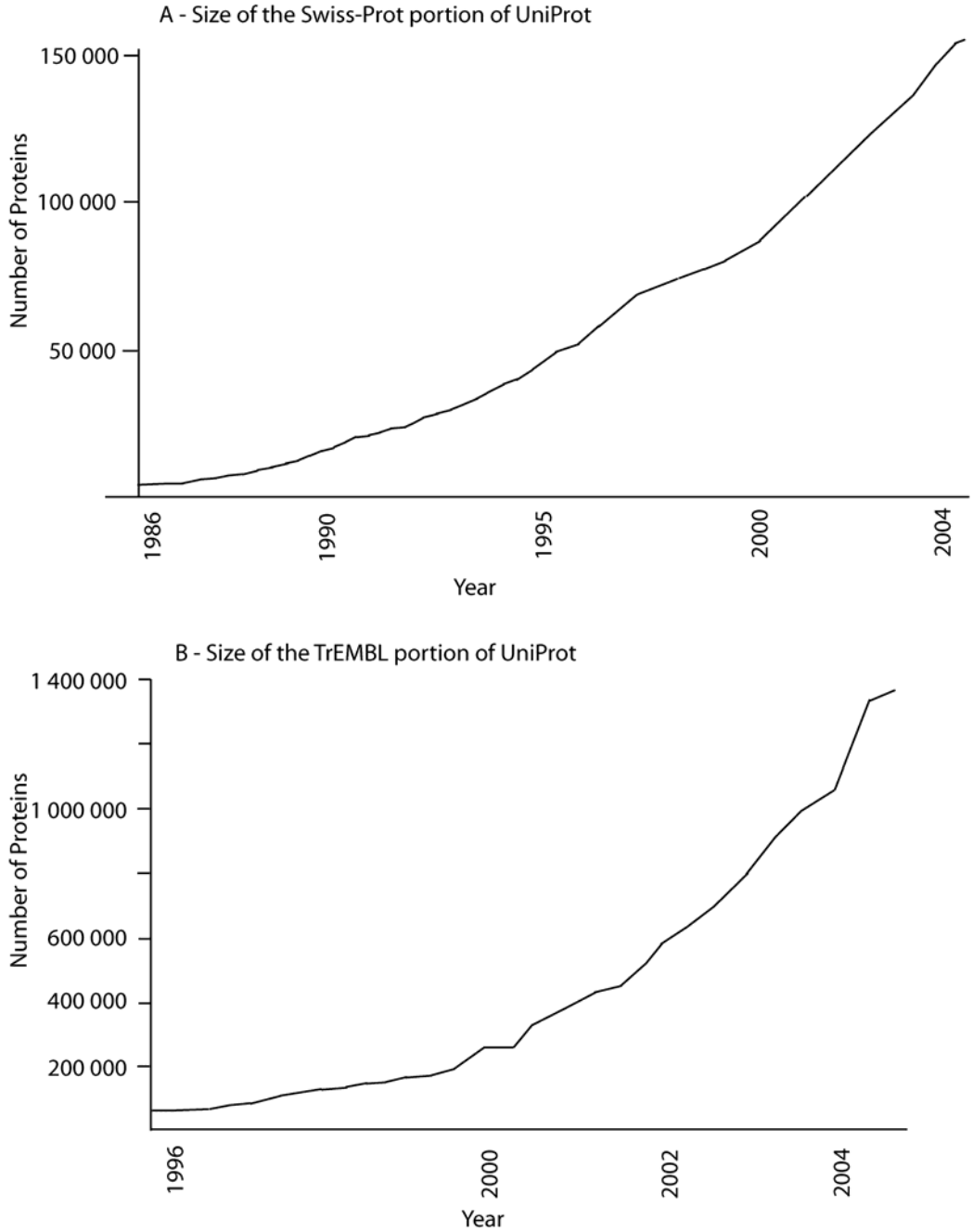

Figure 1.1: Growth of the UniProt database
(A) Swiss-Prot began in 1986 and has grown in a fairly steady fashion to the 150,000 proteins it now contains.

(B) The automatically generated supplement was begun in 1996 to keep up with sequence being generated by the large-scale sequencing projects begun in the 1990s. TrEMBL now contains more than 1,400,000 protein sequences - though there may be a high degree of redundancy relative to Swiss-Prot. Its growth appears to be exponential.

Both of these graphs are based on those presented at http://us.expasy.org/sprot/.

present the results for automatic domain detection, and update them when new sequences become available – for instance ProDom and the derivative Pfam-B (Bateman, Coin *et al*., 2004), or ProtoMap (Yona, Linial *et al*., 1999).

The growth in the power of domain identification from sequence has been driven by the large-scale sequencing projects of the last ten years. Previously the protein sequence databases were small and highly biased towards specific proteins or families of interest. Genome sequencing has led to a much wider range of proteins being sequenced, hence increasing the diversity of domains contained within the sequence database and the diversity of contexts these domains are found in. The increased diversity of sequences found in the protein databases can also allow subtler relationships to be derived, by the introduction of "Stepping Stone Sequences" - see chapter 1.5.2 for an explanation. As a result, not only is it possible to detect recently deposited novel domains, but also it is becoming easier to detect domains that were already present.

1.3 Protein Domains, Repeats, Motifs and Families

Proteins exhibit modular structures, with their overall function or fold being emergent from the modular components they are constructed from. The specific arrangement of modules is called the "domain architecture". All these components can be grouped into three classes - domain, structural repeat, and motif. When it is not possible to assign a component to a particular category, it can be classified as a family. These four types are the same as used by the Pfam (see chapter 1.6.2) database, around which the work in this thesis is based. While there is much discussion on what

constitutes a protein domain, the definition mostly depends on perspective; for a more detailed discussion of the precise differences see the review by Kong and Ranganathan (2004). Since much of the work presented has been done purely on protein sequence, without any available structure models, the most used definition of domain is the second one given below, but it should be noted that all three definitions largely overlap. The effective difference between them is on deciding where to position the edges of the domain within a protein. For instance a functional domain may be equivalent to an evolutionary domain and lie within a structural domain. Figure 1.2 shows examples of domains, motifs and repeats.

Three Common Definitions of a Protein Domain

- Structural: An independently folding unit in a polypeptide chain, which forms its own hydrophobic core.
- Evolutionary: A segment of amino acid sequence that is conserved in differing surrounding sequence contexts.
- Functional: The minimum sequence required to encode a function in a protein, as determined by experimentation.

Definition of a Structural Repeat

A repeat is a conserved sequence that only forms a stable structure when present in more than one copy. Each repeat is not independently stable but all contribute to a final stable structure. Examples are the WD40 repeats (Neer, Schmidt *et al*., 1994) and TPR repeats (Goebl and Yanagida, 1991). The number of repeats that make up the final structure may or may not be restricted to a range: WD40 repeats occur in sets of 6-8 and form a single propeller-like structure; the approximately 35 residue TPR repeats can occur anywhere between 2 and 50 times and form a solenoid structure.

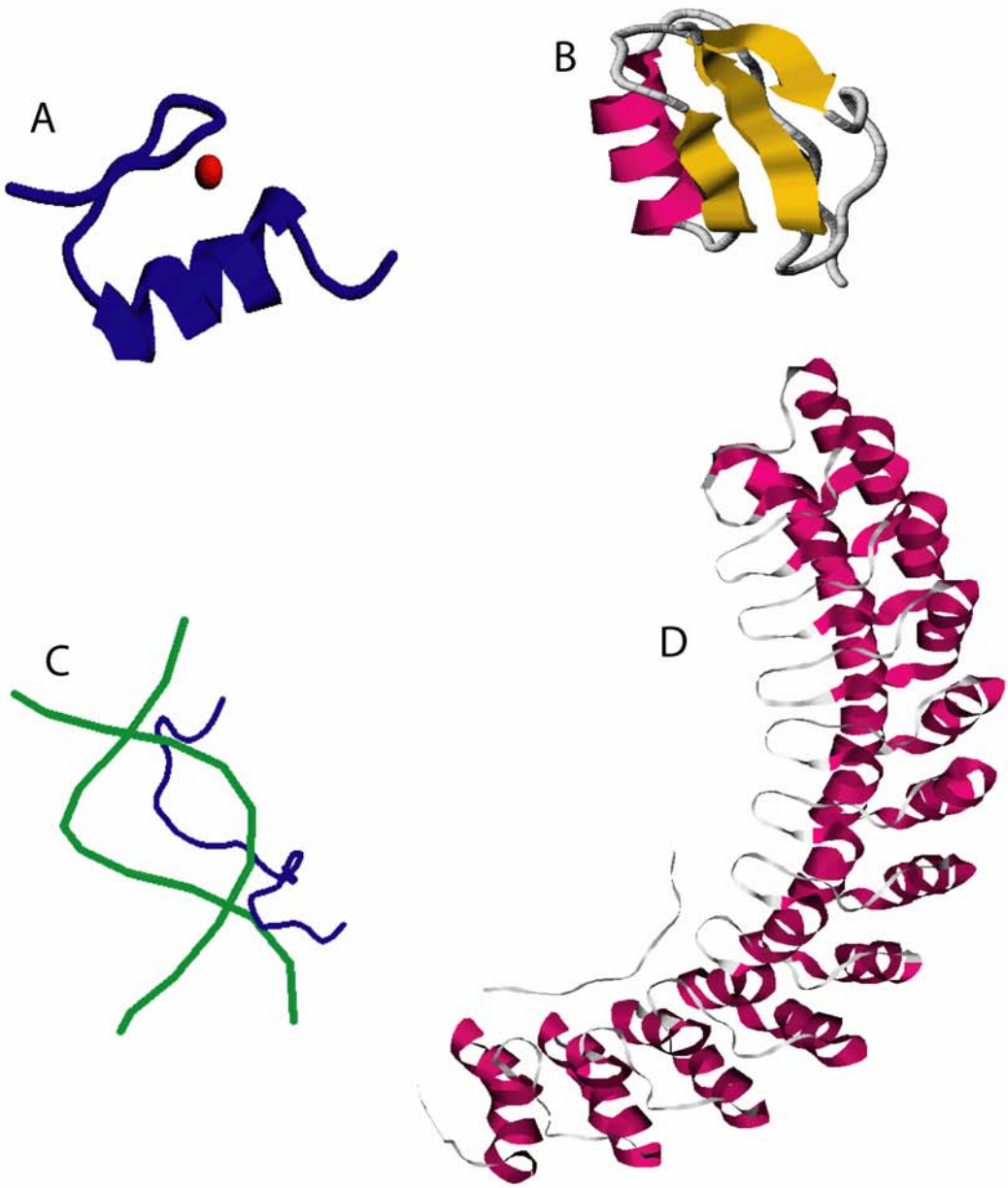

Figure 1.2: Examples of different protein structure types.
(A) **Zinc Finger domain:** Zinc fingers are able to form a stable tertiary structure in the presence of a stabilising zinc cation (red sphere). Image derived from PDB:1KLR.

(B) **PASTA domain:** PASTA domains form independently stable tertiary structures and can be considered as a classic structural domain. Image derived from PDB:1QMF.

(C) **AT-hook motif:** The AT-hook motif (blue) is a protein sequence that preferentially binds AT rich DNA (green) but has no stable structure. Image derived from PDB:2EZE.

(D) **Ankaryin repeats:** Several ankyrin repeats group together to form a higher order structure. Image derived from PDB:1N11

Definition of a Sequence Motif

A motif is an amino acid sequence that does not form an independently stable globular structure but has a specific function that is conserved between related sequences. For example the AT-hook motif is a short motif of around 13 residues that binds AT rich DNA (Nissen, Langan *et al*., 1991). Although it is believed to form a particular secondary structure (Huth, Bewley *et al*., 1997) its short size and lack of stabilising ligands means that it can not form a stable tertiary structure itself.

Definition of a Sequence Family

A family is a group of sequences that have been shown to be related using sequence comparison, but may consist of more than one domain, motif, repeat or combination thereof.

1.4 Characteristic Properties of a Protein Domain

As described above there are several ways of defining a protein domain, with the definition used being the one appropriate to the type of investigation. However, no matter the definition there are several common characteristics that typify what would be considered a domain. As discussed above, domains are the modular units of proteins, and so modularity would be expected. This can be expressed in several ways: Ideally the domain will be found in multiple architectures, as this demonstrates that it is independent of the surrounding sequence. Experimental evidence can also indicate modularity. For instance proteolytic degradation of the PulD protein (Nouwen, Stahlberg *et al*., 2000) revealed the same N-terminus for the Secretin domain as the sequence based prediction made in chapter 3.3 (Secretin_N domain). Of course, there are exceptions to this apparently straight forward rule. In some proteins a domain may be dependant on another for correct folding. An example is the

strand swapping between homologous TOBE domains from different *Escherichia coli* ModE proteins (Hall, Gourley *et al*., 1999; Koonin, Wolf *et al*., 2000).

The second and possibly simplest property is that domains almost always measure between 50 and 400 residues in length (see Figure 1.3). The lower limit probably reflects the minimum number of residues required to form a stable structure. Stable structures usually are generally globular with a hydrophobic core. There are some exceptions in which strong stabilising interactions have allowed the formation of smaller stable structures. An example is the Zinc finger family, in which a Zn^{2+} ion stabilises a 22 amino acid structure (Miller, McLachlan *et al*., 1985; depicted in Figure 1.2). Other interactions may include disulphide bridges and hydrogen bonding. If a region has been experimentally determined to be a functional domain, then it may be disordered – it has no stable tertiary structure – and maybe provides an electrostatic charge or some flexibility to the structure (i.e. SMC_hinge). Also transmembrane domains may not fold correctly until inserted into the membrane (i.e. Voltage-dependent K^+ channels; Jiang, Lee *et al*., 2003). At the other end of the spectrum there are some giant domains - for instance the lipoxygenase domain is apparently a non-dividable structure of over 500 residues (Boyington, Gaffney *et al*., 1993).

The reason for the lack of folds found that are larger than a few hundred residues in length is not clear. It is possibly due to several reasons rather than any particular one. For a start there may be a lack of unique structures beyond this threshold, with most possible stable forms being a composite of several smaller domains. Also larger

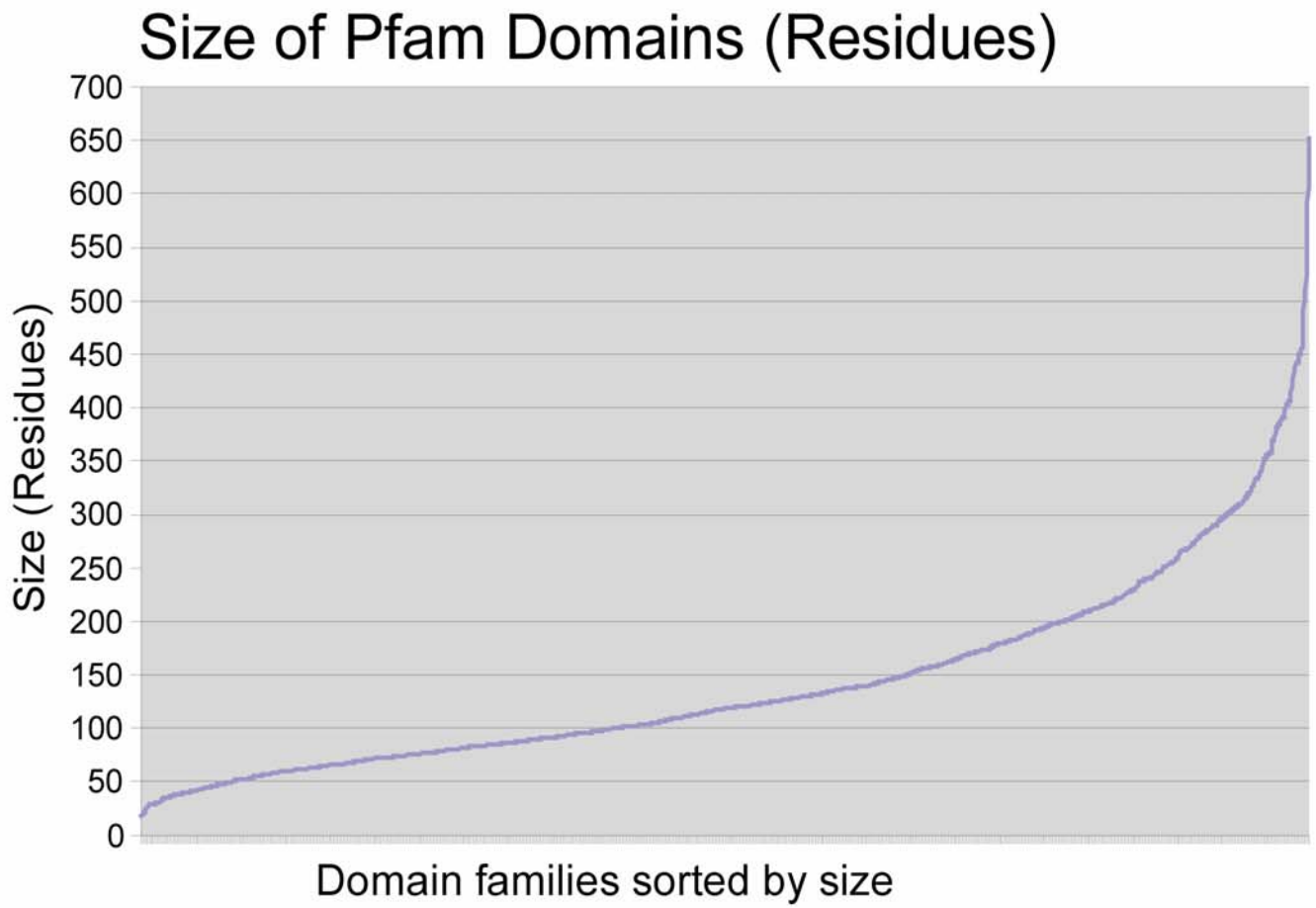

Figure 1.3: Graph displaying the average size of domains recognised by Pfam.
For each domain family in Pfam 14 the average length was calculated, after discarding all fragment matches. Displayed are the results sorted by size along the X-axis. As can be seen the large majority of sequence domains are between 50 and 400 residues in length.

domains require more sequence to encode; to a certain extent natural selection minimises genome size, as evidenced by the general lack of intergenic space in bacterial genomes (average gene density = 86%, data from Genome Atlas; Pedersen, Jensen *et al*., 2000). Furthermore an analysis by Lipman and co-workers (2002) found evidence for significant selection of shorter proteins. It is possible that longer domains would be selected against if there is a smaller domain that can carry out the same function, though this has not been observed.

The third property is that two related domains will also share function. So any member of, for instance, of the Transpeptidase domain family can be predicted to be a transpeptidase provided the catalytic residues are present. However, the extent to which this information transfer can take place varies for different domains and the form it will take can be subtle. The different transpeptidases may have slightly variant substrate specificity, but the basic reaction can be easily described for any. In contrast, the Ig domain shows a huge range of functions, and variants of the domain are able to bind nearly any chemical – one of their biological roles is forming the recognition sites in the immune system immunoglobulins. In this case the domain acts as a scaffold upon which functional motifs, which determine the specific function, can be hung. This mechanism of creating functional diversity is also commonly seen with structural repeats – for instance β-propellers (Murzin, 1992) and CASH repeat proteins (Ciccarelli, Copley *et al*., 2002) show a similar range of functional diversity as the Ig domain.

So, although the basic concept of a domain is clear and straightforward, as always with biological systems, there are caveats that must always be borne in mind.

1.5 The Limitations and Difficulties of Domain Hunting

1.5.1 Domain Boundary Identification

Between 60 and 80% of proteins in a genome can be expected to consist of more than one domain (Teichmann, Park *et al*., 1998; Gerstein, 1998). Hence when presented with a single amino acid sequence, the first problem in identifying novel domains is identifying the edges. Correctly identifying the edges of a domain can significantly alter the power of a predictive domain model (e.g. a profile HMM, see chapter 1.6), and lead to large expansions in the number of identified family members. An example from within this thesis is the PASTA domain. The PASTA model is similar to a previous model called PBP_C built by R. Finn, which correctly identified homologous penicillin-binding protein (PBP) regions, but failed to detect significant similarity to the PknB-like serine/threonine kinases (PSTKs). Subsequent to the creation of the PBP_C model, the crystal structure of PBP2X from *Streptococcus pneumoniae* was determined (Gordon, Mouz *et al*., 2000). From this it was clear that the carboxyl-terminus (C-terminus) consisted of two identical domains and that the model covered the first domain and extended ten residues to the amino-terminus (N-terminus). The boundaries of the PASTA model were found in the sequence using the 'Repeat Hunt Method' described in chapter 2.1.2; it exactly covers one domain and is able to identify many novel homologies. Only a small correction to the model had a dramatic effect on its sensitivity.

Beyond having effects on the sensitivity of the model, correct boundary determination can also affect the quality of information transfer, the effectiveness of structure prediction software, and making crystals for structural analysis (Kong and Ranganathan, 2004). It also can be informative in the evaluation of automated clustering algorithms. A common flaw in many approaches for the automated clustering of protein families is that proteins that are only related by a single domain can be clustered, even though there is no overall functional link and the two proteins are not evolved from a single common ancestor. This type of error can be seen in the genome paper of *Streptomyces coelicolor* (Bentley, Chater *et al.*, 2002), in which the prediction of 44 PSTKs was reported on the basis of single linkage clustering. Using HMMs to predict the domain content shows that there are in fact 34 PSTKs. The discrepancy is caused by single-linkage clustering linking unrelated proteins through domains that they share. This is explained graphically in Figure 1.4.

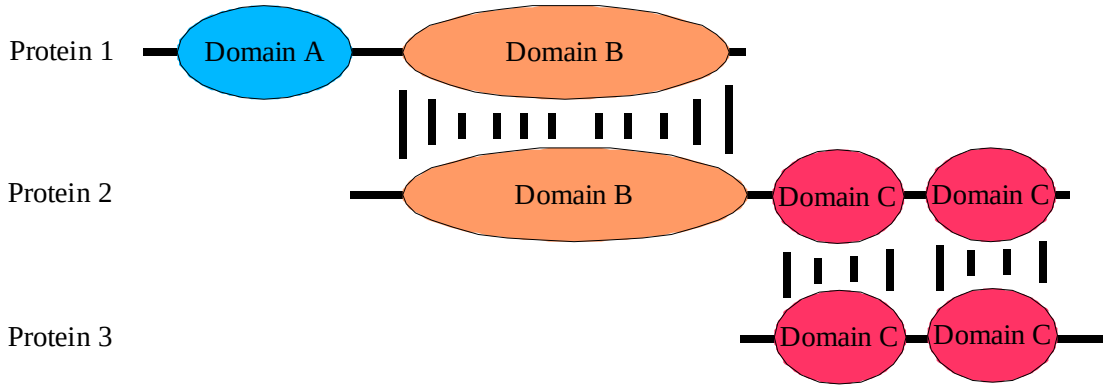

Figure 1.4: A common protein clustering error caused by multidomain proteins. Not knowing the domain structure of a protein under investigation can lead to missannotation. In this case a sequence comparison programme (i.e. BLAST) has identified significant sequence similarity between Protein 1 and Protein 2, as well as between Protein 2 and Protein 3. Naïve interpretation of this result would allow the transfer of information between Protein 1 and Protein 3; However, aided by the knowledge of the domain architectures we can see that there is not likely to be any functional similarity between Protein 1 and Protein 3.

Various different methods have been employed for the recognition of domain boundaries from sequence; I have mostly used manual or semi-automated approaches. These are described in more detail in chapter 2.1. Also there are many researchers developing automated approaches, with two main aims. One is for use over large data sets; the other is for predicting domains from sequence that have no obvious homologues in other proteins. Comparative approaches include mkdom2 (Gouzy, Corpet *et al.*, 1999) - the basis of the ProDom database and an evolution of the original Domainer script (Sonnhammer and Kahn, 1994) - and Gracy and Argos's (1998) pairwise comparison method that underlies DOMO.

Although these methods can be useful for large sets, they have yet to produce the accuracy of results that can be achieved through manual boundary determination – as discussed by Kong and Ranganathan (2004). Recent approaches, such as the combinatorial method developed by Nagarajan and Yona (2004) and the neural network-based method by Liu and Rost (2004) show some promise, and are starting to approach the accuracy of manual detection. The second method also has the advantage that it can take a single sequence and rapidly make a prediction, which can then be refined manually. The current state-of-the-art is reflected in Pfam-A's much higher coverage than Pfam-B despite only consisting of approximately 7,500 families compared to around 100,000 for Pfam-B.

1.5.2 The Stepping Stone Phenomenon

A general rule of thumb in pairwise biological protein sequence comparison is that if two homologous sequences show less than 30% identity (using any measure; May

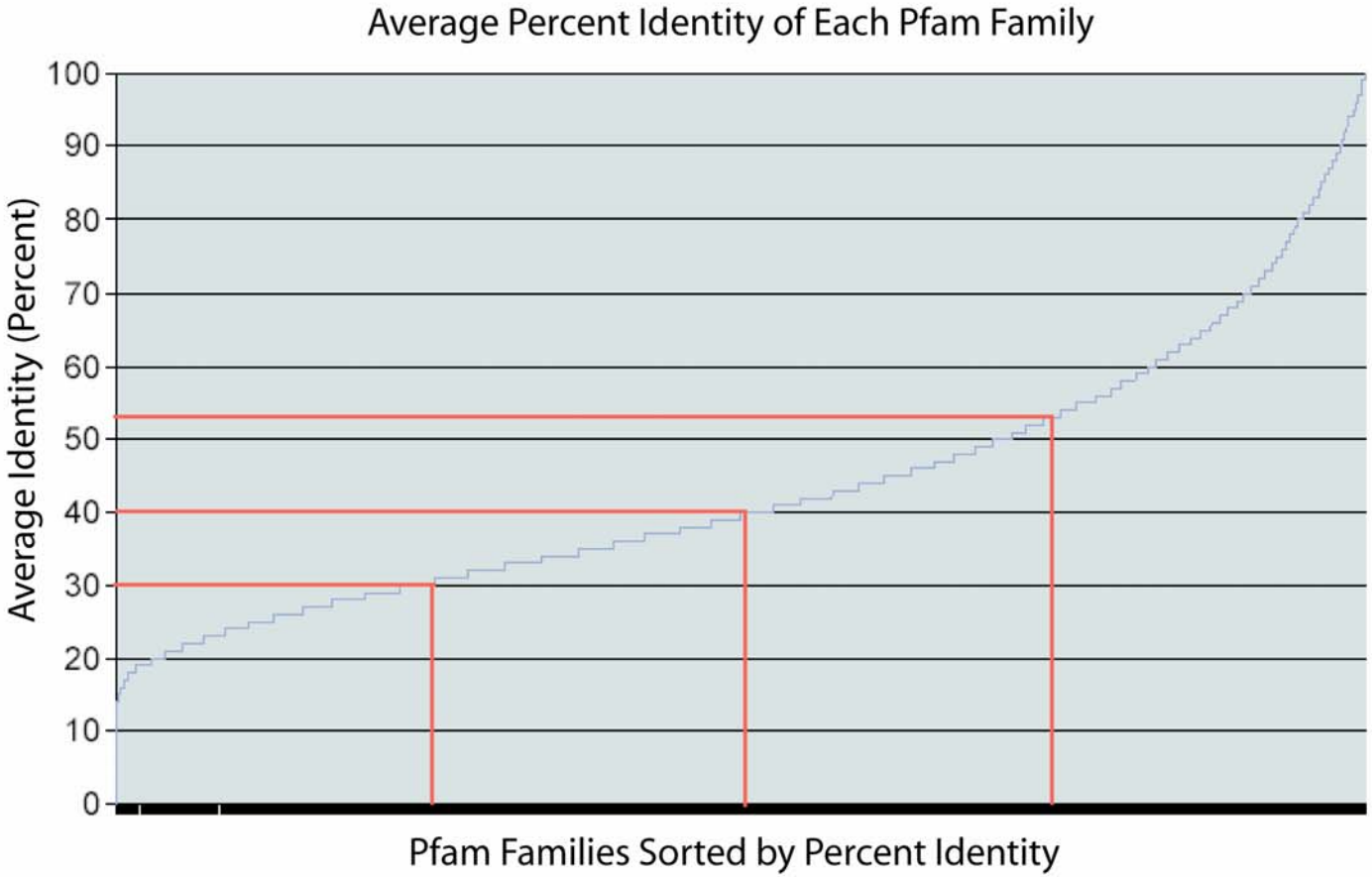

Figure 1.5: Graph of the average percent identity for each Pfam family
The percent identity of each Pfam family was calculated using Sean Eddy's alistat software. The families are sorted on the X-axis according their average percent identity. The red lines mark the quartile divisions. As can be seen half of Pfam families have an average sequence identity below 40%. This illustrates the power of iterative searching.

2004) then identifying the relationship is unlikely. However, 50% of Pfam families exhibit less than 40% average identity and 25% have less than 30% average identity (see Figure 1.5, statistics calculated using S. Eddy's "alistat" software). These distant relationships can be most easily detected by identifying an intermediate or, as they are also known, stepping stone sequence. This is a sequence that shows significant similarity to both distantly related sequences, and so can be used to infer a relationship.

This principal essentially underlies iterative searching: Newly identified homologues are included into the model and hence even more divergent homologues are detected (see Figure 1.6 for a graphical explanation). Prior to the genome sequencing projects, protein sequence databases were often biased towards specific proteins, species or sequence families of interest and so the necessary stepping stones were not present. As this is corrected subtle relationships are becoming apparent, but it also means that searches need regular repetition. As an example the HHE domain was identified in early 2002, and formed a cohesive and internally consistent family (Yeats, Bentley *et al*., 2003). Repeating the searches in 2004 led to the merging of this family with the Hemerythrin family, which had been deposited in Pfam in late 1999. Until recently there was no obvious link between the two because the necessary sequences were not there - such as *Methanosarcina mazei* MM1985 (UniProt:Q8PVH8), a *Streptomyces parvulus* hypothetical protein (UniProt:Q70HY1) and *Shewanella oneidensis* SO3549 (UniProt:Q8EBG9).

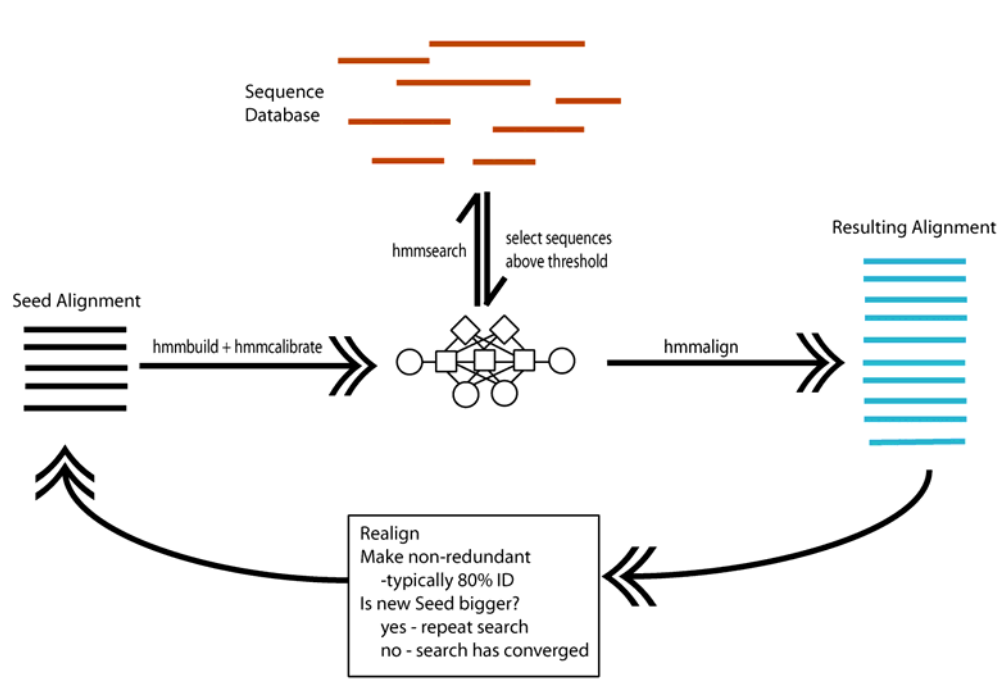

Figure 1.6: The iterative search methodology
This process allows the researcher to start with a small number of sequences that are known to be related and detect homologues with a high degree of sensitivity and selectivity. By the arrows are the components of Sean Eddy's HMMER software that are used in each step - hmmbuild, hmmcalibrate, hmmsearch, hmmalign. At the centre of the diagram is a simplified version of HMMER's Plan 7 HMM architecture (discussed further in Chapter 1.6.1).

1.5.3 Replication of Experiments

Despite the statistical basis and computational nature of domain identification from sequence, assigning a confidence score – a level of certainty that there are no false positives included – to a family is not simple. After each round of searching there are two questions: Are there any false positives included? Are there any members missing? Whilst stepping stone sequences allow iterative searching, as has been mentioned, sometimes the required sequences are not present in a database or may not even exist in nature. In this case it may be necessary to relax the inclusion threshold and incorporate sequences with a low similarity in order to identify distant homologues; concomitantly this increases the risk of including false positives. In this

case it is necessary to use reciprocal searches and other evidence to ensure that the relaxation of the threshold is valid.

Another technique for finding distant homologues is to create a smaller sequence database that is believed to be particularly enriched in the target domain (as discussed in chapter 2.3.4). Commonly used estimates of significance (E-values), including those used by BLAST and Prospero, for evaluating the significance of protein similarity scores are functions of database size: The larger a sequence database is, the more chance you would see an apparent match by chance. So by applying a knowledge-based filter, it is possible to reduce the database size while retaining all the copies of a domain, and hence increase the significance of any potential matches.

Using these different techniques, so as to build up a diverse domain family with a low level of conservation, makes statistical validation difficult. A solution that would have provided internal consistency to this thesis would have been to use a fixed release of UniProt (or Swiss-Prot/TrEMBL, see chapter 1.6.4) so that all searches were equivalent. However, the major protein sequence databases have new releases every couple of weeks, with a size doubling period of around 18 months; by not regularly updating, a vast amount of available information is being ignored and valuable stepping-stones may be missing - as was the case with the HHE/Hemerythrin domain.

There is also heterogeneity in the search tools, with some tools able to find more distant homologues in some families than the other tools. So given a starting sequence or alignment, several different results may be arrived at depending on the search tool

and the sequence database. There are also, of course, many different parameters and weighting systems that can be varied for each search tool, adding an extra layer of complexity. Various search tools are explained and discussed in chapter 1.6.1.

Rather than use a strict system of family building in which specific E-values are rigidly adhered to, the approach I have used in this thesis is to carry out controls that vary databases, search tools, starting points and also depositing the results in a public repository (the Pfam database) for further review.

Sequence Search Controls

(a) Reciprocal searches - varying the search start point.
(b) Vary the N and C-termini of the seed subsequence.
(c) Take sequences falling just below the inclusion thresholds as seeds.
(d) Use a different search tool – PSI-BLAST/BLAST/HMMER.
(e) Vary the sequence database – UniProt/GenBank/Selected sequences.
(f) Publish the family, either in the literature or in a public database, for peer review.
(g) Use different inclusion thresholds.
(h) Careful visual examination of the final alignment to identify inconsistent sequences.
-can be aided by building a Neighbour-Joining Tree to group potential false positives.

The final decision as to whether the identified domain family was genuine, and that as many true members had been identified as possible with few (preferably none) false members included, is subjective but achieved through the consensus of several experiments. It is also important to be conservative in decision making until further

tests support the inclusion of more divergent sequences. Although these tests are not described in detail and are not, in some considerations, complete it is my belief that the families presented are correct. More importantly they are all readily available to the general public via the Pfam database for review and correction. This form of open peer review is probably the best way to ensure that models are as accurate as possible; indeed this open review allowed the realisation that two predicted PPC domains (see chapter 2.2.5) were false positives and they were removed from the alignment.

1.6 Tools

1.6.1 Search Software

HMMER (S. Eddy) and SAM (Hughey and Krogh, 1996)

Over the last decade the applications of Hidden Markov Models (HMMs) have proliferated in biological research. Uses include protein sequence comparison, splice-site prediction (i.e. Henderson, Salzberg *et al*., 1997), transmembrane helix prediction (i.e. Krogh, Larsson *et al*., 2001), signal peptide prediction (i.e. Nielsen and Krogh, 1998), and gene finding (i.e. Burge and Karlin, 1997; Meyer and Durbin, 2002). Their primary relevance to my work is that they underlie the search software I have mostly used - HMMER. HMMER also underpins the Pfam database (see 1.6.3) – around which much the work undertaken is based. In essence HMMER reads in a seed alignment and constructs a profile HMM. The architecture of the HMMER HMM, called 'Plan 7', has a core that consists of a node for each column of the alignment, each node consisting of three states - M, D, I (match, deletion, insert). The core is flanked by a B and an E (begin, end) state. The remaining five states control

algorithm-dependent features of the model, and can be varied to alter the type of model produced (see below).

The emission probabilities for the M state and the transition probabilities of the D state are generated from the multiple sequence alignment. In each column of the multiple sequence alignment the frequency of each amino acid is counted, and hence the emission probability of a particular amino acid appearing at each position can be derived. The transition probabilities of the insert states (I) are based on an internal evolutionary model. Since each node is considered separately, the probabilities assigned at node are independent of the other nodes, and hence higher order information can be lost. However, this seems to be not much of a problem in protein sequences as this type of approach has been successful.

By controlling the algorithm states HMMER can be used to construct two types of HMM – one is known global or 'ls' and the other is the local or 'fs' model'; both are local with respect to the protein sequence. The ls model will only find significant matches that extend over the whole model and will allow multiple non-overlapping hits per sequence. The fs model will report significant alignments that may not extend along the whole HMM, and also will allow multiple hits per sequence. This has an advantage over other methods in that the model itself encodes the fragment or global nature rather than using a different algorithm for searching the same model. One use is that specialised models can be built that capture detailed aspects of specific domains – e.g. a highly variable N-terminus but an absolute requirement for the C-

terminal 10 residues – and then searched against a sequence database using the same algorithm.

Searching the HMM returns a list of bit-scores for each sequence. From the bit score an E-value is calculated. This estimates the number of sequences one would expect to achieve at least that score that would exist by chance in the database or, the number of false positives. This is achieved by best fitting a histogram of scores generated from searching 5,000 random amino acid sequences which approximately reflect the composition and length of UniProt fitted to an extreme value distribution (EVD).

The mathematics that underlie the use of HMMs for sequence searching are well established and are described in detail by Durbin, Eddy *et al*. (1998) and so I do not propose to describe them in detail here. It is enough to know that they work and that the software has been rigorously constructed; HMMER is simple enough to use as a 'black-box' process.

HMMER is just one example of an HMM-based search package. Also popular is the SAM package created by Richard Hughey, Kevin Karplus and Anders Krogh. SAM also includes methods for secondary structure prediction and built-in iterative searching. Comparisons between HMMER and SAM show that at the near zero or zero error rate required for this project there is little difference in the performance of either package - the sequence composition of the seed alignment has a far greater effect on the sensitivity and specificity of the model - and that HMMER is also marginally faster on large sequence databases (Madera, Vogel *et al.*, 2004). The main

reasons for using HMMER were to allow easy interaction with the Pfam database and because it is well understood and supported within the lab.

BLAST/PSI-BLAST (Altschul, Madden *et al*., 1997)

BLAST is a heuristic method for similarity searching that in essence simplifies the Smith-Waterman algorithm. It uses a significant amount of pre-processing and two key assumptions (listed below) so as to reduce the running time. The Smith-Waterman algorithm is derived from the Needleman-Wunsch algorithm for comparing two sequences. The key difference is Needleman-Wunsch compares the entire length of both strings - a global alignment – whereas Smith-Waterman can compare the sub-string of one sequence against any substring in another sequence – local alignment. The BLAST heuristic makes two assumptions:

(1) Most high-scoring local alignments contain one or more high scoring pairs of three letter substrings called 'words'. These locations can be quickly identified and used to grow a longer high-scoring alignment.

(2) Homologous proteins show extensive regions of similarity with no gaps in the sequence. This facilitates extending the words into local alignments.

BLAST is the most widely-used and possibly fundamentally important tool in bioinformatics. It has a very fast running time, which allows it to be used with genome sized datasets. For instance, searching a 65 letter query sequence against a protein database of 1,998,366 sequences (670,625,123 letters) using the NCBI default

gap penalties at the NCBI server, took less than 30 seconds. In contrast HMM-based methods are much slower and, without powerful compute farms, are impractical for large-scale analyses. It is also very adaptable and can be optimised for different types of search fairly easily. BLAST has been reviewed extensively and its uses well documented – for instance Korf, Yandell and co-authors' (2003) book "BLAST".

PSI-BLAST stands for Position Specific Iterated BLAST. It is a development of BLAST that has some similarities to HMMER and SAM in that it creates a profile of the family that it uses to search a sequence database. After starting with a standard BLAST search, the returned alignments are used to generate a Position Specific Score Matrix (PSSM) that is used to search again. This process can be repeated for a set number of rounds or until 'convergence' – when the searches identify the same number of sequences as in the previous round. It essentially uses the BLAST heuristic but is able to take a PSSM as input. It is not as sensitive as SAM or HMMER (Madera and Gough, 2002) and it deals with low complexity sequence less successfully – for a practical example see 2.3, the ALF repeat, and also noted by Chen (2003). On the upside it is much faster; this makes it an ideal tool for carrying out positive controls, or rapidly generating large numbers of seed alignments for refinement using HMMER.

1.6.2 Alignment Software

Dotter (Sonnhammer and Durbin, 1995)

Dotter is a tool for visualising protein to protein comparisons. It compares every amino acid in one sequence with every amino acid in a second. From this it produces

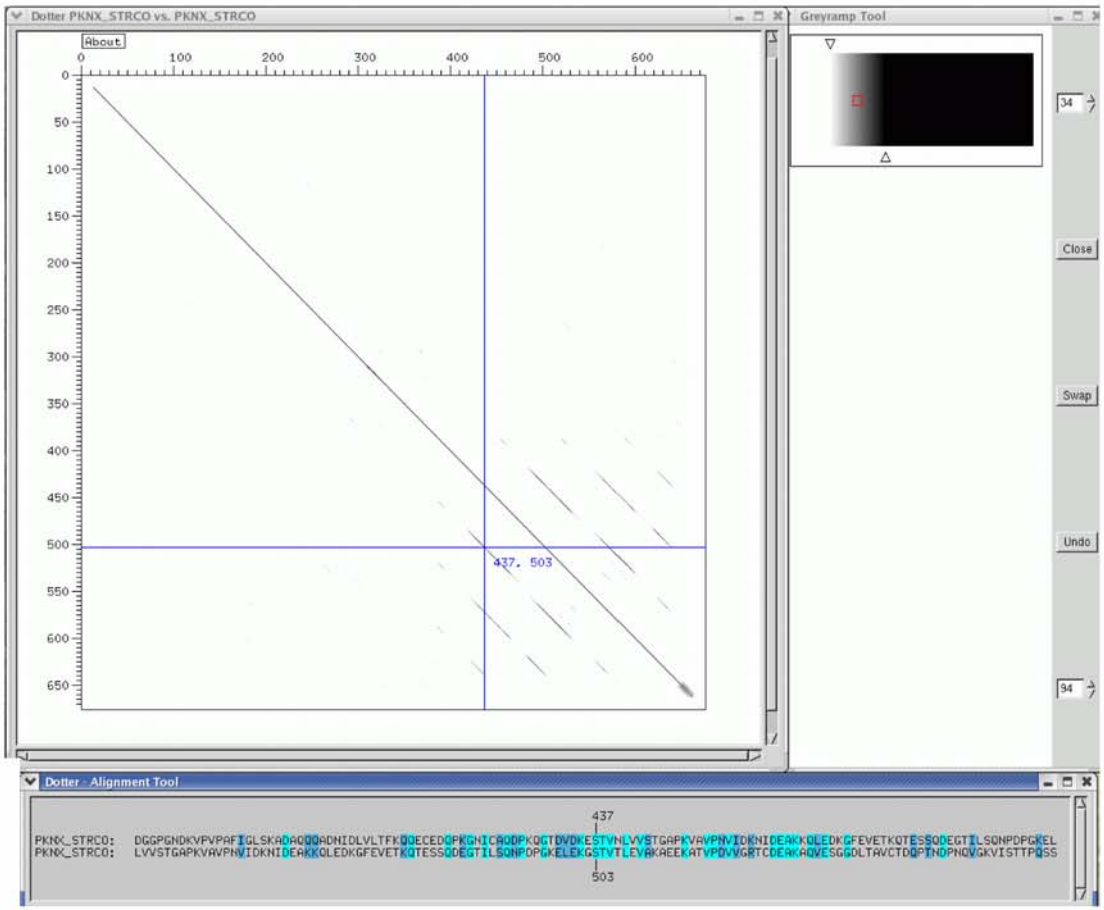

Figure 1.7: Example of the Dotter output
Dotter is a useful tool for identifying repetitive regions within proteins. In this case a self-self comparison of UniProt:Q9XA16 reveals four tandem repeats in the region 400-650; these repeats are four copies of the PASTA domain. The output consists of three parts - the dotplot, a grey ramp tool that alters the similarity threshold for displaying in the window, and an alignment viewer.

a dot plot with one sequence on the X-axis and the other on the Y-axis (see Figure 1.7 for an example). For easier visualisation the scores are averaged over a window that runs along the diagonal; work by E. Sonnhammer has found that 25 residues appears to be the most sensitive window size for identifying repeats and is used as default. This tool was used extensively during the work for this thesis, primarily for self-self comparisons, in order to identify novel repeated regions and to aid in interpretation of the results from Prospero (see below).

Prospero (http://www.well.ox.ac.uk/rmott/ARIADNE/prospero.shtml)

Prospero is part of the Ariadne software created by R. Mott. Prospero generates local alignments using the Smith-Waterman algorithm and then assigns accurate P-values (to within 5%, 95% of the time; Mott, 2000). The P-values are then multiplied by the database size, converting them into E-values. As discussed for HMMER, an E-value represents the expected number of false-positives occurring at that score in a database the size of the one searched. As implicated, the larger the database the greater the number of false-positives one would expect. Therefore, self-self comparison will return the lowest E-value for a particular score and will be more sensitive then searching against a sequence database. This principal underlies the approach used in many of the domain hunts undertaken. A second benefit of Prospero is that the output is easy to parse using computers compared to the graphical output of Dotter. This makes it very simple to carry out very large numbers of self-self comparisons and identify significant alignments, which can then be further processed and used to seed iterative profile-based searches (see chapter 2.1.3.2).

Multiple Sequence Alignment: - ClustalW, T-Coffee and MAFFT

Most of the sequence search software and processes described so far use or produce Multiple Sequence Alignments (MSAs). The sensitivity and specificity of HMMER can be significantly affected by the seed alignment from which it generates the HMM. Furthermore interpreting the patterns of similarity and identifying conserved residues is made much easier when the alignment is accurate. An accurate alignment has all structurally equivalent residues in the same column.

Given the size and number of alignments examined manual alignment is impractical, so three multiple sequence alignment programmes were used - ClustalW, T-Coffee and MAFFT. ClustalW is probably the oldest and most well known of the three (Thompson, Higgins *et al*., 1994). It has the advantages of being fast and reasonably accurate. It is based on the progressive approach proposed by Hogeweg and Hesper (1984) and Feng and Doolittle (1987). To describe the process simply, pair-wise scores are determined for all the sequences by means of a substitution matrix, and are used to grow a Neighbour-Joining (N-J) tree. A series of pairwise alignments are carried out, starting with the most related sequences, then progressing to more distant sequences, and then aligning each of the sub-alignments so as to progressively build up an MSA. ClustalW includes some refinements to this process, which primarily focus on reducing errors in the pair-wise alignments. This type of algorithm is described as a greedy algorithm, and if an error is introduced early in the process its effects will be amplified and may disrupt the overall alignment. Also the global nature of ClustalW means that if one tries to align multidomain proteins that contain unrelated domains there can be deceptive misalignments.

T-Coffee is more recent, and uses a more complex alignment algorithm (Notredame, Holm *et al*., 1998). Instead of using a substitution matrix, as used by ClustalW, it uses a PSSM, termed an "extended library", where the score for each pair of residues depends on their compatibility with the PSSM. The "primary library" is a collection of pairwise global alignments generated using ClustalW and local alignments generated by Lalign (Huang and Miller, 1991). The local alignments are used to create a consistency check, allowing the minimisation of potential errors during the build up of the progressive pairwise alignments. It is also possible to customise the extended library to improve its performance for specific families, or for ensuring that catalytic residues align. In comparison to ClustalW it performs better in general, though is much slower and impractical for alignments more than 200 sequences of length greater than 200 residues (personal observation).

MAFFT is the most recent of the three methods (Katoh, Misawa *et al*., 2002). Although the overall mechanism is similar to ClustalW it transforms the amino acid sequence into a sequence of polarity and volume values; these are aligned using a fast Fourier transformation and a novel scoring scheme. There are two implementations of MAFFT - a progressive method (FFT-NS-2) and an interactive refinement method (FFT-NS-i). I have exclusively used the FFT-NS-i implementation; it is much faster than the other tree programmes described, and also is as accurate.

Comparisons of the three methods have been carried out by various researchers. Presented below in Table 1.1 are the results of a recent test carried out by Edgar (2004), which was used for a comparison with his new sequence alignment

programme MUSCLE. The results from the test against BaliBASE (Thompson, Plewniak *et al*., 1999) are presented below. Three other databases of alignments were also tested against, and similar results were found - PREFAB (Edgar 2004); SABmark (van Walle *et al*., unpublished); and SMART (see below).

In practice all three alignment methods were used. MAFFT was typically used as the default; however, alignments were visually examined and if they did not appear satisfactory the other methods were tried. "Good" alignments are considered to have a minimal number of gaps - especially within secondary structural elements, and conserved motifs are immediately apparent. Bad alignments have unnecessary inserts, e.g. 'gappy', and do not line-up conserved motifs and secondary structural elements. For a trivial example of the difference see Figure 1.8.

Method	***Q***	***TC***	***CPU***
T-Coffee	0.882	0.731	1500
ClustalW	0.860	0.690	170
FFT-NS-i	0.844	0.646	16

Q is the number of correctly aligned residue pairs divided by the number of residues pairs in the reference alignment.

TC is the number of correctly aligned columns divided by the number of columns in the reference alignment.

CPU is total CPU time in seconds.

Table 1.1: Results from Edgar's (2004) comparison of MAFFT, T-Coffee and ClustalW.

Whilst I generally found MAFFT and T-Coffee to be the most accurate, they do tend to push sequences to the ends of the alignment and leave gaps in the centre – MAFFT in particular. This is normally fine, but with short families composed of highly divergent sequences, some very poor alignments were produced (e.g. the FTP motif). I found that ClustalW performed the best with this type of family; T-Coffee was somewhere between the two. The accuracy of the alignment (with regards to the integrity of structural elements) does not overly affect the sensitivity of the HMM (Griffiths-Jones and Bateman, 2002) but it does make identification of conserved regions or residues harder and hence make analysis of the family more difficult. For a good review of the different methods of aligning multiple sequences see (Notredame, 2002).

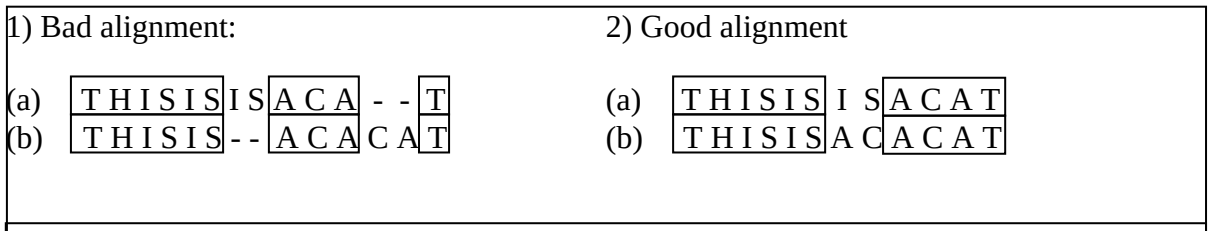

Figure 1.8: A simple example of a "bad" alignment compared to a "good" alignment
The two sequence aligned are "THISISISACAT" and "THISISACACAT"; English words in the strings constitute structural elements. Although simplistic, we can see that alignment 1 unnecessarily used two insertions, and as a result the word "CAT" has been broken up into "CA--T" and "ACA", which is not a word. In 2 no insertions are used, and as a result all English words – and hence structural features - are complete, and we can immediately identify an unconserved loop between "THISIS" and "ACAT".

1.6.3 Databases

Pfam (Bateman, Coin *et al*., 2004)

Pfam is a two tier database for describing proteins. The aim of Pfam is to provide a comprehensive description of the domain content of the protein world, and to provide

tools for querying this data freely to the general research community. Pfam 13 (released April 2004) contains 7426 Pfam-A sequence families, which hit 74% of UniProt at least once (see Figure 1.9 for an example family page from the website). Pfam-A is a searchable database of manually curated sequence families. Each family consists of four primary elements:

(1) A manually inspected SEED alignment of trusted sequences.
(2) A global (ls) and a local (fs) HMM built from the SEED.
(3) A description, including relevant literature.
(4) An ALIGN file created by searching the HMMs against UniProt.

There are four family types in Pfam (see Figure 1.2 for examples). These fit into the definitions given for Domain, Repeat, Motif and Family in chapter 1.2. In Pfam 13 there are 5688 Families, 1464 Domains, 126 Repeats and 38 Motifs. Many of the families may actually represent domains, but a conservative judgement has been taken.

Pfam-B is an automatically generated supplement derived on ProDom (Servant, Bru *et al*., 2002). ProDom is an automatically generated database of predicted domains - an outline of the method is provided in the description of ProDom below. ProDom regions that overlap Pfam-A domains are split or removed, depending on the type of overlap, hence creating an automatic description of homologies not detected by Pfam--A (the process is described by Bateman, Birney *et al*., 2000). Pfam-B contains around 100 000 small families, which hit about 23% of UniProt.

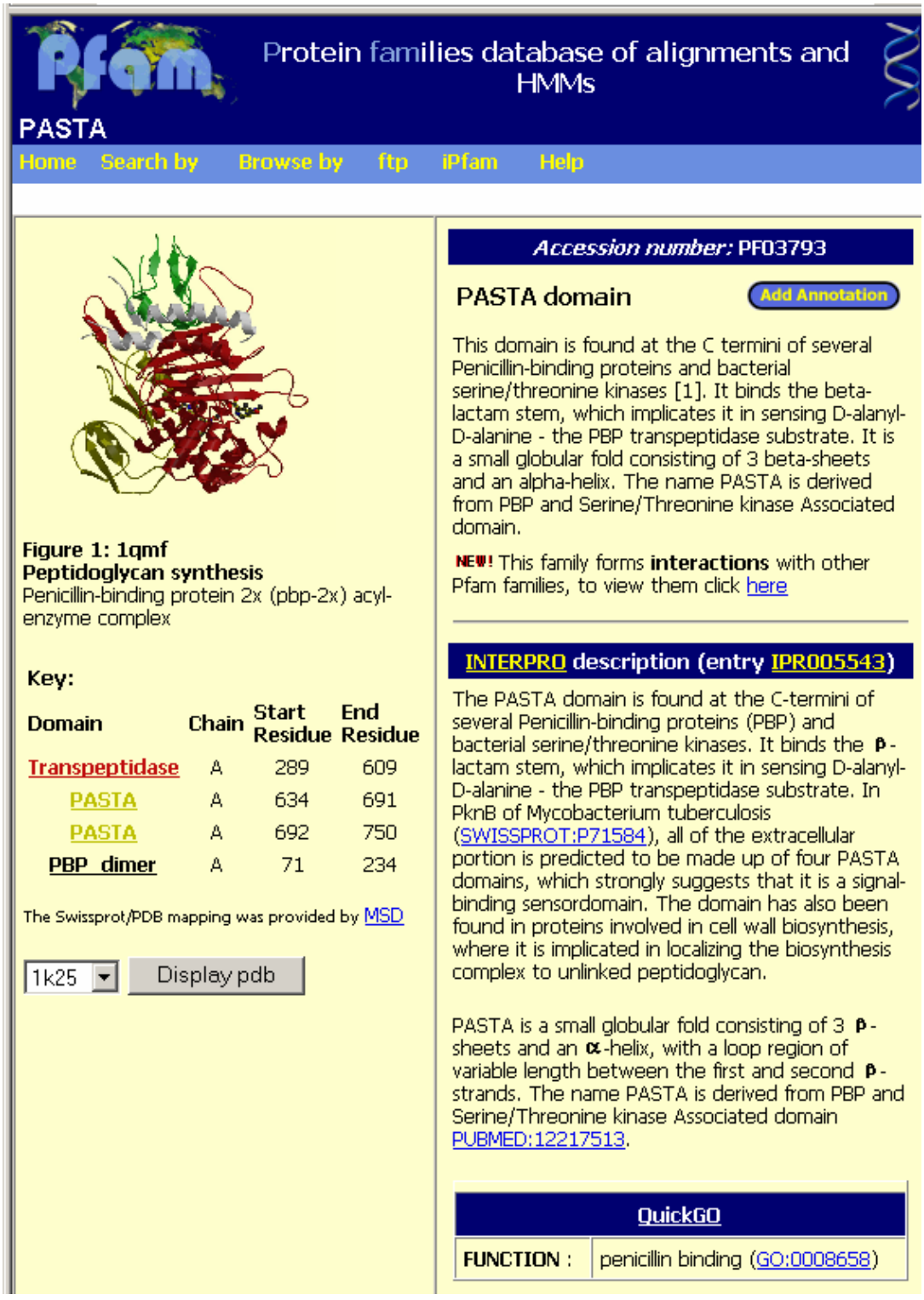

Figure 1.9: Example Pfam Family Page - the PASTA Domain.
Each Pfam family has an automatically generated family page that displays a variety of information about the family. Some of this information is manually entered, while some is imported from other databases (i.e. InterPro), and some is calculated. The links to various tools make Pfam a useful workbench for domain family investigations. In this image the top half of the page is captured, showing annotation and structures. Below are links to graphical representations of the domain architectures, coloured alignments, HMM building information, other databases and cited articles.

InterPro (Mulder, Apweiler *et al*., 2003)

InterPro is a front-end to a collection of databases. InterPro 7.2 (released March 2004) included Pfam (see above), SMART (see below), PROSITE (see below), PRINTS (Attwood, Bradley *et al*., 2003), ProDom (see below), UniProt (see below), TIGRfam (see below), PIR superfamily (Huang and Miller, 1991), SUPERFAMILY (Madera, Vogel *et al*., 2004), CATH (see below), SCOP (see below) and MSD (Golovin, Oldfield *et al*., 2004). It provides facilities for both browsing the data and for searching sequences. The major benefit of InterPro is that it allows you to directly compare the predictions from different domain collections, and also compare these domains against a structural classification from SCOP (if available). Not all these databases were used in the work carried out, so a short description of the relevant ones is given in the section below.

SMART (Letunic, Copley *et al*., 2004)

SMART is similar in form and function to Pfam (see 1.6.3) in its use of HMMs and in its construction of families – though it does not provide the full "ALIGN" files as constructed by Pfam. It is particularly focussed on modelling and describing domains found in signalling, extracellular and chromatin-associated proteins, whereas in other functional categories it is far less comprehensive. As of SMART 4.0 (released March 2004) it contained 667 domains.

PROSITE (Hulo, Sigrist *et al*., 2004)

PROSITE is one of the original collections of sequence patterns (release 1 appeared in 1989). As of release 18.0 it contained "1,639 different patterns, rules or

profiles/matrices" and 1200 documentation entries. This diversity of model types reflects the history of sequence searching during the 1990s. Initially much sequence analysis was carried out using pattern matching techniques such as 'regular expressions'. These patterns tended to take the form "G-x(8,10)-[FYW]-x-G-[LIVM]-x-[LIVMFY]-x(4)-G-K-[NH]-x-G-[STAR]-x(2)-G-x(2)-[LY]-F" (in this case PS00845; CAP_GLY_1). However, profile methods subsequently have come to dominate sequence analysis due to their superior sensitivity, specificity and broader application; as a result PROSITE's earlier models are patterns and their later ones are generalised profiles (Bucher, Karplus *et al*., 1996). PROSITE has detailed documentation for each of its families.

TIGRfam (Haft, Selengut *et al*., 2003)

Release 3 (October 2003) had 1976 families, of which 1004 are "equivalogs", 330 are "other equivalogs" (proposed equivalogs for which the function is not known) and 642 are "other" (families for which it is not known if the function is conserved). Equivalogs are proposed to be families of functional equivalence. The difference in definition to an orthologue is worth noting: orthologues are homologous proteins that have separated due to a speciation event, but the function is not necessarily conserved; in contrast equivologs may be separated by any evolutionary process - such as lateral gene transfer, but the function is conserved. It is a rapidly growing resource – 350 new families were added between release 2.1 and release 3 (about 1 year). The families are more functionally specific than Pfam, allowing for greater confidence in the functional description that accompanies a match, but it is not yet as comprehensive.

ProDom (Servant, Bru *et al*., 2002)

As mentioned above ProDom is an automatically generated domain database, from which Pfam-B is derived. Although automated methods are not as accurate, either in terms of defining the correct domain boundaries or in completeness of the families, ProDom does effectively capture genuine homologies and so can provide a useful starting point for a researcher looking for interesting sub-regions within a protein. The algorithm for its construction is also of interest, as the same principles are behind a method used in this thesis (see chapter 2.1.2). The assumption is made that the shortest amino acid sequence is representative of a domain. This sequence is then searched against UniProt (see below) using PSI-BLAST. Any matching regions and the query sequence are removed from the database and assigned a family number. This process is iterated using the shortest sequence remaining until no sequences with detectable homologies are left. Three filters are applied to the sequence database first; all sequences marked as 'fragment' are removed, low complexity regions are masked using 'seg' (Wootton and Federhen, 1993), and regions shorter than 20 amino acids are excluded.

Swiss-Prot/TrEMBL or UniProt or 'sptr' (Apweiler, Bairoch *et al*., 2004)

The work in this thesis is mostly based on searching HMMs against a sequence database. The sequence database of choice was Swiss-Prot and its supplement TrEMBL. Founded in 1986, Swiss-Prot is a manually curated sequence database, with various functional and structural annotations attached. The increasing rate of DNA sequence production meant that a large volume of data was unavailable between releases, so an automated supplement was created – TrEMBL (Translated EMBL).

Figure 1.1 shows how UniProt has grown between October 2001 and July 2004. It should be noted that much of TrEMBL is redundant; new entries can often already be represented in Swiss-Prot or TrEMBL, or be a fragment of a larger protein. As of 2003 Swiss-Prot merged with the Protein Information Resource (PIR) to form the Universal Protein Knowledgebase (UniProt). This wasn't so much a combining of sequence data, but a merging of resources and infrastructure so as to produce a single high quality database that was able to keep up with the generation of sequence data. As can be seen from the slower growth of Swiss-Prot as compared to the near exponential growth of TrEMBL (see Figure 1.1) this was becoming a problem. As of May 2004 it still consisted of Swiss-Prot and TrEMBL; hence although the later work is done against UniProt rather than Swiss-Prot/TrEMBL, from the researcher's point of view they can be considered interchangeable.

1.6.4 Structural Collections and Classifications

wwPDB (Berman, Battistuz *et al*., 2002)

The Worldwide Protein Data Bank (wwPDB) was established in 1971 as the Protein Data Bank (Bernstein, Koetzle *et al*., 1977) to be “the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data” for the public. It is currently the product of the collaboration between the Japanese PDBj group, the European MSD group, and the American RCSB PDB group - hence "wwPDB" (Berman, Battistuz *et al*., 2002). The other major structural databases – i.e. CATH, SCOP – are all built on top of it. A website provides querying services and an FTP site provides the underlying data freely for download. As of the 4th May 2004 it

contained 25,343 three dimensional structures, including 22,936 proteins, peptides and viruses and representing just under 4000 folds.

PDBSum (Laskowski, 2001)

PDBSum is a web-based interface to summary information contained within the PDB files and from structural analysis software, as well as linking to some relevant structural and sequence data in other databases. The information is presented in a pictorial manner, making it very easy to understand and interpret. It also shows the position of structural domains, as determined by CATH (see below) against the sequence allowing for easy cross-comparison with Pfam.

CATH (Pearl, Bennett *et al*., 2003)

CATH is hierarchical system of protein structure classification based on a combination of automated approaches and manual validation. Proteins are split into domains and the structures characterised. The domains are then described in accordance with eight groups of criteria, which are:

> **Class** – derived according to the secondary structure content: all α; all β; α/β; and "few secondary structures". For example the PASTA domain is α/β (see Figure 4.1), whereas the Hemerythrin domain is all α (see Figure 2.9).
>
> **Architecture** – describes the structure in terms of the orientation of secondary structure elements without reference to their connectivity.

Topology – determined by the order and type of secondary structure elements.

Homologous Superfamily – proteins that are thought to be evolutionarily related and hence homologous.

Sequence Family - groups of structures that show at least 35% sequence identity - as structure is highly conserved at this level.

Non-identical - groups structures that are at least 95% identical; useful for creating non-redundant datasets.

Identical - groups structures that are 100% identical in sequence terms.

Domain - the leaf of the CATH tree; this refers to structural domains as discussed in chapter 1.2.

SCOP (Murzin, Brenner *et al*., 1995)

SCOP (Structural Classification Of Proteins) is another hierarchical system of protein structure classification that categorises domains in terms of their structural elements. The assignments are made based on a variety of evidence, including automated and manual interpretation of the data. The final assignments are determined by expert knowledge; and hence this system is probably the most accurate. There is some delay between a structure being deposited in the PDB and its classification in SCOP - e.g. as of July.9.2004 there were 25977 protein-containing PDB structures, and 20169 classified in SCOP. The classifications are:

Class – The same as CATH's 'Class' (see above), except that SCOP separates the α/β class into two types: α/β, in which the different types of secondary structure are mixed together in the fold; and $\alpha+\beta$, in which the

different types of secondary structure are largely segregated. Also SCOP has a "multidomain protein" class for proteins that consist of several different folds that have no obvious homologues, as well as a membrane protein class and a small protein class; it does not have the "few secondary structures" class.

Fold – groups of structures that have the same major secondary structure elements and topology (same as CATH's 'Topology' above) but show little or no overarching sequence similarity.

Superfamily – groups of structures that are likely to have evolved from a common ancestor, but have significantly diverged in sequence and function.

Family – groups of sequences that can be shown to have evolved from a single ancestor. This is defined by a sequence identity of greater than 30% or high structural and functional conservation.

National Center for Biotechnology Information (NCBI)

The NCBI website provides a simple front end to a range of bioinformatic tools and data resources (Wheeler, Church *et al*., 2004). Of particular relevance is the PSI-BLAST server which searches the "nr" database - a mostly non-redundant composite peptide database made up from compilation of several resources. This provides an analogous system to the HMMER searching of UniProt used in this work, and so is a very useful positive control for the searches carried out. The NCBI also hosts a searchable biological/biomedical literature abstracts database (PubMed), a genetic disease mutations database (OMIM), authoritative taxonomy listings, a BLAST server for partially complete microbial genome sequencing projects and a range of other services.

1.6.5 Presenting Domain Architectures and Alignments

For all the novel families presented in this thesis, three pieces of information are supplied. These are an architecture figure, an alignment figure and a secondary structure prediction. These all conform to the same style discussed here - where there are specific variations these will be noted in the relevant figure caption. The domain architectures are presented in a 'Beads-on-a-String" style of representation. This view represents the protein sequence as a line with features depicted as coloured boxes. The features shown are Pfam-A families, signal peptides (SignalP; Bendtsen, Nielsen *et al*., 2004), transmembrane helices (TMHMM; Krogh, Larsson *et al*., 2001), low complexity regions (seg; Wootton and Federhen, 1993), and coiled-coils (ncoils; Lupas, Vandyke *et al*., 1991). The key to the domain figures is shown below in Figure 1.10, along with a few example architectures. Unless indicated all the images are taken directly, and without alteration, from the Pfam website; this is to ensure that the data shown is publicly available, reviewed and consistent. In general most or all of the different architectures for a family will be shown.

Associated with each protein shown are its UniProt accession, its common name, and the species it is found in. It should also be noted that where possible all the proteins in a figure have been shown on the same scale. However, in some cases members of a domain family can diverge in length by an order of magnitude; in these cases scaled depiction is not realistic. To compensate the lengths are marked by each protein.

The alignments have been drawn in Jalview (Clamp, Cuff *et al*., 2004), using the ClustalX (Thompson, Higgins *et al*., 1994) colouring schema for different amino acid

groups (given below in Table 1.2). The sequences shown are essentially arbitrarily selected but have been picked in order to show the variety in the family as well as its typical form. The colours for each amino acid group are shown in Table 1.2, the colour being chosen according to the residue type and most conserved property in the column. Each sequence is shown with its UniProt accession number and the start/end coordinates of the domain. Another sequence alignment viewer I have commonly used is Belvu by Erik Sonnhammer; however, it does not include the ClustalX colouring scheme and so is not used to create the alignment figure images.

Residue Type	Frequency in Column	Colour	Description
ACFHILMVWY	>60%	Blue	Hydrophobic
DE	>50%	Magenta	Negatively Charged
KR	>60%	Red	Positively Charged
STQN	>50%	Green	Polar Charged
C	>85%	Pink	Cysteine
G	>85%	Orange	Glycine
P	>85%	Yellow	Proline
FYW	>50%	Cyan	Aromatic

Table 1.2: The ClustalX colouring scheme.
This scheme is the one used for the alignment figures shown in this Thesis unless otherwise indicated.

Under each sequence alignment is a secondary structure prediction, unless there is a known three dimensional structure. α-helices are indicated by red cylinders, whereas β-strands are indicated by yellow arrows. These predictions have been made using three programmes: JPred (Cuff and Barton, 2000), PHDsec (Rost, 1996) and PROF (also by B. Rost, but unpublished). Most of the older predictions have been made using JPred, whereas the more recent predictions are made using PROF and PHDsec. The reason for this change is more to do with the development of the servers supplying the service than improvements in accuracy. Whilst in the text for each family it may name either PROF or PHDsec, in reality both methods were run for

each family and it was checked that the results were largely in agreement. The exact output chosen for representation was dependant on how well it agreed with the shape of the alignment. If the two methods showed significant disagreement then the sample alignment was altered and further predictions run. In some cases a transmembrane helix prediction (blue box) takes the place of the secondary structure prediction. The predictions were made using TMHMM (Krogh, Larsson *et al*., 2001).

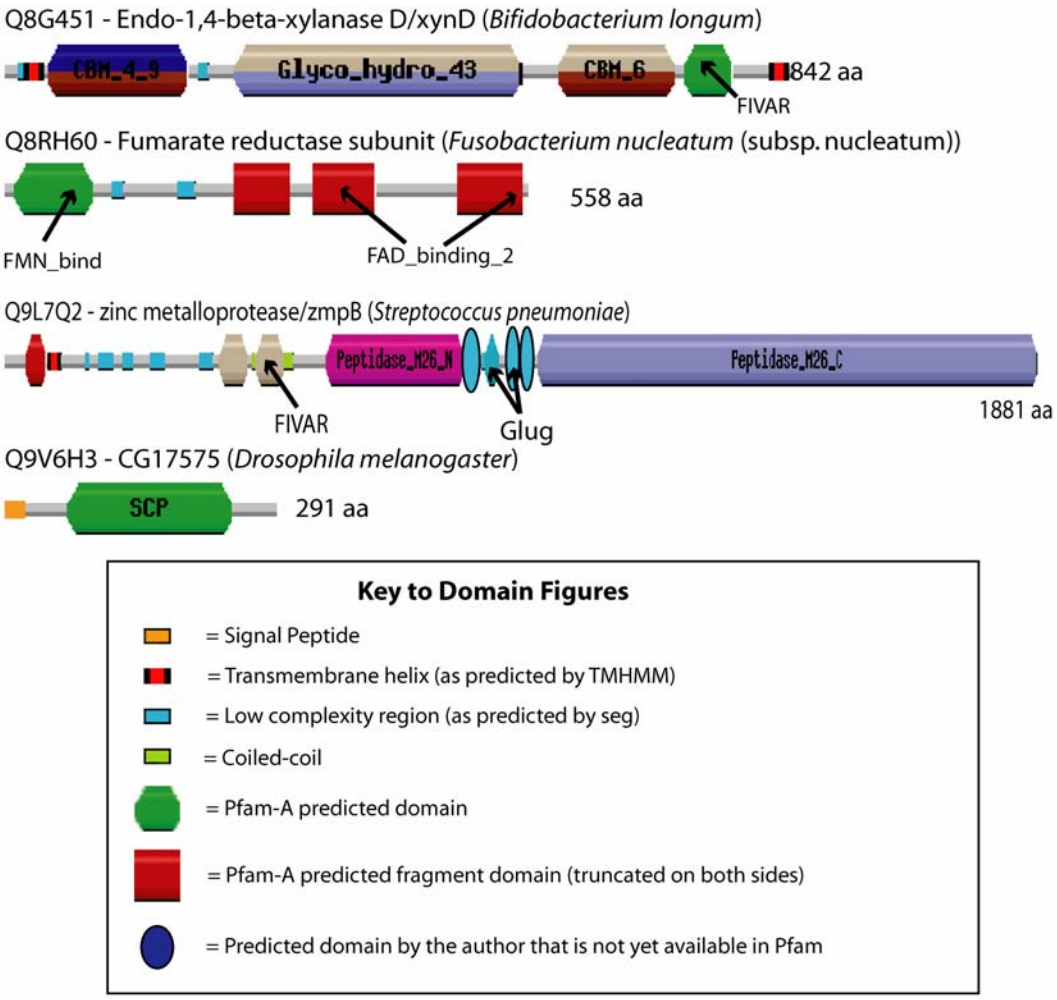

Figure 1.10: Key to the architecture figures and some example architectures
Seven types of features are shown for each protein depicted in the architecture diagrams in this thesis. The methods for predicting each type of feature are noted in the main body of the text. Above each protein is its UniProt accession code, a common gene or protein name and the species in italics. Also near the C-terminus of each protein is its length in amino acids, so as to help relate the size of the proteins. Where possible they have been shown on the same scale.