Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases

Anders Wallqvist1,*, Yoshifumi Fukunishi1,2, Lynne Reed Murphy1, Addi Fadel1 and Ronald M. Levy1,*

1Department of Chemistry, Rutgers University, Wright-Rieman Laboratories, 610 Taylor Rd, Piscataway, NJ 08854-8087, USA

Received on December 15, 1999; revised and accepted on May 23, 2000

Abstract

Motivation: Sequence alignment techniques have been developed into extremely powerful tools for identifying the folding families and function of proteins in newly sequenced genomes. For a sufficiently low sequence identity it is necessary to incorporate additional structural information to positively detect homologous proteins. We have carried out an extensive analysis of the effectiveness of incorporating secondary structure information directly into the alignments for fold recognition and identification of distant protein homologs. A secondary structure similarity matrix based on a database of three-dimensionally aligned proteins was first constructed. An iterative application of dynamic programming was used which incorporates linear combinations of amino acid and secondary structure sequence similarity scores. Initially, only primary sequence information is used. Subsequently contributions from secondary structure are phased in and new homologous proteins are positively identified if their scores are consistent with the predetermined error rate.

Results: We used the SCOP40 database, where only PDB sequences that have 40% homology or less are included, to calibrate homology detection by the combined amino acid and secondary structure sequence alignments. Combining predicted secondary structure with sequence information results in a 8–15% increase in homology detection within SCOP40 relative to the pairwise alignments using only amino acid sequence data at an error rate of 0.01 errors per query; a 35% increase is observed when the actual secondary structure sequences are used. Incorporating predicted secondary structure information in the analysis of six small genomes yields an improvement in the homology detection of ~20% over SSEARCH pairwise alignments, but no improvement in the total number of homologs detected over PSI-BLAST, at an error rate of 0.01 errors per query. However, because the pairwise alignments based on combinations of amino acid and secondary structure similarity are different from those produced by PSI-BLAST and the error rates can be calibrated, it is possible to combine the results of both searches. An additional 25% relative improvement in the number of genes identified at an error rate of 0.01 is observed when the data is pooled in this way. Similarly for the SCOP40 dataset, PSI-BLAST detected 15% of all possible homologs, whereas the pooled results increased the total number of homologs detected to 19%. These results are compared with recent reports of homology detection using sequence profiling methods.

Availability: Secondary structure alignment homepage at http://lutece.rutgers.edu/ssas

Contact: anders@rutchem.rutgers.edu; ronlevy@lutece.rutgers.edu

Supplementary Information: Genome sequence/structure alignment results at http://lutece.rutgers.edu/ss_fold_predictions.

Introduction

Computational biology’s most useful and widely employed contribution to science is the ability to recognize and match DNA and protein structures as well as sequences based on sequence data alone (Henikoff, 1996; Holm and Sander, 1996; Brenner et al., 1997). It is estimated that well over 80% of our biological knowledge concerning protein sequences is inferred from homology (George et al., 1996). With the advent of rapid sequencing and the capability of constructing entire genomes for organisms, protein sequence information has increased

988 © Oxford University Press 2000
much more rapidly than the three-dimensional structural information. Consequently there is considerable interest in the development of improved computational tools (Taylor, 1986; Barton and Sternberg, 1987; Russell and Barton, 1992; Saqi et al., 1992; Holm and Sander, 1993; Vingron and Waterman, 1994; Bryant and Altschul, 1995; Wilmanns and Eisenberg, 1995; Rost et al., 1997; Park et al., 1997; Altschul et al., 1997; Karplus et al., 1998; Park et al., 1998; Geetha et al., 1999; Grundy and Bailey, 1999) to identify the functions and structures of newly sequenced genes (Fischer and Eisenberg, 1997; Rychlewski et al., 1998; Teichmann et al., 1998; Wolf et al., 1999). When the sequence identity of a new protein compared with known protein sequences falls below a threshold value (commonly referred to as the ‘twilight-zone’, Vogt et al., 1995), additional information must be brought to bear on the problem—or it will be necessary to solve the full three-dimensional structure of the new protein.

The development of algorithms to identify folding families and functions of gene products using less information than the full three-dimensional structure of the target protein is desirable because experimentally determining the three-dimensional structure is both time consuming and costly. Precluding a full three-dimensional structure determination, NMR chemical shifts can be used to give secondary structure information directly (Ayers et al., 1999). There is a large repository of protein chemical shifts (Seavey et al., 1991; Wishart and Nip, 1998) and backbone chemical shifts can often be assigned even in large proteins (where it is difficult to accurately determine all the interproton distances).

There are a range of computational techniques that can be used to identify protein functions of newly sequenced genes, which stand between sequence alignments on the one hand, and the complete three-dimensional structure determination on the other. A promising technique is the use of protein secondary structure information to detect similar folds and functions. Sequences which are distantly related to each other but which have similar functions, tend to have highly conserved patterns of secondary structure (Russell and Barton, 1994).

In attempting to align sequences using information about secondary structure assignments, it is assumed that the sequential arrangement of secondary structure elements along the sequence is correlated with the three-dimensional arrangement of the secondary structure (Smith-Brown et al., 1993) and therefore highly correlated with the protein folding families. There are two approaches to including secondary structure information in the analysis of aligned sequences, either using it as primary information (Sheridan et al., 1985; Russell et al., 1996; Di Francesco et al., 1997a,b; Aurora and Rose, 1998) or including it as additional constraints in a general fold recognition scheme primarily based on amino acid alignments (Fischel-Ghodsian et al., 1990; Liithy et al., 1991; Fischer and Eisenberg, 1996; Alexandrov et al., 1996; Rice and Eisenberg, 1997; Rice et al., 1997; Rost et al., 1997; Jaroszewski et al., 1998). The subject of this paper is the analysis of an approach to protein fold recognition based on augmenting amino acid alignments with secondary structures represented as strings of letters corresponding to the secondary structure designation of each residue in the sequence (sequence/structure alignment). Our work differs from that of previous studies in our use of an iterative search of the template database while systematically varying the weights on the sequence and structure dependent terms, and in the extent of the statistical analysis of fold prediction reliability which allows us to calibrate the method using a large and relevant database.

The probability of identifying the correct structural fold of a query sequence depends on many features of the underlying database used to test the recognition method, including the basis for the proposed clustering of the proteins into families, and the comprehensiveness of the database with respect to both the total number of clusters (families) and the distribution of proteins across clusters (Park et al., 1997). In this work we use the SCOP40 database (Murzin et al., 1995; Brenner et al., 1998) which includes a representative sample of all protein structures in the Protein Data Bank (PDB) (Abola et al., 1987). This database is filtered to remove all homologous sequences which have a similarity above 40%. This ensures that we are using only low homology sequences while at the same time the database is large enough to ensure that the detection of false positives remains a challenging problem.

In this paper the additional information contained in the secondary structure is used to identify protein folding families and compare the results with those based on standard amino acid alignment methods to determine the maximum amount of information that can possibly be derived from the secondary structure assignments in a fold recognition test. We note that the present analysis of fold detection is similar in spirit to recent reports based on one-dimensional pattern matching using both amino acid and secondary structure sequence information (Fischer and Eisenberg, 1996; Russell et al., 1996; Di Francesco et al., 1997a,b; Rice and Eisenberg, 1997; Rost et al., 1997; Aurora and Rose, 1998). Our results are compared directly with those of Fischer and Eisenberg (1996) and Rice and Eisenberg (1997). The method presented here differs from previous work in the use of an iterative search of a structural database to identify protein folds and the construction of the secondary structure similarity matrix, and in the calibration of the reliability of both the sequence/structure alignment method and the amino acid sequence only alignment methods using the SCOP40 clustered database. The calibration step is necessary in
order to correlate expectation values calculated from the alignments with errors in homology assignments and gives an error rate that can be used to compare different homology detection methods (Brenner et al., 1998; Gerstein and Levitt, 1998; Park et al., 1998). Using the idea that the error rate in fold detection can be calibrated, we accumulate folds detected with the same reliability by iteratively searching through a database, varying the weights on the sequence and structure information with each iteration. Calibration with SCOP40 establishes a benchmark test and shows the effects of different weights, ranging between 0 and 100%, of the amino acid sequence and secondary structure terms used in the alignment. Homology detection within the SCOP40 database is investigated using both the true and predicted secondary structures. The performance of the sequence/structure alignment method is then compared to that of PSI-BLAST on the SCOP40 database. Because we calibrate both methods to be at the same error rate, we are able to combine the sequence/structure and PSI-BLAST results to produce increased coverage relative to either method alone.

We further tested the effects of incorporating secondary structure sequence information into homology detection in six bacterial genomes and compared the results with several popular alignment procedures. At an error rate of 1% we were able to assign ~40% more sequences to folding families than could be assigned using FASTA (Smith and Waterman, 1981; Pearson, 1991) or BLASTP (Altschul et al., 1997) and 20% more assignments than SSEARCH (Smith and Waterman, 1981; Pearson, 1991). A comparable number of sequences were assigned to folding families using PSI-BLAST alignments (Altschul et al., 1997) as with the sequence/structure alignment method. However, because the sequence/structure alignment and PSI-BLAST procedures produce different alignments and the error rates can be calibrated, it is possible to combine the results of both searches. An additional 25% improvement in the number of genes identified at an error rate of 0.01 is observed when the data is pooled in this way.

Materials and methods

Databases used in homology detection

The classification of proteins into groups of proteins with similar structure and/or function is central to making the connection between sequences and structural families. We have used the SCOP classification scheme (Murzin et al., 1995) to define homology of the proteins in the Protein Data Bank (PDB) (Abola et al., 1987). SCOP classifies protein domains based on class, fold, superfamilies, families, and domains. Homologous proteins (those thought to have arisen from a common evolutionary ancestor) are grouped together at the superfamily level, i.e. the class, fold, and superfamly of two sequences coincide. In order to concentrate on distantly related proteins it is desirable to remove sequences which are closely related (Brenner et al., 1998). The SCOP40 database is a subset of SCOP in which sequence pairs have less than or equal to 40% amino acid sequence identity. SCOP40 was formed by Brenner and co-workers by first sorting all SCOP domains by the quality of their structure (resolution) and making a list. The best structure was taken for inclusion in SCOP40 and removed from the list, and domains of greater than 40% sequence identity to it were discarded. The process was then repeated until the list was empty. For SCOP this resulted in a final representative sample of low homology sequences from the PDB of 1434 protein sequences (SCOP40 release 1.37) containing a total of 8022 ordered pairs of homologs. The SCOP40 dataset can be downloaded from the SCOP web-server (http://scop.mrc-lmb.cam.ac.uk/scop) or directly from the Sequence and Structure Searching Site (http://sss.berkeley.edu). A copy of the data can also be obtained directly from us.

For fold recognition in genomes, the open reading frames (ORF) from the following genomes were analyzed, Mycoplasma genitalium (MG) (Fraser et al., 1995), Treponema pallidium (TP) (Fraser et al., 1998), Methanothermus jannaschii (MJ) (Bult et al., 1996), Borrelia burgdorferi (BB) (Fraser et al., 1997), Haemophilus influenzae (HI) (Fleischmann et al., 1995), and Helicobacter pylori (HP) (Tomb et al., 1997). This sequence data was downloaded from the TIGR web-server (http://www.tigr.org). For the fold detection within these genomes we searched the non-redundant PDB sequence database (PDBAA). This database contained at the time of investigation 5569 sequences and represents all non-redundant structures deposited in the PDB.

Secondary structure sequences

For the known structures we collected the secondary structure information from the DSSP (Kabsch and Sander, 1983) library, available from the DSSP WEB-server (http://www.sander.embl-heidelberg.de/dssp). For sequences in the SCOP database that are based on domains, the corresponding secondary structure elements were assembled according to the deposited amino acid sequence. The true secondary structure was only assigned in the cases that full atomic coordinates had been deposited in the PDB. Structures consisting of only Cα atoms were assigned a secondary structure according to the DEFINE_S program (Richards and Kundrot, 1988). In the cases where the secondary structure is not known, or we want to gauge the effect of predicting the secondary structure, we used the PREDATOR algorithm (Frishman and Argos, 1996, 1997a,b) to convert an amino acid sequence to a secondary structure sequence. For the 29 sequences that could not be processed with PREDATOR because of minimal length requirements, we used the DSC procedure.
In order to predict the secondary structure each sequence was initially aligned with the non-redundant sequence database at NCBI (340,186 sequence in December 1998) using FASTA (Pearson, 1991) and homologous sequences were passed on to the PREDATOR/DSC program. For the secondary structure sequences predicted with DSC a multiple sequence alignment using CLUSTALW (Thompson et al., 1994) was initially performed. PREDATOR contains an option to copy the secondary structure assignment directly from the PDB database if the query sequence is found in its database of PDB chains with less than 30% sequence identity. With the option of using database sequences in the PREDATOR program turned off the overall accuracy of the secondary structure prediction was 68.5%, with the individual helix (H), sheet (E) and loop (L) predictions at 67.8, 49.9 and 78.3%, respectively. These predictions are in agreement with the recent benchmarks by Cuff and Barton (1999) on the PREDATOR program in which they report an accuracy of 68.6% for a carefully selected set of 396 proteins. Using the option to include database sequences in the prediction program gave an enhanced overall accuracy of 77% with predictions at 76, 68 and 83% for the H, E, and L elements respectively. This option was chosen for the prediction of the secondary structure elements of the genomes studied.

**Secondary structure similarity matrix**

We have based the evaluation of the secondary structure similarity matrix on the 3D.ali data bank collated by Pascarella and Argos (1992) and Pascarella et al. (1996). This database contains 455 proteins arranged in 86 structural families and was downloaded from the EMBL web-server (http://www.embl-heidelberg.de/argos/ali/ali.html). Each protein group contains two or more aligned structures. The proteins themselves are collected from the PDB (Abola et al., 1987). Both x-ray and NMR structures are included among the 455 proteins. The secondary structure elements were assigned by Pascarella and Argos according to the definitions of Kabsch and Sander (1983). Only three distinct secondary structure elements were retained for our analysis: helical segments (H) including the regular α-helix as well as the 3_{10}-helix, β-sheets (E); all other elements were placed by us in a ‘loop’ category (L), regardless of assigned structure.

The similarity matrix reflects the probability of occurrence of secondary structural elements paired in the three-dimensional alignment, e.g. how often a helical residue is aligned with another helical residue. The probability of occurrence is calculated as outlined by Henikoff and Henikoff (1992) for the calculation of similarity matrices for amino acids. The probability of finding paired structural elements i and j in an alignment of sequence A and B is denoted $P_{ij}$. The similarity matrix elements are normalized by the probability of finding the same pair in a randomly aligned sequence $P_{ij}^{ss}$,

$$M_{ij}^{ss} = 2 \log_2 \left( P_{ij} / P_{ij}^{ss} \right).$$

The matrix elements are thus a measure of how often such a pairing occurs relative to the random case. A positive value of the matrix element, $M_{ij}$, indicates a favorable pairing; a negative value indicates that the structural elements are unlikely to be found together.

### Table 1. The derived secondary structure similarity matrix $M_{ij}^{ss}$.

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>E</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-15</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>-4</td>
<td>-4</td>
<td>2</td>
</tr>
</tbody>
</table>

The similarity score, $w$, between two aligned sequences A and B is then formulated as

$$w_{\alpha\beta} = \sum_k m_{ab} \left( \alpha M_{\alpha k, \beta k}^{aa} + \beta M_{\alpha k, \beta k}^{ss} \right) + N_{\alpha} g_{\alpha} + N_{\beta} g_{\beta}.$$  

$M^{aa}$ corresponds to the amino acid similarity matrix (BLOSUM 50) (Henikoff and Henikoff, 1992), and $M^{ss}$ corresponds to the secondary structure similarity matrix defined above. $m_{ab}$ is the number of paired elements in the alignment between sequences A and B. $\alpha$ and $\beta$ determine the weighted importance of the amino acid and secondary structure sequence respectively. In the expression above equation (2) $a_k, b_k$ denotes the kth amino acid pair or secondary structure element pair of the aligned sequences.
\[ A \text{ and } B \text{. The number of gap openings } N_{p} \text{ is multiplied} \]
\[ \text{by the gap opening penalty } g_{o}, \text{ and the number of gap} \]
\[ \text{elongations } N_{e} \text{ is multiplied by the gap extension penalty} \]
\[ g_{e}. \]

**Score evaluation and quality of homology detection**

In order to evaluate the accuracy of an alignment we need to evaluate how many non-homologous sequence pairs have the same or better score. The non-homologous scores generated from the alignment are distributed according to the extreme value distribution (Karlin and Altschul, 1990; Altschul et al., 1994; Pearson, 1995, 1996, 1998). This provides us with a consistent way of determining how many errors we are expected to make when investigating a database of size \( N_{\text{dbs}} \). The expectation value, \( E \) or e-value, is the total number of non-homologous sequences in the database which have the same or better score,

\[ E(N_{\text{dbs}}, x) = N_{\text{dbs}} P(x), \tag{3} \]

where \( P(x) \) is the normalized probability of finding a non-homologous sequence pair with the same or higher score \( x \) drawn randomly from the extreme value distribution. If a query sequence is aligned against a database of sequences and one alignment has a score that results in an e-value of 1, we can expect by chance that within this database there is at most one non-homologous sequence pairing that has a score that is at least as large as this alignment score.

In order to evaluate an alignment method for detecting homologous proteins we need to determine how many homologous pairs can be detected at a specified error rate. This is accomplished using the following quantities obtained from the alignment: \( M_{\text{true}}^{\text{pos}}(m) \), \( M_{\text{true}}^{\text{neg}}(m) \), \( M_{\text{false}}^{\text{pos}}(m) \), \( M_{\text{false}}^{\text{neg}}(m) \). The superscripts true/false refer to whether or not the pair of sequences are homologous, and the subscripts pos/neg refer to the correct/incorrect identification of the homologs with the given method. These quantities are then used to calculate the coverage, specificity, and errors per query, which give a measure of how many of the total homologs are detected and the reliability of the detection.

Let \( N_{\text{dbs}} \) be the number of sequences in a given database; some of these sequences are homologous to other sequences in the database. The homologies can be established independent of any sequence alignment algorithm, i.e. from structural and functional characteristics. Let \( M_{\text{hom}} \) be the number of true homologous pairs.

In evaluating the results from the alignment we count the number of protein pairs that are actually homologous which we have detected; they are the true positive homologs, \( M_{\text{true}}^{\text{pos}}(m) \), where \( m \) is a measure such as an alignment score or expectation value. Likewise, we count the number of protein pairs that are correctly identified as not homologous; they are the true negative homologs, \( M_{\text{true}}^{\text{neg}}(m) \). One can also determine false positives and false negatives. The coverage, i.e. the fraction of true homologs detected as a function of \( m \), is the number of true homolog pairs detected \( M_{\text{true}}^{\text{pos}}(m) \) divided by the total number of homologous pairs in the database,

\[ \text{coverage}(m) = \frac{M_{\text{true}}^{\text{pos}}(m)}{M_{\text{hom}}} \tag{4} \]
\[ = \frac{M_{\text{true}}^{\text{pos}}(m)}{(M_{\text{true}}^{\text{true}}(m) + M_{\text{false}}^{\text{neg}}(m))}. \tag{5} \]

The coverage is thus parametrically dependent on the measure \( m \) that we choose to use to select homologous proteins from alignment results. The goal is to have as high a coverage with as little error as possible for a given measure \( m \) to detect the homologs.

The error in homology detection made as we try to increase the coverage can be quantified in two different ways: (1) how many errors are made for each query against a database of sequences (Brenner et al., 1998); and (2) the fraction of true homologs of all assigned homologs (specificity) (Rice and Eisenberg, 1997). These quantities are defined as,

\[ \text{errors per query}(m) = \frac{M_{\text{false}}^{\text{pos}}(m)}{N_{\text{query}}} \tag{6} \]
\[ \text{specificity}(m) = \frac{M_{\text{true}}^{\text{true}}(m)}{(M_{\text{true}}^{\text{true}}(m) + M_{\text{false}}^{\text{true}}(m))}, \tag{7} \]

where \( N_{\text{query}} \) is the number of query sequences submitted for alignment. Thus the errors per query (EPQ) should be as low as possible while having as large as possible coverage, while specificity should be as high as possible for a given coverage.

The errors per query (EPQ) gives information about what fraction of the putative homologs identified by alignments are false. Thus, choosing a threshold value \( m \) to identify homologs which have been calibrated to achieve an EPQ equal to 0.01, if the database is queried with 1000 sequences, a total of ten false positive answers are expected. The specificity, or fraction of all the pairs which are identified as homologs which are true homologs, is used to determine the confidence in the results of the alignment, thus a specificity of 0.90 indicates that 90% of all alignments with a score greater than or equal to \( m \) returned from the alignments are true. For the databases investigated in this work we have chosen an e-value threshold to give us a homology detection error rate of 0.01 errors per query.

**Calibration with the SCOP database**

Using the SCOP classification of proteins (Murzin et al., 1995; Brenner et al., 1998) as a benchmark we can evaluate the different alignment methods, i.e. since we know the ‘true’ homologs based on the classification,
we can explicitly calculate which sequence relations are correctly identified by the procedure (Park et al., 1997; Brenner et al., 1998; Gerstein and Levitt, 1998). We have used the SCOP40 1.37 dataset to calculate the errors per query accumulated during the ‘all-against-all’ alignments of the sequences in the SCOP40 database using both PSI-BLAST and the sequence/structure alignment method as a function of stated e-values. The PSI-BLAST runs were performed using the entire non-redundant sequence database in excess of 300,000 entries to construct position specific scoring matrices. By aligning all sequences of the SCOP40 dataset for different e-values and calculating the errors per query we can calibrate the e-values with EPQ. Adopting a value of 0.01 errors per query in this work, the expectation value threshold used in the sequence/structure alignment method using variable weights on the amino acid sequences for these datasets (at high specificity), beyond what can already be determined using the amino acid sequences for these datasets (at high specificity).

Fig. 1. Homology detection as a function of specificity for either pure amino acid sequence or pure secondary structure sequence alignments. In this graph we compare the SCOP40 (Brenner et al., 1998) database and the database assembled by Rice and Eisenberg (1997). Our calculations based on the SCOP40 database employed a range of expectation values to parameterize the coverage versus specificity curve, whereas the data taken from Rice and Eisenberg employed z-scores. To interpret the graph we can look at a point for a particular expectation value or z-score that corresponds to a specificity of 90% and a coverage of 20%, this implies that there is a 90% probability that a sequence pair with score greater than this expectation value or z-score is homologous; however, 80% of the homologous pairs have scores less than this and would fail to be detected at this threshold.

In Figure 1 we compare the coverage/specificity for two different databases using either amino acid sequences or pure secondary structure sequences. We have calculated all possible alignments for the SCOP40 database and have also included the published result reported by Rice et al. (1997) in the figure. The results of the amino acid alignments carried out using the SCOP40 database shown in Figure 1 correspond closely to the results presented by Brenner et al. (1998), these authors compared several different scoring schemes and statistical measures for constructing coverage/specificity plots.

For specificities greater than 20%, the coverage of the SCOP40 database using amino acid sequence alignments is always higher than for the secondary structure alignments (see Figure 1). This indicates that in order to find homologs with high specificity when searching large databases, recognition based on the alignment of secondary structure sequence patterns alone is not effective. None of the alignments using only secondary structure sequences provided new information about homologs beyond what can already be determined using the amino acid sequences for these datasets (at high specificity).
Even though it may be true that the organization of secondary structure elements in space determines the fold of a protein, the matching of these elements between sequence families when represented as a one-dimensional string is not sufficient to differentiate between sequence families any better than amino acid sequences. This is due in part to the use of a three letter secondary structure sequence alphabet versus a twenty letter amino acid alphabet. We have investigated the behavior of fold recognition in the SCOP40 database using a number of amino acid alphabets of reduced sizes (Murphy et al., 2000), and we found that an alphabet reduced to three letters retains almost no coverage relative to the twenty letter alphabet. Interestingly however, the three letter alphabet contains more coverage of SCOP40 than the reduced three letter amino acid alphabet.

Rice et al. (1997) have published an analysis of coverage/specificity plots for different one-dimensional alignment models containing varying amounts of structural information. They include in their work coverage/specificity plots for 'pure' amino acid sequence alignments and 'pure' secondary structure sequence alignments separately, and their results can be directly compared with ours as shown in Figure 1. Our results are significantly different from theirs. Using a protein fold database of their own construction based on the SCOP classification, they find that at very low specificities (e.g. 0.1), the coverage has already decreased to very low values (∼0.1) when the database is searched for homologies using either amino acid sequence comparison or secondary structure sequence comparison as the search tool. Thus, they report a much steeper degradation of coverage with increasing specificity than we observe for either alignment scheme, amino acid sequence or secondary structure sequence alignment. The differences are likely due to differences in the underlying protein databases used to test the fold recognition algorithms; apparently the Rice and Eisenberg database contains many more distant homologs as a fraction of the total when compared with SCOP40.

This points to the essential role of the protein sequence databases used to construct the coverage/specificity plots in the assessment of the apparent accuracy of the query method. The data in Figure 1 provide an illustration of the dependence of coverage/specificity plots on such features of the underlying protein database as: the total number of entries, the number of protein families, and the clustering among and within families, i.e. the distribution of cluster sizes and the filtering of high sequence identity pairs within the clusters. With the availability of a large clustered database representative of the full PDB such as SCOP40, more robust tests of homology detection methods are possible.

Synergy between amino acid sequence and secondary structure information

It has been recognized for some time that secondary structure information can be useful as an adjunct to sequence data for aligning sequences and for fold recognition. This has lead to the construction of expanded similarity matrices which incorporate information about secondary structure propensities of amino acids in the alignment (Fischel-Ghodsian et al., 1990; Fischer and Eisenberg, 1996; Alexandrov et al., 1996; Rice and Eisenberg, 1997; Rice et al., 1997; Rost et al., 1997). Aligning sequences by both their amino acid similarity and secondary structure similarity, separately and in combination, makes it possible to analyze the synergistic effects of these two alignment procedures for fold recognition.

The sequence/structure alignment technique employed here consists of running eleven different sets of alignments for each query with evenly spaced weights (pairs of values of α and β; see Section Materials and methods) on amino acid and secondary structure sequences, starting from a 100% weight on the amino acid sequence. The output from these alignments is then scanned for results which have e-values below a given threshold, and the unique pairs are collected as the final output. Only novel homologies arising from the use of secondary structure information in the alignment are gathered once the first alignment pass, based only on the amino acid sequence, is carried out.

The procedure used to calibrate and compare the fold detection results based on mixing amino acid and secondary structure sequence information in different proportions relies on first determining the error rate as a function of expectation value and then collating results at the desired error rate. We have constructed plots of coverage versus error rates for different values of α and β in equation (2). The resultant errors per query versus coverage (parametrically dependent on the expectation values) is shown in Figure 2 for some selected weights. Figure 3 displays the individual and the cumulative coverage of detected homologs as a function of increasing secondary structure weight for a given error per query fixed at 0.01. This information provides a way of estimating the synergistic effects of combining amino acid and secondary structure information when aligning sequences.

Fold detection based on the use of amino acid sequence similarities is clearly better than the results obtained using only secondary structure similarities, but an examination of Figures 2 and 3 reveals that when both amino acid and secondary structure sequence information is included in appropriate amounts it is possible to identify homologous proteins for sequences that could not be assigned homologs based on the alignments of amino acid sequences. In addition Figure 3 displays the results obtained when using predicted secondary structure instead
of the true structure. In this case the coverage is a decreasing function of the amount of secondary structure incorporated into the alignments. Using only the predicted secondary structure, and with an e-value threshold set so that the error per query is 0.01, there is no homology detection for pure secondary structure alignments beyond those detected by aligning amino acid sequences.

Because the alignments depend on both the amino acid and secondary structure content of each sequence, different homologs are detected as the weights on the sequence versus structural data in the similarity score are changed. If the datasets are scanned with different combinations of amino acid and secondary structure weights in the alignments, the unique homologs detected at the same fixed error rate can be collected cumulatively. This is shown in Figure 3; there is an increase of the total coverage as more weight is added to the secondary structure similarity score (when weights of more than 50% secondary structure information are used the increase in the cumulative coverage is negligible). For the sequences that were aligned using weighted combinations of amino acid sequence and secondary structure data based on the true secondary structure sequence, it is possible to collect additional homologs that were not detected by the amino acid sequence alone. For example at an errors per query rate of 0.01, there is 13% coverage using sequence data alone, compared with 17% coverage when additional structural data is used. This corresponds to a 35% relative enhancement of fold detection when the structural data is used together with the sequence data. In contrast there is only a 8–15% relative increase in the detection of homologs in this database when the predicted secondary structure is used instead of the true secondary structure in the alignments. The lower estimate reflects the use of a small secondary structure database in conjunction with PREDATOR, see Section Materials and methods. These results suggest that currently the full potential of the sequence/structure alignment method cannot be achieved without additional experimental information to augment predicted secondary structure (Ayers et al., 1999; Geetha et al., 1999).

The results of all-against-all alignments of the SCOP40 database using the PSI-BLAST method are indicated
in Figure 3 in order to compare that method with the sequence/structure alignment method. At an EPQ level of 0.01 there is a 15% coverage using PSI-BLAST. PSI-BLAST performs better than pairwise amino acid alignments (13%) and the sequence/structure alignment method using predicted secondary structure (14%), but the sequence/structure method performs better than PSI-BLAST when the true secondary structure is used (17%). Because the sequence/structure alignment and PSI-BLAST methods produce different alignments and the error has been calibrated, the results of both searches can be combined to yield the total number of unique homologs detected. This combined result is also indicated on the figure; the combined result (19%) gives increased coverage over that of either method alone.

To illustrate how the use of secondary structure helps in the identification of homologs with low amino acid sequence identity, Figure 4 shows two hemoglobins (PDB codes 1ash and 2lhb), represented as ribbon diagrams with segments of alignments between the two sequences indicated in the figure. Both proteins are all-α-helical globins containing six helices which function in the storage and transport of oxygen through binding of a heme group. 1ash is Ascaris hemoglobin (domain one) from the pig roundworm with SCOP 1.37 classification 1.1.1.34, and 2lhb is obtained from the sea lamprey with SCOP 1.37 classification 1.1.1.33 (Murzin et al., 1995). Due to the low sequence identity (15%), amino acid alignment does not identify the sequences as homologs at an error level of 0.01 EPQ, although visual examination of the structures shows their obvious similarity. The sequences do have high secondary structure identity (83%), and when 50% predicted secondary structure is used in the alignments they are correctly identified as homologs with the AA/SS method. PSI-BLAST misses this pair of homologs at 0.01 EPQ. At this stringent error level there are 85 ordered pairs of sequences detected as homologs using 50% secondary structure; two of these are pairs of distant homologs (sequence identity equal to or less than 15%), with the remaining pairs having between 15 and 40% sequence identity. The two pairs of domains with equal to or less than 15% amino acid identity in the SCOP40 database that are identified as homologous at the 50% secondary structure iteration are the above example and the DNA clamp proteins proliferating cell nuclear antigen (1plq region 127–258) and DNA polymerase III beta subunit (2pol region 245–366). PSI-BLAST does not identify either of these homologous pairs.

The SCOP40 benchmark test can be compared with the results of Fischer and Eisenberg (1996). In this test set 68 probe sequences were used against a library of 301 known target structures with a maximum sequence identity of 30% between probes and target sequences. Fold detection was studied for a variety of similarity matrices, including ones containing secondary structure information. The SCOP40 dataset contains at least one member of each protein superfamily as defined by SCOP—8022 homologous pairs are contained in this dataset among a total of more than two million sequence pairs. Whereas the general trend of improved fold recognition by incorporation of structural data is similar between the two datasets, around +35% for top ranked scores, there is lower fold recognition within the SCOP40 benchmark based on secondary structure alone than the Fisher and Eisenberg benchmark set. This discrepancy is a reflection of the very much larger number of non-homologous sequences that have to be discriminated amongst in the SCOP40 dataset, and the increased likelihood of sequences having similar secondary structure patterns in different homologous superfamilies. Other recent studies of homology detection (Jaroszewski et al., 1998; Ayers et al., 1999; Geetha et al., 1999) using secondary structure information find similarly that secondary structure augments fold recognition, although they employ much smaller datasets than studied here.

Geetha et al. have also recently reported a comparison of protein sequence-based methods with predicted secondary structure-based methods for identifying remote homologs (Geetha et al., 1999). They compared existing sequence comparison methods, including local amino acid sequence similarity by BLASTP, and hidden Markov models (HMMs) of sequences of protein families, with HMMs based on amino acid sequence motifs and secondary structure motifs. The test set was relatively small, consisting of 45 proteins from nine structural families in the CATH database (Orengo et al., 1997). These authors find, similar to our results, that pure secondary structure pattern recognition does not improve upon pure amino acid sequence based homolog detection overall (they did not study the effects of combining the amino acid and secondary structure pattern recognition within the same search). For the most remote homologs, however, with sequence identities less than 15% they did observe a clear advantage to using secondary structure to identify homologs, when the actual secondary structures are used for the pattern recognition. Interestingly, they also observed significant degradation in coverage and specificity when predicted secondary structure patterns are substituted for the actual patterns.

We have used the detection of homologs within the SCOP40 database to calibrate the combined sequence/structure alignment method on a clustered database of proteins with low sequence identity that is representative of the entire PDB. In the next section, homology detection via sequence/structure alignment is compared with other alignment procedures for a more practical fold recognition test, i.e. finding structural homologs to all sequences in six complete bacterial genomes.
Fold recognition in genome databases using secondary structure information

The sequences from six small complete genomes were used to search for structural homologs in the non-redundant sequence database of all PDB structures (PDBAA). Predicted secondary structure sequences were used together with amino acid sequence information to search for structural homologs in the non-redundant sequence database. The genomes investigated range from 479 predicted open reading frames (ORF) of MG to 1771 ORF of MJ, representing a small selection of different bacterial and archaebacterial organisms (Fraser et al., 1995; Fleishmann et al., 1995; Bult et al., 1996; Fraser et al., 1997; Tomb et al., 1997; Fraser et al., 1998). For the purposes discussed here a genome sequence is defined to be homologous to a database sequence when at least one sequence in the database has an expectation value that is lower than a threshold value. The results for amino acid sequence searches using four different popular alignment programs (BLASTP, PSI-BLAST, Altschul et al., 1997, and FASTA, SSEARCH, Smith and Waterman, 1981; Pearson, 1991) and for the sequence/structure alignment technique described here are summarized in Table 2.

The number of structural homologs identified in these genomes is relatively small at the very low error rate of 0.01, ranging from about 15–30%, indicating that a large fraction of these proteins cannot be positively identified as related to sequences in the PDB when a stringent cutoff is applied. The use of sequences together with predicted secondary structure data significantly increases the number of structural homologs detected compared to other pairwise alignment methods such as BLASTP, FASTA, and SSEARCH—the relative increase ranges between 12 and 51% for different genomes, with an increase of 21% when compared with the results of SSEARCH averaged over all the genomes.

The superiority of the PSI-BLAST method as compared to pairwise amino acid alignment methods lies in its use...
of a position-specific scoring matrix constructed from multiple sequence information. In Figure 5 the number of structural homologs detected in the six genomes by sequence/structure alignment and by PSI-BLAST searches are compared. The number of homologs detected by the sequence/structure alignment and PSI-BLAST methods is very similar. However, because the expectation values obtained from the alignments using these methods have been calibrated using the SCOP40 database, it is possible to combine the unique homologs assigned using sequence/structure alignment with those assigned using PSI-BLAST. When the results are pooled, the average increase in sequences assigned for the genomes shown in Figure 5 is about 20% above those assigned using either sequence/structure alignment or PSI-BLAST alone.

Several groups have reported fold predictions for the MG genome (Fischer and Eisenberg, 1997; Bork et al., 1998; Huynene et al., 1998; Rychlewski et al., 1998; Teichmann et al., 1998). It is difficult to directly compare these results because of the different definitions of ‘significance’ used by the authors when assigning homologs, as well as differences in the sizes of the sequence and structure databases that were searched for homologs in these studies. In an early report of fold predictions in MG by profiling, Fischer and Eisenberg reported the functional identification of 22% of the genes in this genome. Subsequently, Huynene et al. (1998) used an iterative PSI-BLAST search in combination with several filters to predict the function for at least one domain in 37% of the genes. Rychlewski et al. (1998) reported positive identification of 38% of all genes in the MG genome using a profiling algorithm and position-specific probability distributions derived from homologous sequences. Teichmann et al. (1998) used multiple sequence alignment programs to match all or substantial portions of 191 sequences, to yield an overall prediction rate of 41% of the proteins in the genome. The cutoff value used in their PSI-BLAST searches was set to $1.0 \times 10^{-5}$ from a calibration of an edited version of the SCOP95 dataset to remove those sequences that are likely to produce false matches with low e-values.

Summarizing the results in Table 2, the sequence/structure alignment technique identified a total of 1990 sequences at an error rate of 0.01 in the six genomes, as compared with 1657 sequences identified by the best pairwise
alignment method, and 2051 sequences identified by PSI-BLAST at the same error rate. Comparing the sequence/structure alignment method with other pairwise sequence alignment procedures, there was a 40% increase on average in new fold designations over BLASTP and FASTA, and a 20% increase over SSEARCH results. When the results using the PSI-BLAST search are pooled with those based on sequence/structure alignment, an additional 466 structural homologs are detected at the same error rate based on the same sequence information. This results in an over 50% increase on average in fold recognition over the best pairwise alignment techniques, and a 25% increase on average over either PSI-BLAST or sequence/structure alignments used alone. The combined PSI-BLAST and sequence/structure procedure yields a fold assignment for an average of 32% of all genes in the six genomes studied at an error rate of 0.01 errors per query. The complete dataset is available from the URL: http://lutece.rutgers.edu/ss_fold_predictions.

Concluding remarks
Sequence alignment techniques provide an extraordinarily powerful way to find structurally and functionally homologous proteins using sequence data alone. Advances in the methodology, including multiple sequence alignments, will make it possible to detect homology relations among proteins over even greater evolutionary distances than is possible today. Even so, there exists a far larger number of structurally homologous proteins than have been detected using current sequence alignment algorithms. By one recent estimate (Brenner et al., 1998) over 60% of the structural homologs have diverged to the point where evolutionary relationships can no longer be detected through amino acid sequence comparisons alone. Thus significant effort is being devoted to the incorporation of additional structural, biological, and chemical information into 1D sequence comparisons—this is the basis of so called ‘sequence profiling’ methods (Bowie et al., 1991; Rice and Eisenberg, 1997). The incorporation of secondary structure information into sequence profiles is a natural idea that has been exploited by several groups (Fischer and Eisenberg, 1996; Russell et al., 1996; Di Francesco et al., 1997a,b; Rost et al., 1997; Aurora and Rose, 1998), but the question of how far it is possible to go towards predicting protein structural and functional homologies based on secondary structure sequence comparisons has received less attention (Geetha et al., 1999). Through a better understanding of how well homologs may be detected using secondary structure sequence information to supplement the information contained in the amino acid sequence, we hope to develop protocols for exploiting this information in two ways—by using experimental NMR secondary structure information to detect homologs through secondary structure sequence comparisons, and by further improvements in the use of secondary structure information as a component of sequence/structure alignment.

In order to critically assess the performance of homolog detection by sequence or sequence/structure alignment methods and make a comparison of the different homology detection methods, it is necessary to calibrate the error level at which the alignments are performed. The calibration step is necessary in order to correlate expectation values calculated from the alignments with errors in homology assignments and gives an error rate that can be used to compare different methods. Calibration of the alignment methods also requires using a protein database for which the homologies are already known. We used a large representative database that covers all protein structures known in the PDB and which has filtered out highly similar sequences. The SCOP40 database (Murzin et al., 1995; Brenner et al., 1998) used in the present study provides a good benchmark test for homology detection using a combination of amino acid and secondary structure sequence information. The identification of protein families based on the addition of secondary structure information to amino acid sequence in sequence alignments provides extra information beyond what can be achieved using amino acid sequence alone. These fold identifications represent new information that can be used for investigating novel sequence relationships and as a starting point for refined homology modeling that directly incorporates the structural information in the alignments themselves.

Acknowledgements
This project has been supported by the National Institutes of Health Grant GM-30580 and by the High Performance Computing Project at Rutgers University.

References
Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-


