doi:10.1006/jmbi.2000.4036 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 303, 61–76

JMB

Analysis and Prediction of Functional Sub-types from Protein Sequence Alignments

Sridhar S. Hannenhalli[1] and Robert B. Russell[2]*

[1]*Bioinformatics Research Group, SmithKline Beecham Pharmaceuticals Research & Development, 709 Swedeland Road, King of Prussia PA 19406, USA*

[2]*Bioinformatics Research Group, SmithKline Beecham Pharmaceuticals Research & Development, New Frontiers Science Park (North), Third Avenue, Harlow, Essex CM19 5AW, UK*

**Corresponding author*

The increasing number and diversity of protein sequence families requires new methods to define and predict details regarding function. Here, we present a method for analysis and prediction of functional sub-types from multiple protein sequence alignments. Given an alignment and set of proteins grouped into sub-types according to some definition of function, such as enzymatic specificity, the method identifies positions that are indicative of functional differences by comparison of sub-type specific sequence profiles, and analysis of positional entropy in the alignment. Alignment positions with significantly high positional relative entropy correlate with those known to be involved in defining sub-types for nucleotidyl cyclases, protein kinases, lactate/malate dehydrogenases and trypsin-like serine proteases. We highlight new positions for these proteins that suggest additional experiments to elucidate the basis of specificity. The method is also able to predict sub-type for unclassified sequences. We assess several variations on a prediction method, and compare them to simple sequence comparisons. For assessment, we remove close homologues to the sequence for which a prediction is to be made (by a sequence identity above a threshold). This simulates situations where a protein is known to belong to a protein family, but is not a close relative of another protein of known sub-type. Considering the four families above, and a sequence identity threshold of 30 %, our best method gives an accuracy of 96 % compared to 80 % obtained for sequence similarity and 74 % for BLAST. We describe the derivation of a set of sub-type groupings derived from an automated parsing of alignments from PFAM and the SWISSPROT database, and use this to perform a large-scale assessment. The best method gives an average accuracy of 94 % compared to 68 % for sequence similarity and 79 % for BLAST. We discuss implications for experimental design, genome annotation and the prediction of protein function and protein intra-residue distances.

Keywords: protein function; protein structure; prediction; sequence alignment

Introduction

Multiple sequence alignments are central to protein classification and analysis. When protein sequences are aligned, it becomes possible to see sequence conservation patterns that are indicative of, for example, enzyme active sites and secondary structure types (e.g. Zvelebil *et al.*, 1987; Casari *et al.*, 1995; Lichtarge *et al.*, 1996a). With such patterns, it is possible to derive motifs that encapsulate the features defining the protein family. Moreover, the aligned sequences can be used to construct sensitive profiles (e.g. Gribskov *et al.*, 1987; Birney *et al.*, 1996), or hidden Markov models

Present address: R. B. Russell; EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany.

Abbreviations used: URL, Universal Resource Locator; SCOP, Structural Classification Of Proteins; PFAM, Protein FAMilies; SMART, Simple Modular Architecture Resource Tool; NCBI, National Center for Biotechnology Information; SB, SmithKline Beecham; HMM, Hidden Markov Model; BLAST, basic local alignment search tool; HSP, high-scoring segment pair; RGS, regulator of G-protein signalling; SH2/SH3, *src* homology 2/3. The standard one and three-letter abbreviations for the amino acid residues are also used throughout..

E-mail address of the corresponding author: russelr1@mh.uk.sbphrd.com

0022-2836/00/010061–16 $35.00/0

(HMMs; e.g. Eddy, 1998; Krogh *et al*., 1994) that can be used to detect further, more remote members of a protein family. These techniques and others have aided greatly the detection of protein families and the associated construction of protein alignment databases, such as SMART (Schulz *et al*., 1998) and PFAM (Bateman *et al*., 1999), which are of growing importance in the analysis of data from large scale genome sequence projects.

However, the detection and alignment of sequences from diverse protein families creates new problems. Among these is the fact that homologous proteins frequently evolve different functions, which we hereafter refer to as a sub-type. It is common for proteins to evolve slightly different functions, such as different substrate specificities, or activities. In extreme cases, both enzymes and effector molecules (i.e. non-enzymes) can reside in the same homologous superfamily (e.g. Murzin, 1993), and ultimately proteins with similar folds can perform completely different functions (e.g. Russell *et al*., 1998). If a protein is of unknown function, but is found to belong to a diverse protein superfamily, or fold, with multiple functions, then determining functional sub-type becomes of great importance.

Often a perfect division of a protein family into sub-types can be accomplished by a simple phylogenetic analysis. In other words: sub-type correlates exactly, and it is clear that with the branches of a phylogenetic tree, therefore making the prediction of sub-type simply a matter of deciding into which branch a protein belongs. It is not surprising that most previous attempts to classify proteins have been very reliant on phylogenetic trees.

However, the division of proteins into functional sub-types cannot always be accomplished by phylogeny. If much time has passed since the evolution of different sub-types, then sequences may have diverged beyond the point where phylogeny can easily give a clear division. In addition, proteins usually have multiple features that co-evolve, such as differing affinities for more than one substrate, variations in sub-cellular location (e.g. membrane attached *versus* cytosolic) or the interaction with other proteins that differ across paralogues, even if other details, such as catalytic mechanism, remain unchanged. Finally, there remains the possibility that details of molecular function may evolve convergently (e.g. Makarova & Grishin, 1999). This is particularly likely in instances where specificity is conferred by only a handful of residues, or even a single position (e.g. Wu *et al*., 1999).

Various methods have been developed previously that attempt to address the problem of the analysis and prediction of protein sub-types from protein sequence alignments. Livingstone & Barton (1993) developed a method to annotate protein sequence alignments with the aim of highlighting positions of residue conservation. They made use of amino acid properties similar to those of Taylor (1986) and ‘‘sensible groups’’ provided from sequence similarity, functional or evolutionary criteria to highlight positions in the alignment conferring the unique features of a sub-group. The method was demonstrated graphically by analysis of SH2 and annexin domains, but to our knowledge, it has not been applied to the problem of predicting sub-types.

Casari *et al*. (1995) used a principle component analysis of a vector representation of sequences in space to develop an elegant method for identifying functional residues on proteins based on a multiple sequence alignment. Analysis of various dimensions in the vector sequence space gave both positions that are conserved across the whole protein family, in addition to residues specific to sub-types, either specified in advance, or determined from analysis of the sequence space itself. The method was successful at identifying positions determining specificity in the Ras-Rab-Rho superfamily, SH2 domains and cyclins. Subsequent studies have applied this method to alcohol dehydrogenases (Atrian *et al*., 1998), the ran-RCC1. interaction (Azuma *et al*., 1999), effector recognition by GTP-binding proteins (Bauer *et al*., 1999), and other families.

Lichtarge *et al*. (1996a) developed the ‘‘Evolutionary Trace’’ method, to determine important positions on protein sequences and structures that were of functional importance. Their method combined knowledge of protein structures with an evolutionary history derived from a phylogenetic tree to extract functionally important residues to identifying functional interfaces on protein surfaces. They made a distinction between positions conserved across all sequences, and those that vary only between subgroups (class-specific). In this way they were able to identify positions on protein structures that were important, both for features of the family as a whole, as well as for particular sub-types. The method has been applied to several protein families, including SH2, SH3, nuclear hormone receptors (Lichtarge *et al*., 1996a), G-proteins/G-protein coupled receptors (Lichtarge *et al*., 1996b), zinc binding domains (Lichtarge *et al*., 1997) and the RGS/G-protein interaction (Sowa *et al*., 2000).

Sjolander (1998) developed a method of Phylogenetic Inference specifically designed for protein super-family analysis. Here, a phylogenetic tree is built for the input sequences based on nearest neighbour heuristics. The nodes in the tree are represented by a sequence profile of the sequences under that node, and the distance between two nodes is computed in terms of symmetric relative entropy, together with Dirichlet mixture priors. The method ensures that the highly conserved sites have higher weights while computing distances between nodes. The method was applied to SH2-domain containing proteins, resulting in new sub-family assignments for two proteins.

Here, we present another approach for studying protein sub-types associated with sequence alignments. Rather than attempt to define sub-types, we focused on the problems of identifying regions that confer specificity of sub-types already known (e.g. from experimental studies), and of predicting sub-types for "orphan" sequences (i.e. those where no sub-type is known).

Given a multiple sequence alignment and a classification of different sub-types (e.g. differences in enzyme specificity), the method exploits the differences between hidden Markov model profiles to highlight positions on the sequences that are most discerning of each sub-type. The method permits conservative substitutions, and tolerates missing data by combining alignments with amino acid exchange matrices *via* the construction of an HMM (Eddy, 1998). For new sequences known to be homologous to an existing family, but of unknown sub-type, the method can exploit the known sub-type classifications and associated profiles to predict sub-type. We demonstrate the method first by application to four well characterised protein families. We then perform a large scale assessment of sub-type prediction by applying the method to automatically derived sub-type groupings for 42 alignments from PFAM (Bateman *et al.*, 1999). We discuss implications for experimental design, prediction of protein function, prediction of inter- and intra-protein distances, and applications to genome annotation.

Algorithm

Assessing the discerning value of amino acid positions

This procedure locates positions in a protein alignment that are best able to discriminate between different sub-types. Essentially this involves finding positions that are conserved within each sub-type, but that vary between the different sub-types.

Given an alignment A of sequences in family F, and the sub-types $S_1, S_2, \cdots, S_k$ of the sequences, we extract the sub-alignment A^j from A, corresponding to the sequences of sub-type S_j. We use the *hmmbuild* program of the HMMER 2.1.1 (http://hmmer.wustl.edu) to build profile P^j of the alignment A^j. In profile building, the issues of small sample sizes and bias in the sample are important. By default HMMER uses Dirichlet priors and G/S/C sequence weighting scheme to address these issues. We refer the reader to Durbin *et al.* (1998) for the details of these methods.

We represent the profile of A^j at position i of the alignment by P_i^j, and the profile value for amino acid x at position i of the alignment by $P_{i,x}^j$. We convert the score profile in the hmmbuild output into a probability profile such that:

$$\sum_{\text{for all } x} P_{i,x}^j = 1$$

for each alignment position i. For a sub-type s, we use $\bar{s}$ to denote the union of all the sub-types excluding s. To estimate the role of an alignment position (or, site) i in determining the sub-type s, we compute the relative entropy (Shannon & Weaver, 1963; Durbin *et al.*, 1998) of the position i for sub-type s with respect to the entropy of that position for the sub-type $\bar{s}$. Let RE_i^s be relative entropy of P_i^s with respect to $P_i^{\bar{s}}$:

$$RE_i^s = \sum_{\text{for all } x} P_{i,x}^s \log \frac{P_{i,x}^s}{P_{i,x}^{\bar{s}}}$$

Notice that RE is greater than or equal to zero and is exactly zero when the two distributions are identical (Durbin *et al.*, 1998). To estimate the role of an alignment position i in determining the sub-types, we define cumulative relative entropy CRE_i as:

$$CRE_i = \sum_{\text{for all sub-types s}} RE_i^s$$

The cumulative relative entropies for all the positions are converted into Z-scores based on the distribution of entropies for an alignment. Let μ and σ be the mean and the standard deviation of cumulative relative entropies of all positions, then the Z-score for position i is computed as:

$$Z_i = \frac{CRE_i - \mu}{\sigma}$$

We expect a position with high Z-score to be important in determining the sub-types. Inspection of alignments together with knowledge of residues determining specificity *via* experiment suggested that Z-scores > 3.0 correlated well with preconceptions of positions known to determine specificity. We use this value in the discussion of individual families below.

Once we identify the important sites, we identify the residues responsible for the low entropies at those sites. Given a position i, we compute the ratio of the probability of observing amino acid x in sub-type s to that for sub-type $\bar{s}$. Let $PR_{i,x}^s$ be value of this ratio for amino acid x in the profile of $P^{\bar{s}}$ with respect to $P_i^{\bar{s}}$:

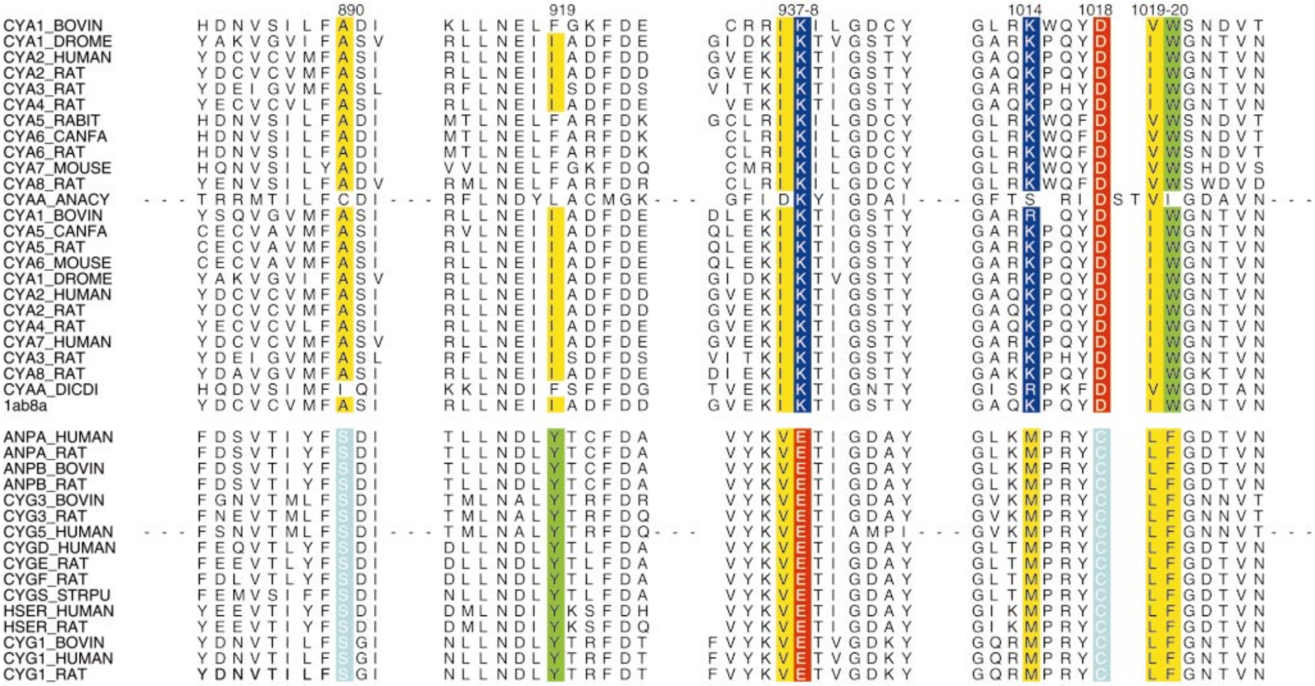
$$PR_{i,x}^s = \frac{P_{i,x}^s / P_{i,x}^{\bar{s}}}{\sum_{\text{for all y}} P_{i,y}^s / P_{i,y}^{\bar{s}}}$$

Inspection shows that single amino acid residues (or groups with similar properties) having $PR_{i,x}^s \geqslant 0.5$ agreed with our prior knowledge of determinants of sub-type specificity. We thus use this value when highlighting amino acid residues in Figures 1 through 6.

Predicting protein sub-types

Sequence similarity method

If a particular sequence X of unknown sub-type has a high sequence similarity to a sequence with known sub-type, then X can often be assigned



```
CYA1_BOVIN        HDNVSILFADI    KLLNELFGKFDE       CRRIKILGDCY    GLRKWQYD    VWSNDVT
CYA1_DROME        YAKVGVIFASV    RLLNEIIADFDE      GI DKIKTVGSTY   GARKPQYD    IWGNTVN
CYA2_HUMAN        YDCVCVMFASI    RLLNEIIADFDD      GVEKIKTIGSTY    GAQKPQYD    IWGNTVN
CYA2_RAT          YDCVCVMFASI    RLLNEIIADFDD      GVEKIKTIGSTY    GAQKPQYD    IWGNTVN
CYA3_RAT          YDEIGVMFASL    RFLNEIISDFDS      VI TKIKTIGSTY   GARKPHYD    IWGNTVN
CYA4_RAT          YECVCVLFASI    RLLNEIIADFDE        VEKIKTIGSTY   GAQKPQYD    IWGNTVN
CYA5_RABIT        HDNVSILFADI    MTLNELFARFDK      GCLRIKILGDCY    GLRKWQFD    VWSNDVT
CYA6_CANFA        HDNVSILFADI    MTLNELFARFDK       CLRIKILGDCY    GLRKWQFD    VWSNDVT
CYA6_RAT          HDNVSILFADI    MTLNELFARFDK       CLRIKILGDCY    GLRKWQFD    VWSNDVT
CYA7_MOUSE        HQNVSILYADI    VVLNELFGKFDQ       CMRIKILGDCY    GLRKWQYD    VWSHDVS
CYA8_RAT          YENVSILFADV    RMLNELFARFDR       CLRIKILGDCY    GLRKWQFD    VWSWDVD
CYAA_ANACY   ---  TRRMTILFCDI --- RFLNDYLACMGK ---  GFI DKYIGDAI --- GFTS  RIDSTVIGDAVN ---
CYA1_BOVIN        YSQVGVMFASI    RLLNEIIADFDE      DLEKIKTIGSTY    GARR QYD    IWGNTVN
CYA5_CANFA        CECVAVMFASI    RVLNEIIADFDE      QLEKIKTIGSTY    GARKPQYD    IWGNTVN
CYA5_RAT          CECVAVMFASI    RLLNEIIADFDE      QLEKIKTIGSTY    GARKPQYD    IWGNTVN
CYA6_MOUSE        CECVAVMFASI    RLLNEIIADFDE      QLEKIKTIGSTY    GARKPQYD    IWGNTVN
CYA1_DROME        YAKVGVIFASV    RLLNEIIADFDE      GI DKIKTVGSTY   GARKPQYD    IWGNTVN
CYA2_HUMAN        YDCVCVMFASI    RLLNEIIADFDD      GVEKIKTIGSTY    GAQKPQYD    IWGNTVN
CYA2_RAT          YDCVCVMFASI    RLLNEIIADFDD      GVEKIKTIGSTY    GAQKPQYD    IWGNTVN
CYA4_RAT          YECVCVLFASI    RLLNEIIADFDE      GVEKIKTIGSTY    GAQKPQYD    IWGNTVN
CYA7_HUMAN        YDCVCVMFASV    RLLNEIIADFDE      GVEKIKTIGSTY    GARKPQYD    IWGNTVN
CYA3_RAT          YDEIGVMFASL    RFLNEIISDFDS      VI TKIKTIGSTY   GARKPHYD    IWGNTVN
CYA8_RAT          YDAVGVMFASI    RLLNEIIADFDE      DI EKIKTIGSTY   GAKKPQYD    IWGKTVN
CYAA_DICDI        HQDVSIMFIQI    KKLNDIFSFFDG      TVEKIKTIGNTY    GI SRPKFD   VWGDTAN
1ab8a             YDCVCVMFASI    RLLNEIIADFDD      GVEKIKTIGSTY    GAQKPQYD    IWGNTVN

ANPA_HUMAN        FDSVTIYFSDI    TLLNDLYTCFDA       VYKVETIGDAY    GLKMPRYC    LFGDTVN
ANPA_RAT          FDSVTIYFSDI    TLLNDLYTCFDA       VYKVETIGDAY    GLKMPRYC    LFGDTVN
ANPB_BOVIN        FDSVTIYFSDI    TLLNDLYTCFDA       VYKVETIGDAY    GLKMPRYC    LFGDTVN
ANPB_RAT          FDSVTIYFSDI    TLLNDLYTCFDA       VYKVETIGDAY    GLKMPRYC    LFGDTVN
CYG3_BOVIN        FGNVTMLFSDI    TMLNALYTRFDR       VYKVETIGDAY    GVKMPRYC    LFGNNVT
CYG3_RAT          FNEVTMLFSDI    TMLNALYTRFDQ       VYKVETIGDAY    GVKMPRYC    LFGNNVT
CYG5_HUMAN   ---  FSNVTMLFSDI --- TMLNALYTRFDQ ---  VYKVETIAMPI --- GVKMPRYC    LFGNNVT ---
CYGD_HUMAN        FEQVTLYFSDI    DLLNDLYTLFDA       VYKVETIGDAY    GLTMPRYC    LFGDTVN
CYGE_RAT          FEEVTLYFSDI    DLLNDLYTLFDA       VYKVETIGDAY    GLTMPRYC    LFGDTVN
CYGF_RAT          FDLVTLYFSDI    DLLNDLYTLFDA       VYKVETIGDAY    GLTMPRYC    LFGDTVN
CYGS_STRPU        FEMVSIFFSDI    NLLNDLYTLFDA       VYKVETIGDAY    GLTMPRYC    LFGDTVN
HSER_HUMAN        YEEVTIYFSDI    DMLNDIYKSFDH       VYKVETIGDAY    GI KMPRYC   LFGDTVN
HSER_RAT          YEEVTIYFSDI    DMLNDIYKSFDQ       VYKVETIGDAY    GI KMPRYC   LFGDTVN
CYG1_BOVIN        YDNVTILFSGI    NLLNDLYTRFDT      FVYKVETVGDKY    GQRMPRYC    LFGNTVN
CYG1_HUMAN        YDNVTILFSGI    NLLNDLYTRFDT      FVYKVETVGDKY    GQRMPRYC    LFGNTVN
CYG1_RAT          YDNVTILFSGI    NLLNDLYTRFDT      FVYKVETVGDKY    GQRMPRYC    LFGNTVN
```

Figure 1. Alscript (Barton, 1993) Figure showing an alignment of representatives of nucleotidyl cyclases with positions predicted to confer specificity to adenylate or guanylate highlighted by the method. The alignment only shows regions that contain positions predicted to confer specificity, deleted regions are indicated by dashes (---). Positions are coloured only if $PR^{s}_{i,x} \geqslant 0.5$. Colours are according to the residue conservation: hydrophobics, yellow; small residues, light blue; positive residues, dark blue; negative residues, red; polar residues, magenta. Note that positions sharing the same colour across both groups may have subtle differences that are discussed in more detail in the text. Numbers above the alignment refer to positions discussed in the text, and correspond to the PDB structure 1ab8.

(accurately) to the same sub-type as the most similar sequence. For comparison to the methods described below, we devised a simple sequence similarity method on this principle. Given a sequence of unknown sub-type, we assign it the same sub-type as the sequence of known sub-type with the highest percent sequence identity, calculated by ignoring gaps, and by leaving the sequence aligned as they were in the original alignment.

BLAST method

Another means to assign sub-type *via* sequence similarity is to perform a database search. This would be a typical strategy adopted when given a new gene known to belong to a large homologous family, particularly in the absence of pre-computed multiple sequence alignments. To test this approach, we performed a BLAST search (Altschul *et al*., 1990) using the query sequence, *X*, and assigned it to the sub-type of the best HSP score.

HMM method

Since we use HMMER to compute the profiles for each sub-type, another approach is to use the search program *hmmsearch* to align the sequence of unknown sub-type to all HMMs and assign it to the sub-type yielding the maximum (*Viterbi*) alignment score.

Profile (HMM-derived) method

Frequently, the alignment resulting from *hmmsearch* is slightly different from the original alignment. For many protein families, hand-editing is performed to give alignments that are generally better than those generated automatically (this is true for both SMART and seed alignments within PFAM). Slight changes to the alignment introduced by an alignment algorithm (i.e. *hmmsearch*) might thus affect the prediction accuracy adversely. To avoid this potential problem we devised a profile method, where instead of aligning the sequence to HMM using *hmmsearch*, we assume the original alignment, and compute the score as described below.

We assume that a sequence with unknown sub-type is aligned to the other members of the family so that the length of the aligned sequence is the same as the length of the profile. For a sequence $X = x_1\ x_2\ \cdots\ x_n$, and profile P, the score of X with respect to P is computed as:

$$p(X|P) = \prod_{i=1}^{n} P_{i,x_i}$$

Given sub-types $S_1, S_2, \cdots, S_k$, with profiles $P_1, P_2, \cdots, P_k$, respectively, X is assigned to sub-type S_i that maximises $p(P^i \mid X)$, which is the same as maximising $p(X \mid P^i)$ using Bayes rule and assuming equal *a priori* probabilities of various sub-types.

Note we could potentially use the known sizes of sub-types to compute their *a priori* probabilities. However we feel that this would unfairly favour the profile method.

Sub-profile method

This method is a slight variant of the profile method. The difference is that only those positions in the alignment with a positive relative entropy Z-score are used when computing the score of a sequence against a profile. More exactly:

$$p(X|P) = \prod_{\mathrm{i}=1, Z_{\mathrm{i}}>0}^{\mathrm{n}} P_{\mathrm{i}, \mathrm{x}_{\mathrm{i}}}$$

Essentially, this removes contributions of the non-discriminating alignment positions to the score, thus filtering out noise.

The sequence similarity and BLAST methods attempt to simulate predictions of sub-type that might be made by a simple sequence database search. We acknowledge that it does not necessarily compare to a rigorous phylogenetic analysis (e.g. Sjolander, 1998).

Evaluation of sub-type prediction accuracy

As mentioned above, we believe that if a sequence X, with unknown sub-type, has a high degree of sequence similarity to a sequence with known sub-type, then X can be assigned the same sub-type with confidence. The methods proposed in this paper are aimed at predicting sub-types in the absence of very similar sequences of known sub-type. Therefore, before predicting sub-type for sequence X, we first eliminated all the sequences highly similar to X. We defined sequence similarity as percentage sequence identity (ignoring gaps) and we varied the threshold for sequences to ignore when making a prediction (see Results).

For the sequence similarity and BLAST methods, we assign sequence X to the sub-type of the most similar sequence in the reduced set (after removing the close homologues). For the HMM method, we construct HMMs using *hmmbuild* for the reduced set of each sub-type, align the sequence to each of the HMMs using *hmmsearch* and assign the sequence to the sub-type yielding the maximum score. The profile and the sub-profile methods are similar to the above, but instead of aligning the sequence against the HMMs using *hmmsearch*, we simply score the sequence against the profiles derived from the HMMs, as described before, leaving the original alignment of the sequence unperturbed. For the sequence similarity, profile, and sub-profile- methods we did not adjust the alignment in any way from that found in PFAM or PKR.

Aligned sequence data and sub-types

To test and demonstrate the method initially, we chose four examples of large enzyme families with clear sub-types: nucleotidyl cyclases, eukaryotic protein kinases, lactate/malate dehydrogenases and trypsin-like serine proteases. For all of these families laboratory experiments (e.g. site-directed mutagenesis or crystallography) or manual analysis of the alignments have been used previously to determine details regarding specificity. We also sought examples where phylogeny or simple sequence comparison would not easily lead to a correct prediction of catalytic sub-type. For all four of these protein families, we generated trees using the Clustal W package (1000 bootstrap trials, excluding positions with gaps, and not correcting for multiple substitutions). For all this procedure failed to separate the sub-types into distinct clades (results not shown). Note that division into sub-types may still be possible *via* other methods of tree construction (e.g. for the cyclases see Danchin, 1993).

The aim for these four protein families was to see if previously identified positions were found by the method, and check for additional insights that might have been missed during previous studies. For this reason, we required that all of the examples contained at least one protein of known 3D structure, to allow inspection of spatial proximity of amino acid residues thought to be important. Unless otherwise stated, alignments were taken from PFAM (Bateman *et al*., 1999), and groupings from inspection of SWISSPROT (Bairoch & Apweiler, 1999) annotations, or prior biochemical knowledge.

For a rigorous assessment, ideally one would require a carefully curated set of sub-groupings for a large set of alignments (e.g. from PFAM or SMART). It is unfortunate that this would involve a vast literature investigation that would be beyond the scope of this paper. It is also problematic to construct a large set of test examples automatically, since details regarding molecular function of a particular protein are not easily extracted from any database currently available. However, certain resources do provide some capacity to derive such a large set of alignments and sub-type groupings. Here, we divided proteins within the PFAM database (version 2.0, 1465 alignments) by considering functional details described in SWISSPROT (Bairoch & Apweiler, 1999). We first sought to use all keyword data (KW), however these produced ambiguities that lead to a vast number of meaningless groups. We thus chose to focus on details of enzymatic activity. We extracted activities by searching for the string "CC -!- CATALYTIC ACTIVITY" for each SWISSPROT entry in PFAM alignments, and then grouped sequences in alignments accordingly. After ignoring groups with fewer than ten sequences, we constructed all possible group combinations, and ignored those where sequences were contained in more than one group.

The above procedure initially produced 96 groups from 50 alignments. Inspection ruled out 18 (e.g. different names for the same activities; knowledge that enzymatic function was not conferred in the domain considered, etc.). Another 16 were ignored because the pairwise sequence identities indicated that the divisions based on catalytic activity would be easily discernable by sequence comparison. 62 groupings from a total of 42 alignments remained. These are described in Table 1. To avoid biasing due to multiple groups from a single alignment, we randomly chose one grouping for each alignment. This resulted in a set of 42 groupings over 42 alignments.

A problem with the above procedure is the assignment of catalytic activity to the correct domain. For example, the groupings for SH2 domains were discarded as we knew that the kinase catalytic activity was not localised in this domain. There is no simple way of doing this by parsing SWISSPROT, therefore, the results of analysing these data must be considered with some caution.

All of the sequence alignment and sub-type data are available from the authors.

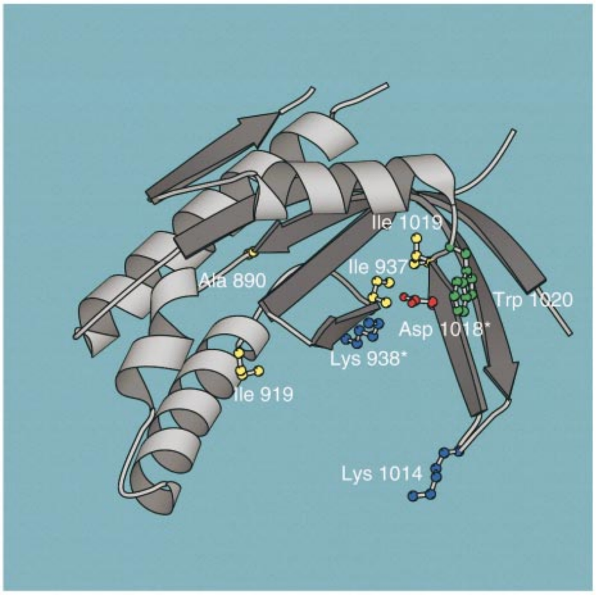


Figure 2. Rasmol (Sayle & Milner-White, 1995) Figure showing the structure of adenylate cyclase (PDB accession 1ab8, chain A), with positions found to confer specificity for adenylate or guanlyate by the method. Those that are starred (*) were reported to switch the specificity from guanylate to adenylate by mutagenesis (Tucker *et al*..,1998). More details are given in the text.

Results

Nucleotidyl cyclases

Nucleotidyl cyclases are a family of membrane attached or cytosolic domains that catalyse the reaction that forms a cyclic nucleotide monophosphate from a nucleotide triphosphate. The known cyclases act either on GTP (guanalyate cyclase) or ATP (adenylate cyclase). Mutations of two residue positions from Glu-Lys and Cys-Asp are known to be sufficient to change the specificity of the enzyme from GTP to ATP (Tucker *et al*., 1998). Mutations of several other residues near to the key Cys-Asp change were shown not to have any significant effect on specificity or enzymatic activity.

Figure 1 shows an alignment of nucleotidyl cyclases highlighting positions that have an entropy $Z > 3.0$. These positions are shown on the known 3D structure of adenylate cyclase (Zhang *et al*., 1998; PDB code 1ab8) in Figure 2. The first and third best positions are the Asp-Cys (residue 1018 in 1ab8, $Z = 6.5$) and Lys-Glu (938, $Z = 4.0$) changes identified by Tucker *et al*. (1998) that can be changed to modify cyclase specificity. Positions 1019 ($Z = 2.6$) and 1020 ($Z = 3.9$) were also identified by Tucker *et al.* but the change from Leu,Phe in guanylate cyclase to Ile,Trp (as is seen in most adenylate cyclases) in concert with the changes above actually lead to a poorer adenylate cyclase activity. The Trp-Phe (1020) change implies that a larger side-chain is needed in the adenylyl cyclases, and changes Ile-Val (937, $Z = 3.0$) and Val/Ile-Leu (1019) appear to be involved in subtle positioning of the residues that are adjacent on the sequence. Positions 937 and 1019 pack against one another in the known structures, implying a complementary change. The Val/Ile-Leu (1019) change is also interesting in that it suggests that the adenylyl cyclases require a branched C^{β} residue (i.e. two non-hydrogen substituents on the beta carbon, as is only seen in valine, isoleucine or threonine) instead of the non-C^{β}-branched leucine found in the guanylyl cyclases. Inspection of the structure suggests that this may have to do with adopting a slightly different main-chain conformation: branched C^{β} residues are slightly more restricted in the backbone psi/phi conformations that they can adopt (Swindells *et al*., 1995). These observations may help to explain why the mutants involving position 1019 in guanylate cyclase (substituting Leu with Ile; in concert with the Glu-Lys and Cys-Asp changes mentioned above) performed by Tucker *et al.* (1998) lead to poorer adenylate cyclase activities. If the changes had been made in concert with the appropriate mutation at position 937, then activity may have been closer to the wild-type.

The method also identified additional positions. The second best scoring position was the Lys/Arg-Met substitution (1014, $Z = 4.6$), which is also far away from the others in space if one considers a single cyclase subunit. However, inspection of the dimeric structure of adenylate cyclase (PDB code 1azs) shows that the equivalent position from the adjacent subunit is in the same location as the other positions discussed above. Changes Ala-Ser

Table 1. Groupings for PFAM alignments extracted from SWISSPROT

No.	PFAM name	Substrates
1	2-Oxoacid_dh	Acetyl-coA/succinyl-coA
2	ATP-gua_Ptrans	Creatine/L-arginine
3	Aconitase_C	***3-Isopropylmalate***/malate/***citrate***
4	Epimerase	dTDP-glucose/UDP-glucose
5	FGGY	D-xylulose/glycerol
6	GATase	1-(2-carboxyphenylamino)-1-deoxy-D-ribulouse-5-phosphate/ ***2ATP+glutamine/*** ***ATP+xanthosine-5′-phosphate+L-glutamine/*** ***chorismate+L-glutamine***
7	GATase_2	***5-Phospho-β-D-ribosylamine+L-glutamate/*** ***ATP+L-asparate+L-glutamine/*** ***L-glutamine+D-fructose-6-phosphate***
8	GHMP_kinases	D-Galactose/L-homoserine
9	HMA	ATP/HG+NADP(+)+H(+)
10	OTCace	Asparate/ornithine
11	Orn_DAP_Arg_deC	***L-Arginine/L-ornithine***/meso -2,5diaminoheptanedioate
12	PDEase	(Adenosine/Guanosine) -3′,5′-cyclicphosphate
13	PGAM	***2-phosphoglycerate+2,3-diphosphoglycerate/*** ATP+D-fructose-6-phosphate/ ***D-fructose-2-6-bisphosphate***
14	PGM_PMM	α-D-glucose-1-phosphate/D-mannose-1-phosphate
15	Pribosyltran	IMP/orotidine-5′-phosphate
16	Rhodanese	Protein-Tyr-phosphate/thiosulphate+cyanide
17	Rieske	Plastoquinol-1+2 oxidised plastocyanin/ QH_2-+2 ferricytochrome *c*
18	SQS_PSY	2-Farnesyldiphosphate/prephytoenediphosphate
19	S_T_dehydratease	***L-theronine/O-acetyl-L-serine/*** ***O-phospho-L-homoserine***
20	Semialdhyde_dh	L-Asparate-semialdehyde/ N-Acetyl-L-glutamate-5-semialdehyde
21	aakinase	L-Aspartate/***L-glutamate/L-glutamate-5-semialdehyde***
22	aconitase	3-Isopropylmalate/citrate
23	adh_zinc	Alcohol/cinnamylalcohol
24	aminotran_1	L-Asparate+2-oxoglutarate/*S*-adeosylmethionine
25	aminotran_3	***(S)-4-amino-5-oxopentanoate/*** ***4-aminobutanoate+2-oxoglutarate/*** ***L-ornithine+A 2-oxoacid/*** ***N-2-acetyl-L-ornithine+2 oxoglutarate***
26	cytochrome_b_N	Plastoquinol-1+2 oxidised plastocyanin/ QH_2-+2 ferricytochrome *c*
27	guanylate_cyc	ATP/GTP
28	malic	NAD/NADP
29	isodh	***3-Carboxy-2-hydroxy-4-methylpentanoate/*** ***Isocitrate+NAD/*** ***Isocitrate+NADP***
30	ldh	L-lactate/L-malate
31	ligase-CoA	ATP/GTP
32	lyase_1	L-Argininosuccinate/L-malate
33	oxidored_nitro	Reduced ferredoxin+6H++N2 +NATP/ chlorophyllidea+NADP
34	oxidored_q1	Plastoquinone/ubiquinone
35	oxidored_q1_N	Plastoquinone/ubiquinone
36	pyr_redox	***dihydrolipoamide/*** ***Hg+NADP+H^{+}/NADPH+glutathione***
37	pyridoxal_deC	L-glutamate/L-tryptophan
38	tRNA-synt_1	L-Isoleucine/***L-leucine/L-methionine/L-valine***
39	tRNA-synt_1b	***L-tryptophan/L-tyrosine***
40	tRNA-synt_2	***L-aspartate***/L-lysine/L-histidine/***L-asparagine***
41	tRNA-synt_2b	***L-Histidine/L-proline***/L-serine/***L-threonine***
42	tyrosinase	2-Catechol/L-tyrosine+L-dopa

Substrates for different sub-types are separated either by a/character or a line-break. Names in boldfaced italic text indicate those that form the randomly chosen group in instances where more than one grouping was available.

(890, $Z = 3.1$) and Ile-Tyr (919, $Z = 3.6$) are in the same approximate spatial location as the others, but are not in direct contact with any of the positions mentioned above. Inspection of the structure suggests that these changes may be responsible for subtle shifts in secondary structures that may help accommodate different substrates. Alternatively, these could be evolutionary relics, reflecting the likely divergence of adenlyate and guanylate cyclases (e.g Danchin, 1993). These positions may simply have not yet been subject to genetic drift that may have occurred at the majority of positions within the cyclases that are not involved in function. Only further experimental studies on the cyclases can reveal the meaning of these changes.

Protein kinases

Proteins serine, threonine and tyrosine kinases form one of the largest protein families known, estimated to make up between 1-2 % of proteins from metazoan genomes (e.g. Chervitz *et al*., 1998). They function to attach a phosphate group to a hydroxyl moeity on a particular amino acid side-chain. A major division is that between serine/threonine and tyrosine kinases. Serine and threonine are quite similar in size and shape, with a hydroxyl group attached to the C^{β} carbon; the only difference is a methyl group in threonine in place of a hydrogen in serine also attached to the C^{β} carbon. In tyrosine, however, the hydroxyl group is attached to a six-membered aromatic ring, making both the chemistry of the reaction and the size of the substrate substantially different. Certain positions are known to confer this specificity. Within sub-domain VI of protein kinases (Hanks *et al*., 1988), the consensus sequence RDLKPEN is usually found in serine/threonine kinases, whereas the sequence RDLAARN is typical of tyrosine kinases (Hanks & Hunter, 1996). Analysis of the first kinase crystal structures also identified other regions (Taylor *et al*., 1995). We used the alignment (295 sequences) and divisions from the protein kinase resource (Smith *et al*., 1997). The three major categories of Ser/Thr kinases were grouped into a single type, and the category "other protein kinases" (OPK, of unknown or ambiguous sub-type) were ignored.

Figures 3 and 4 show an alignment and 3D structure highlighting the ten positions with the highest entropy Z score ($Z > 3.04$). All of these positions lie in the C-terminal (catalytic domain), and most of the top scoring positions lie in two regions of the sequence and adjacent in space. One of these regions, containing positions 168 to 170 (in PDB code 2cpk, $Z = 5.7$, 3.4, 3.9) are in the Hanks *et al*. sub-domain VI region known to determine kinase specificity (discussed above), in the catalytic loop (Lys,Pro,Glu-Ala, Ala, Arg). The second contains substitutions Thr/Ser-Pro (201, $Z = 5.5$), Trp-Lys/Arg (203, $Z = 3.2$), Tyr - Trp (204, $Z = 6.5$) and Ile/Val/Leu-Ser (209, $Z = 3.2$) from sub-domain VIII,

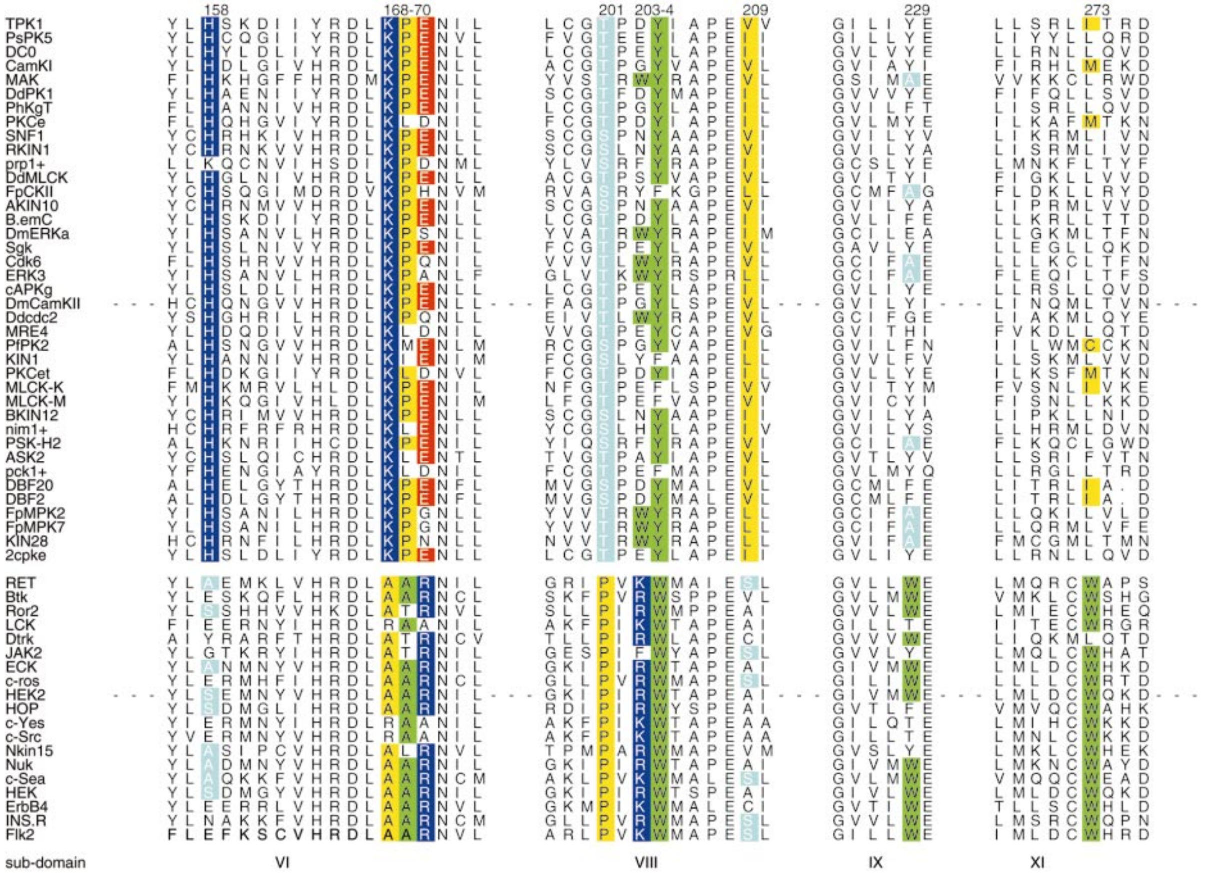


Figure 3. Alscript (Barton, 1993) Figure showing an alignment of representatives of protein kinases with positions predicted to confer specificity to Ser/Thr or Tyr highlighted by the method. Numbers above the alignment refer to positions discussed in the text, and refer to the PDB structure 2cpk. Labels below the alignment refer to the kinase "domain" nomenclature of Hanks *et al*. (1988). Other details are as for Figure 1.

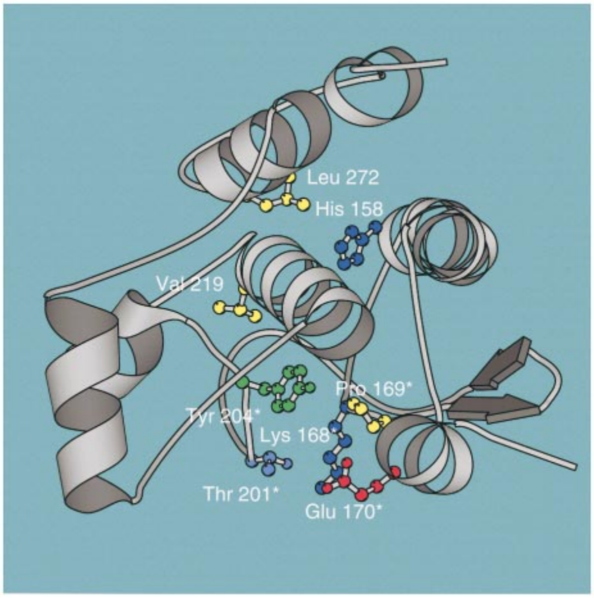


Figure 4. Rasmol (Sayle & Milner-White, 1995) Figure showing the structure of cAMP-dependent serine threonine kinase (PDB accession 2cpk, chain E), with positions found to confer specificity for serine/threonine or tyrosine by the method. Those that are starred (*) are taken from the literature (Hanks & Hunter, 1996; Taylor *et al*. 1995) and are thought to confer serine/threonine *versus* tyrosine specificity. More details are given in the text

within (or near to) the P + 1 loop. Most of the positions within these regions were identified by Taylor *et al*. (1995) as those that are most characteristically distinct for the Ser/Thr and Tyr sub-types.

The other positions shown in Figure 4 are His - Ala/Ser (158, subdomain VI), Ala-Trp (229, subdomain IX, $Z = 3.1$) and Leu/Met/Cys-Trp (273, subdomain XI, $Z = 3.3$), and numerous other positions with $Z >= 3.0$ are near to these in space (results not shown). None of these positions are close enough to interact directly with those residues discussed above. However, as for the cyclases, inspection suggests that they may be involved in aiding subtle conformational changes of the structure to accommodate differing substrates. Several other differences between protein kinase A (Ser/Thr) and insulin receptor tyrosine kinase (IRK) were reported by Taylor *et al*. (1995), though not detected during this study. Inspection of the alignment shows that the positions are either not conserved across the sub-types, or show substantial overlap between the Ser/Thr and Tyr sub-types when one considers all homologous sequences (results not shown).

Lactate/malate dehydrogenases

Dehydrogenases that act on lactate and malate are part of a larger superfamily of Rossmann fold (nucleotide-binding domain) containing enzymes (e.g. Rossmann & Argos, 1976; Russell & Barton, 1992). Lactate and malate dehydrogenases (LDH, MDH) form a large sub-set of this family, and share the additional common feature of a similar substrate binding domain. They are found across all kingdoms of life and are thus highly divergent, meaning that it is difficult to distinguish between lactate and malate sub-types. A key mutation Gln-Arg (position 102 in pig LDH, PDB code 9ldta) is known to switch the specificity from lactate to malate (Wilks *et al*., 1988), and is known to be involved in distinguishing lactate from malate. In addition, all possible variants of postions 101 and 102 have been analysed (Hawrani *et al*., 1996). These variants were used to determine residues conferring specificities for many other substrates known to bind to this large family of enzymes, including phenyl-lactate, hydroxyisocaproate and 4-phenyl-2-hydroxy-butanoate, though we do not consider these substrates here.

Figure 5 shows an alignment illustrating the six positions with the highest entropy Z score. The position with the highest entropy ($Z = 4.0$) is the Gln-Arg (102) change identified by experiment. With the exception of the Tyr - Pro change (position 190, $Z = 3.4$) all positions are near to the Gln - Arg position and surround the experimentally determined location of NAD. They are thus likely to be involved in the lactate/malate distinction. All have $Z > 3.0$ with the exception of the Glu-Gly change at position 194 ($Z = 2.9$), which is shown as it also appears to have a role in determining substrate specificity.

Serine proteases

Trypsin-like serine proteases are a large family of enzymes involved in the hydrolysis of peptide bonds. Although they all act *via* a similar catalytic mechanism, they have different preferences for the amino acids that they prefer to cleave. Trypsin cleaves C-terminal to arginine or lysine, chymotrypsin next to large aromatic residues, and elastase cleaves next to small, uncharged amino acid residues, and it was proposed long ago that the distinction is conferred by key changes in a specificity pocket (e.g. Fersht, 1985). Three positions were proposed originally to define the pocket. An aspartic acid found in trypsin (Asp189) is usually replaced by a small residue in chymotrypsins (Ser) and elastases (Gly). Two positions adjacent to this in space were originally described as defining substrate differences in these three families (e.g. Fersht, 1985). Positions 216 and 226 (in trypsin) are generally glycine in chymotrypsins and trypsins, but replaced by valine and threonine in elastases.

Figure 6 shows the positions with $Z >= 3.0$ identified by the method for the trypsin-like serine proteases when grouped into elastase, chymotrypsin and trypsin sub-types. The top two scoring positions (position 189, in bovine trypsin, PDB code 5ptp, $Z = 5.6$ and 226, $Z = 3.9$) correspond to two

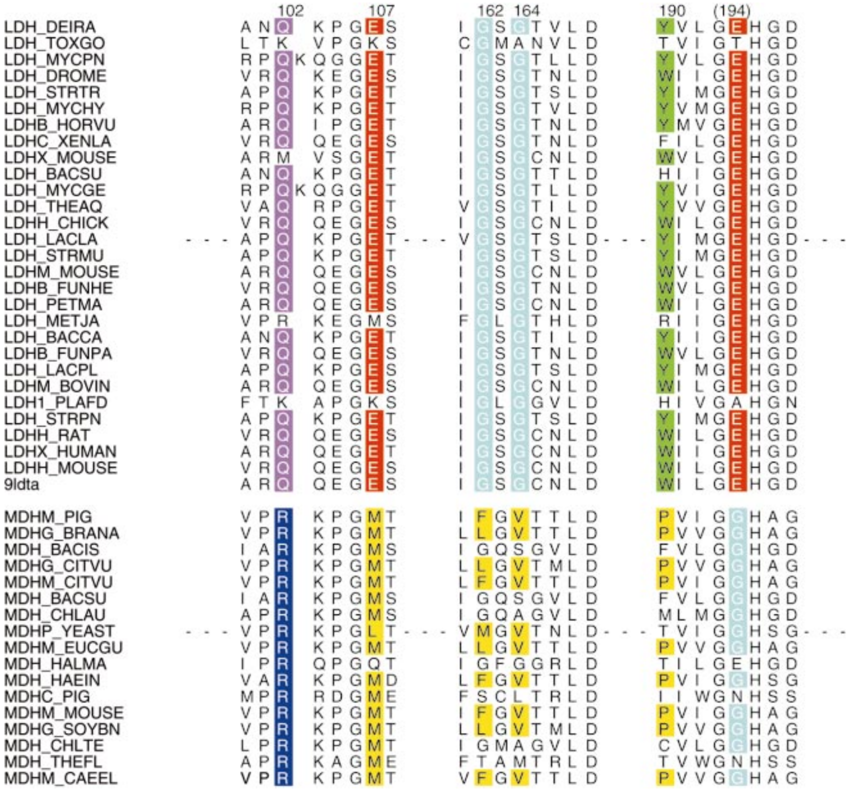


Figure 5. Alscript (Barton, 1993) Figure showing an alignment of representatives of lactate and malate dehydrogenases with positions predicted to confer specificity for lactate or malate highlighted by the method. Numbers above the alignment refer to positions discussed in the text, and refer to the PDB structure 9ldt. Other details are as for Figure 1.

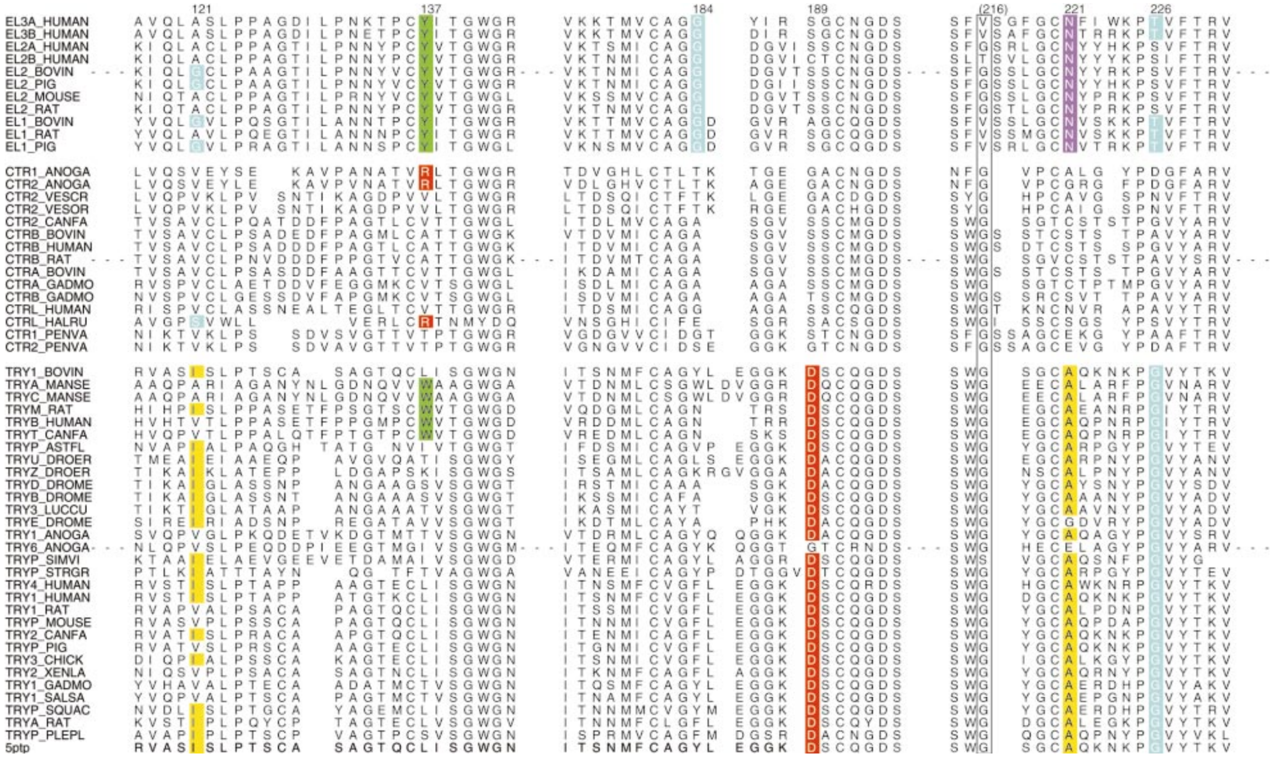


Figure 6. Alscript (Barton, 1993) Figure showing an alignment of representatives of tyrpsin-like serine-proteases with positions predicted to confer specificites known for trypsins, chymotrypsins or elastases highlighted by the method. Numbers above the alignment refer to positions discussed in the text, and refer to the PDB structure 5ptp. The box shows a position not highlighted by the method that is thought to be involved in enzyme specificity. Other details are as for Figure 1.

of the pocket positions above. The third pocket position (216) has a low Z score (1.0). Inspection of the alignment (see boxed position in Figure 6) shows that glycine is frequently tolerated in the elastases sub-type, giving a low Z score. The third best scoring position (221, $Z = 3.6$) is an Asn residue in the elastases, and generally an Ala residue in trypsins, and is near to the specificity pocket discussed above. Of the other three positions identified only position 184 ($Z = 3.1$) is near to the other pocket forming residues in space. Here glycine, which is preferred in the elastases may aid the recognition of small side-chains in elastase substrates. The other two positions with $Z > 3.0$ (121, $Z = 3.5$; 137, $Z = 3.3$) are not near to the pocket in space, though like the cyclases and kinases (above), it is possible that they are involved in any subtle conformational adjustments to accommodate differing substrates.

Prediction accuracies for cyclases, kinases, dehydrogenases and serine proteases

Table 2 reports the prediction accuracies for the four families discussed above (see legend for details). The accuracies of the methods vary greatly according to the percentage sequence identity threshold used to include sequences in the alignment.

The sequence similarity method is consistently better than the BLAST method for these four families. When all but the very distant homologues of the predicted sequence are removed (i.e. 30 % threshold) the HMM and profile methods clearly out-perform these two methods. The distinction between the methods diminishes as the threshold is raised, and when sequences sharing identities of 50 % or greater with the predicted sequence are included in the alignment, the predictive accuracies

Table 2. Prediction accuracies of all methods for four protein families

	Cyclase (2) 72		Kinase (2) 293		Dehydrogenase (2) 103		Trypsin (3) 101	
Sequence identity threshold (%)	Pred.	Acc. (%)	Pred.	Acc. (%)	Pred.	Acc. (%)	Pred.	Acc. (%)
20	6	SP: 0 HMM: 33 P: 0 SS: 0 B: 0	191	SP: 100 HMM: 94 P: 96 SS: 60 B: 57	103	SP: 47 HMM: 78 P: 47 SS: 24 B: 56	0	SP: N/A HMM: N/A P: N/A SS: N/A B: N/A
30	43	SP: 100 HMM: 91 P: 63 SS: 21 B: 9	293	SP: 100 HMM: 98 P: 97 SS: 98 B: 94	103	SP: 88 HMM: 88 P: 88 SS: 63 B: 56	37	SP: 76 HMM: 65 P: 68 SS: 54 B: 39
40	63	SP: 90 HMM: 63 P: 54 SS: 56 B: 37	293	SP: 100 HMM: 100 P: 100 SS: 100 B: 100	103	SP: 90 HMM: 92 P: 91 SS: 87 B: 85	89	SP: 96 HMM: 85 P: 83 SS: 66 B: 75
50	72	SP: 100 HMM: 99 P: 99 SS: 97 B: 88	293	SP: 100 HMM: 100 P: 100 SS: 100 B: 100	103	SP: 93 HMM: 95 P: 91 SS: 97 B: 92	101	SP: 97 HMM: 91 P: 85 SS: 94 B: 90
60	72	SP: 100 HMM: 100 P: 100 SS: 99 B: 97	293	SP: 100 HMM: 100 P: 100 SS: 100 B: 100	103	SP: 96 HMM: 96 P: 98 SS: 98 B: 97	101	SP: 100 HMM: 98 P: 98 SS: 97 B: 97
100	72	SP: 100 HMM: 100 P: 100 SS: 99 B: 97	293	SP: 100 HMM: 100 P: 100 SS: 100 B: 100	103	SP: 100 HMM: 100 P: 100 SS: 99 B: 98	101	SP: 100 HMM: 100 P: 100 SS: 99 B: 100

The number of sub-types for each family is shown in parenthesis along with the family names in the first row. The number of sequences in the alignment is shown in the first row under the family name for each family. To simulate the situation where a close homologue is not available, for each sequence (assumed to be of unknown sub-type) all other sequences with percentage identity greater than a threshold were ignored. The first column shows the similarity threshold used to eliminate sequences. For certain sequences (e.g. trypsins at 20 %), the elimination process removes all sequences of the sub-type. In these situations, the sub-type of such a sequence is considered unpredictable. Pred gives the number of sequences for which a sub-type prediction was made. Acc gives the percentage accuracy of prediction for those sequences predicted. The four numbers given in each column are the accuracies for the sub-profile (SP), HMM (HMM), profile (P), sequence similarity (SS), and BLAST (B) methods. For example, for the 191 kinases at threshold of 20 %, the values are 100 % (sub-profile), 94 % (HMM) 96 % (profile) and 60 % (sequence similarity). N/A in the accuracy column means that no predictions could be made.

of the methods are indistinguishable as expected. The HMM method is almost always superior to the profile method. This may indicate that the risk of wrong alignment by *hmmsearch* is well compensated by the fact that *hmmsearch* pays careful attention to the gap penalties which are not considered in the profile method where we only incorporate the ''match'' states in the profile. It may also reflect that only one of the four alignments studied here (the kinases) was hand-curated.

The sub-profile method performs best of all. Removing contributions from positions that do not discriminate between the sub-types has a dramatic effect. This is perhaps not surprising, as non-discriminating positions will be expected only to contribute noise to the overall score. The inherent variability of the noise would be expected to produce incorrect predictions. At a 20 % threshold, the poor accuracy of the sub-profile method for cyclases (0 %) and dehydrogenases (47 %) is due to the fact that the removal of close homologues leaves very few sequences with which to build a profile (fewer than three). The *Z*-scores based on so few sequences do not capture the important positions in this situation and hence the performance of the sub-profile method is not significantly better than the profile method. This highlights the fact that the method can only work efficiently with a sufficient number and diversity of sequences of a particular sub-type.

PFAM alignments

The large set of alignments and groups extracted automatically from PFAM provides a useful set for assessing the predictive accuracy of the methods above. Based on the results for the four families discussed above, we chose only two similarity thresholds (20 % and 30 %) for the automatically generated PFAM alignments grouped by SWISSPROT, and we only applied the sequence similarity, BLAST and sub-profile methods. Inspection showed that the results did not differ greatly from those for the four families discussed above. We expect, in particular, that the HMM method would also be effective in discerning sub-types.

Out of a total of 2593 sequences in the 42 alignments/groupings, sub-type predictions could be made for 1520 and 2204 sequences at the 20 % and 30 % thresholds, respectively (no prediction could be made on the remaining since removing the close homologues removed all the sequences within a group). At the 20 % threshold, the accuracies for the sequence similarity, BLAST, and sub-profile methods were 51.5, 69.8 and 91.2 respectively. With a 30 % threshold the figures were 68.1, 78.2 and 94 %. Considering the percentage accuracies averaged for each of the 42 families: with a 20 % threshold the values were 46, 62.6, and 82 %; with a 30 % threshold, the values were 68, 79.2 and 94 %. It is clear that the sub-profile method is providing highly successful predictions of protein sub-types, even when only very distant homologues are present in the alignment.

Table 3 shows details calculated for each of the 42 PFAM alignments with a sequence identity threshold of 30 %, for sub-type groupings specified in Table 1. In no instance do the sequence similarity or BLAST methods significantly out-perform the sub-profile method, though there are two instances (3′5′-cyclic nucleotide phosphodiesterase, PDEase; CoA ligases, ligase-CoA) where the three methods essentially perform randomly (i.e. when there are two groups of approximately equal size, one expects about a 50 % prediction accuracy). In neither case does it appear that specificity is located in a different domain, implying that these are genuine failures of the method to discern the respective sub-types. Apart from these, the sub-profile method gives good and in many instances perfect predictions.

There is insufficient space to discuss the analysis on PFAM alignments in detail. However, inspection of select families is illustrative of the general applicability of the method. One example, where the sub-profile method greatly outperforms the sequence similarity method (100 % compared to 24 % and 28 %), is for the pyridine nucleotide-disulphide oxidoreductases (class I; PFAM name pyr_-redox). Here the method has identified 14 positions within the alignment predicted to confer specificity between dihydrolipoamide, mercury (Hg^{+}) and glutathione. Of these positions, 12 are found near to the experimentally determined location of the co-factor FAD in dihyrolipoamide dehydrogenase from *Azotobacter vinelandii* (PDB code 3lad; Mattevi *et al*., 1991). Residues Pro13, Tyr16, Lys34, Gly104 Met324, Ala326 and His327 are located within the FAD binding domain (residues 1 to 158 and 278 to 348) whereas residues Ser389, Gly390, Ala449, Ala456 and Glu459 are within the dimerisation domain (according to SCOP, Murzin *et al*., 1995). It is interesting that these last four residues are near (in sequence and space) to two catalytic residues (His450 and Glu455; Mattevi *et al*., 1991). Inspection of single chains from the dimeric structure did not show proximity of residues from the two domains. However when one considers the active homodimer, both sets of residues are near to the bound FAD molecule (in two sites within the homodimer). It is clear that the method has successfully identified positions likely to confer specificity within this diverse group of enzymes, and which could be the subject of site-directed mutagenesis or other experiments to probe enzymatic function. Moreover, by focussing attention on these positions, the method is able to predict the correct specificity perfectly, even if all homologues with >30 % sequence identity are removed.

A similar picture is seen for a family of carbohydrate kinases (PFAM FGGY). Here, the method has identified 11 positions that are predicted to distinguish between D-xylulose and glycerol. Within the known structure of glycerol kinase from *Escherichia coli* (PDB code 1glf; Feese *et al*., 1998) six

Table 3. Prediction accuracies for automatically derived groups for PFAM alignments

PFAM Family	*N*pos	*N*pos (Z > = 3.0)	*N*seq	Pred.	Acc. SS (%)	Acc. B (%)	Acc. SP (%)
2-oxoacid_dh (2)	257	7	27	10	40	50	90
ATP-gua_Ptrans (2)	429	10	38	4	100	100	100
Aconitase_C (2)	218	2	50	50	100	100	100
Epimerase (2)	797	16	37	24	100	82	100
FGGY (2)	431	6	26	26	50	88	100
GATase (3)	388	6	63	54	69	100	100
GATase_2 (3)	268	6	49	19	89	95	100
GHMP_kinases (2)	93	2	28	28	100	96	100
HMA (2)	30	1	67	56	16	88	98
OTCace (2)	461	16	73	73	90	96	100
Orn_DAP_Arg_deC (2)	605	6	29	29	34	79	100
PDEase (2)	290	6	44	34	38	26	53
PGAM (2)	278	4	28	28	43	100	96
PGM_PMM (2)	1160	12	31	31	61	84	90
Pribosyltran (2)	268	4	41	41	100	100	100
Rhodanese (2)	201	1	63	63	71	89	98
Rieske (2)	175	1	31	30	57	100	100
SQS_PSY (2)	349	6	30	30	80	97	100
S_T_dehydratase (3)	551	7	54	54	87	93	96
Semialdhyde_dh (2)	623	11	34	34	100	100	100
aakinase (2)	316	3	35	35	94	100	100
aconitase (2)	604	18	59	59	75	81	95
adh_zinc (2)	1133	28	111	91	87	85	100
aminotran_1 (2)	738	14	48	35	57	97	94
aminotran_3 (4)	604	7	55	32	16	38	72
cytochrome_b_N (2)	409	6	306	233	55	55	100
guanylate_cyc (2)	385	7	98	67	46	100	100
isodh (3)	586	11	79	79	14	42	89
ldh (2)	436	5	103	103	63	56	93
ligase-CoA (2)	169	1	32	32	44	38	47
lyase_1 (2)	542	0	32	32	38	100	100
malic (2)	615	10	28	4	100	50	100
oxidored_nitro (2)	722	13	74	74	78	96	95
oxidored_q1 (2)	519	6	295	295	79	79	86
oxidored_q1_N (2)	71	0	92	56	84	84	100
pyr_redox (3)	797	16	51	29	24	28	100
pyridoxal_deC (2)	455	3	27	17	12	0	65
tRNA-synt_1 (3)	2210	37	59	59	92	68	93
tRNA-synt_1b (2)	614	13	33	33	100	100	100
tRNA-synt_2 (2)	1001	17	38	37	81	68	92
tRNA-synt_2b (3)	983	12	66	66	97	100	100
tyrosinase (2)	418	4	29	18	83	100	100

Details of applying our method to 42 groups derived automatically for PFAM alignments. The value given in parentheses after the PFAM name is the number of groups considered (the exact groups are specified in Table 1). *N*pos indicates the number of positions in the alignment *N*pos (Z > = 3) the number where the relative entropy Z score is 3 or higher. *N*seq gives the number of sequences in the alignment, *P*red the number for which predictions were possible (given the sequence identity threshold of 30 %), Acc SS , Acc B and Acc SP give, the prediction accuracies for the sequence similarity, BLAST and sub-profile methods respectively.

of these residues (Arg188, Gln246, Gly259, Trp356, Asp409, Leu418) are near to the experimentally determined location of bound glycerol or ATP and two others (Thr86, Val165) appear to be interacting with these residues. As for the pyridine nucleotide-disulphide oxidoreductases (above), the method appears to have identified residues conferring specificity, and has predicted sub-type accurately even in the absence of close homologues.

Discussion

We have presented and evaluated a method for assigning and analysing sub-types within protein sequence alignments. For four examples we have shown that the method is able to detect positions known to confer specificity in close agreement with experiment. Both on these four examples, and the 42 groupings derived from PFAM/SWISSPROT, the method is shown to predict protein sub-types with remarkable success, even in the absence of closely related sequences of the same sub-type, and predictions are much better than those made by a simple sequence comparison.

There are obvious similarities between the method presented here and those of Livingstone & Barton (1993), Casari *et al.* (1995) and Lichtarge *et al.* (1996a). One difference encoded in the method here is a careful handling of non-identical positions by way of the construction of a hidden Markov model and incorporation of amino acid exchange matrices. Incorporation of exchange matrix data will permit amino acids not seen in the current set of known sequences from a sub-type if they have sufficiently similar physicochemical properties.

The method presented here is perhaps most similar in spirit to that of Sjolander (1998) with one important difference: we make no attempt at phylogenetic reconstruction, and no attempt to define sub-type groupings. Our method is most useful in analysing super-families where there are relatively few functional sub-types and there is some knowledge of members belonging to sub-types in the literature. This method takes advantage of the knowledge by constructing profiles of the sub-types explicitly, and also by an explicit analysis of the positions to detect functionally important sites.

Like other methods, that described here has many potential applications within studies of large protein sequence families. Prior knowledge of the functions of the sub-types can be used to extract regions on the protein sequence that are the best candidates for laboratory experiments to elucidate function. If an uncharacterised protein shares only a weak degree of sequence similarity with a large protein family, then the method can identify the correct sub-type. This is likely to be of greatest use when there are multiple orphan members of a protein family, and where some priority or rank order of experiments (e.g. ligand binding assays, etc.) must be assigned to keep laboratory experimental effort to a minimum.

The method may also be applied to genome annotation. Newly sequenced genes that are only weak matches to large protein families with different functions can be tested and highlighted if they contain amino acid residues that determine a particular functional sub-type. In this way, it might be possible to avoid ambiguities that arise when a weak sequence similarity score cannot distinguish between two or more different sub-types (e.g. amino acid permeases; see Figure 2(a) of Brenner, 1999).

Another application is for the prediction of spatial proximity of residues within proteins of unknown 3D structure. There have been several studies attempting to correlate intra-protein distances with correlated mutations (or compensating changes, or cooperative subsitutions, or correlated changes; or complementary changes; Gobel *et al.*, 1994; Taylor & Hatrick, 1994; Neher, 1994; Olmea & Valencia, 1994; Russell & Barton, 1994). Although techniques vary, the common theme is to identify positions within a protein sequence that are co-varying during evolution. Residues involved in conferring sub-type are frequently near to each other in space, even when they are far apart on the protein sequence. Methods that identify positions that confer sub-type (this study; Livingston & Barton, 1993; Casari *et al.*, 1995; Lichtarge *et al.*, 1996a; Sjolander, 1998) are thus likely to be of use in predicting inter and intra-protein distances (e.g. Pazos *et al.*, 1997).

A problem in extending the analysis described in this paper is the lack of any large source of data regarding sub-types. A large database of groupings extracted from the literature would be a time-consuming, but rewarding exercise for many further analyses. Such studies might be aided by recent attempts to extract textual data from the biological literature (e.g. Andrade & Valencia, 1998; Andrade, 1999).

The current availability of dozens of complete genomes provides a wealth of data that will require many computational analyses. Methods like that described here and others (e.g. Casari *et al.*, 1995; Lichtarge *et al.*, 1996a; Sjolander, 1998; Pellegrini *et al.*, 1999; Marcotte *et al.*, 1999; Enright *et al.*, 1999; Goh *et al.*, 2000) will be of great importance in attaching biological information to orphan sequences prior to time-consuming and costly laboratory experiments.

Acknowledgements

We thank Jim Fickett and David Searls (SB) for encouragement and support. We are indebted greatly to Pankaj Agarwal (SB) for detailed help in the design of the method and the *Z*-score computation, in addition to other useful comments. Thanks also to Richard Copley (SB/EMBL Heidelberg), Malcolm Duckworth (SB), Mark Hurle (SB), Chris Larminie (SB), Chris Ponting (NCBI/Oxford) and Richard Jackson (University College, London) for helpful discussions. We are grateful to two referees for helpful and constructive comments.

References

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

Andrade, M. A. (1999). Position-specific annotation of protein function based on multiple homologs. *ISMB*, **7**, 28-33.

Andrade, M. A. & Valencia, A. (1998). Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics,* **14**, 600-607.

Atrian, S., Sanchez-Pulido, L., Gonzalez-Duarte, R. & Valencia, A. (1998). Shaping of *Drosophila* alcohol dehydrogenase through evolution: relationship with enzyme functionality. *J. Mol. Evol.* **47**, 211-221.

Azuma, Y., Renault, L., Garcia-Ranea, J. A., Valencia, A., Nishimoto, T. & Wittinghofer, A. (1999). Model of the ran-RCC1 interaction using biochemical and docking experiments. *J. Mol. Biol.* **289**, 1119-1130.

Bairoch, A. & Apweiler, R. (1999). The SWISSPROT protein sequence data bank and its new supplement TrEMBL in 1999. *Nucl. Acids Res.* **27**, 49-54.

Barton, G. J. (1993). ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.* **1**, 37-40.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, ? (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.* **27**, 260-262.

Bauer, B., Mirey, G., Vetter, I. R., Garcia-Ranea, J. A., Valencia, A., Wittinghover, A., Camonis, J. H. &

Cool, R. H. (1999). Effector recognition by the small GTP-binding proteins Ras and Ral. *J. Biol. Chem.* **274**, 17763-17770.

Birney, E., Thompson, J. D. & Gibson, T. J. (1996). PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucl. Acids. Res.* **24**, 2730-2739.

Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet.* **15**, 132-133.

Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171-178.

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

Danchin, A. (1993). Phylogeny of adenylyl cyclases. *Advanc. Sec. Mess. Phosphoprot. Res.* **27**, 109-162.

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids*, Cambridge University Press, Cambridge, UK.

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

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature,* **402**, 86-90.

Feese, M. D., Faber, H. R., Bystrom, C. E., Pettigrew, D. W. & Remmington, S. J. (1998). Glycerol kinase from *Escherichia coli* and an Ala65 → Thr mutant: the crystal structures reveal conformational changes with implications for allosteric regulation. *Structure,* **6**, 1407-1418.

Fersht, A. R. (1985). *Enzyme Structure and Mechanism*, 2nd edit., Freeman and Company, New York.

Gobel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309-317.

Goh, C., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283-293.

Gribskov, M., Luthey, R. & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol.* **183**, 146-159.

Hanks, S. K. & Hunter, T. (1996). The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* **9**, 576-596.

Hanks, S. K., Quinn, A. M. & Hunter, T. (1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science,* **241**, 42-52.

Hawrani, A. S., Sessions, R. B., Moreton, K. M. & Holbrook, J. J. (1996). Guided evolution of enzymes with new substrate specificities. *J. Mol. Biol.* **264**, 97-110.

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

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

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996b). Evolutionarily conserved G-alpha-beta-gamma binding surfaces support a model of the G protein-receptor complex. *Proc. Natl Acad. Sci. USA,* **93**, 7507-7511.

Lichtarge, O., Yamamoto, K. R. & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.* **274**, 325-337.

Livingstone, C. D. & Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745-756.

Makarova, K. S. & Grishin, N. V. (1999). The Zn-peptidase superfamily: functional convergence after evolutionary divergence. *J. Mol. Biol.* **292**, 11-17.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature,* **402**, 83-86.

Mattevi, A., Schierbeck, A. J. & Hol, W. G. (1991). Refined crystal structure of lipoamide dehydrogenase from *Azotobacter vinelandii* at 2.2 angstroms resolution. A comparison with the structure of glutathione reductase. *J. Mol. Biol.* **220**, 975-994.

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

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

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA,* **91**, 98-102.

Olmea, O. & Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.* **2**, S25-35.

Pazos, F., Hlmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511-523.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA,* **96**, 4285-4288.

Rossmann, M. G. & Argos, P. (1976). Exploring structural homology in proteins. *J. Mol. Biol.* **25**, 75-95.

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

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

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

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

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

Shannon, C. & Weaver, W. (1963). *Mathematical Theory of Communication*, University of Illinois press, Champaign, IL.

Sjolander, K. (1998) Phylogenetic inference in protein superfamilies: analysis of SH2 domains. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* (Glasgau, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D. & Sensen, C., eds) pp. 165-174, AAAI Press, Menlo Park, CA.

Smith, C. M., Shindyalov, I. N., Veretnik, S., Gribskov, M., Taylor, S. S., Ten Eyck, L. F. & Bourne, P. E. (1997). The protein kinase resource. *Trends Biochem. Sci.* **22**, 444-446.

Sowa, M. E., He, W., Wensel, T. G. & Lichtarge, O. (2000). A regulator of G protein signaling interaction surface linked to effector specificity. *Proc. Natl Acad. Sci. USA*, **97**, 1483-1488.

Swindells, M. B., MacArthur, M. W. & Thornton, J. M. (1995). Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nature Struct. Biol.* **2**, 596-603.

Taylor, S. S., Radzio-Andzelm, E. & Hunter, T. (1995). How do protein kinases discriminate between serine/threonine and tyrosine? Structural insights from the insulin receptor protein tyrosine kinases. *FASEB J.* **9**, 1255-1266.

Taylor, W. R. (1986). The classification of amino acid conservation. *J. Theoret. Biol.* **119**, 205-218.

Taylor, W. R. & Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341-348.

Tucker, C. L., Hurley, J. H., Miller, T. R. & Hurley, J. B. (1998). Two amino acid substitutions convert a guanylyl cyclase, RetGC-1, into an adenylyl cyclase. *Proc. Natl Acad. Sci. USA*, **11**, 5994-5997.

Wilks, H. M., Hart, K. W., Feeney, R., Dunn, C. R., Muirhead, H., Chia, W. N., Barstow, D. A., Atkinson, T., Clarke, A. R. & Holbrook, J. J. (1988). A specific, highly acitve malate dehydrogenase by redesign of a lactate dehydrogenase framework. *Science*, **242**, 1541-1544.

Wu, G., Fiser, A., ter Kuile, B. & Miklos, M. (1999). Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc. Natl Acad. Sci. USA*, **96**, 6285-6290.

Zhang, G., Liu, Y., Ruoho, A. E. & Hurley, J. H. (1998). Structure of the adenylyl cyclase catalytic core. *Nature*, **386**, 247-253.

Zvelebil, M. J. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957-961.

Edited by J. Thornton

(Received 12 April 2000; received in revised form 7 July 2000; accepted 11 July 2000)