INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 1758 - 1766
RECEIVED : 05.03.2017 PUBLISHED : 20.03.2017
March 23, 2017

ANALYSIS OF PROTEIN SEQUENCE USING PROFILE ALIGNMENT METHOD

[1]R. Sangeethapriya, [2]B. Kalaiselvi
[1,2]Assistant Professor, Mahindra Engineering College for Women, Tamilnadu, India

Abstract

The accuracy of an alignment between two protein sequences can be improved by including other detectably related sequences in the comparison. To optimize an approach that relies on aligning two multiple sequence alignments, each one including one of the two protein sequences. Thirteen different protocols for creating and comparing profiles corresponding to the multiple sequence alignments are implemented in the SALIGN command of MODELLER. A test set of 200 pair wise, structure-based alignments with sequence identities below 40% is used to benchmark the 13 protocols as well as a number of previously described sequence alignment methods, including heuristic pair wise sequence alignment by BLAST, pair wise sequence alignment by global dynamic programming with an affine gap penalty function by the ALIGN command of MODELLER, sequence-profile alignment by PSI-BLAST, Hidden Markov Model methods implemented in SAM and LOBSTER, pair wise sequence alignment relying on predicted local structure by SEA, and multiple sequence alignment by CLUSTALW and COMPASS. The alignment accuracies of the best new protocols were significantly better than those of the other tested methods. The new method is currently applied to large-scale comparative protein structure modeling of all known sequences.

Keywords: protein sequence alignment, sequence profiles, comparative protein structure modeling

1. INTRODUCTION

Nucleic acid and protein sequence alignments are central to many problems in biology, including gene assignment, phylogeny construction, protein structure modeling, protein design, and functional annotation of proteins. An alignment between two sequences of residues is usually calculated by optimizing an alignment scoring function. The two common ingredients of the scoring function are a gap penalty function and a matrix of substitution scores for matching every residue in one sequence to every residue in the other sequence. The alignment score is usually a sum of the gap penalties that depend linearly on the gap lengths and the pair wise substitution scores that depend on the matched residue types. The original and still widely used optimization method for sequence alignment is based on dynamic programming. The scoring function and its optimization by dynamic programming have been improved for alignment accuracy and speed, and applied to a variety of alignment problems.

1.1 BLAST and SATCHMO algorithm

One of the most significant improvements in alignment accuracy was achieved through the use of multiple sequence alignments and the corresponding sequence profiles. For proteins, a sequence profile lists a preference for the 20 standard amino acid residue types at each position in a given multiple sequence alignment. The PSI-BLAST program relies on the BLAST algorithm to collect homologs of a query sequence and construct its profile by iteratively scanning a sequence database is a comparison of the query sequence profile with each sequence in the database. A multiple sequence alignment can also be transformed into a Hidden Markov Model (HMM), a class of probabilistic models that are generally applicable to a time series or linear sequences. A particularly successful method in this class is implemented in the SAM package that outperforms other sequence-based methods for fold recognition.

The SATCHMO algorithm in the LOBSTER package simultaneously constructs a similarity tree and compares multiple sequence alignments of each internal node of the tree using HMMs. The CLUSTALW program compares two multiple sequence alignments by scoring an alignment of two positions, one from each profile, as the average of all pair wise substitution scores for the amino acid residues in the two profiles. The LAMA program aligns two multiple sequence alignments by first transforming them into profiles and then comparing the two to each other by the Pearson correlation coefficient. Similarly, the FFAS program was developed to align two sequence profiles with each other. A related approach was also used to construct the ProtoMap database of protein sequence families. ProtoMap database combined multiple structure and sequence comparisons to improve the accuracy of alignments of SH2 domains. The COMPASS program was developed to locally align two multiple sequence alignments with assessment of statistical significance. These methods compare two profiles by constructing a matrix of scores for matching every position in one profile to each position in the other profile, followed by either local or global dynamic programming to calculate the optimal alignment. It was noted previously that profile profile alignment methods are capable of detecting more remote relationships compared to the sequence-profile methods, such as PSI-BLAST. Another significant improvement of the alignment accuracy in the low similarity range was achieved by considering protein structure information for one of the sequences in a pair wise comparison. The methods in this class include threading and 3D template matching.

2. MULTIPLE SEQUENCE ALIGNMENT

For each sequence in a pair of sequences to be aligned, a multiple sequence alignment with its homologs was prepared by scanning the no redundant protein sequence database at NCBI with the program PSI-BLAST. The scanning was performed without filtering out compositionally biased segments, was run for up to 20 iterations, and included all matches with an e-value smaller than 0.0005. Up to 1000 sequences with the most significant e-values were retained in the multiple sequence alignment. The default values were used for all other parameters. The multiple sequence alignment and the profile were saved after each iteration. The PSI-BLAST multiple sequence alignment of a sequence was defined to be the sequence-profile alignment with the most significant e-value from any of the iterations. It proceed by defining the 13 profile–profile alignment protocols in terms of four alternative schemes for transforming a multiple sequence alignment into a profile or a matrix and six alternative measures for comparison of two profiles.

Another approach is implemented in the SEA program, which aligns a pair of remotely related sequences by optimizing a match between the predicted conformations of their short segments. The resulting alignments were more accurate than the pair wise sequence alignments by BLAST and ALIGN, as well as the profile-profile alignments by FFAS. For closely related protein sequence pairs with sequence identity over 40%, an accurate alignment is almost always trivial to obtain. In contrast, despite the methodological advances listed above, alignments in the so-called “twilight zone” of less than 30% sequence identity still contain many errors. Some pairs of related proteins have almost no correctly aligned positions when aligned by sequence-based alignments methods. Alignment accuracy in the twilight zone is crucial for several applications, including comparative protein structure prediction. To calculate an accurate comparative model, it is necessary to identify and correctly align at least one template structure to the target sequence. An incorrect alignment invariably leads to an inaccurate model, because none of the existing comparative model building methods can generally recover from an incorrect alignment.

2.1 Sequence weighting

Sequence weighting is part of the calculation of a sequence profile from a multiple sequence alignment, and is used to compensate for no uniform distribution of the homologs in the alignment. There are two different weighting schemes. First, we tested the often used position-based sequence weighting that assigns low weights to overrepresented sequences and high weights to unique sequences.

$$W_j^{(1)} = \sum_i \frac{1}{r_i \cdot n_{ij}}$$

where r_i is the number of different residue types at position i and $n_{i,j}$ is the frequency of the residue type in sequence j at position i. Second, we also tested our variation of the position-based sequence weighting that increases the weights of those sequences that are more similar to the query sequence.

$$W_j^{(2)} = \sum_i \frac{O_{a(i,1),b(i,j)}}{r_i \cdot n_{ij}}$$

Where $O_{a(i,1),b(i,j)}$ are the Blosum62 odds ratios for matching the residue type a in the query sequence with the residue type b in sequence j.

2.2 Sequence Profile

A sequence profile of a given set of similar sequences specifies a preference for each of the 20 standard amino acid residue types at each of the residue positions in the set. A number of different estimation schemes have been suggested, because a multiple alignment may not contain a sufficiently large number of homologs to calculate a statistically robust profile solely from the occurrence of each residue type in the multiple alignments. They generally depend on prior or expected probabilities of residue occurrences and residue-residue substitutions.

First, profiles generated by pseudo-counting as implemented in the PSI-BLAST program. The use of pseudo-counting for profile generation was chosen for its simplicity of implementation and comparable performance to other tested approaches.Second, profiles generated by pseudo-counting as implemented by us in the MODELLER-7 program. The probability of a residue type a to occur at position i in a multiple alignment is estimated by

$$P_{i,a} = \frac{N_i}{N_i + B_i} \cdot \frac{n_{i,a}}{N_i} + \frac{B_i}{N_i + B_i} \cdot \frac{b_{i,a}}{B_i}$$

$$B_i = m \cdot r_i$$

$$b_{i,a} = B_i \cdot \sum_{a=1\ b=1}^{20} \frac{n_{i,a}}{N_i} \cdot \frac{M_{a,b}}{M_a}$$

N_i is the sum of the weights $W_j^{(1)}$ (eq. 1) for the sequences that do not have a gap at position i. $n_{i,a}$ is the sum of the weights $W_j^{(1)}$ for the sequences with residue type a at position i. B_i is the total number of pseudo-counts at position i and depends on the parameter m that is set to the optimal value of 5. $b_{i,a}$is the number of pseudo-counts for residue type a at position i. M_a is the probability of residue type a in the background distribution that is obtained from the Blosum62 matrix. $M_{a,b}$ are the Blosum62 probabilities for matching the residue type a in the query sequence with the residue type b in sequence j. Both $n_{i,a}/N_i$ and $b_{i,a}/B_i$ are estimates of $P_{i,a}$, based on the observed and pseudo-counts, respectively. Correspondingly, $P_{i,a}$ is a weighted sum of the two estimates, with the contributions determined by N_i and B_i. IfN_i is larger than B_i, $P_{i,a}$ is dominated by the observed counts, whereas if B_i is larger than N_i, $P_{i,a}$ is dominated by pseudo-counts. Third, our variation of the Henikoff and Henikoff schema with sequences weighted proportionally to their similarity to the query sequence, using $W_j^{(2)}$ (eq. 2) instead of $W_j^{(1)}$.

2.3 Profile–Profile substitution scores

An optimal alignment of two profiles P and Q would be obtained by relying on a matrix of probabilities $S_{i,j}$ that any pair of profile positions P_i and Q_j are equivalent. It is not clear what the best definition of equivalent is and how to calculate such a probability of equivalence, given two profile distributions P_i and Q_j. As a result are forced into a parametric approach, whereby we calculate a substitution score that approximates the probability of equivalence. Such substitution scores, together with a gap penalty function, can then be used to obtain an optimal alignment of two profiles by dynamic programming. Six recipes for calculating profile–profile substitution scores $S_{i,j}$ for each pair of profile positions i and j were tested.

First, the dot product between two distributions P_i and Q_i at profile positions i and j, respectively.

$$S_{i,j}^{(1)} = \sum_a (P_{i,a} \cdot Q_{j,a})$$

Second, the correlation coefficient between two distributions P_i and $Q_{j.}$

$$S_{i,j}^{(2)} = \frac{\sum_a (P_{i,a} \cdot Q_{j,a})}{\sqrt{\sum_a (P_{i,a} \cdot P_{i,a}) \cdot (Q_{j,a} \cdot Q_{j,a})}}$$

Third, the Euclidean distance between two distributions P_i and Q_i.

$$S_{i,j}^{(3)} = \sqrt{\sum_a (P_{i,a} \cdot Q_{j,a})^2}$$

Fourth, a substitution score based on the Jensen-Shannon divergence measure D^{JS} for two distributions

$$S_{i,j}^{(4)} = D^{JS}(P_i, Q_j) = \lambda \cdot D^{KL}(P_i,\ R) + (1 - \lambda) \cdot D^{KL}(R, Q_j)$$

$$R = \lambda \cdot P_i + (1 - \lambda \cdot Q_j)$$

$$D^{KL}(P_i,\ Q_j) = \sum_a P_{i,a} log_2 \frac{P_{i,a}}{Q_{j,a}}$$

The R vector can be seen as the most likely parent distribution of P_i and Q_j. D^{KL} is the Kullback-Leibler distance, also called the “cross-entropy measure” in information theory. λ is a parameter between 0 and 1, set to 0.5 in this study. λ and its complement (1−λ) are the weights given to the P_i and Q_j distributions, respectively.

Fifth, for each position in a multiple sequence alignment, a pair wise residue substitution probability matrix was calculated as a weighted sum of the Blosum62 substitution probability matrix and the matrix of relative residue substitution frequencies observed at the given position in the multiple sequence alignment. Next, the substitution score for two multiple alignment positions i and j was calculated by averaging over these residue substitution probabilities for all pairs of residues containing a residue from each of the two compared positions.

$$S_{i,j}^{(5)} = \sum_{a=1}^{20}\sum_{b=1}^{20} f_a^{(i)} \cdot f_b^{(j)} \cdot (M_{a,b}^{(i)} + M_{b,a}^{(j)})$$

$$M_{a,b}^{(i)} = \omega_1 \cdot M_{a,b} + \omega_2 \cdot f_{a,b}^{(i)}$$

$$\omega_1 = \frac{1}{1 + \frac{n}{\sigma}}$$

$$\omega_2 = 1 - \omega_1$$

where $f_a^{(i)}$ is the observed frequency of residue type a at position i in the first multiple alignment corrected for sequence weights as defined above (using equation 1), $M_{a,b}^{(i)}$ is the substitution probability matrix for residue types a and b at position i in the first multiple alignment, $M_{a,b}$ is the Blosum62 substitution probability matrix for residue types a and b, and ω1 and ω2 are scalar weights. Variable n is the number of the pair wise residue-residue substitutions within the multiple alignments at position i, and σ is a smoothing parameter.

Sixth, the score $S_{i,j}^{(6)}$ was defined as the correlation coefficient between the corresponding values in two posterior substitution matrices $M_{a,b}$ $^{(i)}$ and $M_{b,a}$ $^{(j)}$ for positions i and j in the first and second multiple alignments, respectively. After the substitution scores were computed according to one of the six recipes above, they were scaled to fit the range from 0 to 1000.

3. Alignment methods

The testing pairs of sequences were aligned by (1) heuristic pair wise sequence alignment as implemented in BLAST 2.1.2 , (2) pair wise sequence alignment by global dynamic programming with an affine gap penalty function as implemented in the ALIGN command of MODELLER-7 , (3) sequence-profile alignment as implemented by PSI-BLAST 2.1.2 , (4) Hidden Markov Model (HMM) as implemented in SAM 3.3.1 and LOBSTER, (5) pair wise sequence alignment based on matching predicted local structure as implemented in the SEA Web server, (6) multiple sequence alignment by CLUSTALW 1.81, (7) profile–profile alignments as implemented by COMPASS 1.24 , and (8) the 13 schemes of profile–profile alignment by global dynamic programming with an affine gap penalty function as implemented by the SALIGN command of MODELLER-7.

For BLAST, a high e-value threshold of 100 was used for accepting an alignment between two sequences. Otherwise, the pair of sequences was ignored. The e-value is increased the threshold relative to the commonly used value of ~10^{-4} to produce the maximum number of pair wise alignments obtained from the BLAST program. All other parameters were kept at their default values.

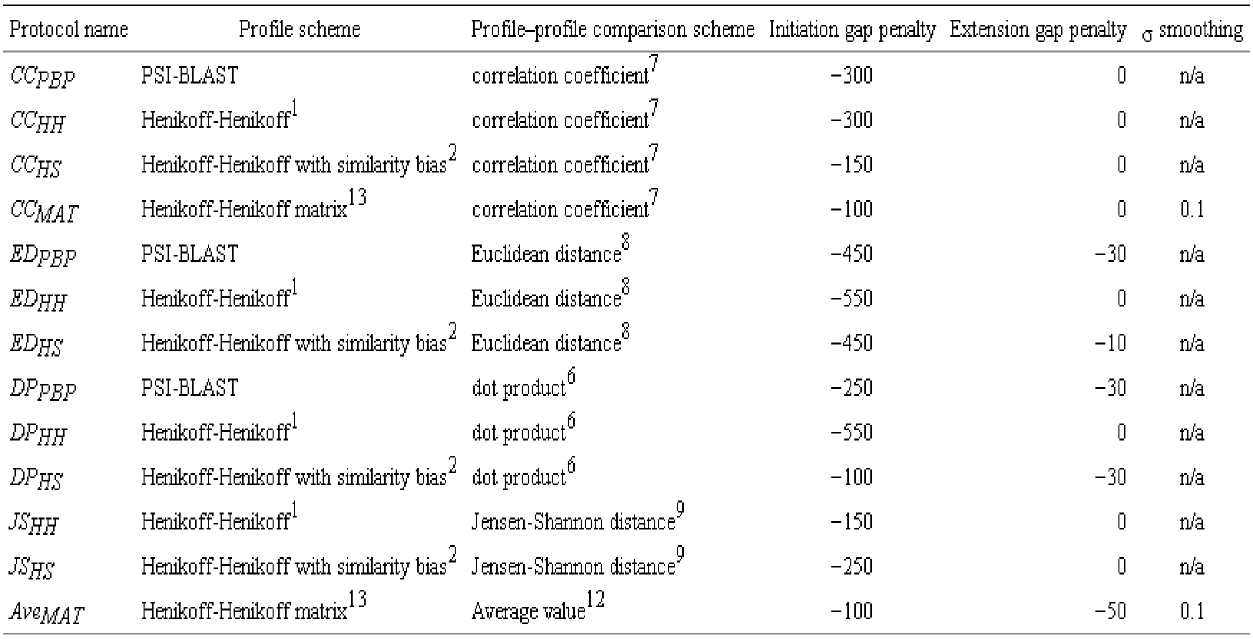
For ALIGN, the default parameters were used. They include the AS1 residue type similarity matrix calculated from the reference structure alignments, the initiation gap penalty u of −450, and the extension gap penalty v of −50, the penalty for a gap of n residue positions is u + v n. For PSI-BLAST, multiple sequence alignments of each one of the two sequences were calculated as described above. The sequence-profile alignment with the most significant e-value from any of the iterations with either of the two sequences as queries was used as the PSI-BLAST alignment.

For SAM, the following protocol was used (R. Karchin, pers. comm.). The w0.5 script with default parameters was applied to build HMMs for the target and template sequences, using their PSI-BLAST multiple sequence alignments. Next, the program hmmscore in the SAM package was employed to align the HMM of the target and the template with the template and the target sequences, respectively, resulting in two generally different template-target alignments. The alignment with the most significant e-value as reported by the hmmscore program was selected.

For LOBSTER, the COACH algorithm was used through the -coach option to align a multiple sequence alignment against a Hidden Markov Model. First, the program was used to build HMMs for the target and the template sequences, using their PSI-BLAST multiple sequence alignments. Next, we aligned the HMMs of the target and the template to the template and target sequences, respectively, resulting in two generally different template-target alignments. The alignment with the higher bit score as reported by LOBSTER was selected.

For CLUSTALW, the profile alignment option with the default parameters was used. This option over the multiple sequence alignment option to benchmark CLUSTALW using the same profiles as for the other tested programs. For COMPASS, the default parameters were used to align the target and template multiple sequence alignments.

For SALIGN, the 13 different protocols were tested, combining three different ways to construct a profile with four different ways to score a match between two profile positions, as well as two protocols based on posterior substitution probability matrices. The PSI-BLAST profiles cannot be used with the Jensen-Shannon scheme for calculating the profile–profile substitution scores because this scheme relies on probabilities P_i and Q_j that are not reported in the PSI-BLAST output.The alignment of two multiple sequence alignments by SALIGN requires approximately 40 sec for ~250 sequences with about ~250 residues in each of the two profiles on a typical Pentium 4 computer. The total CPU time is dominated by the computing of the scoring matrix, rather than the dynamic programming step. This CPU time is approximately proportional to the product of the numbers of sequences in the two profiles and the profile lengths.

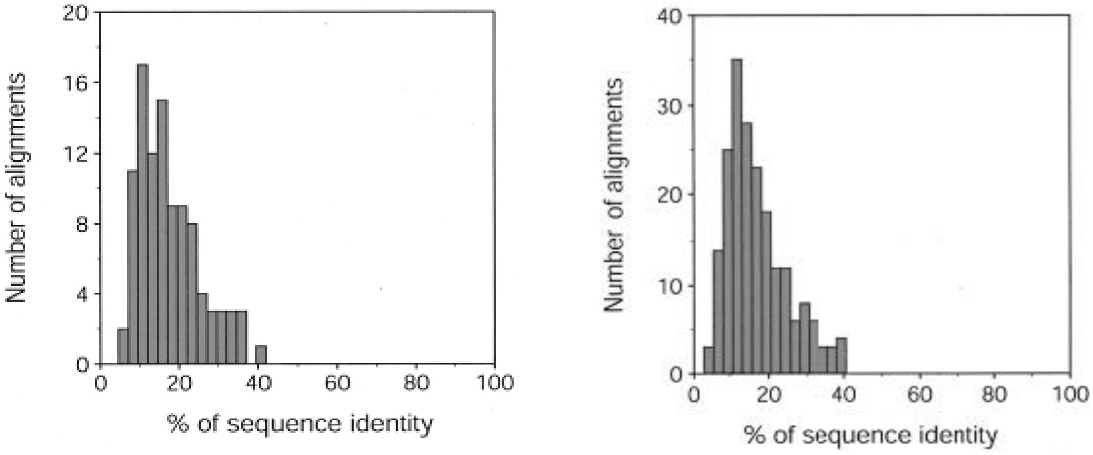


Protocol name	Profile scheme	Profile–profile comparison scheme	Initiation gap penalty	Extension gap penalty	σ smoothing
CC_{PBP}	PSI-BLAST	correlation coefficient[7]	−300	0	n/a
CC_{HH}	Henikoff-Henikoff[1]	correlation coefficient[7]	−300	0	n/a
CC_{HS}	Henikoff-Henikoff with similarity bias[2]	correlation coefficient[7]	−150	0	n/a
CC_{MAT}	Henikoff-Henikoff matrix[13]	correlation coefficient[7]	−100	0	0.1
ED_{PBP}	PSI-BLAST	Euclidean distance[8]	−450	−30	n/a
ED_{HH}	Henikoff-Henikoff[1]	Euclidean distance[8]	−550	0	n/a
ED_{HS}	Henikoff-Henikoff with similarity bias[2]	Euclidean distance[8]	−450	−10	n/a
DP_{PBP}	PSI-BLAST	dot product[6]	−250	−30	n/a
DP_{HH}	Henikoff-Henikoff[1]	dot product[6]	−550	0	n/a
DP_{HS}	Henikoff-Henikoff with similarity bias[2]	dot product[6]	−100	−30	n/a
JS_{HH}	Henikoff-Henikoff[1]	Jensen-Shannon distance[9]	−150	0	n/a
JS_{HS}	Henikoff-Henikoff with similarity bias[2]	Jensen-Shannon distance[9]	−250	0	n/a
Ave_{MAT}	Henikoff-Henikoff matrix[13]	Average value[12]	−100	−50	0.1

Table 3.1: Thirteen protocols implemented in the SALIGN command in MODELLER-7

4. Training and testing alignment sets

To improve the accuracy of comparative protein structure modeling, the reference alignments were pair wise, structure-based alignments. They were extracted from our comprehensive database of pair wise structure-based alignments, DBAli. The alignments in DBAli were calculated by superposing all pairs of proteins of known structure in the Protein Data Bank that are classified into the same H class in the CATH database using the program CE. There are 33,920 such alignments with a Z-score higher than 3.8 .



A) Training Set B) Testing Set

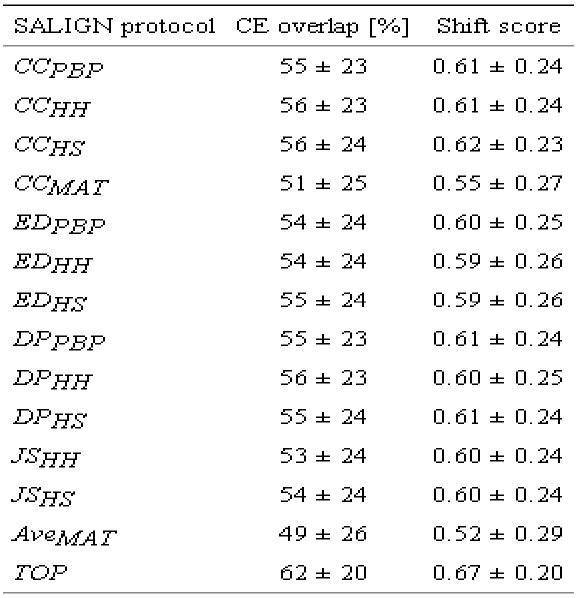
Figure 4.1: Composition of the 300 reference alignments that constitute the training and testing sets. (A) Distributions corresponding to the 100 alignments in the training set. (B) Distributions corresponding to the 200 alignments in the testing set.

First, 387 alignments were extracted from DBAli by requiring up to 40% sequence identity, at least 100 aligned residues, at least 50% of the residues aligned, and that at least 90% of the residues of one chain are covered in the alignment. Second, structure pairs that did not have at least 50% of the residues in the shorter chain aligned by MAMMOTH were also eliminated, resulting in the final set of 300 reference alignments. These 300 alignments were randomly divided into the training and testing sets of 100 and 200 alignments, respectively. The training set of alignments was used to optimize the gap initiation and gap extension penalties for all of our alignment protocols and the parameter σ for the two posterior substitution probability matrix protocols, and the testing set was used to assess the performance of all examined alignment methods. The PDB chain identifiers, chain lengths, percentage sequence identities, root-mean-square deviations (RMSDs) for the aligned C_α atoms, average percentages of the aligned C_α atoms, and percentages of structurally equivalent residues (below) are listed separately for the training and testing alignments.

4.1 Measures of alignment accuracy

The accuracy of an alignment was measured by relying on the aligned native structures extracted from the PDB. First, the RMSD between the corresponding C_α atoms in the two structures was calculated upon rigid-body least-squares superposition of all the C_α atoms, as implemented in the SUPERPOSE command of MODELLER.

Second, the percentage of structurally equivalent positions was defined as the percentage of the C_α atoms in the shorter of the sequences that are within a certain cutoff of the corresponding atoms in the superposed structure. The structure overlap quoted is the average over all cutoffs. Additionally, the alignment methods were assessed by the percentage of alignments with the structure overlap higher than 30%, structure pairs with at least as much overlap have the same fold.



SALIGN protocol	CE overlap [%]	Shift score
CC_{PBP}	55 ± 23	0.61 ± 0.24
CC_{HH}	56 ± 23	0.61 ± 0.24
CC_{HS}	56 ± 24	0.62 ± 0.23
CC_{MAT}	51 ± 25	0.55 ± 0.27
ED_{PBP}	54 ± 24	0.60 ± 0.25
ED_{HH}	54 ± 24	0.59 ± 0.26
ED_{HS}	55 ± 24	0.59 ± 0.26
DP_{PBP}	55 ± 23	0.61 ± 0.24
DP_{HH}	56 ± 23	0.60 ± 0.25
DP_{HS}	55 ± 24	0.61 ± 0.24
JS_{HH}	53 ± 24	0.60 ± 0.24
JS_{HS}	54 ± 24	0.60 ± 0.24
Ave_{MAT}	49 ± 26	0.52 ± 0.29
TOP	62 ± 20	0.67 ± 0.20

Table 4.1.1: Accuracy of the SALIGN protocols

The accuracy of the alignment comparison through with the CE structure-based alignment. First, the fraction of correctly aligned positions was defined as the percentage of positions in the tested alignment that were identical to those in the CE structure-based alignment, the residue-gap matches are ignored in this calculation.

Secondly, the shift score, which ranges from e for two completely different alignments to 1 for identical alignments, was also calculated. The optimal gap initiation and extension penalties for the 11 profile–profile alignment protocols were identified by maximizing the average percentage of correctly aligned positions for the training set of sequence pairs. The maximization scanned all combinations of the initiation penalties from −1000 to 0 in steps of 50 and the extension penalties from −200 to 0 in steps of 10. The gap initiation, gap extension, and the σ parameters for the two posterior substitution probability matrix protocols were optimized on a 3D grid, with σ ranging from 0.001 to 10.

CONCLUSION

In this study, we optimized alignments specifically for comparative protein structure prediction. We begin by describing 13 profile–profile alignment protocols, the training and testing alignment sets, and measures of alignment accuracy. There are 13 variations in the calculation of the profiles and the profile–profile substitution scores. The opening and extension gap penalties as well as the σ parameter were optimized separately for each one of the 13 protocols, by relying on the 100 training alignments. To assess SALIGN and a variety of other alignment methods are used the 200 reference structure-based alignments. First, we assessed the differences in accuracy between the 13 different SALIGN protocols. Next to compared two of the SALIGN protocols for profile–profile alignment by global dynamic programming to a heuristic pair wise sequence alignment (BLAST), a pair wise sequence alignment by global dynamic programming (ALIGN), a heuristic sequence-profile alignment (PSI-BLAST), two HMM methods (as implemented in SAM and LOBSTER), a pair wise sequence alignment by matching predicted local structures (SEA), and two profile–profile alignment methods (CLUSTALW and COMPASS). Finally, to illustrate the utility of our method of comparative protein structure modeling that benefit from profile–profile alignment. The results quantify the significant improvement in the accuracy of sequence alignment that is achieved by the use of multiple sequences. For this analysis, the alignment accuracy of a method was measured independently by the average shift score and CE overlap, both calculated for the 200 testing pairs of sequences.

REFERENCES

1. Al Lazikani, B, Sheinerman, F.B, and Honig.B ”Combining multiple structure and sequence alignments to improve sequence detection and alignment: Application to the SH2 domains of Janus kinases” Proc. Natl. Acad. Sci. 98 14796–14801, 2001.
2. Bairoch, A. and Apweiler, R. “The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000” Nucleic Acids Res. 28 45–48, 2000.
3. Baker, D. and Sali, A. “Protein structure prediction and structural genomics” Science 294 93–96, 2001.
4. Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., et al. “The Protein Data Bank” Acta Crystallogr. D. Biol. Crystallogr. 58 899–907, 2002.
5. Blake, J.D. and Cohen, F.E. Pairwise “sequence alignment below the twilight zone. J. Mol. Biol” 307721–735, 2001.
6. Bonneau, R., Strauss, C.E., and Baker, D “Improving the performance of rosetta using multiple sequence alignment information and global measures of hydrophobic core formation” Proteins 43 1–11, 2001.
7. Cuff, J.A. and Barton, G.J “Application of multiple sequence alignment profiles to improve protein secondary structure prediction” Proteins 40 502–511, 2000.
8. David, R., Korenberg, M.J., and Hunter, I.W. “3D-1D threading methods for protein fold recognition” Pharmacogenomics 1 445–455, 2000.
9. Edgar, R.C. and Sjolander, K “SATCHMO: Sequence alignment and tree construction using hidden Markov models” Bioinformatics 19 1404–1411, 2003.
10. Madera, M. and Gough.J ”A comparison of profile hidden Markov model procedures for remote homology detection” Nucleic Acids Res. 30 4321–4328, 2002.
11. Marti-Renom, M.A., Madhusudhan, M.S., Fiser, A., Rost, B., and Sali, A “Reliability of assessment of protein structure prediction methods” Structure 10 435–440, 2002.
12. John, B. and Sali, A “Comparative protein structure modeling by iterative alignment, model building and model assessment” Nucleic Acids Res. 31 3982–3992, 2003.
13. Panchenko, A.R “Finding weak similarities between proteins by sequence profile comparison” Nucleic Acids Res. 31 683–689, 2003.
14. Ye, Y., Jaroszewski, L., Li, W., and Godzik.A “A segment alignment approach to protein comparison” Bioinformatics 19 742, 2003.