

Multiple Protein Structure Alignment by Deterministic Annealing

Luonan Chen

Osaka Sangyo University, Nakagaito 3-1-1, Daito, Osaka 574-8530, Japan. chen@elec.osaka-sandai.ac.jp

1. Introduction

Although many algorithms have been developed so far for structure alignment problem based on either distance-based methods or vector-based methods, such as iterative dynamical programming, Monte Carlo simulation, graph theory, and mean field equations, only a few methods are published for multiple structure alignment, such as progressive structure alignment and Monte Carlo optimization.

In this paper, we propose a novel method for solving multiple structure alignment problem, based on mean field annealing technique. We define the structure alignment as a mixed integer-programming (MIP) problem with the inter-atomic distances between two or more structures as an objective function[1]. The integer variables represent the matchings among structures whereas the continuous variables are translation vectors and rotation matrices with each protein structure as a rigid body. By exploiting the special structure of continuous partial problem, we transform the MIP into a nonlinear optimization problem (NOP) with a nonlinear objective function and linear constraints, based on mean field equations. To optimize the NOP, a mean field annealing procedure is adopted with a modified Potts spin model[2]. Since all linear constraints are embedded in the mean field equations, we do not need to add any penalty terms of the constraints to the error function. In other words, there is no "soft constraint" in our mean field model and all constraints are automatically satisfied during the annealing process, thereby not only making the optimization more efficiently but also eliminating unnecessary parameters of penalty that usually require careful tuning dependent on the problems.

2 Pairwise Structure Alignment

Let n_1 and n_2 be the atom numbers of two proteins, and $X_j^i = (x_{j,1}^i, x_{j,2}^i, x_{j,3}^i) \in \mathcal{R}^3$ ($i = 1, j = 1, \dots, n_1; i = 2, j = 1, \dots, n_2$) be the atom coordinates of protein chains, which correspond to C_α atoms along the backbones in this paper. A square distance metric between the chain atoms is adopted, i.e. $d_{ij}^2 = |X_i^1 - X_j^2|^2 = \sum_{k=1}^3 (x_{i,k}^1 - x_{j,k}^2)^2$. The coordinate transformation of a rigid body is generally expressed by a translation vector $A \in \mathcal{R}^3$ and a rotation matrix $R \in \mathcal{R}^{3 \times 3}$, i.e., $\hat{X}_i^k = A + R^k X_i^k$ for the atom i of the chain k , where there are three independent variables for the translation vector and the rotation matrix respectively. For pairwise structure alignment, we fix the coordinates of the second protein chain, which is assumed to be longer than the first chain. Therefore, a square distance between the chain atoms is

$$d_{ij}^2 = |A + RX_i^1 - X_j^2|^2 \quad (1)$$

where $A = A^1$ and $R = R^1$. We define binary variables s_{ij} to describe marching of two atoms for $i = 1, \dots, n_1; j =$

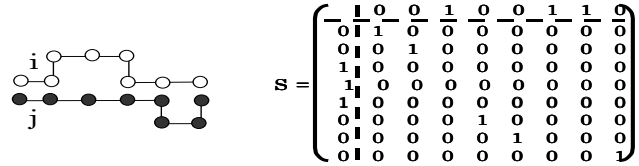


Figure 1. Two protein chains and assignment matrix S

$1, \dots, n_2$:

$$s_{ij} = \begin{cases} 1 & \text{if atom } i \text{ marches } j \text{ in the other chain} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

We also consider the gaps in the protein chains by

$$s_{i0} = \begin{cases} 1 & \text{if atom } i \text{ marches a gap in the other chain} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

In the same way, s_{0j} is defined for $j = 1, \dots, n_2$.

Since each atom in one chain must match one atom in the other (including gap), the following conditions are satisfied.

$$\sum_{i=0}^{n_1} s_{ij} = 1 \text{ for } j = 1, \dots, n_2 \quad (4)$$

$$\sum_{j=0}^{n_2} s_{ij} = 1 \text{ for } i = 1, \dots, n_1 \quad (5)$$

Figure 1 is an example illustrating the notation.

We assume that gap penalties have affine expression for consecutive gaps. Let S represent all binary variables s_{ij} . Then we have an error function for the pairwise structure alignment problem[1],

$$E(S, A, R) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} s_{ij} |A + RX_i^1 - X_j^2|^2 + \sum_{i=1}^{n_1} \lambda_i^1 s_{i0} + \sum_{i=2}^{n_1} (\pi - \lambda_i^1) s_{i-1,0} s_{i0} + \sum_{j=1}^{n_2} \lambda_j^2 s_{0j} + \sum_{j=2}^{n_2} (\pi - \lambda_j^2) s_{0,j-1} s_{0j} \quad (6)$$

where λ_i^1 and λ_j^2 are position-dependent penalties for opening a gap while π is a position-independent penalty for gap extension proportional to the gap length. The first term is total square distances between two protein chains. The second and third lines are gap penalties corresponding to the first and the second chains respectively.

Therefore, pairwise structure alignment is a nonlinear mixed integer programming with objective function eqn.(6) and constraints eqns.(4)-(5) for binary variables s_{ij} , ($i = 0, 1, \dots, n_1; j = 0, 1, \dots, n_2; (n_1 + 1)(n_2 + 1) - 1$ discrete variables) and continuous real variables (A, R : six continuous variables).

3 Mean Field Equations

We adopt mean field annealing[2] to solve the MIP problem by approximating real variable $v_{ij} \in [0, 1]$ to the binary variable $s_{ij} \in \{0, 1\}$. Define a free energy function

$$\begin{aligned} F(V, A, R) &= E(V, A, R) - TH(V) \\ &= E(V, A, R) + T \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} v_{ij} \log v_{ij} \end{aligned} \quad (7)$$

where T is a temperature for the annealing process, and $H(V)$ corresponds to the entropy function.

Then the MIP is changed to the following nonlinear programming problem

$$\text{minimize } F(V, A, R) \quad \text{for } V, A, R \quad (8)$$

$$\text{subject to } \sum_{i=0}^{n_1} v_{ij} = 1 \quad (j = 1, \dots, n_2) \quad (9)$$

$$\sum_{j=0}^{n_2} v_{ij} = 1 \quad (i = 1, \dots, n_1). \quad (10)$$

The conditions of optimality for the Karush-Kuhn-Tucker (KKT) are

$$\frac{\partial E}{\partial A} = 0; \quad \frac{\partial E}{\partial R} = 0 \quad (11)$$

$$\frac{\partial E}{\partial v_{ij}} + T(\log v_{ij} + 1) + \alpha_j + \beta_i = 0 \quad (i = 0, \dots, n_1; j = 0, \dots, n_2) \quad (12)$$

$$\sum_{i=0}^{n_1} v_{ij} = 1 \quad (j = 1, \dots, n_2); \quad \sum_{j=0}^{n_2} v_{ij} = 1 \quad (i = 1, \dots, n_1) \quad (13)$$

where α_j and β_i are Lagrange multipliers corresponding to eqn.(9) and eqn.(10) respectively, and $\alpha_0 = \beta_0 = 0$. Note that there are three independent variables in R , and $\frac{\partial E}{\partial R}$ are the partial derivatives for these three variables.

From eqn.(11), there exist differentiable functions f and g for a given V

$$A = f(V); \quad R = g(V) \quad (14)$$

provided that the Jacobian matrix of eqn.(11) for A and R is not singular. Define

$$u_{ij} = -\frac{1}{T} \frac{\partial E}{\partial v_{ij}} \quad (15)$$

for $i = 0, \dots, n_1; j = 0, \dots, n_2$. From eqns.(12) and the second eqn. of (13), we get

$$v_{ij} = \frac{e^{u_{ij}} / w_j}{\sum_{k=0}^{n_2} e^{u_{ik}} / w_k} \quad (16)$$

for $i = 0, \dots, n_1; j = 0, \dots, n_2$, where $w_j = e^{\alpha_j / T}$. Note $w_0 = 1$ due to $\alpha_0 = 0$.

Substituting eqn.(16) into the first eqn. of eqn.(13), for $i = 0, \dots, n_1; j = 0, \dots, n_2$

$$w_j = \sum_{i=0}^{n_1} \frac{e^{u_{ij}}}{\sum_{k=0}^{n_2} e^{u_{ik}} / w_k} \quad (17)$$

Therefore, we obtain the mean field equations (14)-(17), which are equivalent to KKT conditions. All constraints are automatically satisfied provided that eqns.(16)-(17) hold. Besides, it is easy to show that eqns.(16)-(17) are scale invariant for w_j , i.e. $w_j \rightarrow kw_j$ does not affect v_{ij} . Notice that E can be expressed as $E(V) = E(V, A(V), R(V))$ due to eqns.(14).

4 Deterministic Annealing and Algorithm

From the analysis of the previous section, we get the following algorithm straightforward.

- STEP-0: initialization. Set $T(0), \gamma$ ($0 < \gamma < 1$: cooling coefficient), λ, π , and all initial values of variables v_{ij}, A, R . Let iteration index $t = 1$.
- STEP-1: solve $u_{ij}(t)$ at iteration t by iterative equations

$$u_{ij}(t) = -\frac{1}{T(t-1)} \frac{\partial E(V(t-1))}{\partial v_{ij}} \quad (18)$$

- STEP-2: solve w_j either by Newton method from eqn.(17) or by the following iterative equation until converged. That is, iteratively calculate

$$w_j^{new} = \sum_{i=0}^{n_1} \frac{e^{u_{ij}(t)}}{\sum_{k=0}^{n_2} e^{u_{ik}(t)} / w_k^{old}} \quad (19)$$

If converged, rescale $w_j(t)$ to be $\sum_{j=0}^{n_1} w_j = 2$.

- STEP-3: solve eqn.(16) at iteration t for $V(t)$.
- STEP-4: If V is converged, terminate the computation. Otherwise, reduce the temperature T by $T(t) = \gamma T(t-1)$, and let $t \rightarrow t + 1$ and then goto STEP-1.

Generally, when converged with a sufficient low T , v_{ij} will eventually approach either 0 or 1 due to eqns.(15)-(16), which are then taken as s_{ij} . On the other hand, (A, R) are calculated by eqn.(14). Since all constraints are embedded in the mean field equations, there is no "soft constraint" in our mean field model and all constraints are automatically satisfied during the annealing process, which implies that the proposed approach is much efficient and can be used for exact structure alignment.

In addition, we can extend the algorithm as well as the equations to the multiple structure alignment with a few changes in the same manner, by defining additional binary variables corresponding to the multiple chains.

5 Conclusion

This paper developed a new method for solving protein structure alignment problem, based on mean field annealing technique. The proposed model is general and treats the structure alignment in a more exact manner with implicit complete exploration of the entire space. We have tested our approach to several small-size alignment problems for preliminary study, which verified the efficiency and effectiveness of our algorithm.

References

- [1] M. Ohlsson, C.Peterson, M.Ringner, R.Blankenbecler, "A Novel Approach to Structure Alignment," *LU TP 00-07, SLAC-PUB-8429*, 2000.
- [2] K.Urahama, "Analog method for solving combinatorial optimization problems," *IEICE, E77-A*, pp.302-308, 1994.