New Approaches to Facilitate Genome Analysis

Philip Scordis

Sequence Analysis Group

Biomolecular Structure and Modelling Unit

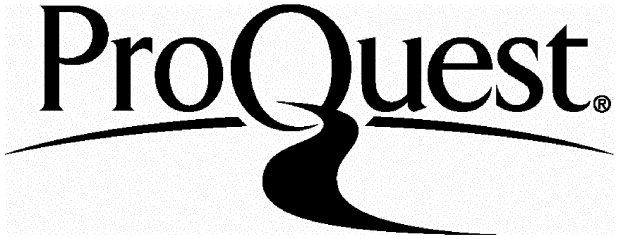
Department of Biochemistry and Molecular Biology

University College London

A thesis submitted to the University of London

for the degree of Doctor of Philosophy

September 2000



Abstract

In this era of concerted genome sequencing efforts, biological sequence information is abundant. With many prokaryotic and simple eukaryotic genomes completed, and with the genomes of more complex organisms nearing completion, the bioinformatics community, those charged with the interpretation of these data, are becoming concerned with the efficacy of current analysis tools. One step towards a more complete understanding of biology at the molecular level is the unambiguous functional assignment of every newly sequenced protein. The sheer scale of this problem precludes the conventional process of biochemically determining function for every example. Rather we must rely on demonstrating similarity to previously characterised proteins via computational methods, which can then be used to infer homology and hence structural and functional relationships. Our ability to do this with any measure of reliability unfortunately diminishes as the pools of experimentally determined sequence data become muddied with sequences that are themselves characterised with "in silico" annotation.

Part of the problem stems from the complexity of modelling biology in general, and of evolution in particular. For example, once similarity has been identified between sequences, in order to assign a common function it is important to identify whether the inferred homologous relationship has an orthologous or paralogous origin, which currently cannot be done computationally. The modularity of proteins also poses problems for automatic annotation, as similar domains may occur in proteins with very different functions. Once accepted into the sequence databases, incorrect functional assignments become available for mass propagation and the consequences of incorporating those errors in further "in silico" experiments are potentially catastrophic. One solution to this problem is to collate families of proteins with demonstrable homologous relationships, derive a pattern that represents the essence of those relationships, and use this as a signature to trawl for similarity in the sequence databases. This approach not only provides a more sensitive model of evolution, but also allows annotation from all members of the family to contribute to any assignments made.

This thesis describes the development of a new search method (FingerPRINTScan) that exploits the familial models in the PRINTS database to provide more powerful diagnosis of evolutionary relationships. FingerPRINTScan is both selective and sensitive, allowing both precise identification of super-family, family and sub-family relationships, and the detection of more distant ones. Illustrations of the diagnostic performance of the method are given with respect to the haemoglobin and transfer RNA synthetase families, and whole genome data.

FingerPRINTScan has become widely used in the biological community, e.g. as the primary search interface to PRINTS via a dedicated web site at the university of Manchester, and as one of the search components of InterPro at the European Bioinformatics Institute (EBI). Furthermore, it is currently responsible for facilitating the use of PRINTS in a number of significant annotation roles, such as the automatic annotation of TrEMBL at the EBI, and as part of the computational suite used to annotate the *Drosophila melanogaster* genome at Celera Genomics.

Acknowledgements

I thank my supervisor Terri Attwood for all of the support, some of the criticism and all the time and effort she had to expend to get me through this. I am grateful to my industrial supervisor Darren Flower and I acknowledge my sponsors Astra Charnwood for funding this research. I must also thank my present employers for their patience.

Thanks also go to my good friends and colleagues at UCL and Manchester: Denise Henriques, Duncan Milburn, Adrian Shepherd, Jane Mabey, Julian Selley, Will Wright, who have been with me most of the way. And some who were only around for part of the ride but who certainly helped keep me sane, Karen Eilbeck, Crispin Miller, Harriet Watkin and Andrea Edwards.

Heartfelt thanks must go to Vicki Kitchener, Jennifer and Michael Scordis, because without their continual support and optimism I would not have survived.

Contents

Abstract 2

Acknowledgements 3

1 Introduction 19

1.1 Biosequences 20

1.2 Databases 24

1.3 Sequence Analysis 24

2 Primary databases 27

2.1 Biological Sequences 28

2.2 Nucleic Acid Databases 29

2.2.1 Other nucleic acid sequence resources 31

2.3 Protein Sequence Databases 32

2.3.1 PSD 32

2.3.2 SWISS-PROT 33

2.3.3 Summary 35

2.4 Pairwise Sequence Analysis 37

2.4.1 Sequence similarity 38

2.4.1.1 Identity 38

2.4.1.2 Amino acid side chain similarities 41

2.4.2 Scoring the alignment 45

2.4.2.1 Counting and scoring identity 45

2.4.2.2 Counting and scoring similarity 48

2.4.3 Identifying the optimal alignment 52

2.4.3.1 Global alignment algorithms 54

2.4.3.2 Local alignment algorithms 58

2.4.4 Assessing the significance of the alignment 61

2.4.4.1 Scoring 61

2.4.4.2 Probability values 63

2.4.4.3 Adjusting probabilities for database searches. . . . 66

2.5 Problems of pairwise sequence analysis 67

3 Secondary Databases **71**

3.1 Gene families 72

3.2 The Multiple Sequence Alignment 75

3.2.1 Creating a multiple sequence alignment. 76

3.3 Multiple Sequence Analysis 78

3.3.1 Scanning a sequence 80

3.3.1.1 The sliding window 80

3.3.2 The Motif 81

3.3.3 Selecting a motif 82

3.3.4 Encoding a motif 83

3.3.4.1 The regular expression 83
3.3.4.2 The frequency matrix. 85
3.3.4.3 The 'profile' motif. 86
3.3.5 Motif Scoring 88
3.3.5.1 Scoring a sequence - the frequency matrix 90
3.3.5.2 Scoring a sequence - the 'profile' motif 91
3.3.6 Multiple Motifs 92
3.3.7 Whole Alignments 93
3.3.7.1 Profile Methodology 93
3.3.7.2 Profile-HMMs 94
3.3.8 Summary 95
3.4 Pattern and Family Databases 96
3.4.1 Pattern Databases 96
3.4.1.1 PROSITE 96
3.4.1.2 PRINTS 97
3.4.1.3 Blocks 98
3.4.1.4 Meta-MEME 98
3.4.1.5 IDENTIFY 99
3.4.1.6 Profiles 101
3.4.1.7 Pfam 102
3.4.2 A summary of secondary databases. 103
3.4.3 A composite pattern database - InterPro 108
3.4.4 Family or clustered sequence databases 108
3.4.5 PSI-BLAST 110

4 PRINTS 112

4.1 Introduction 113

4.2 The development of fingerprints 113

4.2.1 Alignment 114

4.2.2 Motif Extraction 116

4.2.3 Iteration 116

4.2.3.1 Scanning and matching sequences 116

4.2.3.2 Iterating 120

4.2.4 Annotating 122

4.2.5 Analysis of the fingerprinting method 122

4.2.5.1 Some 'true' sequences fail to match *all* of the motifs in a fingerprint 124

4.2.5.2 Some 'false' sequences match all motifs of a fingerprint 128

4.2.5.3 It is difficult to distinguish false from true 131

4.3 Creating fingerprints for the PRINTS database 133

4.3.1 Haemoglobin 133

4.3.2 tRNA synthetases 134

4.4 Using fingerprints 137

5 Methods 141

5.1 Aim 142

5.2 The development of a new search tool for PRINTS 142

5.2.1 The scanning process 143

5.2.1.1 The sliding window 143
5.2.1.2 The frequency matrix. 144
5.2.1.3 The profile matrix. 145
5.2.2 Scoring a sequence 146
5.2.2.1 Scoring a sequence - the frequency matrix 146
5.2.2.2 Scoring a sequence (the 'profile' matrix) 148
5.2.3 Dealing with matches 149
5.2.3.1 Identifying fingerprint context 152
5.2.4 The PathFinder algorithm 153
5.2.4.1 Finding the longest path 153
5.2.4.2 Partially matching fingerprints 156
5.2.4.3 Motif positions 157
5.2.4.4 Inter-motif distances 158
5.2.4.5 Summary 159
5.2.5 Scoring fingerprints 161
5.2.6 Summary 163
5.3 Interfaces 164
5.3.1 FPScan 164
5.3.2 GRAPHScan 166
5.3.3 MULScan 172
5.3.3.1 FPScan Multiple Sequence Analysis (FPScanMSAn) 172
5.4 Summary 177

6 Results and Applications **178**

6.1 Introduction 179

6.2 A comparison of scoring schemes 180

6.2.1 Summary 191

6.3 Multiple motifs 191

6.3.1 Summary 196

6.4 Genome Analysis 198

6.4.1 Summary 199

6.5 Sensitivity 200

6.5.1 Summary 203

6.6 Applications 203

6.6.1 TrEMBL 204

6.6.2 InterPro 205

6.6.3 Other applications 209

7 Discussion and Conclusions **211**

List of abbreviations **215**

Bibliography **217**

List of Figures

2.1 The growth of GenBank. 30

2.2 An example entry from the PIR international Protein Sequence Database. 34

2.3 An example entry from the SWISS-PROT protein sequence database. Highlighted regions indicate keywords, and an example of ambiguity in the naming of the family, to which the protein belongs. 36

2.4 An alignment of bactrian camel and human β-haemoglobin. 39

2.5 A pairwise alignment of human α and β-haemoglobin. 40

2.6 A pairwise alignment of human α and β-haemoglobin showing identities between sequences. 40

2.7 A pairwise alignment of human α- and β-haemoglobin showing similarities between sequences. 42

2.8 A classification of the collective properties of the amino acids. 42

2.9 A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physico-chemical properties. 43

2.10 A coloured alignment of human α- and β-haemoglobin. 44

2.11 A coloured alignment of human α- and β-haemoglobin that emphasises commonality. 45

2.12 The identity matrix. 46

2.13 An aligned block of residues, and the pairwise substitutions that are observed between its constituent sequences. 50

2.14 An aligned block of residues clustered at 80%, and the effect on the pairwise substitutions that are observed. 51

2.15 The dot-matrix 53

2.16 The Needleman-Wunsch global alignment calculation. 56

2.17 The Needleman-Wunsch alignment matrix 57

2.18 The extreme value distribution of MSS scores (S). The score of an observed MSS is represented by x. 64

3.1 Long exposure photography is analogous to multiple sequence alignment. 74

3.2 An alignment of the most modern sequences of each lineage. 76

3.3 A graphical representation of the alignment process. 77

3.4 The acylphosphatase family unaligned and aligned (PRINTS: ACYLPHPHTASE). 78

3.5 The ancestral relationships of three sequences (A, B and C) 79

3.6 An MSA can lead to the identification of functionally preserved motifs 79

3.7 Scanning a sequence with a fixed window. 81

3.8 A motif, including notation that describes its generalisation. 82

3.9 A simple motif, and its RE. 84

3.10 A simple motif with a single divergent position, and its RE. 84

3.11 A frequency matrix, and its normalised form, based on the example in figure 3.8. Frequencies for the occurrence of each residue in each column are normalised for the number of sequences. 87

3.12 An example motif, frequency matrix and profile matrix (derived) from the PRINTS database. 89

3.13 Scoring the query sequence "SACDEKGHI" against the normalised frequency matrix of figure 3.11. 90

3.14 The Profile-HMM is characterised by its match, delete and insert states and the allowed transitions between them. 95

3.15 Substitution groups are sets of amino acids found to occur together in columns of aligned sequences. Arranging these hierarchically provides an opportunity to describe relationships between the residues in the groups, and provides a clear representation of each of the overlapping sets. 100

3.16 An RE forced to represent a divergent relationship may ultimately be too unselective; however, by defining two, more specific, REs this region can be described more effectively. 101

3.17 The seed alignment for the generation of a Profile-HMM of the α-haemoglobin family. The figure shows a section (the first 50 residues from each sequence) of the file used as input for the hmmbuild program (from the HMMer suite). 106

3.18 Searching a SWISS-PROT/TrEMBL composite database of sequences (version 37_9) with the Profile-HMM generated from the alignment shown in figure 3.17, produced the following result. 107

4.1 An overview of the fingerprinting process. 114

4.2 The selection of sequences is a critical step in the description of an alignment. 115

4.3 The fingerprint ALPHAHAEM was derived from this alignment of α haemoglobin sequences. The boxes highlight the sub-sequences extracted to become the initial set of motifs. (Motif 1 is shown, in full, in figure 4.4) 116

4.4 Motif 1 from the ALPHAHAEM fingerprint (figure 4.3). 117

4.5 N-single scoring of motif matching sub-sequences. 118

4.6 A hit-list of matches to a single motif. 119

4.7 The scanning process can identify new sequences, which can be used to augment the original motifs. 121

4.8 Part of the ALPHAHAEM fingerprint, including annotation and external database links 123

4.9 Augmenting the original motifs with new sequences can potentially introduce a bias into the process. 125

4.10 If a number of sequences fail to match all motifs after a number of iterations, this can indicate a poorly chosen motif. 127

4.11 The simple examples shown in figures 4.9 and 4.10 may occur in combination, and can result in the poor representation of more than one outlier group. 128

4.12 An illustration of two alignments (families A and B) is shown, each containing six motifs. The shared motifs (A:5 and 6, B:1 and 2) fall in a common domain. Selecting all six motifs, when describing either A or B will result in partial identification of the other. Selecting only the shared motifs, independently, from either alignment will result in a fingerprint that makes no distinction between the families. 131

4.13 Partially matching sequence can blur the distinction between true and false. 132

4.14 α- and β-haemoglobin share significant similarity. In order to define motifs to describe either individually it is necessary to produce an alignment containing both (an 'α-,β-haemoglobin' family fingerprint) and select motifs from regions where there is less commonality within the family alignment and more in the sub-family. 135

4.15 An alignment of the shared 'KMSKS' domain of three families of the tRNA synthetases. 138

4.16 The set of tRNA synthetase fingerprints in PRINTS. 139

5.1 A query sequence is not evenly represented by a fixed width window. 144

5.2 The result of padding the edges of a sequence is an even representation. 145

5.3 Sorting a list of matches by each of the scoring schemes. 151

5.4 An example list of matches. 153

5.5 The PathFinder process 154

5.6 Details of the PathFinder process 155

5.7 Derivation of the inter-motif limits 160

5.8 The FPScan submission form. 166

5.9 An example result of a FPScan search, using OPSD_SHEEP as the query sequence. 167

5.10 The top level of FPScan results 167

5.11 The second level: the ten top matches. 168

5.12 The final level: the ten top matches detailed by motif 169

5.13 The GRAPHScan output describing the matches to a single motif (motif 1 of the GPCRRHODOPSN fingerprint), plotted along the length of a sequence (SWISS-PROT:OPSD_SHEEP). 170

5.14 The GRAPHScan output describing the matches to all motifs in the GPCRRHODOPSN fingerprint, plotted along the length of a sequence (SWISS-PROT:OPSD_SHEEP). 171

5.15 A comparison between the match scoring schemes, demonstrated using the GRAPHScan tool. 173

5.16 FingerPRINTScan results are placed into a relational database to enable complex queries to be formed over the data. Below is the database schema, which describes the tables and attributes, as well as indicating the relationships between the entities represented in the database . . . 176

6.1 The first ten sequences from each of the two result datasets. 181

6.2 A plot of average AW-PID scores for sequences matching fingerprints from PRINTS 27.0. 'True' sequences are members of the true27 sequence database, while 'False' members are derived from the rand27 database. 182

6.3 A plot of summed AW-PID scores for sequences matching fingerprints from PRINTS 27.0. 'True' sequences are members of the true27 sequence database, while 'False' members are derived from the rand27 database. 183

6.4 A plot of summed AW-PID scores for sequences matching fingerprints from PRINTS 27.0, highlighting the region of cross-over between true and false scores. 184

6.5 A plot of profile scores for sequences matching fingerprints from PRINTS 27.0. 'True' sequences are members of the true27 sequence database, while 'False' members are derived from the rand27 database. 185

6.6 A plot of profile scores for sequences matching fingerprints from PRINTS 27.0, highlighting the region of cross-over between true and false scores. 186

6.7 A plot of p-values for sequences matching fingerprints from PRINTS 27.0. 'True' sequences are members of the true27 sequence database, while 'False' members are derived from the rand27 database. 187

6.8 A plot of p-values for sequences matching fingerprints from PRINTS 27.0, highlighting the region of cross-over between true and false scores. 188

6.9 For each of the scoring schemes a threshold value was set. Each scheme has a different scale (e.g., the average AW-PID ranges from 0-100%, while the p-value scale ranges from 1- $\sim 1e^{-200}$), in order to express thresholds defined on these scales, each is expressed as the percentage of false positive assignments it creates (10%, 5%, 1% and 0.01%). Tabulated are the corresponding percentages of false negatives produced by each threshold. 189

6.10 The p-value scoring scheme provides the best performance of all the scoring schemes for any given threshold. Each p-value (or e-value) threshold corresponds to a percentage of false positives and false negatives. A new threshold is introduced into this table that has an e-value threshold of $1e^{-4}$, which is the default value used by FPScan to indicate significant results. The value sits approximately midway between the 1% and 0.01% values, and effectively represents a compromise between selectivity and sensitivity. 190

6.11 Both graphs show the distribution of p-values against the number of motifs in a fingerprint. Frequency is plotted vertically, while p-value is plotted along the x-axis. Each set of fingerprint matches is represented separately based on the the number of motifs each fingerprint contains. The different views of the data clearly show the variation in the distribution of scores over the number of motifs in a fingerprint. . 192

6.12 A comparison between the number of false positive assignments made using the the true27 and true27_full datasets. As no partial matching sequences are represented, any sequences falling below thresholds are false negatives that arise as a consequence of a fingerprint failing to elevate a full matching member above the threshold. Three false-positive threshold values are shown, alongside the corresponding p-value, and the false-negative percentages (from figure 6.10), as well as the true27_full values. 194

6.13 The number of sequences falling below two thresholds analysed by the number of motifs in their families' fingerprints. All sequences falling below the two thresholds 1% and 0.01% false positives were analysed. For each sequence, the number of motifs in its family's fingerprint was extracted. This is presented in the following table as the number of fingerprints containing *n* motifs. 195

6.14 All sequences from fingerprints with 2-3, and 2-4 motifs are expressed as percentages of the total number of sequences falling below the two thresholds. 195

6.15 All of the fingerprints from figure 6.13 were collated, to produce a list of the most frequently occurring fingerprints in the list of false negative assignments (below the 1% and 0.01% thresholds). 197

6.16 Ten genomes from diverse taxa, were selected to observe the effects of variation of the scoring thresholds defined in section 6.2. 199

6.17 Result from the scan of MC3R_RAT against PRINTS 27.0. A) shows the original set of results, and B) shows the effect of supplementing these results with parent-child relationship information from PRINTS-S. 202

6.18 Result from the scan of Q9U320 against PRINTS 27.0. A) shows the original set of results, and B) shows the effect of supplementing these results with parent-child relationship information from PRINTS-S. . . 203

6.19 A GRAPHScan plot of the motif of PRINTS:GPCRMGR against SWISS-PROT:BOSS_DROME. 207

6.20 A GRAPHScan plot of the fingerprint PRINTS:BRIDEOF7LESS against SWISS-PROT:BOSS_DROME. 208

Chapter 1

Introduction

1.1 Biosequences

The quest to unravel the complexities of life on this planet has been pursued by one species in particular for many thousands of years. Today this species has a rough draft of the very essence of its collective existence: the human genome. The acceleration in our understanding of natural processes has always been kept in check by the discovery of yet greater levels of complexity. So, although today we may possess the technology to read the letters of the code of life, tomorrow we must continue to search for its interpretation.

If one attempts to break down a living organism into its components, the result is a set of simple elements, which are common to everything we consider as a living organism and also common to entities whose 'life' is still subject to debate (e.g., viruses). The simple building blocks of the great diversity of structure, function and form of living organisms are, at the most general level, amino acids, sugars, nucleotides and fatty acids.

The monomers of proteins are the amino acids, of which there are 20 commonly occurring variations (alanine, leucine, isoleucine, valine, glycine, aspartic acid, asparagine, glutamic acid, glutamine, serine, histidine, tryptophan, phenylalanine, tyrosine, proline, cysteine, threonine, lysine, methionine, arginine). Of the macromolecules essential to life, proteins are naturally very varied.

The polymers attributed with the storage and transmission of the genetic code are the nucleic acids (Deoxyribose and Ribose Nucleic Acid (DNA and RNA)). Diversity in the monomeric units comes from the combination of ribose or deoxyribose sugars with one of five nitrogenous bases (adenine, guanine, cytosine, thymine and uracil).

Fatty acids play an integral role in forming a bi-lipid membrane, which is fundamental in resisting the force of entropy by keeping all other components from diffusing away from each other.

No one group of elemental units can support life alone: it is the complex interrela-

tionships between these elements, and, more importantly, between the macromolecular structures of which they are components, that builds the simplest component of a living organism, the cell. Proteins perform a myriad of functions within the cell including: the transportation of substrates, signal passing, the maintenance of structural integrity, the construction of components, and cell duplication. DNA forms the templates upon which proteins construct new proteins, and this process is facilitated by structures formed from RNA. Finally, this synergy is encapsulated by the lipid bilayer.

Proteins

The critical importance of the role that proteins play in living organisms has been known for over 200 years (Trifonov, 2000). They exhibit extreme variability in both structure and function. This flexibility is facilitated by the large number of permutations provided by linking the range of amino acid residues together into linear chains. The three-dimensional structure of a protein is a direct consequence of the linear order, or sequence, of its constituent amino acids. As it is clear that the functional role of a protein is dependent on its shape, or the arrangement in space of chemical groups, the study of sequence and structure provides a route to the understanding of function.

Nucleic Acids

While proteins are credited with a wide range of functions, DNA acts entirely as an information storage mechanism. RNA facilitates both the storage and transmission of the genetic code and is often credited with enzymatic activity, but DNA is almost universally used as the repository of biological information due to its stability. Simplistically, every organism has a genome that contains all of the instructions required to construct, maintain and reproduce itself. A genome is subdivided into discrete units, genes, which represent encoded proteins. In higher organisms, genes only occupy a small proportion of the entire length of the genome; the remainder either consists of a

host of regulatory structures involved with gene expression or is of currently unknown function.

The linear sequence of bases within genes dictates the sequence of amino acids in the translated protein, and therefore the structure and function of the gene product. Simple genomes, such as those from prokaryotic organisms, contain genes whose linear structure is directly related to the translated product. However, eukaryotic genes contain both translated (exons) and untranslated (introns) regions. While the function of introns is mostly unknown, their consequences are far reaching. The direct result of the discontinuity exhibited by these genes is the increased difficulty in the identification of exons, which are masked by long non-translated intronic regions. The biological advantages of such a scheme seem to be associated with the observation that exons usually encode distinct protein structures (domains or modules). These domains can be spliced together during protein expression to produce overall functionalities that differ from the functions of the individual units.

Sequencing

The science of biochemistry is firmly based around the study of the effects of proteins in disease phenotypes, normal housekeeping functions, or structural and enzymatic roles. While the techniques of extraction and purification may have been important to protein science, the major landmarks are associated with the first demonstration of the sequence of a protein and the almost parallel discovery of the structure of DNA. In 1951, Sanger and Tuppy sequenced a single chain of the polypeptide hormone insulin (Sanger and Tuppy, 1951) (this followed earlier achievements in sequencing very short polypeptides): this evidence established the linearity of proteins. The structure of DNA, deduced by Watson and Crick (1953), introduced the idea that nucleic acids were also linear. This insight provided a clear link between protein and DNA, and thus the storage and dissemination of genetic information. The concepts of nucleotide

triplets (codons) and of proteins being translation products of the genetic code shortly followed these discoveries.

The inevitable popularity of determining the sequence of proteins led to the emergence of many more protein sequences. Despite the observations of the correlation between gene and protein subunit linearity, sequencing was preferentially performed on polypeptides due to their stability and relative ease of purification. The onset of DNA sequencing was heralded by techniques such as gene cloning, which provided a means of providing large quantities of purified DNA fragments. Once methods for addressing the problems of purification became established, focus turned to nucleic acid sequencing. A simple and rapid method for determining nucleotide sequences was developed by Sanger and Coulson in 1975.

The method of Sanger & Coulson, and a number of subsequent improvements (Wu, 1978), facilitated a rapid growth in sequencing and data accumulation. This growth was enhanced by the automation of methods in the early 1980s. The establishment of genome initiatives throughout the 1980s, leading up to the 1990 announcement of the Human Genome Project (HGP), provided the basis for sequencing on the scale of the whole organism. The first non-viral genome, to be sequenced, *(Haemophilus influenzae)* was announced in May 1995. Now, six years later, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Drosophila melanogaster* genomes represent the most completely sequenced eukaryotic genomes with *Homo sapiens* existing in draft form and a number of other projects underway. Also, over 20 prokaryotic genomes have been completed[1] with many more currently in progress[2].

The expansion of data has had many implications, one of the most significant being the birth of a new field of biological science, bioinformatics, which unites the fields of biology and computer science.

[1] http://www.ebi.ac.uk/genomes
[2] http://www.tigr.org/tdb/mdb/mdb.html

1.2 Databases

Dayhoff and Doolittle were amongst the earliest protagonists for organised storage of biological sequence data. Margaret Dayhoff is well known for her contribution to the development of the resource that would later become known as the Protein Sequence Database (see section 2.3) and for editing the 'Atlas of Protein Sequence and Structure' between 1965 and 1978 (Dayhoff, 1965a). Similarly, Russell Doolittle has long shared a strong interest in the power of applying computing techniques to biological problems (Doolittle, 2000). Many workers have since joined these early advocates and the effects have been widespread: establishing hundreds of biological databases, populated with millions of sequences. The increasing size of resources has: prompted changes to the infrastructure of information exchange, required greater and greater computing power, and instigated wider dissemination of biological data above and beyond the scope of traditional paper journals.

1.3 Sequence Analysis

The deposition of sequences in databases was partially prompted by the need to collect and classify data; however, the major driving force was the study of the effects of evolution on proteins. Connections made between sequence variation and heredity by Zuckerkandl and Pauling, in 1962, (reviewed by Zuckerkandl, 1975) were instrumental in the generation of the field of molecular biology. The principle that evolutionary links could be made from direct observation of sequence was further developed and utilised by Needleman and Wunsch (1970) and Dayhoff (1974). In parallel with the advances of computing, and the growth of databases, the field of sequence analysis arose. Its aim was to apply the principles of information theory to the identification of similarities between sequences in the hope of identifying these evolutionary links.

In more recent history, sequence analysis has become important not only for identi-

fying relationships between known proteins, but for the characterisation of unknown sequences. The demonstration of similarity at the amino acid or nucleotide level can be used to confer functional information derived experimentally for one sequence onto another, based on the hypothesis that sequences sharing significant similarity also share a common ancestor.

This form of analysis was made necessary by the effects that rapid sequencing technologies have had on the practice of molecular biology. At the dawn of rapid DNA sequencing, novel sequences were treated to large scale investigations, which generally resulted in much of the underlying biology being identified in order to support the initial findings (Wheelan and Boguski, 1998). The natural consequence of this was that genes, which were submitted to databases, were accordingly linked to a plethora of information (annotation). Later, as technology advanced, methods such as positional cloning took precedence: an approach in which phenotype directs study through genetic linkage to the identification of the sequence in the genome by position alone. Biochemical characterisation of these proteins, therefore, takes place almost as an after-thought. The consequence of this is a reduction in the annotation accompanying each gene. Currently, whole genomes are sequenced directly with little or no experimental determination of the biochemical functions of resultant sequences. The result is a severe annotation deficit (Boguski, 1999).

This shortage of annotation has pushed the computational analysis of protein and DNA sequences to the forefront and has resulted in the emergence of a wide range of sequence comparison tools. However, despite the revolution in bioinformatics, it is still common to find large proportions of genome data without an assigned function. The goal of many researchers is to find a solution to the annotation deficit: the stumbling block is that it is still incredibly difficult to program computers to 'think biologically'. Approaches range from the analysis of pairwise sequence similarity to the analysis of evolutionarily related families of sequences. However, it is clear that it is not a problem that has yet been solved, or to which there will be an easy solution. Indeed, current

sequence analysis tools are far from perfect, and naive use of these tools can lead to incorrect annotation. In turn, this means that there is always scope for the development of new tools that provide non-overlapping approaches. It is likely that the only way to bridge the gap between sequencing and annotation, will be an increasing use of combinations of multiple analytical techniques. However, the potential effects of misusing sequence analysis can already be seen, as databases continue to be populated with sequences annotated only *in silico*. As these sequences themselves are used more and more as sources of annotation, the potential for an explosion of errors becomes increasingly likely (Karp, 1998).

Chapter 2

Primary databases

2.1 Biological Sequences

The sequencing of polypeptides has been possible since the fifties (Sanger and Tuppy, 1951). The chemical stability and the ease of large scale purification of proteins gave peptide-based sequencing a head-start. However, by the late sixties, DNA sequencing was becoming a viable alternative (Sanger, 1988). A decade later, it had become the norm. Rapid sequencing techniques, such as the gel reading techniques of Sanger and Coulson (1975), heralded the start of the genomic era of large scale nucleic acid sequencing. Automation of these techniques in the 1990s has led to an explosion of sequence data. Currently a single machine can sequence thousands of bases per day, which stands in stark contrast to the situation in the late seventies when Wu wrote: "Today a DNA sequence of 200 nucleotides can be determined within a week" (Wu, 1978).

While sequence data have been analysed and stored in databases for over 30 years by a number of researchers (Dayhoff, 1965b; Doolittle, 2000; Needleman and Wunsch, 1970), the increasing speed of data acquisition has made such efforts ever more necessary. The value of storing and analysing sequences comes from the potential to transfer knowledge from known to unknown sequences via the inference of common ancestry. As more and more sequences from divergent organisms enters these resources, the probability of retrieving useful information increases.

This chapter will initially discuss the storage and organisation of biological sequences in databases. The theme will then turn to the analysis of sequences, and the consequences that database growth has had on the development of tools to facilitate research into the evolutionary relationships between proteins.

2.2 Nucleic Acid Databases

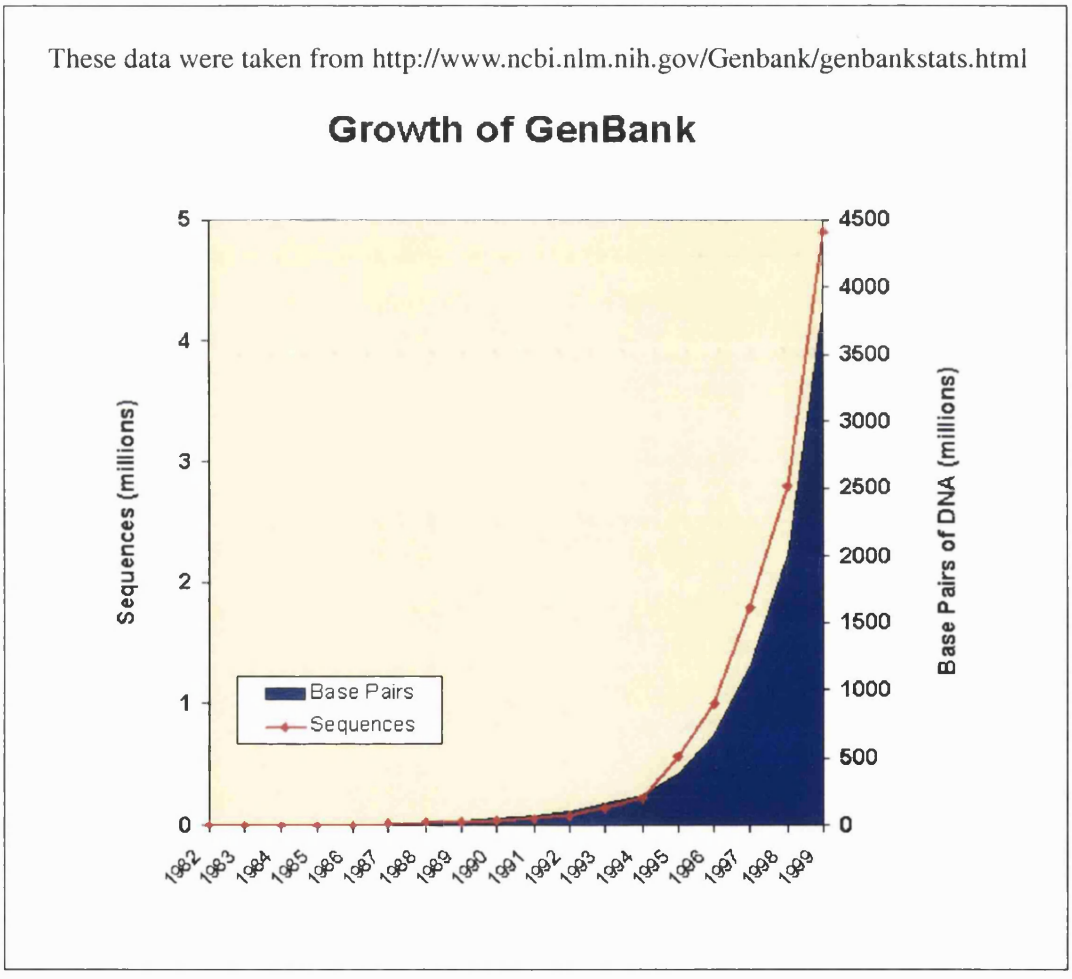
The three databases that provide the foundation for the International Nucleotide Sequence Database Collaboration (INSD), which acts as a repository for all known nucleotide and protein sequences, are:

- The GenBank database (Benson et al., 2000), maintained at the National Center for Biotechnology Information (NCBI), Maryland, United States of America (USA).
- The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (Baker et al., 2000), maintained at the European Bioinformatics Institute (EBI), Cambridge, United Kingdom.
- The DNA Data Bank of Japan (DDJB) (Tateno et al., 2000), maintained at the Center for Information Biology, Mishima, Japan.

In order to ensure that the INSD is the most comprehensive and up-to-date store of biological sequence data in the world, data submitted by researchers to any of the individual sites is propagated nightly to each of the other nodes. Releases dated August 1999 contained approximately 3.4 billion nucleotides from 4.6 million sequences, 63% of which are Expressed Sequence Tags (ESTs) (Benson et al., 2000). Human sequences constitute 56% of the total (34% of all sequences are human ESTs), while other organisms that contribute heavily are *Mus musculus, Caenorhabditis elegans, Drosophila melanogaster* and *Arabidopsis thaliana.* Figure 2.1 shows the growth of GenBank since 1982.

The primary sequence resources represented by the INSD databases are rapidly expanding. The growth of GenBank, which historically doubled in size every 18 months, has accelerated to doubling every 15 months (Benson et al., 2000). Not included in the diagram are the current statistics; the release dated August 2000 contains 9.5 billion

Figure 2.1: The growth of GenBank.



nucleotides from 8.2 million sequence records, which indicates that this rate is still increasing. The rapid accumulation of ESTs is clearly a major factor in this surge: estimates indicate that they account for approximately two thirds of all sequence data.

2.2.1 Other nucleic acid sequence resources

A large number of other centres around the world provide data repositories that are kept up to date via collaboration with INSD. The purpose of this multiplicity is twofold: not only are data stored locally, facilitating faster access, and reducing the burden on the three main distribution points, but most centres also provide local search facilities or specialised services. For example the National Center for Genome Resources (NCGR) situated in Santa Fe (USA), hosts the Genome Sequence DataBase (GSDB) (Harger et al., 2000) and provides access to accelerated implementations of the popular pairwise search tools Smith-Waterman and Frame Search. The Institute for Genomic Research (TIGR) (Quackenbush et al., 2000) maintains a database of high-fidelity, non-redundant transcripts constructed from the vast number of EST sequences in INSD. The TIGR Gene Indices (TGIs) provide valuable organism-specific analyses of ESTs, which are notoriously difficult to use and error prone. The extra level of processing of ESTs in the TGIs generates a resource that is more effective for use in functional and genomic annotation.

Many more groups worldwide maintain databases of biological sequence data. A number of World Wide Web (WWW) resources have recently been established to collate information pertaining to these collections, which give more extensive reviews than it is possible to give here ('The Molecular Biology Database Collection' and 'DBcat': Baxevanis, 2000; Discala et al., 2000). Indeed, the journal Nucleic Acids Research devotes one issue a year exclusively to databases, a practice that has been maintained since 1994. The current issue (January 2000) contains papers from 110 different databases.

2.3 Protein Sequence Databases

Much of the wealth of sequence data in the nucleic acid databases comes from non-expressed nucleic acid sequences (promoters, binding sites, untranslated regions, etc.). However, a large proportion represent genes or Open Reading Frames (ORFs), translation of which yields a protein product. The preponderance of this type of data is clearly responsible for a bias, in biological understanding, towards the mechanisms and functions of these gene products. As a result of the richness of this source of information, there exists a plethora of protein sequence databases, of varying content, maintained by many groups across the world. Such databases are repositories of data derived directly from sequence determination experiments, and, in ever increasing amounts, from translated nucleic acid sequences. These resources invariably contain raw sequence data, and the degree to which this is supplemented with information pertaining to a functional analysis or characterisation of biological role (annotation) varies from resource to resource. The following section will discuss a small collection of the more comprehensive and highly valued databases. Again, other sources are available that provide more extensive listings of sequence databases (Baxevanis, 2000; Discala et al., 2000).

2.3.1 Protein Sequence Database (PSD)

Historically, the PSD was the first collection of sequences to be established. It arose as a direct consequence of Margaret Dayhoff's work, which was disseminated in the 'Atlas of Protein Sequence and Structure'. The PSD currently exists as The Protein Information Resource (PIR) International PSD (Barker et al., 2000). Like the INSD resource (section 2.2), the PSD is now a worldwide collaborative venture, uniting its sequence data with data from Munich Information Center for Protein Sequences (MIPS) and Japan International Protein Sequence Database (JIPID).

The collection is the largest of the annotated protein sequence resources, with the June

2000 release (65) containing 182,096 entries. In addition to the primary sequence data, the resource contains information concerning: the name and classification of the protein and the organism in which it is found, primary literature references, functional and biochemical characteristics of the protein, and sites and regions of biological interest within the sequence. An example of an entry can be seen in figure 2.2. Particular points of interest are the annotation sub-headings, such as the organism, references, keywords (a collection of descriptive words from a restricted alphabet) and features (regions of the protein sequence of functional or structural significance). The database is maintained as four distinct sections (PIR1-4), with each section differing in the annotation level and quality of data. PIR1 contains data with the most complete classification and annotation, while section 4 contains sequences that have been identified as neither naturally occurring nor naturally expressed.

2.3.2 SWISS-PROT

SWISS-PROT (Bairoch and Apweiler, 2000) is a manually maintained protein sequence database. Due to consistent efforts to provide the highest level of annotation (Junker et al., 1999), to reduce redundancy and to provide extensive database integration, SWISS-PROT is considered as the gold standard. The database is maintained as a collaborative effort between EMBL-EBI Outstation and the Swiss Institute of Bioinformatics (SIB). The most recent release (39) contains 86,593 sequences. Each entry is afforded manual analysis to ensure that its annotation is descriptive and as comprehensive as possible. Annotation includes information such as function, post-translational modifications, domains and sites, secondary, tertiary and quaternary structure, similarities, disease characteristics and sequence conflicts or variants. SWISS-PROT entries (see figure 2.3) also contain extensive database cross-references to sources such as: bibliographic references, nucleic acid sequence databases, protein family databases and functional or disease state resources. SWISS-PROT is supplemented by a more comprehensive but less carefully annotated database called TrEMBL (Bairoch and

Figure 2.2: An example entry from the PIR international Protein Sequence Database.

```
ENTRY           OOSH  #type complete
TITLE           rhodopsin - sheep
ORGANISM        #formal_name Ovis orientalis aries, Ovis ammon aries
                #common_name domestic sheep
   #cross-references taxon:9940
DATE            18-Aug-1982 #sequence_revision 30-Sep-1990 #text_change
                07-May-1999
ACCESSIONS      A30407; A90319; A93264; A03155
REFERENCE       A91755
   #authors     Pappin, D.J.C.; Elipoulos, E.; Brett, M.; Findlay, J.B.C.
   #journal     Int. J. Biol. Macromol. (1984) 6:73-76
   #title       A structural model for ovine rhodopsin.
   #accession   A30407
      ##molecule_type protein
      ##residues 1-348 ##label PAP
      ##note no explanation is given for the differences in the sequence
                as seen in this paper from the original reports cited
                below
      ##note peptides and unsequenced residues are ordered by homology
                with bovine rhodopsin
REFERENCE       A90319
   #authors     Brett, M.; Findlay, J.B.C.
   #journal     Biochem. J. (1983) 211:661-670
   #title       Isolation and characterization of the CNBr peptides from
                the proteolytically derived N-terminal fragment of ovine
                opsin.
   #cross-references MUID:83282605
   #accession   A90319
      ##molecule_type protein
      ##residues
                1;40-44;45-86;87-111;144-155;156-163;164-183;184-207;208-2
                41 ##label BRE
REFERENCE       A93264
   #authors     Findlay, J.B.C.; Brett, M.; Pappin, D.J.C.
   #journal     Nature (1981) 293:314-316
   #title       Primary structure of C-terminal functional sites in ovine
                rhodopsin.
   #cross-references MUID:82013638
   #accession   A93264
      ##molecule_type protein
      ##residues 240-348 ##label FIN
REFERENCE       A90324
   #authors     Pappin, D.J.C.; Findlay, J.B.C.
   #journal     Biochem. J. (1984) 217:605-613
   #title       Sequence variability in the retinal-attachment domain of
                mammalian rhodopsins.
   #cross-references MUID:84178280
   #contents    annotation; retinal binding site
REFERENCE       A44548
   #authors     Thompson, P.; Findlay, J.B.C.
   #journal     Biochem. J. (1984) 220:773-780
   #title       Phosphorylation of ovine rhodopsin: identification of the
                phosphorylated sites.
   #cross-references MUID:84279984
   #contents    annotation; phosphorylation sites
CLASSIFICATION  #superfamily vertebrate rhodopsin
KEYWORDS        acetylated amino end; chromoprotein; eye; G
                protein-coupled receptor; glycoprotein; lipoprotein;
                phosphoprotein; photoreceptor; retina; retinal; thiolester
                bond; transmembrane protein; vision
FEATURES
   37-61                #domain transmembrane #status predicted #label
                        TM1\
   74-96                #domain transmembrane #status predicted #label
                        TM2\
   114-133              #domain transmembrane #status predicted #label
                        TM3\
   153-175              #domain transmembrane #status predicted #label
                        TM4\
   203-230              #domain transmembrane #status predicted #label
                        TM5\
   253-276              #domain transmembrane #status predicted #label
                        TM6\
   285-309              #domain transmembrane #status predicted #label
                        TM7\
   1                    #modified_site acetylated amino end (Met) #status
                        experimental\
   2,15                 #binding_site carbohydrate (Asn) (covalent)
                        #status predicted\
   296                  #binding_site retinal (Lys) (covalent) #status
                        experimental\
   322,323              #binding_site palmitate (Cys) (covalent) #status
                        predicted\
   334,338,343          #binding_site phosphate (Ser) (covalent) (by
                        rhodopsin kinase) #status experimental\
   335,336              #binding_site phosphate (Thr) (covalent) (by
                        rhodopsin kinase) #status experimental
SUMMARY         #length 348 #molecular_weight 38891
SEQUENCE
           5         10        15        20        25        30
    1 M N G T E G P N F Y V P F S N K T G V V R S P F E A P Q Y Y
   31 L A E P W Q F S M L A A Y M F L L I V L G F P I N F L T L Y
   61 V T V Q H K K L R T P L N Y I L L N L A V A D L F M V F G G
   91 F T T T L Y T S L H G Y F V F G P T G C N L E G F F A T L G
  121 G E I A L W S L V V L A I E R Y V V V C K P M S N F R F G E
  151 N H A I M G V A F T W V M A L A C A A P P L V G W S R Y I P
  181 Q G M Q C S C G A D Y F T L K P E I N N E S F V I Y M F V V
  211 H F S I P L I V I F F C Y G Q L V F T V K E A A A Q Q Q E S
  241 A T T Q K A E K E V T R M V I I M V I A F L I C W L P Y A G
  271 V A F Y I F T H Q G S D F G P I F M T I P A F F A K S S S V
  301 Y N P V I Y I M M N K Q F R N C M L T T L C C G K N P L G D
  331 D E A S T T V S K T E T S Q V A P A
```

Apweiler, 2000) (Translation of EMBL nucleotide sequence database). This resource exists to bridge the gap between the requirement to store the relentless flow of protein-coding sequence data from genome sequencing projects, and the desire to maintain the level of detailed annotation that SWISS-PROT is known for. TrEMBL, as its name suggests, is based on the translation of coding sequences from the EMBL database. After extensive redundancy checks to ensure that each sequence represents a unique protein, a number of automated sequence analysis steps are performed to identify annotation that can be transfered to these sequences. As this annotation is based purely on *in silico* techniques, it is flagged in each database entry: 'BY SIMILARITY', which serves as a warning that it represents a less conclusive diagnosis than annotation derived via experimental means.

2.3.3 Summary

The importance of annotation-supplemented resources cannot be understated. The databases discussed previously (section 2.2) are comprehensive (i.e., they represent up-to-date collections of all published biological sequence data), but many entries contain little or no annotation. SWISS-PROT and PSD, on the other hand, by no-means represent comprehensive collections of all available protein sequences. However, considerable benefits come from providing links between sequence information and the disparate collection of information regarding both the protein's characterisation, and functional or structural evidence derived from experimental means. A common feature of such databases is the availability of searchable fields, which allow one to identify and collate entries containing relevant information. For example, in order to browse information pertaining to the protein "p53", it is only necessary to use the search facility to look for the word "p53". An alternative to searching all available text is to select relevant keywords from a restricted set of words selected to best describe the protein and its role or function. The keywords for the SWISS-PROT sequence identifier (ID) OPSD_SHEEP are: Photoreceptor, Retinal protein, Transmembrane, Gly-

Figure 2.3: An example entry from the SWISS-PROT protein sequence database. Highlighted regions indicate keywords, and an example of ambiguity in the naming of the family, to which the protein belongs.

```
ID   OPSD_SHEEP     STANDARD;      PRT;   348 AA.
AC   P02700;
DT   21-JUL-1986 (Rel. 01, Created)
DT   01-FEB-1991 (Rel. 17, Last sequence update)
DT   15-JUL-1999 (Rel. 38, Last annotation update)
DE   RHODOPSIN.
GN   RHO.
OS   Ovis aries (Sheep).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovoidea;
OC   Bovidae; Caprinae; Ovis.
RN   [1]
RP   SEQUENCE.
RA   Pappin D.J.C., Elipoulos E., Brett M., Findlay J.B.C.;
RT   "A structural model for ovine rhodopsin.";
RL   Int. J. Biol. Macromol. 6:73-76(1984).
RN   [2]
RP   SEQUENCE OF 1-111 AND 144-239.
RX   MEDLINE; 83282605.
RA   Brett M., Findlay J.B.C.;
RT   "Isolation and characterization of the CNBr peptides from the
RT   proteolytically derived N-terminal fragment of ovine opsin.";
RL   Biochem. J. 211:661-670(1983).
RN   [3]
RP   SEQUENCE OF 240-348.
RX   MEDLINE; 82013638.
RA   Findlay J.B.C., Brett M., Pappin D.J.C.;
RT   "Primary structure of C-terminal functional sites in ovine
RT   rhodopsin.";
RL   Nature 293:314-316(1981).
RN   [4]
RP   RETINAL BINDING SITE.
RX   MEDLINE; 84178280.
RA   Pappin D.J.C., Findlay J.B.C.;
RT   "Sequence variability in the retinal-attachment domain of mammalian
RT   rhodopsins.";
RL   Biochem. J. 217:605-613(1984).
CC   -!- FUNCTION: VISUAL PIGMENTS ARE THE LIGHT-ABSORBING MOLECULES THAT
CC       MEDIATE VISION. THEY CONSIST OF AN APOPROTEIN, OPSIN, COVALENTLY
CC       LINKED TO CIS-RETINAL.
CC   -!- SUBCELLULAR LOCATION: INTEGRAL MEMBRANE PROTEIN.
CC   -!- TISSUE SPECIFICITY: ROD SHAPED PHOTORECEPTOR CELLS WHICH MEDIATES
CC       VISION IN DIM LIGHT.
CC   -!- PTM: SOME OR ALL OF THE CARBOXYL-TERMINAL SER OR THR RESIDUES MAY
CC       BE PHOSPHORYLATED.
CC   -!- MISCELLANEOUS: THIS RHODOPSIN HAS AN ABSORPTION MAXIMA AT 495 NM.
CC   -!- SIMILARITY: BELONGS TO FAMILY 1 OF G-PROTEIN COUPLED RECEPTORS.
CC       OPSIN SUBFAMILY.
DR   PIR; A30407; OOSH.
DR   GCRDB; GCR_0194; -.
DR   PFAM; PF00001; 7tm_1; 1.
DR   PRINTS; PR00237; GPCRRHODOPSN.
DR   PRINTS; PR00238; OPSIN.
DR   PRINTS; PR00579; RHODOPSIN.
DR   PROSITE; PS00237; G_PROTEIN_RECEPTOR; 1.
DR   PROSITE; PS00238; OPSIN; 1.
KW   Photoreceptor; Retinal protein; Transmembrane; Glycoprotein; Vision;
KW   Phosphorylation; Lipoprotein; Palmitate; G-protein coupled receptor.
FT   DOMAIN        1     36       EXTRACELLULAR.
FT   TRANSMEM     37     61       1 (POTENTIAL).
FT   DOMAIN       62     73       CYTOPLASMIC.
FT   TRANSMEM     74     98       2 (POTENTIAL).
FT   DOMAIN       99    113       EXTRACELLULAR.
FT   TRANSMEM    114    133       3 (POTENTIAL).
FT   DOMAIN      134    152       CYTOPLASMIC.
FT   TRANSMEM    153    176       4 (POTENTIAL).
FT   DOMAIN      177    202       EXTRACELLULAR.
FT   TRANSMEM    203    230       5 (POTENTIAL).
FT   DOMAIN      231    252       CYTOPLASMIC.
FT   TRANSMEM    253    276       6 (POTENTIAL).
FT   DOMAIN      277    284       EXTRACELLULAR.
FT   TRANSMEM    285    309       7 (POTENTIAL).
FT   DOMAIN      310    348       CYTOPLASMIC.
FT   CARBOHYD      2      2       BY SIMILARITY.
FT   CARBOHYD     15     15       BY SIMILARITY.
FT   BINDING     296    296       RETINAL CHROMOPHORE.
FT   LIPID       322    322       PALMITATE (BY SIMILARITY).
FT   LIPID       323    323       PALMITATE (BY SIMILARITY).
FT   DISULFID    110    187       BY SIMILARITY.
FT   MOD_RES     343    343       PHOSPHORYLATION (BY RK) (BY SIMILARITY).
SQ   SEQUENCE   348 AA;  38891 MW;  AAFD6F0D6A8BAEE5 CRC64;
     MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIVL GFPINFLTLY
     VTVQHKKLRT PLNYILLNLA VADLFMVFGG FTTTLYTSLH GYFVFGPTGC NLEGFFATLG
     GEIALWSLVV LAIERYVVVC KPMSNFRFGE NHAIMGVAFT WVMALACAAP PLVGWSRYIP
     EGMQCSCGAL YFTLKPEINN ESFVIYMFVV HFSIPLIVIF FCYGQLVFTV KEAAAQQQES
     ATTQKAEKEV TRMVIIMVIA FLICWLPYAG VAFYIFTHQG SDFGPIFMTI PAFFAKSSSV
     YNPVIYIMMN KQFRNCMLTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
//
```

coprotein, Vision, Phosphorylation, Lipoprotein, Palmitate and G-protein coupled receptor. The protein, rhodopsin, is a light absorbing molecule, which is an integral membrane protein belonging to the G-protein coupled receptor super-family. Its specificity is to rod cells, which mediate vision in low light conditions. Selecting just two of these keywords "retinal" and "vision" and searching SWISS-PROT (release 38, containing 80,000 sequences) produces 152 results, the majority of which are opsins and rhodopsins. Linked to each of these 'hits' are references to publications containing a wealth of information about the family of proteins that are "retinal" binding and are involved in "vision".

Clearly the databases represent a massive collection of biological information and, as indicated, attempts have been made to facilitate access to the data through the development of computational tools, such as the text searching facility 'Sequence Retrieval System' (SRS) (Etzold et al., 1996). SRS provides fast access to multiple databases through a number of pre-query processing steps that involve the creation of indices. However, searching any text-field relies on the quality and scope of the annotation. Also, in many instances the vocabulary of biologists is imprecise, which does not lend itself to the rigidity of computational pattern- or string-matching. An example of this ambiguity can be found in the OPSD_SHEEP entry (figure 2.3), where the super-family that it belongs to is known variously as the G-protein coupled receptors, GPCRs and 7tm (seven transmembrane) receptors. Hence, to be certain of retrieving all of the relevant data during a search for these proteins, it would be necessary to include all of the alternative naming conventions.

2.4 Pairwise Sequence Analysis

The cornerstone of sequence analysis is the hypothesis that similarity between two genes (or gene products) indicates a shared ancestral heritage. The relationship between between proteins that have descended with divergence from a common ancestor

is defined as homologous (Fitch, 1970; Wray and Abouheif, 1998). In the light of this, if two proteins are demonstrated to be homologous, an inference can be made that they may also share functional and/or structural characteristics. Confidently identifying homologous sequences, and distinguishing these from unrelated ones, allows experimentally obtained and verified information to be passed on to sequences for which no annotation exists.

The degree to which a homologous relationship is demonstrable from sequence similarity depends on the quantity of identifiable similarities between the proteins. Equally, the identification of similarity in the first place is heavily reliant on the methods available to compare two protein sequences. The following sections describe the principles of sequence similarity and the methods that have been developed to demonstrate it.

2.4.1 Sequence similarity

2.4.1.1 Identity

Similarity between sequences can be measured as the number of residues one sequence shares with another (percentage identity).[1] High identity, as mentioned above, can be used to indicate homology. However, when identity drops below a threshold value (approximately thirty percent) it becomes difficult to assign a relationship with confidence. This threshold is commonly referred to as the Twilight Zone (Doolittle, 1986).

To measure the similarity of two sequences, it is necessary to align those sequences so that equivalent residues are moved into register. The principle behind pairwise alignment relies on the assumption that if two sequences share a common ancestor, then mutational variations between them can be observed more clearly if they are superimposed. Figure 2.4 shows two beta (β) haemoglobin sequences from different organisms (in both organisms, β haemoglobin plays the same role, forming a co-operative

[1] Similarity can be expressed as percentage identity (% ID). One hundred percent identity between two sequences implies that both sequences are identical.

hetero-tetramer, αβαβ, with alpha (α) haemoglobin). Accordingly, the alignment indicates very few differences between the sequences. Indeed, all of the variation is limited to point mutations, with 131 of 146 residues being identical.

Figure 2.4: An alignment of bactrian camel and human β-haemoglobin.

(HBB_CAMDR and HBB_HUMAN).

An alignment highlights the similarity between two sequences more specifically than merely counting the number of residues that the sequences share. These sequences, both being mamalian β-haemoglobins, are so very similar that a homologous relationship is certain.

```
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
VHLSGDEKNAVHGLWSKVKVDEVGGEALGRLLVVYPWTRRFFESFGDLSTADAVMNNPKV

KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
KAHGSKVLNSFGDGLNHLDNLKGTYAKLSELHCDKLHVDPENFRLLGNVLVVVLARHFGK

EFTPPVQAAYQKVVAGVANALAHKYH
EFTPDLQAAYQKVVAGVANALAHRYH
```

The upper sequence, on each of the three rows of the alignment, is the human and the lower is the camel sequence.

Two less-similar sequences are the human α and β haemoglobins, which, while functionally analogous, have been diverging side by side since an ancient gene duplication event. These sequences (figure 2.5) are significantly more divergent than those in the previous example; therefore, the similarity is more challenging to identify.

To facilitate the identification of more distant relationships, the alignment can be augmented by highlighting the shared identities. Figure 2.6 illustrates this process using the sequences from the previous example. Now, however, the relationship is more obvious.

The process of alignment facilitates the comparison of the two sequences and the identification of shared residues. Importantly, the alignment also highlights positions that are not conserved, i.e., those residues that have been subject to mutational drift. All of

Figure 2.5: A pairwise alignment of human α and β-haemoglobin.

(HBA_HUMAN and HBB_HUMAN).

In this alignment the sequences are more distantly related, so much so that they are no longer the same length, which indicates that insertion or deletion events have occured. Gaps are introduced to the sequences to simulate these events.

```
LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF......DLSHGSAQV
LTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV

KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK

EFTPAVHASLDKFLASVSTVLTSKY
EFTPPVQAAYQKVVAGVANALAHKY
```

The upper sequence, on each of the three rows of the alignment, is β and the lower is the α sequence.

Figure 2.6: A pairwise alignment of human α and β-haemoglobin showing identities between sequences.

(HBA_HUMAN and HBB_HUMAN)

An alignment of distantly related sequences can be augmented with a row highlighting those residues that are shared between the two sequences.

```
LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF......DLSHGSAQV
L.P..K..V.A.WGKV.....E.G.EAL.R.....P.T...F..F......D...G...V
LTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV

KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA
K.HGKKV..A.....AH.D........LS.LH..KL.VDP.NF.LL...L...LA.H...
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK

EFTPAVHASLDKFLASVSTVLTSKY
EFTP.V.A...K..A.V...L..KY
EFTPPVQAAYQKVVAGVANALAHKY
```

The upper sequence, on each of the three rows of the alignment, is β and the lower is the α sequence.

these mutations are the combined result of two distinct processes. Firstly, the occurrence of a point mutation in the gene template of a protein that results in the alteration of a single amino acid. Secondly, the acceptance, by the host species, of the mutated protein as the predominant form. Acceptance is usually facilitated by the replacement amino acid side chain being chemically similar to the original, and therefore not adversely affecting the stability or function of the protein. Such mutations are more readily accepted than others. As a consequence of this, looking for similarities between residues, as well as identities, represents a more sensitive approach to determining the relatedness of protein sequences.

A rarer, but more significant mutation, is the insertion or deletion of one or more residues. Its presence in an alignment necessitates the introduction of a gap character in order to keep residues aligned. The significance of gaps in alignments will be explored later.

2.4.1.2 Amino acid side chain similarities

The sequences in figure 2.5 are significantly different, sharing only 42% identity (61/145 residues); however, if amino acid *similarities* are considered (figure 2.7), this value becomes almost 60%.

While all amino acid side-chains have distinct chemical structures, they can be grouped together on the basis of shared physical and chemical properties (figures 2.8 and 2.9).

The concept that certain residues share common properties has implications for the evolutionary tolerance of certain amino-acid substitutions over others. If, for example, the amino acid valine was replaced with leucine, the resultant effect on the stability and viability of the protein may be insignificant. However, a substitution between glycine and tyrosine could constitute a significant modification. Such a mutation may result in the destabilisation of local residue packing, which in turn could adversely affect the catalytic function of the protein. Destabilising mutations are predicted to be accepted

Figure 2.7: A pairwise alignment of human α- and β-haemoglobin showing similarities between sequences.

(HBA_HUMAN and HBB_HUMAN)

The augmentation of an alignment of distantly related sequences can be furthered by adding information about amino acid similarities. Similarities can be based on physical and chemical properties of the amino acids, the alignment below is annotated with '+' for each similar amino acid pair. The criteria by which the amino acids in this example are deemed 'similar' are the groupings defined in figure 2.8.

```
LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF......DLSHGSAQV
L+P.+K+.V.A+WGKV...+.E.G+EAL.R..+.+P.T..+F..F......D+..G+..V
LTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV

KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA
K.HGKKV+.A.+.++AH+D.+......LS+LH+.KL+VDP.NF+LL..+L++.LA.H...
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK

EFTPAVHASLDKFLASVSTVLTSKY
EFTP.V.A...K.+A.V..+L..KY
EFTPPVQAAYQKVVAGVANALAHKY
```

The upper sequence, on each of the three rows of the alignment, is the β and the lower is the α sequence.

Figure 2.8: A classification of the collective properties of the amino acids.

To categorise the amino acids, it is necessary to be flexible due to their overlapping features; however, it is possible to represent common groupings.

Property	Amino acid (single letter code)
Aliphatic	G A V L I P C M
Aromatic	F Y W
Polar/Neutral	S T N Q
Polar/Acidic	D E
Polar/Basic	H K R

Figure 2.9: A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physico-chemical properties.

The relationships between the amino acids are complex and overlapping but categorisations can be made to group residues together.

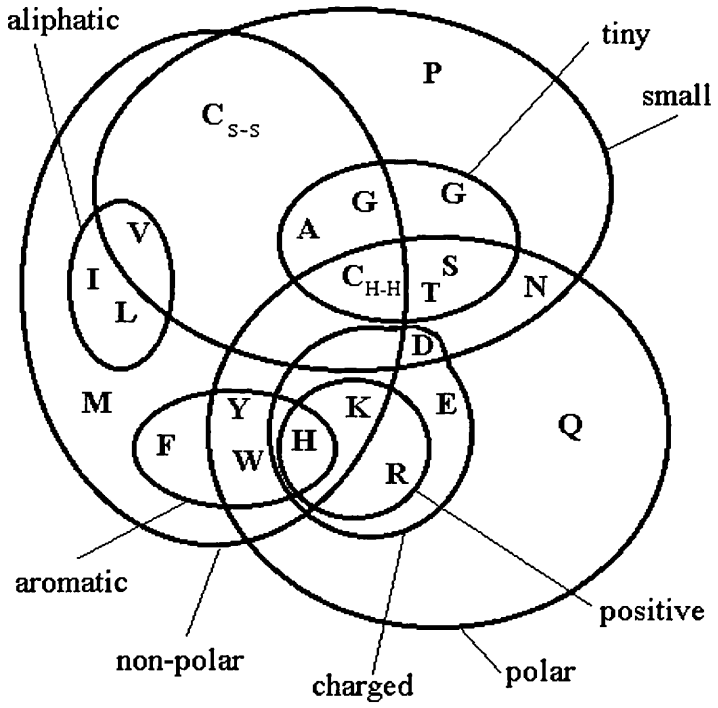


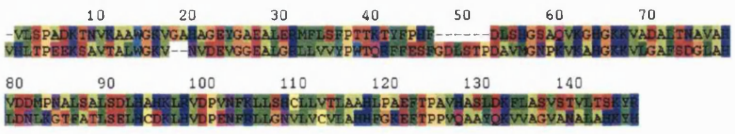
Diagram taken from Taylor (1986).

at a significantly lower frequency than mutations that have little effect on protein structural stabilisation factors (such as packing, salt bridges and hydrogen bonding). The alignments in figures 2.7 and 2.10 are highlighted to emphasise the residue similarities between the proteins. This feature, in both forms, simplifies the diagnosis of a relationship between the two proteins. Labelling alignments with characters indicating identity and similarity (e.g., figure 2.7) is simple and can be achieved without specialist software; however, the advantages of colouring an alignment are obvious (e.g. 2.10 and 2.11). Colour schemes can be devised to either enhance commonality or emphasise deviation, and as a result they are widely used; however, it is most common to see them used with alignments of more than two sequences.

Figure 2.10: A coloured alignment of human α- and β-haemoglobin.

(HBA_HUMAN and HBB_HUMAN).

This alignment represents a different perspective on the alignment in figure 2.4. The distantly related sequences are coloured with respect to the properties of the amino acid side chains such that similar physicochemical properties have similar shades of colour. (e.g., hydrophobicity is represented by the green portion of the spectrum, and hydrophilicity represented by shades of red. This particular colour scheme is attributable to Taylor (1997). Colouring an alignment in this way emphasises the regions of sequence that may have obtained mutations, but in which only mutations between similar residues have been tolerated.



The preceding sections have elaborated on the utility of alignments in the identification of sequence similarity. The demonstrated usage has been purely qualitative, with identification and diagnosis being performed by eye. The following section will discuss the quantitative measurement of sequence similarity that leads to the development of predictive algorithms based on scoring similarity; however, the principles remain the same, with strong reliance on the quality of the alignment.

Figure 2.11: A coloured alignment of human α- and β-haemoglobin that emphasises commonality.

(HBA_HUMAN and HBB_HUMAN).

This alignment again displays the relationship between two related sequences. However, the alignment in figure 2.10 is painted in a colour scheme that, while defining groups of amino acids, uses a range of colours, which means that no two residues have the same colour. This colour scheme attributes a single colour to each group of residues (the default colour scheme used in the alignment editor CINEMA (Parry-Smith et al., 1998)). The result is an alignment that clearly displays the conservation between the two sequences.

2.4.2 Scoring the alignment

The previous section discussed the identification of similarity between two sequences. Now, consider this type of analysis on a different scale. The sheer quantity of data in the sequence databases means that hundreds of thousands of sequences must be compared in order to identify similarities. The naturally time-consuming process of subjective, manual sequence analysis is clearly an inefficient means of performing analyses of this magnitude. In order to identify similarity on this kind of scale, automated diagnoses need to be made. The first step towards automation must be to quantify the measure of similarity or to score it.

2.4.2.1 Counting and scoring identity

Measuring identity, as mentioned in section 2.4.1.1, is one metric that can be used in the task of distinguishing between related and unrelated sequences. This trivial computation can be formalised as scoring the alignment between two sequences with an Identity Matrix (IM) (figure 2.12), in which each of the 20 amino acids is provided with a score for its alignment with any other residue. The 20 by 20 matrix is populated with 0 for any non-self alignment and 1 for alignments between identical residues. A

sequence achieves a score based on the sum of the scores of each of its aligned residues divided by the length of the alignment. This score is known as the percentage identity.

Figure 2.12: The identity matrix.

The identity matrix simply scores an aligned pair of amino acids 1 if they are identical and 0 otherwise.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

This simple metric allows a defined threshold to be determined, above which relationships are deemed to be 'real'. The Twilight Zone is a cut-off point beyond which demonstrating similarity cannot be used to infer homology (Sander and Schneider, 1991). This point is usually measured as $\sim 20-30\%$ identity between sequences of over 100 residues in length. The existence of this threshold stems from the observation that the alignment of random sequences can achieve scores of up to 30% identity. Therefore, scores below this level cannot be deemed to be statistically significant, and therefore cannot be used to infer a homologous relationship. As the Twilight Zone is entered, separation of true relationships from false rapidly moves from a trivial task to a task akin to finding 'a needle in a haystack' (Rost, 1999).

The drawback of the identity scoring scheme is its simplicity. The maintenance of function during natural selection does not directly correspond to the maintenance of *identity* between sequences or residues. A mutation is not necessarily something that abrogates function: it can also be null or neutral. Therefore, a distantly related sequence can become highly populated with changes that make no difference to function. The accumulation of these mutations has no effect on the homologous relationship, yet it can make similarity more difficult to identify, and accordingly, homology more difficult to infer. To detect relationships at greater distances it is necessary to adopt a different scoring scheme.

One suitable scheme, which takes into account the changes that must occur in the codon to affect the protein sequence, is the Genetic Code Matrix (GCM) (Dayhoff, 1978). This matrix accounts for the minimum number of base changes required to alter the codon for one amino acid to that of another. Scores are assigned as follows: 3, for identical amino acids; 2, for those whose codons differ by one base; 1, for those differing by two bases; and 0 for all amino acids whose codon are different in all three positions. In matrix comparisons performed by Dayhoff (1978), using small datasets, the GCM and IM matrices do not perform significantly differently. It is plausible that the assumption upon which the GCM is based (i.e., that one amino acid is *more* likely to be mutated to another if doing so represents a simple genetic event - a single base change rather than three) does not take into account the selective pressure that the protein sequence/structure can be under to maintain functional integrity. However, point mutations that result in unfavourable amino acid changes should be observed disproportionately less frequently in functional proteins, because of their greater likelihood to give rise to deleterious effects (i.e., result in non-functional proteins).

The discussion of these simple matrices leads back to the concepts of observed similarity between amino acids, and the possibility of deriving scoring matrices based on these properties. The following section will describe two of the most popular alignment scoring matrices based on observed mutations.

2.4.2.2 Counting and scoring similarity

The Mutation Data Matrix (MDM)

The MDMs developed by Dayhoff are based on comparisons between closely related sequences and the examination of amino acid exchanges or substitutions. Generating these matrices involved counting the number of occurrences of each of the potential amino acid transitions (i.e., 20 x 20 = 400 transitions). Closely related sequences (> 85% ID) were selected to maximise the chance that every observed amino acid exchange between sequences was due to one event and not successive mutations. The resultant matrix is populated with the observed frequencies of each of the possible substitutions in the dataset. To compute the probability that one amino acid will mutate into another, it is necessary to find the relative mutability of each of the amino acids: the number of times it changes, divided by the number of times it occurs (i.e., the number of times that it was available for mutation). Given this for each amino acid, and the substitution frequencies, a Mutation Probability Matrix (MPM) was calculated. Each element of this matrix expresses the probability that a particular amino acid will be replaced by a second after a given evolutionary interval. The interval is expressed as 1 Point Accepted Mutation (PAM), which represents the time interval in which a single point mutation is accepted per 100 residues. It is often useful to score a relationship between two sequences against a null model that represents the chance occurrence of such a relationship. The null model is the probability of occurrence of the target residue in a sequence by chance. Therefore, the ratio of the observed transition to the target against its chance occurrence represents the relative significance of the observation of such a transition (often referred to as the odds ratio). Simply, this means that an amino acid substitution with an odds ratio greater than 1 occurs more often in related sequences than unrelated, and *vice versa* for those less than 1. The resultant MDM_{78} matrix, which is used throughout sequence analysis today, is produced by taking logs of all the values in the odds matrix. This MDM for 1 PAM can then extrapolated to

greater evolutionary distances by serial multiplication.

Using this, and members of the series of evolutionarily extrapolated PAM matrices, can be more sensitive for identifying distant relationships than the IM and GCM matrices (Dayhoff, 1978). The calculation of the score for an alignment is the same as the procedure of summing the scores from the IM. The log odds matrix contains both positive and negative values. Therefore, alignments that contain substitutions consistently found in related sequences will score highly positively, while the summation of aligned positions between unrelated sequences should rapidly diminish below zero.

While the extrapolations of the PAM series provide a facility for the inference of distant sequence relationships, the original dataset for the construction of the PAM 1 matrix was a very closely related set of sequences (>85% ID). Henikoff and Henikoff (1992) developed a procedure to compute matrices from alignments of protein sequences with greater divergence, producing the BLOSUM series. The process is fundamentally different from the one used to calculate the PAM series, and accordingly there are significant differences in the performance of these matrices for the identification of homologues (Henikoff and Henikoff, 1993).

The Blocks Substitution Matrices (BLOSUM)

The BLOSUM series of matrices are generated from conserved blocks of aligned sequences that comprise the Blocks database (Henikoff and Henikoff, 1991), which will be discussed in detail in the following chapter. The computation of the BLOSUM series is based on calculating substitutions that occur between amino acids in columns of a block (figure 2.13). Blocks can contain sequences at a variety of evolutionary distances; therefore, in order to produce matrices that reflect substitutions occurring at defined distances, sequences within the blocks are clustered. The principle of clustering involves computing pair-wise percentage identities for every pair of sequences (in the block) and the definition of a clustering threshold. Any sequences sharing a

percentage identity greater than (or equal to) the clustering threshold is considered to belong to the same cluster. The process is repeated until all sequences have been assigned to clusters, and any clusters that share 'above-threshold' scores have been merged. The result is a number of clusters of highly similar sequences (how similar is dependent on the threshold value: 100% would collect together all sequence that were identical). Substitutions are measured in the same way as figure 2.13, but contributions from sequences within a cluster are averaged (figure 2.14). This has the effect of identifying residue substitutions that occur between sequences with less identity than the threshold (i.e., sequences that are more divergent than the threshold, and therefore exist in different clusters).

Figure 2.13: An aligned block of residues, and the pairwise substitutions that are observed between its constituent sequences.

```
1, ASEWR
2, ATEYR
3, ASEWK
```

Between sequences 1 and 2, the observed substitutions are: A-A, S-T, E-E, W-Y and R-R; between 1 and 3: A-A, S-S, E-E, W-W and R-K; and between 2 and 3: A-A, T-S, E-E, Y-W and R-K. A matrix containing the observed number of substitutions can be computed:

-	A	E	K	R	S	T	W	Y
A	3	0	0	0	0	0	0	0
E	0	3	0	0	0	0	0	0
K	0	0	0	2	0	0	0	0
R	0	0	2	1	0	0	0	0
S	0	0	0	0	1	2	0	0
T	0	0	0	0	2	0	0	0
W	0	0	0	0	0	0	1	2
Y	0	0	0	0	0	0	2	0

These distant relationships, between clusters, mirror the extrapolations made of the PAM 1 matrix in determining substitution scores for varying evolutionary distances;

Figure 2.14: An aligned block of residues clustered at 80%, and the effect on the pairwise substitutions that are observed.

Sequences 1 and 3 are 80% identical; therefore, they are placed into a cluster, labelled 4.

```
1,  ASEWR
3,  ASEWK
4,  ASEWR/K
2,  ATEYR
```

Only substitutions between 4 and 2 are counted: A-A, S-T, E-E, W-Y, $\frac{1}{2}$R-R and $\frac{1}{2}$R-K. The resultant matrix is shown:

-	A	E	K	R	S	T	W	Y
A	1	0	0	0	0	0	0	0
E	0	1	0	0	0	0	0	0
K	0	0	0	0.5	0	0	0	0
R	0	0	0.5	0.5	0	0	0	0
S	0	0	0	0	0	1	0	0
T	0	0	0	0	1	0	0	0
W	0	0	0	0	0	0	0	1
Y	0	0	0	0	0	0	1	0

however, BLOSUM matrices require no mathematical manipulation as they represent real observations at each distance. Different members of the BLOSUM series are identified by their clustering percentage, and it is possible to find equivalence between PAM distances and clustering thresholds (Henikoff and Henikoff, 1992). The ability of the BLOSUM series of matrices, and of BLOSUM 62 in particular, in searching for a defined set of homologous relationships was demonstrated to out-perform any member of the PAM series (Henikoff and Henikoff, 1993). Consequently, the BLOSUM matrices have become a standard option for scoring alignments in a wide range of pairwise analysis tools. Regardless of the performance of any individual matrix, it must be noted that when investigating an unknown relationship between sequences, one should score the alignment with a range of matrices so as to be sure that the alignment is being scored appropriately.

While the scoring of an alignment is an important issue (and will be returned to) the focus of this chapter will now turn to the task of identifying the optimal alignment of two sequences. The intensive nature of the task of searching a database of sequences to find an alignment has necessitated the development of computational tools and algorithms. This is especially important as the sizes of sequence databases continue to grow exponentially.

2.4.3 Identifying the optimal alignment

In order to identify the best alignment of two sequences, it is necessary to consider every possible permutation of residues and insertions/deletions. However, as the lengths of the sequences increases, enumerating these permutations becomes impractical. Matrix-based methods of sequence comparison can aid the identification of aligned residues without the requirement to calculate each possible arrangement. The most simple example was developed 30 years ago, and it remains in use today throughout a wide variety of implementations. It also forms the basis of the initial processing stages of a

number of the algorithms that followed it.

The dot-plot or dot-matrix method is attributed to the work of a number of authors (Fitch, 1969; Gibbs and McIntyre, 1970). The most elementary type of dot-matrix requires the residues of two sequences to be placed along opposing axes of a two-dimensional matrix (figure 2.15). Where intersecting cells reveal identical residues, a 1 is placed.

Figure 2.15: The dot-matrix

A dot-matrix is constructed by placing the residues of two sequences along the x and y axes of a matrix. A dot, in this case the value 1, is placed at every intersecting cell that reveals an identity in both sequences.

A. This matrix shows the comparison of two identical sequences

	A	C	T	A	G
A	1	0	0	1	0
C	0	1	0	0	0
T	0	0	1	0	0
A	1	0	0	1	0
G	0	0	0	0	1

B. The second matrix represents the effect of an insertion, which is to remove the line of identities from the diagonal

	A	C	T	A	G
A	1	0	0	1	0
G	0	0	0	0	1
C	0	1	0	0	0
T	0	0	1	0	0
A	1	0	0	1	0
G	0	0	0	0	1

The comparison of long sequences using this method benefits from a graphical representation of dots rather than populating a very large matrix with 1's. Two identical sequences are represented as a diagonal unbroken line, while difference between two

similar sequences would be seen as breaks in the continuity of the line. This simple representation of alignment has benefitted from a number of modifications since its inception, yet its strength lies in visual inspection of the plot. In searching large datasets, manual identification is impossible; however, the principles of this method underlie the algorithms subsequently developed to overcome its limitations.

2.4.3.1 Global alignment algorithms

The first algorithms that were developed to efficiently align two sequences were based on Dynamic Programming (DP). The simplest incarnation of the DP algorithm for aligning sequences can be seen as the recursive delineation of a path through a matrix. The matrix resembles the one constructed for the dot-plot (figure 2.15) in that the two sequences are placed on opposing axes. Each intersecting point in the matrix represents the alignment of either the two corresponding residues from each sequence, or the alignment of one residue from either sequence with a gap. Accordingly, the score associated with an alignment between two residues is provided by a suitable substitution matrix, while alignment with a gap is penalised with a defined penalty. If the scores used are log odds ratios (see section 2.4.2.2), then it is to be expected that better alignments will have larger scores (better alignments are ones that represent 'true' alignments). Therefore, to identify the best alignment, an alignment algorithm must calculate the maximally scoring path though the matrix by incorporating as many of the positively scoring residue pairs, while avoiding penalties incurred from mismatches and gaps. An alignment of this type, which attempts to find the best alignment of all the residues from the whole length of both of the sequences, is termed a global alignment.

The scoring of residue matches and mismatches comes directly from the use of scoring matrices such as members of the BLOSUM and PAM series. However, gaps are not catered for and therefore require special treatment. In an alignment, a gap represents an insertion or deletion event that has occurred since divergence from the common ancestor, and as a consequence its occurrence must be penalised by assigning a negative

score. There are a number of models to describe the scoring of gaps, the most common of which are the linear gap penalty and the affine gap penalty (often referred to as the gap extension penalty) (Altschul, 1989, 1998). The linear gap penalty considers the cost associated with an insertion/deletion to be proportional to the length of the gap, and hence the penalty value is multiplied by the number of residues in the gap. The affine score considers the penalty for inserting/deleting to be independent of the actual number of residues inserted/deleted, and is therefore composed of a single 'gap-open' penalty and an extra length-dependent factor.

The Needleman-Wunsch (Needleman and Wunsch, 1970) algorithm is the classical example of a global alignment algorithm. The procedure that will be outlined below is a more efficient modification of this method introduced by Gotoh (1982). The essence of a DP algorithm is a recursive reliance on the solution to a smaller problem. Therefore, the alignment is constructed from the optimal alignments of smaller sub-sequences. First, a matrix for sequences x and y, of lengths I and J respectively, is constructed. This is indexed by i along sequence x, and by j along y (i.e., x_i represents the residue at position i along sequence x). Secondly, the matrix is populated recursively by calculating S(i,j), which is the highest score of the alignment between the segments $x_{1..i}$ and $y_{1..j}$. Starting at the top leftmost cell, the matrix is filled to S(I,J), moving left to right along each row, by identifying the maximum score of each of the preceding alignment possibilities. If we consider a cell S(i,j), then there are three possible ways the amino acids corresponding to this cell could constitute an extension to a previous alignment:

- x_i could align with y_j (that is the i^{th} residue in sequence x and the j^{th} residue in sequence y),
- x_i could align with a gap or
- y_j could align with a gap.

Figure 2.16 describes the calculation performed as each position is considered.

Figure 2.16: The Needleman-Wunsch global alignment calculation.

The three possible cells from which the cell S(i,j) could continue the optimal alignment.

$S_{(i-1,j-1)}$	$S_{(i,j-1)}$
$S_{(i-1,j)}$	$S_{(i,j)}$

Each of the possible paths are scored accordingly, with either the addition of the subsitution matrix lookup (for the two aligned residues) or the addition of a gap penalty. The maximally scoring path is chosen and provides the score for $S_{(i,j)}$ based on the below equation.

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + s(x_i, y_j) \\ S_{i-1,j} + g \\ S_{i,j-1} + g \end{cases}$$

The most important feature of the DP method is that all of the path decisions made during the calculation are recorded (i.e., which of the three cells provide the maximum score). Therefore, when the matrix is fully populated, the optimal alignment is obtained by following the so called traceback route from the bottom right hand corner of the matrix to the top left. An example alignment is described in figure 2.17, and traceback pointers are included as arrows.

This method, as mentioned, generates a global alignment. The traceback path starts at the bottom right and ends at the top left of the matrix. Its objective is to align as many as possible of the residues in the query sequences across their whole lengths. The generation of an alignment is accompanied by a score, which allows one to make decisions about the possibility that the two aligned sequences are related.

This procedure opens the door to database searching. A database contains many sequences that can be scored/aligned in this way with a given query sequence, but only

Figure 2.17: The Needleman-Wunsch alignment matrix

The sequences ADE and ACDE are aligned by constructing a matrix with each on opposing axes. The score for each intersection is provided by the equation in figure 2.16, by considering the three surrounding positions from which an alignment could be constructed. As each score is computed the decision taken is recorded (as arrows in the diagram). When complete the optimal alignment can be found by following the highest scoring path, through the pointers, from the last position in the matrix to the first.

	0	A 1	C 2	D 3	E 4
0	0.0	→ -1d -0.5	→ -2d -1.0	→ -3d -1.5	→ -4d -2.0
A 1	↓ -1d -0.5	↘ (0,0)+3.0 3.0	→ (1,1)-0.5 2.5	→ (1,2)-0.5 2.0	→ (1,3)-0.5 1.5
D 2	↓ -2d -1.0	↓ (1,1)-0.5 2.5	↘ (1,1)-1.0 2.0	↘ (1,2)+3.0 5.5	→ (2,3)-0.5 5.0
E 3	↓ -3d -1.5	↓ (2,1)-0.5 2.0	↘ (2,1)-1 1.5	↓ (2,3)-0.5 5.0	↘ (2,3)+3.0 8.5

The parameters used in scoring this particular alignment were: alignment of identical residues +3, non-identical residues -1 and gap penalty (g) 0.5

truly related sequences should provide large and positive scores. Therefore, by searching a database and ranking the aligned sequences by score, homologues should be identified at the high-scoring end.

2.4.3.2 Local alignment algorithms

With database searching as the objective, the score of an alignment is usually optimised to provide the best distinction between true and false sequence assignment, rather than to provide the optimal alignment. With this consideration, it is often more useful to look at smaller highly-conserved regions of sequences (the local alignment) rather than global alignments. This avoids interference from less well conserved regions of proteins, and allows for the identification of evolutionarily preserved ones. The fact that many proteins are comprised of multiple independent and distinct domains highlights the suitability of local alignment. Globally aligning two proteins that share only a single domain makes no sense; however, a database search that pulls out all sequences that share a single domain can be very informative.

A variant of the Needleman-Wunsch algorithm, which yields the optimal local alignment, is the Smith-Waterman algorithm (Smith and Waterman, 1981). Finding the optimal alignment of two sub-sequences requires two simple but significant alterations to the previously described algorithm. A matrix is constructed in the same way, but each cell can now score zero if there is no higher-scoring path available. Equation 2.1 indicates the four alternative choices to be made in calculating the score at position $S_{i,j}$.

$$S_{i,j} = \max \begin{cases} 0 \\ S_{i-1,j-1} + s(x_i, y_j) \\ S_{i-1,j} - g \\ S_{i,j-1} - g \end{cases} \tag{2.1}$$

Taking the zero scoring option corresponds to the initiation of a new alignment. This means that a sub-alignment can start at any point in the matrix, and does not have to suffer the detrimental effect (from the accumulation of a highly negative score) of any prior poorly aligned residues. The second change allows the alignment to end at any point in the matrix rather than in the final cell.

An important requirement of the scoring scheme, highlighted by the fact that the alignment can end anywhere, is that the expected score for an alignment of random sequences should be negative. If this were not the case, then longer alignments would score more highly, based merely on length, regardless of similarity. There must also be the potential to achieve a positive score from the given scoring matrix, otherwise no cell can score above zero and no alignment will be found. The log odds matrices of the PAM and BLOSUM series conform to these criteria.

Heuristic Methods

The local and global algorithms guarantee to provide the optimal alignment, and therefore constitute the most sensitive search possible. However, speed of search is a important consideration and, with increasing database size, the full DP algorithms rapidly become very time consuming. Heuristic methods (inexact or approximate approaches), while based on the general principles of the local alignment algorithm, attempt to reduce the number of cells searched in the matrix in order to increase the speed of recursive calculation of the alignment. A number of distinct methods exist that use heuristics to provide search tools able to run at high speeds. This enables large data resources to be searched in reasonable time frames. Two such methods are described below.

FASTA

FASTA (Pearson and Lipman, 1988) was developed to exploit the idea that true alignments are liable to contain short stretches of identity or very high similarity. By looking

for these regions, and using them as a starting point for extension into longer alignments, much of the search space is ignored. FASTA first identifies identical word matches between the query and the database sequence. Diagonals in the scoring matrix that contain many word matches are then selected for further attention. These exact matches are extended to identify maximally scoring ungapped alignments. The highest scoring regions are selected by joining the ungapped sub-sequence alignments together by extending a gapped region using gap extension penalties. Finally, the immediate vicinity around the highest scoring regions are realigned using full DP methods.

BLAST

BLAST (Altschul et al., 1990) takes a similar approach to FASTA in identifying local high scoring alignments from which to start a more sensitive alignment process. However, it differs in a number of minor details. A pre-processing stage identifies all potential short words[2] in the query sequence. Searching for exact matches to these short words in a large database can be achieved very quickly and with a minimum of computational effort. When a word match is identified, the alignment process is initiated extending the ungapped alignment around the word match region. BLAST may indicate numerous high-scoring, sub-sequence alignments to a single query, and a measure of significance can be determined from the combined value of their scores. A version of BLAST that provides gapped alignments is commonly used today in preference to the ungapped version (Altschul et al., 1997).

The price paid for speed is that both methods potentially miss real alignments during the initial word match stage. However, the massive growth of databases has made sacrifices like these essential.

[2]Word is used here to refer to a stretch of amino acids (e.g., PS, VEK, LCCM), just as a word conventionally refers to a string of letters of the alphabet.

2.4.4 Assessing the significance of the alignment

The basic sequence analysis task involves aligning sequences, and then asking whether the alignment is more likely to represent a relationship that occurred through descent from an common ancestor than one that arose by chance. The previous sections have discussed the method of deriving alignments from pairwise comparisons of sequences. This section will concentrate on methods that can provide a measure of statistical significance to an alignment score.

2.4.4.1 Scoring

The theory required to convert alignment scores into probability values is derived from the statistics of single sequence scoring (Altschul et al., 1994; Altschul and Gish, 1996; Dembo et al., 1994; Karlin and Altschul, 1990). This section will expand on the theory: from its use in identifying conserved stretches of residues in a single sequence, to its application in sequence alignment.

Let us consider a single sequence, within which we hope to identify a specific feature. The objective of this search would be to identify a region containing amino-acids that are characteristic of this feature. For this to be meaningful, a distinction has to be made between the true representation of this feature and a random sequence. The feature could equally represent a compositional bias, such as a hydrophobic patch, or a true region of pairwise alignment. Taking the former example, while looking for compositional bias in a protein sequence, one would wish to design a scoring scheme that detects the sub-sequence containing the largest number of biased residues. For example, the identification of tracts of hydrophobic residues may be valuable in a search for proteins containing transmembrane helices. In order that high scores identify such sub-sequences, an obvious scoring requirement would be that residues such as alanine, valine, leucine and isoleucine were scored positively. If this is to be the model, then the set of residues $\{A, V, L, I\}$ should be assigned a positive integer (ar-

bitrarily 1.0), with no other residue accruing a score. A given sequence, for example *DIIAVLCDEEGGHEED*, can be scored by selecting sub-sequences that score positively, the highest scoring of which can be termed the Maximally Scoring Subsequence (MSS). Working through the example:

- sub-sequence *DI* would score 0+1 = 1,
- *DII* = 2,
- *DIIA* = 3,
- *DIIAV* = 4,
- *DIIAVL* = 5, and so on.

However, the whole sequence in this example will also score 5, because the model assigns no detrimental effect to identifying residues not in the scoring set. This drawback means that the MSS often tends to be the whole sequence, and hence the aim of identifying the most biased *sub*-sequence fails. This simplistic scoring model, therefore, only succeeds in identifying long sub-sequences, and requires re-evaluation. While it is obvious that the amino acids in the biased set must have positive values (otherwise an additive sum of scores would never result in a positive value), it is equally clear that not assigning any score to the remainder merely allows longer sub-sequences to score highly. By assigning a penalty score to the presence of non-members, the result is a situation in which biased residues are scored positively and all other residues are scored negatively. Indeed, this method works as long as the expected score is negative (i.e., if a random sample of residues is taken, the score is negative).

The example described above outlines the principle of determining a suitable scoring regime for the identification of biased or characteristic stretches of amino acid sequence. The following section covers a generalisation of this method, which leads to the development of a statistical measure of confidence in the fact that a highly scoring region reflects a true (or false) occurrence of such a characteristic.

2.4.4.2 Probability values

Consider a random sequence composed of independently-sampled letters from an alphabet ($A = \{a_1, a_2, ...a_n\}$, where n = 20 - the amino acid alphabet), each with corresponding natural probabilities of occurrence $\{p_1, p_2, ...p_n\}$. A given sequence of these letters $\{a_3, a_7, a_2, a_{10}, a_{20}, a_{13}\}$ can be scored so as to determine its compositional bias towards any given subset of these letters, by attributing an integer s_j to each letter a_j (s_j represents the score for the observation of letter a_j). An MSS can be identified by summing all s for each sub-sequence if, and only if, one of the set of scores is positive, and the expected score ($E = \sum_{j=1}^{n} p_j s_j$) is negative. Extending the concept from the scoring of individual letters (s_j) to the scoring of pairwise alignments is simple: each score s_j can be replaced with s_{ij}, which is the score attributed to the alignment of letters i and j (from sequences I and J). The quest then becomes the identification of the best ungapped alignment, and the criteria for MSS[3] scoring still hold, i.e., the expected score $\sum_{i,j=1}^{n} p_i p_j s_{ij}$ must be negative and the alignment of at least one pair of letters must be positive.

If S is the score of the MSS from the comparison of two random sequences of lengths m and n, then it can be shown that the distribution of S, for many such random sequences, is approximated by the Extreme Value Distribution (EVD) (Dembo et al., 1994) an example of which is shown in figure 2.18.

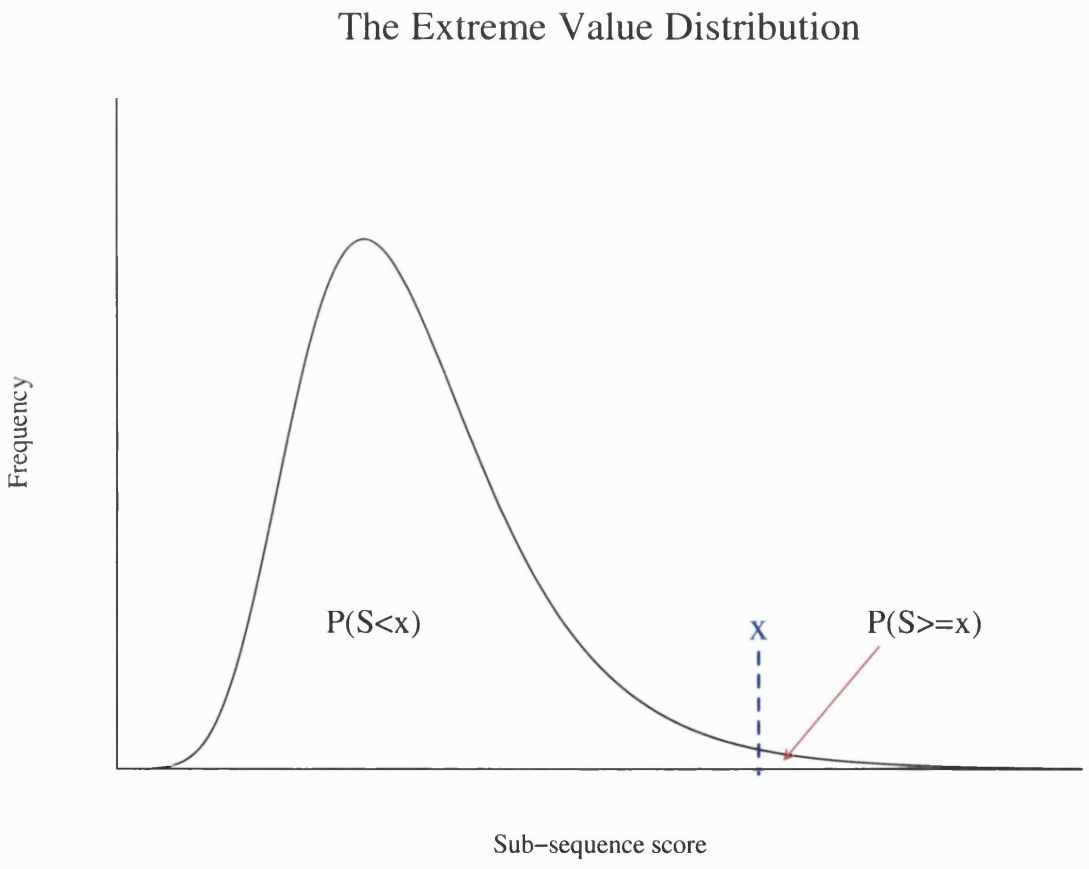
It can be demonstrated that the number of matches scoring greater than a given score x is approximately Poisson distributed. The mean of this distribution is given by equation 2.2, where μ is the average number of random sequences scoring above x, and is often known as the *pairwise* e-value or the expected number.

$$\mu = Kmne^{-\lambda x} \tag{2.2}$$

The distribution is described by two characteristic parameters: lambda and K (Karlin

[3]The Maximally Scoring Sub-sequence now becomes the Maximally Scoring Sub-alignment.

Figure 2.18: The extreme value distribution of MSS scores (S). The score of an observed MSS is represented by x.



and Altschul, 1990). The former is the unique positive solution to equation 2.3, which represents the scale parameter of the distribution (its effect being on the skew); the latter, K, is a constant that can be computed from p_i, p_j and s_{ij}.

$$\sum_{i,j=1}^{n} p_i p_j e^{\lambda s_{ij}} = 1 \tag{2.3}$$

If X is a random variable representing the number of occurrences of MSSs (from N pairs of random sequences) scoring greater than x, because μ is described by a Poisson distribution, the probability of finding exactly n such MSSs can be shown by equation 2.4.

$$P(X = n) \approx e^{-\mu} \frac{(\mu)^n}{n!} \tag{2.4}$$

Therefore, the chance of finding exactly zero such MSSs is $e^{-\mu}$, from equation 2.5:

$$P(X = 0) \approx e^{-\mu} \frac{(\mu)^0}{0!} \tag{2.5}$$

So, by extrapolation, the chance of finding at least one such alignment is one minus the chance of finding none, which gives gives equation 2.6:

$$P(X \geq 1) \approx 1 - e^{-\mu} \tag{2.6}$$

Since the definition of X is the number of MSSs scoring $\geq x$, then this equation can be combined with 2.2 and rewritten as 2.7, i.e., the pairwise p-value:

$$\text{Prob}(S \geq x) \approx 1 - e^{-Kmne^{-\lambda x}} \tag{2.7}$$

Computable from this equation is the *probability* that a score, S, derived from the alignment of two *random* sequences is *at least* x. If a score of x is attributed to the MSS of an alignment between a query sequence (of length m) and a database sequence

(of length n), the probability that a *random* sequence could score as well or better can be given as a measure of the significance of the MSS. A highly significant, non-random, score is likely to be found in a random distribution with a very low probability, and *vice versa*. It is possible to compute this pairwise probability value (p-value) from the score for any alignment given that the scoring scheme conforms to the restrictions mentioned previously.

2.4.4.3 Adjusting probabilities for database searches.

When a *database* search is performed, the p-value must be adjusted for the multiple comparisons made therein. The previously calculated p-value refers only to the comparison of two sequences. An estimate of the probability of observing *at least one* random MSS (where $S \geq x$) in a database search of D sequences is calculated as follows.

The expected number (E) of matches, where $S \geq x$, (encountered in a database search) is the probability of occurrence of one such MSS, when two sequences are compared, p, multiplied by the number of times the database is searched (i.e., the number of sequences in the database D). Therefore, $E = pD$, and from equation 2.6 the probability of observing at least one such MSS is:

$$P \approx 1 - e^{-pD} \tag{2.8}$$

This calculation makes the assumption that all sequences in a database are equally likely to be related to the query. However, it is also valid to assume that only all *equal-length* sub-sequences have a similar likelihood of being related to the query. Therefore, considering a sequence of n residues matching an equal-length sub-sequence in a database of N residues, the estimate is:

$$P \approx 1 - e^{-p\frac{N}{n}} \tag{2.9}$$

This probability is commonly confused with the e-value, due to an approximation rendering them interchangeable at low probabilities (Altschul et al., 1994). To clarify the situation, the following definitions will be made. The probability P is defined as the 'exact probability' of identifying at least one random sequence with $S \geq x$, in a database search of D sequences. The 'expected value', E, is defined as the number of random sequence matches expected to occur with $S \geq x$ in a database search of D sequences. It is E that is correctly termed the e-value.

2.5 Problems of pairwise sequence analysis

The principle process of pairwise sequence analysis involves submitting query sequences to a BLAST or FASTA (or time-permitting Smith-Waterman) search of a primary resource (sections 2.2 and 2.3). The result of this is a score-sorted list of sequences containing sub-sequences that align with a region of the query sequence. The highest scoring sequences are assumed to be related (the associated probability value provides an indication of the mathematical significance of a score), and if there is a consistency in the annotation of these top-scoring sequences, then the hypothesis that 'significant similarity implies homology' allows this annotation to be transferred to the query sequence (a diagnosis of the family membership of the query sequence).

The sequences that match in a search like this can be labelled in such a way that allows the efficacy of a result to be evaluated. Sequences that belong to the same family as the query sequence that score significantly are termed 'true positives'; while non-family member sequences that also have significant scores are termed 'false positives', as these represent false diagnoses. Sequences falling below the significance threshold are termed 'negative' diagnoses; those that belong to the query's family are 'false negatives' and all others are 'true negatives'. A perfect result would provide no 'false' assignments at all; i.e., all true members of the family score above the threshold and all others below. In reality, there is often a balance to be made between avoiding

false positives and false negatives. One search technique may result in a set of significantly scoring sequences that contain no non-family member representatives (no false positives) at the cost of missing a number of distantly related members (some false negatives). An alternative technique may identify distant relatives, but introduce many false positives. The former represents a search that would be termed *selective* while the latter describes a *sensitive* search. Neither result can be described as more effective than the other; however, the different perspectives that they represent can be used to indicate the degree of confidence that a query sequence has been correctly identified.

The highest degree of confidence comes from a *selective* result, which represents a family containing a closely related set of sequences, clearly separated from other families in sequence space. A *sensitive* result is more suited to the process of searching a database with a well known and annotated sequence in the hope of identifying previously unknown related sequences. A challenging result, which falls into neither camp, is the observation of a number of high scoring but seemingly unrelated matches to sequences that belong to different families. A common reason for this may be the existence of multiple domains in the query sequence; therefore, the 'unrelated' matches come from sequences annotated with respect to their functional properties and not to the common domain that is shared with the query sequence. Also, it is of course possible that a query sequence has no close relatives, and therefore produces no significant matches. While, algorithms and computerised searching techniques can be applied to the resolution of these complexities, it is often the case that only the clearest and most significant results can be used to make a confident diagnosis.

Another level of uncertainty can be obtained from searches performed on databases containing variable-quality data. The majority of ORFs that arise from genome sequencing projects are assigned a function *in silico* (via similarity alone) and many are only designated – 'hypothetically expressed'. The move away from the principles of functional cloning (determining function and role first, and sequence later (Boguski, 1999)), has been instrumental in increasing the speed of sequence data accumulation.

However, the detrimental consequences of the move towards the positional cloning era (sequence first and function later) is the ever widening gap between the number of available sequences and the number of those with comprehensive annotation. The development of high-throughput genome analysis has only served to widen this gulf. As the quantity of sequence data, from such sources, increases, so does the likelihood that a pairwise search will reveal matches to query sequences that are themselves unannotated or annotated only *in silico*.

The ramifications of this situation become even more profound if we consider the potentials of transitive annotation. While it may be a leap of faith to apply the experimentally derived functions and characteristics of one protein (A) to another (B), based entirely on inferred homology, this is the central tenet of sequence analysis. However, the application of such information to another sequence (C) based purely on similarity with a sequence, itself annotated via similarity (e.g., B), may lead to an unacceptable increase in incorrectly annotated sequence data (Karp, 1998).

Putting aside the problems of polluted data, it must be mentioned that similarity searches such as these, even ones that demonstrate local similarity, are tools for predicting general similarities between proteins. The drawback of an approach based on generality is simply that the process of evolution can significantly alter a protein's function through relatively minor modifications to specific regions (e.g., very few mutations in, or around, the active site of an enzyme can be responsible for significantly altering substrate specificity). Clearly, the identification of any relationship, however weak, between an unannotated sequence and a known protein may confer a degree of information (e.g., the unknown protein may belong to a related gene family). However, the specific biochemical function of the protein cannot confidently be transfered without further evidence. Despite this caveat, the use of BLAST and FASTA to identify relationships and to assign function, without careful consideration of the implications, is commonplace (Brenner, 1998, 1999).

Pairwise sequence analysis *can* provide evidence that can be used to infer homologous

relationships. However, it should not be blindly assumed that sequence similarity is a *guarantor* of functional homogeneity, since similarity alone cannot distinguish between evolutionary relationship such as *orthology* or *paralogy*. Two sequences share an orthologous relationship, if they are derived from a single protein that has diverged from its common ancestral form *following* a speciation event (Wray and Abouheif, 1998). A paralogous relationship (Fitch, 1970; Henikoff et al., 1997), is the consequence of a gene duplication resulting in two copies of a gene that evolve side by side in an organism. Gene duplications can potentially introduce a copy of a gene that is free of the restrictions of natural selection. If an unaltered copy of the gene remains, mutational events that are deleterious to the function of one copy may no longer produce a phenotypic effect (neutral mutations). This freedom may lead, by chance, to the generation of a diverse functional role, which in turn may provide its own selective advantage. The problem of interpreting the relationship between two paralogous sequences is that, while their functional divergence may have been significant, two paralogous sequences will still share a large degree of generic similarity.

A natural extension of the concept of pairwise analysis is to compare more than two sequences in the hope of revealing regions of commonality that are directly attributable to the preservation of function. Proteins sharing functional and evolutionary relationships can be grouped into families, and it is the analysis of these that is described in the following chapter.

Chapter 3

Secondary Databases

To negate some of the problems associated with the analysis of sequence data, as outlined in the previous section, requires the investment of great deal of time and effort. However, collating individual sequences into families can have beneficial effects above and beyond merely addressing the problems of pairwise sequence analysis:

- The collective knowledge held by a set of proteins is clearly more comprehensive than that available for a single member, and this can compensate for any omissions or errors found in the annotation of individual members.
- The observation of a diverse set of orthologous members of a protein family can provide an insight into the common properties of the family and may, by inference, describe features of their shared ancestry.

3.1 Gene families

A gene family represents a set of proteins that are related via descent from a common ancestor; i.e., those sequences that share a homologous relationship (See section 2.4). Just as the process of pairwise alignment allows conserved regions between two sequences to be identified, the similar procedure of generating a Multiple Sequence Alignment (MSA) can highlight conservations apparent in many members of a gene family. In an MSA, sequences are aligned relative to the alignment itself rather than any given sequence. Gaps are introduced so that regions of conservation are brought into line: an insertion in one member will produce a corresponding column of gaps to be introduced into each of the other sequences so as to preserve the alignment of the other residues. The signature of conserved and divergent regions in an MSA represents the evolutionary history of the protein family: highly mutated regions average out into non informative sections; while structurally or functionally important regions, which are less tolerant to mutation, stand out as islands of calm. An analogy can be drawn between the MSA and a photograph taken with an extended exposure time. In such

a picture, incidental or fast moving objects are seen as averaged blurs, while integral static objects appear clarified in comparison (figure 3.1).

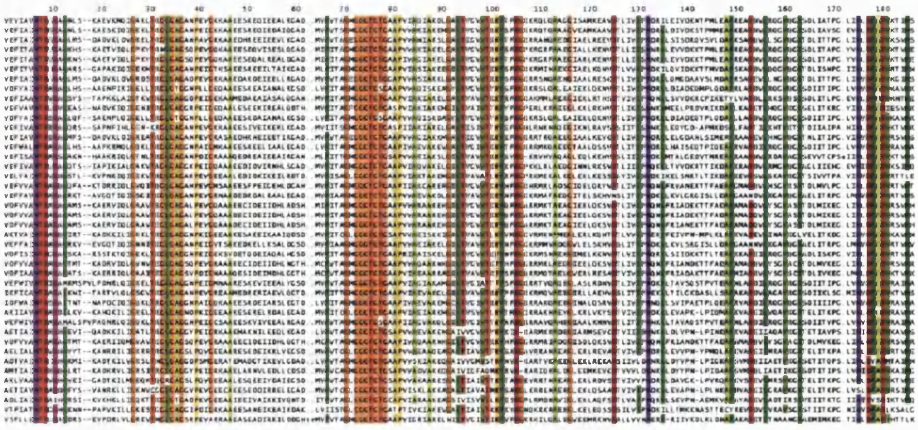
Reconstructing the evolution of a gene family is not trivial: one must attempt to infer information about the ancestry of a protein from a limited number of, usually biased, samples of modern organisms. Mutational events during replication and cell division are rare, especially in higher order eukaryotes. Nevertheless, the time-scales over which comparisons can be made are often huge. It is common to attempt to identify the similarities between proteins from, for example, *Homo sapiens*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*, which represents an ancestral relationship dating back thousands of millions of years. While not all proteins undergo mutational change at the same rate, generally the accumulation of change over such a time period is not limited merely to a few point mutations.

An important point should be made about the concept of a gene family. The above definition is complicated by the phenomenon of modularity: a commonly observed occurrence involving independently evolved domains becoming fused into a single protein. In such a protein, a global MSA demonstrates quite a different evolutionary story to a local alignment of either of the two domains. Therefore, when referring to a gene family, one must be careful to ensure that the comparison is really being made between proteins that share the same ancestor. The definition of a family can therefore be extended to include situations in which quite unrelated proteins share a modular domain (e.g., SRC homology domains), as long as the alignment is only made within the boundaries of the domain.

The principles and process behind the construction of an MSA will be outlined in the following section.

Figure 3.1: Long exposure photography is analogous to multiple sequence alignment.

(a) Taking a photograph over an extended time period captures the relationship between objects and time: transitory events are represented by ghost-like trails on the image; while the clarity of fixed immovable entities is emphasised.



(b) An alignment of bacterial cell-division proteins (ftsZ). The alignment represents the sampling of discrete events over evolutionary time and combining them in a single image. Evolutionarily conserved regions are emphasised; whereas sections containing multiple mutations seem undefined and indistinct. The colouration serves to highlight those residues that are well conserved over the whole alignment.

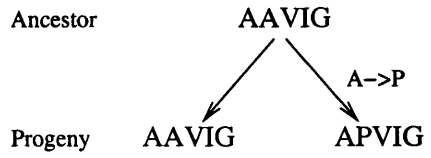
3.2 The Multiple Sequence Alignment

An alignment of sequences from a gene family can be used to reconstruct evolutionary events. It is possible to demonstrate the reconstruction by starting from the perspective of the ancestral sequence. Consider the pairwise comparison of two proteins (A and B) that have only recently diverged:

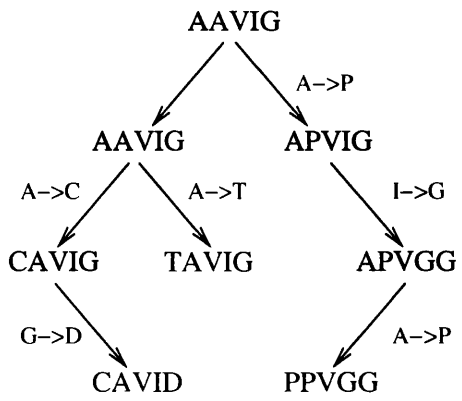
```
Sequence A= AAVIG

Sequence B= APVIG
```

Their inherent relationship can be described by assigning an arbitrary common ancestor; which, as this departure only represents the accumulation of a single point mutation, can be either of the sequences. Therefore the relationship can be described as follows:

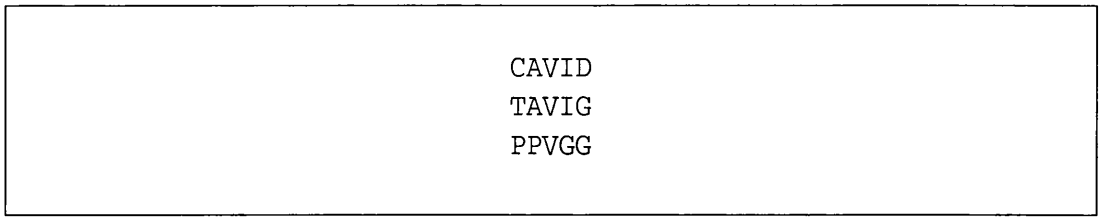


Any subsequent mutations can also be placed in the context of this relationship:



Each additional sequence represents the smallest possible step, i.e., the accumulation of a single point mutation. From an alignment of only those sequences on the tips of the branches, the most modern of each lineage, it is possible to infer the common ancestral sequence (figure 3.2).

Figure 3.2: An alignment of the most modern sequences of each lineage.



Most clearly the ancestor must have been a sequence with valine in the third position. On closer inspection, it is likely that alanine may be found in position 2, isoleucine in 4 and glycine in 5. The consensus is x-A-V-I-G, which, considering the original sequence was AAVIG, is a very good approximation. This simplistic example considers sequences that have not diverged significantly, and therefore the answer was not difficult to obtain. In a real MSA, significantly divergent sequences not only create greater ambiguity in the assignment of the consensus but also make the process of initially aligning the sequences complex. However, the greater the number of samples that are taken into account, the better the reconstruction method performs.

Obtaining an MSA from a set of sequences assumed to belong to a family is not a trivial exercise, and, as a consequence, a multitude of automatic and manual tools exist to facilitate this process.

3.2.1 Creating a multiple sequence alignment.

The starting point for the analysis of a family of related sequences is the creation of a set of members. The set should be as representative and divergent as possible. An alignment editor (SOMAP (Parry-Smith and Attwood, 1991), SEAVIEW (Galtier et al., 1996), CLUSTALX (Thompson et al., 1997), CINEMA (Parry-Smith et al., 1998)) can be used to manually place this set of sequences into the context of the alignment. Sequences are then moved, slid laterally, until residues appear to align, inserting gaps where necessary to model insertion or deletions. This process is repeated for each member of the group until an optimal arrangement is found that appears to re-

flect the mutational history of the family (figure 3.3 shows a simplified representation of an alignment, while figure 3.4 shows an example of a real alignment.).

Figure 3.3: A graphical representation of the alignment process.

The following pictures illustrate the process of aligning members of a gene family (a). As an intermediary step pairwise alignments between sequences sharing significant identity can be performed (b), and then these can be used to direct subsequent alignments between the pairs (c)

a) Four sequences (unaligned).

b) Pairs of sequences sharing high identity are aligned.

c) The result consists of an alignment between the two pairs.

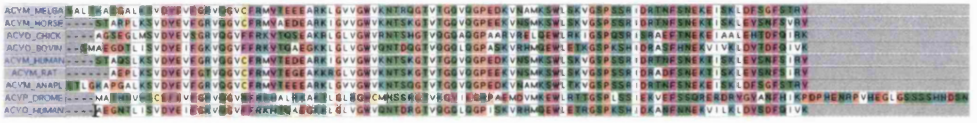
Automation of this procedure suffers from similar problems to those faced by pairwise alignment algorithms: the time taken to achieve an optimal result is dependent on the lengths of the sequences aligned (due to the large numbers of calculations required). In an MSA, this problem is further exacerbated by the need to align more than two sequences. A solution to the pairwise alignment problem (section 2.4.3.2) is to reduce the number of initial observations that are considered in an ensuing alignment, the objective being to increase the speed of the procedure with little reduction in the accuracy of the alignment. Analogously, most MSA algorithms rely on a recursive process of aligning closely related sequences, followed by subsequent alignment of the resultant alignments. This speeds up the process at the cost of fixing alignments at an early stage and ignoring information learned during the remainder of the steps (Thompson et al.,

1994, 1997).

Figure 3.4: The acylphosphatase family unaligned and aligned (PRINTS: ACYLPHPHTASE).

This family of proteins catalyses the hydrolysis of the carboxyl-phosphate bond of acylphosphates, and are found in organisms as diverse as *Homo sapiens* and *Drosophila melanogaster*.

(a) The unaligned sequences of the acylphosphatase family.



(b) The sequences of the same family aligned by shifting members left and right.

3.3 Multiple Sequence Analysis

An MSA, as described in the previous section, can be used to describe the evolutionary relationship between members of a family; for example, as illustrated in figure 3.5, sequence A shares a closer common ancestor with sequence B than with C. A description such as this can form the basis of a phylogenetic analysis.

The construction of an MSA often leads to the identification of specific regions of conservation that are shared between all sequences in the family, which correspond to structural or functional elements of the protein (figure 3.6). These regions or motifs[1] (variously called features or blocks) accordingly represent preservation of function across the population as it diverges from an ancestral organism, the assumption being

[1] A motif is often described as an ungapped sub-alignment.

Figure 3.5: The ancestral relationships of three sequences (A, B and C)

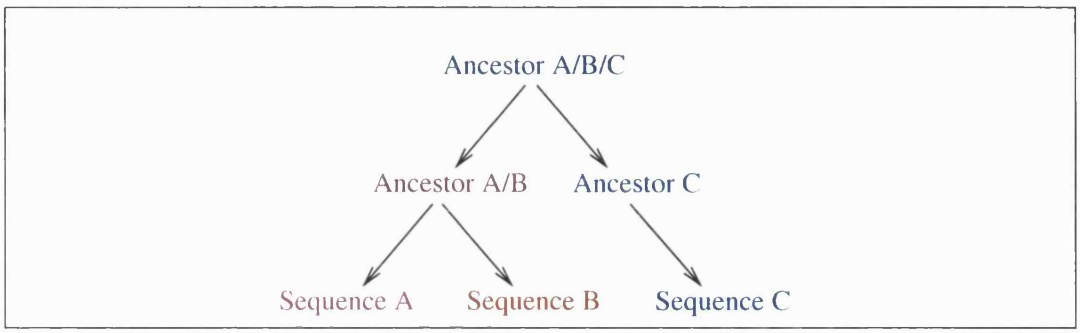
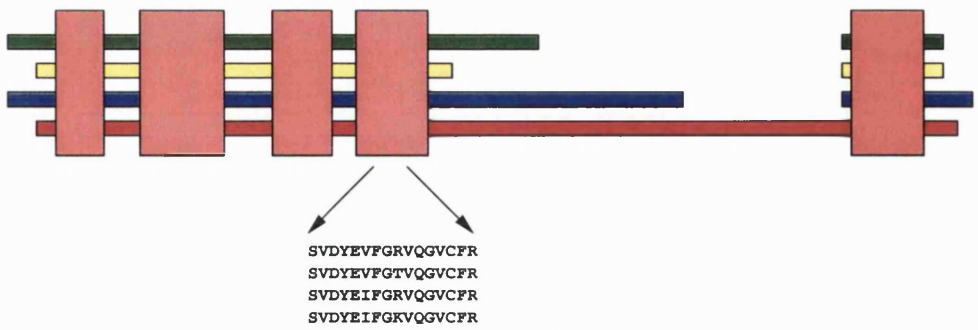


Figure 3.6: An MSA can lead to the identification of functionally preserved motifs

The alignment of the four sequences of figure 3.3 could reveal strongly conserved motifs, which have not accumulated insertions or deletions. An example motif from the acylphosphatase family is shown alongside one of the regions (from the alignment in figure 3.4).



that mutations occurring within these motifs were not tolerated and are therefore not observed in the descendants. An obvious hypothesis is that the essence of the family (i.e., what defines the physical properties of the family) is somehow encoded in these regions. Therefore, if the demonstration of significant similarity between *two* sequences has the ability to confer membership on unannotated proteins, then the MSA must also be able to confer such information. Indeed, as the MSA represents the cumulative history of many sequences, the description of a relationship is more specific than that obtained via a pairwise comparison.

The following sections detail a number of different methods and models that aim to describe a family using an MSA as a starting point. Once a family is described, the resultant descriptor can be used to identify regions of similarity in unannotated sequences. This process is analogous to the pairwise analysis procedures outlined in the previous chapter.

3.3.1 Scanning a sequence

The scanning of a sequence is the process by which a linear string of amino acids is sequentially searched for the existence of a pattern. Each of the following methods produce models that describe either the whole MSA or a representative section of it. A model can be used to search for an alignment between itself and a query sequence. As with pairwise alignment, the demonstration of significant similarity can be used to infer a homologous relationship. The simplistic sliding window approach is commonly used for most models of this type, the exception being the Profile and Profile-Hidden Markov Model, which will be discussed later (sections 3.3.7.1 and 3.3.7.2).

3.3.1.1 The sliding window

The scanning process can be interpreted as the sliding of a fixed-width window along the length of a query sequence. At each position, the window reveals a sub-sequence

(figure 3.7) with which the model can be compared and scored. Scoring such an align-

Figure 3.7: Scanning a sequence with a fixed window.

As the window slides over the sequence it reveals each subsequent overlapping region.

ment is dependent on the methods used to describe the model. The following sections will describe the properties of some of the methods and the scoring schemes that are used to reveal similarity.

3.3.2 The Motif

At the simplest end of the spectrum are those methods that concentrate on the single-most informative region of a family of proteins. This can be defined as a portion of sequence that all members must share, i.e., the most characteristic motif of a family. For example, a suitable motif could be the active site of a catalytic enzyme: the consequences of mutations within this motif may be functionally disastrous, and therefore, it is likely to be well conserved. A motif is a short stretch of residues selected from each of a number of sequences. Its properties include its length and depth; i.e., the number of residues it contains and the number of sequences in which it is observed respectively (figure 3.8).

Figure 3.8: A motif, including notation that describes its generalisation.

A motif is a stack of amino acid strings.

```
ACDEFGH
ACDEKGH
ACRDKGH
ACEERGH
```

It can be generalised as:

$$\begin{array}{l} aa_{1,1}\ aa_{1,2}\ aa_{1,3}\ \cdots\ aa_{1,m} \\ aa_{2,1}\ aa_{2,2}\ aa_{2,3}\ \cdots\ aa_{2,m} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ aa_{n,1}\ aa_{n,2}\ aa_{n,3}\ \cdots\ aa_{n,m} \end{array}$$

Where $aa_{i,j}$ is the amino acid in sequence i, and at position (column) j.

3.3.3 Selecting a motif

As with manual multiple sequence alignment, manual motif selection is facilitated by the use of alignment editors (Galtier et al., 1996; Parry-Smith and Attwood, 1991; Parry-Smith et al., 1998; Thompson et al., 1997). Particular note should be made of colouring schemes, which in this context prove to be invaluable assets. A colouring scheme reflects concepts of sequence similarity (discussed previously, section 2.4.1.2) such that residues that share physical and chemical properties are grouped and coloured similarly (figures 2.8, 2.9 and 2.10). The manual selection of motifs is a subjective process. The use of colouring schemes in alignment editors facilitates the identification of regions that show significant conservation by highlighting the distinction between conservative mutations and mutational drift. As expert identification of motifs is a time-consuming and laborious task, a number of automated methods have been developed that reduce the level of human intervention (Brocchieri and Karlin, 1998; Depiereux and Feytmans, 1992; Henikoff et al., 1995; Smith and Smith, 1992).

Once a motif is extracted, the issue of how it is used in a similarity search is raised.

Motifs are usually encoded so as to provide a suitable score, with which the significance of an alignment can be determined.

3.3.4 Encoding a motif

It can be cumbersome to describe a motif in the way shown in figure 3.8, so a number of simplified representations have been developed to model the motif. A totally conserved motif may be described as the sequence of residues of which it is comprised; e.g., DALIR, which indicates that all sequences in the family contain these amino acids, in that order, at some position along their length. However, total conservation is rare, and it is much more likely that a motif may contain sequences that deviate in one or more positions while maintaining a general pattern (figure 3.8).

3.3.4.1 The regular expression

There are a number of ways of describing a motif without having to detail each member subsequence. A Regular Expression (RE) may be composed that formulates the rules for residues appearing at the various positions of the motif. The motif in figure 3.9 can be described as D-A-[LA]-[IA]-R. In positions 3 and 4 of the RE, the ambiguity is represented by the use of square brackets: in these positions, either of the included residues could have been observed. If, within a motif, a column contains one or more residues that do not conform to any grouping (such as those described in section 2.4.1.2), then the column can effectively be ignored by replacing the allowed list of residues for that column with an x (figure 3.10).

Also included in the RE formalism is the ability to define the exclusion of a particular residue, thereby indicating what is allowed and/or not allowed at a particular position. This feature enables closer modelling of the motif. For example, a motif modeled by D-A-{P}-[LI]-R cannot contain a proline residue in the third column. Proline is an amino acid that significantly affects the conformation of the protein backbone, and

Figure 3.9: A simple motif, and its RE.

motif	D		A		L		I		R
	D		A		A		I		R
	D		A		L		A		R
RE	D	-	A	-	[LA]	-	[IA]	-	R

Figure 3.10: A simple motif with a single divergent position, and its RE.

motif	D	A	L	A	R
	D	A	L	I	R
	D	A	L	Y	R
RE	D	A	L	x	R

certain protein secondary structural elements may not form in its presence. Therefore, its exclusion may be strongly indicative of a particular motif.

A longer RE is less likely to be observed in a random, or non-family, sequence. For example, the short expression D-I-V-L is found in 2,028 sequences of the 260,981 sequences in the OWL sequence database (Bleasby et al., 1994, release 30.3), D-I-V-L-P in 104 and D-I-V-L-P-L in only four. However, an expression that contains flexible positions is more prone to ambiguity, D-[AVI]-V-[AVI]-P is found 657 and the longer D-[AVI]-V-[AVI]-P-[AVI] is found in 126. Therefore, care must be taken to ensure that the pattern derived from an MSA is specific enough to distinguish family members from non-family members.

An extension of this method relies on establishing 'groups' of residues: based on shared physical and chemical properties; e.g., the group of hydrophobic amino acids "IVLAM" can be described as a single token, *h*, which represent the whole group. This allows a familial signature to be defined as a pattern of group types rather than the rigid encapsulations of REs; e.g., D-[AVI]-V-[AVI]-P may be better described as D-h-V-h-P, as this pattern tolerates *any* hydrophobic residue in positions 2 and 4. This approach

is supported by the fact that these residues share similar properties. So, although the sequences selected to represent the family provide no evidence for the existence of the residues L or M, they are residues frequently found to substitute for those in the observed set (see section 2.4.2.2). The added flexibility means that, rather than being rejected due to incomplete sampling of the family, a previously unidentified member may be welcomed into the family; e.g., D-[AVI]-V-[AVI]-P does not identify the sub-sequence DLVLP even though leucine is very similar in nature to alanine, valine and isoleucine. However, the gain in sensitivity made by this approach may be offset by a loss of selectivity: a pattern may become too unspecific and therefore matches sequences that are not related to the family from which it was derived.

A RE is not usually scored against a sub-sequence to produce an alignment score: it merely matches a sub-sequence or not. This apparent simplicity means that searching a database of sequences with patterns like these can be very fast.

3.3.4.2 The frequency matrix.

The frequency matrix is a simple matrix that can be used to describe a motif, and to generate a similarity score for the alignment of a sub-sequence with the motif. Its construction relies on considering each column of the alignment as a single entity, and literally counting the number of amino acids of each type, such that a column of observed frequencies is built. Using the motif representation shown in figure 3.8, equation 3.1 can be used to create a normalised frequency matrix (figure 3.11).

$$F_a = \frac{\sum_{i=1}^{N} \delta_a}{N} \qquad \delta_a = \begin{cases} 1 & \text{if } aa_{i,c} = \text{residue}_a \\ 0 & \text{if } aa_{i,c} \neq \text{residue}_a \end{cases} \tag{3.1}$$

In this calculation, c is the chosen column, a is an amino acid taken from the alphabet of 20 amino acids (A..Y), i is a sequence from the N sequences in the alignment, δ_a is a function returning 1 if a is found in sequence i (in the column in question), and

0 otherwise, and $aa_{i,c}$ is a lookup which identifies the residue at row (sequence) i and column c of the motif.

The frequency matrix is used to score each window of the sequence in the attempt to identify the region of commonality. The scoring process and subsequent steps are detailed below, but first an alternative model for the description of the motif is discussed.

3.3.4.3 The 'profile' motif.

Known variously as a profile (Gribskov et al., 1990), a Position Specific Scoring Matrix (PSSM) (Henikoff et al., 1990) or a weight matrix, this model is constructed from the aligned sub-sequences of a motif using a substitution matrix to provide scores for both observed amino acids and unobserved ones. The application of the 'Gribskov profile' (Gribskov et al., 1987), to the description of a motif requires a minor simplification. A 'Gribskov profile' is a scoring matrix composed of a number of columns, corresponding to allowed sequence tokens (the alphabet of amino acids, or nucleotides, *and* gaps) and a number of rows, corresponding to positions along the alignment. The method includes scores associated with gaps, which highlights the fact that the aim of a profile is to encode more than just the most conserved regions of the alignment. The simplification made for 'motif profiles' is that gap-scoring features are not required.

The 'profile' is constructed from a motif, like the one shown in figure 3.8. For each member of the alphabet of amino acids, the frequency of occurrence in an aligned position is calculated. This produces the frequency matrix, as described in section 3.3.4.2 and figure 3.11. The subsequent calculation, shown in equation 3.2, based on the normalised frequency matrix (taking each element from the matrix; i.e., F_a from equation 3.1) and comparison with a substitution matrix (Dayhoff, 1978; Henikoff and Henikoff, 1992), computes the *score* attributed to the occurrence of each amino acid type (r) at every position (c) in the matrix:

Figure 3.11: A frequency matrix, and its normalised form, based on the example in figure 3.8. Frequencies for the occurrence of each residue in each column are normalised for the number of sequences.

The frequency matrix

rows(r) \ columns(c)	0	1	2	3	4	5	6
A	4	0	0	0	0	0	0
C	0	4	0	0	0	0	0
D	0	0	2	1	0	0	0
E	0	0	1	3	0	0	0
F	0	0	0	0	1	0	0
G	0	0	0	0	0	4	0
H	0	0	0	0	0	0	4
K	0	0	0	0	2	0	0
R	0	0	1	0	1	0	0

The normalised frequency matrix

rows(r) \ columns(c)	0	1	2	3	4	5	6
A	1.0	0	0	0	0	0	0
C	0	1.0	0	0	0	0	0
D	0	0	0.5	0.25	0	0	0
E	0	0	0.25	0.75	0	0	0
F	0	0	0	0	0.25	0	0
G	0	0	0	0	0	1.0	0
H	0	0	0	0	0	0	1.0
K	0	0	0	0	0.5	0	0
R	0	0	0.25	0	0.25	0	0

The alphabet used {A,C,D,E,F,G,H,K,R} represents the subset of the available amino acid alphabet observed in the example motif. In a complete example, all 20 letters of the amino acid alphabet would be represented.

$$\text{Prof}_{(r,c)} = \sum_{a=1}^{20} F_a \text{Subs}(\text{aa}_a, \text{aa}_r) \tag{3.2}$$

This produces a profile (figure 3.12), with *rows* corresponding to members of the alphabet and *columns* representing the positions across the motif, which can be used in the same way as the frequency matrix to provide a score for each sub-sequence.

In the calculation of a profile, the normalised frequency matrix can be populated with weighted frequencies of residues for each position. The weighting function can be used to compensate for redundancy in the motif. When a motif contains many closely related sequences, the weighting function can be used to give the rarer, more distantly related sequences a more balanced representation (equation 3.3).

$$WF_a = \frac{\sum_{i=1}^{N} w_i \delta_a}{\sum_{i=1}^{N} w_i} \qquad \delta_a = \begin{cases} 1 & \text{if } aa_{i,c} = \text{residue}_a \\ 0 & \text{if } aa_{i,c} \neq \text{residue}_a \end{cases} \tag{3.3}$$

In this calculation, c is the column chosen, a is an amino acid taken from the alphabet of 20 amino acids (A..Y), i is a sequence from the N of the alignment, the factor w_i is the weight of i, δ_a is a function returning 1 if a is found in i (in the column in question), and 0 otherwise, and $aa_{i,c}$ is a lookup which identifies the residue at row (sequence) i and column c of the motif.

3.3.5 Motif Scoring

As the sliding window reveals each sub-sequence, the scoring matrices are used to provide a score for each of the matching residues, the total score being the sum of the individual matches between positions of the sub-sequence and the matrix. The differences between the scoring procedures, as well as the actual scores produced and their significances, are discussed in the following sub-sections.

Figure 3.12: An example motif, frequency matrix and profile matrix (derived) from the PRINTS database.

The motif (PRINTS:5HT1BRECEPTR, motif 1) is the first motif in the 5HT1B receptor fingerprint. It is composed of 7 sequences and covers a region of 12 residues.

1	2	3	4	5	6	7	8	9	10	11	12
E	*E*	*Q*	*G*	*I*	*Q*	*C*	*A*	*P*	*P*	*P*	*P*
E	*E*	*Q*	*G*	*I*	*Q*	*C*	*A*	*P*	*P*	*P*	*P*
E	*E*	*Q*	*G*	*I*	*Q*	*C*	*A*	*P*	*P*	*P*	*P*
E	*E*	*P*	*G*	*A*	*Q*	*C*	*A*	*P*	*P*	*P*	*P*
E	*E*	*P*	*G*	*A*	*Q*	*C*	*A*	*P*	*P*	*L*	*A*
E	*E*	*P*	*G*	*A*	*R*	*C*	*A*	*P*	*P*	*P*	*P*
E	*Q*	*P*	*S*	*R*	*L*	*C*	*S*	*P*	*P*	*A*	*S*

The frequency matrix composed from the above motif.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	3	0	0	6	0	0	1	1
B	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	7	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0	0	0
E	7	6	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	6	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	3	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	1	0	0	0	0	1	0
M	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	4	0	0	0	0	0	7	7	5	5
Q	0	1	3	0	0	5	0	0	0	0	0	0
R	0	0	0	0	1	1	0	0	0	0	0	0
S	0	0	0	1	0	0	0	1	0	0	0	1
T	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0

The profile is constructed from the frequency matrix and the BLOSUM 62 (Henikoff and Henikoff, 1992) substitution matrix.

	1	2	3	4	5	6	7	8	9	10	11	12
A	−10	−10	−10	1	11	−9	0	35	−10	−10	−2	0
B	10	8	−11	−8	−22	−7	−30	−17	−20	−20	−22	−17
C	−40	−38	−30	−27	−8	−27	90	−1	−30	−30	−22	−22
D	20	17	−5	−8	−24	−8	−30	−17	−10	−10	−15	−10
E	50	45	2	−17	−17	9	−40	−8	−10	−10	−12	−8
F	−30	−30	−35	−28	−12	−25	−20	−20	−40	−40	−31	−34
G	−20	−20	−20	51	−20	−22	−30	0	−20	−20	−20	−14
H	0	0	−11	−18	−21	−4	−30	−18	−20	−20	−21	−18
I	−30	−30	−30	−37	8	−22	−10	−11	−30	−30	−19	−25
K	10	10	−1	−17	−14	7	−30	−8	−10	−10	−11	−8
L	−30	−28	−25	−37	1	−11	−10	−11	−30	−30	−17	−25
M	−20	−17	−11	−27	−1	1	−10	−10	−20	−20	−12	−17
N	0	0	−11	1	−21	−4	−30	−15	−20	−20	−21	−15
P	−10	−10	35	−18	−20	−14	−30	−10	70	70	44	47
Q	20	24	15	−17	−15	34	−30	−8	−10	−10	−11	−8
R	0	1	−7	−18	−10	11	−30	−10	−20	−20	−18	−17
S	0	0	−5	5	−5	−4	−10	14	−10	−10	−8	0
T	−10	−10	−10	−15	−5	−9	−10	1	−10	−10	−8	−5
V	−20	−20	−20	−28	8	−17	−10	−2	−20	−20	−12	−17
W	−30	−28	−31	−21	−30	−21	−20	−30	−40	−40	−35	−37
X	−10	−10	−15	−8	−5	−9	−20	0	−20	−20	−15	−14
Y	−20	−18	−21	−28	−15	−11	−20	−20	−30	−30	−25	−27
Z	40	38	7	−17	−17	17	−30	−8	−10	−10	−12	−8

3.3.5.1 Scoring a sequence - the frequency matrix

A 'character-match' occurs when a residue in the sub-sequence is cross-matched with one in the normalised frequency matrix in the correct column. Correspondingly, a 'character-mismatch' occurs if a non-scoring, zero-valued element is revealed; i.e., the residue does not reside in the column. The sum of each of the character-matches represents the value given to the total score for that sub-sequence. The sub-sequence and its associated score can be termed the 'motif-match' or simply the match; e.g., figure 3.13 shows three sub-sequences of the query being scored against an example motif.

Figure 3.13: Scoring the query sequence "SACDEKGHI" against the normalised frequency matrix of figure 3.11.

Each sub-sequence, revealed by the sliding window, is scored against the frequency matrix by summing the matches made between residues and columns of the matrix. The sub-sequences of SACDEKGHI are SACDEKG, ACDEKGH and CDEKGHI and are shown on rows 1, 2 and 3 respectively, and the summed scores are shown on rows 4, 5 and 6 respectively.

1)	S	A	C	D	E	K	G
2)	A	C	D	E	K	G	H
3)	C	D	E	K	G	H	I

$A_{1.00}$	$C_{1.00}$	$D_{0.50}$	$D_{0.25}$	$F_{0.25}$	$G_{1.00}$	$H_{1.00}$
		$E_{0.25}$	$E_{0.75}$	$K_{0.50}$		
		$R_{0.25}$		$R_{0.25}$		

4)	0	+	0	+	0	+	0.25	+	0	+	0	+	0	= 0.25
5)	1	+	1	+	0.5	+	0.75	+	0.5	+	1	+	1	= 5.75
6)	0	+	0	+	0.25	+	0	+	0	+	0	+	0	= 0.25

The scores in each column represent the percentage occurrence of each residue, hence by summing these scores across all columns a second normalisation is requires to return the total score for a motif to a comparable form. The summed score is divided by the number of columns to produce the final normalised score, which is similar in nature to the measure of identity provided by the percentage identity to the alignment between two sequences. However, as this score is conceptually different to the pair-

wise percentage identity, it will be referred to as the Weighted Percentage IDentity (W-PID).

W-PID scores of the sub-sequences in figure 3.13 (3% [2], 82% and 3% respectively), indicate the likelihood of the match being correct. One in particular infers the identification of the *true* position of the motif in this sequence. The W-PID score has the advantage of making the score given to a motif match directly comparable to all other matches, which, in turn, facilitates the discrimination between true and false matches. Using these values, a user is able to make a judgment based on the relative magnitudes of two scores and thus differentiate between them: the difference between a match scoring 95% and 5% is immediately obvious. However, not all matches are this clear cut; indeed, while scores of matches in the 0-15% range are usually false, those between 15% and 30% are notoriously difficult to judge (the 'Twilight Zone': Doolittle (1986)).

3.3.5.2 Scoring a sequence - the 'profile' motif

The 'profile' motif, like the frequency matrix, provides a look-up table to score the occurrence of each residue type in each position of the alignment between the matrix and the sub-sequence. However, unlike percentage identity scores, profile scores are not directly comparable in their raw state. These scores are utilised because they conform to a set of conditions that facilitate the use of a statistical model (section 2.4.4.2). The use of the profile scoring method can therefore provide, for each match to each motif, a score and importantly a probability value describing the mathematical significance of that score.

[2] 0.25 / 7 * 100 = 3% (There are seven columns in the scoring matrix)

3.3.6 Multiple Motifs

As motifs represent local, rather than global features, when an MSA is constructed, multiple regions of conservation are often apparent. Therefore, a natural extension of the single motif methods is one, in which many motifs are selected to describe the family.

The use of more than one region inherently endows the model with greater information content. As mentioned previously (section 3.3.2), in a similarity search, longer patterns are less likely to be observed by chance. Therefore, in a search, the combined evidence of matches to a number of discrete motifs may significantly increase selectivity merely by increasing the length of the pattern. An increase in the number of described regions also has an effect on sensitivity (the ability to detect weaker relationships). The use of a single motif provides a simple black and white diagnosis: a sequence either contains a motif or not. However, biological sequences do not usually exhibit such simple relationships. The consequence of an inflexible motif (e.g., a RE) is that it may not match a sequence merely on the basis of a single unforeseen mutation. As a sequence accumulates more and more mutations, even a motif encoded as a frequency matrix or a profile may fail to match it. Therefore, using only a single motif, such divergent sequences may go unidentified. However, using multiple motifs means that levels of stringency *below* the complete match of a particular pattern can be used to indicate family membership. Thereby allowing more deviant sequences to be identified as a member of the family: e.g., a sequence matching five of the six motifs that define a family is very likely to be related. The consequence is that multiple motifs can both add sensitivity and improve selectivity.

There are a variety of implementations of the concept of family identification via multiple motifs (Attwood and Beck, 1994; Grundy et al., 1997; Henikoff et al., 1995). All build upon the MSA and rely on the selection of conserved blocks of alignment; where they vary is in both the procedure of motif selection and the encoding of motifs. One

of the problems that faces the use of multiple motifs is that the method, like the single motif methods, discards information from the alignment, while this loss of information is on a lesser scale, it is nevertheless obvious when attempting to describe very distant relationships.

3.3.7 Whole Alignments

In the natural progression from single to multiple motifs, the next step is to encode the whole alignment, without discarding any sequence information. Two distinct approaches are commonly used to generate familial models from full, or modified MSAs: Profiles (Gribskov et al., 1990) and Profile-Hidden Markov Models (Profile-HMMs) (Eddy, 1996, 1998), the latter being a modification of the former, in which the probabilistic modelling techniques of the Hidden Markov Models (HMMs) have been used to alleviate some of the problems inherent in the Profile methods.

3.3.7.1 Profile Methodology

The encoding of a whole alignment requires some consideration of the scoring or penalising of gaps. The 'Gribskov profile' method (Gribskov et al., 1990, 1987) (discussed in section 3.3.4.3) utilises a system of gap penalty multipliers that rely on the length of a gap to assign two penalty values (one for starting the gap and one for scoring each subsequent gap character, which is similar in concept to the affine gap calculation (discussed in section 2.4.3.1)). The matrix of figure 3.12 is modified slightly to include two extra rows that represent the gap penalty calculation. Alignments between sequences and profiles are scored and generated using DP algorithms (Gribskov et al., 1990), which are necessitated by the requirement to identify gapped regions.

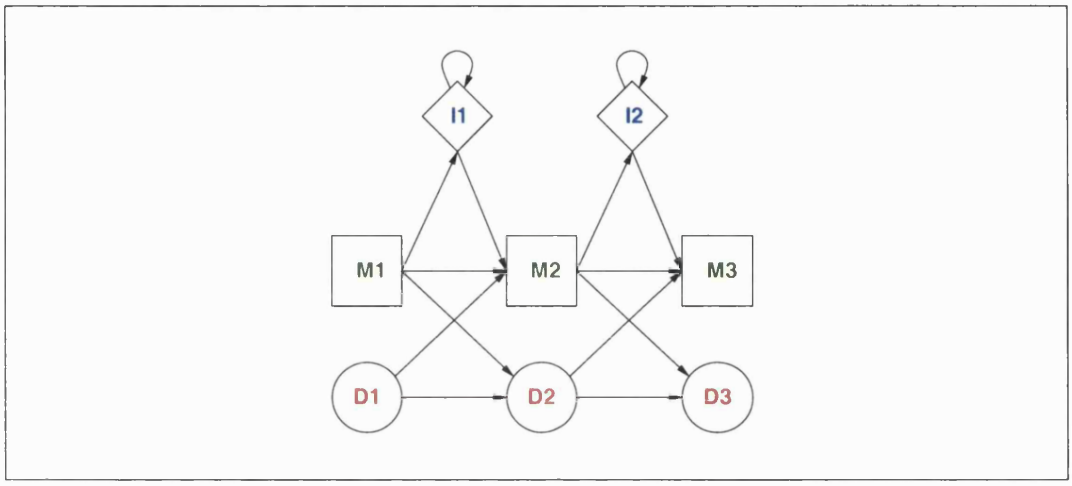
3.3.7.2 Profile-Hidden Markov Models (Profile-HMMs)

One of the drawbacks of the profile methods are their reliance on the use of *ad hoc* gap scoring schemes: while, a coherent statistical theory has been developed to describe the ungapped sequence alignment, the scoring of gapped alignments rely on empirical estimates of this theory. Driven by this, techniques based on the mathematical modelling techniques of the HMM methodology, have recently been introduced to sequence analysis.

HMMs are general probabilistic models that are applicable to the solution of linear problems, e.g., sequences of events or objects, etc. In an MSA, the columns of an alignment are linear events, which can be described by minimal connectivity between neighbouring states. Consider two neighbouring residues in a sequence matching a number of columns of a profile; if residue a matches column n then residue $a+1$ can only do one of three things: match column $n+1$; insert before $n+1$ or skip column $n+1$ altogether. A Profile-HMM constructed from these three states, Match (M), Insertion (I) and Deletion (D), can accordingly describe all possible events in the alignment of a sequence to a profile. Each match state emits symbols (residues) in accordance with emission probabilities (computed from observed residue frequencies and substitution probabilities), and each state is interconnected with transition probabilities (see figure 3.14).

It is possible to model a single sequence using a Profile-HMM; however, it is uninteresting, being merely a collection of match states strung together. Insert a second sequence into the model, and I and D states become populated, as insertion and deletions are made relative to the initial sequence allowing for the alignment of the second. As this process is continued, the overall picture of the MSA is created, with each transition between states being calculated from the addition of sequence data. Rather than the *ad hoc* selection of gap penalties in the profile method, this model extracts probabilities for gaps from the alignment itself, thus creating a more contextual score/penalty for

Figure 3.14: The Profile-HMM is characterised by its match, delete and insert states and the allowed transitions between them.



each gap. As illustrated above, the Profile-HMM does not require the pre-construction of an MSA, as it can be trained from unaligned data.

3.3.8 Summary

The methods and models outlined in section 3.3 provide an overview of the attempts made to describe the biological relationships observed in MSAs. Most of these models have been developed for, and used in, the generation of databases of familial descriptors. Coupled with suitable search tools, these databases provide means for the analysis of novel sequences. However, unlike pairwise analysis, these descriptors inherently provide a more specific description of the essence of a family of proteins. Databases that store these descriptors, alongside annotation compiled specifically about the family in question, further enhance the benefit of making such a search, with the guarantee that any significant result will also be informative. A number of the most popular examples of these databases are described in the following section.

3.4 Pattern and Family Databases

In general, secondary or family databases store data derived from a set of sequences that share a commonality; i.e., a gene family or a domain, as discussed in section 3.1. Such resources can be divided roughly into two camps: pattern databases and sequence-cluster databases. The former represent those that store familial descriptions in the forms mentioned in section 3.3. The latter are represented by databases that use pairwise similarity and multiple sequence alignment to collect together, or cluster, all putative members of a gene family.

3.4.1 Pattern Databases

These resources are characterised by the methods used to encode family membership, the extent to which this is supplemented with annotation, and the search facilities that are provided to identify these patterns in query sequences.

3.4.1.1 PROSITE

The RE is the basis of the encoding of familial patterns in the PROSITE database (Hofmann et al., 1999). Development of such signatures involves careful selection and re-selection of patterns that 'only' characterise members of the family in question. A database search with the pattern at each stage of development is required to ensure that no non-family member also contains this pattern by chance. Once it is determined that the RE is capable of doing its job (i.e., selecting all true members of the family and avoiding false positives and false negatives), a corresponding database entry is created and extensive family-specific annotation is included. A PROSITE database entry contains: a description of the RE pattern; annotation, which provides a concise description of the protein family; and a list of sequences that match the pattern (including an indication of their status; i.e., true-positive, false-positive or false-negative).

Searching PROSITE is facilitated by the ScanProsite Web interface[3], which, when provided with a query sequence, returns a simple list of matching patterns.

3.4.1.2 PRINTS

The PRINTS method (Attwood and Beck, 1994) is the least automated of the multiple-motif based methods. Motif selection, i.e., determination of its start position and width, and its isolation, is performed by hand via manipulation of a seed MSA with an alignment editor (Parry-Smith et al., 1998). Each extracted motif is comprised of a simple block of amino acids (figure 3.8).

The initial seed MSA rarely encompasses the full extent of the biological family that is to be described. So, the set of motifs that are extracted from this alignment are subjected to an iterative process, in which motifs are used to extract further family members from sequence database (currently a composite of SWISS-PROT and TrEMBL (Attwood, 2000)). New sequence information is then used to augment the motifs, and the process is repeated. Cycles of scanning and motif augmentation continue until no more sequences match the motifs. At this convergence point, the motifs can be considered as having reached their full descriptive potential. Known collectively as a fingerprint, this set of motifs is stored in the PRINTS database. Each entry is supplemented with information detailing the biological function or role of the protein family that it describes, including: cross-references to family members in primary resources and to patterns in secondary databases; links to relevant references and articles, and a concise review of functional characteristics and other salient features of the family.

Searching the PRINTS database of fingerprints was originally facilitated by a WWW interface to X-finger (Perkins and Attwood, 1996). This software enabled users to submit a single query sequence, and returned a list of the highest scoring motifs, and fully matching fingerprints.

[3] http://expasy.cbr.nrc.ca/tools/scnpsit1.html

3.4.1.3 Blocks

Blocks (Henikoff and Henikoff, 1991) is a database that uses multiple ungapped sub-alignments, termed blocks, to represent family membership. In the original version of the database blocks were extracted from families identified by PROSITE, while later versions used both PROSITE and PRINTS. Alignment and detection of conserved regions is performed by an automated system called PROTOMAT (Henikoff et al., 1995). Blocks are stored in their raw form in the database, but each sub-sequence is supplemented with additional weighting information. The weight assigned to a sub-sequence represents a measure of its divergence from others in the block. Redundancy, attributable to over representation of particular sequences or residues, can be reduced by giving high weights to sequences that are under represented and low weights to those that are over represented. This weighting is then taken into account when the score, for matching a query sequence to a block, is calculated. Blocks+ (Henikoff et al., 1999) is a recent extension to the Blocks database, which extends family coverage by taking additional blocks from families defined in PRINTS, Pfam-A (Bateman et al., 2000), ProDom (Corpet et al., 2000) and DOMO (Gracy and Argos, 1998).

Blocks can be searched using the BLIMPS search software, which involves the conversion of blocks into PSSMs (Henikoff and Henikoff, 1996) before aligning and scoring a query sequence against each block. Sequences matching two or more blocks are given scores *and* p-values, which indicate the mathematical significance of the given scores.

3.4.1.4 Meta-MEME

Another development of the multiple motif methodology, which uses a different procedure for encoding and identifying motifs is Meta-MEME (Grundy et al., 1997). This method, like Blocks, starts from unaligned sequences and selects motifs using a procedure called MEME. Using this method, motifs and the distances between them, are en-

coded using a single probabilistic model, the HMM. The advantage of using an HMM to describe the combination of residue frequency data and the distances between conserved regions, is that both types of data (residue frequencies, and lengths of gap) can be modelled using identical mathematical tools within the paradigm's framework. This is an important point, as inter-motif distances are characteristics that can aid the differentiation between true and false family members. The previously mentioned multiple motif methods generally either use an arbitrary means of scoring, or penalising, inter-motif gaps that exist beyond the observed norm, or ignore them completely.

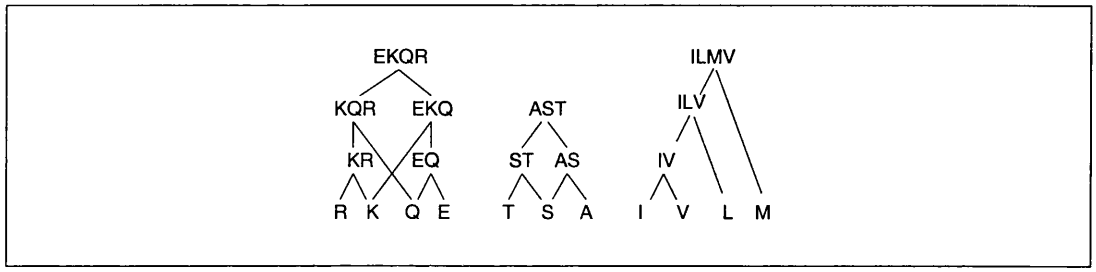
The HMM formalisation has already been outlined (section 3.3.7.2); however, it is important to emphasise that the advantage of their use is that all features of the alignment can be scored using a consistent probabilistic method, including point mutation and insertion/deletion events: *ad hoc* penalties do not need to be derived. The MetaMEME model simplifies the HMM model to only represent match states and insert states. A motif is described by a succession of match states, the number being equal to the number of columns in the motif, the spacing between the 1st motif and the next is captured as the probabilities of transitions into, around and out-of an insert state (see figure 3.14, on page 95, for an illustration of transition states in the more generalised Profile-HMM). Each motif is thus described as an individual chain of match-states, and then the whole model is stitched together, so that an insert state describes the distances observed between each pair of neighbouring motifs in the alignment.

3.4.1.5 IDENTIFY

The IDENTIFY database is generated using the EMOTIF method (Nevill-Manning et al., 1997). Two main principles underlie this method. Firstly, the patterns used to describe motifs are REs, which differ from PROSITE REs in their use of substitution groups. These groups represent sets of residues frequently observed to substitute for one another, while substitutions between, or outside of, groups are observed less frequently (based on alignments from the Blocks and HSSP (Dodge et al., 1998)

databases). Figure 3.15 shows an example of three such substitution groups.

Figure 3.15: Substitution groups are sets of amino acids found to occur together in columns of aligned sequences. Arranging these hierarchically provides an opportunity to describe relationships between the residues in the groups, and provides a clear representation of each of the overlapping sets.



Arranging these groups hierarchically allows a column in a motif to be described by the smallest, most biologically meaningful group possible. For example, a column containing the residues E, K, R and A may be described by [EKRA] (figure 3.16). However, it is likely that the observation of alanine in this position is significant only in indicating that there is little conservation at this position. A better solution would be to model the sequences containing E, K and R using a defined set such as [EKQR] (the smallest group to contain E, K and R from figure 3.15) and describe the other sequences appropriately, i.e., [A].

Indeed, the second feature that characterises the EMOTIF method is its use of multiple specific REs to represent a single motif, in order to describe a relationship containing subdivisions (i.e., a super-family contains families, and a family contains sub-families). This feature also introduces an increased level of selectivity into a sensitive search. A motif that describes a common feature of a super-family may contain many columns that, due to divergence, contain little or no information. An RE can only describe such a position with a wild card, which allows any residue to match. However, as in the previous example, an alternative is to create two REs that together provide full coverage of the super-family, while individually producing a more specific diagnosis of family membership (figure 3.16).

IDENTIFY's REs are created from motifs in the PRINTS and Blocks databases. The

Figure 3.16: An RE forced to represent a divergent relationship may ultimately be too unselective; however, by defining two, more specific, REs this region can be described more effectively.

Sub-family 1	K	P	L
	E	P	L
	E	P	I
	K	P	V
	R	P	A
Sub-family 2	A	P	A
	A	P	A
An RE that describes the whole family is:	[EKRA]	P	[ILVA]
Using only defined substitution groups this yields:	x	P	x
This motif can be described by two, more selective REs:	[EKRQ]	P	[ILV]
	A	P	A

database of REs is searched by the EMOTIF-search program, which provides a facility for the submission of a query sequence and outputs a simple ranked list of RE matching sub-sequences.

3.4.1.6 Profiles

'Profiles' (Bucher et al., 1996) are based on a generalisation of the Gribskov profile (Gribskov et al., 1987), which was described in sections 3.3.4.3 and 3.3.7.1. The profiles database is most commonly used as complementary resource to PROSITE. Profiles are used, due to the enhanced sensitivity of the models, as supplements to PROSITE REs in cases where patterns fail to discriminate, e.g., domains or other regions of high sequence divergence. The Profiles database exists as a distinct entity maintained by the Swiss Institute for Experimental Cancer Research (ISREC) group; however, only those that are verified and annotated, by the ISREC group, are distributed with PROSITE.

Profiles can be searched using the ProfileScan software[4]. Submitting a query sequence to this resource results in a score sorted list of profile matches.

3.4.1.7 Pfam

Pfam (Bateman et al., 2000) contains family MSAs and Profile-HMMs. For each family, a small number of representative sequences are aligned to produce a seed alignments from which Profile-HMMs are built. An iterative process of refinement follows the building of the initial seed HMMs, in which a non-redundant set of sequences 'pfamseq' (Bateman et al., 2000) are searched with the model to establish a complete set of family members. Once the optimal seed alignment is identified (i.e., one that allows the model to identify all members of the family), thresholds are set to establish a cut-off between true and false matches to the model. The above is a description of the development of the Pfam-A component of the resource, complementing this set of annotated and validated alignments (a short description of the family, and a threshold value accompany each of the Pfam-A alignments) is an automatically generated set of Profile-HMMs (Pfam-B). Pfam-B entries are generated from automatically clustered sequence groups derived from the ProDom database.

Pfam is searched using 'hmmsearch[5]' and 'hmmpfam' from the 'hmmer' package[6]. Alignments between sequences and the model are presented in a score-sorted list. A WWW server implementation of the search facility[7] provides graphical representations of results, which includes an overview of the different domains involved in a family (if applicable).

[4]http://www.isrec.isb-sib.ch/software/PFSCAN_form.html

[5]Hmmsearch is a package that allows the user to search a single HMM against a database of sequences, while hmmpfam conversely allows the searching of a single sequence against a database of HMMs.

[6]http://hmmer.wustl.edu/

[7]http://www.sanger.ac.uk/Pfam

3.4.2 A summary of secondary databases.

The 'pattern' clearly holds greater potential to describe the familial relationships inherent between sequences than a single sequence alone. A single sequence represents a distinct point in evolutionary history. The comparison of two points can provide a view of the heritage of individuals; however, this perspective cannot compare to the wealth of knowledge that is made available by the addition of more and more points. As more sequences are added to the model, any apparent conservation reflects the evolutionary preservation of that thing that defines the family of proteins - its structural or functional characteristics. Using patterns to identify relationships, in order to confer functional annotation, can therefore be more precise. The pattern holds the key to the identification of specific features, which, due to the fact that they have been conserved, may be inferred to be vital functional components. Whereas, the identification of general similarities between two sequences cannot achieve such a specific level of diagnosis.

The very act of pattern derivation means that time and effort have to be expended to create familial descriptors. As a consequence of this, the total family coverage in the pattern databases represents a small proportion of the available sequence family data. Unfortunately, due to this discrepancy, the negative result of querying such a resource may be ambiguously interpreted: a query sequence could indeed be novel and have no observed relatives; alternatively, and most likely, it belongs to a family not yet described by the database in question. Most pattern databases contain non-overlapping distributions of family descriptors, and hence, in the case of a negative result, it is important to query as many resources as are available.

While it is true that some families are only represented in individual databases, a degree of overlap does exist. This, however, does not constitute redundancy: each type of pattern (RE, motif, profile, HMM, etc.) has a different range of properties. Consequently, the use of multiple resources can provide independent evidence, which is essential for testing the reality of a diagnosis.

REs, in general, tend to lack sensitivity: they are commonly either detected or not, which leaves little room for ambiguity. Conversely, long REs can be very selective, which means that confidence in their assignments is usually high. PROSITE REs are of varying length and quality, however, so care should be taken to obtain complementary evidence from another source. The IDENTIFY methodology (section 3.4.1.5), clearly shows that, even for REs, sensitivity can be obtained through careful manipulation of the original data.

The use of multiple motifs introduces the potential to identify both sensitive (in which a sequence is allowed to match less than all motifs) and selective results (in which a sequence must match all motifs to be considered a member of the family). The methods adopted by PRINTS and Blocks, produce composite descriptors of familial membership, in which the *black and white* question of "does a sequence match or not?" can be softened to introduce shades of grey that allow distant family-members to match, as well as close members. Also the use of different scoring schemes, such as frequency matrices and PSSMs, to score matches to motifs provide varying improvements in sensitivity; e.g., the introduction of a substitution matrix into the scoring scheme means that residues not observed in the original alignment can still be scored, on the basis of likely substitutions. Families that have become too divergent to contain significant stretches of well conserved alignment, cause motif-based models (especially those which rely on the absence of gaps) to suffer from the absence of conserved columns. Motifs derived from such MSAs are forced to be short, and as a consequence may to fail in distinguishing true from false.

By utilising the potential for the entirety of each member sequence to be used in a description of a family, those resources that describe whole alignments can gain greater levels of sensitivity. In the search for distant family members, it is important to be able to extract as much information as possible from the MSA. Clearly, of benefit in this task is a model that can utilise information-rich regions, which are normally discarded by motif based methods. However, with an increase in sensitivity there is often a

concurrent loss in selectivity. The use of the whole alignment often requires the incorporation of regions so divergent that no meaningful alignment can be made; however, these still contribute to the model. As the proportion of divergent unaligned sequence increases with respect to the conserved aligned regions, the probability of identifying random sequences (false positives) increases. A loss of selectivity usually means that the distinction between true positives and false positives is difficult to define.

The problems associated with making this distinction can be illustrated with the naive derivation of a Profile-HMM to describe the α-haemoglobin family. The first step being to derive a representative set of α-haemoglobin sequences to represent a seed alignment (figure 3.17). If these sequences are then aligned and a Profile-HMM is generated from the alignment, then all members of the seed alignment score highly when probed with the model. Searching a sequence database with the model reveals many high scoring matches, a high proportion of which are α-haemoglobin . However, a significant proportion of these matches come from the closely related δ-, and, less closely related, β-haemoglobins, which do not belong to the seed alignment and should therefore be considered false positives (figure 3.18).[8]

Both profile-based resources, mentioned in the previous section (sections 3.4.1.6 and 3.4.1.7), use a threshold cut-off score that represents the observed limit of family membership (sequences scoring below this value are not considered members of the family), which allows selectivity to be regained. While, the α-haemoglobin example, shows that selectivity can be a problem, it also illustrates the suitability of these models for the description of distant relationships. A number of the low scoring sequences that match the profile, built solely from α sequences, represent distant relationships; including myoglobins and invertebrate globins, as well as the other members of the vertebrate haemoglobin family.

To summarise, the benefits of using patterns over single sequences, in searching for

[8]It will be demonstrated in a later section that it is possible to create a fingerprint, using the PRINTS method, which is capable of making the distinction between the α and β-haemoglobin families.

Figure 3.17: The seed alignment for the generation of a Profile-HMM of the α-haemoglobin family. The figure shows a section (the first 50 residues from each sequence) of the file used as input for the hmmbuild program (from the HMMer suite).

```
HBA1_PLEWA/1-152  --KLTAEDK HNVKAIWDHV KGHEEAIGAE ALYR--MFCC MPTTRIYFPA
HBA_AMBME/1-152   -FKLSGEDK ANVKAVWDHV KGHEDAFGHE ALGR--MFTG IEQTHTYFPD
HBA_CAICR/1-152   --VLSEEDK SHVKAIWGKV AGHLEEYGAE ALER--MFCA YPQTKIYFPH
HBA_SPHPU/1-152   --MLSASDK ANVKAIWSKV CVHAEEYGAE TLER--MFTV YPSTKTYFPH
HBA_MESAU/1-152   --VLSAKDK TNISEAWGKI GGHAGEYGAE ALER--MFFV YPTTKTYFPH
HBA_RAT/1-152     --VLSADDK TNIKNCWGKI GGHGGEYGEE ALQR--MFAA FPTTKTYFSH
HBA_LYNLY/1-152   --VLSAADK SNVKACWGKI GSHAGDYGTE ALER--TFCS FPTTKTYFPH
HBA_PANTS/1-152   --VLSSADK NNVKACWGKI GSHAGEYGAE ALER--TFCS FPTTKTYFPH
HBA1_BOSMU/1-152  --VLSAADK GNVKAAWGKV GGHAAEYGAE ALER--MFLS FPTTKTYFPH
HBA_BOVIN/1-152   --VLSAADK GNVKAAWGKV GGHAAEYGAE ALER--MFLS FPTTKTYFPH
HBA_EQUZE/1-152   --VLSAADK TNVKAAWSKV GGNAGEFGAE ALER--MFLG FPTTKTYFPH
HBA_PHYCA/1-152   --VLSPADK TNVKAAWAKV GNHAADFGAE ALER--MFMS FPSTKTYFSH
HBA_ELEMA/1-152   --VLSDKDK TNVKATWSKV GDHASDYVAE ALER--MFFS FPTTKTYFPH
HBA_CEBAP/1-152   --VLSPADK TNVKTAWGKV GGHAGDYGAE ALER--MFLS FPTTKTYFPH
HBA_HUMAN/1-152   --VLSPADK TNVKAAWGKV GAHAGEYGAE ALER--MFLS FPTTKTYFPH
HBA1_GALCR/1-152  --VLSPTDK SIVKAAWEKV GAHAGDYGAE ALER--MFLS FPTTKTYFPQ
HBA1_TADBR/1-152  --VLSPEDK NNVKAAWSKV GGQAGDYGAE ALER--MFLS FPTTKTYFPH
HBAD_AEGMO/1-152  --MLTADDK KLIQATWDKV QGHQEDFGAE ALQR--MFIT YPPTKTYFPH
HBA_CYPCA/1-152   --SLSDKDK AAVKGLWAKI SPKADDIGAE ALGR--MLTV YPQTKTYFAH
HBA_CARAU/1-152   --SLSDKDK AVVKALWAKI GSRADEIGAE ALGR--MLTV YPQTKTYFSH
HBA_CATCL/1-152   --SLSDKDK ADVKIAWAKI SPRADEIGAE ALGR--MLTV YPQTKTYFAH
HBA1_NOTAN/1-152  --SLSDKDK AAVRALWSKI GKSADAIGND ALSR--MIVV YPQTKTYFSH
HBA1_SALIR/1-152  --SLTAKDK SVVKAFWGKI SGKADVVGAE ALGRDKMLTA YPQTKTYFSH
HBA_SALSA/1-152   --SLTARDK SVVNAFWGKI KGKADVVGAE ALGR--MLTA YPQTKTYFSH
HBA1_XENLA/1-152  --LLSADDK KHIKAIMPAI AAHGDKFGGE ALYR--MFIV NPKTKTYFPS
HBA3_RANCA/1-152  --SLSASEK AAVLSIVGKI GSQGSALGSE ALTR--LFLS FPQTKTYFPH
HBA_LATCH/1-152   --GLTAADK TLIKSIWGKV EKETEAIGVE ALVR--LFKC FPQSKVYFDH
```

relationships between sequences, are obvious. Single sequences merely contain snapshots of the evolution of a family, while an MSA represents a reconstruction of a portion of the evolutionary history of a family. Amongst the many ways of representing the MSA, there are advantages and disadvantages to each method, and, no one pattern database can claim to provide a full coverage of all the protein families available in the primary databases. When using patterns to identify novel sequences, or to hunt for unidentified members of families, the most profitable approach is to use multiple resources. Complementary approaches can provide independent confirmation of unclear results, and provide the most complete coverage of gene families by utilising the fact that these databases contain non-overlapping distributions.

The next sections illustrate attempts made to address some of the deficiencies that have been highlighted in this section, starting with a composite resource, and moving on to more automated approaches at describing familial relationships.

Figure 3.18: Searching a SWISS-PROT/TrEMBL composite database of sequences (version 37_9) with the Profile-HMM generated from the alignment shown in figure 3.17, produced the following result.

Shown below is a selection of the complete set of results which runs to 750 lines. Before the divide, the first 25 results are shown, which are all α-haemoglobin sequences. After the divide, the first non-α sequences appear (highlighted in red) with scores still deemed to be highly significant. Lower still are some remaining α sequences (blue).

```
Sequence                                  score  e-value
SPROT|P14387|HBA_ANTPA|HEMOGLOBIN         391.7 2.3e-113
SPROT|P11755|HBA1_TADBR|HEMOGLOBIN        391.0 3.6e-113
SPROT|P07405|HBA_FELCA|HEMOGLOBIN         388.1 2.7e-112
SPROT|P01927|HBA_ATEGE|HEMOGLOBIN         387.9 3.1e-112
SPROT|P01956|HBA_ROUAE|HEMOGLOBIN         387.1 5.5e-112
SPROT|P01945|HBA_MESAU|HEMOGLOBIN         386.9 6.3e-112
SPROT|P01953|HBA_MELME|HEMOGLOBIN         386.9 6.4e-112
SPROT|P01922|HBA_HUMAN|HEMOGLOBIN         386.8   7e-112
SPROT|P21767|HBA_MACFA|HEMOGLOBIN         386.7 7.3e-112
SPROT|P10892|HBA_LUTLU|HEMOGLOBIN         386.2   1e-111
SPROT|P01937|HBA_NYCCO|HEMOGLOBIN         386.2   1e-111
SPROT|P41327|HBA_LYNLY|HEMOGLOBIN         386.0 1.2e-111
SPROT|P11757|HBA_MYOVE|HEMOGLOBIN         386.0 1.2e-111
SPROT|P18972|HBA_CALAR|HEMOGLOBIN         386.0 1.2e-111
SPROT|P01951|HBA_TALEU|HEMOGLOBIN         385.8 1.3e-111
SPROT|P01924|HBA_PREEN|HEMOGLOBIN         385.7 1.4e-111
SPROT|P01923|HBA_GORGO|HEMOGLOBIN         385.7 1.5e-111
SPROT|P20854|HBA_CTEGU|HEMOGLOBIN         385.6 1.5e-111
SPROT|P01925|HBA_MACMU|HEMOGLOBIN         385.5 1.7e-111
SPROT|P09908|HBA_PHOVI|HEMOGLOBIN         385.3 1.9e-111
SPROT|P23601|HBA_MUSPU|HEMOGLOBIN         385.3 1.9e-111
SPROT|P23600|HBA_MUSLU|HEMOGLOBIN         385.2   2e-111
SPROT|P10885|HBA_PTEBR|HEMOGLOBIN         385.1 2.2e-111
SPROT|P28780|HBA_TAPGE|HEMOGLOBIN         384.9 2.6e-111
SPROT|P01940|HBA_TARBA|HEMOGLOBIN         384.4 3.5e-111
SPROT|P01928|HBA_CEBAP|HEMOGLOBIN         384.2 4.2e-111
⋮
SPROT|P16417|HBA_LIOMI|HEMOGLOBIN         237.0  8.6e-67
SPROT|P06714|HBAT_HORSE|HEMOGLOBIN        215.3  2.9e-60
SPROT|P07408|HBA_SQUAC|HEMOGLOBIN         206.4  1.4e-57
SPROT|P20244|HBA1_TORMA|HEMOGLOBIN        179.3  1.9e-49
TREMBL|Q28932|Q28932|HEMOGLOBIN           177.8  5.6e-49
SPROT|Q03902|HBD_GALCR|HEMOGLOBIN         177.5    7e-49
SPROT|P11517|HBB2_RAT|HEMOGLOBIN          175.3  3.2e-48
SPROT|P02090|HBB_SPAEH|HEMOGLOBIN         175.1  3.6e-48
SPROT|P20245|HBA2_TORMA|HEMOGLOBIN        175.0  3.9e-48
TREMBL|Q03901|Q03901|BETA                 173.7  9.8e-48
SPROT|P02088|HBB1_MOUSE|HEMOGLOBIN        172.7  1.9e-47
SPROT|P02096|HBG_HUMAN|HEMOGLOBIN         172.6    2e-47
SPROT|P02062|HBB_HORSE|HEMOGLOBIN         172.6  2.1e-47
SPROT|P02091|HBB1_RAT|HEMOGLOBIN          171.9  3.2e-47
SPROT|P02063|HBB_EQUHE|HEMOGLOBIN         171.7  3.8e-47
SPROT|P09840|HBB_MACCA|HEMOGLOBIN         171.2  5.3e-47
SPROT|P02060|HBB_SUNMU|HEMOGLOBIN         171.2  5.5e-47
SPROT|P09421|HBB_SPECI|HEMOGLOBIN         170.3  9.8e-47
SPROT|P02066|HBB_CERSI|HEMOGLOBIN         169.4  1.8e-46
SPROT|P02089|HBB2_MOUSE|HEMOGLOBIN        169.4  1.9e-46
SPROT|P08853|HBB_MARMA|HEMOGLOBIN         169.3    2e-46
SPROT|P56691|HBA_DASAK|HEMOGLOBIN         168.8  2.9e-46
SPROT|P02051|HBB_TARBA|HEMOGLOBIN         168.5  3.4e-46
```

3.4.3 A composite pattern database - InterPro

InterPro (Apweiler et al., 2000) is a composite resource that currently provides access to PROSITE, PRINTS and Pfam. It is built at the EBI as a collaborative project between the researchers responsible for its parent resources. An InterPro entry contains no familial discriminator of its own; each merely exists to combine, rationalise and standardise the access to patterns, motifs and models from the parent resources. An entry contains: links to parent databases; family-specific annotation, including literature references (mostly taken from PROSITE and PRINTS); information concerning biological relationships between families described by entries, and a list of all sequence members of the family it describes.

The importance of a resource such as this comes from the advantages inherent in combining a number of different search tools into a single coherent unit. Each database and method, described in section 3.4, provides a non-overlapping perspective on the task of identifying homologous relationships, either in the coverage of different families or in the use of different methods of encoding this information. This diversity means that the result of a similarity search against an individual database can be corroborated or disproved by searching an alternative database. However, each resource is situated in a different geographic location, and while the WWW facilitates seamless transitions between these locales, it is still necessary to perform a number of different searches and manually assemble a coherent result. InterPro enables the disparate models and family descriptions of a number of pattern databases to be accessed and searched using a single set of tools, within one location, and for the results to be displayed in one consistent interface.

3.4.4 Family or clustered sequence databases

Clustered family databases provide a level of coverage of sequence space that cannot be realistically expected from the pattern databases. They are automatically gener-

ated from primary sequence data, and only some benefit from limited manual input. The coverage of a clustered database is limited only by the availability of primary sequence data. Often, resources of this type either use, or contain, references to pattern database entries, and therefore play an important role in maintaining up-to-date familial relationships within sequence databases. The clustering process has already been described in the context of the construction of BLOSUM matrices, and refers to the process by which sequences are collected into clusters on the basis of shared similarity (section 2.4.2.2). Examples of such databases include ProDom (Corpet et al., 2000), SBASE (Murvai et al., 2000) and PIR-ALN (Srinivasarao et al., 1999).

ProDom is automatically generated from SWISS-PROT and TrEMBL using a combination of automated methods including the use of Pfam models and PSI-BLAST (Altschul et al., 1997) to determine the familial membership of clusters of sequences. A number of domains or families described by ProDom are validated by human experts. SBASE consists of sequence clusters generated by demonstrating significant levels of similarity between members, based on BLAST database searches. PIR-ALN is a database of alignments derived from annotation in the PIR database: containing super-family, family and homology domain classifications. Sequences are hierarchically clustered into these classifications and then alignments are generated using a combination of automated and manual approaches.

Together, these databases share the advantage of being able to represent large numbers of families, with the corresponding disadvantage of lacking detailed annotation. To perform a similarity search of these resources requires standard pairwise tools such as BLAST or FASTA. The advantage that this has over a search of a standard sequence database is that the family membership of every sequence is known prior to the search. Therefore, queries matching pre-clustered sequences can be more confidently assigned; i.e., a query sequence matching 10 members of the same family is more likely to represent a true match than one matching sequences from 10 different clusters.

3.4.5 PSI-BLAST

PSI-BLAST (Altschul et al., 1997) is a hybrid approach, which, like BLAST, can search a database of sequences for pairwise similarities. However, based on the sequences identified as high scoring results, PSI-BLAST can generate a profile, which can then be used to perform a subsequent search of the same database. This process can be iterated, by adding new sequences at each step, until a stable set is identified. Unlike conventional profile-based methods, the PSI-BLAST model does not require the prior construction of a seed alignment, and this ability to build a profile 'on the fly' means that this approach is not limited to searching for the subset of families that have been characterised by the pattern databases.

This tool is widely used, as it represents an impressive improvement on the simple BLAST search; however, care must be taken with respect to the iterative process of introducing new sequences to the profile. The inclusion of any particular sequence into the profile dramatically improves the score that it will receive, which is to be expected: if a sequence, such as a divergent member of a family, is weakly matched by a profile, then its inclusion should modify the model in its favour. However, this applies equally for divergent *and* unrelated sequences; hence, the augmentation of a profile with an unrelated sequence can significantly impair its diagnostic ability. In the extreme, this effect can cause the original family to be replaced by the incursion of the unrelated family. Even though this phenomenon is true of all profile methods, its potential is ever more present in an approach like PSI-BLAST, because the construction of the profile is less carefully supervised: subsequent iterations can be automated.

This chapter has described the analysis of multiple sequences, and the potential for MSAs to be used to define patterns capable of describing the evolutionary relationships inherent in protein families. A range of pattern databases have been constructed and patterns are commonly used in the task of *in silico* functional identification of

novel protein sequences. The following chapter will discuss the PRINTS database in greater detail, with particular focus on the research activity required in the generation of patterns based on protein families.

Chapter 4

PRINTS

4.1 Introduction

The following sections detail the development of fingerprints: the familial descriptors of the PRINTS database. Developing fingerprints has contributed to this thesis in two ways. Firstly, as a purely research-led exercise, a number of protein families have been studies and fingerprints have been derived for deposition in PRINTS (especially the haemoglobin and tRNA synthetase families). Secondly, understanding the process that underlies the derivation and use of fingerprints drove the research required for the development of a new PRINTS search tool (Chapter 5).

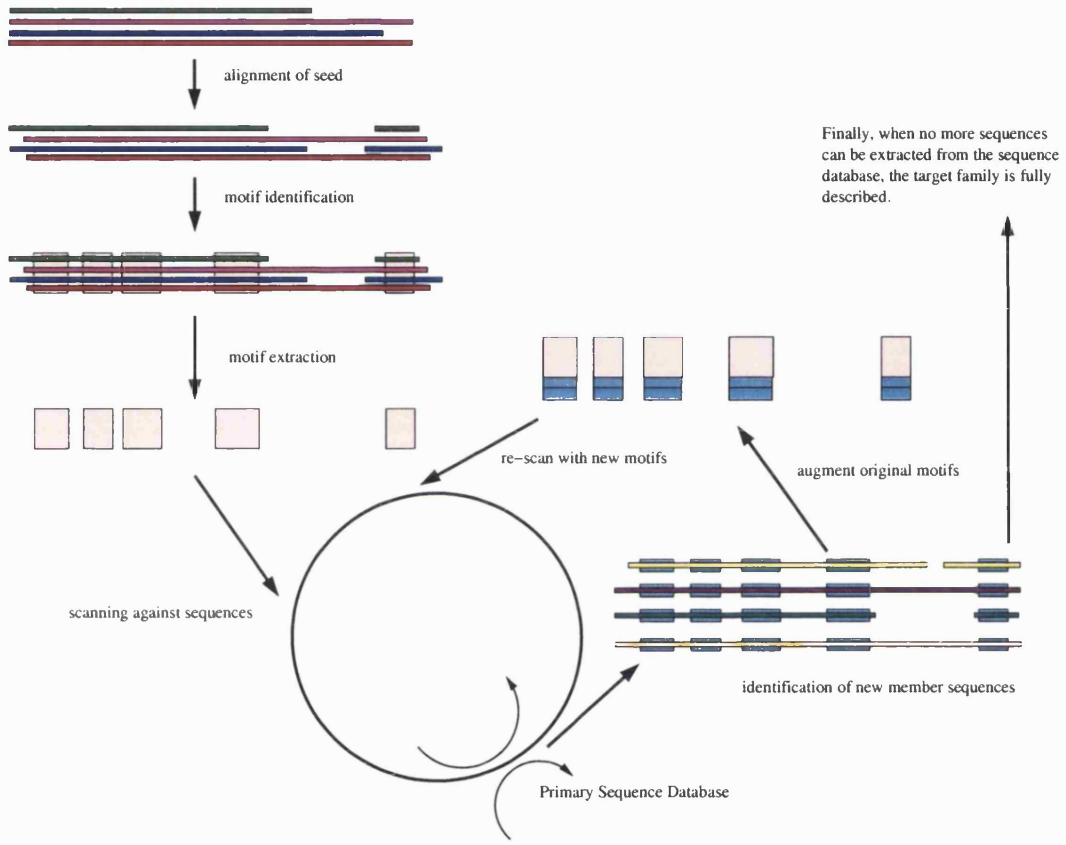
4.2 The development of fingerprints

Fingerprinting a family of proteins (the target family) can be seen as the progression through a number of sequential steps (figure 4.1):

- alignment of a set of sequences that are representative of the target family (i.e., they must be sufficiently diverse to describe the target family),
- regions of conservation are identified in the alignment, and marked as motifs,
- motifs are extracted, and encoded in order that they can be used to identify matching sub-sequences,
- a primary protein sequence data-bank is searched for matches to these motifs,
- sequences identified in the search are evaluated to identify those that match all motifs in the fingerprint,
- the original motifs are augmented with sequence information from these fully matching sequences,
- then the search and augmentation process is repeated (iterated) each time *new* fully matching sequences are identified,

- the process completes when no more new sequences are identified.

Figure 4.1: An overview of the fingerprinting process.



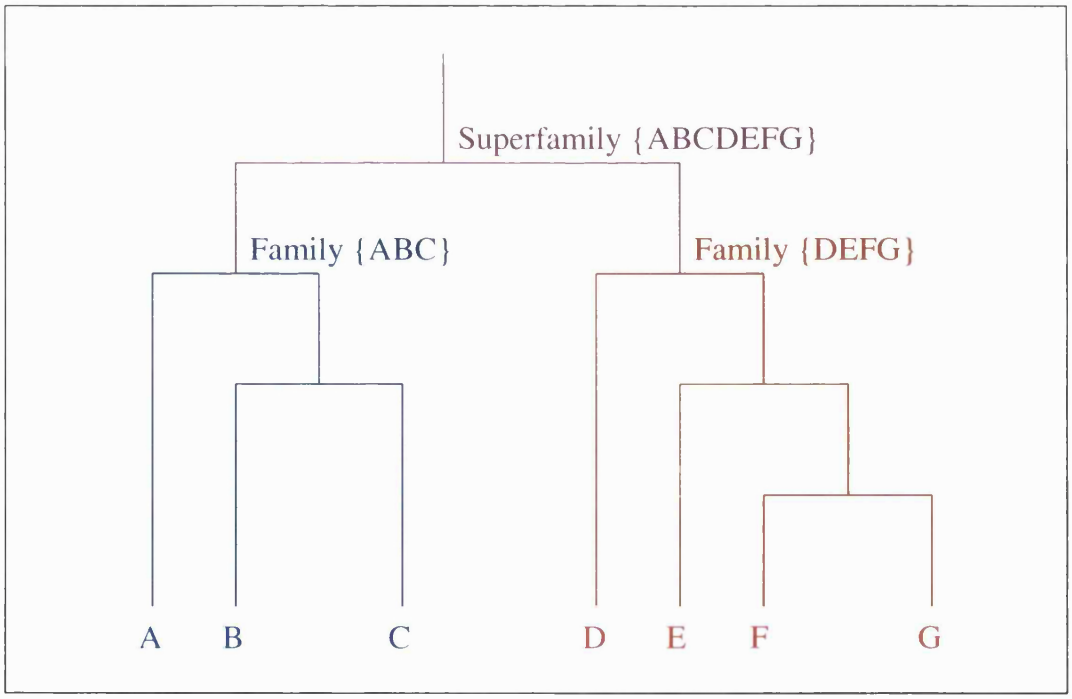
Once a target family has been fully represented, the fingerprint is annotated, and deposited into the PRINTS database. The target family can represent a protein superfamily, a family, a specific sub-family or sub-type, or it can be as diverse as to represent a promiscuous shared domain.

4.2.1 Alignment

The initial alignment of a family of sequences is a critical step in the derivation of a fingerprint. The chosen sequences must be representative and care must be taken to select a set that, while sharing the common features of the family, are not biased towards

a closely related sub-set. Using such a subset can result in a fingerprint that is skewed towards the over-represented group. This considered selection process is particularly important in the derivation of a super-family fingerprint. Modelling a super-family requires sequences to be selected as evenly as possible from each of the different families, and sub-families, that comprise it. In figure 4.2 a super-family is represented as the collection of all sequences from sub-families A-G. Selecting candidate sequences for a super-family MSA should ideally take into account all sub-families. For example, sampling heavily from A, B and C, produces an alignment that favours a diagnosis of the 'ABC' family and not the 'DEFG' family, while sampling from both is more likely to provide a representative seed.

Figure 4.2: The selection of sequences is a critical step in the description of an alignment.



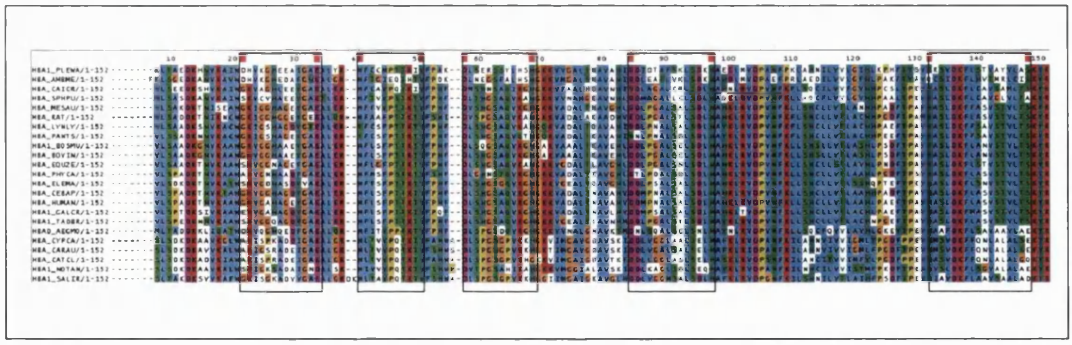
Once a candidate set of sequences has been identified, the MSA is constructed. The manual alignment of sequences essentially represents an incremental pairwise alignment process, which is extended through the addition of distant relatives (to broaden the scope of the MSA) and close relatives (to facilitate the identification of conserved

regions). Manual alignment is often supplemented with judicial use of an automated alignment tool, which is commonly followed by manual re-alignment.

4.2.2 Motif Extraction

The manual alignment process is influenced by the requirement to identify regions of conservation that are suitable for the excision of *ungapped* motifs. Therefore, the objective is to produce alignments that contain blocks of aligned sequence separated by unaligned gapped regions. From this seed MSA, the most conserved regions are selected and extracted to produce the 'initial' set of motifs (figure 4.3).

Figure 4.3: The fingerprint ALPHAHAEM was derived from this alignment of α haemoglobin sequences. The boxes highlight the sub-sequences extracted to become the initial set of motifs. (Motif 1 is shown, in full, in figure 4.4)

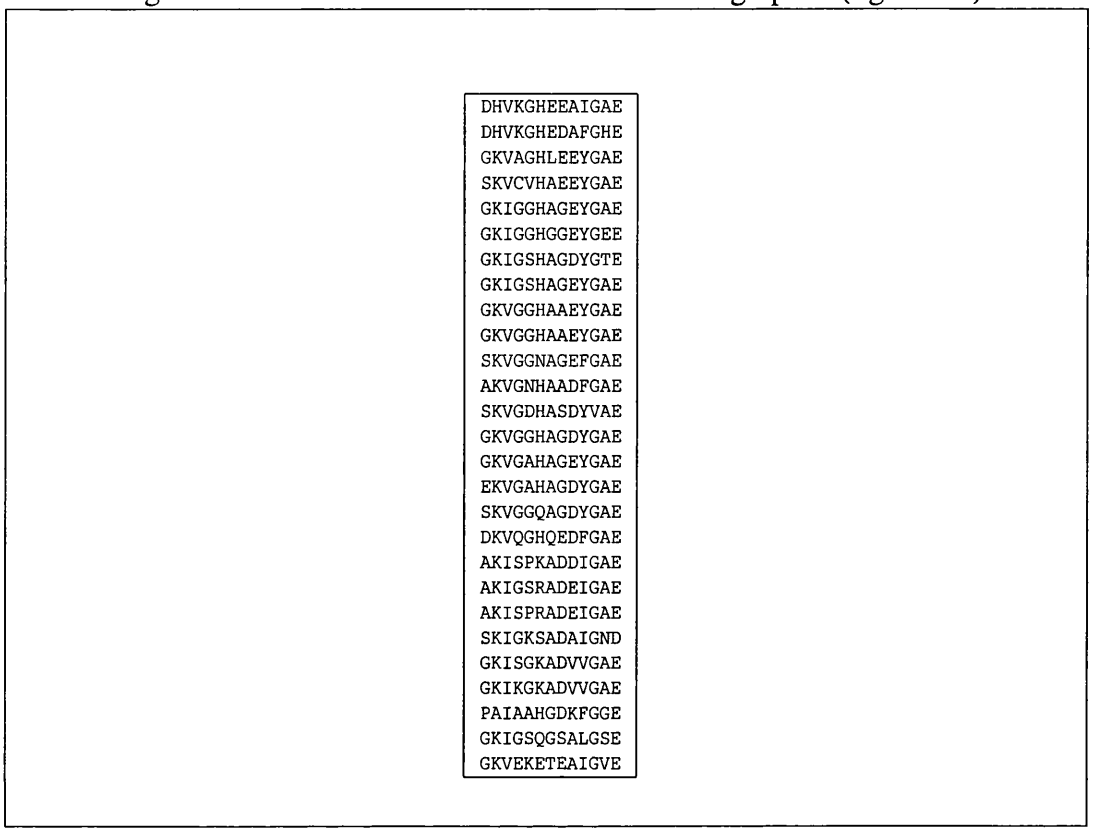


4.2.3 Iteration

4.2.3.1 Scanning and matching sequences

Each motif, in turn, is scanned against all of the sequences in the primary database (currently a SWISS-PROT/TrEMBL composite (Attwood et al., 2000)) to produce a list of potential matches ranked by score. The n-single scoring method, designed to enhance the signal to noise ratio of motif scores (Parry-Smith, 1990; Parry-Smith and Attwood, 1992), is computed for each sub-sequence match to the motif (see figure 4.5).

Figure 4.4: Motif 1 from the ALPHAHAEM fingerprint (figure 4.3).



The scoring and ranking procedure, performed by the 'scan' program (a component of the 'Algorithms and Data Structures for Protein sequence analysis' (ADSP) suite (Parry-Smith, 1990; Parry-Smith and Attwood, 1992)), is repeated for each motif. The result is a sorted list of scoring matches to each motif (figure 4.6) . Analysis of these lists is performed by the 'compare' program (also a component of ADSP[1]). While the fingerprint is a composite structure, which is the sum of its motifs, 'scan' produces individual lists for each motif (hit-lists). Therefore, to identify matches to the fingerprint requires rationalisation of these lists. 'Compare' does this by providing motif-match information from the perspective of the sequence: i.e., the question becomes how many motifs does a sequence match (rather than which sequences are matched by each motif). The rationale is that a random sequence may match a single motif by chance alone, but the probability that it will match two or more diminishes rapidly. So,

[1]The scan and compare programs were originally only made available for the VMS operating system. The updated scan and compare programs that are currently used for the development of PRINTS have been made available by W. Wright (unpublished data).

Figure 4.5: N-single scoring of motif matching sub-sequences.

For each motif, the maximum achievable score is calculated: taking the frequency of occurrence of residues in each column of the motif, and selecting the maximum. Shown below are an example motif and a representation highlighting the frequencies of the residues in each column.

A	C	E	G	F	N	W
A	C	E	G	F	N	W
A	C	E	I	F	N	Y

A_3	C_3	E_3	G_2	F_3	N_3	W_2
			I_1			Y_1

The highest scoring sequence is ACEGFNW ($3+3+3+2+3+3+2=19$) which is then multiplied by the number of residues in the motif (7) ($19*7=133$).

When a query sequence is scored against the motif, its score is given as a percentage of the maximal n-single score. For example, the sequence ACEHFNT ($3+3+3+0+3+3+0=15$) matches in five of the seven possible columns, therefore its score is multiplied by five ($5*15=75$). As a percentage of the maximal n-single score it is $75/133*100=56.4\%$.

Figure 4.6: A hit-list of matches to a single motif.

Below is a list of sub-sequences matching motif 1 of the ALPHAHAEM fingerprint (figure 4.4). Each line in the hit-list contains the following information about each match: its rank in the list, its percentage n-single score, the sequence's identity code and accession, the position of the match within the sequence (start and end points), some parameters relating to the number of times the sequence in question has matched, and finally, the actual sub-sequence that matches the motif. Only the top 15 matches are shown; in this case, the list extends down to 2000 matches, the majority of which represent low scoring chance matches.

HitNo	Score	Id Code	Acc No	Start	End	Toc	Moc	Motif
1	100.00	HBA_MESAU	P01945	15	27	5	1	GKIGGHAGEYGAE
2	99.36	HBA_CERAE	P01926	15	27	5	1	GKVGGHAGEYGAE
3	99.36	HBA_MACAS	P21766	15	27	5	1	GKVGGHAGEYGAE
4	99.36	HBA_MACFA	P21767	15	27	5	1	GKVGGHAGEYGAE
5	99.36	HBA_MACMU	P01925	15	27	5	1	GKVGGHAGEYGAE
6	99.36	HBA_MACNE	P19002	15	27	5	1	GKVGGHAGEYGAE
7	99.36	HBA_MACSI	P21768	15	27	5	1	GKVGGHAGEYGAE
8	98.22	HBA_ANSSE	P01985	15	27	5	1	GKIGGHAEEYGAE
9	98.22	HBA_LARRI	P08260	15	27	5	1	GKIGGHAEEYGAE
10	97.37	HBA_MACSP	P07402	15	27	5	1	DKVGGHAGEYGAE
11	97.37	HBA_MANSP	P08258	15	27	5	1	DKVGGHAGEYGAE
12	97.07	HBA_HORSE	P01958	15	27	5	1	SKVGGHAGEYGAE
13	96.77	HBA_LORTA	P01938	15	27	5	1	EKVGGHAGEYGAE
14	96.47	HBA_CAMDR	P01974	15	27	5	1	GKIGGHAAEYGAE
15	96.47	HBA_ONDZI	P01944	15	27	5	1	GKIGGHAAEYGAE

by identifying sequences matching multiple motifs, 'true' family members should be highlighted. The list of sequences produced by 'compare' details those matching all n motifs of a fingerprint, $n-1$, $n-2$, ..., all the way down to matching two motifs. Sequences matching all motifs are usually members of the seed alignment; however, if new sequences are identified, this indicates that the full extent of the family was not explored by the initial set of motifs.

The following section describes the iterative process required to extend the scope of the fingerprint when new sequences are discovered.

4.2.3.2 Iterating

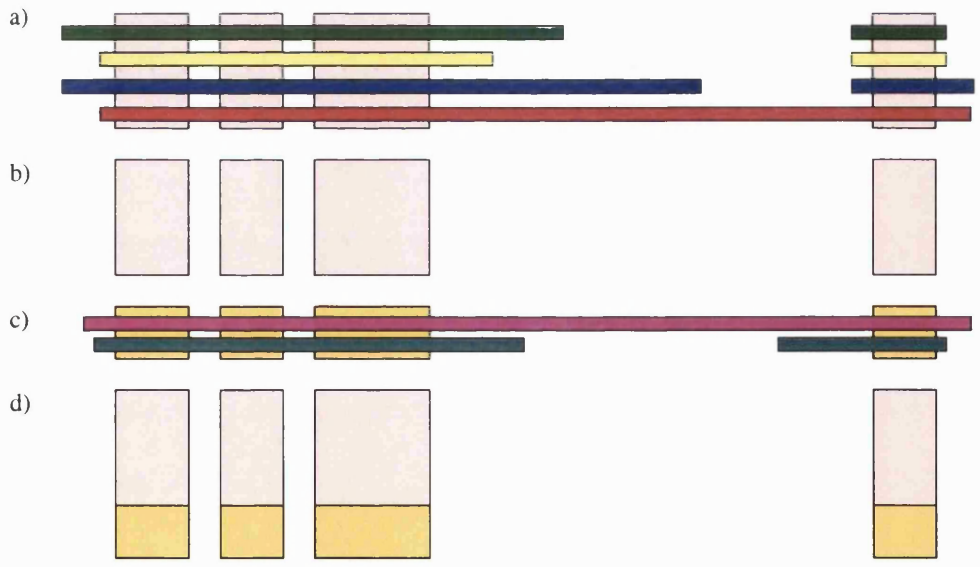
Once a 'scan' and 'compare' cycle has been performed, the resultant list of motif-matching sequences is evaluated. If new sequences are identified as matching all motifs, and their family membership is corroborated by independent evidence, then this indicates that the fingerprint is performing as expected. That is, the fingerprint has identified new family members, rather than matching false sequences. These new sequences can be used to augment the motifs: by adding each matching sub-sequence to the original motifs (figure 4.7).

These augmented motifs become labeled as 'iteration 2' motifs, and can be used in a subsequent sequence database search (a further 'scan' and 'compare' cycle).

The scan-compare-augment process is repeated only if each subsequent round of searching uncovers suitable new sequences that can be added to the motifs. When no further fully matching sequences are identified, the iteration concludes and the current motifs are recorded. At this stage, the motifs are deemed to have achieved their potential, and the database entry can be finalised.

Figure 4.7: The scanning process can identify new sequences, which can be used to augment the original motifs.

An alignment (a) yields a set of motifs (b) that can be used to search a database of sequences in order to identify matching sub-sequences. In this example, two new sequences are identified as matching all four motifs (c). The regions of the new sequences that match the original motifs can be extracted and used to augment the initial motifs (d). The search can then be repeated with the new motifs.



4.2.4 Annotating

A PRINTS database entry consists of more than just the sequence data from which the motifs are derived. Two crucially important features are the annotation, which describes the biological role of the protein family, and the links to other primary and secondary databases. Both elements require research. Annotation is usually derived directly from literature surveys. While, database searching is essential to verify links between a fingerprint's family and relevant descriptions of it in other resources. An example of an annotated fingerprint is shown in figure 4.8 (ALPHAHAEM (PRINTS accession: PR00612)), which describes the α sub-family of the haemoglobin family of oxygen binding and transport proteins. Two distinct sections are shown: the database cross references (indicated in blue) and the family annotation, which contains both literature references (red), and the free-text, family description (magenta).

Annotation is important because it is this information that helps to support a diagnosis and provides the biological context of what family membership actually means. As a consequence, care is taken to provide a concise description of the key functional properties of the family, any super-family, family, or sub-family relationships of relevance and any available biological or biochemical information specific to the motifs (the presence of active sites, binding-sites or structural features etc.).

4.2.5 Analysis of the fingerprinting method

The idealised situation, envisaged above, describes the principles of the fingerprinting process. However, in reality, creating a fingerprint can be a non-trivial exercise. The following discussion aims to describe some of the problems that may be encountered.

A perfect familial discriminator should be totally selective: i.e., no non-family members should match any of the motifs in a fingerprint. However, in each of the ranked lists of motif-matches (the hit-lists produced by 'scan'), many false sequences are matched (usually with low scores). The occurrence of a motif-match in a hit-list is

Figure 4.8: Part of the ALPHAHAEM fingerprint, including annotation and external database links

```
gc; ALPHAHAEM
gx; PR00612
gn; COMPOUND(5)
ga; 20-SEP-1996; UPDATE 23-FEB-1998
gt; Alpha haemoglobin signature
gp; PRINTS; PR00188 PLANTGLOBIN; PR00611 ERYTHCRUORIN; PR00613 MYOGLOBIN
gp; PRINTS; PR00188 PLANTGLOBIN; PR00611 ERYTHCRUORIN; PR00613 MYOGLOBIN
gp; PRINTS; PR00814 BETAHAEM; PR00815 PIHAEM; PR00816 ZETAHAEM
gp; PROSITE; PS01033 GLOBIN
gp; BLOCKS; BL01033
gp; PDB; 1CMY; 1HDA
gp; SCOP; 1CMY; 1HDA
gp; CATH; 1CMY; 1HDA
bb;
gr; 1.  DICKERSON, R.E. and GEIS, I.
gr; Hemoglobin:  Structure, Function, Evolution and Pathology.
gr; THE BENJAMIN/CUMMINGS PUBLISHING COMPANY, 1983.
gr;
gr; 2.  KAPP, O.H., MOENS, L., VANFLETEREN, J., TROTMAN, C.N.A., SUZUKI, T.
gr; and VINOGRADOV, S.N.
gr; Alignment of 700 globin sequences:  Extent of amino acid substitution
gr; and its correlation with variation in volume.
gr; PROTEIN SCIENCE 4 2179-2190 (1995).
gr;
gr; 3.  MOENS, L., VANFLETEREN, J., VAN DE PEER, Y., PEETERS, K., KAPP, O.,
gr; CZELUZNIAK, J., GOODMAN, M., BLAXTER, M. and VINOGRADOV, S.
gr; Globins in nonvertebrate species:  dispersal by horizontal gene transfer
gr; and evolution of the structure-function relationships.
gr; MOL.BIOL.EVOL. 13 324-333 (1996).
gr;
gr; 4.  WHITAKER, T.L., BERRY, M.B., HO, E.L., HARGROVE, M.S., PHILLIPS, G.N.,
gr; KOMIYAMA, N.H., NAGAI, K. and OLSON, J.S.
gr; The D-helix in myoglobin and in the beta subunit of hemoglobin is required
gr; for the retention of heme.
gr; BIOCHEMISTRY 34 8221-8226 (1995).
bb;
bb;
gd; Globins are haem-containing proteins involved in dioxygen binding and/or
gd; transport [1].  At present, more than 700 globin sequences are known [2].
gd; It has been proposed that all globins have evolved from a family of
gd; ancestral, approximately 17 kDa haemoproteins that displayed the globin
gd; fold and functioned as redox proteins [3].  The globin superfamily includes
gd; vertebrate haemoglobins (Hb); vertebrate myoglobins (Mb); invertebrate
gd; globins; plant leghaemoglobins; and bacterial flavohaemoglobins.
gd;
gd; The function of haemoglobins (Hb) is transport of dioxygen in blood plasma.
gd; Hb binds O(2) in the reduced [Fe(II)] state.  The Hb molecule exists as a
gd; tetramer, typically of two alpha- and two beta-globin chains, which form
gd; a well-defined quaternary structure.  Each monomer binds iron protoporphyrin
gd; IX (haem).
gd;
gd; The 3D structures of a great number of vertebrate Hbs in various states
gd; are known.  The protein is largely alpha-helical, eight conserved helices
gd; (A to H) providing the scaffold for a well-defined haem-binding pocket
gd; (Hb alpha subunits lack helix D [4]).  The imidazole ring of the "proximal"
gd; His residue provides the fifth haem iron ligand; the other axial haem iron
gd; position remains essentially free for O(2) coordination.  Conserved "distal"
gd; His and Val residues block an unhindered access to the sixth coordination
gd; site so that a controlled binding of small molecules may result only as a
gd; consequence of side-chain dynamics of the protein [1].  O(2) binding results
gd; in a transition from high-spin to low-spin iron, with accompanying changes
gd; in the Fe-N bond lengths and coordination geometry.  In Hb, these subtle
gd; changes lead to the well-known cooperative effect.  At the quaternary
gd; structure level, O(2) binding induces relative reorientation of the
gd; [alpha-1, beta-1] and [alpha-2, beta-2] dimers.
gd;
gd; Alpha- and beta-haemoglobins are highly similar; the sequence of alpha-
gd; differs in length from that of beta-haemoglobin on average by 5 residues
gd; (actual lengths 141 and 146 residues respectively).  The major structural
gd; difference between alpha- and beta-forms is that beta-haemoglobins contain
gd; an alpha-helix (the D helix) that is missing in alpha-forms.
gd;
gd; ALPHAHAEM is a 5-element fingerprint that provides a signature for alpha-
gd; haemoglobins.  The fingerprint was derived from an initial alignment of 27
gd; sequences:  the motifs were drawn from short conserved sections spanning the
gd; full alignment length, focusing on those regions that characterise the
gd; alpha-haemoglobins but distinguish them from the rest of the globin family -
gd; motif 1 includes the second alpha-helix, leading into helix 3; motif 2 spans
gd; the C-terminus of helix 3 and helix 4; motif 3 includes the N-terminus of
gd; helix 5; motif 4 spans helices 6 and 7, its C-terminal residue being the
gd; invariant proximal His; and motif 5 encodes helix 9.  Seven iterations on
gd; OWL30.1 were required to reach convergence, at which point a true set
gd; comprising 272 sequences was identified.  Numerous partial matches were also
gd; found, all of which are members of the haemglobin family:  most are beta-
gd; haemoglobins that match 2 or 3 motifs.
```

a consequence of either, the true identification of that motif in a related sequence, or a chance match to an unrelated sequence. When considering a single motif only, the hit-list may contain many false matches interspersed throughout the true matches. However, if a comparison is made between the hit-lists for two motifs, it is possible to identify sequences that are represented in both lists (i.e., the sequences contain matches to both motifs). If a motif-match is merely the result of a random event, then the chance of *two* such events occurring in the same sequence is accordingly smaller (the product of the probabilities of each of the independent events). As more hit-lists are compared, accordingly the number of false sequences matching all motifs drops rapidly.

Ideally, all true sequences should be identified as matching *all* motifs, and no other sequences should make partial matches to the fingerprint, until, at the two motif level, chance dictates that some false sequence will occur. Unfortunately, there are frequent exceptions to this ideal. The following is a list of some of the commonly encountered deviations:

- some 'true' sequences fail to match all motifs,
- some 'false' sequences match all motifs
- in some cases it is difficult to distinguish 'true' results from 'false'.

The following subsections discuss these scenarios, and provide insights into these problems and their solutions.

4.2.5.1 Some 'true' sequences fail to match *all* of the motifs in a fingerprint

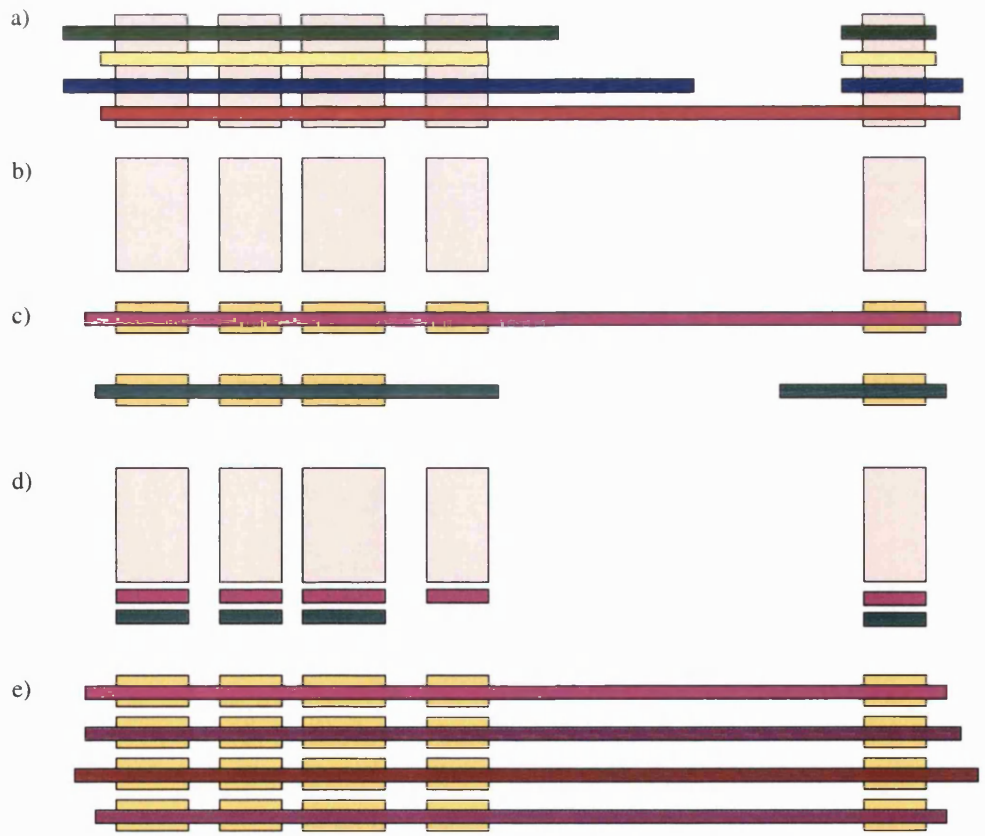
When sequences that would be expected to belong to the family appear as partial matches to the fingerprint the indication is that the model is failing for a particular reason. Some of the reasons and their solutions are explored below.

Re-iterate with augmented motifs

It is often the case, especially after the very first derivation of the seed alignment, that deviations in family outliers are not well represented by the seed. After the first iteration, this is commonly manifested as the apparent loss of motif matches from true family-member sequences (figure 4.9) . If the result of the first 'scan' and 'compare'

Figure 4.9: Augmenting the original motifs with new sequences can potentially introduce a bias into the process.

An alignment (a) yields a set of motifs (b) that can be used to search a database of sequences in order to identify matching sub-sequences. One true family member sequences matches all five, but the other misses a match to the 4th motif (c). This is observed as the apparent loss of motif 4 from this sequence (d). The indication is that the initial alignment was not representative enough to take into accounts the deviance of this particular outlier. Augmenting the motifs with only the fully matching sequence, by disregarding the partial match, could potentially lead to a biasing of the subsequent matches (e).



cycle is the concurrent loss of matches from some sequences *and* the identification of

new sequences that match the whole fingerprint, then it is usual to use the new full-matching sequences to augment the original motifs. However, in this case using these sequences could result in over-training the fingerprint. The effect of over-training is to populate the fingerprint with sequences from a particular sub-division of the family to the exclusion of other members. Consequently, subsequent iterations may force more and more 'deviant' family members out of the fingerprint.

Conversely, adding new sequence data to a set of motifs may provide sufficient diversity to allow the failing sequences to once again match those motifs. The solution is therefore to carefully monitor the effects of re-iterating, and to be aware of the potential signs of over-training (e.g., the dominance of one subtype over another).

Remove motifs

After a number of iterations, or on reaching the final set, if some members of the true family still refuse to match all motifs, this tends to indicate one or more poorly chosen motifs (figure 4.10) . Most often, this observation is confirmed by the absence of the same motif/s from all of the sequences that are failing to match all motifs. A solution for this problem is to remove the offending motif from the fingerprint. Naturally, this action reduces the length of the effective discriminator. However, the assumption is that the poorly selected motif describes a relationship that only a subset of the sequences share, and as a result of this the motif is not performing the role for which it was selected.

More complex manifestations of this situation (true sequences missing motifs) can be observed (figure 4.11) ; the solution generally requires modification of either the number of motifs and/or alterations of the regions of the alignment described by the motif/s that fail to discriminate.

Figure 4.10: If a number of sequences fail to match all motifs after a number of iterations, this can indicate a poorly chosen motif.

When an un-representative alignment (a) yields a set of motifs (b) and they are used to search a database of sequences, the matching sub-sequences may exclude a particular group of true family members. If continued augmentation and iteration fail to identify this group as fully matching members (c and d), the indication is that a motif may have been incorrectly selected.

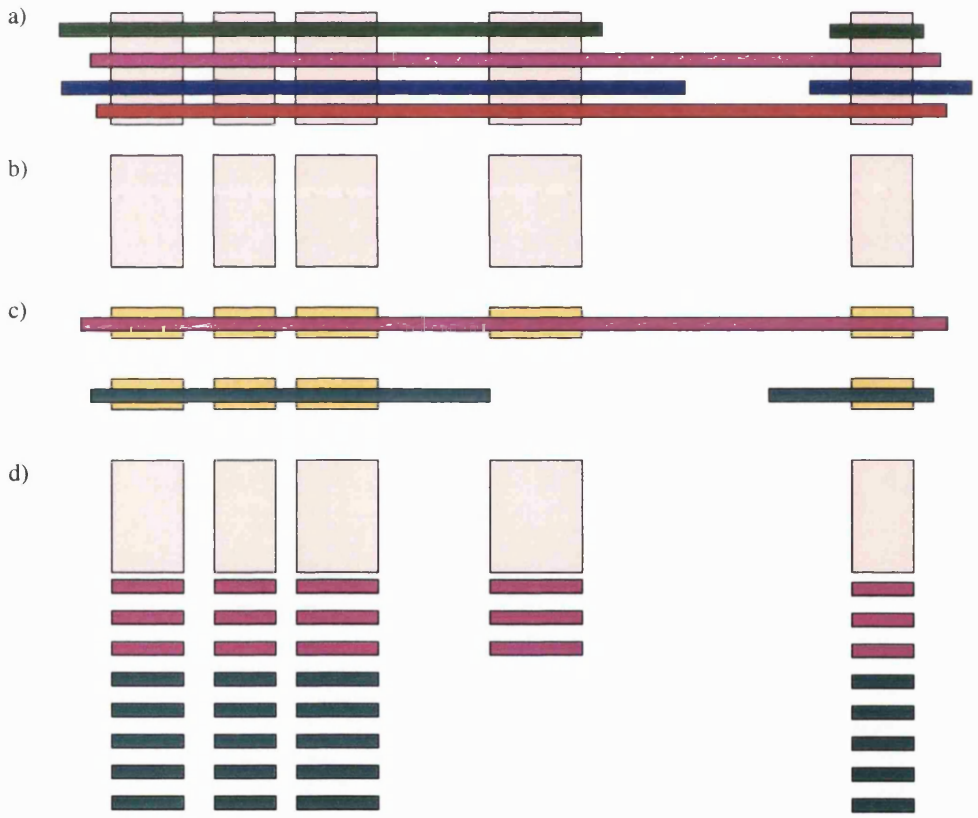
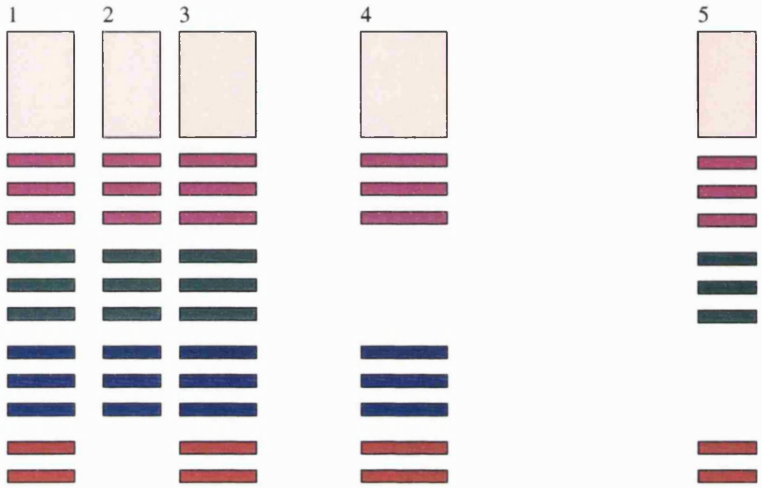


Figure 4.11: The simple examples shown in figures 4.9 and 4.10 may occur in combination, and can result in the poor representation of more than one outlier group.

Complex patterns of poorly discriminating motifs can result from a combination of poorly chosen motifs and over-represented sub-families, often the only solution is to re-evaluate the seed alignment in the light of this new information.



4.2.5.2 Some 'false' sequences match all motifs of a fingerprint

This is always an untenable situation, because the *definition* of a family is based on its constituent sequences being exclusive members of a set that matches *all* motifs. While it may be acceptable for a 'true' member to be missing one or more motifs, it is unacceptable for a non-member sequence to match all motifs. The reduction in diagnostic power that this represents can be attributed to two problems: the number of motifs is too few to discriminate the family over background noise, or the motifs fail to distinguish the family from a closely related family. The solution in all cases is to significantly re-evaluate the alignment and the selection of conserved regions. In most examples, selecting a greater number of motifs will solve the problem of false sequences matching too many motifs. However, if it persists as the number of motifs increases, it may have more complex roots. The problem may stem from one or more of the following failures.

Ill definition of the family

The initial set of sequences (members of the target family) may not be representative of the total extent of the true biological family; i.e., the target family may be larger, or more deviant than expected. If this is indeed the case, then the first iteration will uncover new sequences (which is exactly what happens as part of the standard iterative process). However, iteration can result in the identification of not just members of the target family, but related *families*, or more complex super-family, family and sub-family relationships. If such new relationships are discovered, then the effect on the distribution of matching sequences may be unexpected (the different groups of sequences matching different patterns of motifs in figure 4.11 may well indicate the discovery of such relationships). In this case, what is required is a re-appraisal of the target family. A suitable re-definition might be to include a closely related family so that the fingerprint represents a super-family relationship (motifs 1 and 3 from figure 4.11 are matched by all groups of sequences (potential families), selecting these to define a fingerprint would allow all groups to be identified). Another situation may call for the strict exclusion of co-related regions so as to create a fingerprint that is selective for a single sub-family. This re-evaluation will in most cases require inclusion of new sequences into the alignment process (and into the seed alignment) and re-excision of motifs.

Ill definition of the conserved regions

Regions that appear conserved within an alignment may also occur with a high frequency in totally unrelated sequences. Motifs derived from these regions will consequently perform badly as discriminators. Areas of low complexity, such as poly-amino acid tracts and cysteine rich regions, are notoriously poor regions for defining motifs, due to their strong potential to score highly against random sequences. This situation is not normally responsible for producing false matches to *all* motifs, but if the number

of low complexity regions is high in comparison with the total number of motifs, then it may be a plausible effect.

Ill definition of the problem

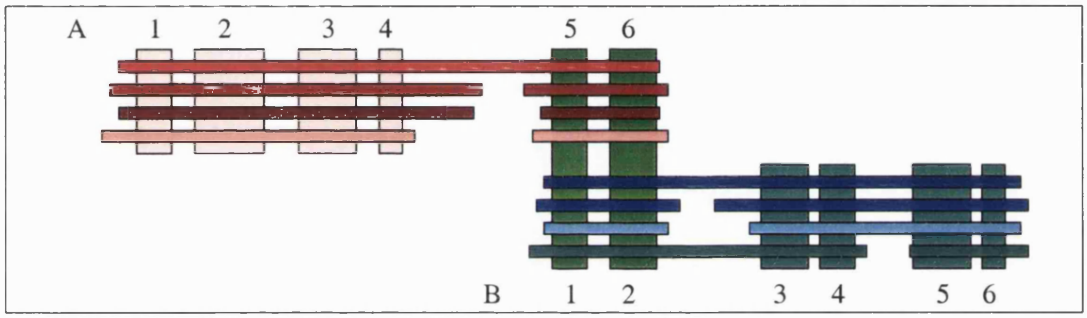
Sometimes the selection of motifs can be misguided by the restricted number of the initial set of sequences. As mentioned previously, a family can be much more extensive than initial research may indicate. Another symptom of this short-sightedness is a lack of understanding of the nature of domain sharing amongst proteins.

The objective of the fingerprinting method is to identify relatives of a family of sequences via the inference of a homologous relationship. However, the path of descent is not always the downhill process that classically defines evolution. Horizontal transfer and gene fusion are just two mechanisms, by which confusion can arise over the definition of a family membership (Fitch, 2000; Gogarten and Olendzenski, 1999). The reuse of domains, within protein evolution, is a clearly observed phenomenon throughout biology (Henikoff et al., 1997; Jacob, 1977). So, attempting to construct a fingerprint without first taking into account the potential mosaic nature of proteins (this can be manifested as the sharing of multiple distinct domains in different numbers and positions in unrelated proteins), can lead to confusing results. After the first iteration, this can be manifested as the identification of 'true' family members (those matching all motifs), that, on the basis of annotation, are considered to be false (i.e., not belonging to the target family). In this case the problem is a lack of comprehension of the scope, and potential, of the fingerprinting method.

In some cases, at the stage of creating a fingerprint to describe a given family, instances of domain reuse may be unknown or have gone unnoticed. Thus, the creation of a set of motifs from conserved regions across the whole alignment may result in the unwitting inclusion of sections of shared domain. In the worst case, the extent of the fingerprint could stretch no further than the boundaries of the shared domain. The

iterative process would obviously fail to discriminate sequences belonging to the 'true' family from sequences that also share the domain, hence, the initial observation would be that many 'false' sequences match the fingerprint (figure 4.12).

Figure 4.12: An illustration of two alignments (families A and B) is shown, each containing six motifs. The shared motifs (A:5 and 6, B:1 and 2) fall in a common domain. Selecting all six motifs, when describing either A or B will result in partial identification of the other. Selecting only the shared motifs, independently, from either alignment will result in a fingerprint that makes no distinction between the families.



The solution to this situation is to reconsider the problem in the light of the new information, i.e., motifs selected from a commonly shared domain will not distinguish members of one family from the set of proteins that also share that domain. Rather than identifying motifs that selectively identify a a particular subset (the original target family), it is equally valid to consider modelling the shared domain so as to create a new target family, in which the defining characteristic is the possession of this common domain; the fingerprinting process can support the definition of either.

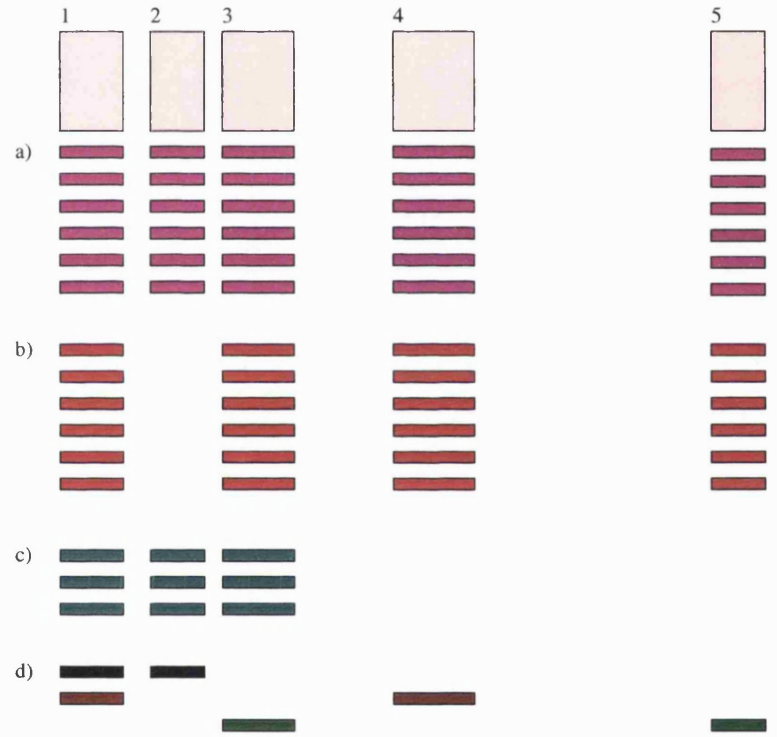
4.2.5.3 It is difficult to distinguish false from true

In attempting to define a fingerprint one of the most commonly observed problems is the difficulty associated with clearly distinguishing target family member sequences from false matches. This is often manifested as 'false' sequences matching more than one motif and 'true' sequences matching less than all motifs, and is apparent for all the reasons discussed in the previous sections (poorly selected motifs, poorly defined target family, cross reaction with related families, etc.). The illustration in figure 4.13

describes such an example, where three types of partially matching sequences blur the distinction between true matches and false matches. Sometimes the perfect result

Figure 4.13: Partially matching sequence can blur the distinction between true and false.

In this example three types of partial match are illustrated. Motifs defined to describe the target family (a) appears to cross-react with members of a closely related family (b). Also some fragmentary sequences from the target family, match only the first three motifs (c). And finally, a number of unrelated sequences make partial matches to two motifs (d).



cannot be obtained, even after many iterations and motif alterations. In these cases, it is necessary to strive to achieve a degree of separation between true and false, even if it cannot be as dramatic as no false sequences matching any motifs and all true sequences matching all motifs.

The concept of the partial match has been discussed before; it represents the identification of a sequence that does not contain sufficient conservation to be described fully by a fingerprint. However, the presence of multiple matching motifs still indicate a significant relationship between it and the family delineated by a fingerprint. The partially

matching sequence may reflect a number of different relationships within this family of proteins; e.g., it could be a member of a distinct sub-family or a member of a closely related sibling family. Because, each of these examples refer to the identification of an homologous relationship, partially matching sequences cannot be ignored, indeed their very presence highlights the sensitivity of the multiple motif approach.

To summarise, the fingerprint clearly has the ability to describe a confident relationship by being selective and identifying only sequences that match all motifs (therefore, maintaining the distinction between true and false members). However, by also considering matches to sequences that do not demonstrate the same degree of confidence in their assignment (partial matches), a level of sensitivity can be achieved.

4.3 Creating fingerprints for the PRINTS database

As an integral part of the research of the fingerprinting process, a number of fingerprints were derived for deposition into PRINTS. This section will concentrate on a few examples that illustrate some of the points made in the previous section.

4.3.1 Haemoglobin

A suitable illustration of the problem of creating fingerprints of closely related families can be illustrated by the example of the haemoglobin family (see figure 4.14; and section 3.4.2: figures 3.17 and 3.18). Haemoglobin's primary function is bind oxygen as it diffuses into the bloodstream from the lungs, and transport it to outlying tissues, where it is released. The most abundant form of this protein exists as a hetero-tetramer of two α- and two β-subunits; these subunits share a similar fold to the monomeric oxygen binding protein myoglobin. These proteins, particularly α- and β-haemoglobin, share significant levels of similarity, as a consequence of their paralogous relationships. At the outset of this research, PRINTS contained a single fingerprint that attempted to de-

scribe the common features of these particular globins. However, because the globin family is extensive and diverse, and the haemoglobin fingerprint contained few motifs, it proved to be a poor discriminator against both non-haemoglobin globins and other unrelated proteins. When distinct families share so much in common, motifs selected for their conservation in one family are often also well conserved in the other families too. The consequence, of deriving a fingerprint using these motifs is a failure to discriminate between members; i.e., the construction a fingerprint from an alignment of α-haemoglobins, may identify myoglobin sequences and β-sequences as well as α. To address this problem, it is necessary to construct an MSA that describes both subfamilies, in order to identify regions of overlap and, more importantly, disagreement. Figure 4.14 illustrates the example with the the α-,β-haemoglobin alignment. In this example, the whole alignment is very short, although a number of suitable motifs can be identified. In particular, where an insertion in the β alignment introduces a gap into the α alignment (α motif 3 and β motif 2 in figure 4.14 [2]), a motif that appears to span the gap can function as a good discriminator against β sequences.

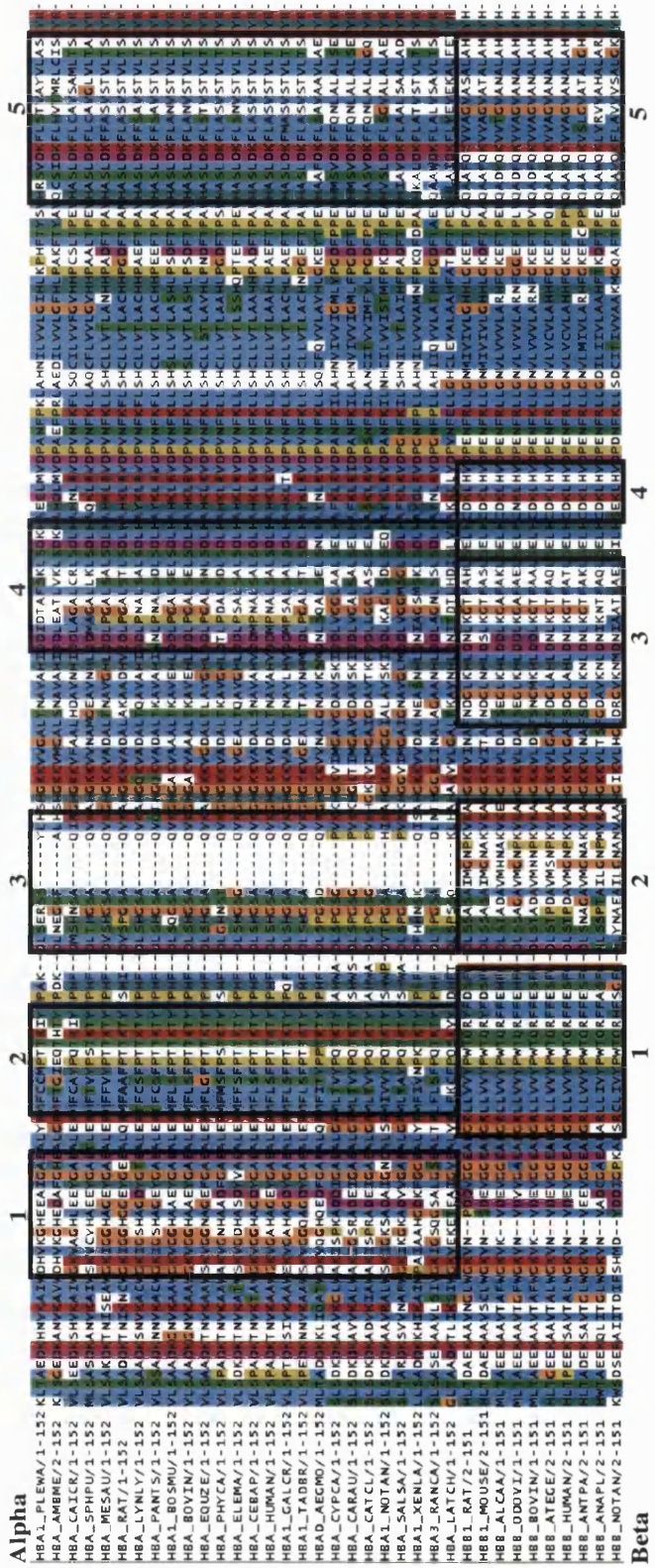
Through, the construction of a number of these alignments, the eventual result was a redefinition of the haemoglobin family, as a set of selective and discriminatory fingerprints (PRINTS:ALPHAHAEM, BETAHAEM, ZETAHAEM, PIHAEM describing the α- and β-haemoglobin families, and the ζ and π paralogues of α, respectively).

4.3.2 tRNA synthetases

Aminoacyl-tRNA synthetases catalyse a two-step reaction resulting in the aminoacylation of transfer-RNA (tRNA): the addition of an amino acid to the 3' end of a tRNA molecule. Each of the 20 amino acids is recognised by a different synthetase, and each synthetase must identify the set of acceptor tRNA substrates that hold the correct

[2] In the alignment, α motif 3 appears to contain gaps. However, this is only necessary in the α,β alignment. The motif is actually selected from the α alignment, hence motif 3 is a ten residue ungapped motif.

Figure 4.14: α- and β-haemoglobin share significant similarity. In order to define motifs to describe either individually it is necessary to produce an alignment containing both (an 'α-,β-haemoglobin' family fingerprint) and select motifs from regions where there is less commonality within the family alignment and more in the sub-family.



anti-codons - a requirement that is defined by the genetic code. The first step involves a condensation reaction between Adenosine Tri-Phosphate (ATP) and an amino acid, followed by a reaction between the aminoacyl-adenylate and the tRNA molecule, releasing Adenosine Mono-Phosphate (AMP) and the aminoacyl-tRNA molecule. The catalytic regions that give rise to these two distinct functions are separated along the length of the sequence in these proteins (Schimmel and Ribas De Pouplana, 2000). It is postulated that during the evolution of these proteins, the functional units evolved as separate entities, which co-operated to provide the dual functionality of the modern proteins. The common role of the aminoacyl-adenylation, and the modularity of all extant synthetases provides a strong argument for the creation of modern tRNA-synthetases via fusion of these two components. The adenylate synthesis domain is considered to represent the ancient functional protein, to which domains have become fused that stabilise and specify interactions with the tRNA molecule (reviewed by Schimmel and Ribas De Pouplana (2000)).

Here we look at the case of fingerprinting the tRNA synthetases, which effectively illustrates the modular nature of these proteins. In defining fingerprints to describe these proteins (and for modular proteins in general, see figure 4.12), care must be taken to consider the evolutionary relationships of each distinct domain separately.

In tRNA synthetases, the catalytic domain plays no role in specifying the binding of the correct amino acid and its cognate anti-codon bearing tRNA. Each of the paralogous sub-types contain domains specialised for this recognition role. To describe the family of proteins that function specifically as valyl-tRNA synthetases, it is essential to describe those features that distinguish it from all other tRNA synthetases. Therefore, a clear understanding of the modularity of these proteins is a prerequisite. Defining motifs from a naive alignment of valyl-tRNA synthetase resulted in the selection of a number of motifs, which crossed the domain boundary, the first iteration produced a telling result. While, several of the results corresponded to the recognition of the correct amino-acid (valine), a large number correspond to proteins sharing the

common domain (data not shown). Clearly, selecting motifs from the shared domain (characterised by its 'KMSKS' motif) of the protein (see figure 4.15) will result in the identification of more than just the valyl-tRNA synthetases. To resolve this, either those cross-reacting motifs should be lost, or if this would result in too few motifs, new motifs should be drawn up. In fact resolution, of this problem was actually better achieved through the construction of an alignment containing members of the cross-reacting families, so as to design motifs that specifically avoided the shared domain.

The final result of this work, was the derivation of a set of amino-acid specific tRNA synthetase fingerprints, each with the inherent ability to identify sequences belonging to its own specific subtype and no other (figure 4.16).

4.4 Using fingerprints

Once the fingerprint has been created and annotation has been researched and put in place, its utility as a discriminator can be realised. The objective is to identify the signature in the sequence of a uncharacterised protein, thereby revealing its relationship to a family of functionally characterised proteins. Using a fingerprint in this way requires software capable of identifying matches to motifs in a query sequence and determining their significance. This role was originally fulfilled by X-finger (Perkins and Attwood, 1995).

Given a query sequence, X-finger both identifies top-scoring matches to individual motifs and top-scoring *full* fingerprint matches. These results are scored and rank ordered using the n-single scoring scheme (section 4.2.3.1 and figure 4.5).

It was made clear in the previous discussion of fingerprinting (section 4.2.5) and single-motif methods (sections 3.3.4.1 and 3.4.2) that single motif matches, while potentially illuminating, often fail to discriminate true from false matches. PRINTS motifs are designed to function as complementary pieces of evidence, which perform as a whole

Figure 4.15: An alignment of the shared 'KMSKS' domain of three families of the tRNA synthetases.

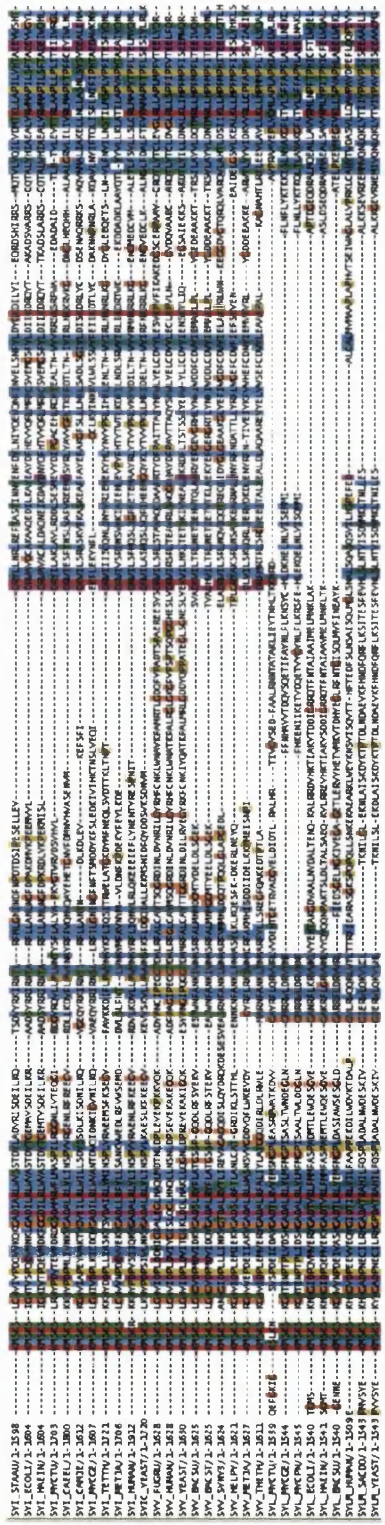


Figure 4.16: The set of tRNA synthetase fingerprints in PRINTS.

PRINTS accession	PRINTS identifier	Family name
PR00980	TRNASYNTHALA	Alanyl-tRNA synthetase signature
PR00981	TRNASYNTHSER	Seryl-tRNA synthetase signature
PR00982	TRNASYNTHLYS	Lysyl-tRNA synthetase signature
PR00983	TRNASYNTHCYS	Cysteinyl-tRNA synthetase signature
PR00984	TRNASYNTHILE	Isoleucyl-tRNA synthetase signature
PR00985	TRNASYNTHLEU	Leucyl-tRNA synthetase signature
PR00986	TRNASYNTHVAL	Valyl-tRNA synthetase signature
PR00987	TRNASYNTHGLU	Glutamyl-tRNA synthetase signature
PR01038	TRNASYNTHARG	Arginyl-tRNA synthetase signature
PR01039	TRNASYNTHTRP	Tryptophanyl-tRNA synthetase signature
PR01040	TRNASYNTHTYR	Tyrosyl-tRNA synthetase signature
PR01041	TRNASYNTHMET	Methionyl-tRNA synthetase signature
PR01042	TRNASYNTHASP	Aspartyl-tRNA synthetase signature
PR01043	TRNASYNTHGLY	Glycyl-tRNA synthetase signature
PR01044	TRNASYNTHGA	Glycyl-tRNA synthetase alpha subunit signature
PR01045	TRNASYNTHGB	Glycyl-tRNA synthetase beta subunit signature
PR01046	TRNASYNTHPRO	Prolyl-tRNA synthetase signature
PR01047	TRNASYNTHTHR	Threonyl-tRNA synthetase signature

in a fingerprint and not individually. Therefore, merely listing top-scoring individual motifs is prone to yield, in all but the clearest of cases, a confused result, particularly, in the presence of motifs that can score highly against a sequence purely by chance. Take the example of a motif designed to represent a cysteine-rich region of a protein. In context, an individual motif detects sequences similar to itself within the confines of its neighbouring motifs, which anchor it into a specific region of a specific family of proteins. Released from these restrictions, the motif is free to identify *any* cysteine rich region. The fingerprint is a composite structure, and therefore, using the combined evidence of its components is clearly a better way to identify sequences with which it may share similarity. Consequently, the view of the whole fingerprint plays an important role in the identification of similarity. Unfortunately, as discussed in section 4.2.5, relationships between sequences are often not clear enough to be described in such 'cut and dry' terms as the requirement for a sequence to match all motifs of a fingerprint. It is highly likely that a sequence, matching 9 of 10 motifs, is a true mem-

ber of the family, and its diagnosis should not be occluded by a strict adherence to a rigid identification policy. By restricting 'true' diagnoses to those matching all motifs of a fingerprint, X-finger risks missing family members, and thus fails to exploit the inherent sensitivity of the fingerprinting method.

The software provides a facility to view the performance of matches to individual motifs in a graphical format; however, there is a drawback to this representation. Matches are plotted on these graphs using the n-single scoring scheme, which renders scores to motif matches as a percentage of the highest achievable score (for that motif). While, this means that different matches to individual motifs can be compared, and that significant scores can be evaluated, comparisons between motifs in a single fingerprint cannot be made, as indeed nor can comparisons between different fingerprints be made. This means that it is difficult to evaluate those very cases where this kind of analysis would be necessary: i.e., determining whether a sequence is more likely to belong to one family rather than another.

In short, X-finger is best suited to the visual comparison of sequences within a family, which makes it a useful tool for use in the development cycle of the fingerprinting process. However, the PRINTS database is of little use to the wider biological community without a suitable analysis tool that can: facilitate rapid and automated analyses of multiple sequences; and provide a means of evaluation that allows results to be interpreted. Furthermore, a tool that could detect, and use, the biological context of motifs within a fingerprint to improve the rejection of false diagnoses, and make use of the increase in sensitivity afforded by the detection of partial matches, would better exploit the potential of fingerprints.

The realisation of this goal has been the main objective of this thesis and has culminated in the development of a new search tool. The following chapter will discuss the development of the PRINTS database searching software.

Chapter 5

Methods

5.1 Aim

Before the inception of this project the PRINTS database lacked a search method capable of doing justice to the powerful contextual information inherent in multiple motifs. The only available search method, X-finger (Perkins and Attwood, 1996), was not able to identify partial matches, and hence missed all but the most specific diagnoses. The challenge therefore was to provide a search method designed from the outset to take into account the multiple motif model.

5.2 The development of a new search tool for PRINTS

The development of a fingerprint searching tool, requires a deconstruction of the methods which underly the process of identifying and scoring a pattern in a sequence. The principle of the overall process, is based on the assumption that if patterns of conservation are observed in an MSA then the identification of these same patterns in an uncharacterised sequence can lead to the inference of an homologous relationship.

Firstly, the process must provide a means of aligning a pattern with the sequence in such a way that all possible alignments are evaluated. In pairwise analysis this is achieved through the construction of the alignment matrix. However, identifying a fixed-length pattern involves a simpler sliding window approach, which scans across the entire length of the sequence, revealing at each step a different, overlapping subsequence. As the fingerprint consists of more than one pattern, this procedure must be repeated for each of the motifs.

Secondly, as each sub-sequence is revealed the alignment made between it and the motif must be scored. This requires the construction of a model suitable for describing the motif and providing a score for its alignment. The simplest model is the identity matrix, which provides a score (+1) for each residue of the sub-sequence that is present in the corresponding column of the motif. As discussed in sections 2.4.2 and 3.3.4 a

variety of scoring schemes exist, which improve on the identity matrix.

Thirdly, once each sub-sequence is scored, the task of identifying matches that correspond to the identification of the fingerprint must begin. Matches are made at the motif level, therefore, it is necessary to place them in the context of the fingerprint; i.e., as motifs are selected from an MSA, matches made to them should also progress linearly across the sequence. An essential factor in the consideration of fingerprint context is that tolerance should be allowed for sequences to fail to match all of the motifs, so as to create provision for the partial match. The importance of the ability to partially match a fingerprint should not be understated, as it is this critical feature of the multiple motif methodology that allows highly selective and/or sensitive diagnoses to be made. The identification of partial matches, however, is a non trivial task that was not addressed by the X-finger search method.

The following sub-sections, review the process of scanning, alignment and scoring of motif matches, which was outlined above, with particular emphasis on the experiments that were important in the evolution of the new PRINTS search tool

5.2.1 The scanning process

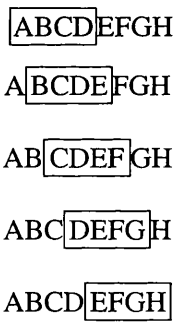
5.2.1.1 The sliding window

The scanning process mimics the effect of sliding a window along the length of a sequence (section 5.2.1.1 and figure 3.7), to reveal sub-sequences. The sliding window process essentially renders every sub-sequence in the query available, in order that every position in the sequence can be probed for the existence of a match to a given motif. However, a fixed width window does not visit every residue with equal frequency, which renders the edges of the sequence a potential source of missed information. (Figure 5.1). This is particularly important when considering fragmentary sequences, which are common in primary databases; therefore, it is important to consider the entirety of the query sequence as a potential site for the identification of a motif, and not

to treat the edges differently. If a fragment is cleaved within a conserved regions, the motif may be missed and the important similarity will go unnoticed. A simple solution is to pad the sequence with non-scoring tokens[1] to allow each sequence element equal chance of identification with the sliding window (Figure 5.2). This token merely acts as a potential amino acid and allows partial motif matches to be recognised. Padding creates $l + m - 1$ motif sized windows from the sequence, for comparison with the motif, where l is the length of the sequence, and m is the length of the motif.

Figure 5.1: A query sequence is not evenly represented by a fixed width window.

The query sequence is unevenly represented in the sub-sequences, with only residues D and E occuring in each position of the window.



Each sub-sequence is scored against a matrix computed from the motif. The encoding of the motif and its scoring of the query sequence can be performed in a multitude of ways, as discussed in section 3.3.4; the following sections will review the two paradigms that became important in the development of the scoring component of the search method.

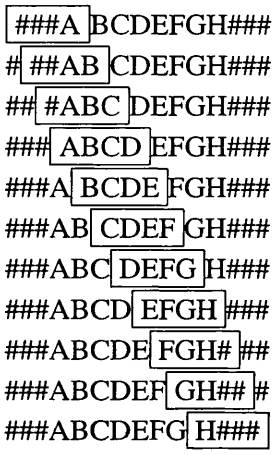
5.2.1.2 The frequency matrix.

In the earliest implementations of scoring methods for the identification of motif matches, the basic frequency matrix was used (its construction was described in section 3.3.4.2).

[1] A non-scoring token is character that does not belong to the alphabet of protein sequences, and therefore can never be scored.

Figure 5.2: The result of padding the edges of a sequence is an even representation.

In this example it is clear that every residue of the query sequence is exposed by the window an equal number of times, and has the opportunity to visit every position of the window. Without padding, partial motif patterns such as XABC or FGHX (where X is any amino acid) would never be identified.



A motif represented in this way comprises a matrix of scores based on the number of occurrences of a particular residue in each column of the motif, expressed as a normalised score (see section 3.3.4.2 and figure 3.11). Scores for matching residues in each column are distributed between 0 and 1, and the matrix is sparse, i.e., most residues have no score.

5.2.1.3 The profile matrix.

Later incarnations of the scoring methods were based on a modified version of the 'Gribskov profile' (Gribskov et al., 1990), (as described in section 3.3.4.3), in which gap-scoring features are not considered. Representing a motif using this model produces a matrix, in which the rows correspond to members of the alphabet, and columns to positions across the motif. Scores are based on the frequency matrix, but with additional weighting from substitution matrices, so they provide log-likelihood values; consequently, positive scores indicate likely residues and *vice versa*.

5.2.2 Scoring a sequence

Scoring a sub-sequence, against each of the scoring matrices was described in detail in section 3.3.5. Again, important consideration that affected the development of the search method are reviewed below.

5.2.2.1 Scoring a sequence - the frequency matrix

The PRINTS methodology has long been based on a weighted scoring scheme known as n-single (see figure 4.5), however, it has one significant drawback. N-single scores provide no basis for comparison between matches to motifs either within the same fingerprint, or between different fingerprints. As one of the objectives of the development of the search is to *clearly illustrate* the significance of fingerprint matches, this renders the use of such a scoring method inappropriate.

For example, let us consider two matches made to a single motif for which the highest possible score is 40.25[2]. The first match is a maximally-scoring sub-sequence, which achieves 40.25 (ACDEKGH); hence, the scaled n-single score will be 100%. The next match is sub-optimal and only achieves 22.5 (ACRDFGI), which when scaled equals 56%. Clearly, using this scoring scheme, matches to the same motif can be compared (the first match is better than the second); however, without prior knowledge of the maximal scores, it makes no sense to compare the scores achieved by matches to different motifs. So, while producing a useful measure of the significance of multiple matches to the same motif, it is not apparent how this can be applied to the context of a whole fingerprint, and for this reason, the use of n-single scores was avoided.

Like the pairwise percentage identity, the W-PID score (derived from the normalised frequency matrix) provides a fixed range over which matches score. A perfect match, to an invariant motif produces a 100% score, while a sub-sequence that contains no

[2]In figure 3.13, the highest score achievable is $1+1+0.5+0.75+0.5+1+1=5.75$, which achieves character-matches in all 7 columns and therefore equals $5.75*7=40.25$.

residues in common with a motif will score 0%. This unique property means that comparisons can be made between the relative magnitudes of scores made by different sub-sequences, and, more importantly, it allows decisions to be made about the likelihood of a sub-sequence representing a true match rather than a false one.

Distinguishing between true matches and false matches within the fixed scale of the W-PID, can however, be a problem. When scanning a sequence with a motif, there is often an abundance of low scoring matches. The problem arises if the difference in magnitudes between true scores and false scores is not significant, which is a particular problem for shorter motifs. The preponderance of low-scoring matches to shorter motifs makes a case for the inclusion of some form of penalty to reduce this abundant noise. To meet this requirement, a scoring adjustment scheme was designed such that it selected *for* good matches and *against* poor ones. In figure 3.13, two of the matches, which are both false, only incur scores in 1/7th of the positions of the matrix, while the true match scores in all 7. Augmenting these scores, through multiplication by the *fraction* of positions incurring a non-zero score, does not affect the true score but reduces the false ones accordingly. Unlike n-single weighting, all sub-sequences that achieve character-matches in all columns are still equally comparable across all motifs (within the same fingerprint or with motifs from other fingerprints), it is only those matches that fail in one or more columns that are down-weighted. The adjusted score, Adjusted W-PID (AW-PID), for each match can be described by equation 5.1:

$$\frac{\text{Sum of matches to columns}}{\text{number of columns}} \times \frac{\text{number of columns matched}}{\text{number of columns}} \times 100 \quad (5.1)$$

The use of W-PIDs, and AW-PIDs is suited to the analysis of small sets of data, where interpretation of the results is facilitated by the judgment of a user. However, there are two obvious problems with the reliance on these scoring schemes: data sets are frequently large (genomes, EST databases, etc.), and the judgment of users varies ac-

cording to their level of expertise. Using a profile matrix instead of a frequency matrix to represent a motif, provides the facility to generate a 'probability value' indicating the likelihood of a correct match. The significance of this value can be set within defined confidence levels, which supports the use of automated systems. Automation, in turn, facilitates the analysis of larger datasets.

5.2.2.2 Scoring a sequence (the 'profile' matrix)

The 'profile' motif also provides a score that can be used to determine the significance of a match. The score, which is based on the summation of log-likelihood ratios, has two beneficial properties that promote its use over and above the PID scores. Firstly, because likely residues score positively and unlikely ones score negatively, summation of these scores provides 'true' motif-matches with highly positive scores, while 'false' matches receive small or negative scores. Secondly, as discussed in section 2.4.4, these scores fit the criteria for their conversion into probability values. The application of the EVD methodology requires that the expected score for the alignment of two sub-sequences, $\sum p_i p_j s_{ij}$, must be negative, and the scoring matrix must produce at least one positive score (see section 2.4.4.2). Applying this methodology to the scoring of sub-sequences against motifs is straightforward, with s_{ij} representing the lookup between residue j in the sequence and the score for its comparison with residue i in the profile matrix.

Therefore, using the profile matrix to generate a score for a match means that less false matches should be identified, and the significance of those remaining matches can be evaluated statistically.

Using these motif-match scores (or probabilities) to determine the significance of a match between a sequence and a fingerprint (where multiple motif matches must be taken into consideration) will be discussed in the following section.

Sections 5.2.3, 5.2.4 and 5.2.5 detail the novel research that was required to address

the unique problem of how to deal with the fingerprint context of multiple matching motifs.

5.2.3 Dealing with matches

Scanning a set of matrices (frequency or profile) against a sequence produces lists of scoring matches (one list from each of the motifs), from which the true fingerprint match must be extracted. The information available for each match is as follows:

- the motif number (each motif in a fingerprint is labeled numerically from the N-terminus),
- the position of the match in the sequence,
- the score (and, for profiles the p-value).

In some cases, the identification of the true fingerprint match cannot be achieved by simply selecting the best scoring matches to each of the motifs. As demonstrated in the discussion of REs (section 3.3.4.1), short motifs have a tendency to be matched by random sub-sequences as readily as they are matched by true sequences: i.e., the distinction between a true match and a false one is not significant. Accordingly, it can be very difficult to predict whether a sequence *truly* matches a fingerprint from mere consideration of matches to individual motifs.

An important effect of the adoption of profile scores, which can be appreciated by examination of the lists of matches, is a clarification of the distinction between true positive and false positive assignments. Figure 5.3 illustrates the effect that changing the scoring regime has on the ability to clearly separate true from false results, in which, the sequence of an Edg-1 orphan receptor (SWISS-PROT:P21453) was scanned against two selected fingerprints. Each of tables show the ten top-scoring matches to motifs of both the rhodopsin-like G-Protein Coupled Receptor super-family (GPCR)

and the unrelated Beta-Lactoglobulin family (BLC) fingerprints[3]. The selection of these particular fingerprints demonstrates the identification of a previously-known relationship and its comparison with a wholly unrelated family; i.e., P21453 is distantly related to GPCR, but not to BLC. The list in the first table of figure is sorted by the W-PID score (the first two columns describe the fingerprint and the motif number that made the match), it is clear that matches to the second motif of the BLC fingerprint dominate the table, pushing true results down below the threshold[4]. Using the AW-PID score (the second table in figure 5.3) suppresses most of the spurious results that overwhelmed the previous example. However, the top scoring sub-sequence still hits the BLC family. In contrast, the third table shows the list sorted by scores generated by the profile method, where the top three matches are true motifs and only two false matches appear in the list.

This less than perfect result highlights the pitfalls of considering single motifs sufficient to describe complex familial relationships, while simultaneously demonstrating the weaknesses of the scoring scheme based on the frequency matrix. The solution to the latter point is clearly to pursue use of the profile matrix. The utility of this approach is further highlighted if the p-values for these scores are calculated. BLC motifs 2 and 5 have p-values of $5 \times 10^{-3}$and 3×10^{-3} respectively, whereas the equivalent values for the top four GPCR motifs range from 2×10^{-8} to 5×10^{-9}, up to six orders of magnitude smaller and thus considerably less likely to have occurred by chance.

The example in the third table, of figure 5.3, also provides an indication of how simple contextual information can help confirm the true result from this list of motifs. Each of the seven motifs of the GPCR fingerprint are present in the list of matches, and each match is sequentially positioned along the sequence, just as the motifs must be in the original alignment. Clearly, if confidence in the scores of matches cannot be provided, the additional information available from the MSA about their number, and relative

[3]The fingerprints are GPCRRHODOPSIN:PR00237 and BLCTOGLOBULN:PR01172, referred to as GPCR and BLC in the tables for brevity.

[4]Here the threshold is arbitrarily the top ten scores, which serves to illustrate this example.

Figure 5.3: Sorting a list of matches by each of the scoring schemes.
The tables contains the ten top-scoring individual motif matches revealed when a set of matches made between the sequence P21453 and the fingerprints GPCR and BLC are sorted using each of the scoring schemes.

Ten top-scoring sequences from results sorted by W-PID score

fingerprint	Motif	W-PID	AW-PID	Profile	Position
BLC	2	53	39	214	127
BLC	2	36	16	61	94
BLC	2	35	16	95	179
BLC	2	35	16	9	266
BLC	2	34	12	53	164
BLC	2	34	12	23	267
BLC	5	32	18	257	192
GPCR	6	32	32	463	253
GPCR	7	32	32	367	292
GPCR	3	31	31	353	124

Ten top-scoring sequences from results sorted by AW-PID score

fingerprint	Motif	W-PID	AW-PID	Profile	Position
BLC	2	53	39	214	127
GPCR	6	32	32	463	253
GPCR	7	32	32	367	292
GPCR	3	31	31	353	124
GPCR	2	31	31	227	80
GPCR	5	23	23	153	202
GPCR	4	22	22	195	159
GPCR	1	21	21	170	47
BLC	5	32	18	257	192
BLC	2	36	16	61	94

Ten top-scoring sequences from results sorted by profile score

fingerprint	Motif	W-PID	AW-PID	Profile	Position
GPCR	6	32	32	463	253
GPCR	7	32	32	367	292
GPCR	3	31	31	353	124
BLC	5	32	18	257	192
GPCR	2	31	31	227	80
BLC	2	53	39	214	127
GPCR	4	22	22	195	159
GPCR	1	21	21	170	47
GPCR	5	23	23	153	202
GPCR	6	15	13	102	47

positions can be used to improve confidence in the diagnosis. Therefore, in order to enhance the detection of meaningful results (from lists of matches), requires identification of the *context* within which motifs reside and, in *suitable* cases, to combine the evidence of multiple complementary matches.

5.2.3.1 Identifying fingerprint context

Fingerprints are derived from aligned families of sequences; therefore, their constituent motifs have the following properties:

- motifs have 'order'; i.e., motif 1 is always followed by motif 2 and never preceded by it.[5]

- motifs have fixed positions within the alignment.

The problem of how to identify the matches that conform to the 'fingerprint context' of motifs must be solved in the most efficient manner, because scanning each motif against a sequence produces numerous matches, most of which, by definition are spurious. The simplest solution is to search all permutations of matching motifs to identify those that best fit the model. However, this is computationally intensive and rapidly becomes time consuming as more matches are encountered. To avoid these difficulties, a single-pass algorithm was developed (one that requires only a single traversal of the list of matches to find the best answer), which uses a trade off of time for memory to facilitate analysis of a list of n matches in linear time $O(n)$. The following section will describe the algorithm and detail its development.

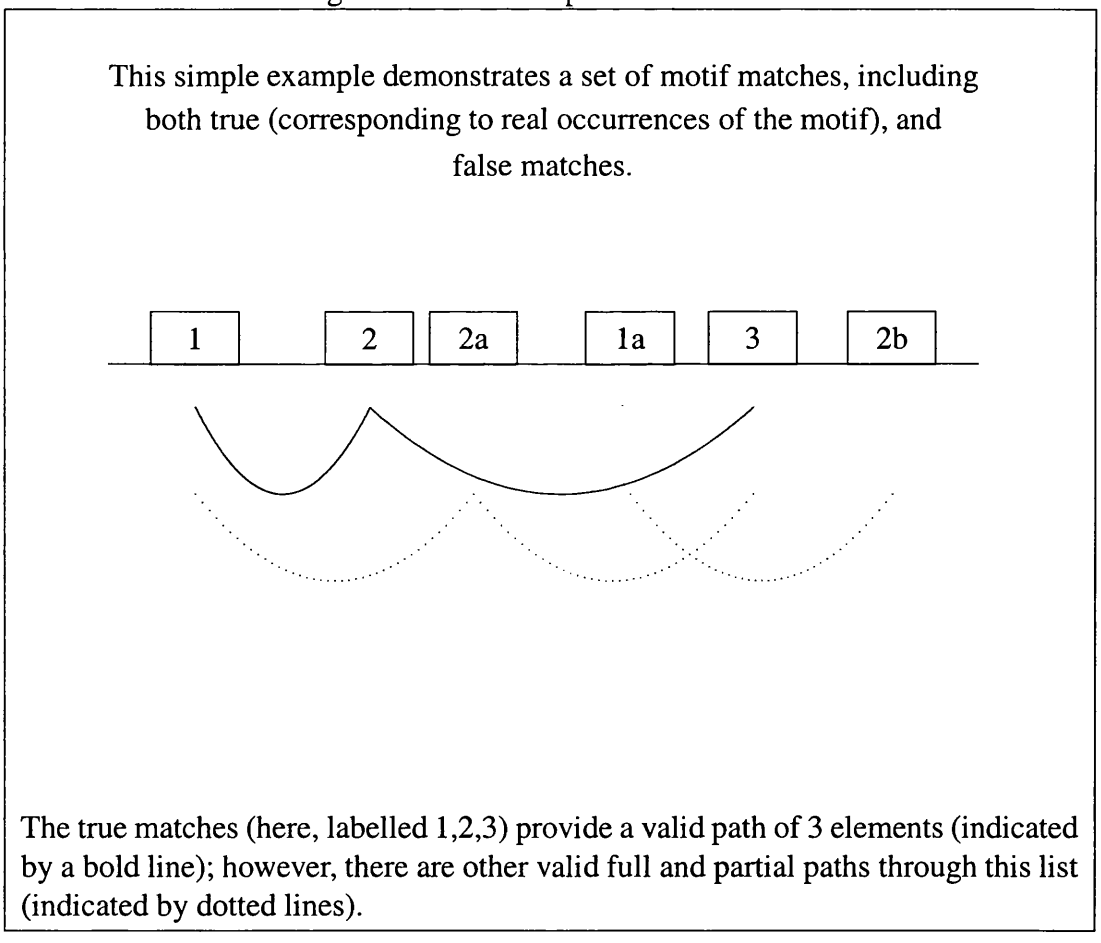
[5] A worthwhile caveat is that this is not strictly true in the case of proteins that contain repeating domains, where a motif from one repeat could be considered out of order if compared with a motif from a different repeat. However, this statement does hold if each repeat is considered as a distinct entity in its own right.

5.2.4 The PathFinder algorithm

5.2.4.1 Finding the longest path

The objective of this algorithm is to identify, for a given query sequence, the longest scoring 'path' through the matches made by a single fingerprint. A path is merely a correctly ordered list of matches, and the hope is that one such path describes the 'true' set of matches that represent the correctly identified fingerprint. The relationship between paths and matches is shown in figure 5.4. The list of matches is sorted on the

Figure 5.4: An example list of matches.



position of the match (first match from the beginning of the sequence first, and so on) and is presented as a list of motif numbers, i.e., matches to motif *n*. Figure 5.5 describes the list of matches and an overview of the process of identifying the longest paths.

Figure 5.5: The PathFinder process

A list of matches contains the elements shown in the table (the list is sorted by position); however, to compute the best path, only the position of the match and the motif number are required.

sub-sequence	position	matching motif	score
AAAAAAAAAA	20	1	595
AAAAAAAAAA	55	2	560
AAAAAAAAAA	80	2	345
AAAAAAAAAA	120	1	395
AAAAAAAAAA	154	3	795
AAAAAAAAAA	189	2	295

During the PathFinder process, the list is traversed by selecting one element at a time. Each element is compared with each of the paths that are available to it, and a decision is made about whether to add it to a growing path or merely to create a new path (with it as the only member).

List	A growing list of paths through the list of matches									
1	1									
2	1	1,2								
2a	1	1,2	1,2a							
1a	1	1,2	1,2a	1a						
3	1	1,2	1,2a	1a	1,3	1,2,3	1,2a,3	1a,3		
2b	1	1,2	1,2a	1a	1,3	1,2,3	1,2a,3	1a,3	1,2b	1a,2b

In this table, multiple occurrences of elements are labelled a or b for clarity; however, they still represent matches to the same motif.

The procedure employed to identify the greatest scoring path is straightforward. Starting with the 1st element (see step *a*, figure 5.6), the list is traversed asking simply 'whether the current motif number is greater than the greatest in each of the paths' (in the case of the first element, there are no paths with which to compare, so this statement is found to be untrue):

Figure 5.6: Details of the PathFinder process

Step	Element	Current paths				New paths to add			
a	1	-	-	-	-	1	-	-	-
b	2	1	-	-	-	1,2	-	-	-
c	2a	1	1,2	-	-	1,2a	-	-	-
d	1a	1	1,2	1,2a	-	1a	-	-	-
e	3	1	1,2	1,2a	1a	1,3	1,2,3	1,2a,3	1a,3
f	2b	1	1,2	1,2a	1a	1,2b	1a,2b	-	-
		1,3	1,2,3	1,2a,3	1a,3	-	-	-	-

- If it is not true, then a new path is created with the current element at the head (so the first path is created with the first element as its only member). For example, in step *d*, motif 1a cannot be added to any of the current paths, and is thus represented as a path in its own right.

- If an element can be added to a path, then the current path is augmented by adding the current element at the head (e.g, in step *b*, motif 2 naturally adds to a path containing motif 1 to create a new longer path (1,2)).

To facilitate the exploration of *all* possible paths it is necessary to duplicate any path before augmenting it, an example of this occurs in the transition between steps *b* and *c*, where both paths now exist (the original '1' and the new '1,2'). This process generates a number of paths that grow as each 'valid' match is encountered until the final match

in the list. The result is a list of all paths (potential fingerprint matches) - the highest scoring of which is expected to be the identification of matches to the real motifs.

The use of this method to remove false-positive matches, both provides a biologically valid solution to the task of computationally identifying true fingerprint matches and reduces the number of chance matches to fingerprints. The probability that motif matches could occur by chance, *and also* be in the correct order may be non-negligible when considering two and (sometimes) three motifs. However, as the number of distinct motifs in a fingerprint increases, the probability that multiple matches will occur by chance diminishes, and accordingly the chance that these will *also* be made in the correct order becomes vanishingly small.

5.2.4.2 Partially matching fingerprints

The phenomenon of a fingerprint matching a sequence is clearly not one that can be described easily with a binary (black or white) result, because biology operates on all the shades of grey between these extremes; i.e., sequences are variable - some parts are more conserved than others. As a consequence of this variation, problems arise when attempting to identify distantly related members of a family.

During the accumulation of mutations in proteins undergoing the selection pressures of evolution, influence is applied such that function is maintained. However, distant relatives, from large gene families, can take on quite divergent roles and even normally conserved regions may become sites of large scale change. Motifs are usually extracted from these *supposedly* stable regions; therefore, fingerprint matches to distant relatives may fail to identify one or more motifs. Following this observation, we can define the the concept of a 'partial match' as any sequence that matches less than the full complement of motifs that make up a fingerprint. By definition, the PathFinder method does not discriminate against partial matches, and through looking at the cumulative scores of each potential path, the most probable result will still be the identification of

the 'true' *partial* match, rather than a chance combination of false, or random, matches.

5.2.4.3 Motif positions

In an attempt to further enhance the power of the PathFinder method, other characteristics of motifs were investigated for their effectiveness at discriminating between true and false assignments. The locations of motifs are obvious candidates for this role as, although less common than point mutations, insertions and deletions are clearly evident in mutational processes. Events such as these are low frequency occurrences, so it is to be expected that inter-motif distances will vary only slightly over short evolutionary periods. However, great variations in these distances can be observed, especially when examining relationships between distant members of super-families. A method that utilises this criterion in the selection of true and false matches must, therefore, be subjected to careful consideration in the selection of suitable thresholds.

The first and simplest criterion that allows for the rejection of a false match is the 'overlapping rule'. During the construction of fingerprints, motifs are selected that are adjacent to each other and are only allowed to overlap by one residue. Therefore, matches that overlap by more than this can be rejected on the grounds that they cannot coexist.

Secondly, the multiple alignment from which a fingerprint is drawn can be seen as a fixed entity about which sequences slide to align the conserved regions. Therefore, as the position of any motif is maintained in the alignment, its position relative to each sequence may vary slightly. This leads to the observation of a range over which motifs *naturally* occur. Taking into account this range can also facilitate the selection of favourable matches. Clearly, a match to a motif found several hundred residues from the N-terminus, when the motif normally occurs within a few positions of this terminus (with very little observed deviation in the natural population), would appear unlikely to represent a true match to this motif.

5.2.4.4 Inter-motif distances

A selection method, based on motif positions, was implemented and utilised in the PathFinder algorithm; however, a significant observation was made, which necessitated its modification. Modularity is a commonplace occurrence in biology, and at the protein level it results in members of protein families playing distinct parts in very different functional roles. This is manifested as the sharing of distinct domains between proteins. As a consequence, the physical positioning of a particular domain in a given sequence can be very varied. Motifs derived from such a domain may occupy many different positions in the sequences of the family (see figure 4.12), which renders the identification the natural variance of these positions unsuitable as a selection criterion. Therefore, the selection of matches based *solely* on absolute position is destined to fail for fingerprints that encode commonly shared domains, of which, due to the gregarious nature of protein modularity, there are many examples.

The aim of designing a fingerprint is to represent core islands of functionally conserved residues (motifs) within an alignment. These motifs are usually identified *within* an expanse of disorder, separated by short regions of relatively minor disorder. The concept of the domain reflects the observation that it is less likely that major rearrangements will occur in an intra-fingerprint position (i.e., between motifs) than outside the conserved region: domains are independent entities. A more consistent measure of fingerprint context can be made, by only considering deviations in the spacing *between* motifs, which avoids the problem of the variable physical location of the domain.

The computation of an inter-motif distance (D) between two motifs n and m is straightforward, and is described in equation 5.2:

$$D = interval_n + \sum_{i=m+1}^{n-1} interval_i + length_i \tag{5.2}$$

where *interval* and *length* are defined as the distance between motif i and $i-1$, and

the length of motif i respectively. To make this approach more flexible, two features are taken into account. Firstly, as there are upper and lower bounds on the position of a motif in an alignment, there must be corresponding limits on inter-motif distances. Secondly, each MSA can only ever be a sample of the evolutionary history of a particular family, and thus may not represent all possible ranges of deviation. The former issue is resolved by recording the upper and lower limits of deviation observed in the alignment, and accepting only matches that are found to occur within this range. The latter issue is a complicated one, but a satisfactory solution can be found empirically. The problem arises because the physical collection of biological sequences has, historically been culled from a relatively biased set of 'experimentally interesting' organisms. Therefore, rejecting matches based on the observations of such a limited dataset is potentially dangerous, and could result in the process becoming *too selective*. A solution is to introduce a variable degree of leniency around the upper and lower limits imposed on inter-motif distances (e.g., plus and minus a user-defined percentage of the average inter-motif distance), which permits selectivity and sensitivity to be modulated. The calculation (detailed in figure 5.7) takes into account the average position of the motif and computes a value based on the user-defined parameter which is taken off the lower limit and added onto the upper.

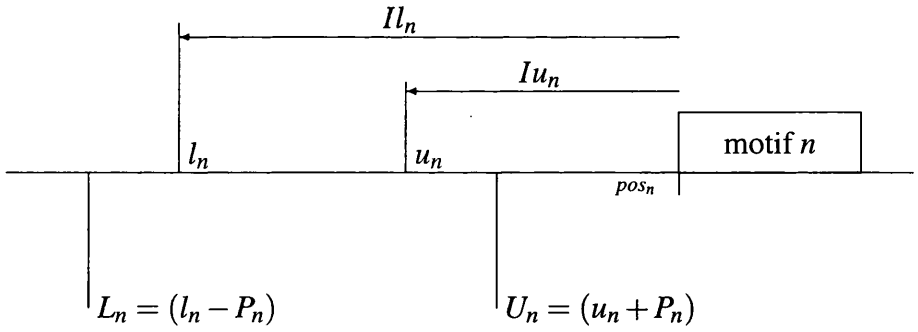
In the PathFinder process, we are interested in whether a match (at position y) can join a path that contains a motif at position x. If the motif falls within the range $L_y < x < U_y$, then the match can indeed be considered a valid addition to the path, and thus potentially a 'true' match to a motif in the fingerprint.

5.2.4.5 Summary

The adoption of selective criteria such as motif position and inter-motif distances facilitates the detection of true matches by allowing the rejection of numerous spurious matches. This in turn simplifies the identification of true matches, which can be seen as an enhancement of the selectivity of the approach. Increases in selectivity notoriously

Figure 5.7: Derivation of the inter-motif limits

The observed upper and lower limits, u_n and l_n respectively, are used to calculate the region ($L_n < m < U_n$) within which the previous motif (m) must have occurred, for the match to motif n to be considered *true*. Iu_n and Il_n are the upper and lower limits for the intervals, or distances, between the previous motif and n. From these intervals an average $I\mu_n$ is calculated, which after multiplication with a scaling factor, is subtracted from l_n and added to u_n.



$I\mu_n = \frac{Il_n + Iu_n}{2}$ the average interval

$P_n = I\mu_n * \%$ multiplying the average interval by the percentage deviance produces the value required to modify the upper and lower limits.

reduce sensitivity. However, by taking into account features such as partial fingerprint matches, variability at the residue level, and the positional variability of motifs in alignments, the sensitivity of the method is maintained.

Once a path has been constructed, the scores of each of the matches are taken into consideration to provide a corresponding path-score (or fingerprint score). In the case of scoring with a frequency matrix, it is the sum of these scores that characterises the total score of the fingerprint. During development, a number of scoring methods were investigated in order to provide the clearest distinction between the score of a chance occurrence, and the score that represents a true match to a fingerprint. The following section will describe the pitfalls of some of these methods and their eventual replacement with product probabilities.

5.2.5 Scoring fingerprints

As previously discussed, a motif is usually too short to prove statistically, or indeed empirically, that identifying it is sufficient to diagnose membership of a family. However, a fingerprint is a collection of motifs and it is the *combined* evidence of a number of motifs that characterises a family.

With scores generated from the frequency matrix method (AW-PID), the simplest method of combining them is to sum the scores; however, as each score represents a percentage identity calculation then it is equally valid to provide an average over all motifs.

Unfortunately, neither are particularly suitable for the role as the both fall short of meeting the criteria required for a scoring scheme; i.e., it must provide good discrimination between true and false scores, and provide a means for the comparative assessment of scores. This highlights the problem suffered by all methods that attempt to make functional or familial assignment based on scoring matrices, i.e., it is difficult to determine a point at which a result can be labelled true rather than false and *vice*

versa. It is the objective of any of these methods to avoid both false positives and false negatives. The perfect method would provide a clear distinction such that it was simple to define a threshold above which scores could be designated true.

The search for a scoring paradigm with which to endow the PathFinder method with such discriminatory power was littered with the experimentation with a number of methods. Each, while empirically providing some help with this decision making step, was nevertheless without mathematical grounding and subject to inconsistency. For example, if one observes that a fingerprint of ten motifs is better at discriminating its family members than one of three motifs, then conceivably the discriminatory power could be some function of the number of motifs. Logically, a score that involved augmentation of the summed score with this value may provide the necessary selectivity. Multiplication of the path-score by the number of matching motifs was indeed able to push scores for true matches higher up on the scale, while false, or poor, matches languored. Another characteristic of a likely match is the length of the model; as mentioned previously, longer motifs are less likely to attract false matches than shorter ones. A scoring scheme taking into account the lengths of motifs, which augments the score for each match by multiplying by the length of the motif, had similar results in expanding the separation between true and false. However, scoring schemes such as these have no basis other than empirical observation, and as such it is difficult to define thresholds and impossible to justify mathematically.

The solution to this problem came from the introduction of the profile matrix, and the subsequent generation of probability values to describe match scores. The raw profile score produces very low, or negative, values when the match score is close to that expected from a random match, and thus can provide a large distinction between those matches that come are true and those that are false. The absolute magnitudes of profile scores are dependent on the parameters of the EVD describing the motif. Therefore, there is no way of comparing the scores of matches to different motifs, and accordingly this makes it difficult to define thresholds. However, the motif p-value

does allow comparisons to be made between diagnoses, even if this is represented as a statement of the likelihood that a match between a motif and a random sequence would produce a particular score (see section 2.4.4.2). Matches with p-values above defined confidence levels can therefore be rejected on the basis that they tend to the random distribution, and are as a consequence likely to represent false matches. P-values of each motif in a fingerprint can be combined, because it is valid to consider each match as an independent observation, thus the product of the motif p-values describes the compound probability of all pieces of evidence (Bailey and Gribskov, 1998a,b).

This compound probability score (p-value, or e-value (the database adjusted p-value), as defined in section 2.4.4.3) provides another angle on the problem of determining the distinction between a true result and a false one. However, it is by no means a perfect solution to the problem. The EVD is an estimated distribution, which attempts to describe the distribution of MSSs (in this case, maximally scoring motif-matches) in random sequences. P-values for scores deemed to indicate significant relationships are taken from the tail of this distribution, i.e., a high scoring match will have a correspondingly low probability of belonging to the random distribution. Therefore, it makes little sense to compare a probability of 1×10^{-236} with one of 1×10^{-235}, other than to say that both results indicate a very *high* probability that a true observation has been made. In this way p-values and profile scores can achieve what the the other scores could not, a means of comparison, albeit estimated, and a good spread of results around the true-false crossover.

5.2.6 Summary

It is clear from the above discussion that no single scoring method can provide all of the answers. It is therefore important to provide as much information as possible so as to facilitate interpretation of the results. As previously mentioned, the objective of the development of these methods was to produce a search tool for PRINTS that could

present the user with sufficient information to supplement any automatic diagnoses made.

Based on the observations made in section 5.2, especially the particular advantages conferred by the used of p-values, it was decided to implement a search tool based on the p-value scoring scheme and using the PathFinder method as a mechanism to identify fingerprint context. The result was the 'FingerPRINTScan' tool (Scordis et al., 1999).

The following section describes a number of software tools that each function as interfaces to fingerPRINTScan and provide distinct analytical roles.

5.3 Interfaces

The dissemination of data is as important as its generation, i.e., a database that lacks the facility to distribute its amassed knowledge is essentially useless other than as a reference tool. FingerPRINTScan was developed to investigate the fundamental biological problem of identifying homologous relationships based on the familial descriptors that are the basis of PRINTS. The following subsections will describe the tools that provide an interface for the FingerPRINTScan software, which allow:

- remote access to the the scanning software (FPScan),
- alternative (graphical) perspectives on the results of a query (GRAPHScan),
- an automated service designed to facilitate the scanning of large numbers of sequences (MULScan).

5.3.1 FPScan

Once FingerPRINTScan made it possible to answer the question "Is it likely that a query sequence shares a relationship with one or more of the fingerprints in the

database?" it was necessary to present this information in a succinct manner. The audience of such a tool is wide and varied, with potential users ranging from molecular biologists to researchers of computational biology. Each category of user has a different agenda, and thus, has a distinct set of requirements pertaining to their goal, all of which have to be satisfied by a tool such as this one. While, one researcher may be attempting to elucidate functional characteristics of an unknown protein, another could be searching genome data for predicted open reading frames. While both are seeking homologous relationships, the former may be only interested in a simple result that supports or contradicts biochemical evidence. However, the latter may, for example, require detailed information of the positions of motifs, as well as access to a comparison between reported high scoring matches and background noise. To approach this problem (i.e., the delivery of information to users with different requirements), a hierarchical reporting system was developed.

Built using a WWW Hypertext Markup Language (HTML) interface, the submission form and results pages are widely accessible and simple to use. The submission form (figure 5.8) allows the input of either a protein-sequence database code or a raw sequence for scanning against the PRINTS database, and results are returned to the WWW browser (figure 5.9). The hierarchical arrangement of results enables the novice user to quickly discover the simplest answer to the query, and removes the necessity to wade through superfluous information. However, a user interested in how the result was obtained is provided increasing amounts of information as they scroll down the page. Three levels of information are presented:

- The top level provides the name of the matching fingerprint and the e-value of its score. All assignments at this level, are ones to which the method has designated the greatest confidence (figure 5.10).

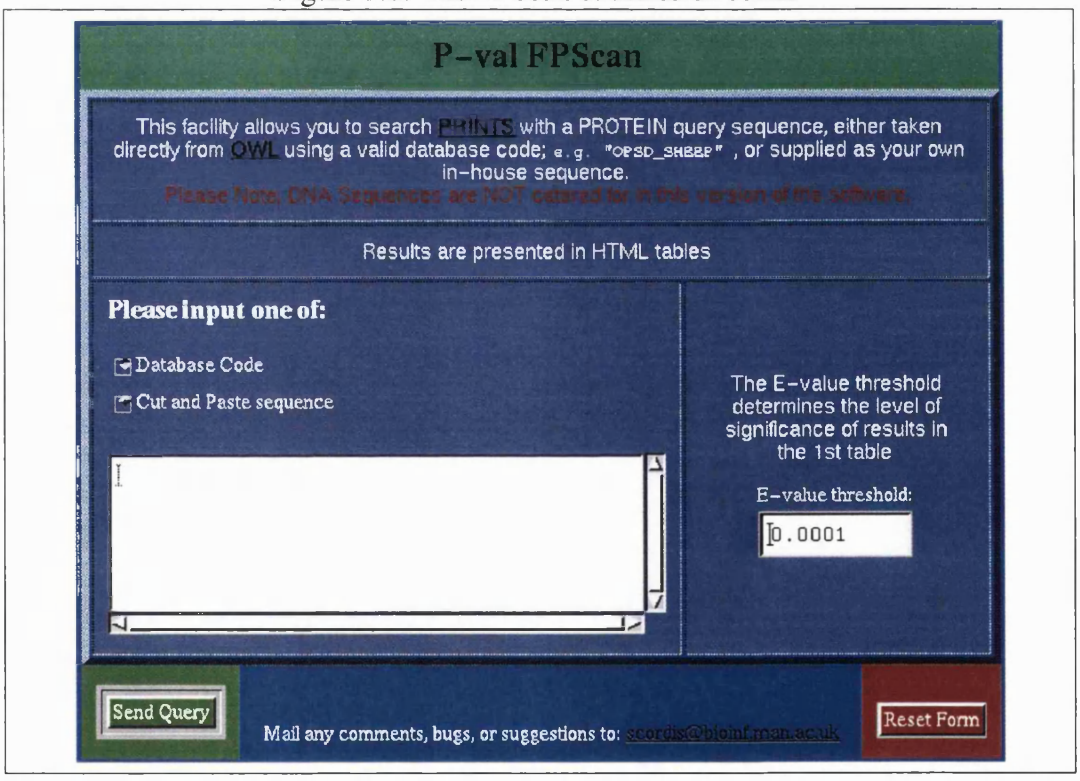
- The second level details the ten top-matching fingerprints. This usually includes the highest scoring false-matches, which are shown by means of comparison

(figure 5.11).

- The final level describes the top ten again, but at the motif level, allowing the matches made with each motif to be analysed separately (figure 5.12).

This organisation and breakdown of the results allows both the descriptive statistical measure and the comparative AW-PID scores to be displayed, and provides the user with sufficient information to supplement the on-screen diagnosis with their own opinion. In situations where the answer is not clear-cut, it becomes invaluable to have extra information at hand to back up a diagnosis.

Figure 5.8: The FPScan submission form.



5.3.2 GRAPHScan

When a distant member of a protein family is matched by a fingerprint, the profile score and corresponding p-value can be unimpressive. In cases such as these, confidence

Figure 5.9: An example result of a FPScan search, using OPSD_SHEEP as the query sequence.

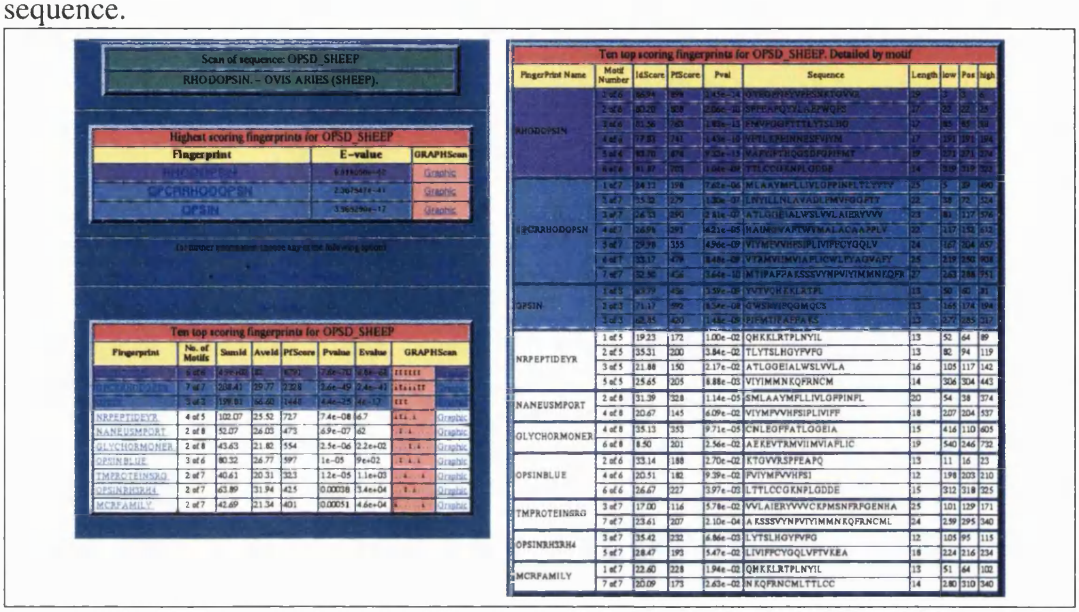


Figure 5.10: The top level of FPScan results

This table describes the top level of the results page. The first column contains a link to the matching fingerprints in PRINTS, which allows access to annotation and database links. The second column contains the e-value (the database size adjusted p-value) of the match score. The final column contains a link to the graphical representation of the corresponding match (described in section 5.3.2)

Highest scoring fingerprints for sequence: OPSD_SHEEP		
Fingerprint	e-value	GRAPHScan
RHODOPSIN	$6.82e^{-62}$	Graphic
GPCRRHODOPSN	$2.37e^{-41}$	Graphic
OPSIN	$3.96e^{-17}$	Graphic

Figure 5.11: The second level: the ten top matches.

This table describes the second level of the results page. The columns left to right contain the following information; a link to the fingerprint; the number of motifs matched (6 of 6, indicate that all available motifs have been matched); the sum of the percentage identity scores for each of the matches; the average percentage identity; the summed profile scores; the product p-value; the corresponding e-value and a link to the graphical representation of the match.

Ten top scoring fingerprints for sequence: OPSD_SHEEP							
Fingerprint	No. of motifs	SumId	AveId	Pfscore	Pvalue	Evalue	GRAPHScan
RHODOPSIN	6 of 6	490	82	4791	$7.6e^{-70}$	$6.8e^{-62}$	-
GPCRRHODOPSN	7 of 7	208	29.8	2328	$2.6e^{-49}$	$2.4e^{-41}$	-
OPSIN	3 of 3	199	66.6	1448	$4.4e^{-25}$	$4e^{-17}$	-
...	..	..	..	....	...	...	-
MCRFAMILY	2 of 7	42	21.2	401	$5.1e^{-4}$	$4.6e^{+4}$	-

Figure 5.12: The final level: the ten top matches detailed by motif

This table describes the final, and most detailed, level of the results page. Each of the ten top-matching fingerprints, listed in the previous table, is shown here broken down into individual motif matches. The information contained in the table is as follows (from left to right): a link to the fingerprint; the motif (to which the match is made); the percentage identity score; the profile score; the p-value; the sequence fragment; the length of the motif; the lowest position in the alignment that this motif is observed; the position of the match and the highest observed position of the motif.

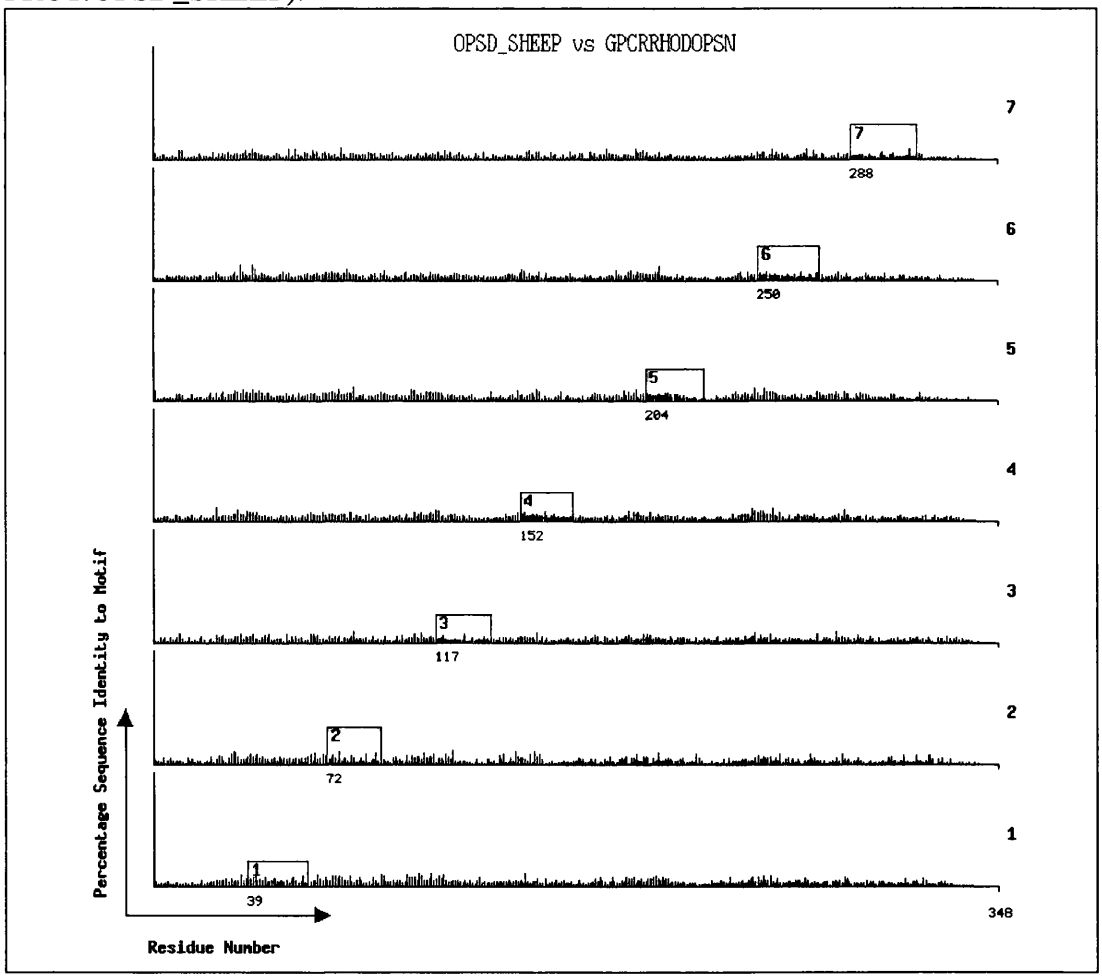
Ten top scoring fingerprints for sequence: OPSD_SHEEP. Detailed by motif.									
Fingerprint	No. of motifs	IdScore	Pfscore	Pvalue	Sequence	length	low	pos	high
RHODOPSIN	1 of 6	86.94	898	$1.45e^{-14}$	GTEGPNFYVPFSNKTGVVR	19	3	3	5
	2 of 6	80.24	808	$2.66e^{-10}$	SPFEAPQYYLAEPWQFS	17	22	22	25
...	...	..	..	...	...	..	..	..	..
	6 of 6	81.87	703	$1.04e^{-9}$	TTLCCGKNPLGDDE	14	319	319	322

can be raised in this assignment by observing the *pattern* of matches made to each of the motifs. Visualisation of the data is facilitated by plotting the score for each matching motif across the length of the sequence, and disregarding the restrictions imposed by the PathFinder method (positions and inter-motif distances). This produces a raw plot of the magnitude and position of each match on the x- and y-axes of a graph respectively (figure 5.13). The results of scanning the sequence with each of the motifs are separated into individual graphs to avoid confusion, and are plotted sequentially on top of each other (figure 5.14). Observing the data like this allows the user to clearly appreciate the distinction between the many false matches, occurring over the entire length of the sequence, and true matches. To further highlight high-scoring matches, all those over a given threshold are displayed as a rectangle rather than a single line, the other dimension being the length of the motif (figures 5.13 and 5.14).

Figure 5.13: The GRAPHScan output describing the matches to a single motif (motif 1 of the GPCRRHODOPSN fingerprint), plotted along the length of a sequence (SWISS-PROT:OPSD_SHEEP).

The observations, made in section 5.2.3, about the effect of varying the motif scoring scheme, can be illustrated through the use of this graphical method, which demonstrates the efficacy of both the scoring methods and the graphical depiction of a scanned sequence. Figure 5.15 shows the result of scanning a sequence (SWISS-PROT: TRB1_YEAST) against its family's fingerprint (PRINTS:PNDRDTASEII) using each of the scoring schemes described previously (i.e., A) W-PID, B) AW-PID and C) the raw profile score). The adjustment made to the W-PID significantly reduces the abundance of false matches, and diminishes the magnitude of those matches that are still apparent. The comparison also shows that by using the profile score (which deviates from positive to negative) the majority of false scores fall below the zero threshold, while none

Figure 5.14: The GRAPHScan output describing the matches to all motifs in the GPCRRHODOPSN fingerprint, plotted along the length of a sequence (SWISS-PROT:OPSD_SHEEP).



of the truly significant scores are affected.

The potential for shorter motifs to produce more spurious matches, was discussed earlier; here it can be demonstrated by observing the difference in the number of false matches made by the short motif, 4, and the long motif, 6. Another advantage of the graphical methods is that it provides a qualitative overview of the quality of individual motifs. GRAPHScan thus allows the user to pinpoint 'good' motifs and 'bad' ones.

5.3.3 MULScan

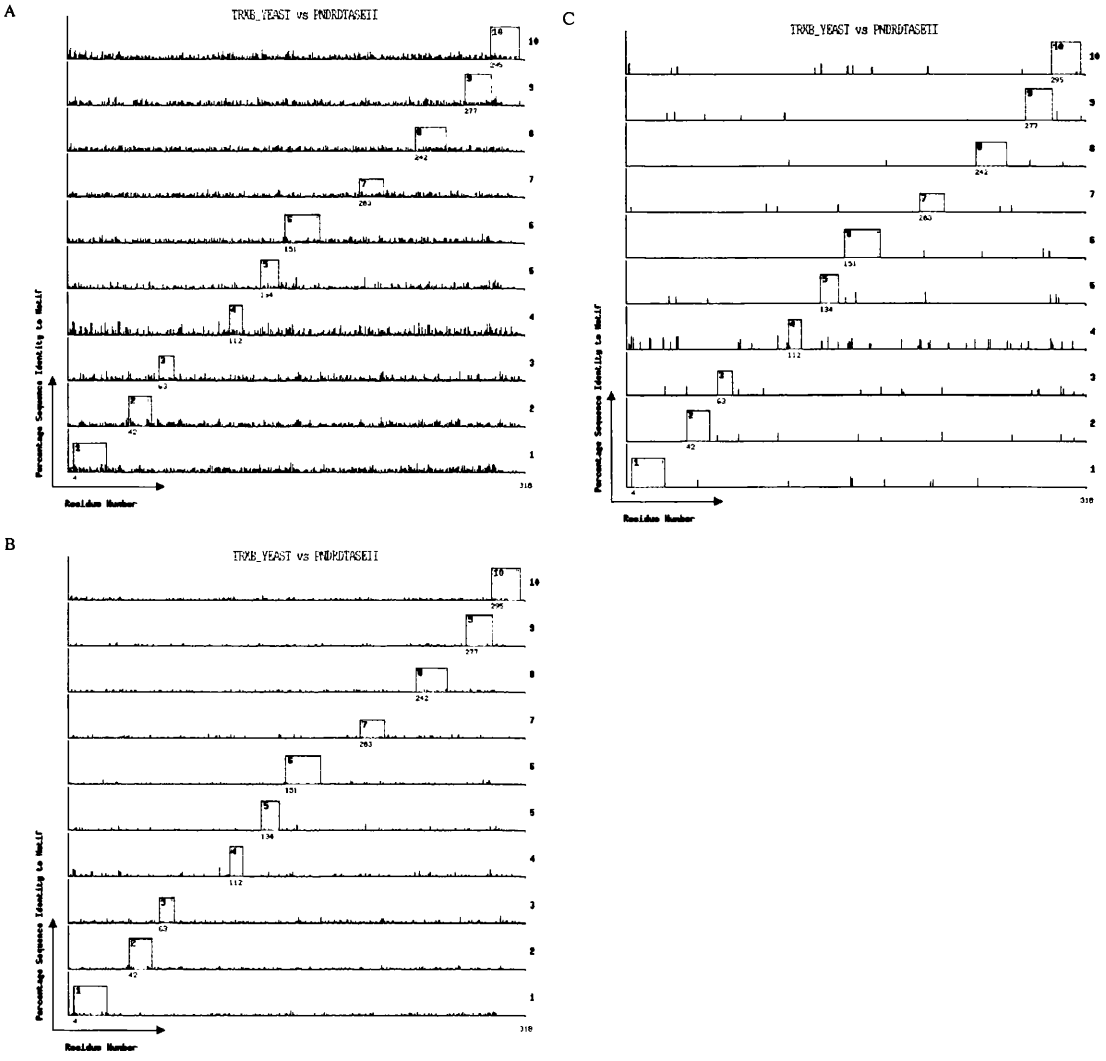
Today sequence data are generated on the scale of the genome; therefore, it is desirable to be able to perform sequence analysis on thousands of sequences at a time. A purely functional addition to the suite of programs is MULScan, which provides the facility to submit multiple sequences for analysis.

The implementation of the FingerPRINTScan software is such that scanning a single sequence is fast enough to provide an interactive WWW based session. To consider the potential of scanning more than one sequence necessitates the development of an alternative solution, which also conserves resources for single sequence scanning users. MULScan provides a WWW interface that enables multiple submissions to be made; these are queued on a server until a time of low load, whereupon they are processed linearly and results are forwarded to the user via electronic mail. The interface does not represent an extension of the software, as the handling of more than one sequence has been a design requirement from the outset; however, what it does represent is an extension of the functionality of the suite of tools.

5.3.3.1 FPScan Multiple Sequence Analysis (FPScanMSAn)

Aside from the WWW, many analyses are performed on personal computers, with local databases and software installations. Working with multiple sequences can ultimately produce large quantities of data. The analyses of these data can be hindered

Figure 5.15: A comparison between the match scoring schemes, demonstrated using the GRAPHScan tool.
All graphs show the matches made (all those scoring above the threshold of zero), to motifs of the PNDRDTASEII fingerprint, by a query sequence that is a member of the same family. Graph A was plotted using the W-PID scoring scheme; graph B used the AW-PID, and graph C used the profile score to determine whether matches score above the threshold and matches are plotted using the percentage identity score, to enable comparisons to be made.



by the practicalities of the task; e.g., searching through 6,000 sets of results by eye is a daunting prospect; nevertheless, scanning the *Saccharomyces cerevisiae* proteome would generate almost exactly this number of results.

The command-line software tool 'FPScanMSAn' was developed to simplify the management of large scale analyses. As mentioned above, the scanning software (fingerPRINTScan) is fast enough to satisfy a user of the WWW interface; i.e., it identifies matches to an average sequence in less than 30 seconds. However, when faced with 13,000 *Drosophila melanogaster* ORFs, time becomes an important factor to consider (even at one sequence every 10 seconds, the task of scanning the fly proteome would take over 36 hours). Also, the results of 13,000 analyses are cumbersome and difficult to process. If multiple proteomes are to be scanned, then this situation becomes increasingly complicated.

FPScanMSAn was designed to automate a number of the steps required to perform such large scale analyses, to provide reporting capabilities for simple queries of the results, and to facilitate the deposition of results into a relational database that supports complex querying.

One of the problems of dealing with large datasets is the practical issue of coping with scanning each and every sequence. While fingerPRINTScan was designed to admit multiple sequences it actually scans then sequentially against a database of fingerprints; however, most computer systems do not have sufficient memory (RAM) to support the submission of very data-sets. FPScanMSAn provides the facility to divide a large file of sequences into smaller, more-manageable units and sequentially submits each of these to fingerPRINTScan. All input files and their results are allocated an individual directory, which means that each is segregated, thus making subsequent analysis of the results more tractable. Dividing the large query set of sequences into smaller subsets also facilitates the use of multiple instances of the software on multiple machines or processors: this affords a primitive, but effective, multiprocessing capability. Consequently, dividing 13,000 sequences over 10 processors, has the potential

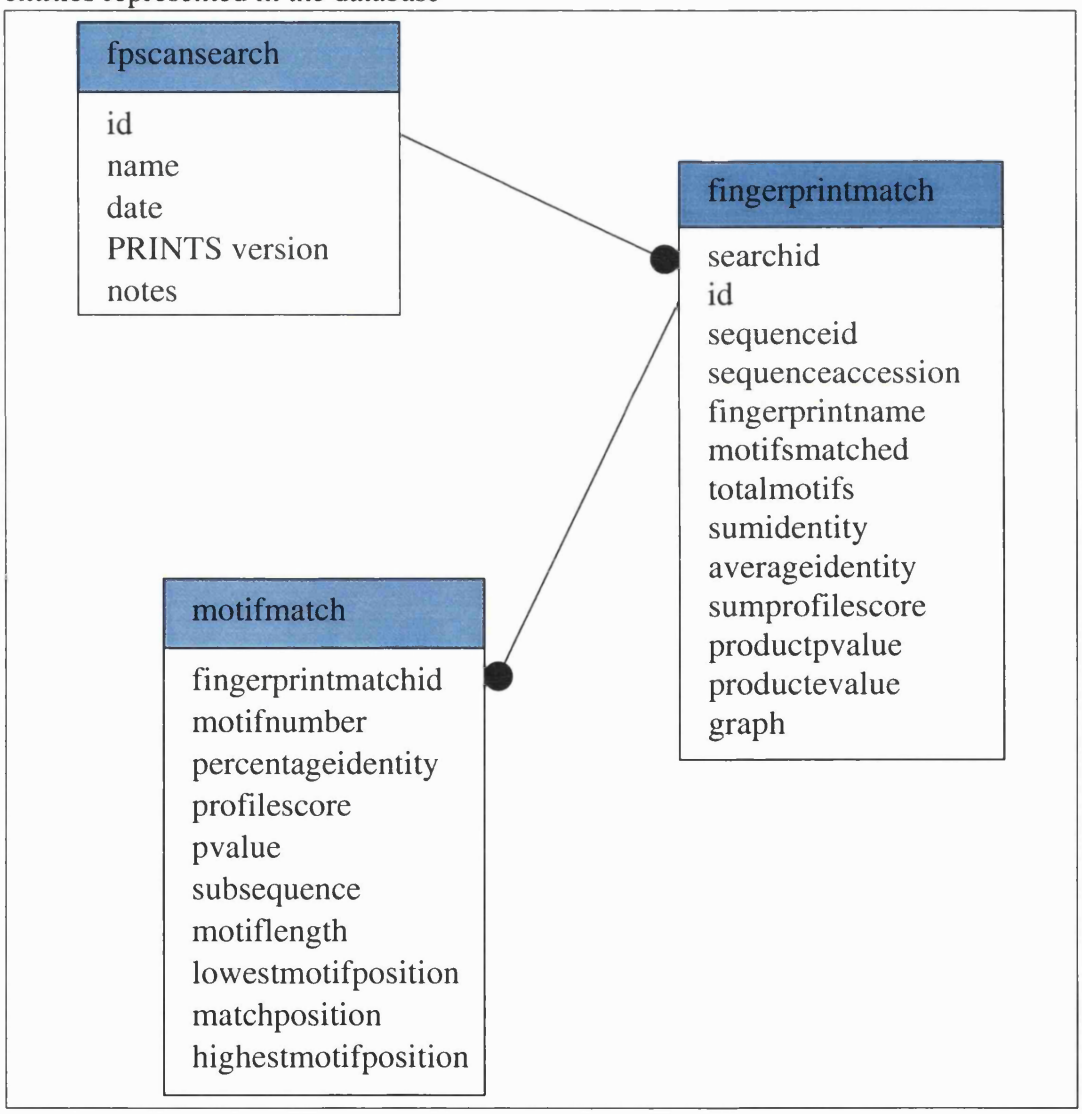
to reduce the processing time from 36 to under 4 hours.

With all the results in place, simple analyses can be performed, such as selecting all sequences that match fingerprints with scores above a given threshold, or identifying the top scoring match to each sequence. Such analysis may be sufficient to satisfy the needs of a researcher hunting proteomes in order to establish ORF identities. Often the requirements for analysis can be more complex: e.g., to find all sequences matching a particular fingerprint, or to cross compare sequences matching particular fingerprints from one proteome with another. To support this, the results of any fingerPRINTScan search can be reformatted for entry into a relational database, which allows complex queries to be phrased in a Structured Query Language (SQL). The advantages of the use of a relational database stem from the ability to define fixed relationships between data, which vastly improve its manageability (figure 5.16 describes the tables used to store the results of each search).

Data produced by sequentially scanning a set of sequences will obviously be sorted by the order of sequences in the original datafile, and this organisation often disregards their biological relationships. By simply sorting the data by the matching fingerprint, previously hidden trends and patterns may be identified. For example, using the results from prokaryotic and eukaryotic proteomes to compare the relative abundances of particular family membership can highlight: conserved, housekeeping protein families, believed to be essential components of all living cells, and protein families that provide specialisations necessitated by the organisms' differences.

As discussed previously, paralogous relationships between sequences in a family can complicate the picture of functional relationships. By looking closely at variations in the scoring patterns of individual motifs from sequences matching the same fingerprint, it may be possible to detect trends that indicate the existence of distinct subtypes (reflecting potential paralogues, the precise functions of which may differ). To construct a query that involves looking at matches to fingerprints and their constituent motif-matches, using a non-database approach is cumbersome and would require writing

Figure 5.16: FingerPRINTScan results are placed into a relational database to enable complex queries to be formed over the data. Below is the database schema, which describes the tables and attributes, as well as indicating the relationships between the entities represented in the database



software to select relevant data from the set of results. However, by defining the relationship between fingerprints and motifs explicitly in the relational database, queries such as these can be posed simply as SQL statements.

While these hypothetical queries are speculative ones based on biological knowledge, the use of databases that support complex querying is a ever growing concern of the biological community; therefore, the provision of a relational database form of fingerPRINTScan results can only be expedient for further analysis.

5.4 Summary

The search tool presented in this chapter has been developed to utilise as much of the biological knowledge inherent in a fingerprint as possible. It has taken the task of identifying potential matches to a group of patterns from a purely numerical approach, to one that is supplemented with the inherent biological context contained within the MSA. The method, exploits the sensitivity of fingerprint models, by providing access to partially-matching results, and facilitates the manual diagnosis of difficult cases, by supplying probability values and differing perspectives (graphical and motif-based). The software is freely available and has been supplied to many research groups. As a consequence, aside from the WWW services that we provide, fingerPRINTScan is in use in many different research activities all over the world. The following chapter will demonstrate a number of the applications of the software.

Chapter 6

Results and Applications

6.1 Introduction

One of the most important issues in the evaluation of the fingerPRINTScan method was to compare the use of the different scoring schemes. This section will discuss how the development of the method was driven, in part, by the requirement to identify a suitably discriminatory scoring procedure.

When using any sequence analysis search tool, it is essential to determine a level of confidence in the familial assignments, or predictions, made. As discussed earlier, some diagnoses are 'black and white': a sequence is identified as either a match or not. A problem with such a method is that it leaves little room for variation, and, as a result, it can be insensitive to the identification of distant family members. However, as provision is made for deviation by using more sensitive approaches, the issue of the distinction between true and false diagnoses is raised. A perfect scenario would be to envisage a score, which both provided a measure of the significance of a result, and definitively drew a line between evolutionary relationships and false positives (Hubbard, 1997). Realistically, scoring schemes do not live up to these criteria, and as a consequence it is profitable to pursue more realistic goals. This can be achieved by balancing the need to avoid false positives with the desire to search sensitively. For example, panning for distant members of a family requires careful study of weak relationships; in such a search, the inclusion of a number of false assignments is to be expected and is therefore tolerable. However, any form of automated analysis must rely on confident diagnoses; therefore, false positives must be minimised. The following section will discuss the application of scoring thresholds to the fingerPRINTScan method that aim to minimise the occurrence of false positives.

6.2 A comparison of scoring schemes

Based on the membership of families in the PRINTS database (version 27.0) a sequence database was constructed (true27). Each fingerprint in PRINTS contains a list of sequence identifiers that are designated as true members of each family. Careful checking has been performed to ensure that these assignments represent true diagnoses, as the integrity of the database relies on the family membership of these sequences. Therefore, these were chosen to represent a set of sequences for which fingerPRINTScan could produce the most confident and verifiable assignments. As a comparison, the same set of sequences were randomly shuffled so as to maintain the amino-acid composition of the sequences, while destroying any positional sequence information (rand27). In this way, scoring artifacts based on compositional bias should be avoided (e.g., motifs that produce high scoring matches based solely on the overabundance of particular residues), and the distinction between matches of random sequence and evolutionarily related sequence can be measured. Randomising the sequences also avoids complications in distinguishing between *real* false assignments (completely unrelated sequences) and matches to related, but distant, family members (which indicate a true evolutionary relationship, but are designated *false* because the sequence does not belong to the fingerprint).

Both sets of sequences (true27 and rand27) were processed by FPScanMSAn, using a database of motif profiles based on fingerprints from version 27.0 of PRINTS. Profiles were generated, as in section 3.3.4.3, using the BLOSUM 62 substitution matrix. The results of both runs were placed into a relational database as described in section 5.3.3.

For each sequence in true27 and rand27, the top scoring match was extracted from the database, and the following data were collated: the matching fingerprint identifier; the sequence (SWISS-PROT or TrEMBL) identifier and accession code; the summed AW-PID score; the average AW-PID score; the summed profile score; the product p-value; and the product e-value. The same ten sequence are shown from each of the

sets of results in figure 6.1[1] (sequences from rand27 are shuffled, so they now merely

Figure 6.1: The first ten sequences from each of the two result datasets.

Sequences from rand27

fingerprint name	sequence identifier	sequence accession	sum AW-PID	average AW-PID	sum profile	product p-value	product e-value
P2X1RECEPTOR	GLU2_ORYSA	P07730	74.07	37.04	404	0.007	1800
BETATUBULIN	GU11_ORYSA	P07728	58.59	29.3	354	0.004	1000
RIBOSOMALP2	GU12_ORYSA	P07729	90	45	446	0.00041	110
MELNOCYTESHR	Q40685	Q40685	66.67	33.33	348	0.0013	330
GPROTEINA12	Q38780	Q38780	76.92	38.46	417	0.0006	150
VP6CAPSID	Q40347	Q40347	68	34	501	3.3e-05	8.5
FMOXYGENASE5	GLUB_ORYSA	Q02898	67.95	33.98	393	0.0022	560
ADENOSINEA1R	Q40689	Q40689	72.47	36.23	406	0.0014	370
FMOXYGENASE	GLUC_ORYSA	Q02897	64	32	500	1.6e-05	4.2
KIR12CHANNEL	GLU4_ORYSA	P14323	84.13	42.06	400	0.0018	470

Sequences from true27

fingerprint name	sequence identifier	sequence accession	sum AW-PID	average AW-PID	sum profile	product p-value	product e-value
11SGLOBULIN	GLU2_ORYSA	P07730	390	56	3923	5.7e-73	1.5e-67
11SGLOBULIN	GU11_ORYSA	P07728	390	56	3898	2.8e-72	7.1e-67
11SGLOBULIN	GU12_ORYSA	P07729	390	56	3898	2.6e-72	6.8e-67
11SGLOBULIN	Q40685	Q40685	390	56	3873	1e-71	2.7e-66
11SGLOBULIN	Q38780	Q38780	390	55	3818	1.1e-70	2.9e-65
11SGLOBULIN	Q40347	Q40347	380	55	3750	1.5e-69	3.7e-64
11SGLOBULIN	GLUB_ORYSA	Q02898	380	55	3724	9.5e-69	2.4e-63
11SGLOBULIN	Q40689	Q40689	380	55	3780	1.2e-69	3.1e-64
11SGLOBULIN	GLUC_ORYSA	Q02897	380	55	3711	2.3e-68	5.8e-63
11SGLOBULIN	GLU4_ORYSA	P14323	380	54	3688	4.8e-68	1.2e-62

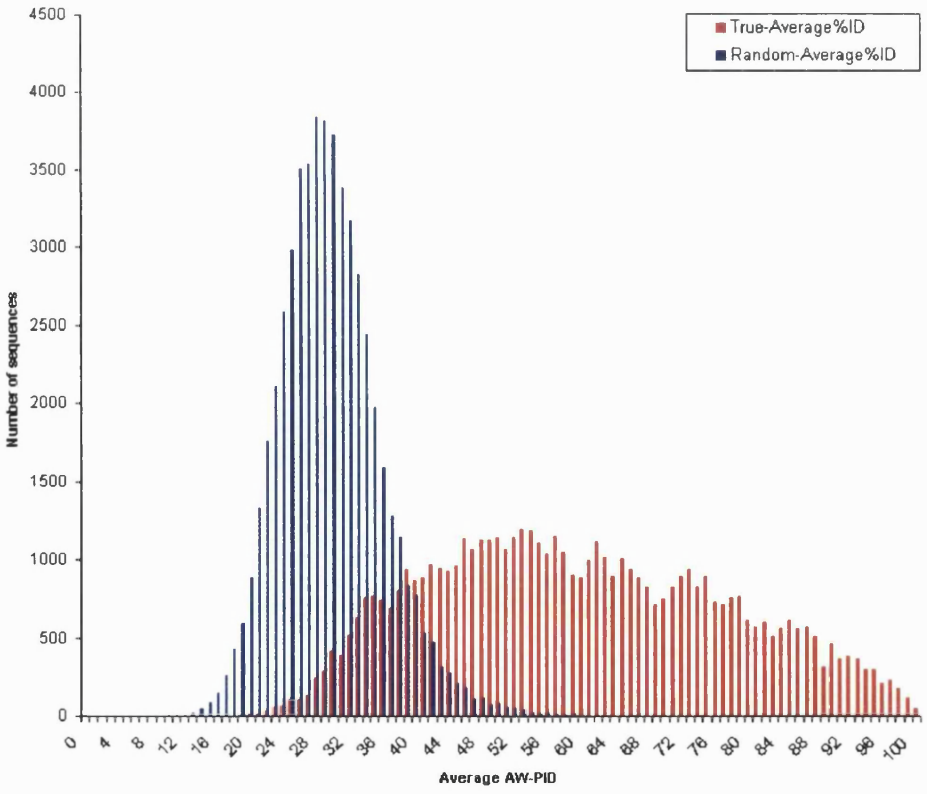
share composition and length with their counterparts in true27). The examples are members of a single fingerprint (as illustrated by the true27 results). Note that the average AW-PID score barely distinguishes the real scores (from the true27 set, which range between 54 and 56%), from the random scores (the maximum of which is 45%).

Performing the same comparison over all of the ~54,000 sequences in true27 and rand27, further highlights the failure of the AW-PID score to effectively discriminate true from false. To demonstrate the efficacy of scoring sequences for each of the four scoring schemes (AW-PID, summed AW-PID, profile and p-value) histograms of the frequency of sequences achieving particular scores were plotted (one for each of the scoring schemes). The comparison between true (true27) and false (rand27) distributions allows for putative thresholds to be set in order to exclude varying percentages of

[1]Throughout this chapter the numerical form $1e^{-10}$ is used in place of 1×10^{-10}.

false positives; i.e., this allows a scoring threshold to be established to produce a selective result (by removing the majority of false positives assignments). A qualitative assessment of the resultant histograms clearly demonstrates the weakness of scoring with AW-PID (figure 6.2). The bulk of the overlap between false and true results occurs between 30 and 40%; however, the range over which these are indistinguishable is extensive (true matches start to appear at around 20%, while false ones do not diminish significantly until 60%).

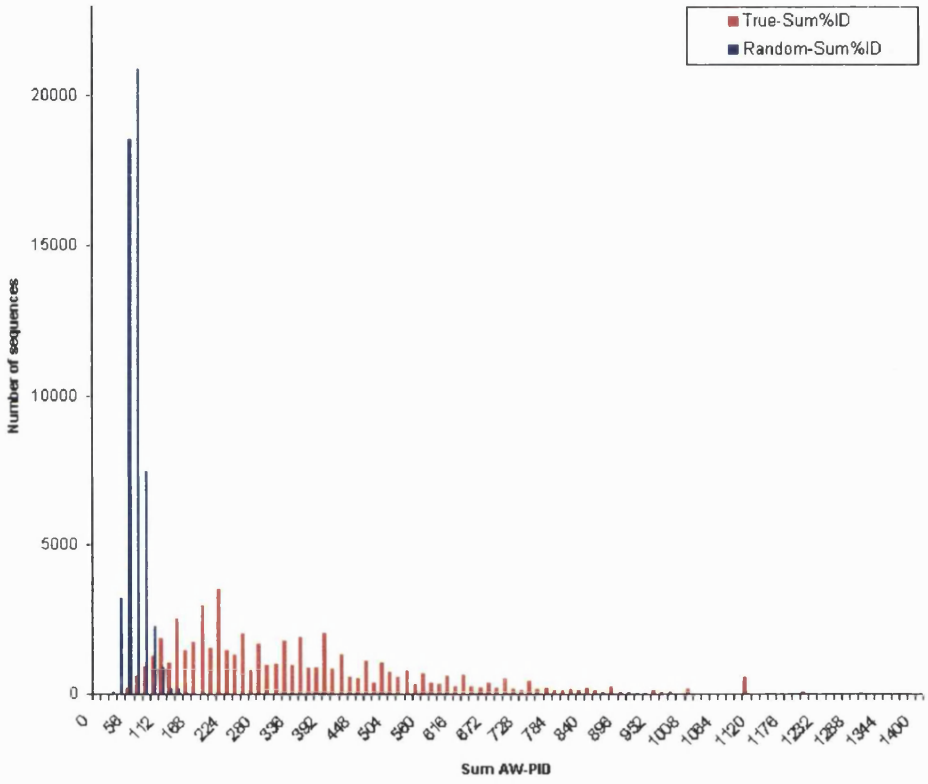
Figure 6.2: A plot of average AW-PID scores for sequences matching fingerprints from PRINTS 27.0. 'True' sequences are members of the true27 sequence database, while 'False' members are derived from the rand27 database.



Qualitatively each of the other scoring schemes provide a better distinction between true and false results (figures 6.3, 6.5 and 6.7); in each, the distribution of the random scores is sharp and tight, which is in stark contrast to the distribution of true scores. The p-value score seems to produce the flattest, more highly variable distribution of

true scores (the overlap of the high-scoring random sequences and low-scoring true sequences is emphasised in figures 6.4, 6.6 and 6.8).

Figure 6.3: A plot of summed AW-PID scores for sequences matching fingerprints from PRINTS 27.0. 'True' sequences are members of the true27 sequence database, while 'False' members are derived from the rand27 database.



Quantitatively, the differences between the scoring schemes are clear. Drawing a vertical line on any of these plots provides a separation between two populations of sequence matches. If we use this line (threshold) to designate family membership, then sequences scoring above the threshold are diagnosed as true, and those below are false. As we can be confident about the family membership of the sequences in true27, the definition of a threshold can be used to compare the performance of the scoring schemes. Sequences from rand27 that score above the threshold represent false positives, and those from the true set that score below are the false negatives. Figure 6.9 shows that by modulating the number of acceptable false positives (by moving the

Figure 6.4: A plot of summed AW-PID scores for sequences matching fingerprints from PRINTS 27.0, highlighting the region of cross-over between true and false scores.

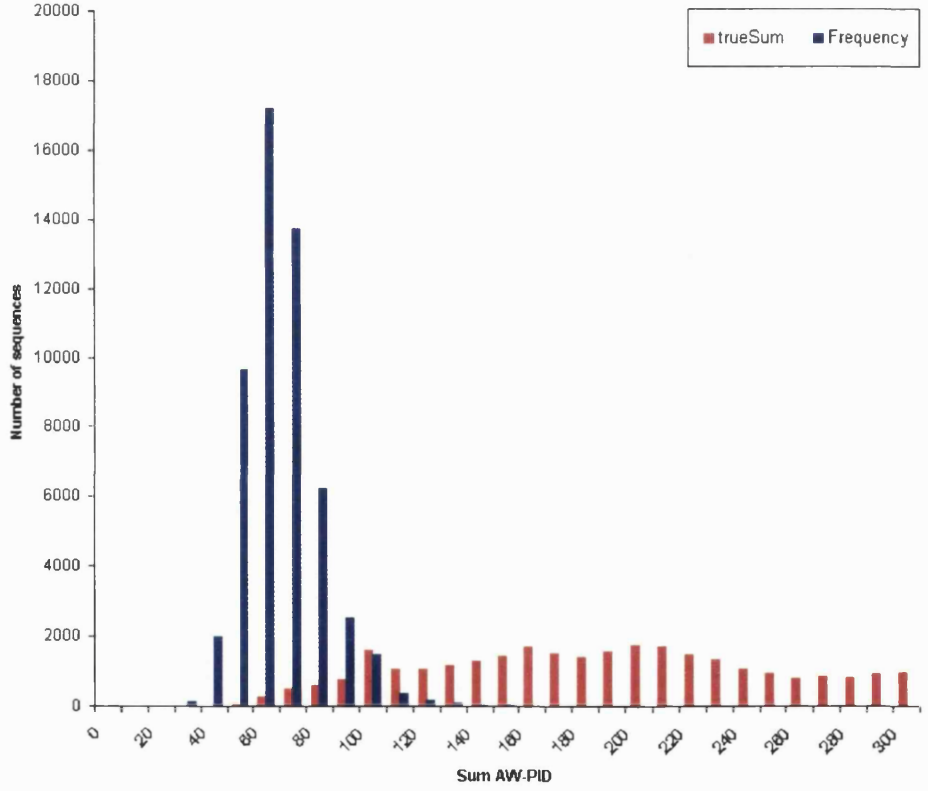


Figure 6.5: A plot of profile scores for sequences matching fingerprints from PRINTS 27.0. 'True' sequences are members of the true27 sequence database, while 'False' members are derived from the rand27 database.

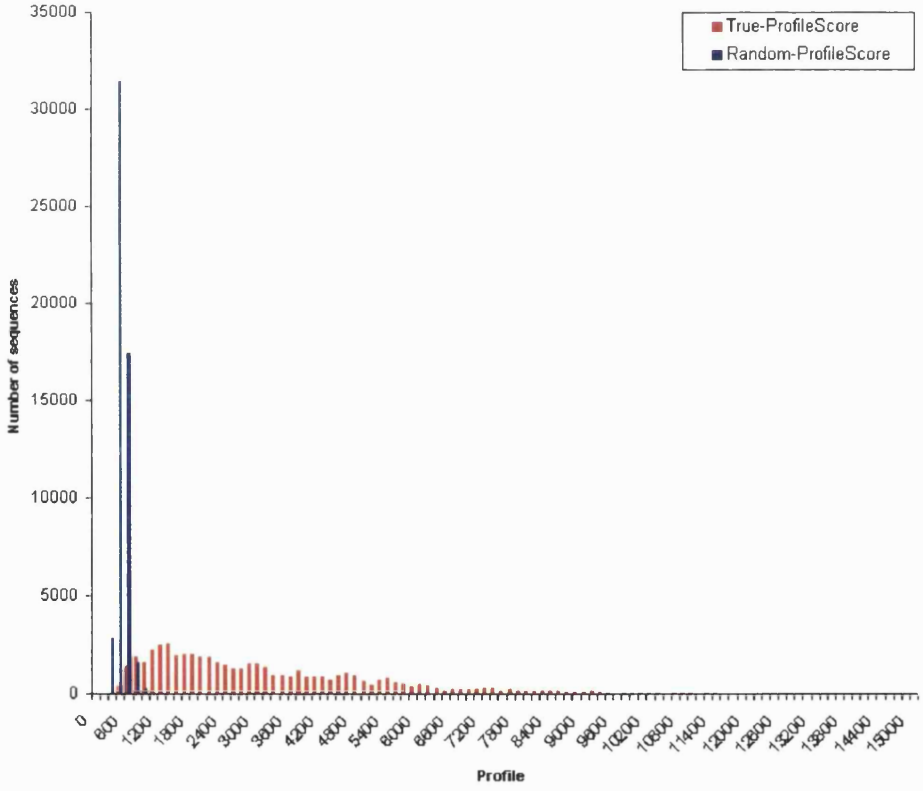


Figure 6.6: A plot of profile scores for sequences matching fingerprints from PRINTS 27.0, highlighting the region of cross-over between true and false scores.

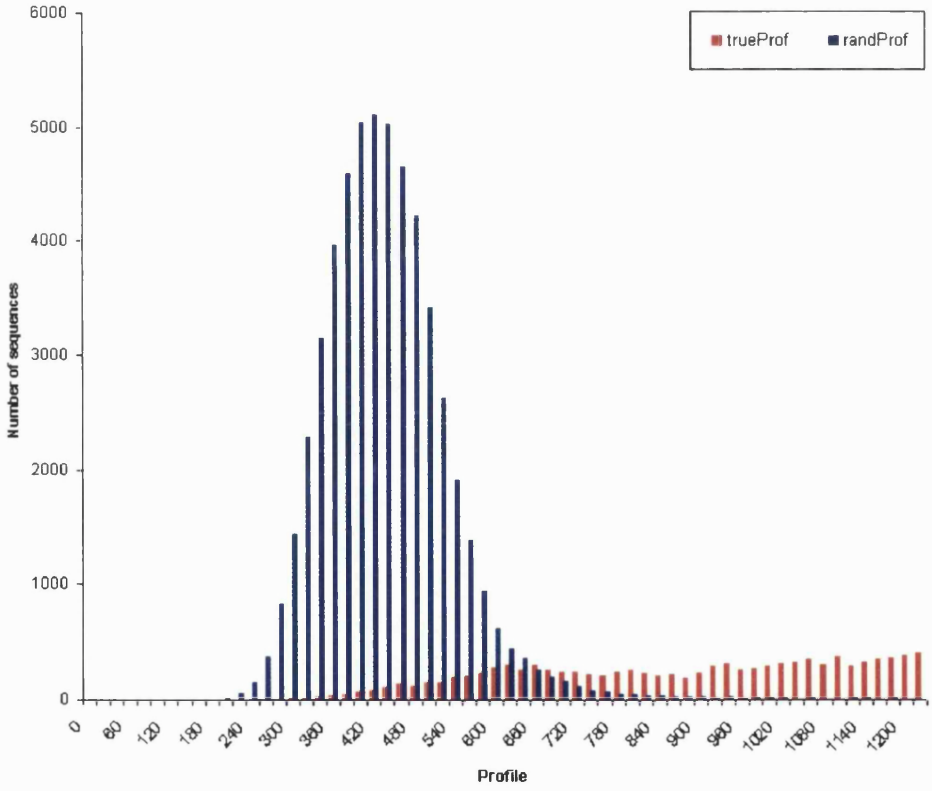


Figure 6.7: A plot of p-values for sequences matching fingerprints from PRINTS 27.0. 'True' sequences are members of the true27 sequence database, while 'False' members are derived from the rand27 database.

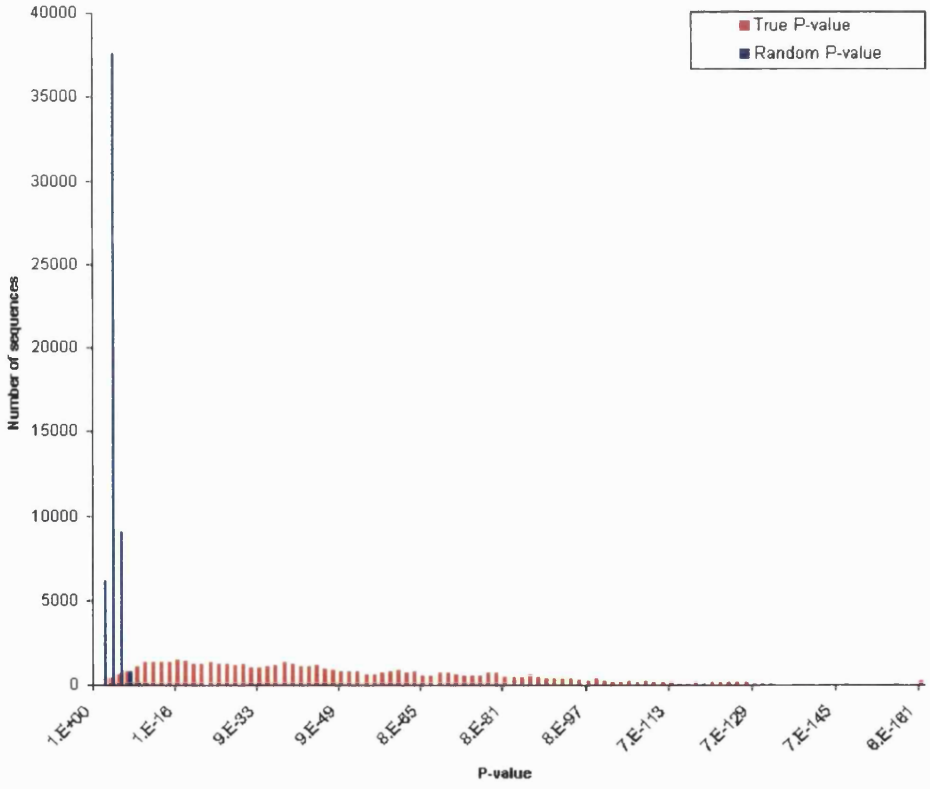
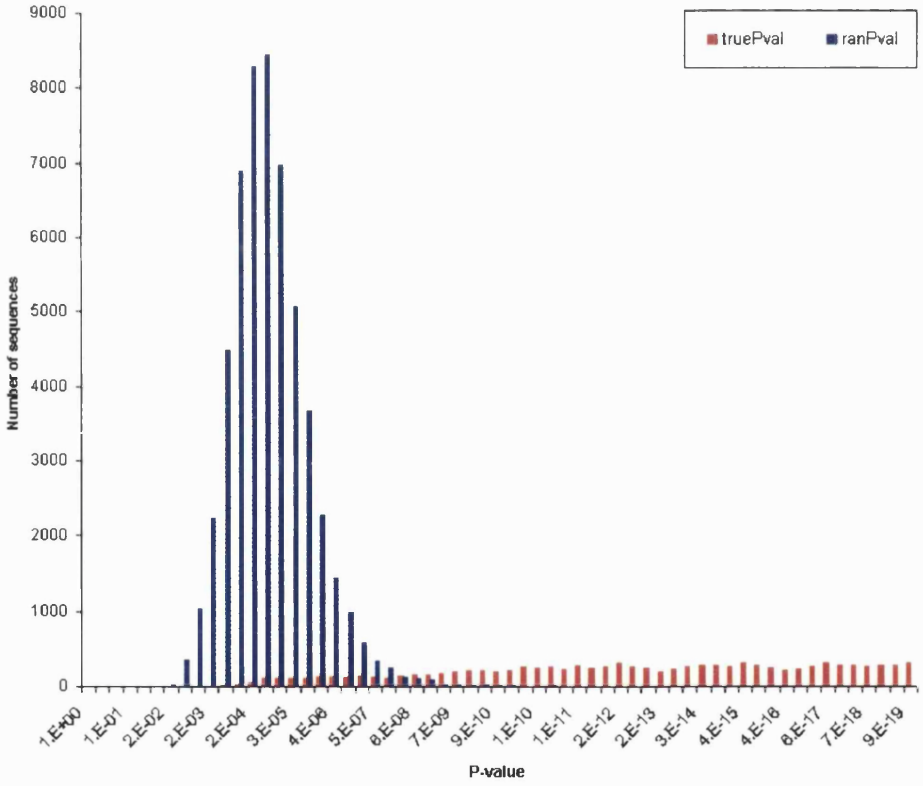


Figure 6.8: A plot of p-values for sequences matching fingerprints from PRINTS 27.0, highlighting the region of cross-over between true and false scores.



threshold) the resultant number of false negative assignments can be calculated.

Figure 6.9: For each of the scoring schemes a threshold value was set. Each scheme has a different scale (e.g., the average AW-PID ranges from 0-100%, while the p-value scale ranges from 1- $\sim 1e^{-200}$), in order to express thresholds defined on these scales, each is expressed as the percentage of false positive assignments it creates (10%, 5%, 1% and 0.01%). Tabulated are the corresponding percentages of false negatives produced by each threshold.

Scoring scheme	False positive thresholds.			
	10	5	1	0.01
Average AW-PID	11.2	16.0	30.0	65.3
Summed AW-PID	2.7	4.0	9.0	33.0
Profile	2.0	2.8	6.3	27.0
p-value	1.0	1.4	2.3	14.3

In order to make any diagnosis, a threshold *must* be defined; however, in drawing this line a compromise must be made between the number of false diagnoses made (false positives versus false negatives). The acceptable levels of either false positives or negatives is strongly dependent on the task at hand. For example, investigation of distant family members requires a minimal loss of true data, and hence the threshold that best supports this role would be defined so as to minimise the number of false negatives. From figure 6.9, the minimum number of false negatives was produced by the p-value scoring scheme (10% false positive assignments results in 1% false negatives), by contrast, if the AW-PID threshold is set at the same level (10% false positives) an order of magnitude more true matches are missed. Here, the compromise required to achieve a more sensitive result is that the number of false positive assignments is high (10%). However, in such a quest for distant family membership, it is to be expected that sequence assignments will be investigated manually; therefore, other evidence can be used to provide a distinction between *real* distant members and high-scoring false positives. Conversely, automating a sequence analysis method means that less false positive assignments are acceptable. Placing a threshold at such a point that it avoids 99.99% of false positives, results in the loss of $\sim$14% of the true matches. Again, a

compromise is made. This time, the requirement to reduce the number of false assignments means that a relatively high number of false negatives must be accepted.

At each point along this scale from the total avoidance of false positives (below the 0.01% threshold) and the avoidance of false negatives (above the 10% threshold), the p-value scheme proves to be the most discriminatory, by consistently providing the least false negatives at every level. The largest deviance in the number of false negatives is seen between the 1% and 0.01% false positive thresholds, which corresponds to the cross-over between the the two distributions (the wane of the random and the rise of the true distribution). Therefore, it is in crossing this range that signifies the move from a sensitive to selective result. As a consequence, the default threshold used for making diagnoses by fingerPRINTScan is set approximately midway between these two values and represents a false positive assignment rate of about 2 in 10 (0.18%), which is an acceptable level for the analysis of single sequences using the WWW facility. The actual p-values thresholds that correspond to each of the false positive percentages are indicated in the following table (figure 6.10).

Figure 6.10: The p-value scoring scheme provides the best performance of all the scoring schemes for any given threshold. Each p-value (or e-value) threshold corresponds to a percentage of false positives and false negatives. A new threshold is introduced into this table that has an e-value threshold of $1e^{-4}$, which is the default value used by FPScan to indicate significant results. The value sits approximately midway between the 1% and 0.01% values, and effectively represents a compromise between selectivity and sensitivity.

false positive percentage	p-value threshold	e-value threshold	false negative percentage
5	$3.8e^{-6}$	0.98	1.4
1	$2.4e^{-7}$	0.06	2.3
0.18	$4.7e^{-10}$	$1.0e^{-4}$	5.4
0.01	$8.9e^{-16}$	$2.3e^{-10}$	14.3

6.2.1 Summary

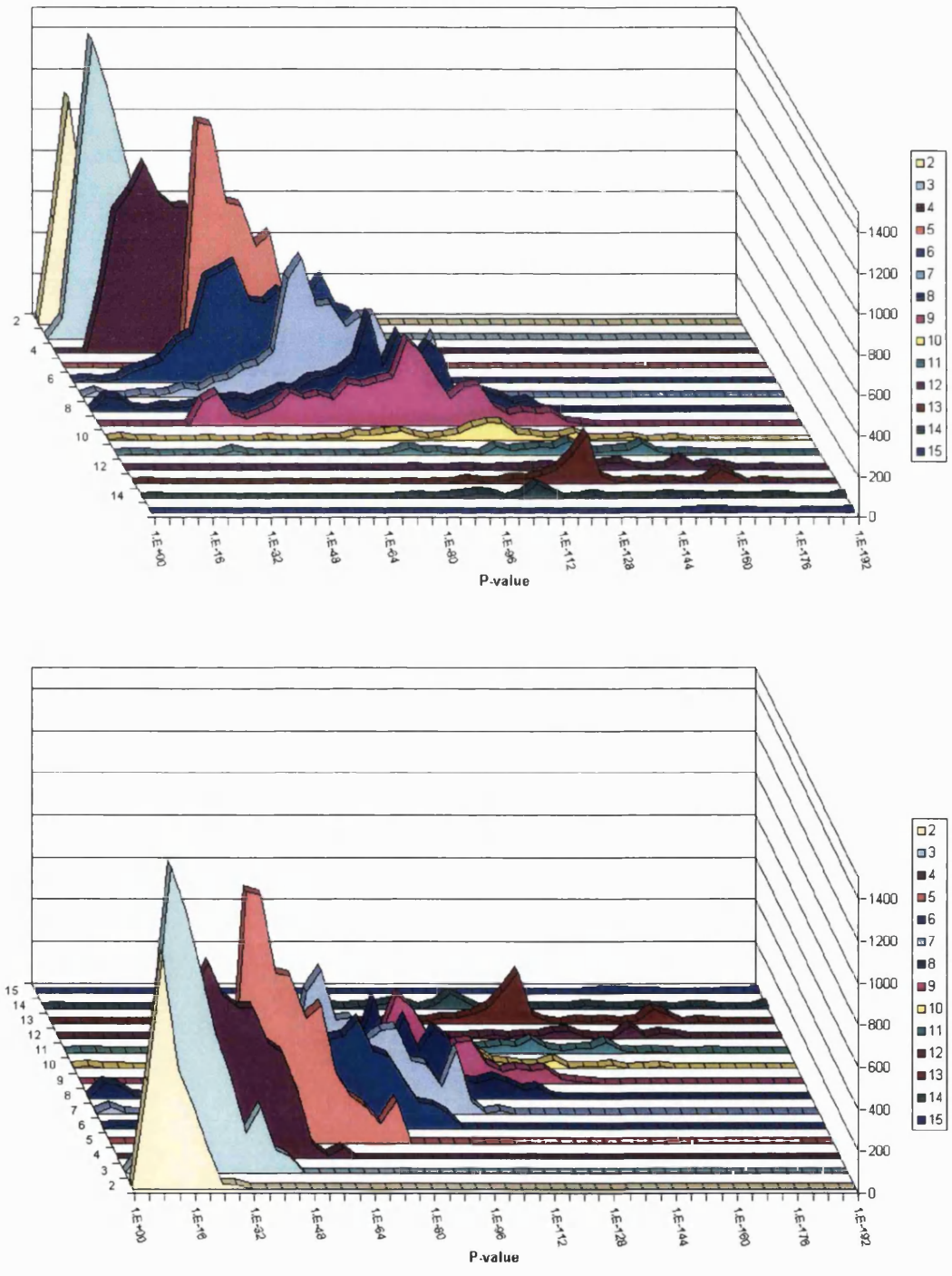
Unfortunately, all of the scoring schemes exhibit the same pattern: a gain in selectivity is offset by a loss of sensitivity and *vice versa*. However, one scoring scheme consistently performs better than the others: the p-value scheme. The merit of using such a scoring scheme is therefore clear. The p-value provides an qualitative measure of the likelihood of a match score, with which it is possible to compare matches to fingerprints in order to determine their significance. Also, the score provides the clearest distinction between random scores and true scores. Consequently, the p-value score is best suited to establishing a threshold, by means of which sequences can be assigned family membership with confidence.

6.3 Multiple motifs

Closer examination of the low-scoring results reveals some interesting features of fingerprints. As discussed previously, shorter patterns fail to discriminate true from false more frequently than longer patterns. This phenomenon can be observed empirically by performing an analysis of the scores achieved by fingerprints with varying numbers of motifs. Plotting the score distributions of fingerprints with 2 to 15 motifs reveals a clear pattern: increasing the numbers of motifs results in higher scores (lower p-values) (figure 6.11). Interestingly, the graphs indicate that a significant proportion of two- and three-motif fingerprints fall below a selective threshold (0.01% false positives, p-value $8e^{-16}$).

The true27 dataset from section 6.2, was used in this experiment to investigate the effect that multiple motifs have on the distribution of scores. It was observed in the previous section that there is a significant amount of overlap between the scores achieved by matches made by true and random sequences, even with the best performing scoring scheme. Any improvement that can be made in the distinction between true and

Figure 6.11: Both graphs show the distribution of p-values against the number of motifs in a fingerprint. Frequency is plotted vertically, while p-value is plotted along the x-axis. Each set of fingerprint matches is represented separately based on the the number of motifs each fingerprint contains. The different views of the data clearly show the variation in the distribution of scores over the number of motifs in a fingerprint.



false results in this region, will significantly improve the confidence that can be placed in any assignments made by fingerPRINTScan.

Initial analysis of these low-scoring matches indicated the presence of both falsely identified true member sequences, and low scoring *partial* matching sequences. Sequences make partial matches with fingerprints for a variety of reasons; e.g., sequences can be incomplete or truncated (fragmentary sequences) or they can be distant members of the family (distant orthologues or paralogues). These accordingly match fingerprints with scores ranging from nearly maximal (e.g., 1 match missing in a 10 motif fingerprint) to minimal (e.g., only matching 2 motifs). While, identifying partial matches is important to provide as much sensitivity as possible, their presence complicates the study of the scoring patterns of n-motif fingerprints. Therefore, in order to define a selective threshold, and to investigate the scoring patterns of fingerprints, a new set of true member sequences was defined to replace true27. True27_full contains only sequences identified as *fully-matching* members of fingerprints: a sequence must match all n of a fingerprint's motifs. It is fair to provide this redefinition of the true dataset as the fingerprinting method considers only these sequences as suitable for extraction of motifs; therefore, they represent the most confidently assigned and verified members of a fingerprint's family.

Using the sequences from true27_full to perform the analysis described in section 6.2, show a reduction in the percentage of false negatives (see figure 6.12). The remaining low-scoring results fall into two categories: those sequences that belong to one fingerprint but are identified by another and those sequences identified by the correct fingerprint, but which score below the threshold. Both sets of sequences represent a failure to diagnose familial membership by the fingerPRINTScan method, and, therefore, require further investigation. The sequences were split into their respective sets by identification of each sequence's fingerprint (all sequences in this set are fully matching members of a fingerprint) and a comparison of this with the fingerprint that actually matched the sequence; those with correct assignments are designated 'false negatives',

Figure 6.12: A comparison between the number of false positive assignments made using the the true27 and true27_full datasets. As no partial matching sequences are represented, any sequences falling below thresholds are false negatives that arise as a consequence of a fingerprint failing to elevate a full matching member above the threshold. Three false-positive threshold values are shown, alongside the corresponding p-value, and the false-negative percentages (from figure 6.10), as well as the true27_full values.

false positive percentage	p-value threshold	false negative % true27	false negative % true27_full
1	$2.4e^{-7}$	2.3	1
0.18	$4.7e^{-10}$	5.4	3
0.01	$8.9e^{-16}$	14.3	11

while those with incorrect matches are termed 'true negatives'[2]. Both sets highlight problems. The fingerprints of the former set correctly identify their members, but fail to provide significant scores, while the latter set are matched by unrelated fingerprints that score higher than their own fingerprint. Observations of the scoring potential of fingerprints with varying numbers of motifs indicated a propensity for smaller patterns (fingerprints with fewer motifs) to produce scores that barely distinguish true matches from random matches (figure 6.11). This observation is clearly reflected in an analysis of the fingerprints that contribute these low scoring matches. The most obvious trend is the preponderance of fingerprints with 2 and 3 motifs falling into this group (figures 6.13 and 6.14).

Closer examination of the low scoring fingerprints reveal particular examples that repeatedly produce scores in this range and, more importantly, a common trend shared between these fingerprints. Clearly, it is short fingerprints, those with 2 to 4 motifs that most consistently fail to promote their own member sequences either above a threshold or above matches to other fingerprints. However, many of these fingerprints share another common feature. Most of the fingerprints in either table of figure 6.15, describe

[2]The 'true negative' matches are also false negatives, because they represent sequences unidentfied by their own fingerprints below the threshold.

Figure 6.13: The number of sequences falling below two thresholds analysed by the number of motifs in their families' fingerprints. All sequences falling below the two thresholds 1% and 0.01% false positives were analysed. For each sequence, the number of motifs in its family's fingerprint was extracted. This is presented in the following table as the number of fingerprints containing n motifs.

Number of sequences scoring below the thresholds											
false negatives											
false positive	p-value threshold	total	number of motifs								
			2	3	4	5	6	7	8	9	10
1	$2.4e^{-7}$	218	204	10	3	1					
0.01	$8.9e^{-16}$	4102	1908	1789	366	30	6	3			
true negatives											
false positive	p-value threshold	total	number of motifs								
			2	3	4	5	6	7	8	9	10
1	$2.4e^{-7}$	289	258	16	15						
0.01	$8.9e^{-16}$	900	619	161	88	10	13	1	2	1	5

Figure 6.14: All sequences from fingerprints with 2-3, and 2-4 motifs are expressed as percentages of the total number of sequences falling below the two thresholds.

false positive percentage	p-value threshold	false negatives		true negatives	
		2 & 3 motif	2-4 motif	2 & 3 motif	2-4 motif
		fingerprints			
1	$2.4e^{-7}$	94.8	100	93.6	98.17
0.01	$8.9e^{-16}$	90.13	99.05	87.22	97.00
		> 90	> 99	> 87	> 97

very divergent relationships. A number of promiscuous domains are represented, most notably the ZINCFINGER (the C2H2-type zinc finger fingerprint), GPROTEINBRPT (the G-protein beta WD-40 repeat) and HOMEOBOX (the homeobox signature) fingerprints. As noted previously, shared domains or highly-divergent families, when described by motif-based models, have a tendency to produce weak patterns, which is due in part to the restrictions on motifs (i.e., that they must be conserved *ungapped* blocks of aligned sub-sequences). When sequences in an MSA become so divergent that extensive gaps are required to maintain alignment, regions from which motifs are derived are frequently shortened and become less abundant. The reduction in the quantity of conserved regions, is usually accompanied with a reduction in the quality of conservation (i.e., more positions in the motif accumulate multiple point mutations); both affect the discrimination power of the fingerprint. Less conservation in columns of the alignment improve the potential for random sequences to attain positively-scoring matches, and fewer, and increasingly polluted, motifs reduce the significance of true member scores, and the context that larger numbers of motifs afford.

6.3.1 Summary

Fingerprints defined from 2-4 motifs clearly affect the diagnostic ability of the fingerPRINTScan method, especially when these a models also contain poorly conserved motifs. This feature, unique to only a subset of the fingerprints in the database, seemingly can not be addressed by an alteration of the methodology employed in the scanning process, and must therefore be solved by other means. The simplest approach would be to remove fingerprints of less than 5 motifs from any but the most sensitive analyses, the unfortunate result would be a loss of the scope usually provided by the absent fingerprints, but selectivity would be improved. Alternatively, by defining scoring thresholds (which measure the score range over which true fingerprint family members vary), rather than a catch-all threshold (based on the estimated number of false positive assignments), a view of the variability of a given family can be provided.

Figure 6.15: All of the fingerprints from figure 6.13 were collated, to produce a list of the most frequently occurring fingerprints in the list of false negative assignments (below the 1% and 0.01% thresholds).

Threshold p-value of $2.3e^{-7}$ (1% false positives)

frequency	fingerprint name
136	ZINCFINGER
69	C2HCZNFINGER
29	GFCYSKNOT
10	4FE4SFRDOXIN
9	GPROTEINBRPT
7	WWDOMAIN

Threshold p-value of $8.9e^{-16}$ (0.01% false positives)

frequency	fingerprint name
453	ZINCFINGER
366	GPROTEINBRPT
329	HOMEOBOX
159	THIOREDOXIN
156	AMPBINDING
132	HTHLYSR, PAPAIN
117	HTHARAC, LEURICHRPT
92	C2HCZNFINGER
86	2FE2SFRDOXIN
84	HTHREPRESSR, NUDIXFAMILY
72	CHITINBINDNG
71	4FE4SFRDOXIN
70	HIVVPRVPX
66	FNTYPEIII, GLUCAGON, SH3DOMAIN

While, this does not provide a clear solution to the problems of distinguishing true from false results with 2-4 motif fingerprints, what it does provide is an indication of the range over which true matches are made. Matches to fingerprints below an arbitrary threshold, which nevertheless fall within the realms of their fingerprint's natural score range, can therefore be assessed differently from matches that lie outside of this range.

6.4 Genome Analysis

The potential for the use of fingerPRINTScan in the analysis of large numbers of sequences has been stated previously. The software can be parallelised, and on average can perform more than 700 sequence scans per hour per processor. The resultant data can be placed into a relational database, which provides the facility for performing analyses of the results. These factors, combined with the multiple motif based approach of the fingerprinting process, are ideally suited for fingerPRINTScan's inclusion in a genome annotation or analysis program. As will be discussed later, fingerPRINTScan has already played a role in the annotation of *Drosophila melanogaster* as a component of InterPro (Rubin et al., 2000).

A set of ten genomes were selected for the following analysis. The genomes represent diverse taxa, ranging from eukaryote, bacteria and archaea. These preliminary analyses are concerned mainly with the effect of the application of the thresholds defined in section 6.2. The analysis of large datasets demands that low-scoring, true diagnoses should be sacrificed so as to avoid as many false positives assignments as possible. Figure 6.16 shows the number of assignments made by fingerPRINTScan at different thresholds, from the analysis of ten genomes.

Figure 6.16: Ten genomes from diverse taxa, were selected to observe the effects of variation of the scoring thresholds defined in section 6.2.

The number of sequences identified above the threshold

Organism	Total ORFs	Threshold 1.0% ($2.3e^{-7}$)		Threshold 0.01% ($8.9e^{-16}$)	
		No.	%	No.	%
Bacillus subtilis	4095	880	21.5%	407	9.9%
Drosphila melanogaster	13615	3859	28.3%	1612	11.8%
Mycoplasma genitalium	483	124	25.7%	81	16.8%
Escherichia coli	4246	959	22.6%	460	10.8%
Methanococcus jannaschii	1772	281	15.9%	165	9.3%
Chlamydia pneumoniae	1052	197	18.7%	119	11.3%
Pyrococcus horikoshii	2061	307	14.9%	126	6.1%
Saccaromyces cerevisiae	6191	1138	18.4%	565	9.1%
Rickettsia prowazekii	834	188	22.5%	118	14.2%
Caenorhabditis elegans	18379	3984	21.7%	1857	10.1%

6.4.1 Summary

These simple results indicate that by using this software a number of assignments can be made, which are in line with the diagnosis rates of other similar automated analysis methods. The confirmation step of automatic annotation for TrEMBL results in 0.07% false positive assignments (Fleischmann et al., 1999) with 10% coverage (sequences assigned). Here, using the 0.01% false positive threshold, a similar level of coverage is achieved. As coverage is a function of the size of the resource, in this case the 1360 fingerprints in PRINTS 27.0, increasing this coverage and improving the confidence of the assignments made is something that can not be achieved alone by any single resource; TrEMBL annotation utilises PROSITE *and* IDENTIFY-like patterns to confirm diagnoses. However, these results indicate that fingerPRINTScan is worthy of inclusion into annotation programmes that *can* utilise the combined efforts of multiple search tools; e.g. the InterPro search facility and the TrEMBL annotation suite.

6.5 Sensitivity

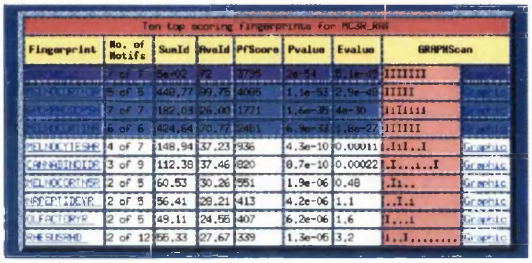
The PRINTS search tool fingerPRINTScan has the ability to report both full *and* partial matches. Providing this flexibility means that fingerPRINTScan can be directed towards the careful (manual) consideration of distant relationships as well as *en masse* diagnoses. So far the only results to be considered have been the top-scoring full fingerprint matches; because high-confidence diagnoses, such as these, are essential for providing functional predictions in large scale analyses. However, the multiple-motif methodology has the potential to identify relationships that are far less clear-cut than the 'full fingerprint match'. Indeed, a partial match to a fingerprint may indicate a novel relationship between the family in question and a distant relative. In cases where an uncharacterised sequence cannot be assigned annotation from pairwise similarity alone, and likewise no membership of a defined family can be confidently prescribed, the identification of any indication of ancestry, however distant, is very important. While a partial match can never be as confident an indicator of familial membership as a full match, it nevertheless provides an assertion to support, or controvert, the accumulation of evidence from other sources. To improve the chance that difficult diagnoses are correctly interpreted a search tool must provide as much information as possible. FingerPRINTScan meets these requirements by displaying differing levels of information relating to the fingerprint match, and then to each of the motif matches; furthermore, it also provides a graphical analysis of motif matches. In order to further supplement and enhance the task of analysing such distant and/or ambiguous results, a new feature has been introduced, which can provide a perspective on the relationships between individual matching fingerprints.

The PRINTS database has recently been supplemented with a resource based on a relational database management system: PRINTS-S (Attwood et al., 2000). This new paradigm provides the facility to describe the relationships between fingerprints, and their components (motifs, sub-sequences, sequences that belong to families, etc.), more

explicitly than it was possible to do in the original. With this development has come the ability to delineate 'parent-child' relationships between individual fingerprints based on their evolutionary heritage. In particular, this hierarchical description of fingerprints, using their super-family, family and sub-family relationships, has proved to be beneficial in the analysis of full and partial matches. Based on the original model of PRINTS, where information regarding relationships *between* fingerprints was only available *within* the annotation of each entry, the observation of a number of matches to a query sequence, may have been ambiguous. By providing a list of scoring matches, as well as access to fingerprint annotation, fingerPRINTScan provided a means for the interpretation of these results. However, by utilising the explicit relationships defined in PRINTS-S, fingerPRINTScan can now provide direct access to this information, which dramatically reduces the effort required to interpret a single result. The beneficial effects of the fact that fingerPRINTScan does not consider fingerprints in PRINTS to be isolated (i.e., they form parent-child relationships) can be clearly observed from the results of matching a sequence to a family with a well defined hierarchy. Scanning the sequence SWISS-PROT:MC3R_RAT (a melanocortin receptor (subtype 3)) against PRINTS 27.0 produces the matches shown in figure 6.17A. The highest scoring match is to the melanocortin receptor family fingerprint; on its own this correctly diagnoses the sequence as a member of the melanocyte-stimulating hormone receptor family (PRINTS:MCRFAMILY). However, there are further lower-scoring matches: MELNOCORTINR, which specifically identifies the melanocortin receptors; MELNOCORTN3R, which is derived from a subtype of the melanocortin receptors (specifically subtype 3); and GPCRRHODOPSN, which describes the rhodopsin-like GPCR super-family. Below the threshold (e-value $1e^{-}4$) there are also a number of partial matches. With the knowledge that the four highest scoring matches share specific parent-child relationships, this result is very clear: the sequence belongs to the rhodopsin-like GPCR super-family, it is specifically a subtype-3 melanocortin receptor, which belongs to the melanocortin specific sub-family of the melanocyte-stimulating

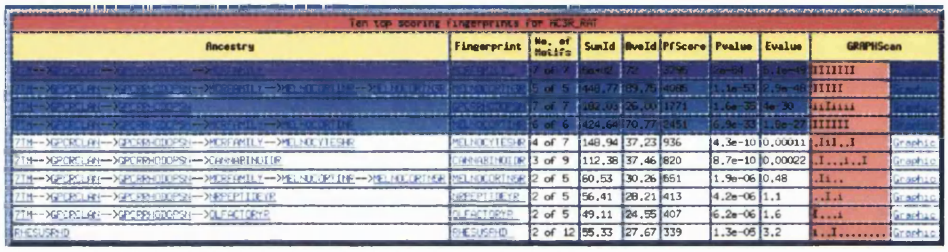
Figure 6.17: Result from the scan of MC3R_RAT against PRINTS 27.0. A) shows the original set of results, and B) shows the effect of supplementing these results with parent-child relationship information from PRINTS-S.

A



Fingerprint	No. of Motifs	SumId	MeId	PfScore	Pvalue	Evalue	GRAPHScan
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	IIIIIII
[illegible]	5 of 5	448.77	99.75	4005	1.1e-53	2.9e-48	IIIII
[illegible]	7 of 7	182.03	26.00	1771	1.6e-25	4e-20	iiIiiii
[illegible]	6 of 6	424.64	70.77	2451	6.9e-33	1.8e-27	IIIIII
MELNOCYTESHR	4 of 7	148.94	37.23	936	4.3e-10	0.00011	.IiI..I
CANNABINOIDR	3 of 9	112.38	37.46	820	8.7e-10	0.00022	.I...i..I
MELNOCORTINSR	2 of 5	60.53	30.26	551	1.9e-06	0.48	.Ii..
NEUROPEPTIDEYR	2 of 5	56.41	28.21	413	4.2e-06	1.1	..I.i
OLFACTORYR	2 of 5	49.11	24.55	407	6.2e-06	1.6	I...i
RHESUSRHD	2 of 12	55.33	27.67	339	1.3e-05	3.2	i..I.....

B



Ancestry	Fingerprint	No. of Motifs	SumId	MeId	PfScore	Pvalue	Evalue	GRAPHScan
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	5 of 5	448.77	99.75	4005	1.1e-53	2.9e-48	IIIIII
[illegible]	[illegible]	7 of 7	182.03	26.00	1771	1.6e-25	4e-20	iiIiiii
[illegible]	[illegible]	6 of 6	424.64	70.77	2451	6.9e-33	1.8e-27	IIIIII
7TM-->GPCRCLAS-->GPCRRHODOPSN-->MELNOCYTESHR	MELNOCYTESHR	4 of 7	148.94	37.23	936	4.3e-10	0.00011	.IiI..I
7TM-->GPCRCLAS-->GPCRRHODOPSN-->CANNABINOIDR	CANNABINOIDR	3 of 9	112.38	37.46	820	8.7e-10	0.00022	.I...i..I
7TM-->GPCRCLAS-->GPCRRHODOPSN-->MELNOCORTINSR	MELNOCORTINSR	2 of 5	60.53	30.26	551	1.9e-06	0.48	.Ii..
7TM-->GPCRCLAS-->GPCRRHODOPSN-->NEUROPEPTIDEYR	NEUROPEPTIDEYR	2 of 5	56.41	28.21	413	4.2e-06	1.1	..I.i
7TM-->GPCRCLAS-->GPCR3GROUP2-->OLFACTORYR	OLFACTORYR	2 of 5	49.11	24.55	407	6.2e-06	1.6	I...i
RHESUSRHD	RHESUSRHD	2 of 12	55.33	27.67	339	1.3e-05	3.2	i..I.....

hormone receptor family (figure 6.17B shows the addition of parent-child relationships to the result in 6.17A). Furthermore, these relationships extend into the low significance scores below the threshold, and all but one of these matches belong to the same super-family.

The extension of similarity, beyond the realms of significance, reflects the importance of providing this degree of annotation to partial matches. Insignificant partially-matching sequences tend to indicate chance events; however, when these partial matches belong to the same super-family, it is possible that they represent real matches that merely lack the mathematical significance to be identified as such. Any such partial match that can find support from insignificant, but related matches, may benefit from the additional information or evidence that this provides. An example is shown in figure 6.18, a sequence TrEMBL:Q9U320 (a predicted ORF from *C. elegans* genome) partially matches GPCRRHODOPSN. The top-scoring match is significant enough to indicate that the sequence is likely to be a member of the super-family, while all other matches are insignificant partials. However, four of the partial matches also belong to the super-family, and although these do not indicate any statistical significance their

biological relevance provides some support for the original diagnosis.

Figure 6.18: Result from the scan of Q9U320 against PRINTS 27.0. A) shows the original set of results, and B) shows the effect of supplementing these results with parent-child relationship information from PRINTS-S.

A

Ten top scoring fingerprints for Q9U320								
Fingerprint	No. of Motifs	SumId	AveId	PfScore	Pvalue	Evalue	GRAPHScan	
[illegible]	6 of 7	1.5e+02	25	1556	7.1e-27	1.8e-21	.iiiiii	
NRPEPTIDEY5R	2 of 12	64.71	32.35	556	3.9e-05	10	....i.....I.	Graphic
PRSINOIDEP4R	2 of 8	54.45	27.23	499	6.8e-05	17	..i...i.	Graphic
NANEUSMPORT	2 of 8	43.58	21.79	365	0.00015	39	.i.i....	Graphic
NEUROKININR	2 of 8	71.33	35.67	361	0.0022	5.7e+02	....i.I.	Graphic
INTEGRINA	2 of 8	46.91	23.46	242	0.0025	6.5e+02	.i...i..	Graphic
GLHYDRLASE20	2 of 8	56.60	28.30	270	0.0026	6.7e+02	i....I..	Graphic
ANGIOTENSINR	2 of 8	56.69	28.35	357	0.0053	1.4e+03	..I...i.	Graphic
FMOXYGENASE2	2 of 9	58.04	29.02	356	0.0059	1.5e+03	I..i.....	Graphic
KIR12CHANNEL	2 of 9	80.81	40.40	330	0.0062	1.6e+03	.I...I...	Graphic

B

Ten top scoring fingerprints for your query									
Ancestry	Fingerprint	No. of Motifs	SumId	AveId	PfScore	Pvalue	Evalue	GRAPHScan	
7TM-->[illegible]-->[illegible]	[illegible]	6 of 7	1.5e+02	25	1536	[illegible]	[illegible]	.iiiiii	
7TM-->GPCRCLAN-->GPCRRHODOPSN-->NRPEPTIDEYR-->NRPEPTIDEY5R	NRPEPTIDEY5R	2 of 12	64.71	32.35	556	3.9e-05	10	....i.....I.	Graphic
7TM-->GPCRCLAN-->GPCRRHODOPSN-->PROSTAGLNDNR-->PRSINOIDEP4R	PRSINOIDEP4R	2 of 8	54.45	27.23	499	6.8e-05	17	..i...i.	Graphic
NANEUSMPORT	NANEUSMPORT	2 of 8	43.58	21.79	365	0.00015	39	.I.i....	Graphic
7TM-->GPCRCLAN-->GPCRRHODOPSN-->NEUROKININR	NEUROKININR	2 of 8	71.33	35.67	361	0.0022	5.7e+02	....i.I.	Graphic
INTEGRINA	INTEGRINA	2 of 8	46.91	23.46	242	0.0025	6.5e+02	.i...i..	Graphic
GLHYDRLASE20	GLHYDRLASE20	2 of 8	56.60	28.30	270	0.0026	6.7e+02	i....I..	Graphic
7TM-->GPCRCLAN-->GPCRRHODOPSN-->ANGIOTENSINR	ANGIOTENSINR	2 of 8	56.69	28.35	357	0.0053	1.4e+03	..I...i.	Graphic
FMOXYGENASE-->FMOXYGENASE2	FMOXYGENASE2	2 of 9	58.04	29.02	356	0.0059	1.5e+03	I..i.....	Graphic
2POREKCHANEL-->KIRCHANNEL-->KIR12CHANNEL	KIR12CHANNEL	2 of 9	80.81	40.40	330	0.0062	1.6e+03	.I...I...	Graphic

6.5.1 Summary

Sensitive searches for distant family members inevitably require manual intervention. In exploiting the sensitivity afforded by the multiple-motif method employed in the construction of fingerprints, fingerPRINTScan can provide support for the analysis of unclear results through the identification of partial matches. The software also attempts to illuminate ambiguous matches by providing access to varying levels of information about any matches made, and to the uniquely rich set of inter-relations between fingerprints in PRINTS.

6.6 Applications

The accessibility of fingerPRINTScan has meant that the software played a larger role than merely the provision of a convenient search tool for the PRINTS WWW site.

From the outset the software was developed with the intent that it would be released freely and openly to the biological community. The paradigm of making software available with minimal restrictions (on the use of either the program itself or redevelopment of its underlying source code) is common today; and high profile software developments like Linux are amongst its most influential proponents. The benefit comes from the freedom that individuals have to reuse and integrate different pieces of software. By being able to use and modify software without restriction, means that time and effort is not wasted on 're-inventing the wheel'; i.e., if the requirement of a component of a sequence analysis package is that it searches PRINTS, then it would be sensible to use fingerPRINTScan and integrate it into the package, rather than developing software from scratch.

As a result of both the above consideration, and its profile as the PRINTS search tool, fingerPRINTScan is currently widely used in numerous commercial, and academic sectors. Usually the incentive to install a private implementation of the software comes from the sensitivity of the data to be analysed or the requirement to analyse large amounts of data, both of which are unsuitable for submission to WWW services. A number of publically available instances of applications of fingerPRINTScan are described in the following sections.

6.6.1 TrEMBL

The TrEMBL database, is an automatically annotated database, which comprises translated coding sequences from the EMBL nucleotide sequence database. The annotation process relies on clustering sequences sequences from an annotated database (SWISS-PROT) using an automated method (e.g., PROSITE). Clusters, or groups, are subsequently analysed to extract annotation that is common among the members. Once annotation has been established for each group TrEMBL sequences are clustered using the same methods. Correspondingly, TrEMBL clusters can be annotated with the

common annotation from correlated SWISS-PROT groups.

Essential for the efficacy of this method is the unambiguous, and consistent, designation of groups of both annotated and unannotated sequences. It is therefore essential to only use the most confident group assignments possible. To determine the confidence or otherwise of groups designated by the PROSITE method, comparisons are made with assignments from the complementary approach implemented in the EMOTIF method. Using this approach 10% of the sequences in TrEMBL can be confidently assigned annotation (Fleischmann et al., 1999). To extend this method requires an increase in the scope of the family coverage provided by the clustering procedure. Both the PROSITE patterns and the confirmatory EMOTIF patterns are derived from the same source (the IDENTIFY database is not used, merely PROSITE patterns are re-encoded using the EMOTIF method), and as a consequence the use of the alternative method only functions as a means of confirming or rejecting the diagnosis. In an effort to increase familial coverage and to increase the potential for utilising overlapping complementary evidence to confirm group designations, the authors express the desire for the inclusion of external databases in this process. Recently, through the integration of PRINTS into InterPro, fingerPRINTScan has provided the means for the use of PRINTS in this annotation process (Steffen Möller, pers. comm.). Currently, over 47,000 sequences in TrEMBL release 14 (June 2000), which contains 351,834 entries, contain links to PRINTS fingerprints.

6.6.2 InterPro

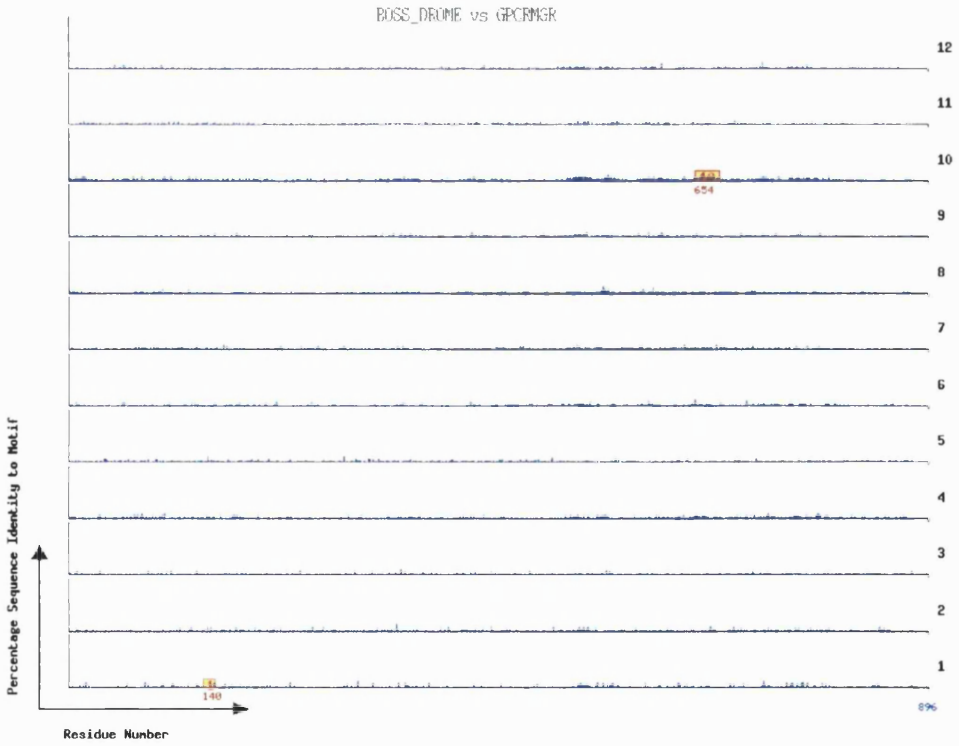
This integrated resource, provides access to patterns from a number of secondary databases. The first release included entries from PROSITE, PRINTS and Pfam, and the latest release also includes ProDom entries. Essential to the construction of InterPro is the ability to merge, and validate, patterns from each of the parent resources that describe a single biological family. InterPro deals with the complexities of the

relationships between entries (either in the same resource or between resources) by defining specific types of parent-child relationships; i.e., sub-type and sub-string relationships. A parent-child relationship recognises the hierarchical structure of the relationship between two entities; e.g., the pattern that describes a particular super-family can be defined as the parent of a number of family patterns, which in turn are parents to sub-family patterns. Parent-child relationships across the pattern databases are complicated by the existence of redundant definitions; e.g., Pfam: 7tm_1 is a profile which characterises the Rhodopsin-like G-protein coupled receptor family, PRINTS: GPCRRHODOPSN provides a similar but non-overlapping characterisation, as does PROSITE: G_PROTEIN_RECEPTOR and G_PROTEIN_RECEPTOR_2. While each entry defines the same family, the diagnostic ability of each resource is quite different. Consequently, the summation of all patterns into a single InterPro entry provides both the largest range of members, and an indication of the most consistent membership (i.e., the intersection of all the sequences matching each of the individual patterns). Patterns that define family membership that resides further down the hierarchy describe more specific diagnoses of function, these are defined with sub-type relationships. Sub-string relationships are shared between patterns that define the same family, but which physically exist within the region described by another pattern; e.g., PROSITE:G_PROTEIN_RECEPTOR is a small RE which covers only a small range of MSA of the family, whereas the HMM Pfam:7tm_1 provides more extensive coverage of the alignment. The inclusion of PRINTS in InterPro has been facilitated by fingerPRINTScan, which functions as component of the integrated search tool and is used in the verification of sequence family membership. The latter is a step akin to the annotation of TrEMBL described above, in which InterPro entries are defined by comparing pattern from its parent resource, and the establishing of overlapping definitions of familial membership. This process ensures that when multiple patterns exist they can be collated into a single InterPro entry; e.g., the entry IPR000276 consists of PROSITE: G_PROTEIN_RECEPTOR, PROSITE: G_PROTEIN_RECEPTOR_2,

Pfam: 7tm_1 and PRINTS: GPCRRHODOPSN.

The differing perspectives that each of the parent resources bring to a diagnosis reflect the utility of InterPro. For example, while Pfam's HMMs provide sensitive diagnoses of distant relationships, the inclusion the extensive sub-family assignments in PRINTS (which can make more specific diagnoses), can yield a more balanced or more precise result. The sequence SWISS-PROT:BOSS_DROME is annotated by Pfam (version 5) as 7tm_3, which indicates its membership of the metabotropic glutamate GPCR family (MGR). This family of proteins is characterised by its seven transmembrane architecture, its involvement in the inositol phosphate calcium signalling pathway and its coupling with G-proteins. However, using fingerPRINTScan to scan BOSS_DROME against the PRINTS MGR family fingerprint (GPCRMGR) shows no significant similarity (figure 6.19). Further investigation using PSI-BLAST shows that

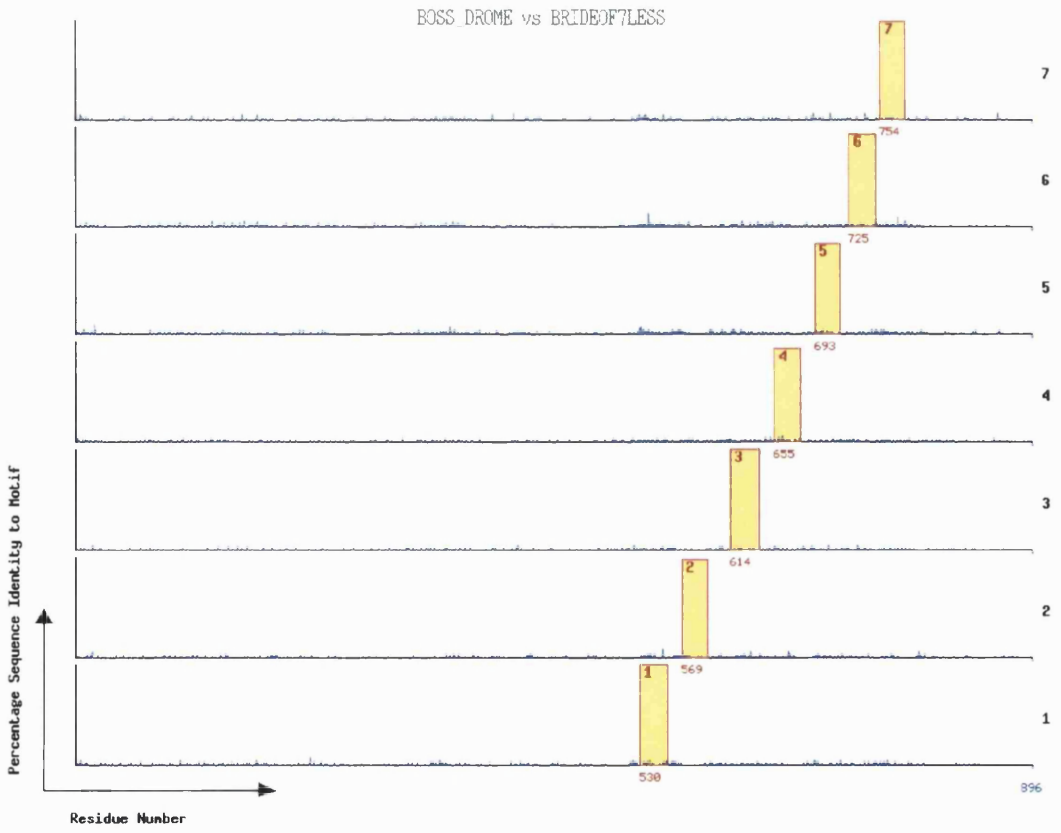
Figure 6.19: A GRAPHScan plot of the motif of PRINTS:GPCRMGR against SWISS-PROT:BOSS_DROME.



BOSS_DROME does identify weak matches to some 'MGR-like' proteins; by integrat-

ing these MGR-like proteins into the PSI-BLAST profile and iterating it is possible to identify some MGR members. However, this kind of transitive assignment does not directly indicate that BOSS_DROME is a member of the MGR family, merely that it shares similarities with proteins that in turn share similarities with MGR members. Close examination of the Pfam seed-alignment reveals a number of these low scoring MGR-like sequences. Hence, the sensitivity of the Pfam model, in this case, reveals a relationship that cannot be clearly verified; and unfortunately this speculative relationship is carried over into a diagnosis of the sequence. The perspective that PRINTS provides on this relationship is clear, there is little evidence to link BOSS_DROME to the MGR fingerprint (figure 6.19); however, its own family fingerprint (BRIDEOF7LESS) provides a clear match (figure 6.20) and an unambiguous diagnosis.

Figure 6.20: A GRAPHScan plot of the fingerprint PRINTS:BRIDEOF7LESS against SWISS-PROT:BOSS_DROME.



The utility of InterPro as a genome annotation tool was demonstrated in a comparative

analysis of three eukaryote genomes performed by Rubin et al. (2000). The authors indicate that using a combination of InterPro analyses and manual inspection, ~54% of the *D. melanogaster*, ~45% of *C. elegans* and ~48% of *S. cerevisiae* proteomes could be assigned.

6.6.3 Other applications

Many researchers, from both the commercial and academic sectors, across the world maintain a personal installation of fingerPRINTScan; a number of the higher profile applications are detailed below.

Researchers from the San Diego Supercomputer Center (SDSC) at the university of California, San Diego, provide WWW-based access to their Biology Workbench. This project is aimed at integrating nucleic acid and protein sequence databases with many of the most popular analysis tools in order to supply a single interface with all of the functionality that a computational molecular biologist may require. This type of integration removes the burden of exchanging data between disparate and incompatible WWW-based search tools and databases. The sequence analysis tools provided by the Workbench range from BLAST and FASTA searches to the use of ClustalW to produce MSAs. Currently the only pattern database analyses that are available are PROSITE and PRINTS (via fingerPRINTScan).

Bionavigator, a commercial project (situated in Sydney, Australia), similarly offers a wide range of biological sequence analysis tools packaged together in a single interface; interest in the use of fingerPRINTScan (FPS) has been clearly stated (Steve Taylor, pers. comm.), but at present the software has not been included in the release version of the product. Likewise, the PRINTS search software is soon to be integrated into the Artemis annotation tool used at the Sanger Centre, Cambridge, U.K. for microbial genome annotation.

To summarise, since its inception as the PRINTS database search tool fingerPRINTScan,

has become a widely used sequence analysis tool; it now provides supports for a wide range of annotation roles, and has emerged as an integral component of the current picture of pattern database driven sequence analysis.

Chapter 7

Discussion and Conclusions

This thesis has shown how the development of fingerprints for the PRINTS database, and the understanding of the processes of pattern based sequence analysis, led to the appreciation of the need for a search tool that could fully exploit the properties inherent in multiple motif methodology. It has also shown how the analysis of these and other methods, has influenced the design and construction of a search tool capable of realising these aims.

The search tool, fingerPRINTScan, attempts to balance the contradictory requirements of sequence analysis (the need for selectivity *and* sensitivity), by providing diagnoses based on the strength of affirmatory multiple motif matches, yet never discarding matches that fail to fully match a fingerprint. Instead of relying purely on statistical calculations to reveal the highest scoring or most likely match, fingerPRINTScan takes into account the biological context of motifs. The additional information supplied by knowledge of the order, or positions, of motifs within the alignment, provide the basis for the identification of the most likely match to a fingerprint, and the rejection of spurious matches.

Ultimately, any sequence analysis tool must rely on a scoring scheme to provide discrimination between sequences that match because they are homologous, and sequences that match due to chance. For any scoring scheme this discrimination is not absolute. Both true and false scores exist in overlapping continuous distributions; and the best that can be achieved is to select the scoring scheme that produces the minimum of overlap. For the analyses performed by fingerPRINTScan the generation of p-values from a profile matrix proved to perform the most effective separation of true and false results.

Where no discrimination is possible (i.e., within the overlap), it was shown that fingerprints of 2 and 3 motifs are most often associated with providing these poor diagnoses. The weakness of these particular patterns stems from an inherent failure of the fingerprinting process itself. The discriminatory power of any pattern is strongly reliant on both the conservation of the MSA and the length of the conserved region. One of

the benefits of multiple motif based patterns is the fact that many smaller-patterns can function as effectively as a single larger one. This means that alignments containing divergent sequences can still be effectively described, even if there are few long stretches of conservation. However, as the relationship described by the MSA becomes more and more distant, the degree of conservation falls. As a consequence, regions suitable for the definition of motifs become smaller and more scarce (motifs must be defined from regions that do not contain gaps). Consequently, fingerprints, suffer from the dual effects of reduced motifs and diminished conservation in the remaining ones. Therefore, beyond a certain point, the ability of a fingerprint to distinguish true members of its family from random matches is removed.

By modulating the threshold p-value that draws the line between true and false diagnoses, both highly selective *and* sensitive searches can be supported. Naturally, a compromise is made; a sensitive search tolerates false positive assignments, while a selective search minimises false positives at the cost of losing a proportion of true matches. The p-value scoring scheme provides the most effective compromise by resulting in the minimal loss of true assignments as selectivity is increased. It was indicated, for a range of proteomes, that fingerPRINTScan is capable of making levels of assignment comparable with the automated analysis tool used for the annotation of the TrEMBL database.

At the other end of the scale, sensitive analyses are supported by the ability to detect partial matches. Such analyses are invariably performed manually, therefore it is necessary to provide as much information to the user as is available. To meet this requirement, and in order to facilitate the interpretation of ambiguous results, fingerPRINTScan supplements each match between a sequence and a fingerprint with additional information. This includes a graphical representation of the match, and details of each matching motif. Furthermore, by revealing the hierarchical relationships that exist between fingerprints, it is possible to illuminate the biological context of any matches that are made.

One of the very clear conclusions that can be drawn from the work presented in this thesis is the identification of the distinct niche that fingerprints occupy. As familial discriminators fingerprints excel in the definition of specific patterns that facilitate the sub-division of families, but produce weak models from alignments containing distant relationships. Conversely models such as Profile-HMMs tend to describe divergent relationships well, but fail to provide very specific diagnoses. InterPro provides a real opportunity for PRINTS to concentrate on providing strong and specific diagnoses, while Pfam and profiles concentrate on the modelling of domains and distant relationships.

This conclusion highlights an important caveat about sequence analysis, which should be reiterated. The warning is that every pairwise or pattern search tool has its own specific strengths and weaknesses; therefore, none should be trusted explicitly or used in isolation. This is especially important in this era of high throughput genomics; as more and more sequences obtain their annotation via similarity alone, the consequences of not heeding this warning could be dire.

The work presented in this thesis has provided a search tool for PRINTS; hopefully it has also created a facility for the dissemination of its unique perspective into the analyses of as many annotation programmes as possible.

List of abbreviations

ADSP 'Algorithms and Data Structures for Protein sequence analysis'

DDJB DNA Data Bank of Japan

DP Dynamic Programming

EBI European Bioinformatics Institute

EMBL European Molecular Biology Laboratory

EST Expressed Sequence Tag

EVD Extreme Value Distribution

GCM Genetic Code Matrix

GSDB Genome Sequence DataBase

HMM Hidden Markov Model

INSD International Nucleotide Sequence Database Collaboration

IM Identity Matrix

ISREC Swiss Institute for Experimental Cancer Research

JIPID Japan International Protein Sequence Database

MDM Mutation Data Matrix

MIPS Munich Information Center for Protein Sequences

MSA Multiple Sequence Alignment

MSS Maximally Scoring Sub-sequence

NCBI National Center for Biotechnology Information

ORF Open Reading Frame

PAM Point Accepted Mutation

PIR Protein Information Resource

PSD Protein Sequence Database

PSSM Position Specific Scoring Matrix

Profile-HMM Profile-Hidden Markov Model

RE Regular Expression

SIB Swiss Institute of Bioinformatics

SQL Structured Query Language

SRS 'Sequence Retrieval System'

TIGR The Institute for Genomic Research

WWW World Wide Web

Bibliography

Altschul, S. (1989). Gap costs for multiple sequence alignment. *J Theor Biol*, 138(3):297–309.

Altschul, S. (1998). Generalized affine gap costs for protein sequence alignment. *Proteins*, 32(1):88–96.

Altschul, S., Boguski, M., Gish, W., and Wootton, J. (1994). Issues in searching molecular sequence databases. *Nat Genet*, 6(2):119–29.

Altschul, S. and Gish, W. (1996). Local alignment statistics. *Methods Enzymol*, 266:460–80.

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

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

Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Birney, E., Bucher, P., Codani, J.-J., Corpet, F., Croning, M., Durbin, R., Etzold, T., Fleischmann, W., Gouzy, J., Hermajakob, H., Jonassen, I., Kahn, D., Kanapin, A., Schneider, R., Servant, F., and Zdobnov, E. (2000). InterPro - An integrated documentation resource for protein families, domains and functional sites. *CCP11 Newsletter*, 3(4).

Attwood, T. (2000). The quest to deduce protein function from sequence: the role of pattern databases. *Int J Biochem Cell Biol*, 32(2):139–55.

Attwood, T. and Beck, M. (1994). PRINTS–a protein motif fingerprint database. *Protein Eng*, 7(7):841–8.

Attwood, T., Croning, M., Flower, D., Lewis, A., Mabey, J., Scordis, P., Selley, J., and Wright, W. (2000). PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res*, 28(1):225–7.

Bailey, T. and Gribskov, M. (1998a). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14(1):48–54.

Bailey, T. and Gribskov, M. (1998b). Methods and statistics for combining motif match scores. *J Comput Biol*, 5(2):211–21.

Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. *Nucl. Acids. Res.*, 28(1):45–48.

Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G., and Tuli, M. A. (2000). The EMBL nucleotide sequence database. *Nucl. Acids. Res.*, 28:19–23.

Barker, W. C., Garavelli, J. S., Huang, H., McGarvey, P. B., Orcutt, B., Srinivasaroao, G. Y., Xiao, C., Yeh, L. S., Ledley, R. S., Janda, J. F., Pfeiffer, F., Mewes, H. W., Tsugita, A., and Wu, C. (2000). The Protein Information Resource (PIR). *Nucl. Acids. Res.*, 28:41–44.

Bateman, A., Birney, E., Durbin, R., Eddy, S., Howe, K., and Sonnhammer, E. (2000). The PFAM protein families database. *Nucleic Acids Res*, 28(1):263–6.

Baxevanis, A. (2000). The molecular biology database collection: an online compilation of relevant database resources. *Nucleic Acids Res*, 28(1):1–7.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Rapp, B., and Wheeler, D. (2000). GenBank. *Nucleic Acids Res*, 28(1):15–8.

Bleasby, A., Akrigg, D., and Attwood, T. (1994). OWL–a non-redundant composite protein sequence database. *Nucleic Acids Res*, 22(17):3574–7.

Boguski, M. (1999). Biosequence exegesis. *Science*, 286(5439):453–5.

Brenner, S. (1998). Practical database searching. *Trends Guide to Bioinformatics, Elsevier*, pages 9–12.

Brenner, S. (1999). Errors in genome annotation. *Trends Genet*, 15(4):132–3.

Brocchieri, L. and Karlin, S. (1998). A symmetric-iterated multiple alignment of protein sequences. *J Mol Biol*, 276(1):249–64.

Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. *Comput Chem*, 20(1):3–23.

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(1):267–9.

Dayhoff, M., editor (1965a). *The Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington, D.C.

Dayhoff, M. (1965b). Computer aids to protein sequence determination. *J Theor Biol*, 8(1):97–112.

Dayhoff, M. (1974). Computer analysis of protein sequences. *Fed Proc*, 33(12):2314–6.

Dayhoff, M. (1978). Matrices for detecting distant relationships. *Atlas Protein Seq. Struct.*, 5(3):353–358.

Dembo, A., Karlin, S., and Zeitouni, O. (1994). Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.*, 22(4):2022–39.

Depiereux, E. and Feytmans, E. (1992). MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. *Comput Appl Biosci*, 8(5):501–9.

Discala, C., Benigni, X., Barillot, E., and Vaysseix, G. (2000). DBcat: a catalog of 500 biological databases. *Nucleic Acids Res*, 28(1):8–9.

Dodge, C., Schneider, R., and Sander, C. (1998). The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res*, 26(1):313–5.

Doolittle, F. (2000). On the trial of protein sequences. *Bioinformatics*, 16:24–33.

Doolittle, R. (1986). *Of URFs and ORFs: A primer on how to analyse derived amino acid sequences*. University Science Books.

Eddy, S. (1996). Hidden Markov models. *Curr Opin Struct Biol*, 6(3):361–5.

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9):755–63.

Etzold, T., Ulyanov, A., and Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*, 266:114–28.

Fitch, W. (1969). Locating gaps in amino acid sequences to optimize the homology between two proteins. *Biochem Genet*, 3(2):99–108.

Fitch, W. (1970). Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113.

Fitch, W. (2000). Homology a personal view on some of the problems. *Trends Genet*, 16(5):227–31.

Fleischmann, W., Moller, S., Gateau, A., and Apweiler, R. (1999). A novel method for automatic functional annotation of proteins. *Bioinformatics*, 15(3):228–33.

Galtier, N., Gouy, M., and Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci*, 12(6):543–8.

Gibbs, A. and McIntyre, G. (1970). The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur J Biochem*, 16(1):1–11.

Gogarten, J. and Olendzenski, L. (1999). Orthologs, paralogs and genome comparisons. *Curr Opin Genet Dev*, 9(6):630–6.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J Mol Biol*, 162(3):705–8.

Gracy, J. and Argos, P. (1998). DOMO: a new database of aligned protein domains. *Trends Biochem Sci*, 23(12):495–7.

Gribskov, M., Luthy, R., and Eisenberg, D. (1990). Profile analysis. *Methods Enzymol*, 183:146–59.

Gribskov, M., McLachlan, A., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, 84(13):4355–8.

Grundy, W., Bailey, T., Elkan, C., and Baker, M. (1997). Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci*, 13(4):397–406.

Harger, C., Chen, G., Farmer, A., Huang, W., Inman, J., Kiphart, D., Schilkey, F., Skupski, M., and Weller, J. (2000). The genome sequence DataBase. *Nucleic Acids Res*, 28(1):31–2.

Henikoff, J. and Henikoff, S. (1996). Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci*, 12(2):135–43.

Henikoff, S., Greene, E., Pietrokovski, S., Bork, P., Attwood, T., and Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*, 278(5338):609–14.

Henikoff, S. and Henikoff, J. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res*, 19(23):6565–72.

Henikoff, S. and Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9.

Henikoff, S. and Henikoff, J. (1993). Performance evaluation of amino acid substitution matrices. *Proteins*, 17(1):49–61.

Henikoff, S., Henikoff, J., Alford, W., and Pietrokovski, S. (1995). Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, 163(2):GC17–26.

Henikoff, S., Henikoff, J., and Pietrokovski, S. (1999). Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, 15(6):471–9.

Henikoff, S., Wallace, J., and Brown, J. (1990). Finding protein similarities with nucleotide sequence databases. *Methods Enzymol*, 183:111–32.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucl. Acids. Res.*, 27:215–219.

Hubbard, T. (1997). New horizons in sequence analysis. *Curr Opin Struct Biol*, 7(2):190–3.

Jacob, F. (1977). Evolution and tinkering. *Science*, 196(4295):1161–6.

Junker, V., Apweiler, R., and Bairoch, A. (1999). Representation of functional information in the SWISS-PROT data bank. *Bioinformatics*, 15(12):1066–7.

Karlin, S. and Altschul, S. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, 87(6):2264–8.

Karp, P. (1998). What we do not know about sequence analysis and sequence databases. *Bioinformatics*, 14(9):753–4.

Murvai, J., Vlahovicek, K., Barta, E., Cataletto, B., and Pongor, S. (2000). The SBASE protein domain library, release 7.0: a collection of annotated protein sequence segments. *Nucleic Acids Res*, 28(1):260–2.

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.

Nevill-Manning, C., Sethi, K., Wu, T., and Brutlag, D. (1997). Enumerating and ranking discrete motifs. *Ismb*, 5:202–9.

Parry-Smith, D. (1990). *Algorithms and data structures for protein sequence analysis.* PhD thesis, University of Leeds.

Parry-Smith, D. and Attwood, T. (1991). SOMAP: a novel interactive approach to multiple protein sequences alignment. *Comput Appl Biosci*, 7(2):233–5.

Parry-Smith, D. and Attwood, T. (1992). ADSP–a new package for computational sequence analysis. *Comput Appl Biosci*, 8(5):451–9.

Parry-Smith, D., Payne, A., Michie, A., and Attwood, T. (1998). CINEMA–a novel colour INteractive editor for multiple alignments. *Gene*, 221(1):GC57–63.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448.

Perkins, D. and Attwood, T. (1995). VISTAS: a package for VIsualizing STructures and sequences of proteins. *J Mol Graph*, 13(1):73–5, 62.

Perkins, D. and Attwood, T. (1996). XFINGER: a tool for searching and visualising protein fingerprints and patterns. *Comput Appl Biosci*, 12(2):89–94.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000). The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res*, 28(1):141–5.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94.

Rubin, G., Yandell, M., Wortman, J., Gabor Miklos, G., Nelson, C., Hariharan, I., Fortini, M., Li, P., Apweiler, R., Fleischmann, W., Cherry, J., Henikoff, S., Skupski, M., Misra, S., Ashburner, M., Birney, E., Boguski, M., Brody, T., Brokstein, P., Celniker, S., Chervitz, S., Coates, D., Cravchik, A., Gabrielian, A., Galle, R., Gelbart, W., George, R., Goldstein, L., Gong, F., Guan, P., Harris, N., Hay, B., Hoskins, R., Li, J., Li, Z., Hynes, R., Jones, S., Kuehl, P., Lemaitre, B., Littleton, J., Morrison, D., Mungall, C., O'Farrell, P., Pickeral, O., Shue, C., Vosshall, L., Zhang, J., Zhao, Q., Zheng, X., Zhong, F., Zhong, W., Gibbs, R., Venter, J., Adams, M., and Lewis, S. (2000). Comparative genomics of the eukaryotes. *Science*, 287(5461):2204–15.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68.

Sanger, F. (1988). Sequences, sequences, and sequences. *Annu Rev Biochem*, 57:1–28.

Sanger, F. and Coulson, A. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*, 94(3):441–8.

Sanger, F. and Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. *Biochem. Journal.*, 49:481–490.

Schimmel, P. and Ribas De Pouplana, L. (2000). Footprints of aminoacyl-tRNA synthetases are everywhere. *Trends Biochem Sci*, 25(5):207–9.

Scordis, P., Flower, D., and Attwood, T. (1999). FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics*, 15(10):799–806.

Smith, R. and Smith, T. (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng*, 5(1):35–41.

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

Srinivasarao, G., Yeh, L., Marzec, C., Orcutt, B., Barker, W., and Pfeiffer, F. (1999). Database of protein sequence alignments: PIR-ALN. *Nucleic Acids Res*, 27(1):284–5.

Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H., and Gojobori, T. (2000). DNA data bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Res*, 28(1):24–6.

Taylor, W. (1986). The classification of amino acid conservation. *J Theor Biol*, 119(2):205–18.

Taylor, W. (1997). Residual colours: a proposal for aminochromography. *Protein Eng*, 10(7):743–6.

Thompson, J., Higgins, D., and Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids. Res.*, 24:4876–4882.

Trifonov, E. (2000). Earliest pages of bioinformatics. *Bioinformatics*, 16(1):5–9.

Watson, J. and Crick, F. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 4356.

Wheelan, S. and Boguski, M. (1998). Late-night thoughts on the sequence annotation problem. *Genome Res*, 8(3):168–9.

Wray, G. and Abouheif, E. (1998). When is homology not homology? *Curr Opin Genet Dev*, 8(6):675–80.

Wu, R. (1978). DNA sequence analysis. *Annu Rev Biochem*, 47:607–34.

Zuckerkandl, E. (1975). The appearance of new structures and functions during evolution. *Journal of Molecular Evolution*, 7:1–57.