# $\boxed{\text{LECTURE \#3}}$

Sept 25 2014. 

1. Architecture Team: Discuss TSP interface.
2. Class: Discussion on TSP. { Go over LP-relaxation }.
3. Optical Mapping:
4. $\boxed{\text{Quiz.}}$
5. Optical Mapping Architecture Team.

## Optical Mapping:

◇ Restriction Map Model:

SMRM (Single Molecule Restriction Map)

A vector with ordered set of rational numbers on the open interval (0,1):

$$\mathcal{D}_j = (s_{1j}, s_{2j}, \ldots, s_{M_j j}),$$

$$0 < s_{1j} < s_{2j} < \cdots < s_{M_j j} < 1. \quad s_{ij} \in \mathbb{Q}$$

◇ Problem

Data: A collection of SMRM vectors:

$$\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$$

Desiderata: Compute a consensus vector

$$H = (h_1, h_2, \ldots, h_N)$$

such that $H$ is "consistent" with each $\mathcal{D}_j$.

$$H^* = \underset{H,j}{\arg\min}\ \mathrm{dist}(\mathcal{D}_j, H).$$

Consensus:

$$H^* = \underset{H,j}{\operatorname{argmin}} \ \operatorname{dist}(D_j, H)$$

$$D_j = (s_{1j}, s_{2j}, \ldots, s_{M_j j})$$

$$\Rightarrow D_j + c = (s_{1j}+c, s_{2j}+c, \ldots, s_{M_j j}+c)$$

$$c \in [0,1) \quad c \in \mathbb{Q} \quad -s_{1j} < c < 1 - s_{M_j j}.$$

$$\operatorname{dist}(D_j, H) = \operatorname{dist}(D_j + c, H)$$

$$D_j^R = (1 - s_{M_j j}, \ldots, 1 - s_{2j}, 1 - s_{1j})$$

$$\operatorname{dist}(D_j^R, H) = \operatorname{dist}(D_j, H).$$

Consensus:

$$H^* = \underset{H,j}{\operatorname{argmin}} \left\{ \operatorname{dist}(D_j, H), \operatorname{dist}(D_j^R, H) \right\}$$

or

$$H^* = \underset{H,j}{\operatorname{arg\,min}} \left\{ \operatorname{dist}(D_j + c, H) \mid -s_{1j} < c < 1 - s_{M_j j} \right\}$$

or

$$H^* = \underset{H,j}{\operatorname{argmin}} \left\{ \operatorname{dist}(D_j + c, H), \operatorname{dist}(D_j^R + c, H) \mid -s_{1j} < c < 1 - s_{M_j j} \right\}$$

Assume some distribution generating $D_j$'s

$\Rightarrow$ Maximum Likelihood formulation $\Rightarrow$.

$$\langle H \rangle = \underset{H}{\operatorname{arg\,min}} \sum_j \min \left\{ \operatorname{dist}(D_j + c, H), \operatorname{dist}(D_j^R + c, H) \mid - s_{1j} < c < 1 - s_{M_j, j} \right\}$$

## TOY EXAMPLE PROBLEM.

(Unknown Orientation:)

Data: A set of ordered vectors with rational entries in the open interval $(0,1)$:

$$D_1, D_2, \ldots, D_\ell, D_{\ell+1}, \ldots, D_m$$

A rational number $p_c \in (0,1)$ and an integer $N$.

An admissible alignment of the data can be represented as

$$D_1', D_2', \ldots, D_\ell', D_{\ell+1}', \ldots, D_m'$$

where

$$\left.\begin{array}{ll} D_j' \in \{D_j, D_j^R\} & (1 \le j \le \ell) \\ \text{and} & \\ D_j' = D_j & (j > \ell) \end{array}\right\} = \text{An Alignment } (A_k)$$

For any rational number $h_i \in [0,1]$, define an indicator variable

$$m_{ijk} = \begin{cases} 1 & \text{if } h_i \in D_j' \\ 0 & \text{otherwise.} \end{cases}$$

Define a characteristic function

$$\chi_k : [0,1] \longrightarrow \{0,1\}$$
$$: h_i \longmapsto \begin{cases} 1 & \text{iff } \sum_j m_{ijk} > p_c m. \end{cases}$$

Desiderata: Find an admissible alignment $A_k$ such that

$$\left| \{ h \in [0,1] \mid \chi_k(h) = 1 \} \right| \geq N.$$

NP-Completeness

Consider an instance of a 3-SAT problem:

With $\ell$ variables:

$x_1, x_2, \ldots, x_\ell$

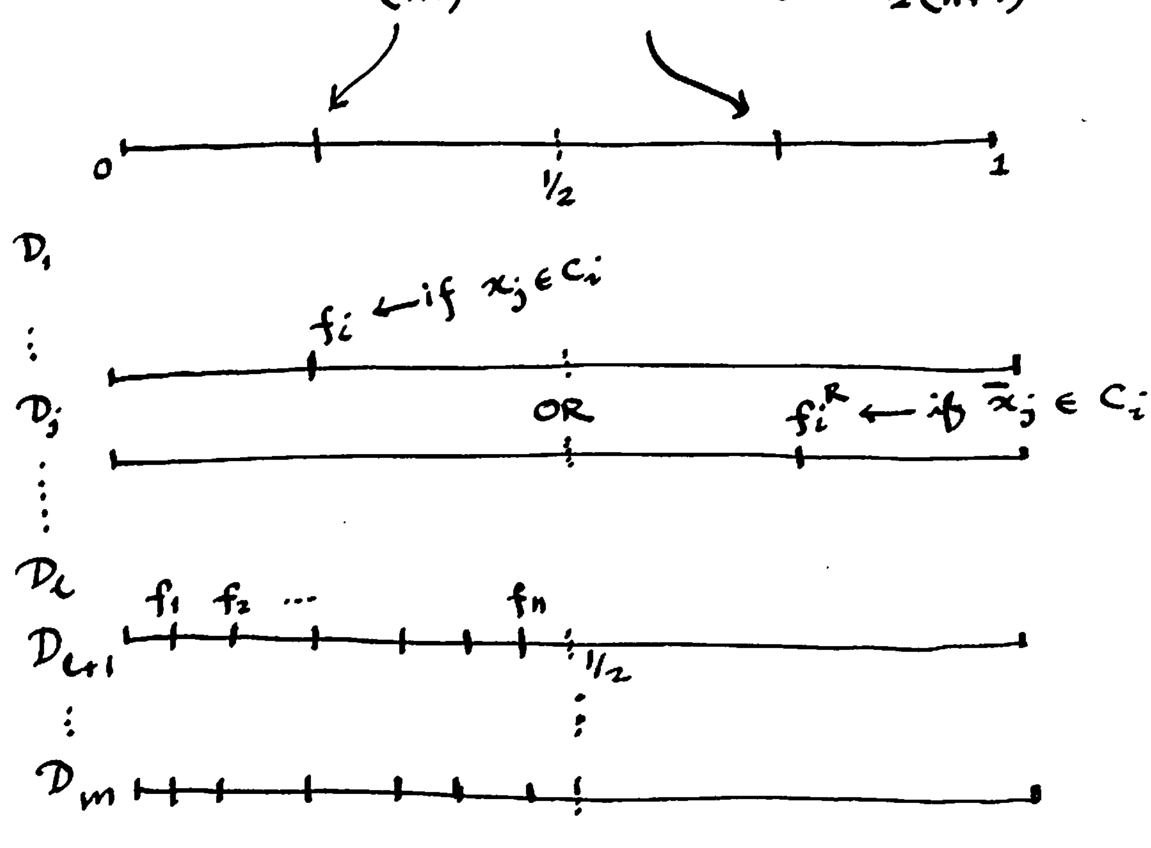
And $n$ clauses:

$C_1, C_2, \ldots C_n \qquad (n \geq \ell)$

- Assume that no clause contains a variable and its negation: $x_j$ and $\bar{x}_j$ (The clause is a tautology $\equiv T$)

- Restriction site associated with a ~~set~~ clause $C_i$.

$$f_i = \frac{i}{2(n+1)} \qquad f_i^R = 1 - f_i = \frac{2n-i+2}{2(n+1)}$$



Create a dataset $D_1, D_2, \ldots, D_\ell, D_{\ell+1}, \ldots, D_m$

with $m = 2\ell - 1$ as follows:

$D_j$ has a cuts at $f_i$ or $f_i^R$, only:

$f_i \in D_j$ iff $x_j \in C_i$

$(f_i^R \in D_j$ iff $\bar{x}_j \in C_i)$

$N \equiv n, \quad p_c = 1/2$

CNF has a satisfying assignment

$\Rightarrow$ Choose an admissible alignment in which

$$D_j' = \left.\begin{cases} D_j & \text{if } x_j = \text{true} \\ D_j^R & \text{if } x_j = \text{false} \end{cases}\right\} 1 \le j \le \ell$$

$$D_j' = D_j, \quad \ell < j \le m.$$

$\therefore$ For every $f_i$, $(1 \le i \le n)$ there are $(\ell-1)$ matches from $D_{\ell+1}, \ldots, D_m$

& at least one more from $D_1', \ldots, D_\ell'$ (since each clause must be satisfied)

$\therefore \forall_{1 \le i \le n}$

$$\sum_j m_{ijk} \ge \ell > \frac{2\ell - 1}{2} = p_c m.$$

$$\Rightarrow \{h \in [0,1] \mid \chi_k(h) = 1\} = \{f_1, f_2, \ldots, f_n\}$$

$$\Rightarrow |h \in [0,1] \mid \chi_k(h) = 1\} = n \ge N.$$

Conversely, if the CNF has no satisfying assignment, then for every admissible alignment there exists an $1 \le i \le n$

$$\forall_k \; \exists_i \qquad \sum_j m_{ijk} = (\ell-1) < p_c m \qquad \text{and}$$

$$\left| \{ h \in [0,1] \mid X_k(h) = 1 \} \right| < n .$$

$\square$.

Problem Generation:

Statistical Model:

- A model or hypothesis $H$.

$$= \{ h_1, h_2, \dots, h_N \}$$

$N \approx 40$.
Distribution for $h_i$'s
Exponential gaps
or uniform gaps.

- $Pr[D_j \mid H]$

$$D_j \sim H.$$

Pairwise Conditional Indep.

$$Pr[D_j \mid D_{j_1}, \dots, D_{j_m}, H] = Pr[D_j, H]$$

- $Pr[\text{bad}]$, $Pr[\text{good}] = 1 - Pr[\text{bad}]$

$$Pr[D_j \mid H] = \frac{1}{2} \sum Pr[D_j^{(k)} \mid H, \text{good}] \, Pr[\text{good}] + \frac{1}{2} \sum Pr[D_j^{(k)} \mid H, \text{bad}] \, Pr[\text{bad}]$$

$(k) \to$ Alignment.

$D_j^{(k)} = D_j$ or $D_j^R$ with equal probability:

$D_j$ = Good $\Rightarrow$

Choose parameters $p_c, \sigma, f$.

$h_i \in H \Rightarrow s_i \sim N(h_i, \sigma)$ with $pr = p_c$.

$s_i$ = absent with $pr = 1 - p_c$.

spurious cuts $\Rightarrow$ ~~Expon~~ Poisson. $e^{-\lambda_f} \dfrac{\lambda_f^{F_{jk}}}{F_{jk}!}$

$D_j$ = Bad $\Rightarrow$

Poisson: $e^{-\lambda_n} \dfrac{\lambda_n^{M_j}}{M_j!}$

$$\Pr\left[D_j^{(k)} \,\middle|\, \begin{matrix} H, \\ \text{good} \end{matrix}\right]$$

$$= \prod_{i=1}^{N} \left[\left(p_c \frac{e^{-(s_{ij} - h_i)^2/2\sigma^2}}{\sqrt{2\pi}\,\sigma_i}\right)^{m_{ijk}} (1 - p_c)^{(1 - m_{ijk})}\right]$$

$$\times e^{-\lambda_f} \lambda_f^{F_{jk}}.$$

$$\Pr\left[D_j^{(k)} \,\middle|\, \begin{matrix} H, \\ \text{bad} \end{matrix}\right] = e^{-\lambda_n} \lambda_n^{M_j}$$