Identifying errors in sequence alignment to improve protein comparative modelling

Danielle Talbot[†] and Andrew C.R. Martin[*]

[†] School of Biological Sciences, University of Reading,
AMS Building, The University of Reading, Whiteknights, Reading,
Berks RG6 6AJ, United Kingdom

[*] Biomolecular Structure and Modelling Unit,
Department of Biochemistry and Molecular Biology,
University College London, Darwin Building, Gower Street,
London WC1E 6BT, United Kingdom
Tel.: +44(0)207 679 7034 Fax: +44(0)207 679 7193
EMail: andrew@bioinf.org.uk –or– martin@biochem.ucl.ac.uk

Running Title: Identifying errors in sequence alignment

Key words: alternative alignments; structure alignment; homology modelling; neural networks; machine learning

Abstract **The difference between the number of known protein sequences and the number of protein structures is vast and comparative modelling offers a way to bridge this gap. Misalignment between target and parent is the largest cause of error in comparative modelling and we define SSMAs (Sequence-Structure MisAlignments) as regions where sequence and structural alignments do not agree.**

We find that most SSMAs are short (< 10 residues) and that there is a strong preference for starting and finishing an SSMA in an unstructured region or a turn. Neural networks were trained to identify regions of sequence likely to be mis-aligned, first using single sequences to predict 'alignability' of homologues with $\leq 35\%$ sequence identity and then combining predictions for single sequences to predict SSMAs in an alignment of two sequences. Predictions of SSMAs in single sequences had positive predictive values up to 89.1% (MCC=0.798) while the alignment predictions had positive predictive values 92.9% (MCC=0.648).

In combination with a program to permute alignments, these networks were applied to comparative modelling of sequences previously submitted to CASP5. The average RMSD of these models improved by some 37% illustrating that the method is likely to be extremely valuable in improving alignment for comparative modelling.

Introduction

The difference between the number of protein sequences held in GenBank[1] and the number of protein structures held by the PDB (Protein DataBank)[2] is vast. Only recently have high throughput methods started to be put in place to solve

```
1ap2A0                          DIVMTQSPSSLTVTAGEKVTM
1igmH0 Sequence alignment       EVHLLESGGNL-VQPGGSLRL
1igmH0 Structural alignment EVHLLESG-GNLVQPGGSLRL
                                        ****
```

Figure 1: An example of an SSMA found between 1igmH0 and 1ap2A0. The SSMA is indicated with asterisks.

protein structure. Comparative modelling[3] offers a way to bridge the gap between the number of sequences and structures.

Comparative modelling generally relies on knowing the structure of a homologous protein and using that as a template to build the structure of a protein of known sequence but unknown structure. The process can be divided into seven major steps: (i) identify homologous 'parent' structures to use in the modelling, (ii) align the target sequence with the parent or parents, (iii) identify structurally conserved regions (SCRs) and structurally variable regions (SVRs), (iv) copy the SCRs from the parent structure(s), (v) build the SVRs either by database search (e.g. SLoop[4, 5]) or *ab initio* methods (e.g. CONGEN[6]), (vi) build the sidechains[7, 8, 9, 10, 11, 12], (vii) optimize (e.g. energy minimization or molecular dynamics using software such as CHARMM[13] or NAMD[14]), evaluate (e.g. using PROSA II [15]) and refine the model. Methods such as COMPOSER[16, 17, 18] and SwissModel[19, 20] automate these steps. Another popular and effective method is MODELLER[21, 22] which combines stages (iii–vi) with optimization using restraints derived from the parents. There are many other methods including 3D-JIGSAW[23], FAMS[24], ESyPred3D[25] and RAPPER[26].

However, the limiting factor in all these methods is obtaining the correct alignment. This is the most important stage of comparative modelling[27, 28], but unfortunately, particularly at low sequence identity, it can be the most difficult to get right. The sequence alignment one wishes to achieve is the alignment that would be obtained by performing a structural alignment and reading off the resulting sequence alignment. This can often differ from the alignment obtained by performing global[29] or local[30] sequence alignment.

There are numerous methods for performing structural alignment which often differ in the precise details of their results (e.g. CE[31], SSAP[32], STRUCTAL[33], DALI[34], MATRAS[35], VAST[36], SSM[37]). There are a number of different ways to superimpose two or more protein structures and, if the proteins are not identical or at least extremely similar in both sequence and structure, then there can be no single optimal superposition[38]. For our purposes, we have chosen the alignment produced by SSAP as the gold standard, 'correct' alignment.

Misalignment between a target and a parent sequence is the largest cause of error in comparative modelling. The most extreme types of misalignment (Misleading Local Sequence Alignments, MLSAs) are areas of protein alignment where structural similarity is clear and where the optimal sequence similarity is substantially higher than that seen in the structure alignment[39]. In other words, the sequence alignment for a region is very clear, yet it does not match the structure-derived alignment. We define less extreme misalignments, where the sequence and structural alignments do not agree, as SSMAs ('Sequence-Structure MisAlignments'). For example, Fig. 1 shows the sequence and structural alignment of a region from 1igmH0 and 1ap2A0 (a human and mouse antibody heavy chain variable region respectively).

In their analysis of the CASP2 comparative modelling section, Martin *et al.*[27] showed that there was a relationship between the percentage of correctly aligned

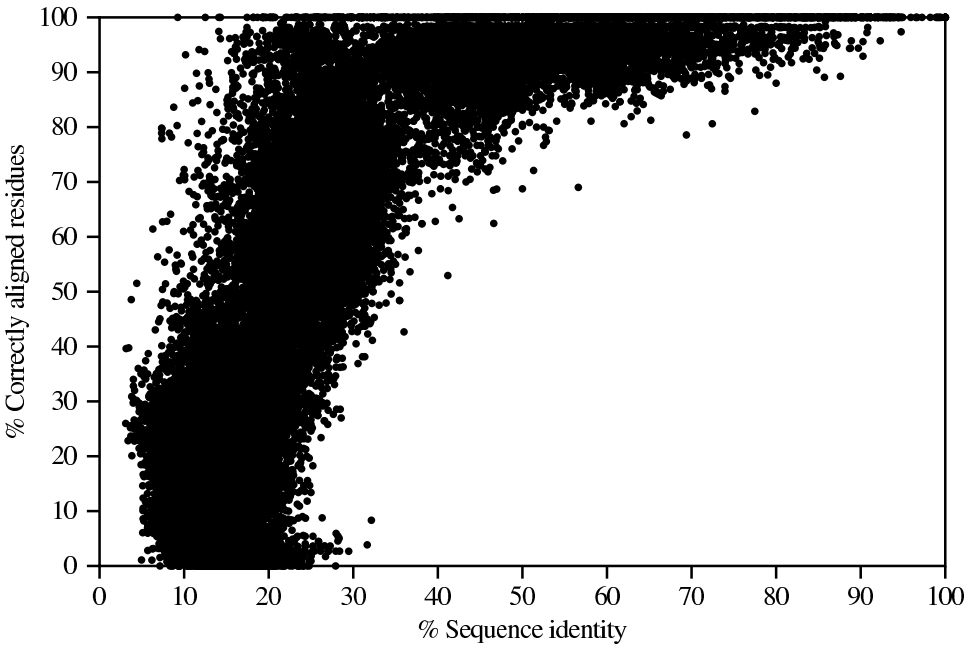

Figure 2: The relationship between percentage sequence identity and the percentage correct sequence alignment. Each pair of NReps in each CATH homologous family has been structurally aligned by SSAP and sequence aligned using a Needleman and Wunsch global alignment. The structural alignment is taken as the correct alignment. Twelve outlying points have been removed after being identified as occurring owing to errors in the CATH database.

residues and the sequence identity (Figure 2 of their paper). We have reproduced that analysis using approximately 56,000 pairs of homologous protein domains from CATH[40, 41], each of which was aligned on the basis of structure using SSAP and on sequence using a Needleman and Wunsch sequence alignment[29]. Fig. 2 clearly shows that if there is a high sequence identity between two sequences then the sequence alignment is likely to match the structural alignment. However as the sequence identity decreases, particularly below 30%, the accuracy of the alignment decreases and can be completely different from the structural alignment. If we can predict regions where mis-alignment occurs then we can hope to improve the alignment in these regions and therefore improve the model.

Materials and Methods

We used a dataset derived from the CATH[40, 41] database. Thus all analysis is performed at the protein domain level as defined in CATH. For each homologous family in CATH, all pairs of near-identical sequence representatives ('NReps') were aligned using SSAP and with the Needleman and Wunsch sequence alignment algorithm. This produced a total data set of approximately 56,466 protein pairs.

Sequence alignment was performed using a local implementation of the Needleman and Wunsch algorithm[29] using the Dayhoff MDM78 matrix, a gap opening penalty of 10 and an extension penalty of 2 (program NW, `http://www.bioinf.org.uk/software/nw/`). Other matrices and alignment programs were also tested and made only minor differences to the ability to replicate the structure-derived alignment.

A program was written in Perl to compare all pairs of sequence and structure alignments and identify regions where they differ (i.e. SSMAs). Secondary structure assignments were calculated using SSTRUC (Smith, D.K. and Thornton, J.M., unpublished) which is a modification of the DSSP algorithm[42]. The 'bend' and undefined secondary structure classes (together with a very small number of π-helix residues) were treated as one 'coil' class. These data were stored together with the SSMA assignments and the comparison of sequence and structural alignments is summarized in Figure 2.

Expected frequency for secondary structure i at the start (or end) of an SSMA region was calculated as:

$$E_{i,s} = \frac{O_{i,d} \times \sum^{i} O_{i,s}}{\sum^{i} O_{i,d}} \tag{1}$$

where $O_{i,d}$ is the observed number of residues with secondary structure i in the whole dataset and $O_{i,s}$ is the observed number of SSMA start (or end) residues with secondary structure i.

Neural networks were implemented using the Stuttgart Neural Network Simulator (SNNS)[43]. Training was always performed for 1000 cycles (early stopping to avoid over-training) using a variety of parameters and training methods. The RProp (Resilient back-propagation) training method[44] was found to work best, generally using default parameters ($\delta_0 = 0.1$, $\delta_{max} = 50.0$, $\alpha = 4$) and one or two hidden layers of 20 nodes per layer. 'Jogging' of networks (adding a small random number between -0.1 and $+0.1$ to each weight at each training epoch) was found to help training by helping to avoid local minima.

Input to neural networks was encoded using a 20-dimensional binary vector to represent each amino acid type (10000000000000000000 = Ala, 01000000000000000000 = Cys, etc.) and a 6-dimensional binary vector to represent secondary structure (100000 = α-helix, 010000 = β-strand, etc.) Output from the network consisted of two nodes to indicate whether the central residue

was an SSMA (10 = No, 01 = Yes).

Outputs for the neural nets are not binary, but are a real value between 0 and 1. Predictions for SSMA and non-SSMA were combined to give a confidence score, C, where -1 indicates a non-SSMA and a $+1$ indicates an SSMA:

$$C = 2 \times \left(\frac{Py}{Py + Pn} - 0.5 \right) \tag{2}$$

where Py = predicted value for the position being an SSMA and Pn = predicted value for the position not being an SSMA.

Results were evaluated using the Matthews' correlation coefficient (MCC):

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{3}$$

and Positive Predictive Value (PPV):

$$\text{PPV} = \frac{TP}{(TP + FP)} \tag{4}$$

where TP, TN, FP, FN are the numbers of true positives, true negatives, false positives and false negatives.

Receiver operating characteristic (ROC) curves[45, 46] were plotted varying the confidence threshold for defining a positive prediction and confirmed that the default cutoff of zero was the best choice (data not shown).

Results

Analysis of SSMAs

Fig. 3a shows the distribution of the number of SSMAs in protein domains. As the figure shows, most comparisons do not show an SSMA; where SSMAs occur, there tend to be only one or two such regions. Figure 3b shows the distribution of SSMA lengths.

We analyzed the secondary structure of the start and end residues of each SSMA, normalizing the results by the expected frequency of each secondary structure class (α-helix, β-strand, 3_{10}-helix, turn, bridge, coil); see Methods. As shown in Fig. 4, the majority of SSMA regions start and end in coils ($(O/E)_{\text{start}} = 2.31$; $(O/E)_{\text{end}} = 1.99$) or turns ($(O/E)_{\text{start}} = 2.31$; $(O/E)_{\text{end}} = 1.99$) and they are particularly unlikely to start or end in α-helices($(O/E)_{\text{start}} = 0.4$; $(O/E)_{\text{end}} = 0.36$). Other classes are also dis-favoured. Considering all residues contained within an SSMA region (rather than the first or last residues), α-helices, bridges, turns and coil are somewhat favoured ($O/E = 1.19, 1.17, 1.13, 1.09$, respectively) while β-strands and 3_{10}-helices are somewhat dis-favoured ($O/E = 0.86, 0.94$, respectively). This non-random distribution of SSMAs in secondary structures suggested that this may be an important factor to consider in predicting their location.

Sequence alignability

Initially we examined sequences to determine whether they contained features which determined their 'alignability' with relatively distant homologues (sequence identity $\leq 35\%$). The dataset of alignments was reduced to consider only the S35 'SReps' from CATH (sequence representatives at the 35% sequence identity level) within each homologous family. A neural network was trained using individual sequences

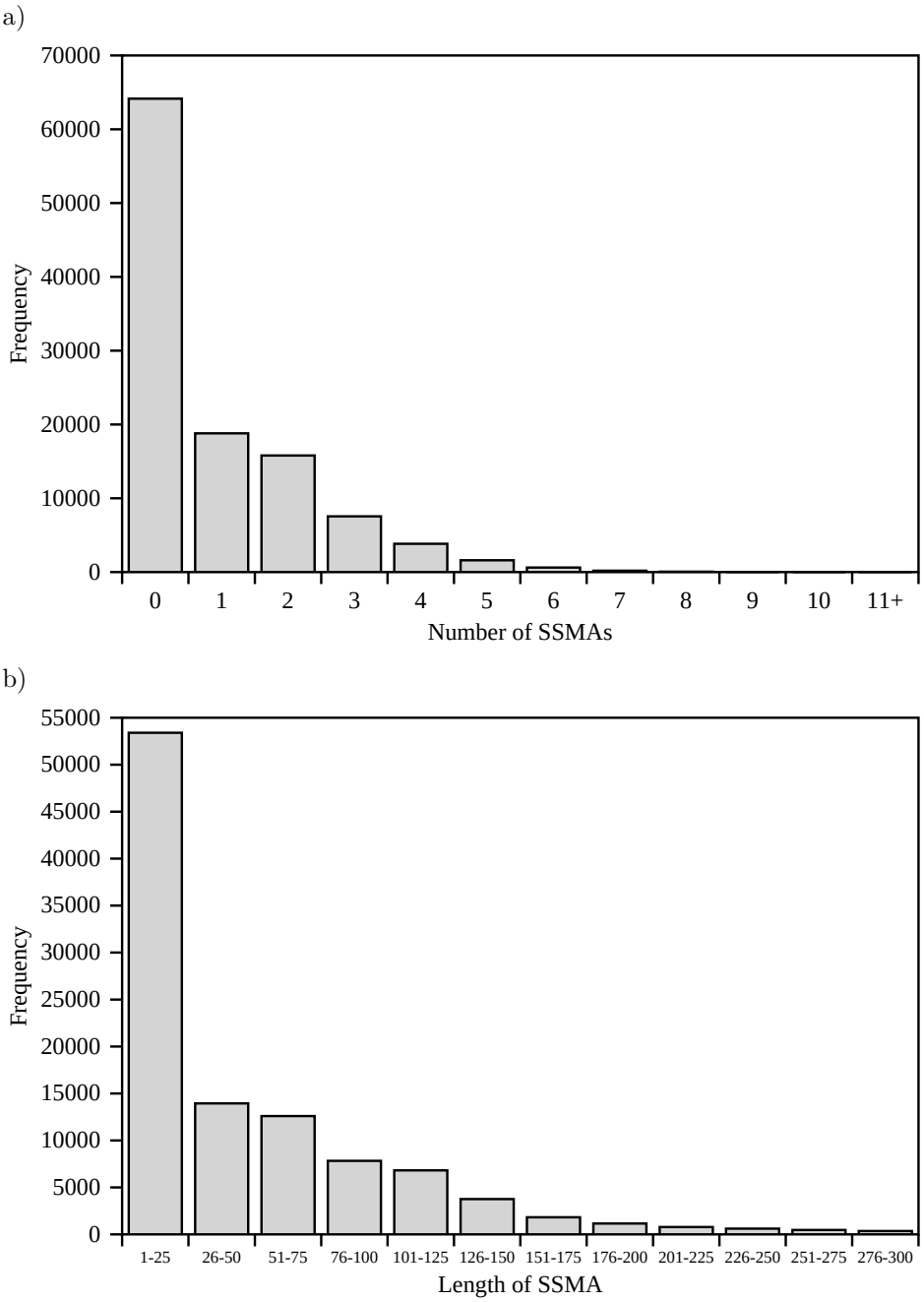

Figure 3: Distribution of a) the number and b) the length of SSMAs within protein domain alignments.

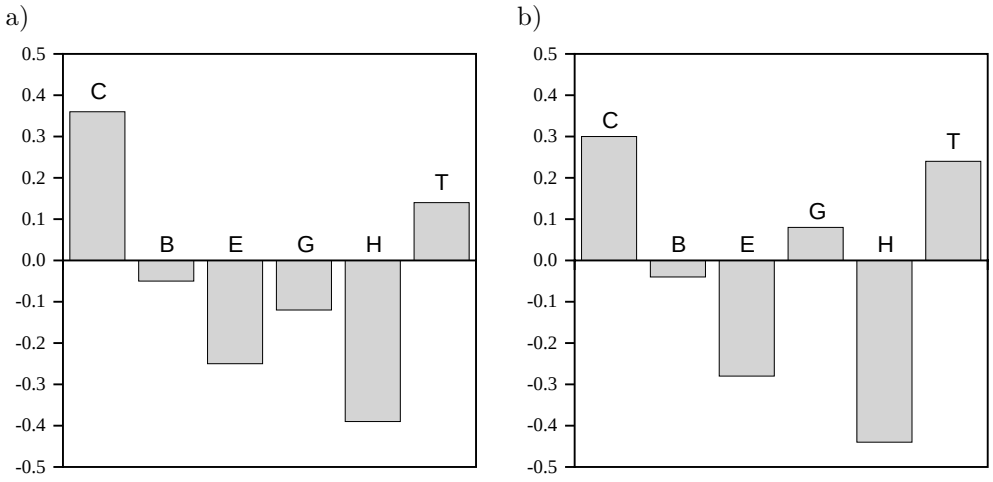

Figure 4: Secondary structures of a) the first and b) the last residue in an SSMA region. (C=coil; B=bridge; E=β-strand; G=3_{10}-helix; H=α-helix; T=turn). Data are expressed as log(observed/expected) frequencies, such that values more frequent than expected are positive and those less frequent than expected are negative.

Table I: Datasets used in training and testing the neural networks.

Dataset	SSMA patterns	Non-SSMA patterns	Total patterns
1	46380	45886	92266
2	46702	46181	92253
3	58070	47059	105129
4	46616	49354	95970

and their secondary structure assignments using a 9-residue window of amino acids and secondary structure assignments.

The total dataset contained 226,812 SSMA residues and 27,981,951 non-SSMA residues. With such large datasets, a program was implemented to select examples at random for use in training and testing. Because the data are so heavily biased towards non-SSMAs, a neural net trained with data in this ratio would do well by predicting everything as a non-SSMA, so the selection program was designed to create approximately equal numbers of SSMA and non-SSMA residues. The data used for the neural networks are summarized in Table I.

Because the dataset was large, jack-knifing or cross-validation was not necessary — in addition to a training set, multiple separate large test sets could be created. Training was performed on approximately 25% of the data (Dataset 1 or 5 of Table I) while, for testing, the MCCs were averaged over the remaining 3 data sets (approximately 75% of the data).

The best performing network, which had a double-hidden layer of 20 nodes per layer and was trained using RProp with recommended settings, had a positive predictive value (PPV) for SSMA prediction of 89.1% (MCC=0.798).

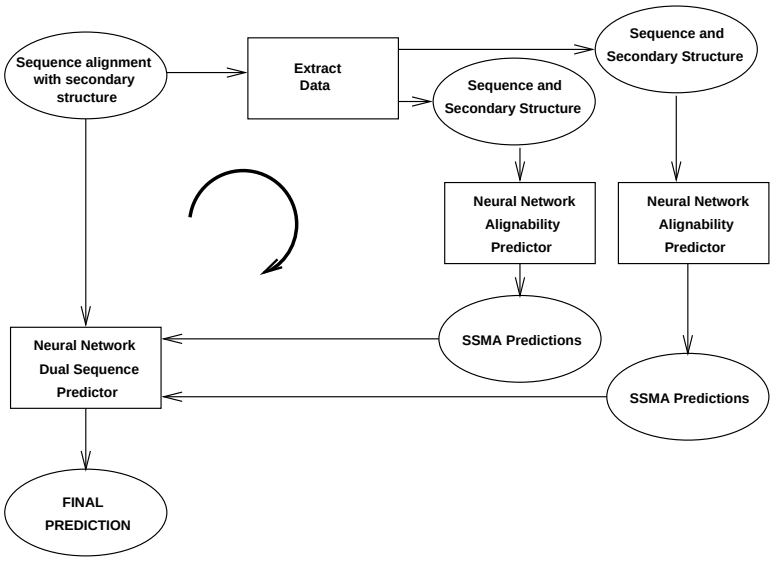

Figure 5: Methodology used for dual sequence prediction. SSMAs are predicted for each sequence individually and combined on the basis of the sequence alignment. The alignment and SSMA predictions are then fed into the second network which makes a final prediction.

Table II: Datasets used in the training and testing of neural networks for dual sequence prediction.

Dataset	SSMA patterns	Non-SSMA patterns	Total patterns
1	53635	49787	103422
2	53356	50165	103521
3	52751	51078	103829
4	53264	49792	103056

Dual Sequence Prediction

The single sequence predictions suggested that there are intrinsic qualities of sequences which affect their ability to be correctly aligned. The next stage, therefore, was to try to use two sequences, aligned on the basis of sequence alone, to predict which aligned residues would match the structural alignments. Given the high quality of the initial 'alignability' networks, initial attempts to create a combined network produced surprisingly poor results (best performance PPV=67.5%, MCC=0.354, other data not shown). Therefore a different approach was taken, based on the successful single sequence networks.

Each sequence was analyzed individually using the best network described above to identify regions in the single sequences likely to be SSMAs. These predictions were then combined using the alignment and fed into a second network to predict the final SSMA locations (Fig. 5).

As before pattern files of approximately 100,000 9-residue windows were used to train and test the networks with no overlap between training and testing sets (Table II).

The best performing second-level network used RProp with recommended settings, a single hidden layer of 20 nodes and a had PPV of 92.9% for prediction of SSMAs with an MCC of 0.648. Another network had a somewhat higher average

```
                                              *                Predicted SSMAs
##############################################|||#############   Confidence
                                                               Actual SSMAs
----------------------PYQVSLNSGY--HFCGGSLINDQWVVSAAHCYKSRIQ   1sluB2
IKGGLFADIASHVCLPPADLQLPDWTECELSGYGKHEALSPFYSERLKEAHVRLYPSSRC   1rtfB1


                    ** ****   ************ ** *******   ****   Predicted SSMAs
##################|||:|#######################################   Confidence
                       ***   ******** *************** ******   Actual SSMAs
VTLGEHNINVLEGNEWFVNAAKIIKHPNFDRKTLNNDIMLIKLSSPVKVATNYVDWIQDT   1sluB2
TSQHLLNRTVTD-NMLCAGDTRSNLH---DACQGDSGGPLVCLNDGRMTLVGIISWGLGC   1rtfB1


 ****                                                          Predicted SSMAs
##############                                                 Confidence
*****                                                          Actual SSMAs
IA-----------*                                                 1sluB2
GQKDVPGVYTKVT*                                                 1rtfB1
```

Figure 6: The SSMA predictions for the alignment pairing 1sluB2 and 1rtfB1. Confidence is calculated as shown in Equation 2 and indicated as '.' $0.2 < |C| \leq 0.4$; ':' $0.4 < |C| \leq 0.6$; '|' $0.6 < |C| \leq 0.8$; '#' $0.8 < |C| \leq 1.0$.

MCC (0.703), but had a lower PPV of 85.1%.

As an example, the alignment between 1sluB2 (anionic N143H, E151H trypsin complexed with A86H ecotin from *Rattus norvegicus*) and 1rtfB1 (two chain tissue plasminogen activator from *Homo sapiens*) is shown in Fig. 6. The alignability predictions were combined with the alignment using the first of the second-level networks described above and confidence scores were calculated (see Methods).

Fig. 7a shows the percentage correct alignment scores for all the original sequence alignments. Figure 7b shows the percentage correct alignment scores for those alignments that the neural networks predicted as not containing any SSMAs. This clearly indicates that those alignments predicted by the networks not to contain SSMAs were, in general, correctly aligned.

Applying Predictions To Modelling

In order to apply the trained neural networks to improving the alignment between protein sequences, a program was written to generate a variety of alternative alignments based on the predicted positions of the SSMA regions. The program initially smooths the SSMA prediction data to remove predictions of single residue SSMAs and merge SSMA regions separated by only one or two residues.

The program then splits the alignment into blocks of either predicted SSMA regions or non-SSMA regions. Each SSMA block is then dealt with individually. Firstly all gaps within the block are removed and the lengths of the remaining sequences compared. If the lengths of each sequence are different then a suitable length gap is reintroduced. The reintroduction of the gap is done at each possible position within the block to create the initial variety within the SSMA block. Also, when the lengths differ, a gap character from within the block may be chosen and moved to a different position at random.

If the two protein sequences within the SSMA block are the same length, then a gap is introduced at random in each sequence. This stage may be repeated for

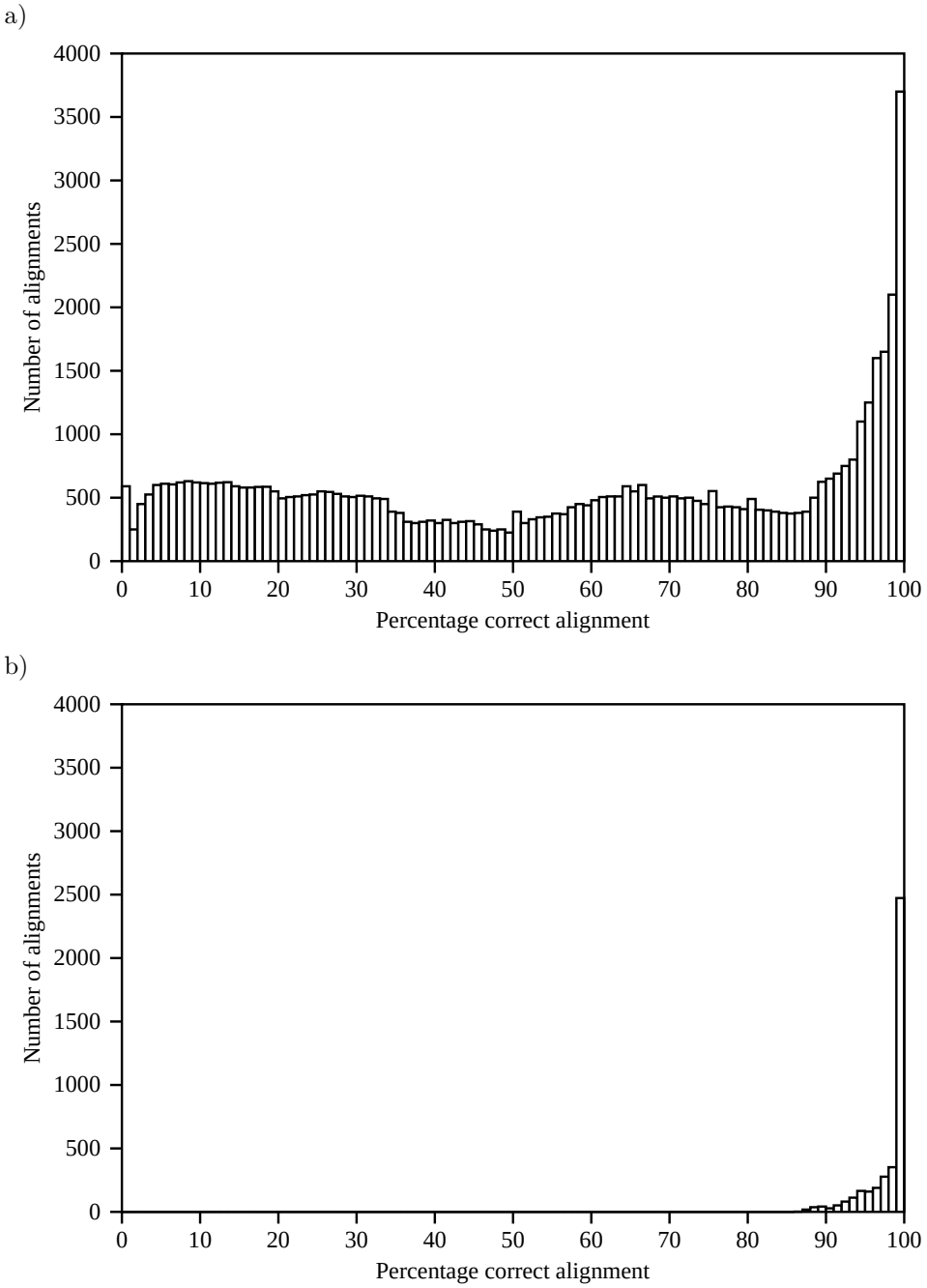

Figure 7: Distribution of the percentage correct alignment scores for a) all domain pairs aligned on the basis of sequence, and b) alignments predicted by the neural networks not to contain any SSMAs.

a)

```
GHYTTILLI-GCCGVNRRK
GGTTTLLLL-ECGGVPHGK
```

b)

```
GHYTTILLIG-CCGVNRRK
GGTTTLLLL-ECGGVPHGK
```

c)

```
GHYTTILL-I-GCGVNRRK
GGTTTLLLLEICGGVPHGK
```

Figure 8: Cleanup stages in generating alternative permuted alignments. a) Removing redundant gaps, b) Removing adjacent gaps c) Removing 'unlikely' alignments.

all the alignments previously created by the program. All the alignments are kept from each stage so as to provide a wide variety of possible permutations.

This is followed by a clean-up stage in which unlikely or repetitious alternative alignments for each predicted SSMA block are removed. Firstly any redundant gaps where a gap is aligned with a gap are removed (Fig. 8a). Secondly adjacent unaligned gaps are closed (Figure 8b). Thirdly permuted SSMA blocks with 'unlikely' alignments are removed. These have a lone non-terminal residue with a gap either side as shown in Figure 8c. Any duplicate alignments are then removed. The final stage is to merge each of the possible SSMA permutations with the original non-SSMA blocks to produce the final selection of alternative alignments.

By altering the number of random permutations generated at each stage it is possible to generate anything from a few tens of permutations up to several million. Clearly the number of possible permutations is also highly dependent upon the number and size of predicted SSMAs blocks, but the program guarantees that there will be a good variety of alternative alignments.

In order to test the effectiveness of the permutation program, large scale testing was done using the large dataset created from the SRep pairs within each H-family of the CATH dataset. Since it would take a great deal of time to run the program for each protein alignment of the approximately 20,000 that made up the dataset, the permutation program was set to produce the minimum number of alternative alignments.

The correct structural alignments were permuted and each permutation was scored using the neural networks to see whether these could correctly select the 'best' of the permuted alignments (Fig. 9). These results should be compared with Figure 7a. Again, it is clear that the alignments selected by the networks as having the fewest SSMAs are generally accurate.

A more thorough test was then performed on a smaller data set, the sequences provided for the CASP5 experiment[47]. We had previously generated alignments and hand-modified these to build models (using MODELLER[21, 22]) for submission to CASP5. These alignments were scored using the networks to predict SSMAs and the alignments were permuted in these regions. The permuted alignments were scored using the networks and the alignments predicted to have the fewest SSMAs were selected. Models were built using MODELLER, based on these alignments.

The RMSD of the original models submitted to CASP5 together with RMSDs of models generated with the alternative alignment generated by this protocol are shown in Table III together with the mean RMSD for each group of models. As

Table III: RMSD values of the CASP5 models created by the different methods of protein alignment together with mean values[a].

Target name	Original model RMSD	Permuted model RMSD
T0130	9.97	8.96
T0130	15.26	9.97
T0133	12.35	9.24
T0133	12.22	8.91
T0137	1.02	1.06
T0142	3.49	2.96
T0149	17.28	6.35
T0149	17.03	7.19
T0149	17.40	7.01
T0150	2.70	2.75
T0150	2.66	2.10
T0153	5.34	6.55
T0153	5.37	6.59
T0154	6.95	6.14
T0154	6.85	4.10
T0155	6.03	2.23
T0160	7.26	3.30
T0160	6.83	2.53
T0160	7.16	2.61
T0167	4.80	5.72
T0171	9.20	5.91
T0171	9.66	5.83
T0179	5.45	2.47
T0179	5.48	2.30
T0182	1.42	1.31
T0184	3.86	3.88
T0188	2.32	2.21
Mean RMSD	7.61	4.82
Mean best RMSD	6.15	4.26
Mean worst RMSD	6.58	4.53

[a]For a number of targets, multiple models were created. These differed in the alignments that were created by manual adjustment of alignments resulting from global sequence alignment. Each of these were used as input to the SSMA prediction and alignment permutation method. The 'Mean RMSD' shows the mean RMSD over all models, while the 'Mean best' and 'Mean worst' RMSDs refer to averages calculated over only the best and worst models in each group.

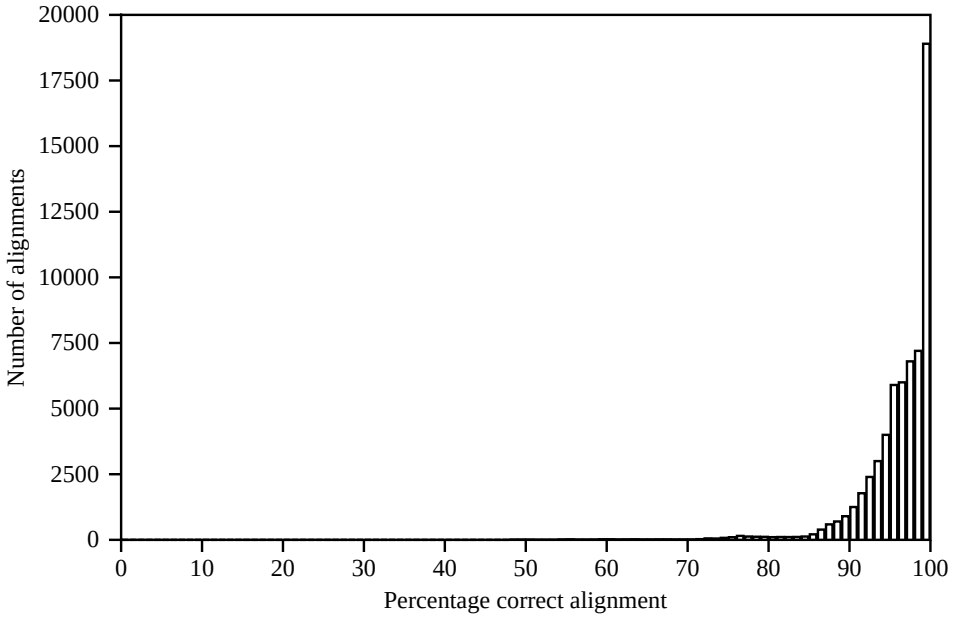

Figure 9: Distribution of percentage correct alignment scores for those alignments selected as 'best' by the neural network from permutations of the original structural alignments.

the table shows, the mean RMSD improved from 7.61Å to 4.82Å, an improvement of approximately 37%. It is important to remember that the networks were trained to recognize errors in alignment where the sequence identity was less than 35%, so cannot be expected to perform well where the sequence identity is high. In the CASP5 examples shown in the table, none of the sequence identities between target and parent is $> 43\%$.

Discussion

Sequence-structure misalignments (SSMAs) are defined as regions in which the sequence alignment does not match the structural alignment. These are less extreme examples of 'misleading local sequence alignments' (MLSAs) previously studied by Saqi *et al.*[39]

The main source of error in comparative modelling is in obtaining the correct alignment. Severe MLSAs are relatively rare, but because, by definition, the sequence alignment is so convincing, they are very difficult to identify when only the sequence alignment is known. SSMAs, however are extremely common. Sequence alignment generates an 'optimum' alignment based solely on a similarity matrix. It fails to take into account factors that may be important once the protein folds into three dimensions: secondary structures, charges, hydrophobicity and distance constraints affecting insertions and deletions.

We studied the occurrence of SSMAs and went on to design neural networks able to predict where they occur. Together with a method for permuting the alignments within regions predicted to be SSMAs, these neural nets were applied to the problem of identifying the most likely alignment and improving comparative modelling.

Examination of SSMAs in CATH domain sequence alignments showed that they varied greatly in length, though smaller SSMAs were more frequent and the number

of SSMAs per sequence alignment was generally small (one or two). There was a strong correlation with secondary structure showing a strong preference to begin and end in a coil or turn; α-helices are particularly dis-favoured. This bias towards beginning and ending in certain types of secondary structure lead to the secondary structure of the sequences being included in the input to neural networks designed to predict the presence of SSMAs.

Initially we used neural networks to predict where a single sequence was likely to be misaligned when aligned with a homologue of $\leq 35\%$ sequence identity. This achieved a surprisingly high Matthews' Correlation Coefficient of 0.798 indicating that there are clearly intrinsic features within sequences which make them difficult to align correctly.

Networks trained directly with two aligned sequences in an attempt to predict misaligned regions performed poorly (data not shown), but by combining the performance of the individual sequence predictions with an alignment we achieved an MCC=0.648.

Permuted alternative alignments were created and tested with this network. The alignment considered likely to have the fewest SSMAs was used to generate models of proteins previously aligned and manually refined for submission to CASP5. The models generated in this way showed a 37% improvement in the mean RMSD from 7.61Å to 4.82Å. However, there is still room for considerable improvement in generating alternative alignments. Depending on the size and number of the SSMAs, several million permuted alignments can be generated and more 'intelligent' generation of permuted alignments (perhaps by exploiting sub-optimal alignments) is needed.

In conclusion, we have demonstrated clear secondary structure preferences for the start and end of SSMAs. We have created neural networks which have proved very successful in identifying SSMAs in both single sequences and aligned protein pairs. These have been applied to real comparative modelling problems from CASP5 and significantly improved the models compared with expert manual adjustment of alignments.

Acknowledgements

DT was funded by a UK MRC priority studentship in Bioinformatics.

References

[1] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nuc. Ac. Res. 2002;30:17–20.

[2] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nuc. Ac. Res. 2000;28:235–242.

[3] Blundell TL, Carney D, Gardner S, Hayes F, Howlin B, Hubbard T, Overington J, Singh DA, Sibanda BL, Sutcliffe MJ. Knowledge-based protein modelling and design. Eur. J. Biochem. 1988;172:513–520.

[4] Donate LE, Rufino SD, Canard LH, Blundell TL. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. Protein Sci. 1996;5:2600–2616.

[5] Burke DF Deane CM. Improved protein loop prediction from sequence alone. Protein Eng. 2001;14:473–478.

[6] Bruccoleri RE Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 1987;26:137–168.

[7] Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins: Struct., Funct., Genet. 2000;41:86–97.

[8] Ponder JW Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 1987;193:775–791.

[9] Liang S Grishin NV. Side-chain modeling with an optimized scoring function. Protein Sci. 2001;11:322–331.

[10] Dunbrack RL Karplus M. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. J. Mol. Biol. 1993;230:543–574.

[11] Bower MJ, Cohen FE, Dunbrack, R. L. J. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J. Mol. Biol. 1997;267:1268–1282.

[12] Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 2003;12:2001–2014.

[13] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. A program for marcromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 1983;4:187–217.

[14] Kale L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K. NAMD2: greater scalability for parallel molecular dynamics. J. Comput. Phys. 1999;151:283–312.

[15] Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins: Struct., Funct., Genet. 1993;17:355–362.

[16] Sánchez R Šali A. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins: Struct., Funct., Genet. 1997;1:50–58.

[17] Sutcliffe MJ, Haneef I, Carney D, Blundell TL. Knowledge based modelling of homologous proteins. 1. Three-dimensional frameworks derived from simultaneous superposition of multiple structures. Protein Eng. 1987;1:377–384.

[18] Sutcliffe MJ, Hayes FRF, Blundell TL. Knowledge based modelling of homologous proteins. 2. Rules for the conformations of sustituted side chains. Protein Eng. 1987;1:385–392.

[19] Peitsch MC. ProMod and Swiss-Model: Internet-based tools for automated comparative modelling. Biochem. Soc. Trans. (London) 1996;24:274–279.

[20] Guex N Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.

[21] Šali A Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 1993;234:779–815.

[22] Fiser A, Do RK, Šali A. Modeling of loops in protein structures. Protein Sci. 2000;9:1753–1773.

[23] Bates PA Sternberg MJ. Model building by comparison at CASP3: using expert knowledge and computer automation. Proteins: Struct., Funct., Genet. 1999;37:47–54.

[24] Ogata K Umeyama H. An automatic homology modeling method consisting of database searrches and simulated annealing. J. Mol. Graph. 2000;18:305–306.

[25] Lambert C, Leonard N, De Bolle X, Depiereux E. ESyPred3D: prediction of proteins 3D structures. Bioinformatics 2002;18:1250–1256.

[26] DePristo MA, De Bakker PIW, Shetty RP, Blundell TL. Discrete restraint-based protein modeling and the Calpha-trace problem. Protein Sci 2003;12:2032–2046.

[27] Martin ACR, MacArthur MW, Thornton JM. Assessment of comparative modeling in CASP2. Proteins: Struct., Funct., Genet. 1997;Suppl. 1:14–28.

[28] Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A. A study of quality measures for protein threading models. Bioinformatics 2001;2:5–5.

[29] Needleman SB Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 1970;48:443–453.

[30] Smith TF Waterman MS. Identification of common molecular subsequences. J. Mol. Biol. 1981;147:195–197.

[31] Shindyalov IN Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 1998;11:739–747.

[32] Taylor WR Orengo CA. Protein structure alignment. J. Mol. Biol. 1989;208:1–22.

[33] Subbiah S, Laurents DV, Levitt M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. Curr. Biol. 1993;3:141–148.

[34] Holm L Sander C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 1993;233:123–138.

[35] Kawabata T. MATRAS: A program for protein 3D structure comparison. Nuc. Ac. Res. 2003;31:3367–3369.

[36] Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. 1996;266:540–553.

[37] Krissinel E Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. 2004;60:2256–2268.

[38] Novotny M, Madsen D, Kleywegt GJ. Evaluation of protein fold comparison servers. Proteins: Struct., Funct., Genet. 2004;54:260–270.

[39] Saqi MAS, Russell RB, Sternberg MJE. Misleading local sequence alignments: implications for comparative modelling. Protein Eng. 1998;11:627–630.

[40] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH–a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.

[41] Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nuc. Ac. Res. 2005;33:D247–D251.

[42] Kabsch W Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.

[43] Zell A, Mamier G, Vogt M, Mache N, Hubner R, Döring S, Herrmann KU, Soyez T, Schmalzl M, Sommer T, Hatzigeorgiou A, Posselt D, Schreiner T, Kett B, Clemente G, Wieland J. Stuttgart neural network simulator, 1995. `http://www-ra.informatik.uni-tuebingen.de/SNNS/`.

[44] Reidmiller M Braun H. A direct adaptive method for faster backpropagation learning: the Rprop algorithm. Proceedings International Conference on Neural Networks 1993;pages 586–591.

[45] Peterson WW. The theory of signal dectectability. IRE Transactions on Information Theory 1954;4:171–212.

[46] Gribskov M Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput. and Chem. 1996;20:25–33.

[47] Tramontano A Morea V. Assessment of homology-based predictions in CASP5. Proteins: Struct., Funct., Genet. 2003;53 Suppl 6:352–368.