

## Problem setting

$$f(M, \mathbf{S}, G) = -\frac{1}{2} \sum_{a_k \in G} \sum_{a_m \in G} \sum_{A_i \in \mathbf{S}} \sum_{A_j \in \mathbf{S}} M_{ki} M_{mj} C(e_{km}, E_{ij}) - \alpha \sum_{a_k \in G} \sum_{A_i \in \mathbf{S}} M_{ki} C(a_k, A_i)$$

$$C(a_k, A_i) = \begin{cases} 0 & A_i = \emptyset \wedge a_k = \emptyset \\ A_i[a_k] - \beta \|T\hat{a}_k - \hat{A}_i\|_2^2 & A_i \neq \emptyset \wedge a_k \neq \emptyset \\ \rho_1 & A_i \neq \emptyset \wedge a_k = \emptyset \\ \rho_2 & A_i = \emptyset \wedge a_k \neq \emptyset \end{cases}$$

In  $\|T\hat{a}_k - \hat{A}_i\|_2^2$ ,  $T$  is the transformation matrix (rotation + translation),  $\hat{a}_k = [x_{a_k}, y_{a_k}, z_{a_k}, 1]^T$  represents the expanded coordinates of the sample node  $a_k$ , and  $\hat{A}_i = [x_{A_i}, y_{A_i}, z_{A_i}]^T$  represents the coordinates of the model node  $A_i$ ,

$$T\hat{a}_k - \hat{A}_i = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} \\ t_{21} & t_{22} & t_{23} & t_{24} \\ t_{31} & t_{32} & t_{33} & t_{34} \end{bmatrix} \begin{bmatrix} x_{a_k} \\ y_{a_k} \\ z_{a_k} \\ 1 \end{bmatrix} - \begin{bmatrix} x_{A_i} \\ y_{A_i} \\ z_{A_i} \end{bmatrix} = \begin{bmatrix} t_{11}x_{a_k} + t_{12}y_{a_k} + t_{13}z_{a_k} + t_{14} - x_{A_i} \\ t_{21}x_{a_k} + t_{22}y_{a_k} + t_{23}z_{a_k} + t_{24} - y_{A_i} \\ t_{31}x_{a_k} + t_{32}y_{a_k} + t_{33}z_{a_k} + t_{34} - z_{A_i} \end{bmatrix}$$

$$\begin{aligned}\|T\hat{a}_k - \hat{A}_i\|_2^2 &= (t_{11}x_{a_k} + t_{12}y_{a_k} + t_{13}z_{a_k} + t_{14})^2 + (t_{21}x_{a_k} + t_{22}y_{a_k} + t_{23}z_{a_k} + t_{24})^2 \\ &\quad + (t_{31}x_{a_k} + t_{32}y_{a_k} + t_{33}z_{a_k} + t_{34})^2\end{aligned}$$

$$\begin{aligned}\frac{\partial \|T\hat{a}_k - \hat{A}_i\|_2^2}{\partial T} &= 2 \begin{bmatrix} t_{11}x_{a_k} + t_{12}y_{a_k} + t_{13}z_{a_k} + t_{14} - x_{A_i} \\ t_{21}x_{a_k} + t_{22}y_{a_k} + t_{23}z_{a_k} + t_{24} - y_{A_i} \\ t_{31}x_{a_k} + t_{32}y_{a_k} + t_{33}z_{a_k} + t_{34} - z_{A_i} \end{bmatrix} \begin{bmatrix} x_{a_k} & y_{a_k} & z_{a_k} & 1 \end{bmatrix} \\ &= 2(T\hat{a}_k - \hat{A}_i)a_k^T\end{aligned}$$

For rigid protein structures, if edges only encode the distance information, we can ignore the first term because of  $\|T\hat{a}_k - \hat{A}_i\|_2^2$  and obtain

$$f(M, \mathbf{S}, G) = - \sum_{a_k \in G} \sum_{A_i \in \mathbf{S}} M_{ki} \left( A_i[a_k] - \beta \|T\hat{a}_k - \hat{A}_i\|_2^2 \right)$$

We can minimize  $f(M, \mathbf{S}, G)$  by iterating the following steps

- Fixing  $\{A_i[a_k]\}$  and the matching  $[M_{ki}]$ , solve the transformation  $T$ .
- Fix  $T$  and  $\{A_i[a_k]\}$ , solve  $M_{ki}$ .
- Fix  $T$  and  $[M_{ki}]$ , solve  $A_i[a_k]$ .

The update equations derived in the following slides are for one protein (i.e., one  $G$ ). To accommodate multiple proteins, we need to sum over the update equations over multiple proteins.

Fixing  $A_i[a_k]$  and the matching matrix  $[M_{ki}]$ , we can solve the transformation  $T$  as below:

$$\frac{\partial f(M, \mathbf{S}, G)}{\partial T} = \beta \sum_{a_k \in G} \sum_{A_i \in \mathbf{S}} M_{ki} (T \hat{a}_k - \hat{A}_i) \hat{a}_k^T$$

$$\begin{aligned} \frac{\partial f(M, \mathbf{S}, G)}{\partial T} = 0 &\Rightarrow T \sum_{a_k \in G} \sum_{A_i \in \mathbf{S}} M_{ki} \hat{a}_k \hat{a}_k^T = \sum_{a_k \in G} \sum_{A_i \in \mathbf{S}} M_{ki} \hat{A}_i \hat{a}_k^T \\ &\Rightarrow T = \left( \sum_{a_k \in G} \sum_{A_i \in \mathbf{S}} M_{ki} \hat{A}_i \hat{a}_k^T \right) \left( \sum_{a_k \in G} \sum_{A_i \in \mathbf{S}} M_{ki} \hat{a}_k \hat{a}_k^T \right)^{-1} \end{aligned}$$

Fix  $T$  and  $\{A_i[a_k]\}$ , solve  $M_{ki}$  by the minimal entropy principle as below

$$f(M, \mathbf{S}, G) = - \sum_{a_k \in G - \{\emptyset\}} \sum_{A_i \in \mathbf{S}} M_{ki} C(a_k, A_i) + \tau \sum_{a_k \in G - \{\emptyset\}} \left( \sum_{A_i \in \mathbf{S}} M_{ki} \log M_{ki} \right)$$

$$\frac{\partial f(M, \mathbf{S}, G)}{\partial M_{ki}} = -C(a_k, A_i) + \tau \log M_{ki} + \tau$$

$$\frac{\partial f(M, \mathbf{S}, G)}{\partial M_{ki}} = 0 \Rightarrow M_{ki} = \frac{1}{Z} \exp \left( \frac{C(a_k, A_i)}{\tau} - 1 \right)$$

$$Z = \sum_{a_k \in G} \exp \left( \frac{C(a_k, A_i)}{\tau} - 1 \right)$$

The smaller the entropy, the more unique the matching.

Fix  $T$  and  $[M_{ki}]$ , solve  $A_i[a_k]$  by the minimal entropy principle as below

$$f(M, \mathbf{S}, G) = - \sum_{a_k \in G - \{\emptyset\}} \sum_{A_i \in \mathcal{S}} M_{ki} \left( A_i[a_k] - \beta \|T\hat{a}_k - \hat{A}_i\|_2^2 \right) \\ + \xi \sum_{a_k \in G - \{\emptyset\}} \left( \sum_{A_i \in \mathcal{S}} A_i[a_k] \log A_i[a_k] \right)$$

Note that we use  $A_i[a_k]$  to denote the probability of  $A_i$  matching with the amino acid denoted by  $a_k$

$$\frac{\partial f(M, \mathbf{S}, G)}{\partial A_i[a_k]} = -M_{ki} + \xi \log A_i[a_k] + \xi$$

$$\frac{\partial f(M, \mathbf{S}, G)}{\partial A_i[a_k]} = 0 \Rightarrow A_i[a_k] = \frac{1}{Z} \exp\left(\frac{M_{ki}}{\xi} - 1\right)$$

$$Z = \sum_{a_k \in G} \exp\left(\frac{M_{ki}}{\xi} - 1\right)$$

# Initialization

- Calculate pair-wise matchings. See the next slide for the details of calculating pair-wise matchings.
- Initialize the model graph by the protein graph that has the best overall match (*i.e.*, sum of the pair-wise match scores) with other protein graphs. In addition, keep the following information:
  - The matchings between the chosen protein graph and all other protein graphs.
  - The transformation matrixes between the chosen protein graph and all other protein graphs.

After the above initialization, update  $A_i[a_k]$ ,  $T$ , and  $M$  by the EM algorithm.

# Initialization

- Given a pair of proteins  $\langle G_m, G_n \rangle$ , perform the following steps:
  - Treat  $G_m$  as the model  $\mathcal{S}$ . Initialize  $A_i[a_k]$  by the BLOSUM62 matrix.
  - Match every subgraph  $g_{m,i}$  centered at the  $i$ -th non-null node of  $G_m$  with every subgraph  $g_{n,k}$  centered at the  $k$ -th non-null node of  $G_n$ . This will generate a local matching score  $s_{m,i,n,k}$ . The matching scores, which involve null-nodes, are defined as the 50% of the distribution of  $s_{m,i,n,k}$ .
  - Initialize the matching between the  $i$ -th node of  $G_m$  and the  $k$ -th node of  $G_n$  as  $M_{m,i,n,k} = \frac{\exp(s_{m,i,n,k})}{\sum_{k'} \exp(s_{m,i,n,k'})}$ .
  - Maximize the matching score  $s(G_m, G_n)$  by iteratively adjusting  $M_{m,i,n,k}$  and  $T_{m,i,n,k}$  as below until converges:

$$T_{m,n} = \left( \sum_{a_k \in G_n} \sum_{A_i \in G_m} M_{ki} \hat{A}_{m,i} \hat{a}_{n,k}^T \right) \left( \sum_{a_k \in G_n} \sum_{A_i \in G_m} M_{ki} \hat{a}_{n,k} \hat{a}_{n,k}^T \right)^{-1}$$

$$M_{m,i,n,k} = \frac{1}{Z} \exp \left( \frac{C(a_{n,k}, A_{m,i})}{\tau} - 1 \right)$$