Sequence patterns derived from the automated prediction of functional residues in structurally-aligned homologous protein families

Ricardo Núñez Miguel

Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, United Kingdom

Received on December 18, 2003; revised and accepted on March 12, 2004
Advance Access publication April 8, 2004

ABSTRACT

Motivation: Most proteins have evolved to perform specific functions that are dependent on the adoption of well-defined three-dimensional (3D) structures. Specific patterns of conserved residues in amino acid sequences of divergently evolved proteins are frequently observed; these may reflect evolutionary restraints arising both from the need to maintain tertiary structure and the requirement to conserve residues more directly involved in function. Databases of such sequence patterns are valuable in identifying distant homologues, in predicting function and in the study of evolution.

Results: A fully automated database of protein sequence patterns, Functional Protein Sequence Pattern Database (FPSPD), has been derived from the analysis of the conserved residues that are predicted to be functional in structurally aligned homologous families in the HOMSTRAD database. Environment-dependent substitution tables, evolutionary trace analysis, solvent accessibility calculations and 3D-structures were used to obtain the FPSPD. The method yielded 3584 patterns that are considered functional and 3049 patterns that are probably functional. FPSPD could be useful for assigning a protein to a homologous superfamily and thereby providing clues about function.

Availability: FPSPD is available at http://www-cryst.bioc.cam.ac.uk/~fpspd/
Contact: Ricardo@cryst.bioc.cam.ac.uk

INTRODUCTION

In principle the availability of a growing number of completely sequenced genomes, including that of human, allows the identification of proteins that may be targets for new drugs. Clues about the functions of proteins identified in this way, and therefore about their value as possible drug targets, can often be obtained from identification of homologues of known function in protein databases, such as the Protein Data Bank (Berman et al., 2000), SWISS-PROT (Bairoch and Apweiler, 1999) and HOMSTRAD (deBakker et al., 2001; Mizuguchi et al., 1998). Unfortunately, in many cases there are no close homologues of known function available, i.e. Psi-Blast does not find confident homologues, and the function of the protein remains undefined.

However, proteins with a common evolutionary history and function can frequently be identified from the occurrence of clusters of conserved residues in their amino acid sequences. Often when the search for global similarity of complete sequences fails to show positive results, such conserved clusters, called variously patterns, motifs or fingerprints, may allow identification of distantly related proteins and thus give clues about function (Andrade and Sander, 1997; Dayhoff et al., 1983; Rastan and Beeley, 1997). Databases such as PROSITE (Hofmann et al., 1999, Kasuya and Thornton, 1999), BLOCKS (Henikoff et al., 1999; Henikoff and Henikoff, 1991), PRINTS (Attwood et al., 1999) and PROF_PAT (Bachinsky et al., 2000) have accumulated sequence patterns for protein families. These sequence patterns include residues responsible for the three-dimensional (3D) structure of the protein as well as those directly involved in its function.

Methods that seek to identify functionally or structurally important residues in proteins of known 3D-structure include those that search for residues with backbone dihedral angles in strained conformations (Herzberg and Moult, 1991), those involving clusters of charged residues (Zhu and Karlin, 1996), and those involved in sizeable cavities or clefts, large enough for substrate binding or protein–protein interaction (Laskowski et al., 1996; Liang et al., 1998). Some methods rely on the measurement of surface properties such as hydrophobicity, planarity, size or shape (Jones and Thornton, 1997a, b), while others depend solely on the energetics of protein structures (Elcock, 2001), on surface mapping of phylogenetic information (Armon et al., 2001) or on the search for small numbers of highly conserved positions, separated by limited variable spacing, PRATT (Jonassen et al., 1995).

Using these ideas databases of enzyme active sites have been developed. For example, PROCAT (Wallace et al., 1997) is a database of 3D enzyme active site templates, which
can be thought of as the 3D equivalent of the 1D templates found in PROSITE and PRINTS, while BALSAMIC (http://www.bioinformatics.leeds.ac.uk/~balsamic/) is a database of active/functional sites of proteins that can be used to compare similarly shaped functional sites on different protein surfaces.

In this paper we ask whether it might be possible to identify further patterns characteristic of protein function from 3D structures of families of homologous proteins. We investigate whether it is possible to distinguish those patterns that are directly involved in the function of the protein from those that are involved principally in maintaining 3D structure. We have developed the Functional Protein Sequence Pattern Database (FPSPD), which uses environment-dependent substitution tables (Overington et al., 1992) and the evolutionary trace (Innis et al., 2000; Licharge et al., 1996) method to identify such residues in the structurally aligned homologous families compiled in HOMSTRAD (deBakker et al., 2001; Mizuguchi et al., 1998). We use solvent accessibility calculations (Lee and Richards, 1971) to estimate the availability of particular residues to be directly involved in functional interactions with co-factors, substrates, intermediates, products and other proteins. Knowledge of the 3D structure has been used to estimate the borders of the patterns. These patterns have been accumulated in a database FPSPD, accessible at the site http://www-cryst.bioc.cam.ac.uk/~fpspd/

**METHODS**

**Pattern positions**

For every HOMSTRAD family a probability matrix has been derived from the environment-dependent substitution tables (Overington et al., 1992) allowing an estimate of the probability that each amino acid could occupy a position, $E_{p,b}$, of the family alignment. Sixty-four possible local environments of amino acids are defined by the combination of three structural features: main-chain conformation/secondary structure (helix/strand/coil/positive phi torsion angle), solvent accessibility (accessible/inaccessible) and hydrogen bonding status (true or false for side- to main-chain NH/side- to main-chain CO/side- to other side-chain) (Overington et al., 1992). To obtain the probability matrices, we calculate $s(a, E \rightarrow b)$, a value related to the probability that amino acid $a$ in environment $E$ is substituted by amino acid $b$ where

$$s(a, E \rightarrow b) = \log \frac{A_{ab}^E}{\sum_c A_{ac}^E}$$

(1)

$A_{ab}^E$ is the (unnormalized) frequency of observing amino acid $a$ in environment $E$ being replaced with amino acid $b$, $q_b$ is the background probability of observing amino acid $b$ in a sequence and can be given by

$$q_b = \frac{\sum_{a \in E} A_{ab}^E}{\sum_{a \in E} \sum_{b \in E} A_{ab}^E}$$

(2)

Let $N_{p,b}$ be the score of replacing amino acid(s) at position $p$ of a structure or structural alignment with amino acid $b$. This score is used in the current version of FUGUE (Shi et al., 2001) (unpublished results). The scoring matrix for every HOMSTRAD family can be derived as

$$N_{p,b} = w_1 \sum_i [f_i \cdot s(a_{p,i}, E \rightarrow b)]$$

$$+ w_2 \sum_j [f_j \cdot s(a_{p,j}, E \rightarrow b)]$$

(3)

The two terms relate to sequences of homologous proteins with known structure ($i$) as well as those with unknown structure ($j$). The substitution tables depend on the local structural environment that is only available for known 3D structures. Nevertheless, the utilization of protein sequences where the 3D structure is unknown provides useful information. The environment $E$ for the unknown structures is considered as the average of that for the known structures at alignment position $p$. It has been found for FUGUE that weights of $w_1 = 70\%$ and $w_2 = 30\%$ for known and unknown structures, respectively give the most useful results. $a_{p,i}$ is the amino acid of the $i$-th structure at alignment position $p$ and $f_i$ is the weight for the $i$-th structure. The weighting scheme used is the Vingron and Sibbald (1993) method and is based on the dissimilarities of the sequences from their centre of gravity. The probability matrices for every HOMSTRAD family are calculated as:

$$E_{p,b} = \frac{e^{N_{p,b}} \cdot q_b}{\sum_b e^{N_{p,b}} \cdot q_b}$$

(4)

The simplest method to define conservation is the addition of the square of the probabilities, which produces higher values for the residues with high probability to be conserved. An average at each alignment position, $V_p$, has been calculated by adding the square of the probabilities of the 20 possible amino acids being found at position $p$.

$$V_p = \frac{\sum_{j=1}^{20} E_{p,j}^2}{20} \cdot \frac{2(n - g)}{2n - g}$$

(5)

The value of $V_p$ was adjusted according to the number of gaps, $g$, present at alignment position $p$, where $n$ is the number of proteins in the family alignment. To avoid differences between families these values are then transformed into percentage values, $v_p$.

$$v_p = \frac{V_p \cdot 100 \cdot l}{\sum_{p=1}^{P} V_p}$$

(6)

where $l$ is the length of the alignment.

The value of $v_p$ is calculated only when at least one protein with known structure exists and at least five proteins, including those with known and unknown structures, in the position of the alignment are available. Thus, for some families, e.g.
‘Arg_repressor_C’, \( v_p \) values are not calculated because the family has only four proteins in total.

For every HOMSTRAD family there is an evolutionary trace file (Lichtarge et al., 1996), implemented using TraceSuiteII (Innis et al., 2000), which indicates whether a residue is conserved in every position of the alignment and therefore in every partition of the evolutionary tree. TraceSuiteII is a series of algorithms that first generate a phylogenetic tree from a CLUSTAL (Tompson et al., 1997) alignment by the Kitsch (Baum, 1989) algorithm and then splits it into 2–10 partitions evenly distributed in evolutionary time. For each partition, protein sequences connected by a common node are clustered in groups. A consensus sequence is generated for each group. Finally, a trace is generated by comparing the aligned consensus sequences and each residue is classified as absolutely conserved, class-specific or neutral. ‘Class-specific’ denotes residues conserved at a position within each class but differing in identity between classes, and ‘neutral’ denotes residues not conserved at least in one of the consensus sequences.

In order to avoid differences arising from the different number of partitions used for each family, a percentage value, \( B_p \), has been obtained from the evolutionary trace file for every position in the family alignment.

\[
B_p = 100 \cdot \frac{m_p}{P} \tag{7}
\]

where \( m_p \) is the number of conservations at the position \( p \) and \( P \) is the number of partitions used for the family. Both values \( v_p \) and \( B_p \) are considered useful for identifying patterns. A multiplication of both, \( b_p \), seems to be the best solution because if one of them is small (poor conservation) the result is small, where

\[
C_p = v_p \cdot B_p / 100 \tag{8}
\]

Those residues with the highest \( C_p \) values and greater than 66 (see Results for discussion of cut-off) are considered as pattern seeds. The number of seeds considered is \( l/17 + 1 \), where \( l \) is the length of the alignment, with the condition that two different seeds that lie within 15 residues of each other in the alignment should be separated by more than 10.5 Å in space. See Results for a discussion of those cut-off values.

**Pattern length**

To determine the length of a pattern the spatial coordinates of those proteins with known structures were used. Distances between every pattern seed and the residues that lie between it and adjacent pattern seeds in the alignment were calculated. For example, in the following alignment there are three seeds Asp2, Tyr8 and Gly19. If Tyr8 is considered a pattern seed, distances between Tyr8 and the residues from Phe3 to Leu18 are calculated in order to obtain the length of the pattern centred on Tyr8.

\[
\begin{align*}
&\text{A D F S A A G Y G M N K P L Y I L G P M V} \\
&1 2 3 5 8 12 18 19
\end{align*}
\]

Residues more than 13.1 Å away from the seed are considered outside the pattern and a maximum of 20 residues are considered on each side of the seed in the alignment.

**Residue accessibility**

Solvent accessible surface area was calculated using the Lee and Richards (1971) definition with a probe sphere of radius 1.4 Å and the Richmond (1984) algorithm. The side-chain and the main-chain solvent accessibilities for every residue of each protein of known structure, together with averages per alignment position, have been calculated by

\[
A_p = \sum_{i=1}^{n_p} a_{p,i} \cdot f_i \tag{9}
\]

where \( a_{p,i} \) is the percentage solvent accessibility of the side-chain or main-chain of the amino acid \( p \) in the sequence \( i \), \( n \) the number of proteins in the alignment and \( f_i \) the sequence weighting (Vingron and Sibbald, 1993) as described for Equation (3). In this way \( A'_p \) and \( A''_p \) values for side-chains and main-chains, respectively have been calculated. In all cases all hetero-atoms were removed from the structure before applying the algorithm. Only single subunits of oligomeric structures were considered in order to expose those residues involved in subunit–subunit contact. In this way, those residues involved in subunit interactions will appear as accessible.

**Functional indices**

In order to identify those patterns which should be considered involved mainly in the function of the protein and those which are only involved in the maintenance of 3D structure, two functional indices \( i_f \) have been calculated for each pattern using Equation (10):

\[
i_f = \frac{\sum_{p=1}^{l_p} A_p/(R_p/5.0)}{l_p} \tag{10}
\]

where \( l_p \) is the length of the pattern and \( A_p \) the solvent accessibility.

From \( A'_p \), the mean accessibility of the side-chains (see Equation (9)) we calculate the ‘side-chain’ functional index \( i'_{f_p} \) and from \( A''_p \), the mean accessibility of main-chains, the ‘main-chain’ functional index \( i''_{f_p} \). \( R_p \) is the number of different residues at position \( p \) and 5.0 (see Results section) is the average number of different residues obtained from analysis of the data. As the functional indexes were calculated for the whole pattern, not only for the seed, conservation had to be included again.

**Family selection method**

In order to improve the constant values used by our method, fifteen families, each with at least two members with known structure and five members in total, were randomly selected from the HOMSTRAD database. The ‘random
method’ consisted of alphabetically sorting the proteins of the 617 families, for which patterns were found, on their names, taking the first protein family for each letter, and placing this in the appropriate cell in Table 1. If the place was already taken this family was not considered. After this process, the cell ‘all-beta/ enzyme’, from Table 2, was empty. The list of all-beta proteins from HOMSTRAD was taken in alphabetic order. The first family that clearly corresponded to an enzyme, related to surface regions in which the substitutions of residues that parameters including the length of the sequence pattern, the distance between residues in three dimensions and values for accessibility for main-chain atoms and side-chains were useful. These parameters were chosen on the bases of an analysis of PROSITE (Hofmann et al., 1999) database, which contains patterns that in the main can be considered to be functional. Table 3 shows selected statistical data from the PROSITE–HOMSTRAD alignments after applying our method and the different initial cut-offs selected for the method used in this work.

The initial cut-off values of various parameters that optimally define functional patterns were defined from 644 PROSITE patterns aligned with 472 HOMSTRAD families. We found that parameters including the length of the sequence pattern, the distance between residues in three dimensions and values for accessibility for main-chain atoms and side-chains were useful. These parameters were chosen on the bases of an analysis of PROSITE (Hofmann et al., 1999) database, which contains patterns that in the main can be considered to be functional. Table 3 shows selected statistical data from the PROSITE–HOMSTRAD alignments after applying our method and the different initial cut-offs selected for the method used in this work.

The initial cut-off values of various parameters that optimally define functional patterns were defined from 644 PROSITE patterns aligned with 472 HOMSTRAD families. We found that parameters including the length of the sequence pattern, the distance between residues in three dimensions and values for accessibility for main-chain atoms and side-chains were useful. These parameters were chosen on the bases of an analysis of PROSITE (Hofmann et al., 1999) database, which contains patterns that in the main can be considered to be functional. Table 3 shows selected statistical data from the PROSITE–HOMSTRAD alignments after applying our method and the different initial cut-offs selected for the method used in this work.

The initial cut-off values of various parameters that optimally define functional patterns were defined from 644 PROSITE patterns aligned with 472 HOMSTRAD families. We found that parameters including the length of the sequence pattern, the distance between residues in three dimensions and values for accessibility for main-chain atoms and side-chains were useful. These parameters were chosen on the bases of an analysis of PROSITE (Hofmann et al., 1999) database, which contains patterns that in the main can be considered to be functional. Table 3 shows selected statistical data from the PROSITE–HOMSTRAD alignments after applying our method and the different initial cut-offs selected for the method used in this work.

Cutoff selection

The initial cut-off values of various parameters that optimally define functional patterns were defined from 644 PROSITE patterns aligned with 472 HOMSTRAD families. We found that parameters including the length of the sequence pattern, the distance between residues in three dimensions and values for accessibility for main-chain atoms and side-chains were useful. These parameters were chosen on the bases of an analysis of PROSITE (Hofmann et al., 1999) database, which contains patterns that in the main can be considered to be functional. Table 3 shows selected statistical data from the PROSITE–HOMSTRAD alignments after applying our method and the different initial cut-offs selected for the method used in this work.

The initial cut-off values of various parameters that optimally define functional patterns were defined from 644 PROSITE patterns aligned with 472 HOMSTRAD families. We found that parameters including the length of the sequence pattern, the distance between residues in three dimensions and values for accessibility for main-chain atoms and side-chains were useful. These parameters were chosen on the bases of an analysis of PROSITE (Hofmann et al., 1999) database, which contains patterns that in the main can be considered to be functional. Table 3 shows selected statistical data from the PROSITE–HOMSTRAD alignments after applying our method and the different initial cut-offs selected for the method used in this work.

The initial cut-off values of various parameters that optimally define functional patterns were defined from 644 PROSITE patterns aligned with 472 HOMSTRAD families. We found that parameters including the length of the sequence pattern, the distance between residues in three dimensions and values for accessibility for main-chain atoms and side-chains were useful. These parameters were chosen on the bases of an analysis of PROSITE (Hofmann et al., 1999) database, which contains patterns that in the main can be considered to be functional. Table 3 shows selected statistical data from the PROSITE–HOMSTRAD alignments after applying our method and the different initial cut-offs selected for the method used in this work.

RESULTS

Strategy

Our approach depends on identifying amino acid residues where the conservation is greater than that predicted on the basis of structure alone. It exploits the development of environment-dependent substitution tables (Overington et al., 1992) to predict patterns of amino acid substitutions on the basis of known structures, and compares this to the observed patterns of substitutions in large families of homologous proteins.

The combination of observed substitutions in the sequences of known family members, which have been structurally aligned, together with a good estimate of like substitution derived from knowledge of the structure, constitutes a powerful approach to identification of residues that are likely to be conserved in other members of the family. These residues are considered seeds of the patterns. The search for these seeds was achieved using probability matrices derived from environment-dependent substitution tables and evolutionary trace analysis carried out using TraceSuiteII. These two calculations are already implemented in HOMSTRAD. Each HOMSTRAD family contains a representative set of members with known structures where the maximum amino acid sequence identity between them is 90%. The lowest average identity in HOMSTRAD corresponds to the family ‘PARATHYROID’ with 8.3%. The use of HOMSTRAD has the advantage that the sequences are aligned on the basis of structural comparisons and the database is carefully curated. Figure 1 shows the probability matrix of the Chitin binding domain (ChtBD) HOMSTRAD family obtained by Equations (1)–(4) (see Methods). Thus, we can see that the probability of finding a glycine in the fourth position of the alignment is 93% (Fig. 1).

To define the length of each pattern the geometries of those proteins with known structures were used. The requirement is that all residues of a functional pattern should be in close spatial proximity.

Those residues in contact with co-factors, substrates, intermediates, products and other proteins are assumed to be accessible. Thus, the functional patterns identified will be related to surface regions in which the substitutions of residues are restrained by binding partners, substrates, co-factors, etc. Two functional indices, $i^f_\text{avg}$ and $i^f_\text{std}$ [Equation (10)] were obtained for each pattern that define how likely it is that the pattern is involved in the function of the protein.

Cutoff selection

The initial cut-off values of various parameters that optimally define functional patterns were defined from 644 PROSITE patterns aligned with 472 HOMSTRAD families. We found that parameters including the length of the sequence pattern, the distance between residues in three dimensions and values for accessibility for main-chain atoms and side-chains were useful. These parameters were chosen on the bases of an analysis of PROSITE (Hofmann et al., 1999) database, which contains patterns that in the main can be considered to be functional. Table 3 shows selected statistical data from the PROSITE–HOMSTRAD alignments after applying our method and the different initial cut-offs selected for the method used in this work.

These PROSITE–HOMSTRAD alignments are available from the HOMSTRAD database (deBakker et al., 2001; Mizuguchi et al., 1998) by choosing the option ‘show WITH homologous sequences’ which is present for a subset of the families. If a PROSITE pattern is available, it will appear aligned with the family alignment. A measure of

---

Table 1. HOMSTRAD family names of the families used for optimizing the FPSPD database

<table>
<thead>
<tr>
<th>Enzymes</th>
<th>Protein/sugar binding proteins</th>
<th>DNA/RNA binding proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-alpha</td>
<td>Zn_dep_PLPC</td>
<td>Hemery</td>
</tr>
<tr>
<td>All-beta</td>
<td>rvp</td>
<td>B_lectin</td>
</tr>
<tr>
<td>Alpha+beta</td>
<td>ACT</td>
<td>LBP_BPI_CETP</td>
</tr>
<tr>
<td>Alpha/beta</td>
<td>KAS</td>
<td>Aabp</td>
</tr>
<tr>
<td>Small</td>
<td>C1</td>
<td>Kazal</td>
</tr>
</tbody>
</table>

Table 2. HOMSTRAD family names of the families used for benchmarking the FPSPD database

<table>
<thead>
<tr>
<th>Enzymes</th>
<th>Protein/sugar binding proteins</th>
<th>DNA/RNA binding proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-alpha</td>
<td>6PGD</td>
<td>DnaJ</td>
</tr>
<tr>
<td>All-beta</td>
<td>Dntpase</td>
<td>Jacalin</td>
</tr>
<tr>
<td>Alpha+beta</td>
<td>Extradiol_dioxy</td>
<td>Protg</td>
</tr>
<tr>
<td>Alpha/beta</td>
<td>Beta_elim_lyase</td>
<td>Tubulin</td>
</tr>
<tr>
<td>Small</td>
<td>Tautomerase</td>
<td>Kunitz</td>
</tr>
</tbody>
</table>
<p>Fig. 1. Probability matrix for the ChtBD HOMSTRAD family. The columns under seq represent the family alignment in this case of five proteins, where each column represents one protein amino acid sequence. The columns give the probability of finding a particular amino acid, given in one letter code at the top of the column.</p>

<p>Table 3. Statistics from the PROSITE-HOMSTRAD alignments for selecting the initial cut-offs</p>

<table>
<thead>
<tr>
<th></th>
<th>Average value (x)</th>
<th>SD (σ)</th>
<th>x + σ</th>
<th>x − σ</th>
<th>Cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>19.7</td>
<td>13.4</td>
<td>33.1</td>
<td>6.3</td>
<td>33 (16)</td>
</tr>
<tr>
<td>Distances</td>
<td>7.3</td>
<td>3.2</td>
<td>10.5</td>
<td>4.1</td>
<td>10.5</td>
</tr>
<tr>
<td>Accessibility SC</td>
<td>20.0</td>
<td>13.6</td>
<td>33.6</td>
<td>6.4</td>
<td>20</td>
</tr>
<tr>
<td>Accessibility MC</td>
<td>10.0</td>
<td>7.8</td>
<td>17.8</td>
<td>2.2</td>
<td>10</td>
</tr>
<tr>
<td>RP</td>
<td>4.3</td>
<td>2.1</td>
<td>6.4</td>
<td>2.2</td>
<td>4.3</td>
</tr>
</tbody>
</table>

Length' means number of residues in a pattern. 'Distance' means distances in A between every residue of a pattern and the seed. 'Accessibility SC' means mean solvent accessibility percentage of the side chain. 'Accessibility MC' means mean solvent accessibility percentage of the main-chain. RP is the number of different residues in the position p.

x < 7 = low, 7 < x < 20 = medium, 20 < x < 34 = high, x > 34 = very high.

x < 3 = low, 3 < x < 10 = medium, 10 < x < 18 = high, x > 18 = very high.

Different residues before the seed + seed + 16 after the seed = 33.

We have classified these families with respect to their fold class: all-alpha, all-beta, alpha+beta, alpha/beta and small proteins and their functional class: enzymes, protein/sugar binding proteins and DNA/RNA binding proteins. In the literature searches, 107 functional patterns were found for these 15 families. The cut-off values were modified until results were optimized for the test set.

Table 4 shows the initial and final values for all parameters used in this work. For the 15 families of the optimization set, and after the cut-off optimization, 139 patterns were obtained compared to the 107 functional patterns found in the literature. Of the 139 patterns, 18 (13%) were not found in the literature. This is expected, as information will often not be available. Indeed protein–protein interactions in general, and oligomer interactions and domain interactions in particular, are only rarely commented in detail, although in some cases comments can be found in the PDB-file. Only 10 patterns (9%) described in the bibliography as functional were not found by FPSPD. Most of these missed patterns are located at the N- and C-termini, probably due to misalignments of the sequences, which happen more often at these positions.

This analysis shows that if the side-chain functional index is lower than 7 and the main-chain functional index lower than 3, the pattern has a low probability of being involved in the function of the protein. If i_f < 7 is in between 7 and 11.2 and i_m < 34, the pattern can be considered to have a medium probability to be functional. If 11.2 < i_f < 34 and 10 < i_m < 18 and i_f < 34, the pattern can be considered to have a high probability of being functional and finally, if i_f > 34 or i_m > 18, there is a very high probability that it is involved in the function of the protein. 

The likely functionality of the pattern is provided by colour annotation, which is based on solvent accessibility and conservation. By selecting the option ‘key to PROSITE patterns’ the colour scheme used in the PROSITE pattern alignment is displayed.

It is possible to obtain the alignment of a PROSITE pattern with the HOMSTRAD family(ies), using the search server available at ‘http://www-cryst.bioc.cam.ac.uk/~ricardo/ prosite_s.html’. The output has links to the PROSITE pattern webpage(s) and the HOMSTRAD family webpage(s).

In order to improve the values of the parameters used to define the patterns, a test set of fifteen families, each with at least two members with known structure and five members in total, were randomly selected from HOMSTRAD (Table 1).
Patterns

7243 patterns were found from 617 structurally-aligned homologous families in HOMSTRAD (11.7 patterns per family). About 3584 (49.5%) of these were considered to have a very high probability of being functional, based on our functional indices (5.8 patterns/family). About 3049 (42.1%) were considered to have a high probability of being functional (4.9 p/f); but some of these are likely to participate in the function. Of the 423 (5.8%) patterns with medium functionality scores (0.7 p/f), we assume that most are not involved in function but rather involved in the maintenance of the structure of the protein. The 187 (2.6%) with low functionality scores (0.3 p/f) are probably only structural. The numbers of patterns obtained with medium and low functionality scores are very small; this is a consequence of our searching only for patterns with high conservation involving some exposed residues. Most structural patterns do not have these characteristics and they are not found in the FPSPD. The average length of the obtained patterns is 19.3 residues.

Benchmark

A further 15 families were randomly selected in order to validate the method (Table 2). In the case of the benchmark set of families, 150 patterns were obtained by FPSPD compared to the 87 functional patterns found in the literature; 25 patterns (17%) were not found in the literature; 7 patterns (8%) found in the literature were not found by FPSPD. Taking the worst results from the optimization set and the benchmark set, we can say that our method is able to find more that 90% of the functional patterns in a family of homologous proteins and that less than 17% of the patterns found by our method are not mentioned in the literature.

In order to define the behaviour of FPSPD in every type of family, results were then obtained for an example of each type of family (Fig. 2). These results show that our method can more easily find enzyme active site patterns (6.2% of the patterns missed) than protein binding patterns (11.5% of the patterns missed). This is probably because residues involved in active site interactions are more conserved than those involved in protein binding interactions. Figure 2 also makes it clear that patterns from all-alpha proteins are more easily found than those from all-beta. The all-alpha and the alpha/beta families missed only 5.7%. It can also be seen that for ‘all-beta’ proteins all patterns predicted by FPSPD as functional are mentioned in the literature while for ‘all-alpha’ more than 25% of the patterns predicted as functional are not mentioned in the literature. For ‘alpha+beta’ and ‘alpha/beta’ intermediate values were obtained. This means that FPSPD can easily overpredict functional patterns from alpha than from beta proteins.

DISCUSSION

The approach described here successfully finds patterns from the structurally aligned homologous families in HOMSTRAD. The $v_p$ values [see Equations (1)–(6)] obtained using the environment-dependent substitution tables reflect the fact that values larger than 100 imply that the conservation at that alignment position is greater than would be expected if there were no functional conservation of the residue.

The solvent accessibility calculations allow us to obtain a functional index using Equation (10), which implies that, if there is only one residue, i.e., $R_p = 1$, at position $p$ of a pattern from the HOMSTRAD alignment, the residue invariance makes this position very important for the pattern. In contrast, if several different amino acids are found at position $p$, the low level of conservation during the evolution at this position implies that it is unlikely to be important for the function or structure of the protein.

Table 4. Initial and final values for the cut-offs

<table>
<thead>
<tr>
<th>Cut-off description</th>
<th>Initial value</th>
<th>Final value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum value of $C_p$ for being considered a seed. $C_p$ is a combination of the substitution table calculations and the evolutionary trace analysis [Equation (8)].</td>
<td>60</td>
<td>66</td>
</tr>
<tr>
<td>Maximum number of seeds allow in a family = (alignment length)/$n$ + 1, where ‘$n$’ is the cut-off used here (see Pattern position in Method section)</td>
<td>25</td>
<td>17</td>
</tr>
<tr>
<td>Minimum number of amino acids in the alignment between two seeds that can be considered to belong to different patterns (see Pattern position)</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Minimum distance ($\AA$) between two seeds that can be considered to belong to different patterns (see Pattern position)</td>
<td>10.5</td>
<td>10.5</td>
</tr>
<tr>
<td>Average number of different residues in each alignment position. $R_p$ [Equation (10)]</td>
<td>4.3</td>
<td>5.0</td>
</tr>
<tr>
<td>Maximum number of amino acids in the alignment between the seed and the borders of the pattern (see Pattern length)</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>Maximum distance ($\AA$) in space between the seed and the borders of the pattern (see Pattern length)</td>
<td>10.5</td>
<td>13.1</td>
</tr>
<tr>
<td>Minimum number of amino acids in the alignment between the seed and the borders of the pattern (see Pattern length)</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>Minimum value of $C_p$ for being considered a border [Equation (8)]</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Threshold value of $r^*$ between ‘medium’ and ‘high functionality’ [Equation (10)]</td>
<td>20</td>
<td>11.2</td>
</tr>
<tr>
<td>Threshold value of $r^*$ between ‘high’ and ‘very high functionality’ [Equation (10)]</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>Threshold value of $r^*$ between ‘medium’ and ‘high functionality’ [Equation (10)]</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Threshold value of $r^*$ between ‘high’ and ‘very high functionality’ [Equation (10)]</td>
<td>18</td>
<td>18</td>
</tr>
</tbody>
</table>
Fig. 2. Results from FPSPD for each type of families in percentages. For ‘enzymes’, ‘protein/sugar binding proteins’, and ‘DNA/RNA binding proteins’ 10 families were used for each. For each ‘all-alpha’, ‘all-beta’, ‘alpha+beta’, ‘alpha/beta’ and ‘small’, 6 families were used. For the ‘optimization set’ and ‘benchmark set’ 15 families were used. And ‘all sets’ includes 30 families.

Example: rvp family
The retroviral proteinase family (rvp) from HOMSTRAD has 6 members with known and 53 members with unknown structures. The amino acid sequence identity values between the 59 members of this family are from 6.25 to 98.99% (the 90% maximum and 8.3% minimum sequence identities in HOMSTRAD is only for proteins with known structure).

For this family, three functional patterns are mentioned in the literature as being involved in peptide binding and four patterns, which include the previous three, as involved in dimerization. Figure 3 shows the residues involved in peptide binding and dimerization for five of the six members, as well as the selected patterns from the literature. Figure 4 shows the FPSPD results for the three patterns obtained for this family.

The second pattern predicted by FPSPD (Fig. 3) is also present in PROSITE (PS00141), although it extends one residue more at each edge. This pattern does not cover the phenylalanine due to the low conservation of this position. The third predicted pattern covers both those residues involved in the active site and those involved in dimerization. FPSPD did not identify the first pattern (Fig. 3) as being functional, probably because it occurs at the N-terminus where the alignment is often poorer.

Five patterns have been found using PROSITE for rvp. Four are patterns with a high probability of occurrence: two phosphorylation sites, one N-myristoylation site and one amidation site. PROSITE recognizes the second pattern of the family, Figure 3, as ‘Asp_protease eukaryotic and viral aspartyl proteases active site’. PRATT recognizes seven patterns for rvp. The only pattern not in the type X-(n,m)-X is the first one, L-x(1,3)-G-x-[ADGPST]. It appears nearly three times per sequence. PRATT predicts this pattern in the following unrelated position of the alignment of Figure 3: 39–45 which is part of one pattern found in the literature; 112–120 which is not mentioned in the literature and 131–137 which corresponds with one of the dimerization areas. The second, third and sixth patterns appear in similar areas as the first. The fourth and fifth patterns appear in five different areas. For the fourth pattern two of the areas correspond to patterns mentioned in the literature, and three areas are mentioned for the fifth pattern. The last pattern appears in areas already predicted by other PRATT patterns.

Example: P450 family
The HOMSTRAD family of cytochrome P450s includes five members with known structures (all non-eukaryotes) and 64 members with unknown structures (mainly eukaryotes). The structure of the rabbit (Williams et al., 2000) cytochrome P450 was included in HOMSTRAD after these calculations had been completed. FPSPD identified ten patterns for this family, of which two have a very high probability of being functional. One of these corresponds to the well-known motif involved in proton transfer and is positioned in the groove where the redox partner protein binds with the cytochrome P450 protein (A/G-G-X-D/E-T-T/S). The second corresponds to the motif containing the cysteine that binds the iron of the haem group and a positive residue that binds one of the carboxyl groups of the haem group (F-X-X-G-X-R-X-C-X-G). This pattern is present in PROSITE with accession number PS0086. Two of the other six patterns with a high probability of being functional contain residues that bind the haem group and two are in a position where the protein interacts with the membrane. The two remaining patterns with medium functionality scores are probably involved in maintaining the structure of the protein.
**CONCLUDING REMARKS**

The main difference between FPSPD and other protein sequence pattern databases is that FPSPD gives a measure of the pattern functionality. It is also fully automatic. As the HOMSTRAD database is updated every week, it is very easy to update FPSPD by running the series of programs that generate it. In addition, it predicts whether a pattern has high or low probability of being involved in the function of the protein or the family of proteins. This is useful for assigning a function to a new protein sequence and assists in the assignment to a family or fold class.

Functional Protein Sequence Pattern Database presents 3584 patterns that are considered functional and 3049 that are probably functional. Thus the number of functional patterns found by FPSPD (dashed bottom lines). The alignment was formatted by JOY (Mizuguchi et al., 1998).

![Fig. 3. Retroviral proteinase family alignment showing the functional residues mentioned in the literature (circles), patterns assigned for the optimization (solid top lines for patterns involved in the active site and hollow top lines for patterns involved in dimerization) and the patterns found by FPSPD (dashed bottom lines). The alignment was formatted by JOY (Mizuguchi et al., 1998).](https://academic.oup.com/bioinformatics/article-abstract/20/15/2380/233141)
Fig. 4. Patterns found by FPSPD for the retroviral protease family. REFERENCE is the id of every pattern found by FPSPD. FAMILY is the name of the family in the HOMSTRAD database. 1PATTERN and SEED VALUE are for computational use. POSITION is the position of the first residue of the pattern in the alignment from HOMSTRAD. LENGTH is the number of residues of the pattern. S.F.INDEX and M.F.INDEX are the side-chain and the main-chain functional indices, respectively. FUNCTIONAL is the functionality predicted by FPSPD for the pattern. STRUCTURES are the pdb-id of the proteins with known structure that appear in HOMSTRAD.

found by FPSPD comprises more than 90% of the total mentioned in the literature. Most of the 10% not identified are positioned at the N- and C-termini where the alignments are poorer. Although the FPSPD has generated some patterns not mentioned in the bibliography (less than 17%), many of these are likely to be involved in function. The method can find more easily enzyme active site patterns than protein binding patterns. And it finds patterns more easily from all-alpha proteins than from all-beta.

Functional Protein Sequence Pattern Database could be useful for assigning a family and thus a likely function to proteins, solely from their amino acid sequences, as well as aiding in evolutionary studies. FPSPD will be the base of a function/family recognition program that is being developed at present. FPSPD also provides potential new functional patterns that have not been mentioned previously in the literature. These new patterns could help in experimental research in areas such as protein–protein interactions, domain interactions and membrane interactions.

ACKNOWLEDGEMENTS
The author thanks Jiye Shi for his help with some programs and Prof. Tom L. Blundell, Nicholas Furnham, Simon Lovell, Kenji Mizuguchi and Lucy Stebbings for helpful advice and discussions.

REFERENCES
patterns used to detect local similarities. Bioinformatics, **16**, 358–366.


