Pairwise Alignment

(or models and algorithms are your friend)

Lecture 2

6.874J/7.90J/6.807

David Gifford

Gene similarities revealed by dot plot

promoter
conservation

Image removed for copyright reasons.

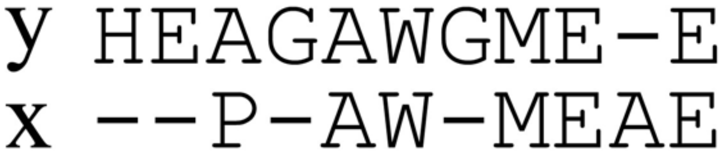
Dot Plots

Align subsequences of *S1* and *S2*; place dot when score is high

Image removed for copyright reasons.

Pairwise Alignment (Global)

Given a query sequence *x*, what is the best alignment to a sequence *y*?



```
y HEAGAWGME-E
x --P-AW-MEAE
```

Protein (20 letters, X, -)

DNA (A, C, G, T, N, -, W, S, R, Y, K, M, B, D, H, V)

RNA (A, C, G, U, -)

q_s probability of symbol *s* occurring at random in a sequence.

Two possible models

- Model R - Random– The sequences are unrelated and were generated by coin flips (biased)
- Model M – Match – The sequences were derived from a common ancestor sequence

A Probabilistic Model of Alignment

$$P(x, y \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

P_{ab}

Joint probability that *a* and *b* have been originally derived from some (unknown) ancestor *c* (might be the same as *a* or *b*)

$$P(x, y \mid M) = \prod_i P_{x_i y_i}$$

The Odds Ratio Statistic (No Gaps)

$$\frac{P(x,y\,|\,M)}{P(x,y\,|\,R)} = \prod_i \frac{P_{x_i y_i}}{q_{x_i} q_{y_i}}$$

$$S = \log\left(\frac{P(x,y\,|\,M)}{P(x,y\,|\,R)}\right) = \log \prod_i \frac{P_{x_i y_i}}{q_{x_i} q_{y_i}}$$

$$S = \sum_i S(x_i, y_i)$$

$$S(a,b) = \log \frac{P_{ab}}{q_a q_b}$$

Example

$$x_1 = C \qquad y_1 = Q$$

$$q_{x_1} = \frac{1}{20} \qquad q_{y_1} = \frac{1}{20} \qquad q_{x_1 y_1} = \frac{1}{400}$$

$$P_{x_1 y_1} = P_{CQ} = \frac{1}{800}$$

$$S_{x_1 y_1} = S_{CQ} = 3\log_2\left(\frac{\frac{1}{800}}{\frac{1}{400}}\right) =$$

$$3\log_2\left(1\!\!\diagup\!2\right) = \quad -3$$

Substitution Matrix

- Logical to think of it in terms of evolutionary time

$$S(a,b\,|\,t)$$

- **PAM (Point Accepted Mutations)**: Based on substitution data from alignment between similar proteins
  - (1% expected substitutions = 1PAM)
  - PAMn = $(1\text{PAM})^n$
- **BLOSUM (BLOck Scoring Matrix)**: Multiple alignment of distantly related proteins
  - BlosumL = Sequences with L% or more of identical residues were clustered to compute log-odds ratio

BLOSUM50

Image removed for copyright reasons.

BLOSUM65

Image removed for copyright reasons.

Gap Penalties

- We can penalize a gap with the function

$$-gd$$

where *g* is the length of the gap

- Typical gap penalties in practice for proteins
  - d=8 third-bits used in Durbin
- We can also add a fixed penalty for opening a gap

Affine gaps

- Assume log odds-ratio of a gap deceases geometrically:

$$f(g) = p(1-p)^{g-1}$$

$$\log f(g) = \log p + (g-1)\log(1-p)$$

$$d = -\log p$$

$$e = -\log(1-p)$$

$$\log f(g) = -d-(g-1)e$$

$$s(g) = -d-(g-1)e$$

Let’s find the best alignment

- To do this we will maximize the score, taking into account our ability to incorporate gaps
- We could enumerate all of the possible alignments…

How many possible alignments exist?

An intercalation of *x* and *y* (discards gaps):

$$\begin{matrix} x_1, x_2, \ldots, x_n \\ y_1, y_2, \ldots, y_n \end{matrix} \Longrightarrow x_1 y_1 x_2 y_2 \ldots x_n y_n$$

$$\binom{2n}{n}$$ intercalations exist…

$$\approx \frac{2^{2n}}{\sqrt{\pi n}}$$

Yikes!

Needleman-Wunsch (global)

- *F(i,j)* = score of best alignment of

$$x_{1...i} \quad \text{and} \quad y_{1...j}$$

- Suppose *F(i-1,j-1),* F(i-1,j), F(i,j-1) are known

$$F(0,0) = 0$$

$$F(i,j) = \max\left\{\begin{matrix} F(i-1, j-1) + s(x_i, y_i), \\ F(i-1, j) - d, \\ F(i, j-1) - d \end{matrix}\right\}$$

Example: Needleman-Wunsch

```
y HEAGAWGME-E
x --P-AW-MEAE
```

Example: Needleman-Wunsch

Image removed for copyright reasons.

Smith-Waterman (Local Alignment)

$$F(0,0) = 0$$

$$F(i, j) = \max \left\{ \begin{array}{c} 0, \\ F(i-1, j-1) + s(x_i, y_i), \\ F(i-1, j) - d, \\ F(i, j-1) - d \end{array} \right\}$$

- Key idea is to look for best alignment between subsequences
- Expected score of random match must be negative

Example: Smith-Waterman

Image removed for copyright reasons.

What does a score mean?

- How can we tell if our match is significant?
- Isn’t this related to the size of the query and the database?

Being Bayesian

- Assume a casino uses a loaded die 1% of the time.
- A loaded die will come up six 50% of the time.
- You pick up a die at the casino and roll it three times, getting three sixes.

- *What is the chance the die is loaded?*

Being Bayesian: II

$$P(X\,|\,Y)P(Y) = P(Y\,|\,X)P(X)$$

$$P(X\,|\,Y) = \frac{P(Y\,|\,X)P(X)}{P(Y)}$$

$$\begin{aligned} P(D_{loaded}\,|\,3\ sixes) &= \frac{P(3\ sixes\,|\,D_{loaded})P(D_{loaded})}{P(3\ sixes)} \\ &= \frac{(0.5)^3(0.01)}{(0.5)^3(0.01)+(\tfrac{1}{6})^3(0.99)} \\ &= 0.21 \end{aligned}$$

Comparing Models (Bayesian)

$$P(M \mid x, y) = \frac{P(x, y \mid M)P(M)}{P(x, y)}$$

$$= \frac{P(x, y \mid M)P(M)}{P(x, y \mid M)P(M) + P(x, y \mid R)P(R)}$$

$$= \frac{P(x, y \mid M)P(M) / P(x, y \mid R)P(R)}{1 + P(x, y \mid M)P(M) / P(x, y \mid R)P(R)}$$

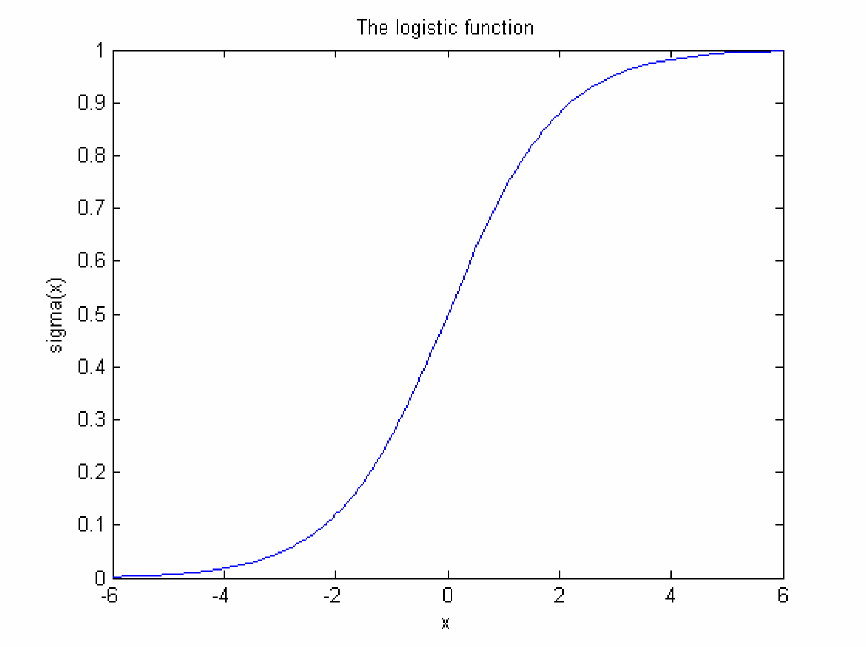
Comparing Models (Bayesian)

$$P(M \mid x,y) = \frac{P(x,y \mid M)P(M) \big/ P(x,y \mid R)P(R)}{1 + P(x,y \mid M)P(M) \big/ P(x,y \mid R)P(R)}$$

$$= \frac{e^{S'}}{1+e^{S'}}$$

where $S' = S + \log\left({}^{P(M)}\!\big/\!{}_{P(R)} \right)$

Comparing Models (Bayesian II)



$$P(M \mid x, y) = G(S')$$

$$G(x) = \frac{e^x}{1+e^x}$$

$$S' = S + \log\left({P(M)}\big/{P(R)} \right)$$

- Global alignment: compare *S* with *log N*
- Local alignment: compare *S* with *log MN*

Classical Approach: Extreme Value Distribution

- Expected number of unrelated matches for a local alignment (*E-value*)

$$E(S) = Kmn2^{-\lambda S}$$

- Used by BLAST

Building Phylogenetic Trees

- Unweighted pair group method using arithmetic averages (UPGMA)

- Clusters sequences based on evolutionary distance

Example: UPGMA

Image removed for copyright reasons.

Parsimony-based Phylogenetic Trees

- Build all possible trees
- Choose tree that uses fewest number of substitutions

Example: Parsimony

Image removed for copyright reasons.

Fin