CS 628 biological sequence analysis

Uri Keich

Why align?

- "Nature is a tinkerer not an inventor" Jacob 1977 (Durbin et al.)
- "New sequence are adapted from pre-existing sequences rather than invented *de novo*."
- Homologous sequences help with:
  - Inferring protein function
  - Annotating functional regions

Why don't you just do it?

- Evolution creates insertions deletions and substitutions
- Databases/sequences are huge
- Example from Nature (Brian Golding, McMaster):
  - ------CCTTCAGAATACAGAATAGGGACATAGAGA
    ATCCCACCCAGCCCCCTGGACCTGTAT---------
- Fitch 1984:
  - CCTTCAGAATACAGAATAGGGACATAGAGA
    ATCCCA---CCCAGCCCCCTGGACCTGTAT

Sequence alignments to a fragment of human α-globin (hemoglobin subunit)

(a) Human β-globin

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

(b) Lupin leghemoglobin

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++           +L+ L+++H+ K
LGB2_LUPLU  NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

(c) Red Herring

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
            GS+ + G +   +D L  ++ H+ D+  A +AL D     ++AH+
F11G11.2    GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Key issues

- Which scoring method?
  - domain knowledge, biology, statistics
- How do we find an optimal alignment?
  - algorithms
- What is the statistical significance?
  - probability/statistics

Scoring alignments

- Score for matches, substitutions, and gaps
- Ignore gaps for now
- How shall we compare different scoring methods?

Is this alignment for real?

- Suppose we are given a (global) gapless alignment of two sequences of length n: $\boldsymbol{x}$, $\boldsymbol{y}$.
- Two *iid* models for generating pairs (x_i, y_i):
  - x_i and y_i are iid samples from the distribution q_a (H_0).
  - The pair (x_i, y_i) is drawn from a distribution p_{ab} (H_1).
- Did the alignment $(\boldsymbol{x}, \boldsymbol{y})$ come from H_0 or H_1?
- Neyman-Pearson: your optimal test statistics is the likelihood ratio (LR):

$$\frac{P\big[(\boldsymbol{x}, \boldsymbol{y})|H_1\big]}{P\big[(\boldsymbol{x}, \boldsymbol{y})|H_0\big]} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}.$$

- The LLR statistics $\sum_i s(x_i, y_i)$ where $s(a, b) = \log \frac{p_{ab}}{q_a q_b}$ is just as optimal . . .

Estimating the LLR

- We don't know q_a and p_{ab}.
- How shall we estimate them?
- How do you estimate the probability that a coin lands h?
- You flip it n times and set $\widehat{p_h} = \frac{n_h}{n}$.
- $\widehat{p_h}$ is a *maximum likelihood* estimator.
- This works by the *law of large numbers*:

$$\frac{n_h}{n} \xrightarrow[n\to\infty]{} p_h.$$

- There are other ways to estimate p_h.

Sample generation

- Estimating p_{ab} is trickier.
- We need to generate pairs of aligned sequences.
- We assume our coin flips are *independent* samples from the *true* distribution.
- Which pairs of proteins should we look at: human-chimp or human-fugu?
- There is no universal underlying distribution.
- Suppose our DB contains 1000 instances of almost identical AA sequences.
- We need to be more creative: PAM Dayhoff et al. 1972-8, BLOSUM Henikoff & Henikoff 1992.

BLOSUM X - Henikoff & Henikoff 1992

- BLOck SUbstitution Matrix
- Start with blocks of aligned sequences:

BABA	CBB
BABC	CBB
AACC	ABC
	AAC

- n_{ab} is the number of times the pair a and b appear in the same column of an aligned block.
- $\widehat{p_{ab}} = n_{ab} / \sum_{cd} n_{cd}$.
- So, $\widehat{p_{AA}} = 4/30$, $\widehat{p_{AB}} = 5/30$, . . .
- How do we account for human-chimp vs. human-fugu?
- Anybody knows what the X stands for?

BLOSUM 75

- Cluster the sequences in each block as follows.
- Build a graph whose vertices are the block's sequences.
- Connect any two vertices whose sequences exhibit $\geq 75\%$ identity.
- The clusters are the connected components of this graph.
- The previous example clustered at 75%:

$$\begin{array}{c} \text{BABA} \\ \text{BABC} \\ \hline \text{AACC} \end{array} \qquad \begin{array}{c} \text{CBB} \\ \text{CBB} \\ \hline \text{ABC} \\ \hline \text{AAC} \end{array}$$

BLOSUM 75 cont.

- Recall

$$n_{ab} = \sum_{k,i \neq j} 1_{\{B^k(i)=a\,,\, B^k(j)=b\}},$$

where B^k is the kth block column.

- Redefine

$$n_{ab} = \sum_{k,l \neq m,i,j} 1_{\{B^k_l(i)=a\; B^k_m(j)=b\}} / (|B^k_m||B^k_l|),$$

where B^k_m is the mth cluster of the kth block column.

- $\widehat{p_{ab}}$ is normalized as before.

- Using the new definition we have $\widehat{p_{AA}} = 2/13$, $\widehat{p_{AB}} = 2/13$, $\widehat{p_{AC}} = 2.5/13 = 5/26$, . . .