Lecture 7

Sunday, September 23, 2018 4:02 PM

Protein Structure Alignment

Let Σ be the alphabet of amino acids, ie $|\Sigma| = 20$.

Let $A \in \Sigma^m$ and $B \in \Sigma^n$.

We are given $d_A : [m] \times [m] \to \mathbb{R}^{\geq 0}$

$d_B : [n] \times [n] \to \mathbb{R}^{\geq 0}$.

We are given a distance threshold $\tau \geq 0$.

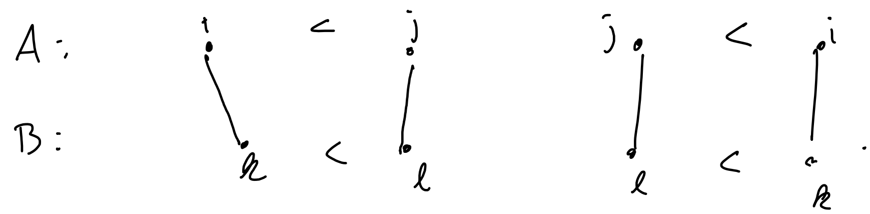
We obtain $C^A = [c^A_{i,j}]$ where

$$c^A_{i,j} = 1 \iff d_A(i,j) \leq \tau .$$

$C^B = [c^B_{k,l}]$ is symmetric.

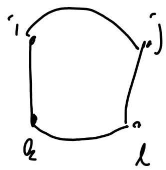
An <u>alignment</u> S of A and B is a set of non-overlapping pairs $(i,k), (j,l) \in S$ with the constraint that

$$(i < j \text{ and } k < l) \text{ or } (j < i \text{ and } l < k)$$



A <u>conserved contact</u> in S is defined by two aligned pairs $(i,k), (j,l) \in S$ s.t.

$C^A_{i,j} = 1$ and $C^B_{k,l} = 1$.



(i,k), (j,l) form a conserved contact

Problem. [Contact Map Overlap (CMO)]

Given contact maps C^A and C^B, find an alignment of A and B with maximum number of conserved contacts.

We can view CMO as a graph problem. There are two graph representations.

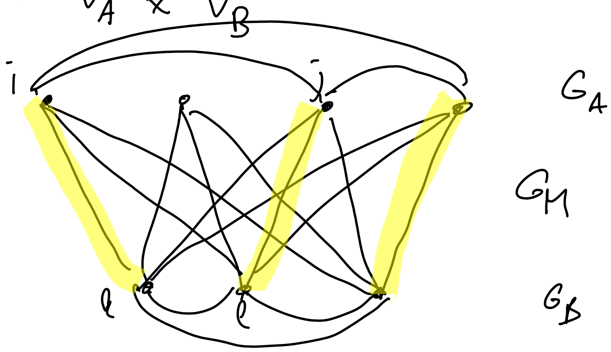
$G_A = (V_A, E_A)$ $\quad |V_A| = m$

$(v_i, v_j) \in E_A \Leftrightarrow C^A_{i,j} = 1$

$G_B = (V_B, E_B)$ is symmetric

① Matching graph $G_M = (V_A \cup V_B, E_M)$

$E_M = V_A \times V_B$



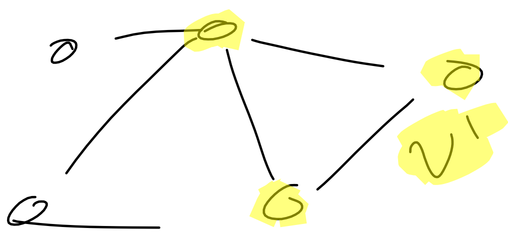
An alignment S is a non-crossing matching in G_M, i.e. for all distinct edges $(v_i, v_k), (v_j, v_l) \in S$ it holds that

$(i < j$ and $k < l)$ or $(j < i$ and $l < k)$.

Given a simple graph $G=(V,E)$

Clique is a subset $V' \subseteq V$ of vertices s.t.

for all distinct $u, v \in V'$ it holds that $(u,v) \in E$



$|V| = n \quad |E| \leq \binom{n}{2} = O(n^2)$

CLIQUE problem

Given $G=(V,E)$ and $k \in \mathbb{N}$,

is there a clique in G of size k?

NP-complete

There is no algorithm

with running time $O(poly(n))$

$\hookrightarrow n^c$ where $c>0$

unless $P = NP$.