Genome 559
Introduction to Statistical and Computational Genomics
Winter 2009

Lecture 13a:

BLAST

Larry Ruzzo

1 minute responses

Pacing was: (a) A little slow (1), (b) great (3) [maybe we don't need semesters after all!], or (c) I was lost/equation-dense (4) (but,I'll try harder to keep up with reading)

Paper slides for note-taking *really* help. *Agreed*

More time for problems helped. *Hopefully again today.*

Is revised hw schedule on web? *Some.*

Liked it, but need some practice problems for it to sink in. *See hw5!*

Fuzzy on purpose of relative entropy; why does it matter. *If motif distribution is like background (low entropy), WMM prediction will be error-prone. Similarly, columns of low entropy may only add noise; at edges, especially, maybe delete them.*

Didn't explain substring matches/match objects (2) *Today*

BLAST:
Basic Local Alignment Search Tool
Altschul, Gish, Miller, Myers, Lipman, J Mol Biol 1990

The most widely used comp bio tool

Which is better: long mediocre match or a few nearby, short, strong matches with the same total score?

- score-wise, exactly equivalent
- biologically, later may be more interesting
- if must miss some, rather miss the former (?)

BLAST is a *heuristic* emphasizing the later

- speed/sensitivity tradeoff: BLAST may miss weak matches, but gains greatly in speed

Heuristic: A method proceeding towards a solution by trial and error, intuition or loosely defined rules. Cf. Algorithm; Smith-Waterman, etc.

A Protein Structure:
(Dihydrofolate Reductase)

BLAST: What

Input:

- a query sequence (say, 50-300 residues)
- a data base to search for other sequences similar to the query (say, 10^6 - 10^9 residues)
- a score matrix $\sigma(r,s)$, giving cost of substituting r for s (& perhaps gap costs)
- various score thresholds & tuning parameters

Output:

- "all" matches in data base above threshold
- "E-value" of each

BLAST: How

Idea: emphasize parts of data base near a good match to some short subword of the query

Break query into overlapping words w_i of small fixed length (e.g. 3 aa or 11 nt)

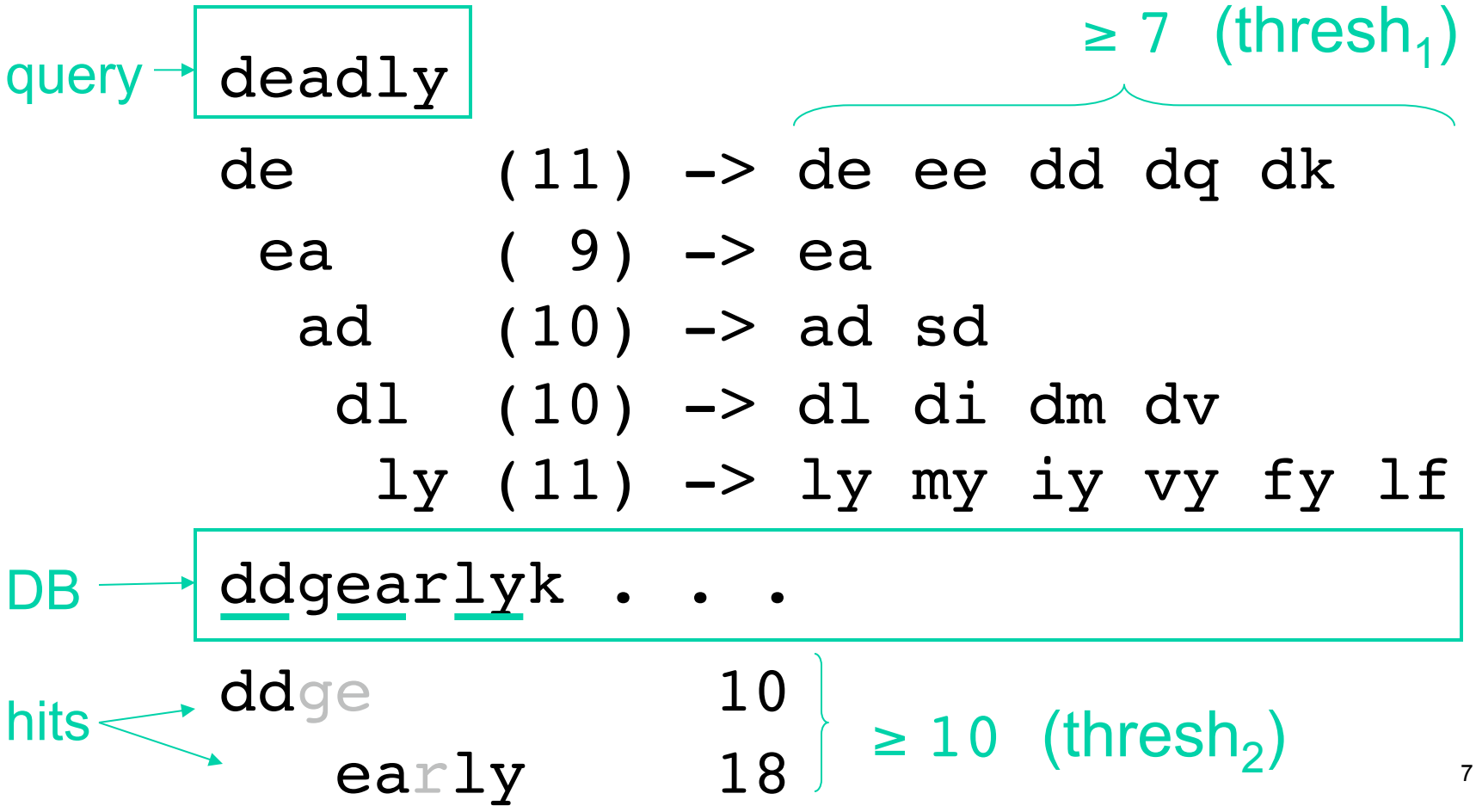
For each w_i, find (empirically, ~50) "neighboring" words v_{ij} with score $\sigma(w_i, v_{ij}) > \text{thresh}_1$

Look up each v_{ij} in database (via prebuilt index) -- i.e., exact match to short, high-scoring word

Extend each such "seed match" (bidirectional)

Report those scoring > thresh_2, calculate E-values

BLAST: Example



BLOSUM 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**4**	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	**5**	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	**6**	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	**6**	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	**9**	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	**5**	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	**5**	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	**6**	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	**8**	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	**4**	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	**4**	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	**5**	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	**5**	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	**6**	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	**7**	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	**4**	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	**5**	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	**11**	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	**7**	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	**4**

BLAST Refinements

"Two hit heuristic" – need 2 nearby, nonoverlapping, gapless hits before trying to extend either

"Gapped BLAST" – run heuristic version of Smith-Waterman, bi-directional from hit, until score drops by fixed amount below max

PSI-BLAST – For proteins, iterated search, using "weight matrix" pattern from initial pass to find weaker matches in subsequent passes (PSI=pos specific iter)

Many others

A Likelihood Ratio

Defn: two proteins are *homologous* if they are alike because of shared ancestry; similarity by descent

Suppose among proteins overall, residue x occurs with frequency p_x

Then in a random ungapped alignment of 2 random proteins, you would expect to find x aligned to y with prob $p_x p_y$

Suppose among *homologs*, x & y align with prob p_{xy}

Are seqs X & Y homologous? Which is more likely, that the alignment reflects chance or homology? Use a *likelihood ratio test.*

$$\sum_i \log \frac{p_{x_i y_i}}{p_{x_i} p_{y_i}}$$

E.g., BLOSUM62: trusted "homologues" = BLOCKS w/ ≥ 62% identity.

ad hoc Alignment Scores?

Make up any scoring matrix you like

Somewhat surprisingly, under pretty general assumptions**, it is *equivalent* to the scores constructed as above from some set of probabilities p_{xy}, so you might as well understand what they are

NCBI-BLASTN: +1/-2 ↔ 95% identity

WU-BLASTN: +5/-4 ↔ 66% identity

** e.g., average scores should be negative, but you probably want that anyway, otherwise local alignments turn into global ones, and some score must be > 0, else best match is empty

Summary

BLAST is a highly successful search/alignment heuristic. It looks for alignments anchored by short, strong, ungapped “seed” alignments

Strengths:

Speed, E-values, well-supported implementation & web server

Weaknesses:

Heuristic search can miss weaker matches