DPANN: Improved Sequence to Structure Alignments Following Fold Recognition

Astrid Reinhardt[1] **and David Eisenberg**[2]
[1]*Faint Signals Pattern Recognition, Los Angeles, California*
[2]*Howard Hughes Medical Institute, UCLA-DOE Institute for Genomics and Proteomics, UCLA, Los Angeles, California*

ABSTRACT **In fold recognition (FR) a protein sequence of unknown structure is assigned to the closest known three-dimensional (3D) fold. Although FR programs can often identify among all possible folds the one a sequence adopts, they frequently fail to align the sequence to the equivalent residue positions in that fold. Such failures frustrate the next step in structure prediction, protein model building. Hence it is desirable to improve the quality of the alignments between the sequence and the identified structure. We have used artificial neural networks (ANN) to derive a substitution matrix to create alignments between a protein sequence and a protein structure through dynamic programming (DPANN: Dynamic Programming meets Artificial Neural Networks). The matrix is based on the amino acid type and the secondary structure state of each residue. In a database of protein pairs that have the same fold but lack sequences-similarity, DPANN aligns over 30% of all sequences to the paired structure, resembling closely the structural superposition of the pair. In over half of these cases the DPANN alignment is close to the structural superposition, although the initial alignment from the step of fold recognition is not close. Conversely, the alignment created during fold recognition outperforms DPANN in only 10% of all cases. Thus application of DPANN after fold recognition leads to substantial improvements in alignment accuracy, which in turn provides more useful templates for the modeling of protein structures. In the artificial case of using actual instead of predicted secondary structures for the probe protein, over 50% of the alignments are successful. Proteins 2004; 56:528–538.** **© 2004 Wiley-Liss, Inc.**

Key words: substitution matrix; distant homology modeling; artificial neural networks

INTRODUCTION

Importance of Fold Recognition

Because of the constantly increasing number of fully sequenced genomes (for reviews see Quackenbush et al.[1] and Peterson et al.[2]) it is important to improve procedures for assigning protein sequences to their proper three-dimensional (3D) folds.[3] Inferring a protein's function from only its sequence remains a challenging problem[3–7] and some approaches require the protein structure to be known.[8–12] Three-dimensional cluster analysis for example exploits the phylogenetic information contained in a multiple sequence alignment to identify residues that vary between subgroups of a protein family, but not within the subgroup. When these residues are mapped onto the protein structure and found to cluster in special proximity, they frequently indicate the functional site of the protein.[8,13–16] But this need for structural information poses a problem, because the structures of most proteins encoded by each genome are unknown.[17–20]

This gives great incentive to improve the prediction of protein structures (for a review see Jones[21] and Skolnick et al.[22]). Homology modeling can yield useful results (for a review see Moult[23]), but the process requires the identification of a structural homolog. Community-wide experiments for the assessment of structure prediction methods (CASP) have shown that the performance of so-called fold recognition (FR) methods, which can be applied in cases where no sequence homology can be detected, have improved significantly over the last decade.[23–27] But even when FR-methods identify the true fold for a sequence, the alignment between the sequence and the identified structure can be far from the actual result yielded by structural superposition[24,26,27] often because the sequence and structure are misaligned. Even for targets eligible for homology modeling problems can arise if the sequences have diverged considerably. If the sequence identity falls below a threshold of around 40–50%, even the structural features can start to show strong variation and as much as 50% of the protein core be different.[28] Some studies have shown success in building low-resolution models based on FR-results[9,11,12] but generally it is impossible to model the structure based on the results of a misaligned fold assign-

Abbreviations: DPANN, Dynamic Programming based on Artificial Neural Networks; ANN, artificial neural network; DASEY, Directional Atomic Solvation EnergY; DAPS, Database of Aligned Protein Structures; ss_aat, amino acid type combined with secondary structure state for the residue; RMSD, root mean square deviation; 2°, secondary structure; FR, fold recognition; H, homologous protein pair related on super-family level of CATH; T, protein pair related on the common structural topology level of CATH; GAPP, gap-penalty.

*Correspondence to: David Eisenberg, Howard Hughes Medical Institute, UCLA-DOE Institute for Genomics and Proteomics, Box 951570, UCLA, Los Angeles, CA 90095-1570.
E-mail: david@mbi.ucla.edu

Received 15 October 2003; Accepted 1 February 2004

Published online 28 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20144

ment, which presents a severe limitation to the usefulness of these methods.

Thus there is a need to find the optimal alignment of each sequence to the known structure it resembles most closely. Commonly used alignment methods use the dynamic programming algorithm, which depends on a substitution matrix to identify the mathematically optimal alignment between two sequences[29] or a sequence to a structure.[30] These matrices are commonly log-odds tables derived from a database of existing alignments.[31] Procedures that rely on the similarity between sequences naturally perform less well for distantly related protein pairs, which are exactly the ones which are of the greatest interest in the structural mining of genomes.[32–35] In an attempt to improve sequence-to-structure alignments, Jaroszewski et al. looked at how well alignments perform in a structural sense that are sub-optimal in the sense of a sequence-to-sequence alignment. They found that sub-optimal alignments based on standard matrices exist, which are improvements over the alignments created in FR. However, there are usually hundreds to thousands of sub-optimal alignments that achieve a better score than even the structural alignment, making it difficult to identify which is the best possible solution.[36] Also, in order to enumerate the sub-optimal alignments, a number of assumptions have to be made, some of which are oversimplifications, such as not allowing insertions in secondary structure elements (SSE),[36] which can prevent the best possible solution from emerging. Other approaches involve extensive Monte-Carlo simulations of lattice-models and are consequently time-intensive.[37,38]

In short, substitution matrices based on amino acid substitution preferences alone perform decreasingly well as the similarity between the two sequences diminishes.[39,40] Consequently we use a substitution matrix based on both amino acid type and secondary structure states (ss-aat-states), based on the observation that protein structure is better conserved than sequence. In order to make as few assumptions as possible, we train an artificial neural network (ANN) on real structural alignments versus decoys and then use the resulting weights of the trained ANN as the substitution matrix.

MATERIAL AND METHODS

We derive a substitution matrix for aligning a sequence to the structure of a similar protein, as identified by fold recognition, from the weights of a two-layer artificial neural network (ANN). We distinguish three secondary structure states (helix, strand, and coil), leading to 60 different ss-aat-types (20 amino acids times three structural states). The "GAP" is introduced as an additional state, bringing the total count to 61 possible ss-aats. The network was trained to distinguish between structural superpositions and decoys based on the fraction of all possible ss-aat-matches and -mismatches occurring in the alignments.

Dataset

We created our dataset as a subset of the in-house Database of Aligned Protein Structures (DAPS),[41] which contains structural alignments of CATH-domains[42] deemed fold-related. CATH is a hierarchically-structured database that classifies how proteins of known 3D structure are related to each other. There are four main levels of similarity: the highest, most general level is the fold class (all-alpha, all-beta, or mixed alpha/beta), followed by the level of same architecture (secondary structure elements have the same orientation). The third level is that of same topology (the secondary structure elements do not only share the same orientations, but they are also connected to each other in the same way). Proteins of the same topology may or may not have a common ancestor. This is followed by the fourth level, which groups together homologous proteins on a super-family (SF) level, which are believed to have evolved from a common ancestor. There are further levels below the fourth, describing increasingly closer relationships, such as family members, which are easily detected as homolog on a sequence level. There are no pairs of family members in our DPANN-dataset, i.e., the closest relationship between aligned proteins is belonging to the same super-family. Furthermore the sequence identity between sequences in a pair is restricted to between 0 and 25%, with 67% of pairs sharing between 5 and 13% sequence identity. Around 30% of the pairs share the same super-family level and 70% are related only by sharing the same topology. Pairs for which the root mean square deviation (RMSD) of the structural superposition of their C_α-atoms exceeds 6 Å are excluded, as are pairs with more than 50% of the alignment consisting of gaps. The entire data set consists of 9921 structural alignments and the full list can be downloaded from the following location http://www.doe-mbi.ucla.edu/Services/DPANN/Supplimentals/DPANN-Downloads.html

The Decoys

The decoys were created based on the fold recognition of the DPANN-dataset. To create a decoy set, the "aligned" ss-aats from each structural alignment were randomly reassigned. One decoy was created from each structural alignment, resulting in the same number of positive and negative examples for the ANN training process. The decoy set is identical to the positive set in all key parameters (i.e., the amino acid and secondary structure composition of both sequences, as well as the number of gaps). Figure 1 illustrates the reshuffling process. Because the ANN encounters only the fraction of matches and mismatches between the various ss-aats, it is not necessary to conserve sequential information.

Independent Dataset

Even though we took a number of precautions to ascertain the networks will not be able to memorize specific information about the training-set, we also evaluated the performance on a completely independent dataset. This dataset was derived from a later release of DAPS, based on the CATH 2.5 release. We assembled this dataset by choosing only protein pairs related either on the T- or H-level, which had a pairwise sequence identity of no more than 25%. We also excluded pairs that were not at least 50% aligned or whose structural superpositions exceeded

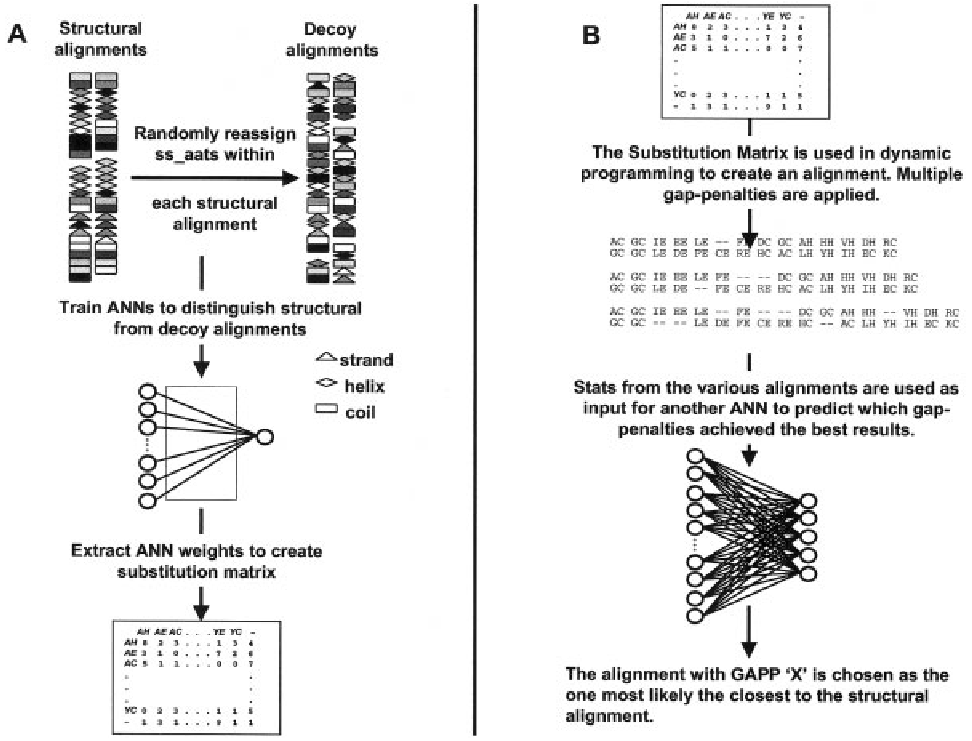
A
Structural alignments
Decoy alignments
Randomly reassign ss_aats within each structural alignment
Train ANNs to distinguish structural from decoy alignments
strand
helix
coil
Extract ANN weights to create substitution matrix
AH AE AC . . . YE YC -
B
AH AE AC . . . YE YC -
The Substitution Matrix is used in dynamic programming to create an alignment. Multiple gap-penalties are applied.
AC GC IE EE LE -- FE DC GC AH HH VH DH RC
GC GC LE DE FE CE RE HC AC LH YH IH EC KC
AC GC IE EE LE FE -- -- DC GC AH HH VH DH RC
GC GC LE DE -- FE CE RE HC AC LH YH IH EC KC
AC GC IE EE LE -- FE -- -- DC GC AH HH -- VH DH RC
GC GC -- -- LE DE FE CE RE HC -- AC LH YH IH EC KC
Stats from the various alignments are used as input for another ANN to predict which gap-penalties achieved the best results.
The alignment with GAPP 'X' is chosen as the one most likely the closest to the structural alignment.

Fig. 1. **A**: Overview of how the DPANN-similarity matrix is derived. The sequences of each structural alignment are split into their component ss-aats (consisting of the amino acid type and the secondary structure state at any given position). Which ss-aat is matched to which ss-aat in the other sequence, is reassigned randomly. An Artificial Neural Network (ANN) is then trained on the fraction of matching and mismatching ss-aat combinations (1890 input nodes) to distinguish between structural alignments and decoys. The weight-matrix of the trained ANN is then used as the values of a substitution matrix. **B**: Gives an overview of how sequence and structure are aligned to each other using the substitution matrix. Using dynamic programming and five different gap-penalty-values (GAPP) ranging from 11 to 15, we generate five different alignments. Statistical data from each alignment are then presented to a further ANN, which was previously trained to identify which is the best of a range of alignments. The alignment with the highest score is then chosen as the one presumed to have the closest resemblance to the structural alignment.

an RMSD of 5 Å. To make sure there is absolutely no overlap with the training-data, we also excluded all pairs for which at least one of the sequences either had a family member or had itself been part of the training set. The final independent dataset contained 23,165 protein pairs. Forty-eight percent (48%) of these were related on the T-level, 52% were related on the H-level, which is significantly different from the 70% T-level and 30% H-level distribution of the original set. The sequence-length distribution was quite similar to the original set for H-level, but not for T-level-related pairs (58% had a shorter sequence longer than 100 residues in the original set, but only 23% fulfilled that criterion in the independent set). As DPANN performs better on H-level-related pairs and shorter alignments, we had to normalize the results in order to make them directly comparable. This was achieved by determining the performance for different groups: H-level-related pairs $\leqslant$ 100 residues, H-level-related $>$ 100 residues, T-level-related $=$ 100 residues and T-level-related $>$ 100 residues. We then scaled the performance by a factor that creates the same distribution as the original set had, thereby making the results directly comparable.

Neural Networks

Networks to derive the substitution matrix

Figure 2 provides an overview of the artificial neural network (ANN) training procedure, as well as the consecutive steps, which ultimately lead to a set of alignments. The input for the network was provided by the fraction of matches between all possible ss-aat-classes, with exception of the impossible GAP-GAP-match. Twenty (20) types of amino acids times 3 different types of secondary structure and the GAP-state equals 61 different states. For a symmetric matrix this amounts to 1890 input nodes (61 times $(61 + 1)/2 - 1$). One output node was used and the ANN was trained assigning a 1 for structural alignments and a 0 for decoys. No hidden layers were introduced, so that it was possible to use the weights from the trained ANN directly as values of a 61×61 substitution matrix. We use the PHD-predicted secondary structures (2°) states[43] to determine the ss-aat-classes. In those cases were we report on actual secondary structure (Experimental 2°), we used the DSSP-assigned states[44] to determine the ss-aat-classes.

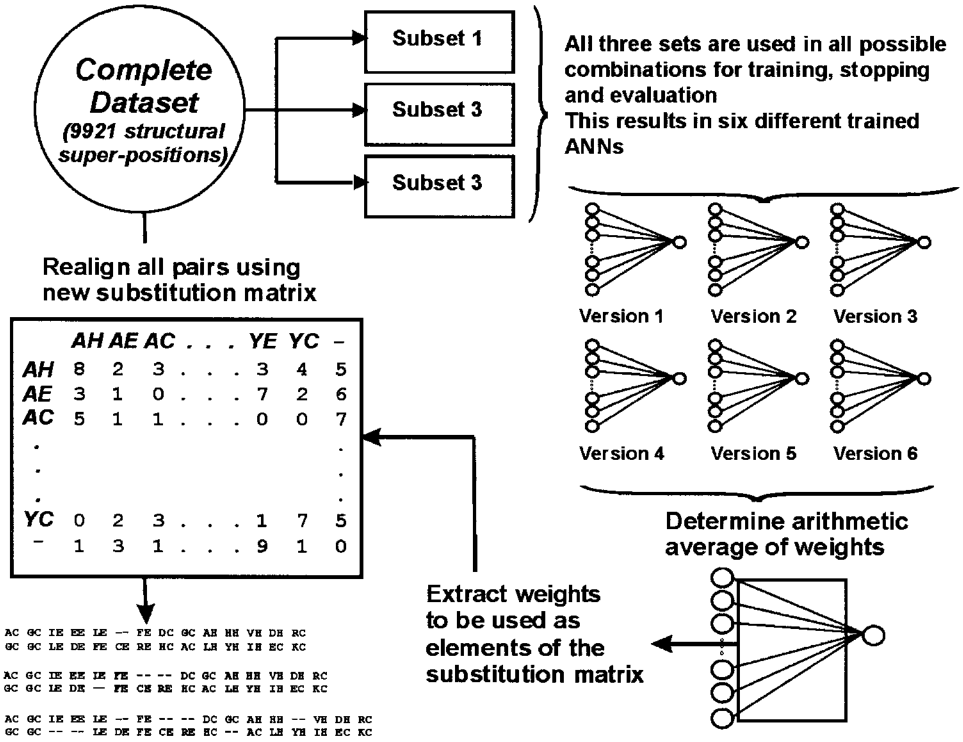
Complete Dataset
(9921 structural super-positions)
Subset 1
Subset 3
Subset 3
All three sets are used in all possible combinations for training, stopping and evaluation
This results in six different trained ANNs
Version 1
Version 2
Version 3
Version 4
Version 5
Version 6
Determine arithmetic average of weights
Extract weights to be used as elements of the substitution matrix
Realign all pairs using new substitution matrix
AH AE AC . . . YE YC –
AH 8 2 3 . . . 3 4 5
AE 3 1 0 . . . 7 2 6
AC 5 1 1 . . . 0 0 7
YC 0 2 3 . . . 1 7 5
– 1 3 1 . . . 9 1 0
AC GC IE EE LE -- FE DC GC AH HH VH DH RC
GC GC LE DE FE CE RE HC AC LH YH IH EC KC
AC GC IE EE IE FE -- -- DC GC AH HH VH DH RC
GC GC LE DE — FE CE RE HC AC LH YH IH EC KC
AC GC IE EE LE -- FE -- -- DC GC AH HH -- VH DH RC
GC GC -- -- LE DE FE CE RE HC -- AC LH YH IH EC KC

Fig. 2. Overview of the artificial neural network (ANN) training procedure. To prevent overtraining an early-stopping approach is applied. The original dataset is split into three subsets. One set is used to train the ANN and another is evaluated after every training step to determine if the performance on non-training data is still improving. Once the ANN stops improving on this test-set, training is stopped. The three sets are used in all possible combinations as training and evaluation sets, resulting in six independent ANNs. The weights of all six networks are averaged, adding additional assurance that the resulting weights will be independent of the data used for training. The averaged weights are then used to align all the pairs in the dataset.

The DPANN-training- and decoy-set were both split in three equally sized subsets, to perform training runs. In early stopping, the network is trained on one dataset, while its performance is measured on an independent dataset after every training step. The training process is stopped as soon as the performance on the second set ceases to improve. This is done to prevent over-training of the network.[45] A standard back-propagation algorithm was used to optimize the weights of the ANN. The activation function mapped the summed output of all weights onto values between 1 and 0, using the following function:

$$O_j = \frac{1}{1 + e^{(0-2\sum)}} \text{ with } \sum = \sum_{i=1}^{n} x_i{*}W_{ij}.$$

The resulting six sets of weights were then numerically averaged and these averages were used as the values for a substitution matrix. This step further ensures that the derived matrix cannot “remember” any specific examples of the dataset and will perform equally well on previously not seen data.

A standard dynamic programming approach was implemented to perform the alignments, but the ss-aat substitution matrix based on the ANN-weights was used instead of a standard amino acid substitution matrix. The ability of the algorithm to create a suitable alignment between sequence and structure depends on the value of the gap-penalty (GAPP). The best values of GAPP varied for different alignments, without a clear correlation to other parameters, such as sequence identity between the proteins or their fold class. Thus various GAPP values were applied to every alignment and key parameters of the resulting alignment were fed into an additional ANN that predicted which of the alignments was closest to the structural superposition. Figure 1(B) gives an overview of the procedure.

Network to predict the best of multiple alignments

GAPP values ranging from 11 to 15 were used to align each pair of domains, which compares to values of the substitution matrix ranging from −21.9 to +16.4. Some general and some alignment-specific parameters where then used as input into a neural network.

The following general parameters were used:

1. The number of residues of the shorter sequence (divided by 250).
2. The residue-length difference between the sequences (divided by 250).
3. The number of (predicted) secondary structure elements (SSE) in the longer sequence (divided by 30).
4. The number of (predicted) SSEs in the shorter sequence (divided by 30).
5. The difference in the number of SSEs (divided by 5).

The following alignment-dependent parameters were used:

1. Fraction of the shorter sequence aligned.
2. Length of longer sequence normalized by length of alignment.
3. The number of gaps in the alignment (divided by 250).
4. Fraction of unnecessary gaps (i.e., number of gaps minus length-difference between both sequences, divided by number of gaps).
5. Fraction sequence identity in the alignment (normalized by the length of the shorter sequence; divided by 20).

The dividers were chosen such that all input units fell between 0 and 1. As the divisions were applied to the values of all data sets equally, they did not change the rank-order or magnitude. This was done to prevent large variations in the magnitude of the input nodes, which can cause problems and delays in the training procedures for the networks.

The neural network therefore has 30 input nodes and five output nodes, where each output node stands for one of the gap-penalties (GAPP) 11–15.

The network was trained through an approach as described above and again the weights of the six resulting networks were averaged, which leads to a marginally improved accuracy in identifying the best possible GAPP. It was also found that in many cases in which the network did not identify the best possible GAPP, the GAPP value chosen created an alignment that was only insignificantly worse than the best possible one.

Determining the RMS-Deviation

In order to assess the quality of the alignments created, we needed to determine the RMSD that resulted from the structural superposition corresponding to the alignment. Given two structures and an alignment as input, the program ProFit,[46] based on the McLachlan algorithm[47] calculates the optimal RMSD based on rigid body superposition. The RMSD given is based only on the residues given as equivalent in the alignment and is therefore dependent on the number of residues aligned. This can be seen as an inversion of the structural superposition problem, where the alignment that gives the best RMSD is identified.

Evaluation Criteria

As the RMSD depends on the number of residues aligned, it alone is not a good measure of whether an alignment is successful or not. Alignments covering less than 90% of the residues that were aligned in the structural superposition were automatically deemed unsuccessful. This was done to make sure that all results are comparable and low RMSD values did not result from a small number of aligned residues. As an additional criterion we used the extent to which the sequence alignment overlaps with the structural alignment. This measure is loosely based on the one used in CASP-experiments as evaluation measure.[48] The following three criteria were used to determine whether an alignment was successful in recovering the structural superposition or not:

1. $RMSD_P - 3.5 \leq RMSD_{DAPS}$
2. $N_P \cdot 0.8 \leq N_{DAPS}$
3. $A_P^{exact} + A_P^{1off} + A_P^{2off} - A_P^{\geq 5off} - A_P^{wrong} \geq N_{DAPS} \cdot 0.5$

With N_P = number of residues aligned and N_{DAPS} = number of residues aligned in the DAPS alignment. A_P^{exact} = The number of residues aligned exactly as in the DAPS alignment. A_P^{1off} = The number of residues misaligned by exactly one,A_P^{2off} = or two residues and $A_P^{\geq 5off}$ = residues misaligned by 5 or more. A_P^{wrong} = The number of residues that were aligned, but are not aligned in the DAPS alignment. $RMSD_P$ = RMSD resulting from the superposition of residues as defined by the alignment. $RMSD_{DAPS}$ = RMSD resulting from the structural alignment in DAPS. An alignment was deemed successful if either criterion 1 applied or both criteria 2 and 3 were fulfilled.

All three criteria were determined by visual inspection of borderline cases of the structural matches resulting from certain alignments and were chosen conservatively to make sure as few as possible "false positives" are included. Consequently some viable alignments, which exceed the set parameters, are discounted.

Fig. 3. Each dot represents one of 6368 pairs of realigned proteins. Pairs for which either the DPANN alignment or the DASEY alignment covered less than 90% of the residues aligned in the structural superposition were discarded. This assures that the RMSD values are comparable and low values are not due to a small number of aligned residues. The X-axis gives the RMSD between the C_α-atoms of the pair of structures given the alignment from DPANN. The Y-axis gives the RMSD given the alignment performed by the fold-recognition method DASEY. If both methods achieve the same RMSD the dot will lie on the diagonal line (black). If the dot lies above the diagonal, the DPANN-alignment achieves a lower RMSD than the DASEY alignment and vice versa if the dot lies in the lower triangle. Red dots indicate cases in which the DASEY alignment achieved an RMSD below 6 Å and the DPANN-alignment exceeded this threshold. The green dots represent the opposite case: The DPANN-alignment achieved an RMSD below 6 Å, while the original DASEY-alignment exceeded that threshold. More dots can be found in the upper triangle, indicating that the DPANN-based alignments are more effective than the DASEY-based alignments. This greater effectiveness is evident in the area where the RMSD of the resulting alignments is small (below 6 Å) where there are a larger number of green dots than red dots.

Fig. 4. Shown are three examples of improved alignments achieved with DPANN, one for each of the major fold types (all-α, all-β and α/β). The superpositions in the first line are based on the alignments of DASEY; those in the second line are based on DPANN alignments. The red structure is in the same orientation in both cases to make the comparison easier. The superpositions are based on the alignments provided by the respective programs. They are used as input into a program, which determined the best superposition, given the alignment. This can be thought of as the opposite of performing a structural superposition, in which case the best superposition is identified and the alignment is then determined from that superposition. All of the shown DPANN-based alignments are very close to the results achieved through structural superposition, in which the equivalent secondary structure elements are correctly aligned to each other. In contrast the DASEY-based alignments frequently align the wrong secondary structure elements to each other and in some cases entire secondary structure elements are left out (see the missing long helix in the blue structure in (**c**). In (**a**), the structures based on the DASEY alignment are shifted against each other so much that the superposition results in a rotation of the structures against each other, a problem that does not occur in the superposition based on the DPANN alignment. The same is true for (**b**), an all-β structure.

RESULTS

Deriving the DPANN Substitution Matrix

We used weights derived from a trained ANN (see Materials and Methods) as the values of a substitution matrix to align a protein sequence to a known structure of a similar fold. We initially attempted to use a log-odds table as the basis for a substitution matrix, which yielded some improvements over the alignments created by the FR-method. However the log-odds derived matrix increased the number of alignments that are suitable for modeling purposes only from around 11% (for alignments coming straight out of FR) to 16% (after realigning with the log-odds matrix).

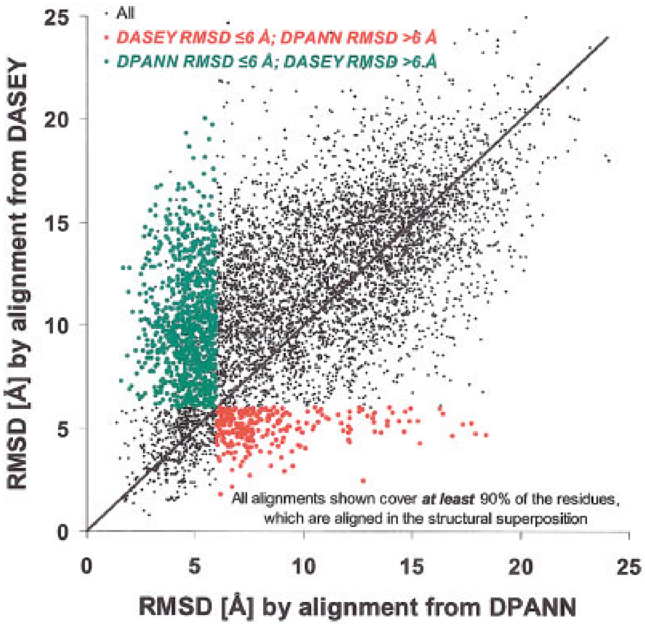

Figure 3.

The resulting weights of the ANN after training were directly used as the elements of a substitution matrix. Their values ranged from −21.9 for matching Valine in strand-state with a gap, to +16.4 for matching an Isoleucine with a Valine, both in strand-state. The Matrix is available for download at http://www.doe-mbi.ucla.edu/Services/DPANN/Supplimentals/DPANN-Downloads.html. Gap-penalties (GAPP) ranging from 8 to 15 were explored and no differentiation was made between a gap-opening- and a gap-extension-penalty. GAPP lower than 11 were found to commonly result in alignments that had less than 50% of their residues aligned and were therefore excluded. GAPP larger than 15 usually resulted in "over-alignment", i.e., structurally non-equivalent residues were aligned to avoid accumulating high penalties, resulting in bad structural superpositions.

Log-Odds Versus DPANN Substitution Matrix

To compare the matrices we normalized both to the same average and standard deviation. A color-coded visualization can be found in the supplemental material (Figure S1; http://www.doe-mbi.ucla.edu/Services/DPANN/Supplementals/MatrixComparison.html). The log-odds based matrix is considerably more structured, showing the strongest positive signals on the diagonal, i.e., for self-substitutions. Most negative values occur for amino acids

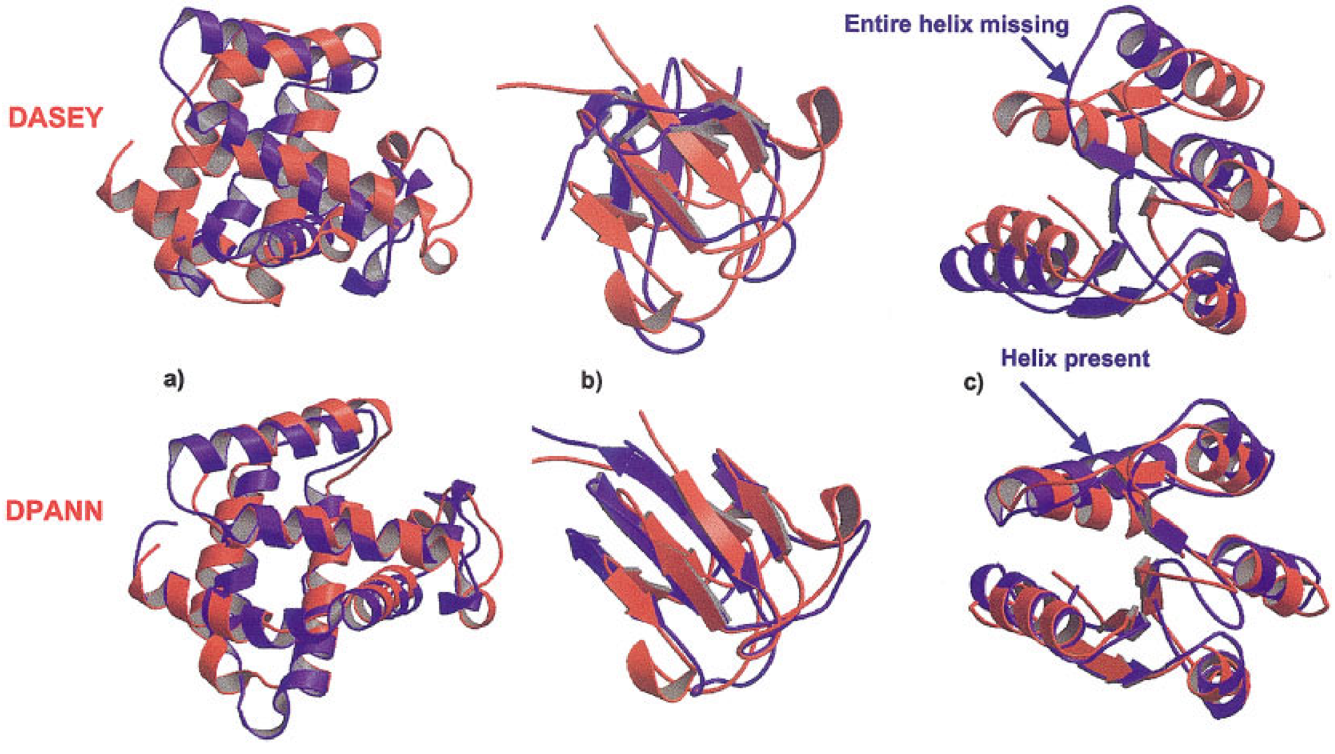

Figure 4.

in a certain structural state being replaced by any other amino acid in a different structural state. Part of this clear structure is due to the fact that many substitutions were not observed at all, in which case we assigned the lowest observed value (368 out of 1890 substitutions were not observed). A log-odd matrix derived from a considerably larger dataset might achieve better results.

The DPANN-matrix accepts most substitutions as "neutral," neither encouraging nor penalizing them, which is in line with the observation that many different sequences can result in the same structure. With exception of Methionine, the substitution of a hydrophobic amino acid by a gap is heavily penalized, while substituting hydrophilic residues with gaps is frequently encouraged. This, too, is expected, as hydrophobic residues tend to occur more frequently in the core of the protein, where insertions and deletions are more difficult to implement, while hydrophilic residues tend to occur on the surface of the protein, where loop-regions with many insertions and deletions are more common. Noteworthy also is that many substitutions in the diagonal are identified as neutral, unlike in the case of the log-odds based matrix. The neutral self-substitutions are for residues in the coil-state and the trend does not extend to hydrophobic amino acids, which always feature highly positive values for self-substitutions. In the off diagonal the strongest positive substitutions are again between hydrophobic residues substituting for each other (Isoleucine, Leucine, Valine and to a lesser extend Phenylalanine), while the most disfavored substitutions are between hydrophobic residues in either helical or strand state substituting for hydrophilic residues or Glycine in coil-state. Generally the ANN-derived matrix agrees well with the intuitive understanding of the malleability of protein structure.

DPANN Alignments Improve Over DASEY Alignments

To assess the resulting alignments we determined the RMSD between the two structures, given the alignment, based on the equivalent C_α-atoms. As this measure is dependent on the number of residues aligned, we also determined how similar the alignment is to the alignment resulting from a structural superposition, i.e., what fraction of aligned residues do both alignments have in common (for details see Materials and Methods). Alignments that covered less than 90% of residues aligned in the structural position were automatically classified as failures.

To determine the usefulness of DPANN, we plotted the RMSD of the alignment created by our fold-recognition method, DASEY, versus the RMSD of the alignment created by DPANN (see Fig. 3). Only alignments that covered at least 90% of residues aligned in the structural superposition are shown. If both methods yielded the same results, all points would lie on a diagonal. However, there is a concentration of points in the upper triangle, indicating that DPANN alignments feature lower RMSD-values than the equivalent alignments of DASEY,[41] our in-house fold-recognition program (http://fold.doe-mbi.ucla.edu/). Of particular interest here are those cases, where the DASEY-alignment would have provided a suitable modeling-base, while the realignment through DPANN does not. Setting the cutoff at an RMSD of 6 Å (red dots), we can see that this occurs rarely (just 3% of the cases). The inverted cases (the DPANN-alignment can provide a modeling-base, while the original alignment does not) are colored in green and make up around 13% of all alignments. Figure 4 shows examples of protein pairs from the three different fold classes (all-α, all-β and α/β) where the DPANN alignment (second row) was successful in improving the alignment generated by DASEY (first row).

How Successful Is the Method Compared to Structural Alignments

Figure 3 provides a rough visual measure of the comparative performances of DASEY and DPANN alignments. But it is ultimately more important how well the newly created alignment agrees with the gold standard, the structural superposition. We judged an alignment to be successful if either $\mathrm{RMSD}_{\mathrm{DPANN}} - \mathrm{RMSD}_{\mathrm{StructuralAlignment}} \leq 3.5$ Å, or if both alignments showed a reasonable overlap as determined by the fraction of correct or only slightly misaligned residues (see Materials and Methods for details). The second criterion was chosen because some alignments for proteins larger than 150 residues were found to be quite good even though they exceeded the 3.5 Å cutoff. This is usually the case when the structural superposition itself has an RMSD larger than 4 Å. The criterion was chosen conservatively and consequently some successful alignments are discounted, to minimize the number of unsuccessful alignments being misclassified as successful. Alignments that aligned less than 90% of the residues aligned in the structural superposition were automatically judged as failures. However, this excluded only a small number of pairs, as DPANN tends to over- rather than under-align protein pairs.

Applying the above criteria, we find 32% of all alignments are successful, while the rest do not meet the requirements for a successful alignment. This raises the question of why the method succeeds on some protein pairs, but fails on others. Because the success of most substitution matrices depends crucially on the sequence identity between the sequences being aligned, we investigated if this is also the case for our matrix. Figure 5 plots the fraction of pairs with a certain sequence identity (seqID) for all alignments that were classified as good (thin black line, empty squares) and there is a clear trend for alignments with lower seqID to be less accurate (a linear regression yields an R-value of around 0.96). It is also possible to split the results according to whether the pair is related on a homologous super-family (H) or only on a topology (T) level. Doing so yields the dark gray line with the black circles for the H-related pairs and the light gray line with gray circles for the T-related pairs. This reveals that H-related pairs perform generally much better than T-related pairs. Overall 62% of H-, but only 20% of T-related pairs are classified as good after realignment. As expected, the curve for all pairs approximates the H-curve for high seqID and the T-curve for low seqID, as the respective types are dominant in those areas. This ac-

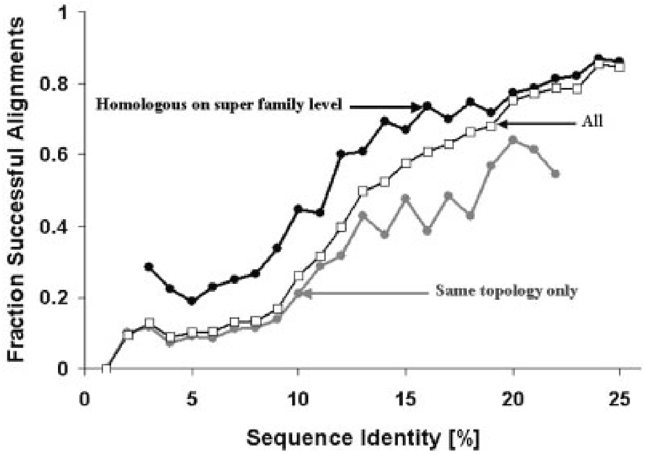

Fig. 5. The success-rate of DPANN depends on the relatedness of the sequence-pair, with homologous super-family (H)-related pairs performing better than topology (T)-related pairs. The thick black line with the open circles plots the binned sequence identity (seqID) of all pairs versus the fraction of all successful alignments in that bin (i.e., the alignment resembles closely the structural superposition). As expected, there is a correlation between low seqID and lower probability of achieving a correct alignment. The dark gray line shows the results achieved when analyzing pairs that are related only on the H-level. The light gray line shows the results for those pairs that are related only on a T-level. It can be seen that H-related pairs are more successfully aligned than T-related pairs across seqIDs. Both also show a tendency for pairs with lower seqID to yield a lower fraction of good alignments, but the trend is somewhat less pronounced than for all alignments. The reason for this is that the overall curve follows the one for H-related pairs in the range of high seqID and the curve for T-related pairs in the range of low seqID, as they are predominant in the respective ranges. Consequently the decline of the curve is steeper than that of either of its components.

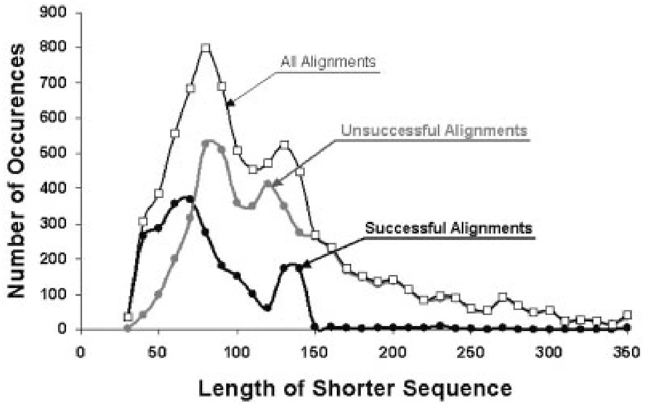

Fig. 6. The alignment performance depends on the length of the shorter sequence in the protein pair. The plot shows how many alignments have a shorter sequence with a certain length. The bin-size was ten residues. The thin black line shows the numbers for all alignments, while the thick black line shows the data only for successful alignments, i.e., alignments that resembled closely the structural superposition. The thick gray line shows the data for unsuccessful alignments. It is apparent that alignments for which the shorter sequence is smaller than 100 residues align quite well, while pairs for which the shorter sequence exceeds 150 residues are not usually aligned correctly. As the data for all alignments show, over 75% of all pairs in the set have a shorter sequence with fewer than 150 residues, indicating that most domains consist of between 50 and 150 residues.

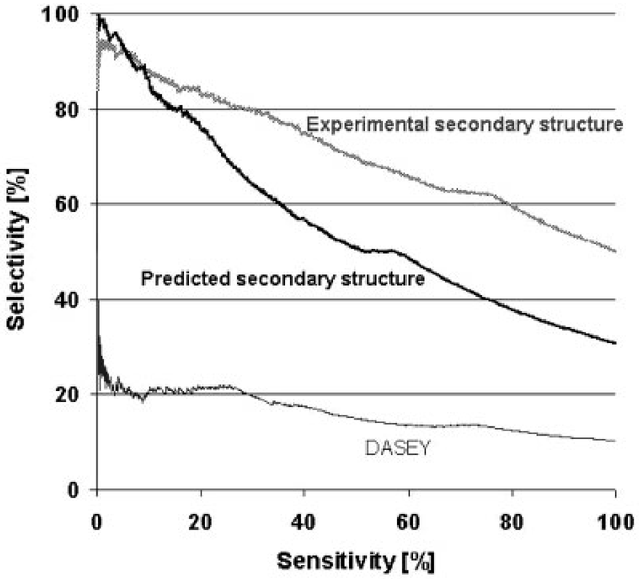

Fig. 7. Sensitivity versus selectivity plots for the length of the shorter sequence of the pair. Shown are the results achieved with the alignments as created by the fold-recognition program DASEY (thin black line), realignment with DPANN using either actual (thick gray line) or predicted secondary structure (thick black line). These plots were obtained by sorting all pairs by increasing length of the shorter sequence, then determining for every position cumulatively which fraction is categorized as well aligned (selectivity) and what fraction of all pairs have been taken into account (sensitivity). With exception of the range of extremely high selectivity (> 95%), the performance is better for the DPANN-alignments based on experimental rather than predicted secondary structure. Even the DPANN alignments based on predicted secondary structure perform between three and four times better than the DASEY alignments.

counts for the even steeper trend of the overall curve. Down to around 20% seqID (in the alignment) around 75% of all protein pairs are successfully aligned. At 13% seqID this value drops to around 50% and falls off steeply at lower seqID.

Another crucial factor for a successful alignment is the length of the shorter sequence in the alignment. Figure 6 shows that shorter alignments are far more likely to be aligned correctly, while alignments in which the shorter of both sequences is 150 residues or longer have only a very small chance of being structurally meaningful. At the same time in the vast majority of all pairs (over 75%), the shorter sequence is 140 residues or shorter as the thin black line with the open squares shows. This length-bias is likely based in the preference for protein domains to be in the range of 50–120 residues. The additional peak at around 130–140 residues contains mainly pairs with a globin fold, which are more easily aligned than other large structures, as they tend have the same number of secondary structure elements. While other folds in the same length-range (such as Jelly-rolls and Rossmann-folds) have on average between four and five secondary structure elements that do not match between both structures, the average for globins is only two. Figure 7 shows a selectivity-versus-sensitivity plot, depending on the number of aligned residues (thick, light gray line). To obtain this plot, we ordered all alignments by the length of the shorter sequence. We then determined for every point a) what fraction where successful alignments (selectivity) and b) how large a fraction of all alignments were covered at that point (sensitivity). The thick black line detailing the performance for predicted secondary structure shows that when over 57% of alignments are covered still around 50% of all alignments successfully reproduce the structural superposition (which covers alignments up to a length of 108 residues for the shorter sequence).

Upper Limits for the Expected Success-Rate Using Actual Secondary Structure

As fold-recognition aims to identify the structure closest to the unknown structure of a given sequence, the true secondary structure of the given sequence is also unknown and must be predicted. In order to determine how well the matrix performs in a real-life situation, we have to test it based on predicted structural features alone. However, knowing how well the method performs using the actual secondary structure gives the upper limit the method can achieve in its current implementation. To obtain comparable data we trained an ANN (as described in Materials and Methods) using DSSP-assigned secondary structure information instead of the predicted one. Then we choose the best alignment, instead of predicting which of the five alignments created for each pair resembles the structural superposition most closely. The lowest RMSD was the prime criterion; however, if an alignment with a higher RMSD also covered a considerably larger fraction of the shorter sequence (20%), the more completely aligned version was chosen. As expected, we observed a considerably better performance. A successful alignment was found for more sequence-pairs than before and the alignment was usually closer to the structural superposition than the one based on predicted secondary structure. Details of the results are discussed below in context of predicting which GAPP achieves the best alignment.

Influence of Predicting the Best GAPP

How does the prediction of the most suitable GAPP influence the performance? To determine this we chose the best alignment from those created using predicted secondary structure instead of predicting which one is the best with the second ANN (see Materials and Methods). As expected, the performance is better than when predicting the best GAPP. This indicates the degree of uncertainty introduced by an additional prediction step. However, the drop here is small compared to the loss of accuracy suffered through the use of predicted instead of the experimental secondary structure.

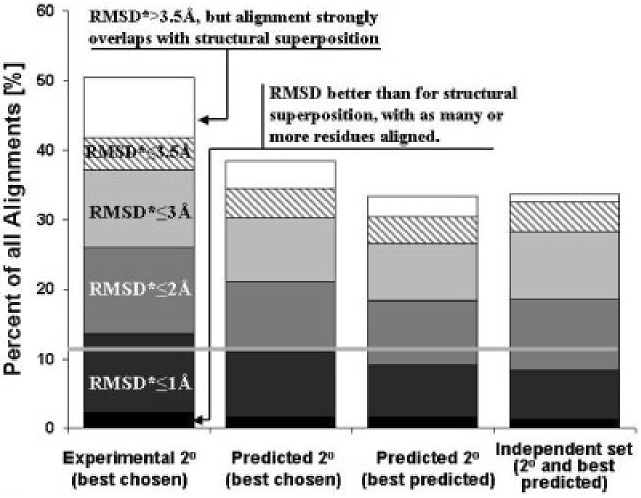

Fig. 8. Comparison of the alignment performance based on experimental and predicted secondary structure as well as a dataset compiled from a later release of CATH. The stacked columns show the percentage of pairs aligned within a given RMSD from the structural superposition. The solid black fraction represents those alignments that were within 1 Å from the structural superposition and the dark gray fraction represents those that were within 2 Å and so on. The exact definition of what was classified as good, despite exceeding the 3.5 Å threshold, can be found in the Materials and Methods section. It is mainly based on good overall overlap with the alignment obtained through structural superposition. The first column shows the results achieved with experimental secondary structure if the best alignment (out of five possible ones generated with varying gap-penalties) was chosen, rather than predicted. The second column shows the results achieved with predicted secondary structure (predicted 2°) if the best alignment is chosen, while the third column gives the results for predicted 2° if the best alignment is predicted with an additional artificial neural network. The differences observed between the first and the second column are indicative of the loss of performance due to the use of predicted 2°, while the differences between the second and third column show the performance loss based on predicting which gap-penalty (out of five tested) will perform best. The black horizontal line indicates the overall performance of the FR-based alignments for a 3.5 Å cutoff. The largest decrease in performance is due to the use of predicted instead of actual secondary structure, indicating that an improvement in the accuracy of secondary structure prediction will likely also increase the performance of this method. The last column shows the performance on a dataset that was compiled after the main study concluded and has no overlap to any data previously used. The new dataset contained a considerably larger fraction of H-level-related protein pairs, which generally perform better than T-level-related pairs, as well as more short proteins. To make the performances more directly comparable, we rescaled the results for this independent set to reflect the performance given the same composition as the original set. The observed results mirror those of the original dataset and confirm that DPANN's performance is robust and independent of the dataset used.

Figure 8 shows that using actual secondary structures, over 50% of all alignments generated successfully reproduce the structural superposition. With predicted secondary structures this value drops to 38% and once we also predict which of the five created alignments is the best, the number falls to 32%. At the same time the fraction of those alignments that are the closest to the structural alignment (with 1 Å RMSD) decreases from 13.6% (actual 2°) to 11% (predicted 2°) to 9% (predicted 2° and best GAPP predicted not chosen). A very small fraction of alignments (2% for actual and 1.5% for predicted secondary structure) outperform even the results of the structural superposition. This is deemed to be the case if the RMSD is lower than that achieved through structural superposition with at least as many residues aligned, or if the RMSD is marginally higher, but a larger number of residues were aligned. Results are also shown for the performance on an additional dataset (termed Independent dataset), which contains no sequence-pairs that were part of the main study. We used predicted secondary structure information and predicted which GAPP achieved the best result. The distribution of H- and T-level-related pairs, as well as long and short sequences was significantly different from the original dataset. In order to make the performances directly comparable, we rescaled the results to those that would have been achieved given the distribution of the original set (see Materials and Methods). Run on a much larger scale (around 23,000 pairs) of protein pairs unrelated to those in the original study, DPANN performs robustly and achieves results comparable to those seen in the evaluation phase of this study. The main difference in performance is the larger number of alignments found within 3 Å.

Figure 7 gives a good indication of the overall quality of DPANN alignments in comparison to alignments resulting from DASEY. It shows a plot of sensitivity versus selectivity based on alignments sorted by their length. When covering 58% of all alignments there is a 50% chance that the sequence and structure will be aligned correctly (up to a length of 151 residues in the shorter sequence; thick black line for predicted secondary structure information), while this selectivity is never achieved with the DASEY alignments (thin dark-gray line). At the same coverage, the selectivity using actual secondary structures is as high as 67% and does not fall to 50% until virtually all protein pairs are included (thick light-gray line). At any given point realignments achieved with predicted secondary structure (2°) perform at least three but up to four times better than the DASEY-alignments, while realignments based on actual secondary structure (Experimental 2°) perform four to five times better.

Performance of the Fischer-Eisenberg Benchmark Set

We also ran DPANN on the Fischer-Eisenberg benchmark set[49] (FE-set) for a rough comparison with current work such as John and Sali, who developed a method that uses repeated alignment-adjustment and full atom model-building and evaluation to improve the alignment quality of a subset of the FE-set.[50] We examined only those pairs of the FE-set, which have an RMSD of no more than 6 Å, which reduced the set from 68 to 49 pairs. In all cases the coverage compared to the structural superposition was at least 95% while on average it was 99%. On average 42% of all residues were aligned exactly as in the structural superposition, which compares to 45% reported by John and Sali,[50] but at the same time their average coverage is below 90. It should also be kept in mind that the approach reported by John and Sali requires much more computational resources: Refinement of one sequence pair takes around a day of computational time, while DPANN takes less than a minute for the same task.

DISCUSSION

The substitution matrix developed here leads to a threefold improvement over the alignments coming straight out of our FR-method, DASEY. In a small number of cases (< 2%) the DPANN-based alignments represent an improvement even over the structural superposition. These yield either a smaller RMSD while aligning as many or more residues or have an only marginally larger RMSD while aligning a considerably larger number of residues. As expected the use of actual versus predicted secondary structure yields consistently better results.

Our results also show that protein pairs related on a homologous super-family (H) level (but without readily recognizable sequence identity) are around three times more likely to be aligned correctly than are those related only on the same topology level (T) and this observation raises an interesting question. H-relations are defined as those where a direct family-relation cannot be identified on a sequence-level. Two sequences-families can be grouped into the same H-group only once additional structural and/or functional information demonstrates that they are actually related through a common ancestor. As a result it happens occasionally that new structures or experiments provide the information needed to group previously "unrelated" families together. Consequently the H-level is a conservative measure: Families grouped on an H-level are proven related, but families on an T-level may related on an H-level as well (although proof is missing). Given that the matrix developed here performs more than twice as well on pairs related on H-level, it is possible that the T-pairs doing well in this experiment might in fact be also related on the H-level, but no structural or experimental proof can be provided for this as yet. Perhaps the DPANN matrix can be used to identify potential H members and suggest experimental work to verify such relationships.

Frequently two H-related proteins will have too low sequence identity to be aligned with standard alignment algorithms but after they have been structurally aligned an increased sequence identity becomes apparent. This raises another question: Are these "hidden" sequence-signals between H-related members picked up during the training of the ANN? Should this be the case, it is possible that these signals overshadow any more general signal provided by same topology-only pairs, leading to a much worse performance on the T-related pairs.

While the current procedure to generate Decoy-alignments has proven to be useful, there might be even better approaches. Randomly moving the existing gaps around would be one way, to better simulate what actually occurs in a misalignment and might result in further improvements of the performance.

Scalability of the Method

The main computational effort lies in the training of the ANNs for extraction of weights and prediction of the optimal GAPP value. The alignment itself is created using a dynamic programming approach based on the substitution matrix derived from the weights of the ANN. For each query protein, five alignments have to be created, using GAPP values ranging from 11 to 15. The following prediction as to which alignment resembles the structural alignment the most is simply based on the evaluation of the ANN trained for this purpose. This makes the method highly scalable and consequently the speed of any fold-recognition method, which necessarily has to be run beforehand, is normally the speed-limiting step. The method is publicly available at the following URL: http://www.doe-mbi.ucla.edu/cgi-bin/DPANN/index.cgi.

ACKNOWLEDGMENTS

The authors thank Parag Mallick and Stuart Sievers for discussions and for the Database of Aligned Protein Structures (http://www.doe-mbi.ucla.edu/DAPS/), which served as the basis for the dataset used in this study, Gary Kleiger for discussions, and DOE BER, HMMI and NIH for support.

REFERENCES

1. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J. The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 2001;29:159–164.
2. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The comprehensive microbial resource. Nucleic Acids Res 2001;29:123–125.
3. Baxter SM, Fetrow JS. Sequence- and structure-based protein function prediction from genomic information. Curr Opin Drug Discov Devel 2001;4:291–295.
4. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nature Struct Biol 1995;2:171–178.
5. Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE. Genome annotation assessment in Drosophila melanogaster. Genome Res 2000;10:483–501.
6. Pawlowski K, Jaroszewski L, Rychlewski L, Godzik A. Sensitive sequence comparison as protein function predictor. Pac Symp Biocomput 2000:42–53.
7. Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002;318:595–608.
8. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
9. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 1998;281:949–968.
10. Zhang B, Rychlewski L, Pawlowski K, Fetrow JS, Skolnick J, Godzik A. From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. Protein Sci 1999;8:1104–1115.
11. Fetrow JS, Siew N, Di Gennaro JA, Martinez-Yamout M, Dyson HJ, Skolnick J. Genomic-scale comparison of sequence- and structure-based methods of function prediction: does structure provide additional insight? Protein Sci 2001;10:1005–1014.
12. Di Gennaro JA, Siew N, Hoffman BT, Zhang L, Skolnick J, Neilson LI, Fetrow JS. Enhanced functional annotation of protein sequences via the use of structural descriptors. J Struct Biol 2001;134:232–245.
13. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
14. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.
15. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002;18(Suppl 1):S71–7.
16. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 2003;326:255–261.
17. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
18. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
19. Buchan DW, Shepherd AJ, Lee D, Pearl FM, Rison SC, Thornton JM, Orengo CA. Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. Genome Res 2002;12:503–514.
20. Siew N, Fischer D. Twenty thousand ORFan microbial protein families for the biologist? Structure (Camb) 2003;11:7–9.
21. Jones DT. Protein structure prediction in the postgenomic era. Curr Opin Struct Biol 2000;10:371–379.
22. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotechnol 2000;18:283–287.
23. Moult J. Predicting protein three-dimensional structure. Curr Opin Biotechnol 1999;10:583–588.
24. Levitt M. Competitive assessment of protein fold recognition and alignment accuracy. Proteins 1997;Suppl 1:92–104.
25. Marchler-Bauer A, Levitt M, Bryant SH. A retrospective analysis of CASP2 threading predictions. Proteins 1997;Suppl 1:83–91.
26. Sippl MJ, Lackner P, Domingues FS, Koppensteiner WA. An attempt to analyse progress in fold recognition from CASP1 to CASP3. Proteins 1999;Suppl 3:226–230.
27. Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. Proteins 2001;Suppl 5:55–67.
28. Lesk AM. Molecular evolution. Introduction to protein architecture. First ed., Vol 1. Oxford: Oxford University Press; 2001. p 176–186.
29. Waterman MS, Vingron M. Sequence comparison significance and poisson approximation. Statistical Science 1994;9:367–381.
30. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
31. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
32. Elofsson A. A study on protein sequence alignment quality. Proteins 2002;46:330–339.
33. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol 2001;307:721–735.
34. Panchenko AR, Bryant SH. A comparison of position-specific score matrices based on sequence and structure alignments. Protein Sci 2002;11:361–370.
35. Marchler-Bauer A, Panchenko AR, Ariel N, Bryant SH. Comparison of sequence and structure alignments for protein domains. Proteins 2002;15:439–446.
36. Jaroszewski L, Li W, Godzik A. In search for more accurate alignments in the twilight zone. Protein Sci 2002;11:1702–1713.
37. Skolnick J, Kolinski A, Kihara D, Betancourt M, Rotkiewicz P, Boniecki M. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. Proteins 2001;Suppl 5:149–156.
38. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 2001;44:133–149.
39. Lassmann T, Sonnhammer EL. Quality assessment of multiple alignment programs. FEBS Lett 2002;529:126–130.
40. Jaroszewski L, Rychlewski L, Godzik A. Improving the quality of twilight-zone alignments. Protein Sci 2000;9:1487–1496.
41. Mallick P, Weiss R, Eisenberg D. The directional atomic solvation energy: an atom-based potential for the assignment of protein sequences to known folds. Proc Natl Acad Sci USA 2002;99:16041–16046.
42. Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo Conte L, Thornton JM. The CATH database provides insights into protein structure/function relationships. Nucleic Acids Res 1999;27:275–279.
43. Rost B. Phd - predicting one-dimensional protein-structure by profile-based neural networks. Methods Enzymol 1996;266:525–539.
44. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
45. Baldi P, Brunak S. Learning algorithms: miscellaneous aspects: . In: Thomas Dietterich, editor. Bioinformatics: the machine learning approach, adaptive computation and machine learning. London: The MIT Press; 1998. p 88.
46. Martin ACR. http://www.bioinf.org.uk/software/profit/.
47. McLachlan AD. Rapid comparison of protein structures. Acta Crystallogr A 1982;38:871–873.
48. Venclovas C, Zemla A, Fidelis K, Moult J. Some measures of comparative performance in the three CASPs. Proteins 1999; Suppl 3:231–237.
49. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. Pac Symp Biocomput 1996:300–318.
50. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–3992.