CSE 527 10/9/06
Scribe: David Langworthy (dlan@microsoft.com)

--Course Home Page
Update “change your subscriptions options” if you do not use an @U address.
Lecture notes and slide numbering will be out of sync.
New resource links.

--Schedule
See new readings.
HW1 — trying to sort out wiki permissions.

Lecture Slide Numbers 4&5

Alignment is widely used, but speed is a problem.
Most used tool is BLAST
Discuss scoring significance

--1
Some pieces of a protein matter more than others. Some parts are active and some parts are scaffolding.

--2
Nothing in biology makes sense except in light of evolution.

Most changes make a protein worse, some wont matter, and a few make it better.

Nothing about the process of change is focused on making a protein (or organism) better.

Changes are less tolerated in many interactions.

--3
Basic Local Alignment Search Tool (BLAST)

Could run Smith Waterman, but it’s slow and the database is growing faster than computers are getting faster, which is pretty fast.

Small good matches are more meaningful than long mediocre matches.

BLAST is a heuristic. It may miss some long weak matches.

--4

E-value is a measure of statistical confidence.

--5

There are 8000 possible amino acid sequences of length 3.
Build an un-gapped alignment based on seed matches.

--6

Substrings of length 2 for convenience.

:20

It’s heuristic. A lower first threshold might find a longer higher overall score.

--7
BLOSUM 62
Default score matrix for BLAST.
Entries on diagonal are positive. Most but not all off diagonal entries are negative.
Diagonal entries are not all equal. Vv is 4, ww is 11. V is more common than W.
The matrix is symmetric [sigma commutative].
PAM is the other accepted score matrix.

62 means the designers eliminated sequences that were greater than 62% similar.
BLOSUM 50 only used sequences with less than 50% similarity.
This is to prevent extrapolation from close relatives.

--8

BLAST Refinements
'97 -- Databases growing larger. Need more selectivity in filter. Require two "close" non-overlapping hits.
Allow for gaps. Run bidirectional Smith Waterman until the score drops below some dynamic threshold.
More flexible than just searching a fixed number of diagonals off the main i=j diagonal.

Blast does not do well with weaker matches.
Position Specific Iterated (PSI-) BLAST uses iterated search to boost results for distant matches by building a similarity matrix from initial hits and requerying using this as a weight matrix.

--9
Is 42 a good score?

-- Board
Hypothesis testing

Coin either fair P(H) = .5 or unfair P(H) = .66666
Gather data say HHHHH
Null Model (hypothesis) M0: coin is fair
Alternate: M1: coin is biased

Prob(ObservedData | M0) == (.5)^5 == 1/32
P(Data | M1) == (2/3)^5 == 32/243

These are small numbers and with more trials, the numbers will necessarily get smaller.
Typically use "likelihood" ratio P(D|M1) / P(D|M0) = (32/243) / (1/32) = 1024/243 ~ 4
Math works nicely for 5 heads, but works for any sequence.

Neyman-Pearson Theorem
You can come up with what ever means of testing you want, but you are not going to come up with anything better than likelihood ratio test.

:55

Often convenient to look at log likelihood. It makes the math easier. Threshold tests work out the same.

P-value: probability, given that M0 is true, that you see data as or more extreme than observed. This is the probability of one of the two possible errors -- rejecting the null when it's actually true.

Suppose 100 flips and 80 heads.

In a simple hypothesis like the coin flip, we can calculate the P-value. In other scenarios the P-value is not amenable to analytic solution. It could be some complex process with feedback. Can do simulations in this case.

--Back to slides

--10

BLAST tells if two proteins are similar. Want to know if they are homologous—similar because they evolved from a common ancestor. Want to know if they are so similar they could not have evolved independently.

--11

Where numbers in BLOSUM matrix come from.

-- Back to BLOSUM 62 table
1:15

--12

Any matrix you come up with reflects some expectation of PXY.

This is one of the reasons for the success of BLAST. The authors worked out the probabilistic significance of a match. Not just some ad hoc match score.