PROTEIN CLUSTERING AND CLASSIFICATION

ori Sasson[1] and Michal Linial[2]

[1]The School of Computer Science and Engeeniring and [2]The Life Science Institute, The Hebrew University of Jerusalem, Israel

1. Introduction

Proteins are the building blocks of all organisms. The name *protein* is derived from the Greek word ‘protos’ which means first or primal. Indeed proteins are the most fundamental substance of life, as they are the key component of the protoplasm of all cells. In addition to their role as the building blocks of cells and tissue, proteins also play a role in executing and regulating most biological processes. Enzymes, hormones, transcription factors, pumps and antibodies are examples for the diverse functions fulfilled by proteins in a living organism.

Proteins are macromolecules, and consist of combinations of amino acids in peptide linkages, that contain carbon, hydrogen, oxygen, nitrogen, and sulfur atoms. There are only 20 different types of amino acids, and they can be combined to generate an infinite number of sequences. In reality, only a small subset of all possible sequences appears in nature.

In the study of proteins, there are three important attributes of proteins: sequence, structure, and function. The *sequence* is essentially the string of amino acids which comprises the protein. The *structure* of the protein is the way the protein is outlaid in the three dimensional space. Perhaps most important, yet most elusive, is the protein *function*. The protein function is its actual role in the specific organism in which it exists. Understanding the protein function is critical for most applications, such as drug design, genetic engineering, or pure biological research.

The advent of advanced techniques for sequencing proteins in the last two decades, spur the explosive growth witnessed today in protein databases. Due to this rate of growth, the biological function of a large fraction (between one third and one half, depending on the organism) of sequenced proteins remains unknown.

The difficulty of assigning a certain function to a particular protein stems from the fact the function of many proteins is defined by its context in term of protein partner and even its localization. In addition, a protein string may be subjected to large number of modifications that may affect its function and fate. In this survey we will not address any of the dynamic properties of the proteins and will only address the protein as a predetermined string of amino-acids.

A common way to tackle the complexity of protein function prediction uses database searches to

find proteins similar to a new protein, thus inferring the protein function. This method is generalized by protein *clustering* or *classification*, where databases of proteins are organized into groups or families in a manner that attempt to capture protein similarity.
We survey the field of protein clustering and classification systems. Such systems use the protein sequence, and at times structure, to classify proteins into families. The classification may be leveraged towards function inference.
The structure of this survey is as follows. We start by describing the most commonly used algorithms for sequence similarity, and the way they can be used directly for protein classification. The following sections describe classification systems based on the methodology used: motif-based classifications, full-sequence analysis classifications, phylogenetic classification, and structure based classifications, aggregated classifications making use of the results of other classification systems. The last section provides a summary.
It is important to note that vast amounts of research were made in the field of proteomics which are related to the topic of this survey. Herein, we focus primarily on publicly available software systems for protein classification. Furthermore, due to space constraints we describe only a selected subset of systems and methods available. An attempt was made to represent the full spectrum of methods and directions.

2. Sequence similarity

One of the most common approaches towards classifying proteins is using sequence similarity. Sequence similarity is a well studied subject, and numerous software packages suited for biological sequences are available. Such packages (e.g. BLAST) are probably the most widely used software in the fields of biology and bioinformatics.
Sequence similarity algorithms (and software) take as input two sequences and provide a measure of distance or similarity between them. Note that these quantities are related in the sense that the higher the distance the lower the similarity, and vice versa.
The notion of distance between sequences has been formalized by Levenshtein (1965), who has introduced a dynamic programming algorithm for determining this distance, referred to as ‘edit distance’. The ‘edit distance’ between two strings is defined as the number of insertions, deletions, and replacements of characters from the first string required to obtain the second string. Some variants of the edit distance allow for reversals of sub-strings.
The edit distance problem is strongly related to the problem of string alignment. The problems are essentially equivalent, as the alignment can be easily produced from a set of insertions and deletions of characters. In the context of biological sequences, similarity and alignment were first studied by Needleman and Wunsch (1970). The Needleman-Wunsch sequence similarity and sequence alignment are usually referred to as *global* sequence alignment. In other words, this is alignment of full length sequences. In practice, only extremely similar sequences can be nicely globally aligned. However, many proteins exhibit strong *local* similarity. The local alignment problem was studied by Smith and Waterman (1981).

Algorithms for calculating either local or global similarity for protein sequences do not give equal weight to all amino-acids. Instead, scoring matrices are used to achieve an alphabet-weighted similarity. An alphabet-weighted edit distance can also be defined using such scoring matrices, giving different weight to replacement of different characters. The most commonly used scoring matrices are BLOSUM (Henikoff and Henikoff, 1992), and the earlier PAM (Dayhoff *et al*., 1978). The most popular algorithm and software package used for global and local similarity calculation is BLAST, along with a variant of it called PSI-BLAST. Other algorithms are FASTA and Smith-Waterman's algorithm.

2.1. Smith-Waterman

The Smith-Waterman method (Smith and Waterman, 1981) searches for local alignment. In other words, instead of looking at the entire length of each sequence, it compares substrings of all possible lengths. The Smith-Waterman algorithm is based on *dynamic programming*. This is an algorithmic technique where a problem is solved by caching the solutions of sub-problems, and these are used in later stages of the computation.

In the case of sequence alignment, the idea is to calculate the best local alignment score at a certain location along the string based on the best scores in the previous locations. This process is typically described in a table, where the rows correspond to one sequence and the columns to another sequence (see Figure 1).

The alignment scores are based on the notion of 'edit distance', counting the number of transformation one sequence is required to obtain the second sequence. Transformations include substituting one character for another, insertion of a string of characters, or deletion of string of characters. The actual score is calculated using score matrices which associate a weight with each pair of characters. The algorithm uses two penalties, one for opening gaps and another for extending them. The formula shown in Figure 1 combines these penalties into one penalty g_k which is the penalty for opening a gap of length k. In the same figure, $s(a,b)$ denotes the score matrix entry associated with the amino acids a and b.

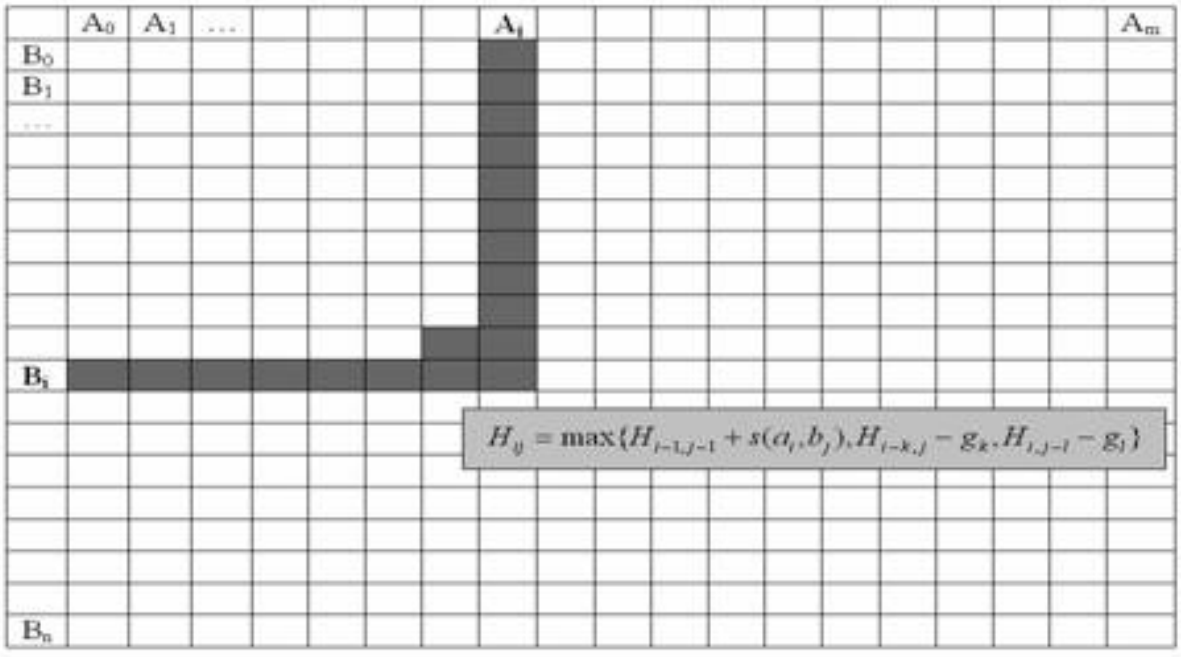


Figure 1: Smith-Waterman Algorithm Dynamic Table

2.2. FASTA

The Smith-Waterman algorithm provides a good measure of local alignments, but at a cost. In a naïve implementation, its running time is cubic in the lengths of the compares sequences. As sequence database grew, the need for a more efficient comparison method emerged.
FASTA (Pearson, 1990) is a heuristic method which provides an approximation of the local alignment score. It is based on the following observation: good local alignments typically stem from identities in sequences. The FASTA algorithm constructs a lookup table which stores all instances *k*-tuples of amino-acids appearing in the sequence. The typical value of *k* used is k=2, which means the lookup table stores all instances of pairs of proteins. Using this lookup table, the best regions with the highest density of identities are identified. These identities, viewed on full dynamic programming table, can be considered as k-length diagonals. With these diagonals at hand, FASTA searches for the best 'diagonal runs', which are sequences of consecutive identities on a single diagonal. Finally, a dynamic programming algorithm is used for these areas only (using the Needleman-Wunch algorithm).

2.3. BLAST

BLAST (Altschul *et al.*, 1997) is another heuristic method for local alignment. Instead of looking for identical k-tuples in sequences, it looks for *similar* k-tuples. It looks for k-tuples (the typical value used for k in protein comparison is 3) in one sequence that score at least T when aligned with the other sequence, again using a scoring matrix. Such local similarities are the extended in both directions in an attempt to find locally optimal un-gapped (i.e. continuous) alignments, with a score of at least S. These alignments are called high scoring pairs (HSPs), and are combined to provide the best local alignment of the two strings. The neighborhood threshold T and the score threshold S

are tunable parameters.
BLAST is even faster than FASTA, as it does not use dynamic programming. It is considered to be as sensitive (and thus as accurate) as FASTA, and for this reason it is currently the most popular sequence search and comparison tool, both for amino acid and nucleic acid sequences.
BLAST outputs the similarity score, sequence alignments, and the statistical significance of the similarity, referred to as E-score. The latter provides an estimate of the probability of having similarity of this quality with a random string.

2.4. PSI-BLAST

PSI-BLAST (Altschul *et al.*, 1997) is a variant of BLAST which performs databases searches in an iterative fashion. Given a query sequence and a database of sequences, BLAST is used to find the sequences in the database that are most similar to the query sequence. For these sequences a multiple alignment is constructed, and based on this multiple alignment, a profile is generated. The profile is a position specific scoring matrix which holds the probability of having each amino-acid in each on of the positions in the sequence. This profile is now compared to the protein database, seeking for local alignments, using an algorithm similar to BLAST. A possibly new set of sequences in the database are found, which match the profile. This process can be repeated for this set and so on, until convergence occurs (i.e. the BLAST result is identical to the set of sequences from which the profile was constructed) or for a fixed number of iterations.
PSI-BLAST is considered the program of choice for detecting remote homologues. Yet, the exhaustive iteration scheme often results in considering non-related sequences among the correct hit list.

2.5. Classification with Sequence Similarity

Several inherent challenges exist in classifications proteins that are based on their sequences similarities. The current protein databases combine proteins having different evolutionary history. For example, the sequences of many proteins that were evolved from a common ancestor were already diverged beyond detection by any of the search programs described above. Other proteins may still exhibit extremely high degree of conservation in their sequences despite very long evolutionary distances. Thus, it is evident that varying the distance matrix selected, defining the penalty for opening and extending gaps in the alignments and choosing the preferred statistical and computational parameters are fundamental for any sequence-based classification.
The Sequence similarity measures such as BLAST and Smith-Waterman are used as the building block for some classification systems, detailed in Section 4. Such measures can be used directly for classification, using the well-known “Nearest-Neighbor” paradigm for supervised learning. In other words, a new protein is associated with the protein nearest to it in the database of proteins with known function. The BLAST software package even provides a clustering program, called `blastclust`.

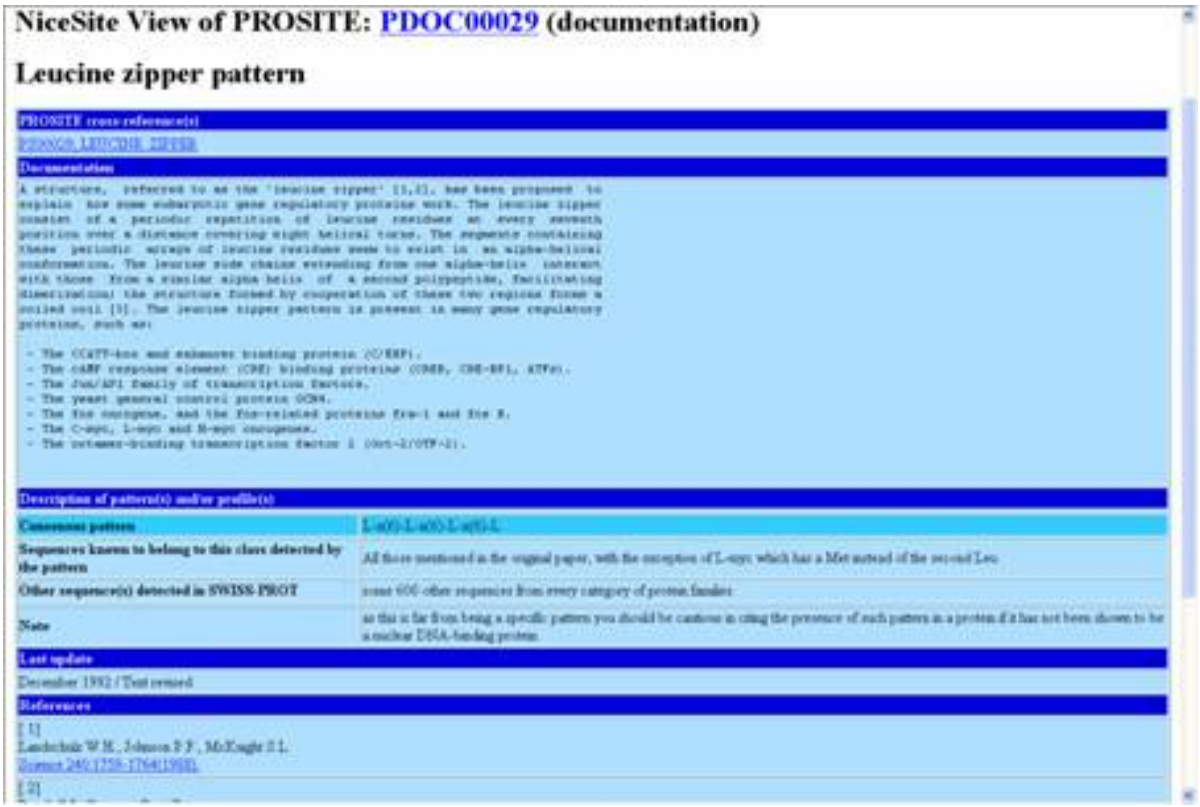
Similarity based classification is typical done by searching for all proteins similar to a given protein up to a specified threshold. One caveat of such classification is the choice of threshold. If the threshold is chosen to be too low (i.e. too restrictive), it is possible no matches are found. Alternatively, if the threshold is chosen to be high (i.e. permissive) it is likely that a lot of non-related sequences, or 'false positives' shall be retrieved. This problem frequently surfaces when classifying sequences which have diverged during evolution, and thus the similarity is difficult to detect, at least without having a large fraction of false positives.

3. Classification based on domain and motif Analysis

While sequence comparison is the most widely used tool for classifying proteins, it is not sufficient in all cases, as explained above. An alternative approach is based on the notion that perhaps domains form the building block of proteins and not single amino acids. Based on this notion, proteins with similar domains are associated with each other, even if they have low similarity scores.
A variety of classification systems were developed which are based on domain and motif analysis. Such systems use multiple sequence alignments to detect conserved regions in sequences. The differentiating factor between the different classifications based on domain and motif analysis is the underlying computational representation of a motif or domain.

3.1. PROSITE

PROSITE (http://www.expasy.ch/prosite, Falquet *et al.*, 2002) is the oldest motif-based classification of proteins. The goal set out by PROSITE is to identify and represent all biologically significant signatures (or 'fingerprints'). Signatures are described either as *regular expressions* or as profiles (similar to the profiles described above for PSI-BLAST). Release 17.31 dated December 15th 2002 contains 1,156 families described by 1585 different patterns, rules, and profiles.


NiceSite View of PROSITE: PDOC00029 (documentation)

Leucine zipper pattern

PROSITE cross-reference(s)

PDOC00029; LEUCINE_ZIPPER

Documentation

A structure, referred to as the 'leucine zipper' [1,2], has been proposed to explain how some eukaryotic gene regulatory proteins work. The leucine zipper consist of a periodic repetition of leucine residues at every seventh position over a distance covering eight helical turns. The segments containing these periodic arrays of leucine residues seem to exist in an alpha-helical conformation. The leucine side chains extending from one alpha-helix intersect with those from a similar alpha helix of a second polypeptide, facilitating dimerization; the structure formed by cooperation of these two regions forms a coiled coil [3]. The leucine zipper pattern is present in many gene regulatory proteins, such as:

- The CCAAT-box and enhancer binding protein (C/EBP).
- The cAMP response element (CRE) binding proteins (CREB, CRE-BP1, ATFs).
- The Jun/AP1 family of transcription factors.
- The yeast general control protein GCN4.
- The fos oncogene, and the fos-related proteins fra-1 and fra 2.
- The C-myc, L-myc and N-myc oncogenes.
- The octamer-binding transcription factor 1 (oct-1/OTF-1).

Description of pattern(s) and/or profile(s)

Consensus pattern	L-x(6)-L-x(6)-L-x(6)-L
Sequences known to belong to this class detected by the pattern	All three mentioned in the original paper, with the exception of L-myc which has a Met instead of the second Leu
Other sequence(s) detected in SWISS-PROT	some 600 other sequences from every category of protein families
Note	as this is far from being a specific pattern you should be cautious in citing the presence of such pattern as a protein if it has not been shown to be a nuclear DNA-binding protein

Last update

December 1992 / Text revised

References

[1]
Landschulz W.H., Johnson P.F., McKnight S.L.
Science 240:1759-1764(1988).

[2]

Figure 2: PROSITE Entry PS00029

A regular expression is a concise way to describe a family of strings. Regular expressions are commonly used when searching for files (e.g. the DOS command '`DIR *.EXE`' is using a simple regular expression). Regular expressions in PROSITE are described using the following rules:

1. Standard one-letter codes for amino-acids are used
2. The symbol 'x' is used as a wildcard, in a position where any amino acid is accepted.
3. At positions where one of several amino acids may be used, square parentheses are used. For example [AT] stands for 'A' or 'T'.
4. At positions where most amino acids can be used, curly brackets are used. For example {AT} stands for any amino-acid other than 'A' or 'T'.
5. Elements in a pattern are separated by a hyphen ('-').
6. Repetition of an element is indicated by a number in parenthesis. For example x(3) means x-x-x and x(2,3) means x-x or x-x-x.
7. '<' and '>' indicate when a pattern is restricted to either the N- or C- terminal of a sequence respectively.
8. Patterns end with a dot ('.').

For example, the PROSITE entry PS00029 (see Figure 2) describes the *Leucine zipper* pattern as L-x(6)-L-x(6)-L-x(6)-L. This reads as 'L' followed by any 6 amino acids, followed by another L with 6 more amino acids, follows by an L with 6 more amino acids, and a final L.

PROSITE is essentially a manually maintained database. This allows each entry to be associated with all relevant literature references. In fact, many of the entries are generated via published multiple-alignments. In addition, cross references to PDB (Berman *et al.*, 2000) are provided when

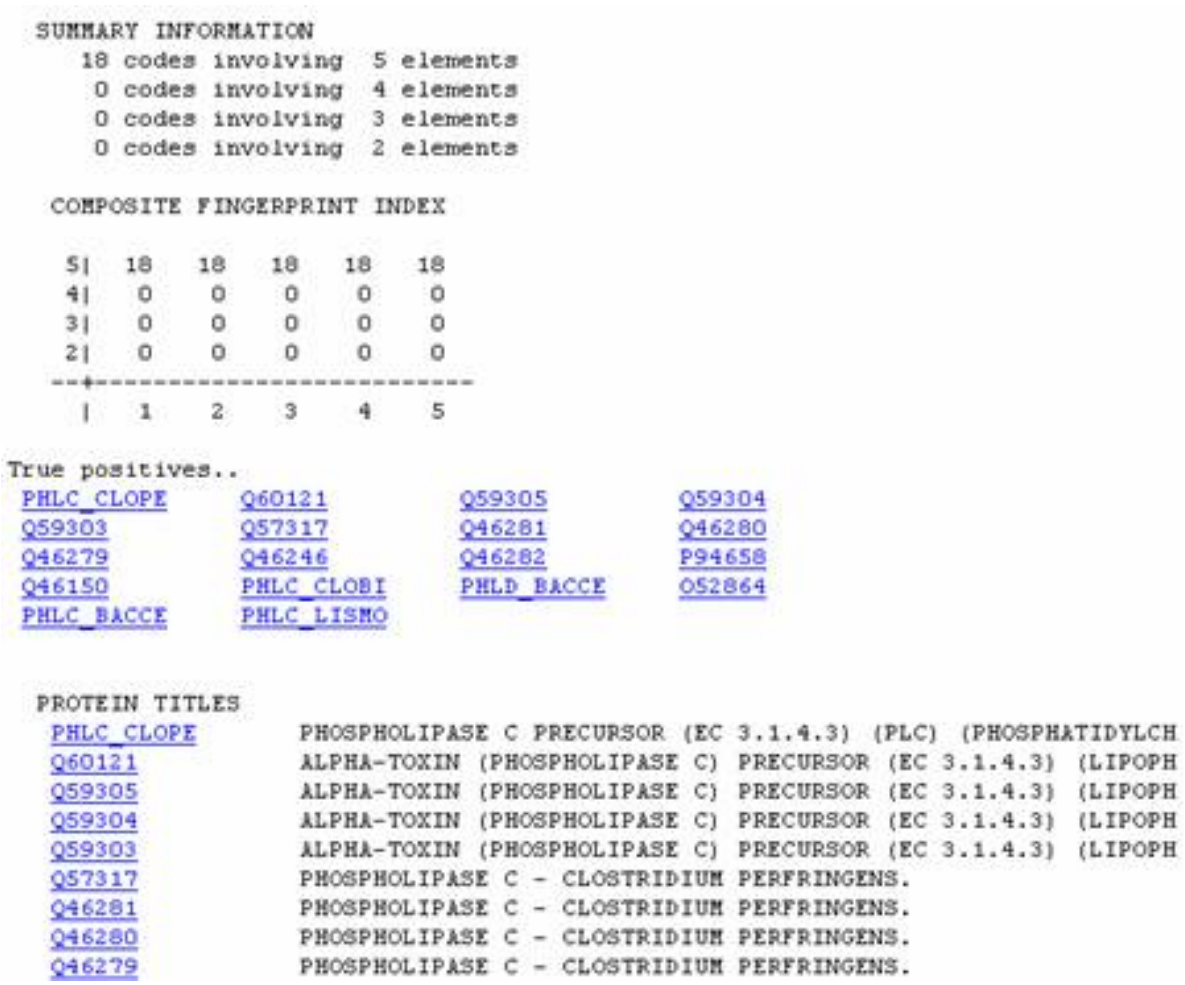
applicable. PDB is a database of proteins for which three-dimensional structures are known. Regular expressions are not suitable for classifying families whose members are highly diverged. To this end, PROSITE was extended to include profiles, thus extending its coverage.
While PROSITE provides a dictionary of motifs and domains it has several drawbacks as a classification system. One problem is that of missing patterns. As patterns are detected mostly manually, not all patterns existing in real-world proteins have been detected. Another problem is that of low-information patterns. Since pattern lengths can vary from a few amino-acids to hundreds of amino-acids, some patterns (the shorter ones) have little value in classifying a protein.

3.2. BLOCKS

BLOCKS (http://www.blocks.fhcrc.org, Henikiff *et al.*, 2000) is a database of highly conserved protein regions. A ‘block’ is defined as a contiguous segment corresponding to the most conserved regions of proteins. Such blocks are automatically detected. In contrast to PROSITE, blocks are not associated with function or with known literature, as the process is completely automated. Blocks are derived by performing multiple alignments of protein families as defined in InterPro, and searching for segments with a high number of identities.
Version 13.0 of the Blocks database (dated August 2001) consists of 8656 blocks, representing 2101 groups documented in InterPro 3.1. These blocks are keyed to SwissProt 39.17.
It is important to note that the PROSITE pattern is not used to build to Blocks database, and a Block entry may or may not contain the PROSITE pattern corresponding to the InterPro group from which the entry was derived.
An important application of BLOCKS is generated amino acid substitution matrices. Such matrices are used for the sequence similarity algorithm mentioned above (Smith-Waterman, FASTA, and BLAST). The BLOSUM (Henikoff and Henikoff, 1992) matrices are derived from the BLOCKS database, and are currently the substitution matrices most widely used for sequence comparison.

3.3. PRINTS

PRINTS (http://www.bioinf.man.ac.uk/dbbrowser/PRINTS, Attwood *et al.*, 2002) is a database of domain-family fingerprints. A fingerprint is defined as a group of conserved motifs used to characterize a protein family. Fingerprints are more powerful constructs than single motifs, in the sense that they can narrow down families more effectively.



```
SUMMARY INFORMATION
   18 codes involving  5 elements
    0 codes involving  4 elements
    0 codes involving  3 elements
    0 codes involving  2 elements

 COMPOSITE FINGERPRINT INDEX

  5|  18   18   18   18   18
  4|   0    0    0    0    0
  3|   0    0    0    0    0
  2|   0    0    0    0    0
 --+-------------------------
   |   1    2    3    4    5

True positives..
 PHLC_CLOPE     Q60121         Q59305         Q59304
 Q59303         Q57317         Q46281         Q46280
 Q46279         Q46246         Q46282         P94658
 Q46150         PHLC_CLOBI     PHLD_BACCE     O52864
 PHLC_BACCE     PHLC_LISMO

 PROTEIN TITLES
  PHLC_CLOPE       PHOSPHOLIPASE C PRECURSOR (EC 3.1.4.3) (PLC) (PHOSPHATIDYLCH
  Q60121           ALPHA-TOXIN (PHOSPHOLIPASE C) PRECURSOR (EC 3.1.4.3) (LIPOPH
  Q59305           ALPHA-TOXIN (PHOSPHOLIPASE C) PRECURSOR (EC 3.1.4.3) (LIPOPH
  Q59304           ALPHA-TOXIN (PHOSPHOLIPASE C) PRECURSOR (EC 3.1.4.3) (LIPOPH
  Q59303           ALPHA-TOXIN (PHOSPHOLIPASE C) PRECURSOR (EC 3.1.4.3) (LIPOPH
  Q57317           PHOSPHOLIPASE C - CLOSTRIDIUM PERFRINGENS.
  Q46281           PHOSPHOLIPASE C - CLOSTRIDIUM PERFRINGENS.
  Q46280           PHOSPHOLIPASE C - CLOSTRIDIUM PERFRINGENS.
  Q46279           PHOSPHOLIPASE C - CLOSTRIDIUM PERFRINGENS.
```

Figure 3: PROSITE *fingerprint for Prokaryotic zinc-dependent phospholipase C*

The fingerprinting method used in PRINTS relies on performing an iterative process of multiple sequence alignments. The process starts with only a small number of proteins, and once a set of conserved regions forming a finger print is identified, the full database is scanned to find matching proteins. The possibly larger set of proteins is then analyzed to generate more motifs. The whole process is repeated with the possibly larger set, until convergence.

PRINTS shares the main drawback of other domain-based systems, which is the lack of full coverage of the protein space. Another problem is that fingerprints are at times too restrictive. Thus it is possible that a new protein will only match a portion (several motifs) in a fingerprint. In such cases, the protein might belong to a sub-family, but there is no clear-cut classification for it.

Figure 3 shows an example of the fingerprint for *Prokaryotic zinc-dependent phospholipase C*.

PRINTS provides a visualization tool for multiple alignments called CINEMA, as shown in Figure 4.

PRINTS release 35.0 dated July 2002 contains 1,750 database entries, relating to 10,626 motifs.

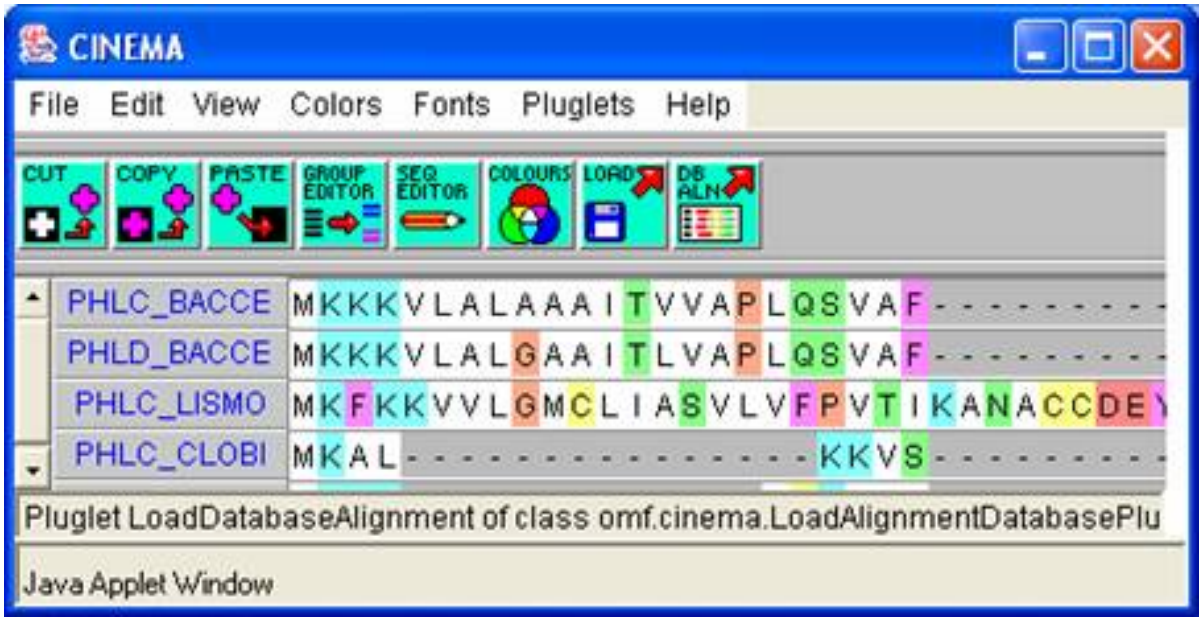


Figure 4: PRINTS Multiple Alignment with CINEMA

3.4. PFAM

Pfam (http://www.sanger.ac.uk/Software/Pfam, Bateman *et al.*, 2000) is a database of protein alignments and HMMs. HMM stands for 'Hidden Markov Model'. Version 7.8 dated November 2002 contains 5049 families.
A hidden Markov model is an abstraction used to statistically describe the consensus sequence for a

protein. Such a model consists of a sequence of node, with a designated begin state and end state. Each node in an HMM has three ‘invisible’ states associated with it (hence the name ‘hidden’). These are a match state (M), insert state (I), and delete state (D). Each has a transition probability associated with it. This probability is node-specific, and hence is position specific in terms of the sequence. The match state has a probability of matching a particular amino-acid. This probability is referred to as ‘emitting’ probability. Similarly, in the insert state there is a probability associated with each amino-acid. The probability of no amino-acid associated with a node is captured by the probability of transitioning into the delete state.

HMMs can be automatically generated from multiple alignments. One of the most popular software packages for generating HMMs, calls HMMER (Sonnhammer, *et al.*, 1998), which is the one used in Pfam.

Given a new sequence, it is possible to evaluate the probability of that sequence belonging to the family modeled by certain HMM. A similarity score is associated with a new sequence based on the most probable path through the HMM which generates the input sequence.

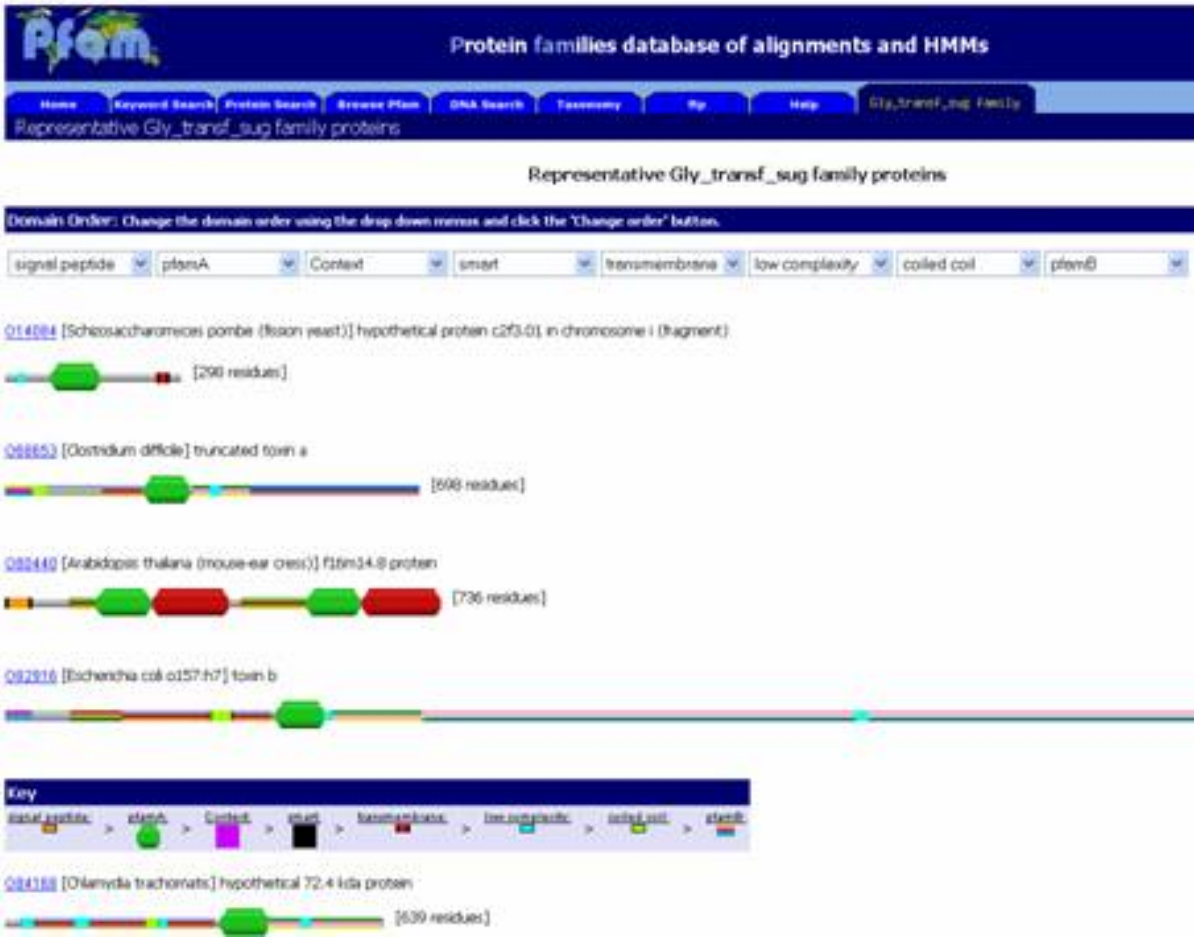


Figure 5: Pfam Graphical Representation of a Protein Family

Pfam comes in two flavors: Pfam-A is a manually curated version of Pfam, and Pfam-B is an automating clustering of the remaining proteins in SwissProt and TrEMBL. Pfam-A covers roughly 65% of the SwissProt database.

The Pfam-A database is generated in a semi-automated process, starting from a seed multiple alignments based (either from the literature or from other databases such as Prosite or ProDom). After manual inspection an HMM profile is built, and used to search the database. Members are added to the seed alignment and the process is repeated. Pfam-A HMMs do not overlap.

Pfam offers a comprehensive Web-based interface which provides links to InterPro (see below) and other systems. Each entry includes the HMM itself, and structural information if available. Pfam also offers a graphical representation of families. For example, Figure 5 shows the graphical representation of the *Gly_transf_sug* family proteins.

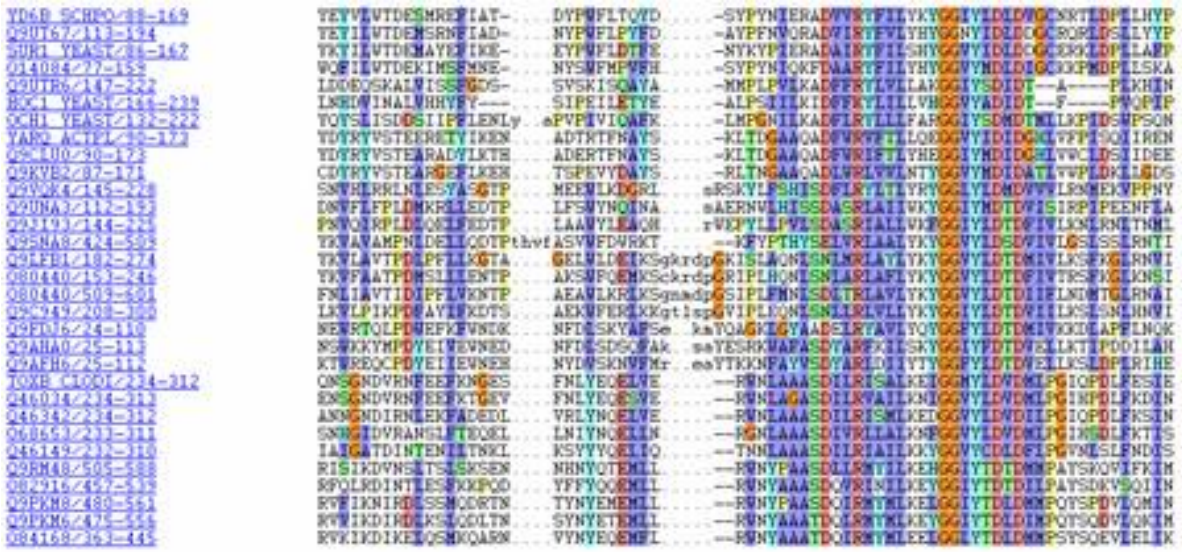


Figure 6: Pfam Multiple Sequence Alignment

Figure 6: Pfam Multiple Sequence Alignment

Figure 6 shows a multiple alignment for the same family. Other visualization methods include a Phylogenetic tree and domain organization by species.
The advantage of HMMs is the fact they implement position-specific scoring, in contrast to ordinary sequence similarity measures. This allows more accurate modeling of families, with less chance similarities occurring. The disadvantage of HMMs (compared to regular expressions for example) is that they are typically quite large and thus difficult to understand. Another caveat is that it is typically difficult to detect subfamilies.

3.5. PRODOM

Figure 7: ProDom Entry PD000089

ProDom (http://protein.toulouse.inra.fr/prodom, Corpet *et al.*, 2000) is a database of automatically generated protein domains.

ProDom identifies domains using BLAST search. Originally it used plain BLAST to identify domains, and now it used PSI-BLAST. ProDom groups all SwissProt and TREMBL sequences into domains, and generated 365,172 families in version dated May 2002.

ProDom clustering is generated under the assumption that the shortest full-length sequence is a single-domain protein. PSI-BLAST is used to find homologous domains, which are then clustered together with that sequence. The process is repeated for the remaining sequences.

The ProDom Web-site provides various search facilities, and the output provided for a ProDom entry is by default a multiple alignment. For example, Figure 7 shows the multiple alignments for the ProDom entry PD000089, *Heat Shock ATP-Binding chaperone*. Figure 8 shows an alternative representation as a phylogenetic tree.

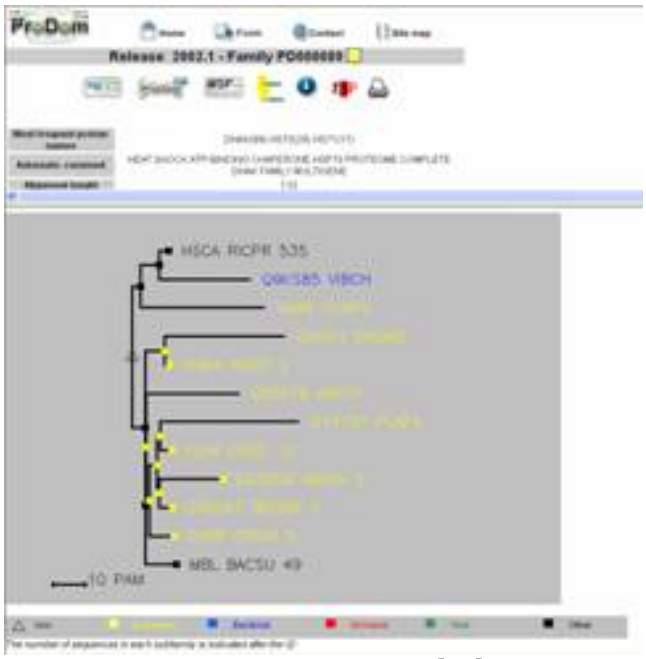


Figure 8: ProDom Phylogenetic Tree Representation

3.6. DOMO

DOMO (http://www.infobiogen.fr/services/domo, Gracy and Argos, 1998a) is a protein domain database generated in a fully automated fashion. The current version of DOMO, dated December 2002, includes 8877 multiple sequence alignments and 99058 protein domains. The database was generated from 83054 non-redundant protein sequences taken from SwissProt and PIR.
DOMO clustering is based on iterative sequence similarity search followed by multiple sequence alignments. Global similarities are detected from the pairwise comparison of amino acid and dipeptide compositions of each protein.
One representative sequence is chosen from each of the generated clusters. For these representatives, a suffix tree is constructed. This suffix tree is then used to detect local sequence similarities. Finally, proteins with similar sequence sequences are multiply aligned, and the alignments are analyzed to detect domain boundaries.
A comparative study (Gracy and Argos, 1998b) indicated that the performance of DOMO (in terms of correctness of classification) is significantly better of that exhibited by ProDom, and falls only slightly short of PROSITE (which is manually maintained).

3.7. SMART

ProDom (http://smart.embl-heidelberg.de, Schultz *et al.*, 2000) is a database of domains. SMART version 3.4 dated December 2002 contains 654 HMMs.
SMART builds HMMs based on multiple sequence alignments of hand-picked family members. PSI-Blast profiling is used to build HMMs which are then used to retrieve more matching sequences. This process is repeated until convergence.

SMART provides a facility for searching for domains in a query proteins. Matching proteins are displayed graphically as shown in Figure 9 below.

Figure 9: SMART Domain Graphical Display

3.8. S-BASE

S-BASE (http://www.icgeb.org/sbase, Vlahovicek *et al.*, 2002) is a protein domain library. S-BASE provides clustering of functional and structural domains.
S-BASE uses a fairly simple strategy to construct domains. It performs BLAST searches against a manually annotated database of subsequences. S-BASE thus heavily relies on good annotation of domains.
S-BASE provides the facility for BLAST of PSI-BLAST searches against the domains database.

3.9. TIGRFAMS

TIGRFAMs (http://www.tigr.org/TIGRFAMs, Haft *et al.*, 2001) is a database of protein families based on HMMs. TIGRFAMs is built using manually collected multiple sequence alignment, along with associated functional information. TIGRFAMs places an emphasis on protein function, and is biased towards microbial genomes.
The TIGRFAMs Web-based interface shows both the name and literature reference for each HMM

generated. The TIGRFAMs database also includes Pfam HMMs, and provides reference to both Pfam and InterPro families.

3.10.EMOTIF

eMOTIF (http://dna.stanford.edu/identify, Huang *et al.*, 2002) is a database of motifs derived from the multiple sequence alignments in BLOCKS (Version 13.0; August, 2001) and the PRINTS database (Version 31.0; June, 2001).
The eMOTIF Web-site allows users to input a query sequence and look for instances of the motifs in the sequence.

4. Classification based on full protein sequence

Domain-based classifications are somewhat limited in the sense that many proteins have several domain appearances, and that some proteins do not have any domain (or at least not a recorded domain). In addition, for small families (e.g. with 2 members) it is not possible to define domains. A different approach to classification is offered by several systems which classify proteins based on the full sequence. This is typically achieved by sequence comparison.
The basic tenet of most full-sequence based classifications is that of homology transitivity. The idea is that homology (the relation between two proteins which have evolved from the same protein) is a transitive relation, whereas similarity is not necessarily a transitive relation. The main caveat of all such classifications is chance similarities, which result in misclassification, and multi-domain proteins which may be related to several families.

4.1. ProtoMap

ProtoMap (http://protomap.cs.cornell.edu, Yona *et al.*, 2000a) provides a fully automated hierarchical clustering of the protein space. ProtoMap takes a graph-based approach, where the sequence space is represented by a directed graph where vertices are protein sequences and edges represent similarity between proteins. The weight associated with an edge measures the similarity, or the significance of the relationship. ProtoMap uses a combination of Smith-Waterman, FASTA, and BLAST to determine similarity.
The ProtoMap algorithm constructs a partitioning of the protein database at different levels of granularity. At first only the most significant relationships are considered. The subgraph induced by all edges with high similarity (e.g. E-value of 1e-100 or less) is used, and each connected components is considered a single cluster. These clusters are then iteratively merged in agglomerative hierarchical clustering, using an average-link where the average is geometric mean of sequence similarity scores.

Hierarchical clustering, as the name suggests, is a clustering techniques which generates a hierarchy of clusters, where at each level clusters are generated by merging clusters of a lower level, and at the bottom level each cluster is a single entity (e.g. a single protein in our case). There are two fundamental paradigms for hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). The former starts from the bottom, from single entities, and repeatedly cluster pairs of clusters into larger clusters. At each step the pair of clusters with the highest similarity is merged. In the divisive approach, the whole set of data items is considered, and is recursively split into smaller clusters.

In practice, when using hierarchical clustering for proteins, usually agglomerative clustering is used. ProtoMap uses average-link clustering, which means that the similarity of two clusters is defined as the average of pairs where one element is taken from the one cluster and another element form the other cluster.

Coming back to ProtoMap, once the hierarchical clustering procedure is completed, the threshold used is increased. The same process is repeated for 1e-95, 1e-90, etc, up to 1e0. Each time the threshold is increased, strongly connected clusters are merged and hierarchical clustering applied again.

Figure 10: ProtoMap Cluster Details

ProtoMap offers a Web-based interface which allows browsing the clusters both textually and graphically. ProtoMap also maintains SwissProt annotation and keywords, as well as links into BioSpace (see 6.4 below). Figure 10 shows a sample cluster page in ProtoMap (for cluster #1, at level 1e-0). The cluster page details each member protein, along with its SwissProt keywords. ProtoMap provides individual protein pages linking each protein to all clusters containing it. Each cluster also has a summary page with some additional information such as PROSITE families and taxonomy. ProtoMap provides a graphical display of cluster relationships (see Figure 11), as well as a tree-like representation.

Release 3.0 of ProtoMap covers 365,174 proteins taken from SwissProt/TREMBL.

ProtoMap was used as a platform for target selection for mapping 3D structures of proteins (Portugaly *et al*., 2002).

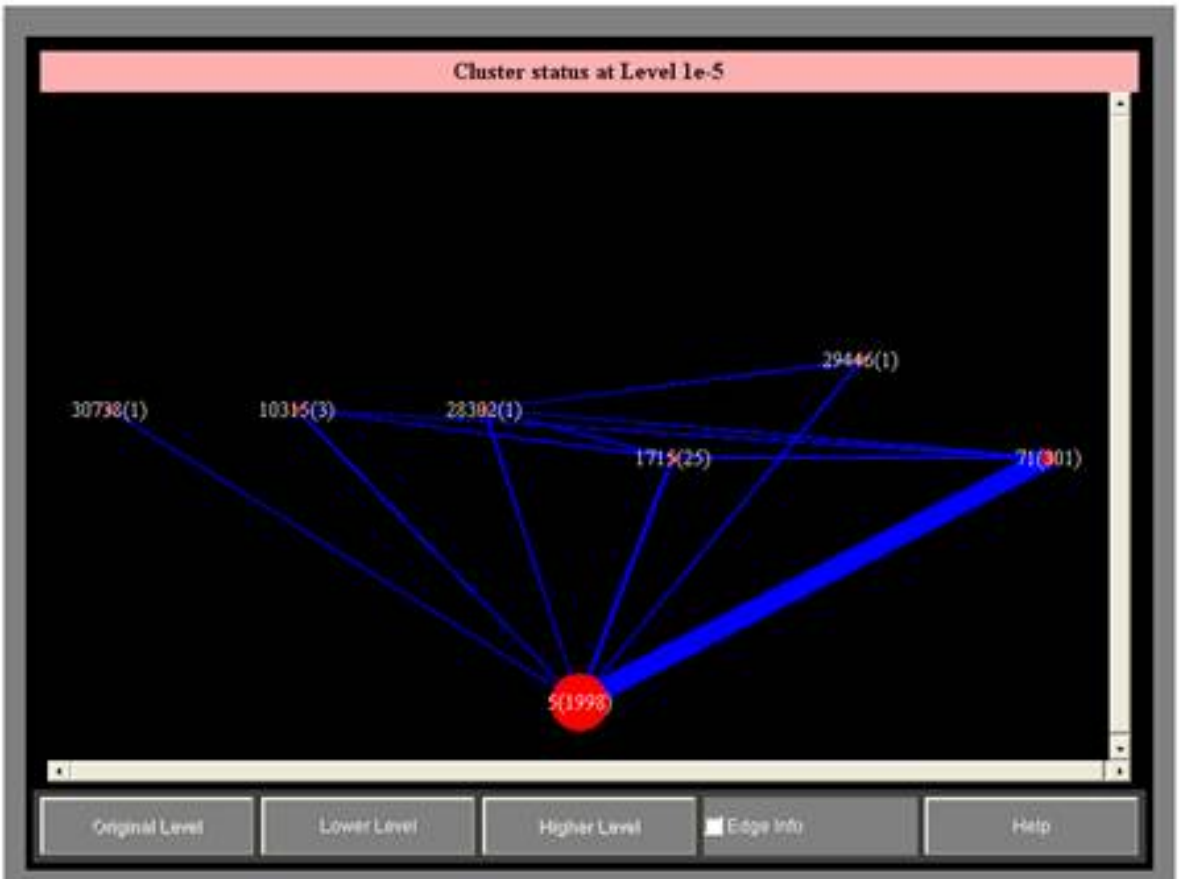


Figure 11: ProtoMap Cluster Relationships Diagram

4.2. ProtoNet

ProtoNet (http://www.protonet.cs.huji.ac.il, Sasson *et al.*, 2003) provides a fully automated hierarchical clustering of the SwissProt proteins. ProtoNet implements an average-link hierarchical agglomerative clustering algorithm. The novelty of ProtoNet is the use of several averaging methods. The averaging methods provided are arithmetic mean, geometric mean, and harmonic mean.

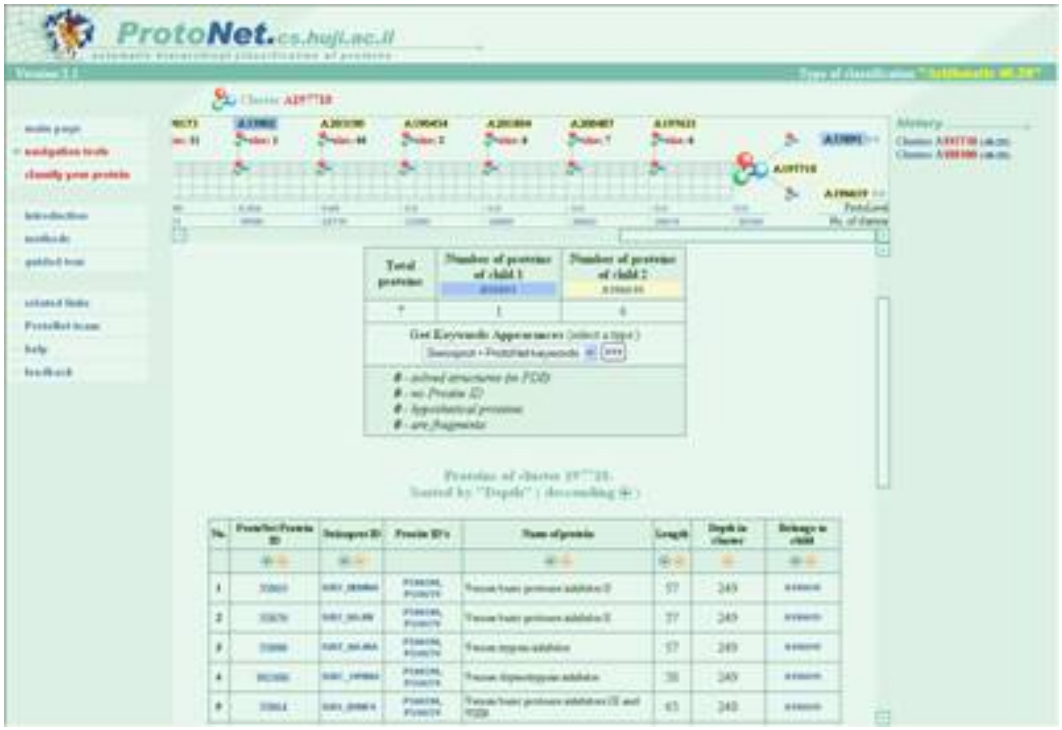


Figure 12: ProtoNet Cluster Card

ProtoNet offers a variety of methods to traverse through clusters and analyze their contents. The hierarchical clustering is fully traversable, thus providing users with varying granularity, based on individual requirements.

The ProtoNet Web-based interface provides detailed information in the cluster level, as shown in Figure 12. The top portion of the cluster details shown a graphical representation of the sequence of

merges taking place for the creation of this cluster and subsequent clusters.
Each protein recorded in the ProtoNet database is associated with detailed taxonomical representation as well as a variety of keywords (including InterPro and PROSITE entries). Protonet also provides motif and domain information taken from PFAM, Prints, ProDom, Prosite, and SMART, when such information is available. For example see Figure 13 below.
ProtoNet provides a unique feature by which users may classify their own proteins. In contrast to other systems, where users may classify only a single sequence at a time, ProtoNet stores sequences classified by users, and specifically, multiple sequences may be classified concurrently.
ProtoNet provide statistical measures for the purity and the sensitivity for each merging in the process as compared with other type of classifications.
ProtoNet version 2.1 dated December 2002 covers 114,000 proteins from SwissProt release 40.28.
The hierarchical nature of ProtoNet allows analyzing the protein space at different level of granularity. At a level of 10,000 clusters (of which ~2500 are singletons), very good correspondence can be observed between ProtoNet clusters and the InterPro entries.
Distances in the graphs of ProtoNet and ProtoMap are used for target selection in the scope of Structural Genomics initiatives (Linial and Yona, 2000).

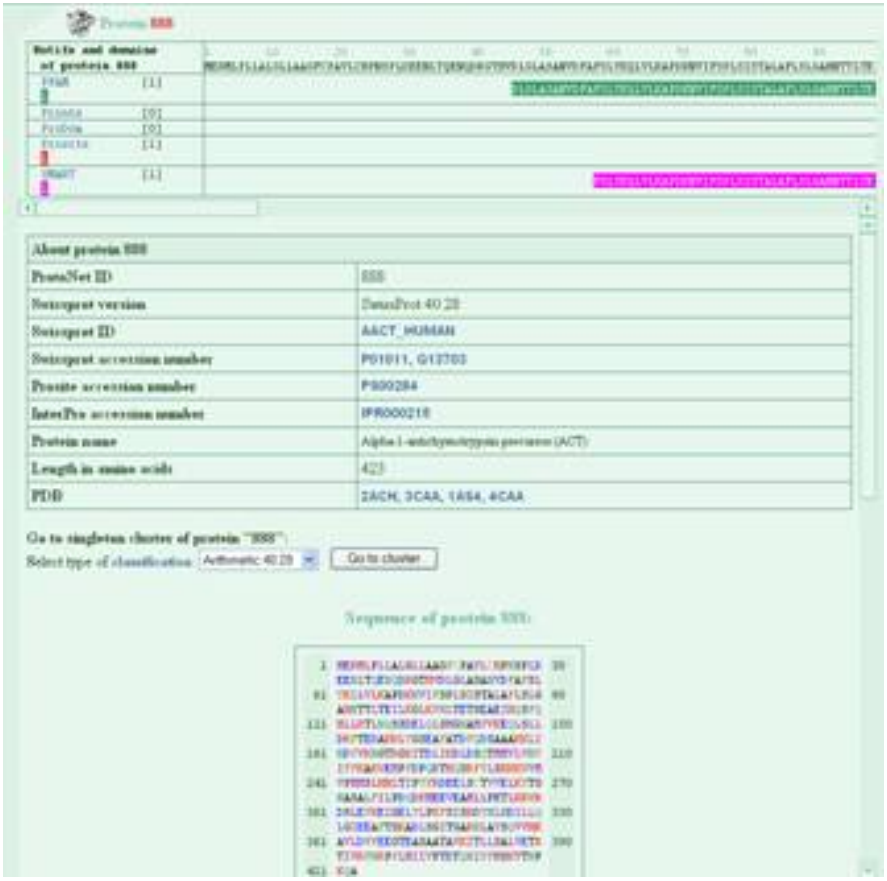


Figure 13: ProtoNet Protein Card with Domain and Motif Layout

4.3. PIR-ALN

PIR-ALN (http://www-nrbf.georgetown.edu/pirwww/dbinfo/piraln.html, Srinivasarao *et al., 1999)* provides a simple classification of protein sequences based on sequence alignment. Classification is into three categories. Families include sequences which are at least 45% identical. Superfamilies are multiple alignments containing sequences from different families. Homology domain alignments contain homologous subsequences (segments) from different proteins. The PIR-ALN version release in December 2000 contains 3508 alignments includes 994 superfamilies and 386 homology domain alignments.

4.4. Systers

Systers (http://systers.molgen.mpg.de, Krause *et al.*, 2002) is classification of the protein space based on single linkage agglomerative clustering. Single-linkage clustering (in contrast to the average-linkage clustering used in ProtoMap and ProtoNet) defined the similarity between two clusters as the highest similarity between pairs from the two clusters.
Systers clustering is based on Smith-Waterman all-against-all sequence comparison. To avoid problems resulting from asymmetric alignment scores, a symmetric E-value is computed for each pair with significant similarity.

Figure 14: Systers Sample Cluster Page

The agglomerative clustering generated a single linkage tree. This tree is the result of successive merging of any two clusters which are associated with at least one significant similarity. From this tree, superfamilies are derived, and this is done automatically without user intervention (and specifically without the need to define an arbitrary cut-off value), in an attempt to choose the optimal level of detail. Once superfamilies are identified, they are broken down into 'family' clusters. This is done by reviewing the similarities graph for proteins within the superfamily, and looking at the connected components generated by omitting all edged below a certain threshold. Systers clusters are annotated with Pfam domains, and links to PROSITE and PDB where applicable. Figure 14 above shows a sample Systers cluster page.

Release 3 of Systers includes 290,811 non-redundant sequences (as well as annotations for 583,448 redundant sequences) taken from SwissProt, TREMBL, PIR, FlyBase, Wormpep 20, and MIPS Yeast protein translations. These proteins are sorted into 82,450 disjoint clusters. Of these 55,182 are singleton (i.e. single sequence) clusters, and the remaining 235,629 sequences are contained in 27,268 clusters ranging in size from 2 to 8,821 non-redundant sequences per cluster.

4.5. ProClust

ProClust (http://promoter.mi.uni-koeln.de/~proclust, Pipenbacher *et al.,* 2002) is a protein clustering derived using a graph-based approach. A graph of proteins is constructed where edge-weights are based on Smith-Waterman similarity scores scaled with respect to self-similarity. After removing all edges with insignificant similarities (based on some arbitrary threshold), clustering proceeds by seeking maximal strongly connected components in the graph. In other words, this means that every two proteins in the cluster are connected in the graph in both ways (the graph here

is directed, and the requirement is for connectivity in both directions).
ProClust covers SwissProt proteins in release 39 (June 2000).

4.6. CluSTr

CluSTr (http://www.ebi.ac.uk/clustr, Kriventseva *et al., 2001)* provides a classification of SwissProt/TREMBL proteins. The clustering is based on single-linkage hierarchical clustering (similar to Systers), where the underlying scores are based on Smith-Waterman. The scores are processed to generate a 'Z-score' representing the statistical significance of the similarity. Unlike E-values generated by BLAST, the Z-score is independent of the sequence database and thus facilitates easier updating of the database.

4.7. Picasso

Picasso (http://www.ebi.ac.uk/picasso, Heger and Holm, 2001) is a global classification of protein sequences. The classification is based on generating a minimal set of family profiles covering all known protein sequences.
Picasso uses a clustering algorithm similar to hierarchical clustering. Clusters are merged based on a threshold E-value, and the threshold is gradually increased.
Picasso has been successfully used for selecting targets for mapping 3D structures of proteins.

4.8. Tribe-MCL

TribeMCL (http://www.ebi.ac.uk/research/cgg/tribe, Enright *et al.*, 2002) is a protein clustering software package. TribeMCL takes as input the results of a BLAST calculation, and output a clustering of proteins into families.
TribeMCL uses a novel clustering method called Markov Clustering. This clustering algorithm addresses the difficulties in protein clustering such as multi-domain proteins, protein fragments, and promiscuous domains.
The TribeMCL software is available for download, but there is no Web-based classification of common databases into families.

5. Phylogenetic classification

COGS (http://www.ncbi.nlm.nih.gov/COG, Tatusov *et al.*, 2001) provides a clustering of proteins derived from 43 complete genomes. COGs applies single linkage hierarchical agglomerative clustering to get clustering of orthologous proteins or orthologous sets of paralogs. The former

refers to genes from different special which have evolved from the same protein, and the later to genes from the same genome which are related by duplication.
Each cluster consists of at least three species. The clustering process starts by forming a minimal COG with three elements, and then proceeds by merging COGs sharing an edge. The COGs clustering algorithm has the capability of splitting COGs which were incorrectly merged.

6. Classification based on protein structure

Full-sequence analysis helps in classifying proteins with known sequences. However, the common perception is that protein function stems from protein structure and not so much from sequence. This gives rise to the idea of classifying proteins based on structure. The drawback of structure-based clustering is that the number of sequences for which the structure is available (or the structure is 'solved') is relatively small, due to the complexity of obtaining high-resolution structural information.
Structure based clustering takes into account the three-dimensional structure of a protein. Several different algorithms and software packages are available for measuring the similarity between two protein structures. Examples are CE (Shindyalov and Bourne, 1998), Dali (Holm and Sander, 1998), PrISM (Yang and Honig, 2000), VAST (Gibrat *et al.*, 1996) and STRUCTAL (Orengo *et al.*, 1999). A full description of these methods is beyond the scope of this survey.

6.1. SCOP

SCOP (http://scop.berkeley.edu, Lo Conte *et al.*, 2002) is a structural classification of proteins into a four-level hierarchy.

- **Family** – Proteins with significant sequence similarity (typically 30% or greater identity), and with clear evolutionary relationship.
- **Superfamily** – Proteins with low sequence similarity, but with structural and functional features suggesting a common evolutionary origin.
- **Fold** – Superfamilies with major structural similarity.
- **Class** – High level classification (e.g. All-alpha, All-beta, Alpha/Beta, Alpha+Beta, Membrane proteins, etc.)

SCOP is organized as a tree, and on top of all classes there is the SCOP "root". The SCOP classification is based on manual analysis by experts.

6.2. CATH

CATH (http://www.biochem.ucl.ac.uk/dbbrowser/cath, Orengo *et al.*, 1999) is a structural classification into a different 4-level hierarchy:

- Homologous superfamily (H level) – Sequences with high similarity. Several conditions are defined combining sequence and structure (e.g. at least 35% sequence identity and at least 60% structural similarity).
- Topology (T level) – Structure comparison is used to group together protein structures into fold families.
- Architecture (A level) – Structures are grouped based on the overall shape of the domain structures.
- Class – Structures are determined according to secondary structure composition, similar to SCOP classes.

The name of the system is based on the level names (Class, Architecture, Topology, Homology). Assignment of structures to families and topologies is automatic, based on similarity. At the architecture and class levels, manual considerations are included into the classification. Figure 15 shows a sample page (for domain 1cuk03) in CATH.

Figure 15: CATH Domain Page

6.3. FSSP

FSSP (http://www.ebi.ac.uk/dali/fssp, Holm and Sander, 1996) provides fold classification using structure-structure alignment of proteins. The classification is based on an exhaustive all-against-all structure comparison using Dali (Holm and Sander, 1998) structure comparison.
FSSP provides a Web-based interface which facilitates 3D superimposition and multiple alignments of structures.
The recent version of FSSP (dated June 2002) contains 3242 sequence families representing 30,624 protein structures (taken from PDB).

6.4. BioSpace

BioSpace (http://biospace.cornell.edu, Yona *et al.*, 2000b*)* is a protein classification system combining sequence and structure information.
BioSpace uses a novel clustering algorithm which gives preference to structural similarity over sequence similarity. This is based on the common perception that structural similarity implies strong functional similarity.
The BioSpace Web interface allows browsing the various clusters, providing multiple sequences alignment and 3D models where available (for example, see Figure 16).

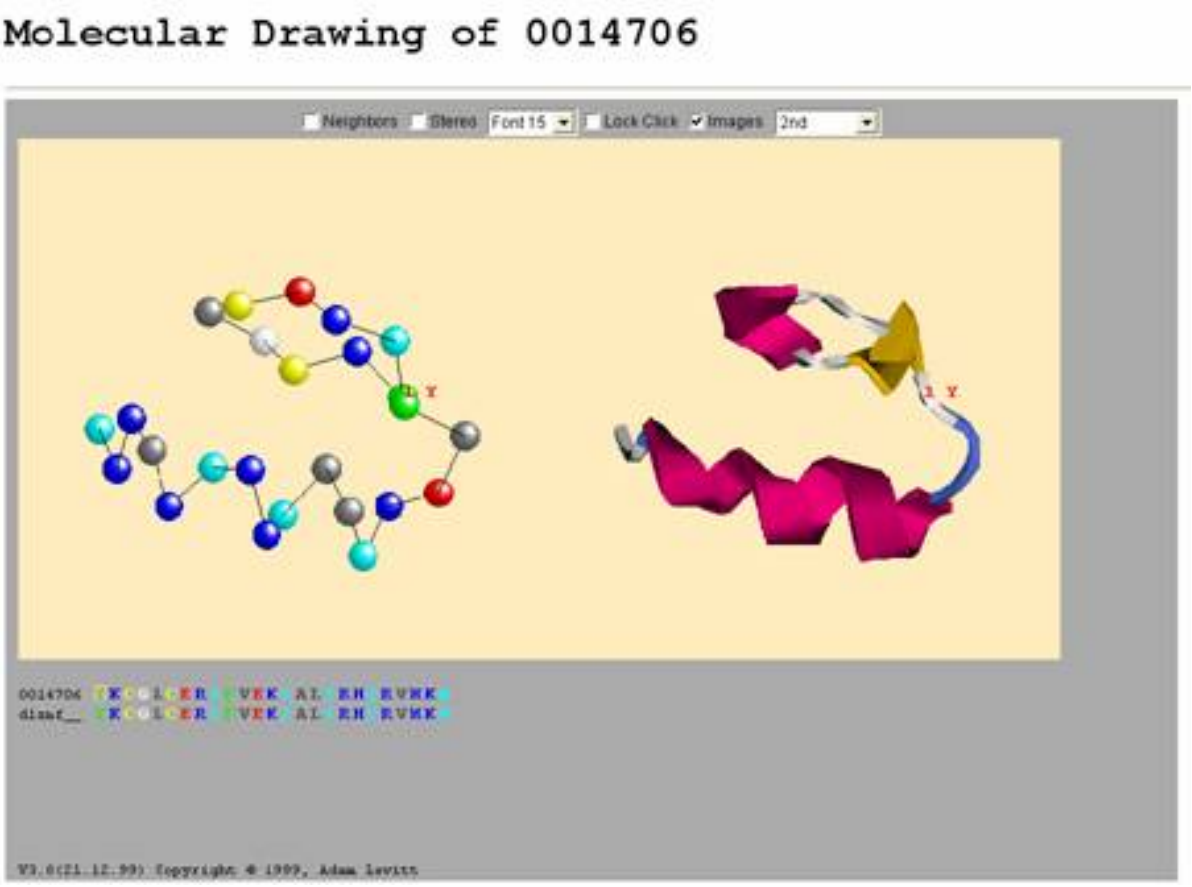


Figure 16: BioSpace Sample 3D Structure Drawing

Once particularly appealing aspect of BioSpace is the fact is facilitates target selection for mapping 3D structures of proteins (Linial and Yona, 2000).

BioSpace revision 1 (dated January, 2000) covers most protein sequences known at the time of release. While BioSpace offers a promising approach to protein classification, it has not been updated recently.

6.5. Superfamily

Superfamily (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY, Gough and Chothia, 2002) is an HMM library which models structure. It focuses on the so called ‘superfamily’ level. In this level, all proteins with any structural evidence for a common evolutionary ancestor are grouped together. SUPERFAMILY uses a library of 1073 SCOP families (taken from SCOP 1.59), each represented with a group of HMMs. HMMs are generated using the SAM (Karchin and Highey, 1998) software. Superfamily provides sequence searching, alignments, and genome assignments.

7. Aggregated systems

Each classification system has its own strengths as well as its own weaknesses. In order to leverage on the strength of several systems, it is possible to combine several systems to generate one ‘aggregated’ classification. Such a classification is appealing in the sense that it might be more ‘valid’. Several attempts were made to combine multiple classifications.

7.1. InterPro

InterPro (http://www.ebi.ac.uk/interpro, Apweiler *et al.*, 2001) provides a unified domain database, based on Pfam, PRINTS, PROSITE, ProDom, SMART, and TIGRFAMs.

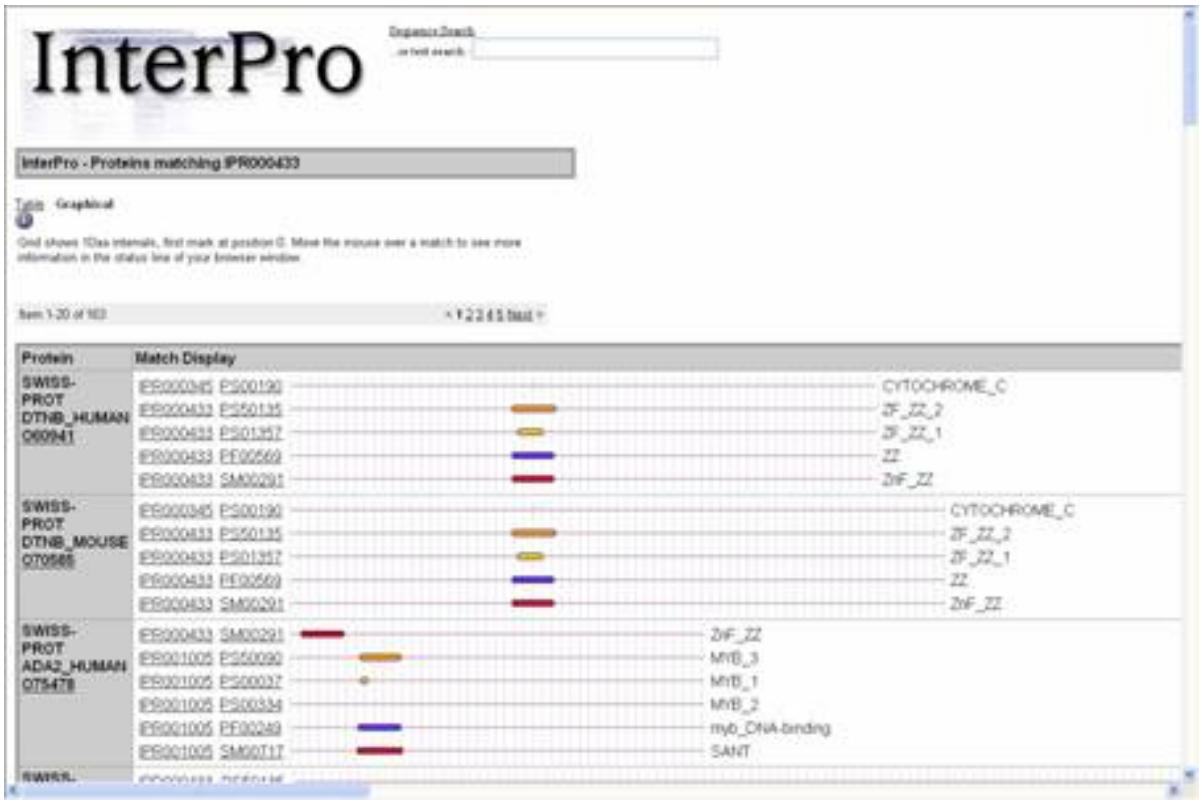


Figure 17: InterPro Graphical Representation of Domain

InterPro version 5.3, dated November 2002, includes 6725 entries, representing 1,453 domains and

5121 families. The InterPro Web-site provides for each InterPro entry (corresponding to a domain or family) the list of member proteins, literature references, references to the underlying databases, and an abstract of the biological context for the entry. A graphical representation of domains is available, as shown in Figure 17.
InterPro is considered the 'state of the art' in terms of protein classification. The classification provided by IntePro is based on a carefully selected set of rules and considerations, which balance the different underlying classifications.
InterPro versions are released in conjunction with SwissProt and TrEMBL releases on a routine basis, and thus it is continuously updated. This stands in contrast to other domain-based systems which are only infrequently updated.

7.2. MetaFam

MetaFam (http://metafam.ahc.umn.edu, Silverstein *et al.*, 2001) is a protein clustering system obtained by seeking maximal agreement between several clustering systems, including BLOCKS, DOMO, Pfam, PIR-ALN, PRINTS, ProDom, PROSITE, Protomap, Systers, and SBASE. Metafam automatically creates supersets of overlapping families. It covers a non-redundant protein set from SwissProt and PIR.

7.3. iProClass

iProClass (http://pir.georgetown.edu/iproclass, Wu *et al.*, 2001) is an integrated database linking PIR-ALN, Prosite, Pfam, and BLOCKS. It contains a set of non-redundant SwissProt and PIR proteins, classified into 28,000 families. iProClass is an extension of an older system called ProClass which provided classification based on PROSITE patterns and PIR superfamlies. iProClass provides a Web-based interface which allows searching by protein name or sequence. Release 2.12, dated December 2002 contains 877,360 entries. The database consists of non-redundant PIR and SwissProt/TREMBL sequences organized into more than 36,200 PIR superfamilies, 3880 domains, and 1300 motifs.

7.4. CDD

CDD Conserved Domain Database (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml, Marchler-Bauer *et al.*, 2002) is a database of domains consolidating SMART, Pfam and some NCBI contributions (e.g. from COGs).
CDD uses a set of specially constructed score matrices prepared for each conserved domain (using a multiple alignment), and relies on BLAST for searches.

8. Summary

The field of protein classification provides a varied landscape. Diversely different approaches are used to generate results which are quite impressive considering the complexity of the problem. There are numerous protein classification systems, usually applied only to a single family or superfamily of proteins, which were excluded from this survey. Detailed accounts of such systems are available in each on Nucleic Acid Research database issues (published every January).
While immense work was done by numerous different research groups, the 'holy grail' of protein classification has yet to be found. In practice, the best approach to protein classification is combining the use of several systems, preferably trying to leverage the advantages of each system. An important point to keep track of is the frequency in which classifications are updated.
The primary objective of protein classification efforts is finding a mechanized shortcut in the laborious task of functional annotation and functional prediction. However, a very important byproduct of those efforts is global analyses of the protein space, both sequence and structure. The classification of protein families has become an essential building block in genomic comparative studies, in tracing evolutionary processes and in systematic methods for prediction.
The current challenges in the field of protein classification are providing firm indications of the validity of classifications (possibly via cross-validation with manually maintained classifications), as well as consolidating the different considerations (e.g. structure and sequence). On a more technical note, the issue of correctly dealing with multi-domain proteins will become more crucial as complete genomes of higher organisms become available. Another technical issue is tacking the very different rate of evolution in different protein families. Current method usually use different score matrices for different purposes (e.g. BLOSUM62 for homology searches, and BLOSUM45 for 'remote homologue' searches).
We witness two distinct directions in which this field is currently progressing. On one hand, there are systems and algorithms specializing in unique genomes, systems, and protein families. Such systems are typically limited in their scope of application, but are highly reliable in the context for which they were designed. On the other hand, there is a trend of combining non-sequence based information sources into the sequence and structure data. Examples of such 'external' information sources are DNA profiling, protein-protein interaction, and phylogenetic profile. The most ambitious effort undertaken to map all protein-related information is the Gene Ontology (GO) database (Ashburner *et al.*, 2000). The objective of the GO consortium is to provide an all-encompassing dynamic controlled vocabulary of all known biological processes, molecular functions, and cellular components. Going forward, GO holds tremendous potential both for cross-validating classification systems and for improving classification systems.

Similar to other fields in bioinformatics, the field of protein interaction is still in its infancy. Only time will tell how the limitations of classification based on sequence and structure can be overcome in the quest for automatically identifying protein function and understanding protein machines in the cell.

9. References

Altschul, S. F., Madden, T. L., Schäffer, 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.

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.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25-29.

Attwood, T. K., Blythe, M. J., Flower, D. R., Gaulton, A., Mabey, J. E., Maudling, N., McGregor, L., Mitchell, A. L., Moulton, G., Paine, K., and Scordis, P. (2002). PRINTS and PRINTS-S shed light on protein ancestry. Nucleic Acids Res *30*, 239-241.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000). The Pfam protein families database. Nucleic Acids Res *28*, 263-266.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. (2000). The Protein Data Bank and the challenge of structural genomics. Nat. Struct. Biol. *7 Suppl*, 957-959.

Corpet, F., Servant, F., Gouzy, J., and Kahn, D. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res *28*, 267-269.

Dayhoff, M.O., Schwartz, R. M. and Orcutt, B.C. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5, 345-352.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res *30*, 1575-1584.

Falquet L., Pagni M., Bucher P., Hulo N., Sigrist C. J., Hofmann K. and Bairoch A. (2002). The PROSITE database, its status in 2002. Nucleic Acids Res. *30*, 235-238.

Gibrat, J-F., Madej, T., Bryant, S.H. (1996) Surprising similarities in structure comparison. Current Opinion in Structural Biology *6*: 377-385.

Gough, J., and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res. *30*, 268-272.

Gracy, J., and Argos, P. (1998a). DOMO: a new database of aligned protein domains. Trends Biochem Sci *23*, 495-497.

Gracy, J., and Argos, P. (1998b). Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities. Bioinformatics *14*, 174-187.

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T., and White, O. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res *29*, 41-43.

Henikoff, S. and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Academy Science *89*, 915-919.

Henikoff, J. G., Greene, E. A., Pietrokovski, S., and Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. Nucleic Acids Res. *28*, 228-230.

Holm, L., and Sander, C. (1996). The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Res *24*, 206-209.

Holm, L., and Sander, C. (1998). Touring protein fold space with Dali/FSSP. Nucleic Acids Res *26*, 316-319.

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

Heger, A., and Holm, L. (2001). Picasso: generating a covering set of protein family profiles. Bioinformatics *17*, 272-279.

Karchin, R., and Hughey, R. (1998). Weighting hidden Markov models for maximum discrimination. Bioinformatics *14*, 772-782.

Krause, A., Haas, S. A., Coward, E., and Vingron, M. (2002). SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. Nucleic Acids Res *30*, 299-300.

Kriventseva, E. V., Fleischmann, W., Zdobnov, E. M., and Apweiler, R. (2001). CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. Nucleic Acids Res *29*, 33-36.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR *163*, 845-848.

Linial, M. and Yona, G. (2000). Methodologies for target selection in structural genomics. Progress in Biophysical and Molecular Biology *73*, 297-320

Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res. *30*, 264-267.

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y., and Bryant, S. H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res *30*, 281-283.

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.

Orengo, C. A., Pearl, F. M., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L., and Thornton, J. M. (1999). The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res. *27*, 275-279.

Pearson W.R (1990). Rapid and Sentive Sequence Comparison with PASTP and FASTA. Methods Enzymol. 183, 63-98.

Pipenbacher, P., Schliep, A., Schneckener, S., Schonhuth, A., Schomburg, D., and Schrader, R. (2002). ProClust: improved clustering of protein sequences with an extended graph-based approach. Bioinformatics *18 Suppl 2*, S182-191.

Portugaly, E., Kifer, I., and Linial, M. (2002). Selecting targets for structural determination by navigating in a graph of protein families. Bioinformatics *18*, 899-907.

Sasson O, Vaaknin A, Fleischer H, Portugaly E, Bilu Y, Linial N, Linial M. (2003). ProtoNet: hierarchical classification of the protein space. Nucleic Acids Res. *31*, 348-352.

Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P., and Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res *28*, 231-234.

Shindyalov, I.N., Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering *11* (9) 739-747.

Silverstein, K. A., Shoop, E., Johnson, J. E., and Retzel, E. F. (2001). MetaFam: a unified classification of protein

families. I. Overview and statistics. Bioinformatics *17*, 249-261.

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

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. *26*, 320-322.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res *29*, 22-28.

Srinivasarao, G. Y., Yeh, L. S., Marzec, C. R., Orcutt, B. C., and Barker, W. C. (1999). PIR-ALN: a database of protein sequence alignments. Bioinformatics *15*, 382-390.

Vlahovicek, K., Murvai, J., Barta, E., and Pongor, S. (2002). The SBASE protein domain library, release 9.0: an online resource for protein domain identification. Nucleic Acids Res *30*, 273-275.

Wu, C. H., Xiao, C., Hou, Z., Huang, H., and Barker, W. C. (2001). iProClass: an integrated, comprehensive and annotated protein classification database. Nucleic Acids Res. *29*, 52-54.

Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structure alignment and quantitative measure for protein structural distance. J Mol Biol. *301*(3), 665-678.

Yona, G., Linial, N., and Linial, M. (2000a). ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Res. *28*, 49-55.

Yona, G. and Levitt, M. (2000b) A unified sequence-structure classification of protein sequences: combining sequence and structure in a map of protein space. The proceedings of RECOMB, 308-317.