# Protein Structure Alignment using Mean Field Annealing

Mattias Ohlsson

Complex Systems Division, Department of Theoretical Physics
Lund University, Sölvegatan 14A, S-223 62 Lund
Sweden
mattias@thep.lu.se http://www.thep.lu.se/complex/

*Abstract:* - Unraveling functional and ancestral relationships between proteins as well as structure prediction procedures require powerful protein alignment methods. This paper describes the use of fuzzy alignments when matching protein structures. The method use mean-field annealing optimization of fuzzy alignment variables, based on a cost expressed in terms of distances between aligned atoms and of gaps. The approach performs well when compared to other methods, requires modest CPU consumption, and is robust with respect to choice of iteration parameters for a wide range of proteins.

*Keywords:* - protein structure alignment, dynamical programming, fuzzy assignment, mean field annealing.

## 1 Introduction

Comparative analysis of protein structures is a subject of utmost relevance. It enables the study of functional relationships between proteins and is very important for homology and threading methods in structure prediction. Furthermore, grouping protein structures into fold families and subsequent tree reconstruction may unravel ancestral and evolutionary issues.

Pairwise structure alignment amounts to matching two 3D structures such that potential common substructures, e.g. $\alpha$-helices, have priority (see Fig. 1). The latter is accomplished by allowing for gaps in either of the chains. At first sight, the problem may appear very similar to sequence alignment, as manifested in some of the vocabulary (gap costs etc.). However, from an algorithmic standpoint there is a major difference. Whereas sequence alignment can be solved within polynomial time using dynamical programming methods [1], this is not the case for structure alignment since rigid bodies are to be matched. Hence, for all structure alignment algorithms the scope is limited to high quality approximate solutions. This paper will describe the use of *fuzzy alignments* [2] as an approach to pairwise structure alignment. We will also briefly look at an extension into the problem of multiple structure alignments.

## 2 Methods

Consider two proteins with $N_1$ and $N_2$ atoms that are to be structurally aligned. We denote by $\mathbf{x}_i^{(1)}$ $(i = 1,...,N_1)$ and $\mathbf{x}_j^{(2)}$ $(j = 1,...,N_2)$ the atom coordinates of the first and second chain, respectively. The phrase "atom" is here used in a generic sense – it could represent individual atoms but also groups of atoms. In our applications it will mean

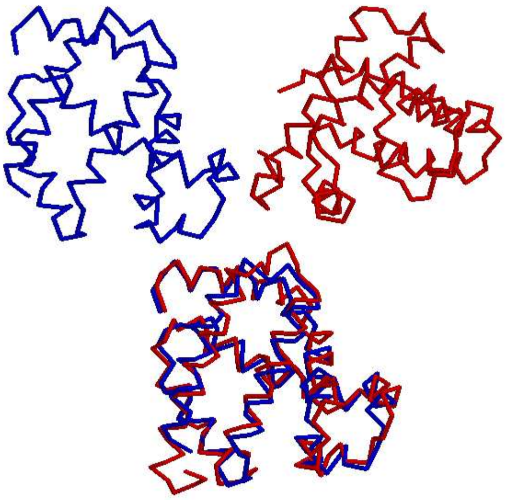

Figure 1: An example of pairwise structure alignment. The two proteins 1ECD (upper left) and 1MBD (upper right) are to be structurally matched to each other. A possible alignment is shown in the lower part. The proteins are in the backbone representation.

$C_\alpha$-atoms along the backbone. A square distance metric between the chain atoms is used,

$$d_{i,j} = |\mathbf{x}_i^{(1)} - \mathbf{x}_j^{(2)}|^2 \qquad (1)$$

## 2.1 Pairwise alignment procedure

The pairwise alignment between the two proteins will be accomplished using a series of weighted rigid body transformations of one of the chains, keeping the other one fixed. This method is similar to the dynamical programming method for global sequence alignment [1], but with two important differences. First, instead of using a score between aligned atoms, a cost formulation is used. This cost, which depends on the distances between the atoms and on the number of gaps and their locations, is changing throughout the alignment procedure. Second, in the original Needleman–Wunsch algorithm an optimal alignment path is calculated, whereas fuzzy alignment paths are computed here. It can be summarized in a 2-step iterative procedure as follows:

- Calculation of a *fuzzy assignment* matrix $\mathcal{W}$, where element $\mathcal{W}_{i,j} \in [0,1]$ is the probability that atom $i$ in the first chain is matched to atom $j$ in the second.

- Rigid body transformation of one of the chains using a fixed $\mathcal{W}$.

While iterating step 1 and 2 above, an annealing of a *temperature* parameter $T$ is performed. The $T$ parameter controls the fuzziness of the assignment matrix $\mathcal{W}$.

## 2.2 Rigid body transformation

The aim with the rigid body transformation is to minimize the following chain error function,

$$E_{\text{chain}} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \mathcal{W}_{i,j} \left( \mathbf{a} + \mathcal{R}\mathbf{x}_i^{(1)} - \mathbf{x}_j^{(2)} \right)^2 , \qquad (2)$$

where $\mathcal{R}$ is a rotation matrix and $\mathbf{a}$ is a translation vector. This minimization problem can be solved exactly with closed-form expressions for $\mathcal{R}$ and $\mathbf{a}$ that minimizes $E_{\text{chain}}$ [3].

## 2.3 The fuzzy assignment matrix

The structure alignment of two proteins is carried out in an annealing procedure, controlled by a *temperature* parameter $T$. Let $\mathcal{D}_{i,j}$ denote a fuzzy generalization of the optimal alignment cost at node $(i,j)$ in the *dot-matrix*, used to represent all possible alignments of two proteins (see Fig. 2).

$D_{i,j}$ is given by,

$$\mathcal{D}_{i,j} = \sum_{l=1}^{3} v_{i,j;\,l}\, \widetilde{\mathcal{D}}_{i,j;\,l} \ , \qquad (3)$$

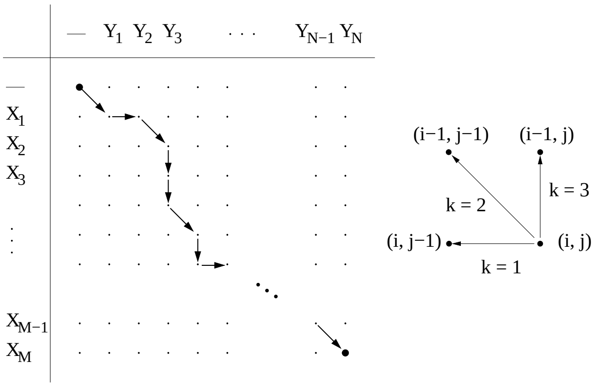

Figure 2: Aligning two chains. $(A)$ The alignment matrix for an alignment between the two chains $\mathbf{X} = (X_1 X_2 \dots X_M)$ and $\mathbf{Y} = (Y_1 Y_2 \dots Y_N)$. $(B)$ Unit vectors connecting to the three possible predecessors to a dot $(i, j)$.

where $\widetilde{\mathcal{D}}_{i,j;\ l}$ is the corresponding generalized fuzzy alignment cost if the alignment path is forced to pass through the preceeding node given by $l$ (see Fig. 2(B). In the Needleman–Wunsch algorithm only the optimal direction $l$ is used, which implies that $v_{i,j;\ l}$ are integers and $\sum_l v_{i,j;\ l} = 1$. This restriction is relaxed in the fuzzy alignment method where $v_{i,j;\ l} \in [0, 1]$, but still sum up to unity. These so called *mean field* variables are calculated according to,

$$v_{i,j;\ l} = \frac{e^{-\widetilde{\mathcal{D}}_{i,j;\ l}/T}}{\sum_{l'} e^{-\widetilde{\mathcal{D}}_{i,j;\ l'}/T}} \ . \tag{4}$$

The generalized fuzzy alignment costs $\widetilde{\mathcal{D}}_{i,j;\ l}$ are calculated using the following recursive relation,

$$\begin{aligned} \widetilde{\mathcal{D}}_{i,j;\ 1} &= \mathcal{D}_{i,j-1} + \lambda_j^{(2)}(1 - v_{i,j-1;\ 1}) + \lambda_{\text{ext}} v_{i,j-1;\ 1} \ , \\ \widetilde{\mathcal{D}}_{i,j;\ 2} &= \mathcal{D}_{i-1,j-1} + d_{i,j} \ , \\ \widetilde{\mathcal{D}}_{i,j;\ 3} &= \mathcal{D}_{i-1,j} + \lambda_i^{(1)}(1 - v_{i-1,j;\ 3}) + \lambda_{\text{ext}} v_{i-1,j;\ 3} \ . \end{aligned} \tag{5}$$

Here, $\lambda_a^{(n)}$ is the penalty for matching atom $a$ in chain $n$ to a gap and $\lambda_{\text{ext}}$ is the gap extension penalty.

At each iteration in the annealing procedure of lowering $T$, a fuzzy assignment matrix $\mathcal{W}_{i,j}$ is calculated as

$$\mathcal{W}_{i,j} = P_{i,j} v_{i,j;\ 2} \ , \tag{6}$$

where $P_{i,j}$ is the probability that node $(i, j)$ is part of the optimal path and $v_{i,j;\ 2}$ is the probability that atoms $i$ and $j$ are locally matched. In other words, the probability for matching atom $i$ in the first chain and atom $j$ in the second chain is the product of the probability that $(i, j)$ is part of the optimal path and the probability that this pair is locally matched.

$P_{i,j}$ can be calculated with a similar recursive relation as for $\mathcal{D}_{i,j}$. With the obvious initial value $P_{M,N} = 1$, one has

$$\begin{aligned} P_{i,j} &= v_{i,j+1;\ 1} P_{i,j+1} \\ &+ v_{i+1,j+1;\ 2} P_{i+1,j+1} \\ &+ v_{i+1,j;\ 3} P_{i+1,j} \ . \end{aligned} \tag{7}$$

## 2.4 Multiple structure alignment

The idea of fuzzy alignments can be used for the problem of multiple structure alignment [4]. In this approach a virtual *consensus* chain is constructed and each of the $K$ proteins is aligned to this consensus chain using the pairwise alignment method described above. This multiple structure alignment method is performed in a similar fashion as for the pairwise case:

- Fuzzy pairwise structure alignment of each of the $K$ chains to the common consensus chain.
- Weighted rigid body rotations and translations for each of the $K$ chains.
- Calculation of a new consensus chain.

The coordinates for the consensus chain are calculated using the coordinates of all the $K$ proteins and the $K$ fuzzy assignment matrices from each of the pairwise alignments (see [4]).

## 3 Results

To test the quality of the pairwise alignment algorithm, we have compared alignments of protein pairs with results from other automatic procedures. Figure 3 shows a comparison between YALE ALIGNMENT SERVER[*] [5], DALI[†] [6] and CE[‡] [7] on a subset of the protein pairs tested in [2]. Here the algorithm was run with varying sizes of the gap penalty parameter, which resulted in alignments with different number of aligned atoms $N$ and root mean squared distance ($rms$) between aligned atoms. Comparing to the other methods we typically found a lower $rms$ for the same $N$.

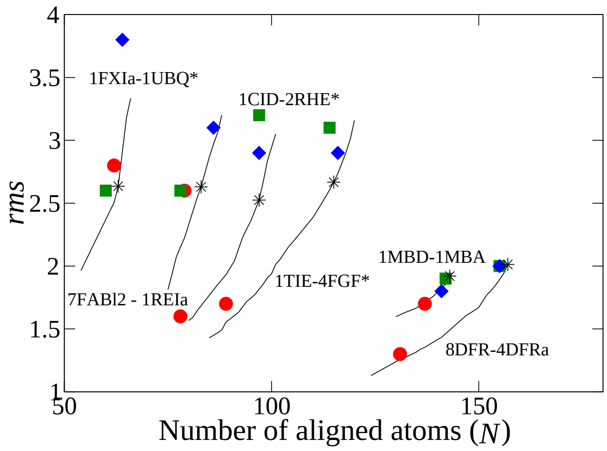

Figure 3: Alignment results for a set of protein pairs in terms of $rms$ and number of aligned atoms ($N$). The results from YALE (red circles), DALI (green squares), CE (blue diamonds), and our method (black asterisks) are plotted. For comparison, different $rms$-$N$ pairs for our algorithm, obtained by varying the size of the gap penalty parameter are shown (solid lines).

An example when using the multiple structure alignment method is shown in Fig. 4, where the triosephosphate isomerase family [8] is aligned. This family consists of 10 proteins belonging to the $\alpha/\beta$-barrel structure class.Comparing to the manual alignment from the HOMSTRAD database [9], we align 236 of the total 260 columns correctly. A detailed investigation shows that all $\alpha$-helix and $\beta$-sheet structures are correctly aligned. Misalignments occurs in loop regions between $\alpha$-helix and $\beta$-sheets.

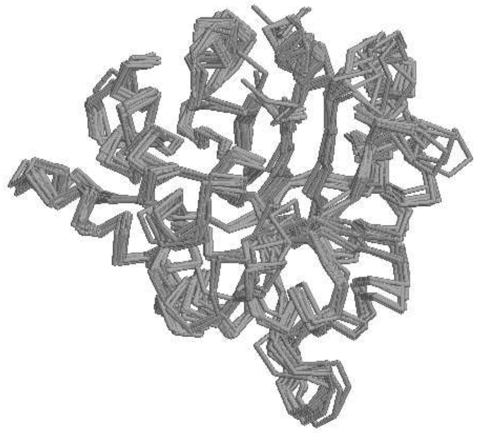

Figure 4: Alignment of the 10 proteins in the triose phosphate isomerase family. Each protein is optimally aligned to the consensus chain (which is not shown). The proteins are: 1amk, 5timA, 1htiA, 1timA,1ypiA, 1treA, 1ydvA, 1aw2A, 2btmA and 1tcdA.

## 4 Summary

In summary, with the use of fuzzy alignment paths we have developed a new approach to structure alignment of proteins. In addition to very good

[*] http://bioinfo.mbb.yale.edu/align/
[†] http://www.ebi.ac.uk/dali/
[‡] http://cl.sdsc.edu/ce/ce_align.html

performance this approach can provide a probabilistic interpretation of the result without tedious stochastic simulations. This was used in [10] were the fuzzy alignment paths were used to obtain a measure of local reliability in protein sequence alignments.

Furthermore, the fuzzy pairwise alignment method can easily be extended to handle more detailed chain representations (e.g. side chain orientation) and additional user-provided constraints of almost any kind.

*References:*

[1] S.B. Needleman and C.D. Wunsch, A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *Journal of Molecular Biololgy* **48**, 1970, pp. 443-453.

[2] R. Blankenbecler, M. Ohlsson, C. Peterson and M. Ringnr, Matching Protein Structures with Fuzzy Alignments, *Proceedings of the National Academy of Sciences*, **100**, 2003, pp. 11936-11940.

[3] K.S. Arun, T.S. Huang and S.D. Blostein, Least-Squares Fitting of Two 3-D Point Sets, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9**, 1987, pp. 698-700.

[4] H. Haraldsson and M. Ohlsson, A Fuzzy Matching Approach to Multiple Structure Alignment of Proteins, *LU TP 03-17* (submitted manuscript).

[5] L. Holm and C. Sander, Protein Structure Comparison by Alignment of Distance Matrices, *Journal of Molecular Biololgy*, **233**, 1993, pp. 123-138.

[6] M. Gerstein and M. Levitt, Using Iterative Dynamic Programming to Obtain Accurate Pairwise and Multiple Alignments of Protein Structures, *Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology*, 1996, pp. 59-67.

[7] I. N. Shindyalov and P. E. Bourne, Protein structure by incremental combinatorial extension of the optimal path, *Protein Engineering*, **11**, 1998, pp. 739-747.

[8] E. Lolis and T. Alber and R.C. Davenport and D. Rose and F.C. Hartman and G.A. Petsko, Structure of yeast triosephosphate isomerase at 1.9-A resolution, *Biochemistry*, **29**, 1990, pp. 6609-6618.

[9] K. Mizuguchi and C.M. Deane and T.L. Blundell and J.P. Overington, HOMSTRAD: a database of protein structure alignments for homologous families, *Protein Science*, **7**, 1998, pp. 2469-2471.

[10] M. Schlosshauer and M. Ohlsson, A Novel Approach to Local Reliability of Sequence Alignments, *Bioinformatics*, **18**, 2002, pp. 847-854.