Optimal contact map alignment of protein–protein interfaces

BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 20 2008, pages 2324–2328
doi:10.1093/bioinformatics/btn432

Structural bioinformatics

Optimal contact map alignment of protein–protein interfaces

Vinay Pulim[1], Bonnie Berger[1,2,*] and Jadwiga Bienkowska[1,3,*]

[1]Computer Science and Artificial Intelligence Laboratory, MIT, [2]Mathematics Department, MIT, Cambridge and [3]Biomedical Engineering Department, Boston University, Boston, USA

Received on April 4, 2008; revised on July 16, 2008; accepted on August 14, 2008

Advance Access publication August 18, 2008

Associate Editor: Anna Tramontano

ABSTRACT

The long-standing problem of constructing protein structure alignments is of central importance in computational biology. The main goal is to provide an alignment of residue correspondences, in order to identify homologous residues across chains. A critical next step of this is the alignment of protein complexes and their interfaces. Here, we introduce the program CMAPi, a two-dimensional dynamic programming algorithm that, given a pair of protein complexes, optimally aligns the contact maps of their interfaces: it produces polynomial-time near-optimal alignments in the case of multiple complexes. We demonstrate the efficacy of our algorithm on complexes from PPI families listed in the SCOPPI database and from highly divergent cytokine families. In comparison to existing techniques, CMAPi generates more accurate alignments of interacting residues within families of interacting proteins, especially for sequences with low similarity. While previous methods that use an all-atom based representation of the interface have been successful, CMAPi's use of a contact map representation allows it to be more tolerant to conformational changes and thus to align more of the interaction surface. These improved interface alignments should enhance homology modeling and threading methods for predicting PPIs by providing a basis for generating template profiles for sequence–structure alignment.

Contact: bab@mit.edu; jbienkowska@gmail.com

Supplementary information: Supplementary data are available at http://theory.csail.mit.edu/cmapi

1 INTRODUCTION

Structure-based protein–protein interaction (PPI) prediction is an emerging area with vast potential to impact systems biology, genomics, molecular biology and therapeutics. Success would greatly improve data mining from genome sequencing, structural proteomics and other large-scale experiments that probe networks. It would also provide leads for experiments and drug design.

Akin to the prediction of single chain protein structure, homology modeling and threading techniques (Bienkowska and Lathrop, 2005; Bowie *et al*., 1991; Dunbrack, 2006; Jones *et al*., 1992; Pieper *et al*., 2006; Xu *et al*., 2003) are likely to be effective for predicting PPIs (Lu *et al*., 2002; Singh *et al*., 2006). The first step in any such method is to generate profiles for templates from an alignment of sequences. However, the construction of optimal profiles for PPI prediction is particularly challenging because interactions between the two protein sequences must be taken into account, preventing their treatment as independent alignments. Indeed, this is likely the reason that threading algorithms that align single chains independently do not perform well when extended to complexes (Pulim *et al*., 2008), largely due to the weighting of non-interacting residues equally with residues critical to the interaction. We have previously achieved more accurate threading by first aligning the interface regions as a whole via contact map representations (Fig. 1) and subsequently generating sequence profiles from the resulting alignment. We have shown that accurate structural alignment of interfaces leads to successful PPI prediction (Pulim *et al*., 2008).

There has been limited previous work on protein–protein interface alignment. A number of different representations have been used to describe protein structure and thus protein–protein interfaces, from contact maps to all-atom representations. We have previously developed the LTHREADER program (Pulim *et al*., 2008) that uses a contact map representation of protein interfaces and generates an accurate alignment of binding interfaces for cytokines, a medically important class of protein families with very low sequence similarity. When multiple structural complexes are available for a ligand-receptor family, LTHREADER performs interface alignment in two stages: first, it identifies interaction core regions by clustering contacts within a specified distance threshold; then, it aligns contact maps by maximizing the overlap between the submatrices defined by the core regions. A limitation of this method is that cores are defined before they are aligned which has the potential to generate sub-optimal contact map alignments. The other existing algorithm for interface alignment is MAPPIS, which uses an all-atom representation of protein interfaces and optimizes the alignment of interface regions with similar physico-chemical properties (Shulman-Peleg and Shatsky, 2005). MAPPIS is useful for certain applications such as function prediction that require recognition of conserved structural patterns of physico-chemical interactions. However, since MAPPIS uses a physical, all-atom-based representation of interfaces, it may be sensitive to small differences caused by conformational changes in the interface surface.

In this article, we introduce a polynomial-time algorithm for optimal pairwise contact map alignment of protein interfaces (CMAPi) using two-dimensional (2D) dynamic programming. For multiple alignment, we apply a neighbor joining algorithm akin to that used for multiple sequence alignment (Higgins and Sharp, 1988). We evaluate our algorithm on the SCOPPI database (Winter *et al*., 2006), which classifies all protein–protein interfaces into similarity classes, and measure its performance according to the

*To whom correspondence should be addressed.

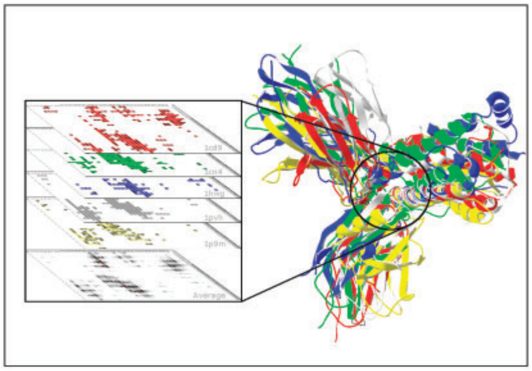

Fig. 1. Contact map representation of the aligned interfaces of five four-helical bundle cytokine complexes. Each color represents a complex and the gray map at the bottom is the average contact map of the aligned interfaces.

percentage of interacting residues aligned correctly (see Section 2). We demonstrate that CMAPi produces more accurate alignments than existing methods, such as MAPPIS and MUSCLE (Shulman-Peleg and Shatsky, 2005; Edgar, 2004), especially for protein sequences with low similarity. Compared to LTHREADER, CMAPi is faster, automated and as accurate, allowing large-scale application. Moreover, our new approach aligns entire contact maps without having to first identify core regions. Instead, cores are automatically determined by the algorithm as a post-alignment step and then are used to generate sequence profiles of the interaction cores. In the future, these profiles will be used to predict new PPIs as described previously for LTHREADER (Pulim *et al*., 2008).

2 METHODS

2.1 Algorithm

CMAPi finds alignments of similar protein–protein interfaces using a contact map representation. First, we generate optimal pairwise interface alignments and then use a version of the neighbor-joining algorithm to align multiple interfaces.

The contact map representation is a 2D matrix X indexed by the residues $i \in L$ and $j \in R$ from the interacting proteins L and R. Entry $X_{i,j}$ in contact map X is defined as

$$X_{i,j} = \min_{h_i \in i, h_j \in j} \left(d_{h_i h_j} \right), \qquad (1)$$

which is the minimum distance between all heavy atoms, h_i and h_j, of residues i and j. In our contact maps, we include all the residues that have at least one contact with the minimum distance <10 Å. The residues with contacts more than 10 Å are not included in the contact maps. Formally, their distance is treated as infinite and stretches of residues with contacts more than 10 Å are represented as one entry in the contact map.

The 10 Å cutoff is much more generous than the conservative 4.5 Å threshold used for defining contacts in single complexes (Lu *et al*., 2003; Pulim *et al*., 2008) that Lu *et al*. optimized for development of statistical scoring functions. The 10 Å cutoff was selected without any optimization and with the sole purpose of avoiding dependence of the contact definition on the conservative distance cutoff, which would make the contact definition sensitive to small differences in distance. The more generous initial cutoff allows for alignment of contacts that may pass the conservative cutoff in one complex but not others. Later, when determining the interaction cores from multiple contact map alignments, we use the conservative cutoff for defining contacts (see below).

Given two contact maps matrices, C and D, our goal is to find the alignment of C and D that maximizes the overlap between interacting residues. Our alignment algorithm uses 2D dynamic programming to optimize the alignment score (Smith and Waterman, 1981). We allow for gaps in the maps by assigning a gap penalty that penalizes gap insertions between highly interacting residues. The justification for this penalty is that adjacent residues that are highly interactive should be part of the same interaction core and therefore should not be split.

The first step in the dynamic programming approach is to create a 4D scoring matrix M, where $M_{i,j,k,m}$ is the maximum score at position *i, j, k, m* [$0 \le i < width\ (C)$, $0 \le j < height\ (C)$, $0 \le k < width\ (D)$, $0 \le m < height\ (D)$]. Entry $M_{i,j,k,m}$ is then determined from previously solved sub-problems as follows:

$$M_{i,j,k,m} = \begin{cases} 0 & if\ i.j.k.m = 0 \\ \text{Max} \begin{pmatrix} M_{i-1,j,k-1,m} + S(i,j,k,m), \\ M_{i,j-1,k,m-1} + S(i,j,k,m), \\ M_{i,j,k-1,m} + w_c(C,i), \\ M_{i,j,k,m-1} + w_r(C,j), \\ M_{i-1,j,k,m} + w_c(D,k), \\ M_{i,j-1,k,m} + w_r(D,m) \end{pmatrix} & \text{otherwise} \end{cases} \qquad (2)$$

where $w_c(X,i)$ is the gap penalty for inserting a gap at column i in contact map matrix X and $w_r(X,j)$ is the gap penalty for inserting a gap at row j in contact map matrix X. In order to ensure that clusters of interacting residues are not split, we assign a high penalty for gap insertions in rows and columns containing a high number of interactions. Specifically, we used the following gap penalty functions:

$$w_c(X,i) = -\sum_j \frac{1}{X_{i,j}^2}, \qquad (3)$$

$$w_r(X,j) = -\sum_i \frac{1}{X_{i,j}^2}$$

$S(i,j,k,m)$ is the similarity score between the interaction at i,j in contact map C and interaction k,m in contact map D. We use the following similarity function:

$$S(i,j,k,m) = \frac{1}{C_{i,j} D_{k,m}}, \qquad (4)$$

Although here we use a simple similarity function based on inter-residue distance within an interaction, one can define a more complex similarity function that incorporates physical and chemical properties of the interacting residues. We note that both the scoring function and gap penalty functions are defined in the same units of inverse square of the distance.

Once all values of M are computed using (2), the optimal alignment of contact maps is determined by backtracking through the scoring matrix as in standard dynamic programming. Movements within the scoring matrix correspond to the following alignment actions:

Change in (i,j,k,m)	Contact map C	Contact map D
(+1, 0, +1, 0)	align column i	align column k
(0, +1, 0, +1)	align row j	align row m
(0, 0, +1, 0)	gap at column i	
(0, 0, 0, +1)	gap at row j	
(+1, 0, 0, 0)		gap at column k
(0, +1, 0, 0)		gap at row m

An optimal alignment ${}_{it}A$ of contact maps is a mapping $A(i,j) = (a(i), b(j))$ of the pair of (i,j) indices from a complex C onto the (k, m) pair in a complex D where $(k, m) = (a(i), b(j))$.

Multiple alignment of contact maps is accomplished using the same neighbor-joining algorithm as in CLUSTALW (Higgins and Sharp, 1988), but with similarity of contact maps as a distance metric:

$$d_{C,D} = \frac{1}{ij} \sum_{i,j,k,m} S\left(i,j,k = A(i), m = A(j)\right) \qquad (5)$$

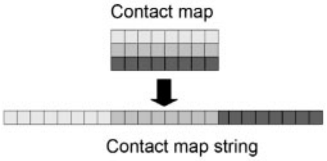

Fig. 2. Mapping of contact map entries onto a string.

The final step of our algorithm identifies core regions within each of the interface sequences. We consider two consecutive residues in one sequence to be part of the same core if they both interact with the same residue in the second sequence. For a given SCOPPI family consisting of complexes of proteins $\{L\}$ and $\{R\}$, let i and j denote the aligned positions among all contact maps between the $\{L\}$ and $\{R\}$ proteins. The residue positions i and $i+1$ from a set of aligned 'ligand' sequences $\{L\}$ belong to the same interaction core if for some residues j from the 'receptor' sequences $\{R\}$, the contact map distance is <4.5 Å for some complexes in the family. That is, each of the adjacent residues i and $i+1$ from $\{L\}$ have to be in contact with the same residue j in $\{R\}$ in at least one complex, but not necessarily in the same complex. A similar definition is applied to define interaction cores in $\{R\}$ sequences. Thus the interaction cores consist of contiguous stretches of aligned residues within the $\{L\}$ and $\{R\}$ protein sequences.

2.2 Performance analysis

The CMAPi algorithm explores the entire space of possible alignments between contact maps and aligns two contact maps by optimizing the similarity function (4). Thus we can claim that CMAPi is optimal for pairwise alignment provided that the residues are assumed to be ordered sequentially. This is indeed the case for our contact maps, which are constructed sequentially. In fact, by concatenating the rows in a contact map and creating a 1D sequence of contacts, the CMAPi algorithm can be mapped to a specific case of 1D sequence alignment with a complex, position-dependent gap penalty structure as defined by Equation (3). Figure 2 illustrates the mapping of a 2D contact map onto a 1D string.

The gap penalty function would record the features of the contact map, i.e. the number and distances of contacts. In the case of multiple alignment of contact maps, although the neighbor-joining method is not optimal, it has been shown in practice to perform nearly as well as an optimal, exhaustive search for multiple sequence alignment (Gascuel and Steel, 2006). In terms of computational complexity, while single-chain contact map alignment [introduced by Godzik *et al*. (1992)] has been shown to be NP-hard by Goldman *et al*. (1999), PPI interface alignment is tractable because gap insertions in the two interacting protein sequences defining the contact maps are independent. Thus, Equation (2) does not apply to the single chain case since a gap in the first sequence L would imply the same gap in the second sequence R, identical to the first $L=R$ in that case. Furthermore, multiple alignment is also tractable since we are using the polynomial-time neighbor-joining algorithm. Specifically, pairwise alignment takes time $O(k^4)$, where k is the number of interface contacts in a protein complex, and multiple alignment, $O(k^4m^2)$ given the m contact maps. In the SCOPPI database, we found that on average $k \cong 43$ and $m \cong 4$: See Figure 3 for the distributions of k and m.

2.3 Evaluation of interface alignments

The interface alignment accuracy is measured using the IRACC function defined in Pulim *et al*. (2008):

$$\mathrm{IRACC} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{iracc}_{ij} \qquad (6)$$

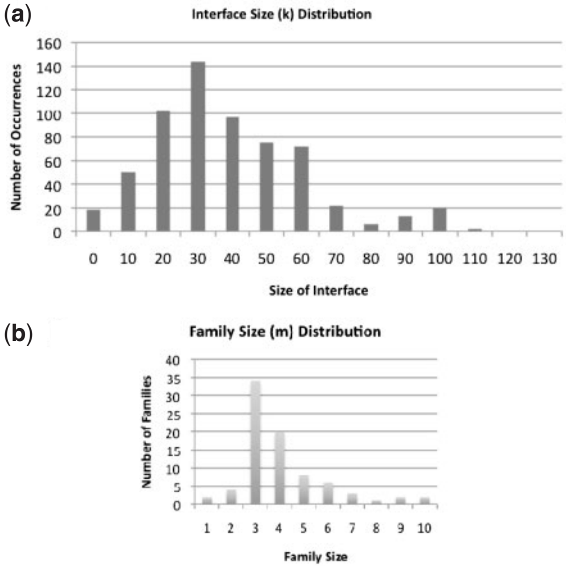

Fig. 3. (**a**) Distribution of k (interface size). The average interface size is 43 contacts. (**b**) Distribution of m (family size). The average family size is 4.1 complexes.

where iracc_{ij} is the alignment accuracy for a pair of template complexes i,j and is defined as

$$\mathrm{iracc}_{ij} = \frac{n_{\mathrm{align}}(i,j)}{n_{\mathrm{min}}(i,j)}.$$

$n_{\mathrm{align}}(i,j)$ is the number of aligned contacts between two complexes and $n_{\mathrm{min}}(i,j)$ is the minimum number of contacts in complexes i and j. The complex with the smaller number of contacts determines how many contacts can be aligned in a best case scenario. Here, the contacts are interacting residues with a distance between any pair of heavy atoms $X_{i,j} < 4.5$ Å. We note that contact maps are aligned by optimizing the distance-dependent scoring function defined by Equations (3) and (4), which is different than the distance-independent IRACC measure used for evaluation of the alignments: the scoring function encodes the shortest distance between any two heavy atoms of contacting residues. Furthermore, the contact map alignment includes all contacts up to 10 Å, while evaluation includes only contacts up to 4.5 Å.

3 RESULTS

To evaluate the accuracy of our dynamic programming algorithm, we compared our results to those produced by MAPPIS (Shulman-Peleg and Shatsky, 2005) and LTHREADER (Pulim *et al*., 2008). We also compared our algorithm to purely sequence-based alignments generated by MUSCLE (Edgar, 2004) as a baseline for our test. MUSCLE was chosen from among many different sequence alignment algorithms for development of the SCOPPI database.

CMAPi generates the most accurate interface alignments for cytokine families, one of the more challenging cases previously investigated by LTHREADER. In this case, since LTHREADER generates different core boundaries than our algorithm, we measured the accuracy of CMAPi and MAPPIS using LTHREADER's core definitions. CMAPi has accuracy almost identical to LTHREADER, while not requiring predefined interaction cores (Table 1).

In addition, CMAPi demonstrates an improvement over MAPPIS of 5% for four-helical bundles and 6% for TNF-like cytokines and

Table 1. Comparison of alignment accuracy, IRACC, for various alignment methods for the four-helical bundle and TNF-like cytokine families

	4-Helical Bundles	TNF-Like
CMAPi	0.84	0.72
LTHREADER	0.85	0.70
MAPPIS	0.80	0.64
MUSCLE	0.73	0.62

In both cytokine families, our CMAPi algorithm achieves higher alignment accuracy than MAPPIS and MUSCLE and the same accuracy as LTHREADER.

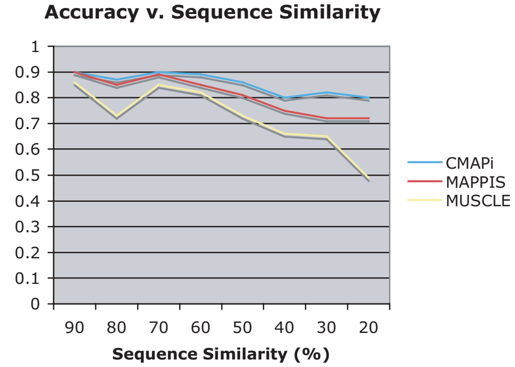

Fig. 4. Comparison of alignment accuracy (IRACC) versus the similarity of sequences within complexes using the CMAPi, MAPPIS and MUSCLE algorithms. Out of 67 SCOPPI complex families only the 47 that have been aligned by MAPPIS are used for comparison.

much higher accuracy than MUSCLE (over 11% for four-helical bundles and 10% for TNF-like cytokines). The use of structural information by both MAPPIS and CMAPi leads to significantly better alignments when compared to MUSCLE, which uses only sequence information.

We also investigated alignment accuracy as a function of sequence identity and demonstrated CMAPi's superior performance as sequence identity declines (Fig. 4) using complexes classified in SCOPPI database. Sequence identity was calculated over the full length of alignments generated using MUSCLE. To evaluate the influence of sequence identity on the performance of different algorithms (MUSCLE, MAPPIS CMAPi), we aligned complexes from PPI families listed in the SCOPPI database. Only families containing at least three complexes were chosen to ensure that enough structure information was available to generate alignments. The current release of SCOPPI contains 63 such families. Results from this evaluation are shown in Figure 4, and individual alignments are included in the supplementary website.

While both methods generate significantly better alignments than MUSCLE at all sequence similarities of SCOPPI families, CMAPi performs better than both MAPPIS and MUSCLE when sequence similarity is below 75%. Furthermore, CMAPi improves alignments over MAPPIS by about 5% for structures that are typically considered for homology modeling, where sequence similarity is 50–70%.

We also verified that the IRACC measure of the interaction core alignment generated by CMAPi on SCOPPI correlates well with the more standard RMSD measure (Fig. 5a). Note that RMSD is an independent measure of the accuracy of the alignments and has not been optimized by CMAPi. Moreover, we showed that the RMSD for CMAPi alignments is not appreciably worse than the RMSD for MAPPIS alignments, even though MAPPIS partially optimizes for low RMSD and CMAPi does not (Fig. 5b). Notably, there are some families where CMAPi generates better RMSD than MAPPIS (2 Å versus 3–4 Å), while for several families MAPPIS yields better RMSD (2 Å versus 4 Å). This result indicates that CMAPi may also be useful as a first step in building detailed homology models of protein interfaces from multiple complex structures.

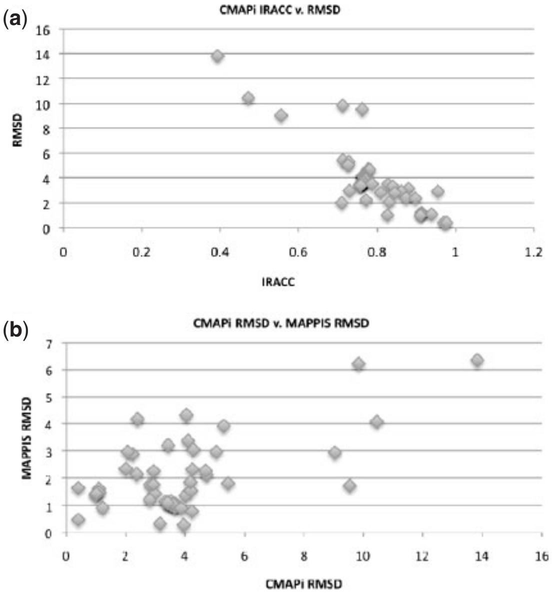

Fig. 5. (**a**) Comparison of CMAPi IRACC versus RMSD. (**b**) Comparison of RMSD of interface alignments using CMAPi and MAPPIS. The interface is defined by CMAPi's core regions. Out of 67 SCOPPI complex families only the 47 that have been aligned by MAPPIS are used here.

4 DISCUSSION

We have shown that the alignment accuracy of our CMAPi algorithm is higher than other existing interface alignment algorithms and in particular MAPPIS. Our algorithm is optimal for pairwise alignment of contact maps and near-optimal in practice for multiple alignment, while having polynomial-time complexity. We believe our method generates better alignments of interacting residues due to its use of a contact map representation of protein interfaces instead of the all-atom-based representation used by MAPPIS. The all-atom representation is helpful in situations where the fine details of the structure can be predicted with high confidence, such as homology modeling of very similar proteins. However, in the case when fine details cannot be predicted accurately, representations using coarser features, such as contact maps, lead to better predictions. CMAPi is tolerant to conformational changes and thus aligns more of the interaction surface.

In future work, we intend to use the CMAPi alignment algorithm to build profiles for every family of interacting proteins defined in the publicly available SCOPPI database. For each family, we will use the multiple alignment of contact maps corresponding to each PPI complex within the family and generate aligned core regions within each sequence pair. The aligned cores will then be used to derive sequence profiles that will be used for PPI prediction as described in LTHREADER. The improvements in the alignment of interacting residues for sequences with 50–70% similarity indicate that CMAPi could also be helpful in building better homology models of protein–protein interfaces when multiple complexes having different interface conformations are available as templates. In this work, we have used a pre-existing classification of protein–protein binding modes provided by SCOPPI. In the future, we will investigate if CMAPi can be used to classify protein binding modes based on contact map similarity.

ACKNOWLEDGEMENTS

We would like to thank Jinbo Xu for helpful discussions.

Funding: National Institute of Health (1R01GM081871-01A1).

Conflict of Interest: none declared.

REFERENCES

Bienkowska,J. and Lathrop,R. (2005) Threading algorithms. In: Dunn,M. *et al*. (ed.). *Encyclopedia of Genetics, Genomics, Proteomics, and Bioinformatics*. Wiley, Hoboken, Nj, USA.

Bowie,J.U. *et al*. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Dunbrack,R.L. Jr. (2006) Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 374–384.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Gascuel,O. and Steel,M. (2006) Neighbor-joining revealed. *Mol. Biol. Evol.*, **23**, 1997–2000.

Godzik,A. *et al*. (1992) Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.*, **227**, 227–238.

Goldman,D. *et al*. (1999) Algorithmic aspects of protein structure similarity. In *40th Annual Symposium on Foundations of Computer Science, 1999.* IEEE Computer Scoiety, Piscataway, NJ, USA, pp. 512–521.

Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.

Jones,D.T. *et al*. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Lu,H. *et al*. (2003) Development of unified statistical potentials describing protein-protein interactions. *Biophys. J.*, **84**, 1895–1901.

Lu,L. *et al*. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350–364.

Pieper,U. *et al*. (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.

Pulim,V. *et al*. (2008) LTHREADER: prediction of extracellular ligand-receptor interactions in cytokines using localized threading. *Protein Sci.*, **17**, 279–292.

Shulman-Peleg,A. and Shatsky,M. (2005) MAPPIS: multiple 3D alignment of protein-protein interfaces. In *Computational Life Sciences: First International Symposium, CompLife 2005, Konstanz, Germany, September 25-27, 2005: Proceedings*. *Spinger Lec. Notes in Comp. Sci.*, Konstanz, Germany.

Singh,R. *et al*. (2006) Struct2Net: integrating structure into protein-protein interaction prediction. *Pac. Symp. Biocomput.*, **13**, 403–414.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Winter,C. *et al*. (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.

Xu,J. *et al*. (2003) RAPTOR: optimal protein threading by linear programming. *J. Bioinform. Comput. Biol.*, **1**, 95–117.