The role of pattern databases in sequence analysis

Terri K. Attwood

Date received (in revised form): 21st October 1999

Terri Attwood
holds a Royal Society University Research Fellowship, a concurrent Senior Lectureship in the School of Biological Sciences at the University of Manchester and a Visiting Fellowship at the European Bioinformatics Institute. Her research interests include development of software and databases to facilitate protein sequence alignment and pattern recognition.

Abstract

In the wake of the numerous now-fruitful genome projects, we are entering an era rich in biological data. The field of bioinformatics is poised to exploit this information in increasingly powerful ways, but the abundance and growing complexity both of the data and of the tools and resources required to analyse them are threatening to overwhelm us. Databases and their search tools are now an essential part of the research environment. However, the rate of sequence generation and the haphazard proliferation of databases have made it difficult to keep pace with developments. In an age of information overload, researchers want rapid, easy-to-use, *reliable* tools for functional characterisation of newly determined sequences. But what are those tools? How do we access them? Which should we use? This review focuses on a particular type of database that is increasingly used in the task of routine sequence analysis – the so-called pattern database. The paper aims to provide an overview of the current status of pattern databases in common use, outlining the methods behind them and giving pointers on their diagnostic strengths and weaknesses.

Keywords: *bioinformatics, protein sequence, sequence alignment, similarity search, pattern recognition, function annotation*

Introduction

There are now hundreds of databanks around the world housing information that floods from the genome projects. The endeavour to store and analyse these vast quantities of data has required increasing levels of automation. However, automation carries a price. For example, although software robots are essential to the process of functional annotation of newly determined sequences, they pose a threat to information quality because they can introduce and propagate mis-annotations.[1] Although the curators strive to improve the quality of their resources, databases nevertheless carry the indelible scars of time and are far from perfect. To get the most from current biological databases it is thus important to have an understanding both of their powers and of their pitfalls.

Automation carries a price

To characterise a new sequence, the first step usually involves trawling a sequence database with tools such as BLAST[2] or FASTA.[3] Such searches quickly reveal similarities between the query and a range of database sequences. The trick then lies in the reliable inference of homology (the verification of a divergent evolutionary relationship) and, from this, the inference of function. Ideally, a search output will show unequivocal similarity to a well-characterised protein over the full length of the query, providing sufficient information to make a sensible diagnosis. Sometimes, however, an output will reveal no significant hits or, more commonly, will furnish a list of partial matches to diverse proteins, many of which are uncharacterised, or possess dubious or contradictory annotations.[4]

There are several reasons why such searches might not give direct answers. For example, the growth of sequence databases and their population by greater numbers of poorer-quality partial sequences makes it increasingly likely that high-scoring matches will be made to a query simply by chance. Low-complexity matches, in particular, may swamp search outputs – these are parts of a sequence that have high densities of particular residues (eg poly-GxP, such as occurs in sequences

T. K. Attwood,
School of Biological Sciences,
University of Manchester,
Oxford Road,
Manchester M13 9PT, UK

Tel: +44 (0) 161 275 5766
Fax: +44 (0) 161 275 5082

Modules and domains cause different problems

like collagen, or poly-glutamine tracts that occur in Huntingdon's disease protein etc). Although it is possible to mask such sequences, this can also create complications. The modular and domain nature of many proteins also causes problems on different levels. When matching multi-domain proteins, it may not be clear which domain or domains correctly correspond to the query. Even if the right domain has been identified, it may not be appropriate to transfer the functional annotation to the query because the function of the matched domain may be different, depending on its biological context. Similar issues arise with the existence of multi-gene families, because database search techniques cannot differentiate between orthologues (usually the functional counterparts of a sequence in another species) and paralogues (homologues that perform different but related functions within the same organism).

Searching pattern databases is more selective

Searches don't differentiate orthologues and paralogues

Correct functional assignment difficult to achieve

Given these complexities, correct functional assignment from searches of sequence databases alone can be difficult or impossible to achieve. As a result, it is now customary also to search a range of 'pattern' databases, so-called because they distil patterns of residue conservation within groups of related sequences into discriminators that aid family diagnosis. Searching pattern databases is thus more selective than sequence database searching because discriminators are designed to detect particular families. Different analytical approaches have been used to create a bewildering array of discriminators, which are variously termed regular expressions, profiles, fingerprints, blocks, etc.[5] – these terms are summarised in Figure 1. The different descriptors have different diagnostic strengths and weaknesses and different areas of optimum application, and have been used to generate different pattern databases, which also differ in content! The aim of this paper is to provide an overview of pattern databases in common use and to offer pointers on how best to use them. As this is a fast moving area, a list of web addresses is given in Table 1 to allow readers to obtain current information on the resources discussed.

Table 1: Web addresses of pattern and alignment databases in common use. For a more exhaustive list, refer to the annual database issue of *Nucleic Acids Research* (http://www3.oup.co.uk/nar/)

PROSITE	http://www.expasy.ch/prosite/
Blocks	http://www.blocks.fhcrc.org/
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
IDENTIFY	http://dna.Stanford.EDU/identify/
Profiles	http://www.isrec.isb-sib.ch/software/PFSCAN_form.html
Pfam	http://www.sanger.ac.uk/Software/Pfam/
ProDom	http://www.toulouse.inra.fr/prodom.html
SBASE	http://www.icgeb.trieste.it/sbase/
PIR-ALN	http://www-nbrf.georgetown.edu/pirwww/search/textpiraln.html
PROT-FAM	http://vms.mips.biochem.mpg.de/mips/programs/classification.html
DOMO	http://www.infobiogen.fr/~gracy/domo/
ProClass	http://pir.georgetown.edu/gfserver/proclass.html
ProtoMap	http://www.protomap.cs.huji.ac.il/
PIMA	http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html
InterPro	http://www.ebi.ac.uk/interpro/

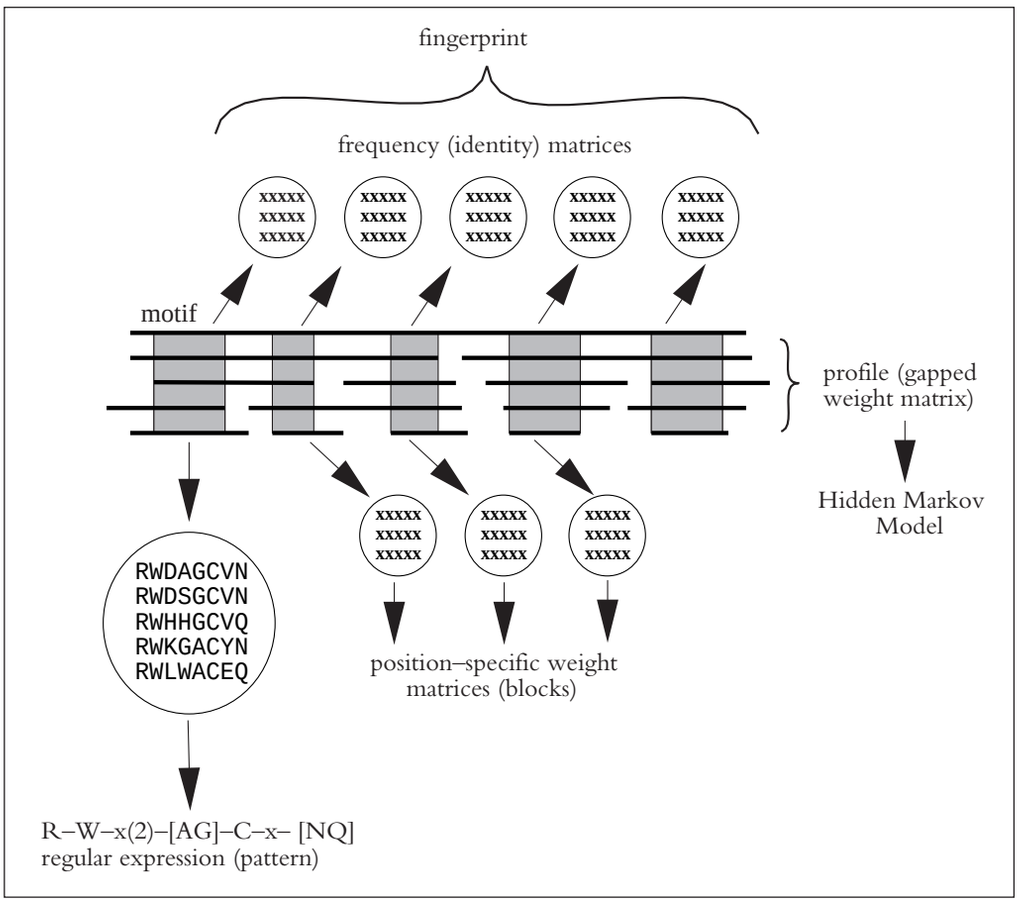

Figure 1: At the heart of sequence analysis methods is the multiple sequence alignment. Application of these methods involves the derivation of some kind of representation of conserved features of the alignment, which may be diagnostic of structure or function. Various terms are used to describe the different types of data representation, as shown. Within a single conserved region (motif), the sequence information may be reduced to a consensus expression (a regular expression), often simply referred to as a pattern. In this example, square brackets indicate residues that are allowed at this position of the motif and x denotes any residue, the (2) indicating that any residue can occupy *consecutive* positions in the motif. The term used to describe groups of motifs in which all the residue information is retained within a set of frequency (identity) matrices is a fingerprint, or signature. Adding a scoring scheme to such sets of frequency matrices results in position-specific weight matrices, or blocks. Using information from extended conserved regions that include gaps (usually referred to as domains) gives rise to profiles; and probabilistic models derived from alignment profiles are termed hidden Markov models

THE METHODS BEHIND THE DATABASES

Alignment underpin pattern databases

At the heart of the analysis methods that underpin pattern databases is the multiple sequence alignment. When building an alignment, as more distantly related sequences are included, insertions are often required to bring equivalent parts of adjacent sequences into the correct register, as illustrated schematically in Figure 2.[6] As a result of this gap insertion process, islands of conservation emerge from a backdrop of mutational change. These regions, usually termed motifs or blocks, are typically around 10–20 residues in length and tend to correspond to the core structural or functional elements of the protein.

The conserved nature of motifs effectively provides us with a set of

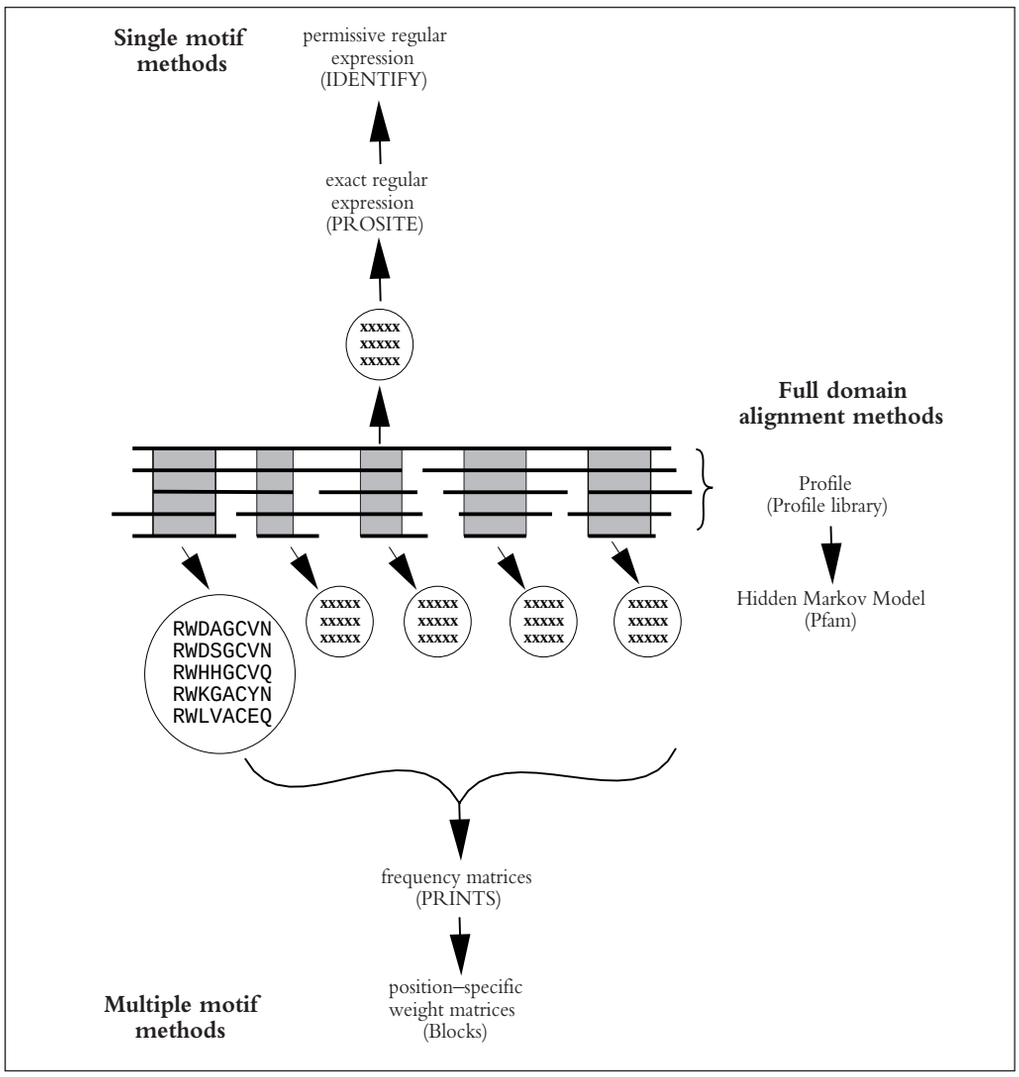

Figure 2: Illustration of the three principal methods for building pattern databases: ie using single motifs, multiple motifs and full domain alignments. Single-motif (regular expression pattern) approaches have given rise to the PROSITE and IDENTIFY databases; multiple-motif methods have spawned the Blocks and PRINTS databases; and domain alignment methods have resulted in the Profiles and Pfam resources

Conserved motifs provide family blueprints

familial blueprints, and different techniques have evolved to exploit this fact. As shown in Figure 2, the methods fall broadly into three categories, depending on whether they use single motifs, multiple motifs or full domain alignments. All of these methods involve the derivation of some kind of discriminatory representation of aspects of the alignment, providing a characteristic signature for the family that can be used to diagnose future query sequences.

Minimise overlap of time-positives and true-negatives

The diagnostic success of the different methods depends on how reliably true family members (true-positives) can be distinguished from non-family members (true-negatives). In practice, there is a crucial balance between the number of incorrect matches that are made (false-positives) and the number of correct matches that are missed (false-negatives) at a given scoring threshold. As shown in Figure 3, for a given search, this requires the distribution of true-positive matches to be resolved from that of the true-negatives, such that the overlap between

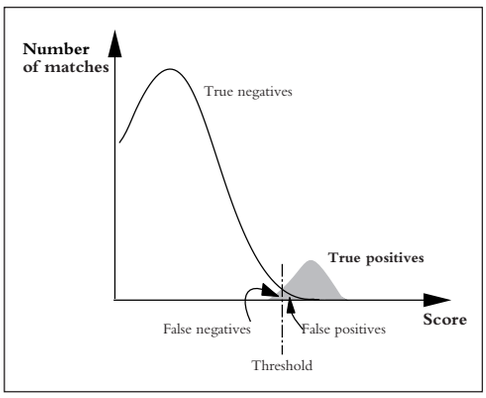

Figure 3: Resolving true and false matches. In a database search, the desire is to establish which sequences are related (true-positive) and which are unrelated (true-negative). At a given scoring threshold, it is likely that several unrelated sequences will match a search pattern erroneously (so-called false-positives), and several correct matches will fail to be diagnosed (false-negatives). In sequence analysis, the challenge is to improve diagnostic performance by capturing all (or the majority) of true-positive family members, including no (or few) false-positives, and minimising or precluding false-negatives

Regular expressions perform poorly with divergent families

Difficult to determine which matches are correct

Mathematical significance is not biological proof

Reduce a motif to a consensus expression

them is minimised or eliminated. This is important because, for matches in the overlapping area, it can be difficult or impossible to determine which are correct (statistical approaches are used to assign confidence levels to matches in this area, but mathematical significance does not give biological proof). The different analytical methods that have been designed to tackle these issues are outlined below.

Single-motif methods

Of the various approaches, single-motif (regular expression pattern) methods are easiest to understand. The idea is that a particular protein family can be characterised by the single most conserved, often functionally important, region (eg an enzyme active site) observed in a sequence alignment. The motif is reduced to a consensus expression in which all but the most significant residue information is discarded. For example, the expression D-x-{KR}-[NQ] means that a conserved aspartic acid (D) residue is followed by an arbitrary residue (x) and any residue *except* lysine (K) or arginine (R), and finally a polar residue, which may be asparagine (N) or glutamine (Q). No other residues or residue combinations are tolerated by the expression; matches to it must therefore be exact, or will be disregarded.

So rigid is this syntax that regular expression patterns do not perform well when used to represent highly divergent protein families. For example, such patterns will fail to match significant sequences if they contain a *single* amino acid difference. The sequence DARN is thus a mis-match, in spite of matching the above expression in all but one position (it has a forbidden arginine as its third residue). Conversely, a pattern will match *anything* that corresponds to it exactly, regardless of whether it is a true family member. The problem is that matches to single motifs lack biological context – a match to a pattern is just a match to a pattern, and may well only be fortuitous. To assess the likelihood of a match being 'real', it must be verified with corroborating evidence, whether via other database searches, the literature or experiment.

An approach that addresses the strict nature of exact regular expression matching is to assign amino acid residues to distinct, but overlapping, substitution groups corresponding to various biochemical properties (eg charge and size), as shown in Table 2. This is biologically sensible because each amino acid has several properties and can serve different functions, depending on its biochemical context.[7] However, although the technique is more flexible, its inherent permissiveness has an inevitable signal-to-noise trade-off, ie resulting patterns not only have the potential to make more true-positive matches, but they will consequently also match more false-positives. For example, the sequence EVEN, which would be

Table 2: Overlapping sets of amino acids and their properties. These are used to create the permissive regular expressions used as the basis of the IDENTIFY resource

Residue property	Residue groups
Small	Ala, Gly
Small hydroxyl	Ser, Thr
Basic	Lys, Arg
Aromatic	Phe, Tyr, Trp
Basic	His, Lys, Arg
Small hydrophobic	Val, Leu, Ile
Medium hydrophobic	Val, Leu, Ile, Met
Acidic/amide	Asp, Glu, Asn, Gln
Small/polar	Ala, Gly, Ser, Thr, Pro

excluded by the exact regular expression above, would be matched by the permissive one (because Asp and Glu belong to the same group), even if aspartic acid were biologically mandatory at the first position of the motif.

Multiple-motif methods

Intricate scoring system

In response to these problems, diagnostic techniques evolved to exploit multiple motifs. Within a sequence alignment, it is common to find several motifs that characterise the aligned family. Diagnostically, it makes sense to use many or all such regions to build a family signature. In a database search, there is then a greater chance of identifying a distant relative, whether or not all parts of the signature are matched. For example, a sequence that matches only four of seven motifs may still be diagnosed as a true match if the motifs are matched in the correct order in the sequence and the distances between them are consistent with those expected of true neighbouring motifs. The ability to tolerate mis-matches, both at the level of individual residues within motifs, and at the level of motifs within the complete signature, makes multiple-motif matching a powerful diagnostic approach.

Multiple-motif matching is a powerful diagnostic approach

Different multiple-motif methods have arisen, depending on the technique used to detect the motifs and on the scoring method employed. Probably the simplest to understand is the technique of fingerprinting.[8] Here, groups of conserved motifs are excised from a sequence alignment and used to create a series of frequency (identity) matrices – no mutation or other similarity data are used to weight the results. The scoring scheme is thus based on the calculation of residue frequencies for each position in the motifs, summing the scores of identical residues for each position of a retrieved match. However, the simplicity of this approach is both its strength and its weakness. In other words, because the method exploits observed residue frequencies, the scoring matrices are sparse and thus perform cleanly (with little noise) and with high specificity; at the same time, their absolute scoring potential is limited by the nature of the observed data. For richly populated families, this is not a problem because the resulting matrices will reflect the constituent sequence diversity; but for poorly populated families, the matrices may be too sparse and may not encode sufficient variation to be able to detect distant relatives reliably, if at all.

One way to address this problem is to use mutation or substitution matrices to weight non-identical residue matches. Commonly used scoring matrices include the PAM[9] and BLOSUM[10] series. The former is based on the concept of the point-accepted mutation (PAM). PAM 250 is often used as a default matrix in comparison programs because it gives similarity scores equivalent to 20 per cent matches remaining between two sequences, the Twilight Zone[11] of similarity. The BLOSUM matrices, which are derived from observed substitutions in blocks of aligned sequences from the Blocks database, were designed to detect distant similarities more reliably than the Dayhoff series, which can only *infer*

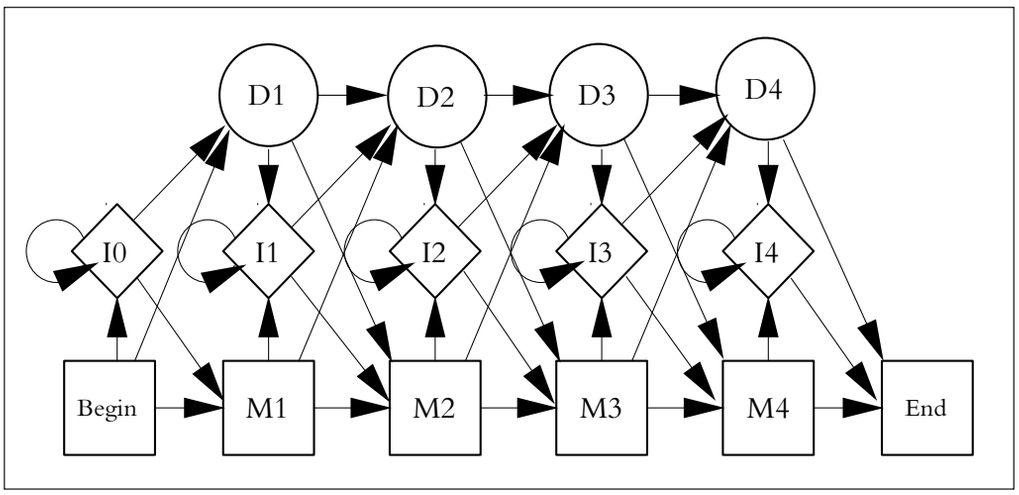

Figure 4: Linear hidden Markov model (HMM). Each position of an alignment is represented as a match (M), an insert (I), or a delete (D) state in the HMM. This allows a query sequence to be aligned by assigning the most probable state transition to each of its residues

Evolutionary distance; profile; weight matrices; hidden Markov models

remote relationships because their substitution rates were derived from sets of highly similar sequences. Whatever the approach, however, similarity matrices are inherently noisy because they indiscriminately weight both random matches and weak signals. Thus care should be taken to select a scoring matrix appropriate to the evolutionary distance at which relationships are being sought. For practical purposes, this means using a range of different matrices.

Profile methods

An alternative philosophy to motif-based approaches takes into account the variable regions between conserved motifs, which also contain valuable information. Here, the complete conserved portion of the alignment (including gaps) effectively becomes the discriminator. The discriminator, termed a profile, defines which residues are allowed at given positions, which positions are highly conserved and which degenerate, and which positions can tolerate insertions. The scoring system is intricate and may include evolutionary weights and results from structural studies, as well as data implicit in the alignment. In addition, variable penalties may be specified to weight against insertions and deletions occurring within core secondary structure elements.[12,13] Profiles (sometimes referred to as weight matrices) provide a sensitive means of detecting distant sequence relationships where only very few residues are well conserved.

Just as there are different ways of exploiting motifs, so there are different ways of using alignments to build family discriminators. An extension of the concept of profiles lies in the application of hidden Markov models (HMMs).[14] These are probabilistic models consisting of a number of interconnecting states – they are essentially linear chains of match, delete or insert states that attempt to encode the sequence conservation within aligned families. A match state is assigned to each conserved column in a sequence alignment, an insert state allows for insertions relative to the match states, and delete states allow positions to be skipped, as illustrated in Figure 4. Probabilities or costs (negative log-probabilities) are associated with each omission and each transition between states. To align a sequence is to find the highest-probability (lowest-cost) path through the HMM.

Although capable of providing precise

descriptors for particular families, as with all methods there are drawbacks. One problem arises from the specificity of profiles and HMMs. For example, they may be well trained for a given family, but an outlier that was not included in the training set may be missed if features of its sequence are incompatible with the model. Another problem may arise where the nature of HMM training is automatic and iterative. In such circumstances, without adequate supervision, the process may include false-positive matches, which may ultimately corrupt the model. Of course, this problem is by no means tied particularly to HMM training, as any automatic iterative method risks inclusion of false members.

Training problems

PATTERN DATABASES

The different methods of analysing sequences and encoding protein families have given rise to different pattern databases, as shown in Table 3. Despite their differences, pattern databases have arisen from the same principle, ie homologous sequences share conserved motifs, presumably crucial to the structure or function of the protein, which provide a signature of family membership. A new sequence that matches these predefined characteristics may then be assigned to a family. If the structure and function of the family are known, searches of pattern databases thus theoretically offer a fast track to the inference of biological function. Because these resources are derived from multiple sequence information, searches of them are often better able to identify distant relationships than are searches of the sequence databases. However, none of the pattern databases is yet complete. They should therefore be used to augment sequence searches, rather than to replace them. The status of some of the commonly used pattern resources is outlined below.

Pattern databases

Homologous sequences share conserved motifs

Fast track to inference of function

PROSITE

PROSITE stores motifs in the form of regular expression patterns,[15] which are derived from searches of SWISS-PROT. Entries are deposited in the database in two distinct files: (i) a structured data file that houses the pattern and lists all matches in the parent version of SWISS-PROT, as shown in Figure 5; and (ii) a free-format documentation or annotation file, which provides details of the characterised family and, where known, a description of the biological role of the chosen motif/s and a supporting bibliography. A number of features of the data file are worthy of note. The identifier (ID) and description (DE) lines identify the characterised family, and the DR lines list all true (T), possible (P), false (F) and missed/negative (N) matches to the pattern – these results are summarised in the NR lines. In the example shown in Figure 5, 29 matches are made to the

Table 3: Some of the major pattern databases in common use. In each case, the primary source is noted, together with the type of pattern stored (regular expression, fingerprint, HMM, etc.)

Pattern database	Data source	Stored information
PROSITE	SWISS-PROT	Regular expressions (patterns)
PRINTS	SWISS-PROT/TrEMBL	Raw aligned motifs (fingerprints)
Profiles	SWISS-PROT	Gapped weight matrices (profiles)
Pfam	SWISS-PROT/TrEMBL	Gapped domain alignments (HMMs)
Blocks	PROSITE/PRINTS	Weighted aligned motifs (blocks)
IDENTIFY	Block/PRINTS	Permissive regular expressions (patterns)

pattern, all of which are true, and there are two false negative matches. If we inspect one of these (LOX4_SOYBN) by retrieving its sequence from SWISS-PROT, we find that a disallowed serine in the ninth position of the motif (HQIISHWL**S**THAIVE) is the reason for the mis-match – referring back to the pattern (PA) line, we see that only members of the group [NQRC] are allowed at this point.

The quality of a pattern can thus immediately be ascertained from the NR lines, which are therefore probably the most important lines to inspect when first viewing a PROSITE entry. In some cases, there are numerous false positives and false negatives (especially for large super-families with substantial numbers of divergent sequences, such as G-protein-coupled receptors, lipocalins, etc.). Such patterns are diagnostically unreliable and are a limitation to the diagnostic potential of the database. PROSITE release 16 (September 1999), with updates to 18th October 1999, contains 1,035 entries characterised by 1,375 patterns.

More important is the number of blocks

Blocks

Blocks, which is based on families already identified in PROSITE, stores motifs as aligned, clustered blocks, which are derived from searches of SWISS-PROT.[16] The format of entries is PROSITE-compatible, but details of matches to a given block are not provided. Nevertheless, the diagnostic power of a block is given in terms of a strength value (which is reported on the BL lines). The strength is a normalised quantity, allowing the diagnostic performance of individual blocks to be compared. Strong blocks are more effective than weak blocks (strength less than 1100) at separating true-positives from true-negatives. In searching the database, however, more important than the strength of individual blocks is the *number* of blocks matched. High-scoring matches to single blocks seldom have biological

```
ID  LIPOXYGENASE_1; PATTERN.
AC  PS00711;
DT  DEC-1992 (CREATED); NOV-1997 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE  Lipoxygenases iron-binding region signature 1.
PA  H-[EQ]-x(3)-H-x-[LM]-[NQRC]-[GST]-H-[LIVMSTAC](3)-E.
NR  /RELEASE=36,74019;
NR  /TOTAL=29(29); /POSITIVE=29(29); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR  /FALSE_NEG=2; /PARTIAL=0;
CC  /TAXO-RANGE=??E??; /MAX-REPEAT=1;
CC  /SITE=4,iron; /SITE=9,iron;
DR  Q06327, LOX1_ARATH, T; P29114, LOX1_HORVU, T; P16050, LOX1_HUMAN, T;
DR  P38414, LOX1_LENCU, T; P12530, LOX1_RABIT, T; P37831, LOX1_SOLTU, T;
DR  P08170, LOX1_SOYBN, T; P27479, LOX2_BOVIN, T; P18054, LOX2_HUMAN, T;
DR  P29250, LOX2_ORYSA, T; P14856, LOX2_PEA , T; P16469, LOX2_PIG , T;
DR  Q02759, LOX2_RAT , T; P09439, LOX2_SOYBN, T; P09918, LOX3_PEA , T;
DR  P09186, LOX3_SOYBN, T; P09917, LOX5_HUMAN, T; P51399, LOX5_MESAU, T;
DR  P48999, LOX5_MOUSE, T; P12527, LOX5_RAT , T; P38415, LOXA_LYCES, T;
DR  P27480, LOXA_PHAVU, T; P38416, LOXB_LYCES, T; P27481, LOXB_PHAVU, T;
DR  P38418, LOXC_ARATH, T; P38419, LOXC_ORYSA, T; P55249, LOXE_MOUSE, T;
DR  P39654, LOXL_MOUSE, T; P24095, LOXX_SOYBN, T;
DR  P38417, LOX4_SOYBN, N; P39655, LOXP_MOUSE, N;
3D  1YGE; 2SBL;
DO  PDOC00077;
//
```

Figure 5: Example PROSITE entry, showing the data file for the lipoxygenase pattern. When viewing PROSITE on the Web, accession numbers are hyperlinked, allowing direct access to the corresponding SWISS-PROT entry for each sequence matched. Similarly, the documentation file for a given pattern can be accessed via the hyperlinked PDOC accession number at the bottom of the file

Full potency derives from context of motif neighbours

significance; conversely, matches to sets of blocks from the same family are unlikely to have arisen by chance, and a probability value is calculated to reflect that likelihood. Release 11.0 contains 4,034 blocks, representing 994 groups from PROSITE release 15.

Recently, a number of other Blocks databases have been made available. For example, in Blocks+, supplementing the entries derived from PROSITE families are blocks from families in PRINTS that are not already in Blocks, and then successively, any additional blocks from Pfam, ProDom and DOMO. Blocks+ is thus comprehensive, containing 10,070 blocks from 2,235 sequence groups. Complementing this resource is a version of PRINTS in which block-scoring methods have been exploited.[16] Note, however, that PRINTS' motifs tend to be deeper than those in Blocks because its source database is larger; the diagnostic performance of entries in the two resources can therefore differ, Blocks-format-PRINTS tending to be more prone to problems of noise. Because the Blocks databases are derived automatically, their entries are not annotated, but links are made to the corresponding PROSITE and PRINTS documentation files.

PRINTS

PRINTS stores motifs in the form of fingerprints,[17] which are derived from searches of a non-redundant SWISS-PROT/TrEMBL composite. Each entry is manually annotated with descriptions of the family, details of the structural or functional relevance of the motifs where known, cross-references to related databases, bibliographic references, etc. In addition, a full list of matching sequences is provided, including those that fail to match one or more motifs. Fingerprint diagnostic performance is indicated via a summary that tabulates the numbers of complete and partial matches. The fewer the partial matches, the better the fingerprint. The full potency of the method derives from the mutual context provided by motif neighbours. The more motifs in a fingerprint, the better able it is to identify distant relatives, even when parts of the signature are absent; conversely, the fewer the motifs, the poorer the diagnostic performance. Fingerprints with only two motifs are diagnostically little better than single motifs and are therefore more likely to make false-positive matches. When searching PRINTS, probability- and expect-values are calculated to assign a measure of confidence to both complete and partial matches. Release 24.0 (September 1999) contains 1,210 entries (7,241 motifs), making it currently the most comprehensive manually annotated pattern database.

Fingerprint diagnostic performance; bio-chemical properties

IDENTIFY

IDENTIFY stores motifs in the form of regular expressions and is derived automatically from motifs in Blocks and PRINTS.[7] The program used to create the database constructs consensus expressions from the motifs, adopting a permissive approach in which different residues are tolerated according to a set of prescribed groupings (Table 2). These groups correspond to various biochemical properties, theoretically ensuring that the resulting expressions have sensible biochemical interpretations. However, because in practice the approach may lead to an increase in noise, when searching the resource, different levels of stringency are offered from which to infer the significance of matches. The approach is thus diagnostically more powerful than exact pattern matching.

Profiles

Profiles stores the conservation encoded in gapped domain alignments in the form of weight matrices,[15] which are derived from searches of SWISS-PROT. As a result of their potency, profiles are used to complement some of the poorer regular expressions in PROSITE, or to

Separate documentation

provide a diagnostic alternative where extreme sequence divergence renders the use of regular expressions inappropriate. Each profile has separate PROSITE-compatible data and documentation files. This allows results that have been validated and annotated to an appropriate standard to be made available as an integral part of PROSITE. As before, diagnostic performance can be ascertained from the DR and NR lines. Profiles are less prone to make false matches than are regular expressions, but the numbers released via PROSITE are only small (48 in September 1998). A further 301 profiles that have not yet achieved the necessary standard of validation and annotation are available in a pre-release distribution.

Pfam

Pfam stores the conservation encoded in gapped domain alignments in the form of HMMs,[19] which are derived from searches of SWISS-PROT and TrEMBL. The resource is based on two distinct classes of alignment: hand-edited seed alignments, which are deemed to be accurate; and an automatically clustered set derived from ProDom families. The seed alignments are used to build HMMs, to which sequences are automatically aligned to generate final full alignments. The collection of seed and full alignments, coupled with minimal annotations (often no more than a description line), related database and literature cross-references, and the HMMs themselves, constitute Pfam-A. Pfam-B is then generated automatically by filtering out from existing ProDom clusters all sequences that have been matched in Pfam-A. Although the methods and parameters used to create the full automatic alignment are noted (including Noise- and Trusted-Cutoff values, which indicate the size and location of the score gap between true and false members), no indication is given of the diagnostic performance of a given HMM in terms of the numbers of true and false matches made and the number of true matches missed. Direct visualisation of the final alignment is thus probably the best indicator of how sound its HMM is likely to be. Pfam release 4.3 (September 1999) encodes 1,815 domains.

HMMs; trusted-cutoff values; family-cluster alignment databases

FAMILY-CLUSTER DATABASES

In addition to the resources discussed above, several 'family-cluster' alignment databases are available for searching via the Web, some of which are listed in Table 4. The construction of alignment and pattern databases is based on different principles, so the two types of resource should not be confused. The

Table 4: Some of the major alignment databases. In each case, the primary source is noted, together with the level of information stored (ie whether domain, family or superfamily alignments)

Alignment database	Primary source	Stored information
ProDom	SWISS-PROT	Domains
SBASE	SWISS-PROT	Domains
ProtoMap	SWISS-PROT	Families
PIR-ALN	PIR	Superfamilies, families and domains
PROT-FAM	PIR	Superfamilies, families and domains
ProClass	SWISS-PROT/PIR	Superfamilies, families and domains
DOMO	SWISS-PROT/PIR	Domains and repeats
PIMA	Entrez	Domains

main difference between them is that alignment databases tend to be derived simply by automatic clustering of sequence databases. This allows them to be more comprehensive than pattern resources because they do not depend on manual crafting and validation of family discriminators. However, searches of alignment databases are often less sensitive because they tend to be based on implementations of BLAST. Typical resources include ProDom,[20] SBASE,[21] ProtoMap,[22] PIR-ALN,[23] PROT-FAM,[24] ProClass,[25] DOMO[26] and PIMA.[27]

Infer homology; validation of results

WHICH DATABASE IS BEST?

Bewildering choices

The plethora of available databases presents bewildering choices to the sequence analyst. Which is diagnostically most reliable? Which has the most useful annotations? Which is the most comprehensive? Which should I use? It is difficult to assess the quality of particular resources: each has different diagnostic strengths and weaknesses, each offers different family coverage and different levels of annotation – each has good points and bad. Nevertheless, some general points bear consideration.

Different strengths and weaknesses

Initially, the clustered family resources appeal because they are so comprehensive, yet they suffer certain limitations. Automatic clustering is based on pre-set scores and the resulting clusters need not have precise biological correlations. Furthermore, the search tools tend to involve flavours of BLAST or FASTA, which are good at highlighting generic similarities but cannot pinpoint differences (eg such as between highly similar but functionally disparate receptor subtypes).

Structural and functional contexts

Perhaps the biggest failing of automatically generated pattern and cluster databases is that they carry no annotations. The advantage of searching them is that they are more comprehensive than their manually derived counterparts. The disadvantage is that there may be no way to ascertain the biological significance of a match, if indeed it has any (that a match has been made does not mean an evolutionary relationship necessarily exists). This is important to understand – automatic methods can only *detect similarities*, but it is for the user to *infer homology* from supporting biological evidence.

Among pattern databases, single-motif methods that use exact regular expression pattern-matching have known diagnostic limitations. These methods tolerate no similarity, so will fail to diagnose sequences that contain subtle changes not catered for by the pattern. Moreover, single motifs offer no biological context within which to assess the significance of a match – each has therefore to be verified individually. Multiple motif approaches inherently offer improved diagnostic reliability by virtue of the mutual context provided by motif neighbours. Thus, if a query fails to match all the motifs in a signature, the pattern of matches formed by the remaining motifs still allows the user to make a confident diagnosis.

Pattern resources derived from existing databases have the limitation that they offer no further family coverage. Nevertheless, they have the advantage of implementing different analytical methods from their source databases, thus offering different scoring potentials on the same data, and furnishing important opportunities to diagnose relationships missed by the original implementations.

Finally, manually annotated databases are set apart from their automatically created counterparts by virtue of (i) attempting to provide *validation* of results, and (ii) offering detailed information that helps to place conserved sequence information in structural or functional contexts. This is vital for the user, who not only wants to discover whether a sequence has matched a pre-defined motif, but also needs to understand its biological significance.

A COMPOSITE PATTERN DATABASE

Error catastrophe

Today, comprehensive sequence analysis requires accessing a variety of disparate databases, gathering the range of different outputs and arriving at some sort of consensus view of the results. In the future, however, this process should become more straightforward. The curators of PROSITE, PRINTS, Pfam, Profiles and ProDom are creating a unified database

Unified family database

of protein families, termed InterPro. The aim is to provide a single family annotation resource, based on existing documentations from PRINTS and

Functional clues

PROSITE, and on the minimal annotations in Pfam, with each InterPro document linking back to the relevant entries in the satellite pattern databases. This will simplify sequence analysis for the user, who will thereby have access to a central resource for protein family diagnosis.

CONCLUSION

Creating and searching pattern databases are activities that lie at

Fallible chain of events

different ends of a fallible chain of events. We begin with a sequence alignment, we create some kind of scoring function to encode the conservation within the alignment (a scoring matrix, HMM, etc.), we store the discriminators in a database, and we

Informed assessment of function

search them with different algorithms. Problems arise if unrelated sequences have crept into the alignment, which in turn lead to errors in the discriminators, which then give ambiguous or incorrect search results. Alternatively, the discriminators may be sound, but the search algorithms may not be sufficiently sensitive to allow unequivocal diagnosis, leading the user to false conclusions of family ties. If the user has performed this experiment on a newly determined sequence and submits the results to one of the

False conclusions

sequence databases, the annotation error becomes available for mass propagation.

Recently, there has been doom-mongering in the literature about the quality of our databases, some harbingers of misfortune predicting a future error catastrophe. At the same time, claims of success for some approaches to family classification and function prediction have been equally overdone. A more balanced view recognises that our databases and search routines are not perfect, but with the right approach we can avoid the pitfalls of jumping to over-pessimistic or over-zealous conclusions.

Until we have sufficient experimental data available, pattern and sequence databases are probably the best tools we have for accessing the functional and evolutionary clues latent in the sequences flooding from the genome projects. Pattern databases offer several benefits: (i) by distilling multiple sequence information into family or domain descriptors, trivial errors in the underlying sequences may be diluted; (ii) annotation errors may be quickly spotted if the description of one sequence differs from that of its family; and (iii) they allow specific diagnoses, placing individual sequences in domain or family contexts for a more informed assessment of possible function. By contrast, searches of sequence databases tend to reveal only *generic* similarities, making precise pinpointing of a particular biological niche more difficult.

While there is some overlap between them, the contents of the pattern databases differ. Together they encode about 2,200 families, including globular and membrane proteins, modular polypeptides, and so on. It has been estimated that the total number of families might be in the range 1,000 to 10,000, so there is a long way to go before any of the databases can be considered complete. Thus, in building a search strategy, it is good practice to include all available pattern resources, to ensure that the analysis is as comprehensive as possible and that it takes advantage of a

variety of search methods. Where there is consensus, diagnoses can be made with greater confidence.

Unfortunately, creating and annotating family descriptors is time-consuming, so pattern databases have not kept pace with the deluge of sequence data and are consequently still very small. Nevertheless, as they become more comprehensive, as the volume of sequence data expands and search outputs become more complex, their diagnostic potency ensures that pattern databases will play an increasingly important role as the post-genome quest to assign functional information to raw sequence data gains pace.

Pattern databases play an important role

Acknowledgements
The author is grateful to the Royal Society for a University Research Fellowship.

References

1. Doerks, T., Bairoch, A. and Bork, P. (1998), 'Protein annotation: detective work for function prediction', *Trends Genetics,* Vol. 14, pp. 248–250.
2. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25(17), pp. 3389–3402.
3. Pearson, W. R. (1998), 'Empirical statistical estimates for sequence similarity searches', *J. Mol. Biol.*, Vol. 276(1), pp, 71–84.
4. Hofmann, K. (1998), 'Protein classification and functional assignment', in 'Trends Guide to Bioinformatics', Elsevier Science, New York, pp. 18–21.
5. Attwood, T. K. (1997), 'Exploring the language of bioinformatics', in 'Oxford Dictionary of Biochemistry and Molecular Biology', Stanburg, H., Ed., Oxford University Press, Oxford, pp. 715–723.
6. Attwood, T. K. and Parry-Smith, D. J. (1999), 'Introduction to Bioinformatics', Addison Wesley Longman, Harlow.
7. Nevill-Manning, C. G., Wu, T. D. and Brutlag, D. L. (1998), 'Highly specific protein sequence motifs for genome analysis', *Proc. Natl Acad. Sci., USA*, Vol. 95, pp. 5865–5871.
8. Parry-Smith, D. J. and Attwood, T. K. (1992), 'ADSP – a new package for computational sequence analysis', *Comput. Appl. Biosci.*, Vol. 8(5), pp. 451–459.
9. Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978), 'A model of evolutionary change in proteins', in 'Atlas of Protein Sequence and Structure', Vol. 5, Suppl. 3, Dayhoff, M. O., Ed. National Biomedical Research Foundation, Washington, DC, pp. 345–352.
10. Henikoff, J. G. and Henikoff, S. (1992), 'Amino acid substitution matrices from protein blocks', *Proc. Natl Acad. Sci., USA*, Vol. 89, pp. 10915–10919.
11. Doolittle, R. F. (1986), 'Of URFs and ORFs: A Primer On How to Analyse Derived Amino Acid Sequences', University Science Books, Mill Valley, CA.
12. Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987), 'Profile analysis: Detection of distantly related proteins', *Proc. Natl Acad. Sci., USA,* Vol. 84(13), pp. 4355–4358.
13. Luthy, R., Xenarios, I. and Bucher, P. (1994), 'Improving the sensitivity of the sequence profile method', *Protein Sci.*, Vol. 3(1), pp. 139–146.
14. Hughey, R. and Krogh, A. (1996), 'Hidden Markov models for sequence analysis: extension and analysis of the basic method,' *Comput. Applic. Biosci.*, Vol. 12(2), pp. 95–107.
15. Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999), 'The PROSITE database, its status in 1999', *Nucleic Acids Res.*, Vol. 27(1), pp. 215–219.
16. Bairoch, A. and Apweiler, R. (1999), 'The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999', *Nucleic Acids Res.*, Vol. 27(1), pp. 49–54.
17. Henikoff, J. G., Henikoff, S. and Pietrokovski, S. (1999), 'New features of the Blocks Database servers', *Nucleic Acids Res.*, Vol. 27(1), pp. 226–228.
18. Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J. and Wright, W. (1999), 'PRINTS prepares for the new millennium', *Nucleic Acids Res.*, Vol. 27(1), pp. 220–225.
19. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. and Sonhammer, E. L. L. (1999), 'Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins', *Nucleic Acids Res.*, Vol. 27(1), pp. 260–262.
20. Gouzy, J., Corpet, F. and Kahn, D. (1999), 'Recent improvements of the ProDom database of protein domain families', *Nucleic Acids Res.*, Vol. 27(1), pp. 263–267.
21. Murvai, J., Kristian Vlahovicek, K., Barta, E., Szepesvári, C., Acatrinei, C. and Pongor, S. (1999), 'The SBASE protein domain library, release 6.0: A collection of annotated protein sequence segments', *Nucleic Acids Res.*, Vol. 27(1), pp. 257–259.

22. Yona, G., Linial, N., Tishby, N. and Linial, M. (1998), 'A map of the protein space – an automatic hierarchical classification of all protein sequences', in 'Proceedings of 6th International Conference on ISMB', AAAI Press, Menlo Park, CA, pp. 212–221.

23. Srinivasarao, G.Y., Yeh, L.-S. L., Marzec, C. R., Orcutt, B. C., Barker, W. C. and Pfeiffer, F. (1999), 'Database of protein sequence alignments: PIR-ALN', *Nucleic Acids Res.*, Vol. 27(1), pp. 284–285.

24. Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker S. and Frishman, D. (1999), 'MIPS: a database for genomes and protein sequences', *Nucleic Acids Res.*, Vol. 27(1), pp. 44–48.

25. Wu, C. H., Shivakumar, S. and Huang, H. (1999), 'ProClass protein family database', *Nucleic Acids Res.*, Vol. 27(1), pp. 272–274.

26. Gracy, J. and Argos, P. (1998), 'DOMO: a new database of aligned protein domains', *Trends Biochem. Sci.*, Vol. 23(12), pp. 495–497.

27. Smith, R. F. and Smith, T. F. (1992), 'Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling', *Protein Eng.,* Vol. 5(1), pp. 35–41.