COMMUNICATION

MANIFOLD: protein fold recognition based on secondary structure, sequence similarity and enzyme classification

Eckart Bindewald1,2, Alessandro Cestaro3, Jürgen Hesser1, Matthias Heiler4 and Silvio C.E. Tosatto1,3,5

1Computer Science V, University of Mannheim, B6, 26, D-68131 Mannheim, Germany, 2Center of Excellence in Bioinformatics, State University of New York at Buffalo, 901 Washington Street, Buffalo, NY 14203-1199, USA, 3CRIBI Biotechnology Centre, University of Padua, V.le G. Colombo 3, I-35121 Padova, Italy and 4Chair for Computer Vision, Graphics and Pattern Recognition, University of Mannheim, L13, 19, D-68131 Mannheim, Germany

*To whom correspondence should be addressed, at the address in Italy.
E-mail: silvio@cribi.unipd.it

We present a protein fold recognition method, MANIFOLD, which uses the similarity between target and template proteins in predicted secondary structure, sequence and enzyme code to predict the fold of the target protein. We developed a non-linear ranking scheme in order to combine the scores of the three different similarity measures used. For a difficult test set of proteins with very little sequence similarity, the program predicts the fold class correctly in 34% of cases. This is an over two-fold increase in accuracy compared with sequence-based methods such as PSI-BLAST or GenTHREADER, which score 13–14% correct first hits for the same test set. The functional similarity term increases the prediction accuracy by up to 3% compared with using the combination of secondary structure similarity and PSI-BLAST alone. We argue that using functional and secondary structure information can increase the fold recognition beyond sequence similarity.

Keywords: classification/enzyme code/protein fold recognition/secondary structure

Introduction

Protein fold recognition is a classification problem. Given an amino acid sequence, one wants to identify known protein structures (templates) that belong to the same fold class as the target protein. The most important feature for choosing the possible template structures is sequence similarity, expressed as the quality of the alignment between target and template sequence. If the pairwise sequence similarity between target and template sequence is higher than ~30% (the threshold depends also on the length of the sequences), they will most likely belong to the same fold class (Rost, 1999). There are, however, cases of protein sequences that have a very low sequence similarity (say <20%), with structures still belonging to the same fold class (Rost, 1999). It is the objective of fold recognition programs to identify such proteins with similar structure, even if their sequence similarity is low.

In 1991, Bowie and co-workers developed a fold recognition method; instead of scoring the compatibility of a sequence by comparing it with a sequence of known structure, the sequence was compared with the structural information of a known three-dimensional structure. This method was termed structural 3D–1D profile (Bowie et al., 1991). Since then, a variety of fold-recognition methods have been published (Godzik et al., 1992; Jones et al., 1992; Sippl and Weitckus, 1992; Bryant and Lawrence, 1993; Russel et al., 1996; Karplus et al., 1998; Jones, 1999a; Thiele et al., 1999; Fischer, 2000; Kelley et al., 2000; Rychlewski et al., 2000) and several reviews on these methods have appeared (Sternberg et al., 1999; Jones, 2000).

With the availability of sensitive sequence alignment methods, the two-step PDB-BLAST protocol (Rychlewski et al., 2000) is increasingly used as the baseline method for defining easy and hard fold recognition targets. This protocol uses PSI-BLAST (Altschul et al., 1997) to build a sequence profile from the non-redundant database of known sequences. The profile generated is then used in a second PSI-BLAST round to search for related sequences of known structure. It was shown to perform well during the LiveBench-2 experiment (Bjunicki et al., 2001). Because PDB-BLAST is used to classify target difficulty, it is safe to say that for the prediction of more difficult fold recognition targets, sequence similarity alone is unlikely to yield satisfactory results. This perception has also recently been emphasized by McGuffin and Jones in the context of discriminating ‘novel’ and ‘known’ folds (McGuffin and Jones, 2002). They suggest that the use of secondary structure alignments is significantly more accurate at identifying ‘novel’ folds than sequence-based methods. Owing to its speed it is also suitable for genome-wide analysis (McGuffin and Jones, 2002).

It is widely accepted that protein function is related to its secondary structure (Bartlett et al., 2003). Several classification schemes have been proposed for protein function. Enzymes are generally easier to classify than non-enzymes because they catalyze distinct chemical reactions, which can be easily enumerated. The enzyme code scheme (NC-IUBMB, 1992) is the best developed and most widely used functional classification scheme. Hegyi and Gerstein found statistically significant correlations between structural class and enzyme code (Hegyi and Gerstein, 1999). The two most versatile folds had 29 (TIM barrel) and 17 (αβ hydrolase fold) different (to the third level) enzyme codes associated with them. However, all other folds are restricted to five or fewer enzyme codes (Hegyi and Gerstein, 1999; Bartlett et al., 2003). This functional correlation can therefore be used in fold recognition to reduce the number of possible folds for proteins which have been previously characterized as enzymes.

Our system, called MANIFOLD (MANnhelm FOLD recognition), is a hybrid of different approaches. Instead of relying on sequence similarity alone, it also uses secondary structure and functional information to improve fold recognition. For a given target protein sequence, the program computes several different similarity scores, which represent the similarity in sequence, secondary structure and function, between the target...
protein and a template protein from the fold library. The new aspect of our system is the use of functional information about the query protein in form of an enzyme code. Also, as described below, we developed a non-linear ranking scheme in order to combine the scores of the three different similarity measures used.

Materials and methods

Our approach of protein fold recognition is to use a set of target–template similarity measures and combine them in a single score value which predicts whether the target and the template structure belong to the same fold class.

The program generates a set of features for each template structure of the fold library. In this study a vector of three features is used: (i) a score for a global alignment of the secondary structural elements of the target and template protein; (ii) the negative logarithm of the PSI-BLAST e-value; and (iii) the logarithm of our structure–function compatibility score (described below). For each feature vector a score is computed, which in effect is a prediction of whether a particular template belongs to the same fold class as the target protein.

Each feature score is an input into a small two-layer neural network; the output of the neural network is the similarity score of the fold recognition program (see Ranking method for more details).

Using enzyme codes and structure–function relationships

We use the enzyme code system (NC-IUBMB, 1992) for describing the function of a protein. It is a four-level hierarchy that classifies different aspects of chemical reactions catalyzed by enzymes. The first digit denotes the class of reaction: oxidoreductase, transferase, hydrolase, lyase, isomerase or ligase. The two subsequent digits classify the type of bond involved, cofactors and other specificities. The fourth digit identifies different substrates involved in the same general reaction. The function of non-enzymes is not modeled in this study (we represent the function of a non-enzyme using the enzyme code 0.0.0.0). Some functions have only been observed with a single fold class so far, but in some cases, proteins belonging to different fold classes can perform the same or a similar function. Given the enzyme code of a protein, what is its probability of having a certain structure?

In order to model the ‘known’ structures and functions, we prepared a data file that contains annotated enzyme codes of all non-identical CATH codes [version 2.4 (Orengo et al., 1997)]. CATH domains belonging to the same homology class (H-level, fourth hierarchy level in CATH) as any member of the test set were eliminated. This parameter-file serves as our structure–function database.

We used the supplementary data files provided by the SwissProt and the Enzyme database (Bairoch and Apweiler, 2000) in order to annotate the enzyme code of the proteins of the data sets. If the PDB code of a protein is associated with several different enzyme codes, the annotation was set to the status of an ‘unknown’ function (modeled by enzyme code number 0.0.0.0).

The structure–function score models the probability that the target protein with function $f_{\text{tar}}$ and template protein with function $f_{\text{temp}}$ belong to the same fold class. This score is computed as follows. We define an a priori probability of a protein belonging to a certain fold class to be a constant $p_{\text{min}} = 1/650$. This assumes that all fold classes are equally likely if one knows nothing about the target protein and that all fold classes are equally populated. This assumption oversimplifies the uneven distribution of different fold classes of proteins in the PDB. However, the exact value of $p_{\text{min}}$ has little influence on the overall score, because small values of the function scores are all scaled to a value close to zero with the non-linear scaling of the neural network (see Ranking method).

The enzyme codes of the template and the target are compared. If the similarity between the enzyme code is less than three identical levels, the algorithm stops and the similarity score is set to the logarithm of $p_{\text{min}}$. Otherwise, a set $\varepsilon$ of proteins with non-identical CATH codes are collected from the prepared structure–function database, which are associated with the three or four (depending on the enzyme-code similarity between target and template) common enzyme code hierarchy levels. Let $v$ be the number of non-identical CATH codes (four levels) and let $n_{\text{fold}}$ be the number of different folds (three levels for CATH) of that set. As the resulting score of the algorithm we use a convex combination between a maximum and a minimum score:

$$x_{\text{func}} = \log(p_{\text{func}}) = \log[(1 - a) \times p_{\text{min}} + a n_{\text{fold}}]$$

The factor $a$ describes how sparse the known structure–function data are for that target–template pair. We set $a = v/(v + m)$, with $v$ being the number of different CATH codes (counting four levels) found for this enzyme code. In the limit of a perfect structure–function relationship (i.e. the same enzyme code is always associated with the same fold class, many structures are known for proteins with this function), we obtain $a \to 1$, $p_{\text{func}} \to 1$. In the limit of very sparse data or very diverse structure–function relationships, we obtain $a \to 0$ or $1/n_{\text{fold}} \to 0$, hence $p_{\text{func}} \to p_{\text{min}}$. We set $m = 3$. The particular values of the constants $m$ and $p_{\text{min}}$ have little influence on the performance of the structure–function scoring term.

PSI-BLAST

We also use the e-value computed with PSI-BLAST (Altschul et al., 1997) in order to have a sequence-based similarity measure. We are searching the non-redundant sequence database provided by the NCBI. The database was processed with a filter from the PSIPRED program suite (Jones, 1999b) in order to remove low-complexity regions, trans-membrane regions and coiled-coil segments. The performed search consists of four rounds of PSI-Blast on the non-redundant ‘nr’ database with standard parameters. The resulting position-specific scoring matrix is used to search the available sequences of the template library with BLAST. This protocol, called PDB-BLAST, improves fold recognition beyond standard PSI-BLAST and was first introduced by Rychlewski et al. (Rychlewski et al., 2000). The negative logarithms of the obtained e-values are used as sequence similarity scores.

Secondary structure similarity

As secondary structure similarity measure we used the global alignment of secondary structure elements. Our algorithm is a slightly simplified version of the secondary structure alignment procedure described by McGufin et al. (McGufin et al., 2001) and Przytycka et al. (Przytycka et al., 1999).

In short, each secondary structure element is represented by a letter for the three states helix, strand or coil and a number corresponding to the length of the region. Unlike the method described by Przytycka et al. (Przytycka et al., 1999), we do...
not split secondary structural elements in order to obtain better matches.

As secondary structure prediction we use either the consensus procedure of Albrecht et al. (Albrecht et al., 2003) or the PSIPRED predictions (Jones, 1999b) for both the target and template proteins. For the dataset provided by McGuffin and Jones (McGuffin and Jones, 2002), we use the PSIPRED predictions as given in their downloadable data.

**Data sets**

As a test set for evaluating our fold recognition method, we used the protein data set of 542 non-redundant domains developed by McGuffin and Jones (McGuffin and Jones, 2002). They describe a subset of 252 ‘known’ folds, which had at least one other match in the data set, and 290 domains with folds unique with respect to the data set. From the set of unique folds we had to take out the templates with the CATH names 1GCB01, 1GCB02, 1BPLA1, 2CGPC1 because their domain definition has changed or was not part of CATH version 2.4. As targets for the fold recognition we used the set of known folds and as the template library we used the complete set of 538 domains.

In addition, we tested our method on the protein test and training set provided by Ding and Dubchak (Ding and Dubchak, 2001). This dataset is based on SCOP (Murzin et al., 1995) and consists of 386 SCOP domains for the testing set and 285 SCOP domains for the training set. We compile a structure–function database, which annotates the enzyme codes of the SCOP (version 1.53) domains from the SwissProt and the Enzyme database. We leave out all homologous proteins from that database which result in a Blast e-value of <0.001 with any protein of the protein test set defined by Ding and Dubchak.

**Ranking method**

A non-linear ranking scheme, which is in effect a simple two-layer neural network, is used.

For each feature $i$ the score $x_i$ is weighted with a non-linear function $g_i(x_i)$:

$$g_i(x_i) = a_i/\{\tanh[c_i(x_i - b_i)] + 1/2\}$$

The terms $a$, $b$ and $c$ correspond to a weight, an offset and a rescaling factor of the feature score, respectively. For a feature score $x_i << b_i$, the term $g_i(x_i)$ is close to zero; for $x_i >> b_i$ the term $g_i(x_i)$ approximates the weight $a_i$.

The total score of a template is the sum of the weighted scores of the features used:

$$S = \sum g_i(x_i)$$

**Evaluation**

In order to evaluate our fold recognition method, we generated 20 random partitions of the McGuffin and Jones dataset of ‘known’ folds into 80% training and 20% test set. For each partition, we ran a 200-step Monte Carlo optimization of the three weights $a_i$ for the three features. The parameters $b_i$ and $c_i$ are set to fixed values. For each partition and training run, the optimized weights are used to rank the proteins of the test set. If a feature is not used, its weight is set to zero during both training and testing. For the Ding and Dubchak data set, the test and training sets are fixed. Here for each feature combination we ran 20 different 200-step Monte Carlo optimizations.

**Results and discussion**

The average (and standard deviation) of the number of correct first hits (in percent) is shown in Table I for different combinations of features used for the McGuffin and Jones dataset. Table II shows analogous data for the Ding and Dubchak data set.

The alignment of secondary structure elements is the most important feature for the prediction of the fold class, because in both cases it is part of all top four feature combinations. The McGuffin and Jones dataset is difficult, because PDB-BLAST has a success rate of only ~13%. The performance of another widely used sequence-based fold recognition method, GenTHREADER (Jones, 1999a), is benchmarked at about 14% correct first hits (McGuffin and Jones, 2002). Using PDB-BLAST in addition to secondary structure alignment does not increase the prediction accuracy significantly (compare Table I, lines 3/4 and 1/2). Using the enzyme code similarity in addition to the secondary structure alignment increases the prediction accuracy.
Table III. Examples of proteins for which a correct template structure could only be found if the functional similarity term is used

<table>
<thead>
<tr>
<th>Target</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Template</td>
<td>Sec</td>
</tr>
<tr>
<td>1CEM00</td>
<td>1CLC02</td>
<td>1.50.10.30</td>
</tr>
<tr>
<td>1.50.10.10</td>
<td>3.2.1.4</td>
<td>.</td>
</tr>
<tr>
<td>3.2.1.4</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>153L00</td>
<td>119L00</td>
<td>1.10.530.40</td>
</tr>
<tr>
<td>1.10.530.10</td>
<td>3.2.1.17</td>
<td>.</td>
</tr>
<tr>
<td>3.2.1.4</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1CHD00</td>
<td>1WAB00</td>
<td>3.40.50.1110</td>
</tr>
<tr>
<td>3.40.50.180</td>
<td>3.1.1.61</td>
<td>.</td>
</tr>
<tr>
<td>3.1.1.61</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

The targets and templates correspond to CATH domains taken from the dataset described by McGuffin and Jones (McGuffin and Jones, 2002). The columns marked A (B) correspond to predictions, in which the functional similarity term is switched on (off); Target, CATH domain i.d., CATH code and EC number of the target protein; Template, top-ranking template protein; Sec, secondary structure similarity score; Func, score of functional similarity term (A only); Score, total score after non-linear weighting.

accuracy by about 2.5%, to 32.7%. Using all three features results in the highest prediction accuracy of 34 ± 0.9%. McGuffin and Jones report about 30% as highest rate of correct top predictions for the secondary structure alignment using PSI-PRED predicted secondary structure (McGuffin and Jones, 2002). They also report the theoretical upper limit for secondary structure alignment, using DSSP assignments (Kabsch and Sander, 1983) of PDB structures, at about 33% correct top hits. Our results are therefore in good agreement with those of McGuffin and Jones for the single features and indicate the value of combining the three features in MANIFOLD.

PDB-BLAST is more successful for the dataset provided by Ding and Dubchak, with about 59.1% correct first hits, because the test and training set contain some distantly homologous protein pairs. As can be seen in Table II, the highest prediction accuracy is achieved if all three similarity features are used (74.9%). An improvement in accuracy over the second best combination (secondary structure plus sequence) of nearly 1.8% is observed. These values can be directly compared with the prediction accuracy reported by Ding and Dubchak (Ding and Dubchak, 2001) for this test set. They obtained a prediction accuracy of 56% for the test and training set underlying the statistics of Table II.

A novel aspect of the present method is the use of functional information in the form of enzyme codes. The effect of this term is to reduce the number of false positives by sieving out improbable folds. In Table III we show some examples of protein targets for which a correct template is predicted with the highest score only if the functional similarity scoring term is used. The examples are taken from the McGuffin and Jones dataset. In all cases shown in Table III, the PSI-Blast e-values for the templates with the target as a query sequence are greater than the cutoff value and do not influence the ranking. The examples were chosen to highlight three different situations. The first example shows a case was a completely false prediction (wrong class) is replaced by a correct one with a perfect (four-digit) enzyme code match. In the second example, discrimination between enzymes versus non-enzymes allows one to assign the correct topology. The third example in this table is especially interesting: The target protein (CATH code 3.1.1.61) has only three digits of the enzyme code in common with the top-scoring template protein (CATH code 1WAB00, an acetylhydrolase with enzyme code 3.1.1.47). However, because of the functional similarity term, a correct template structure is predicted (three CATH levels), even though it has a slightly lower secondary structure similarity score compared with the template 1LUCA0.

The use of the enzyme code to predict the fold class of a protein is only examined for proteins for which the enzyme code is available through Swiss-Prot. For the test sets this amounted to 56% (McGuffin and Jones) and 51% (Ding and Dubchak) of the targets being annotated with an enzyme code. We have not examined the extent to which hypothetical functions based on sequence similarity can be useful to increase the accuracy of the fold recognition system.

The method, which uses a combination of secondary structure similarity and sequence similarity, is well suited for fold assignment on a genome-wide scale. Provided that the secondary structure prediction and the PDB-BLAST result of the target sequence are given, the computing time for predicting a fold from sequence using a template library of 4030 SCOP domains typically takes less than 1 min on a Pentium IV 2.0 GHz computer.

We show that by using knowledge about the function, in addition to the predicted secondary structure of a protein, it is possible to increase the fold recognition prediction accuracy beyond the values achieved by sequence alignment-based tools such as PSI-BLAST or GenTHREADER. We conclude that the secondary structure and the enzyme class of a protein contain additional information with respect to sequence similarity and this should be used extensively by fold recognition systems.

Acknowledgements

We thank Heinz Kredel for the opportunity to use the parallel computing cluster at the University of Mannheim. We thank Reinhard Manner and Giorgio Valle for supporting this work. E.B. thanks the Boehringer Ingelheim Fonds for funding his research. A.C. and S.C.E.T. are supported by Telethon (Italy), grant number B057-I. S.C.E.T. is also funded by a ‘Rientro dei cervelli’ grant from the Italian Ministry for Education, University and Research (MIUR).

References
