MIMM4750G

Score matrices and BLAST

Search a sequence database

- So far we have learned about querying the Genbank database using keywords.
- What if we only have a nucleotide or protein sequence to work with?
- An unknown species has no keywords to search by.
- One approach would be to *align* the sequence against every other sequence in the database, and take whatever aligns best.
- This would take too long! (More on alignment later.)

Dot plots

- A simple visualization tool for comparing two unaligned sequences.
- Make a table with one sequence along the top, and a second down the left.
- Fill in cells where both sequences contain the same residue.

	g	a	t	c	g	a	a	c	t	g	g
t			·						·		
g	·				●					·	·
a		·				●	·				
a		·				·	●				
c				·				●			
g	·				·					●	
g	·				·						●

INCA Q1

- Fill out this dot plot!

	C	A	G	A	A	G	A	A	T	C
G										
A										
G										
A										
A										
G										
C										

(Portions of 16S rRNA from Vibrio cholerae and S.typhimurium.)

BLAST

- Basic Local Alignment Search Tool
- Developed by Stephen Altschul, Walter Gish and David Lipman at the NCBI.
- Local similarity = search for conserved intervals.
- This requires some way to measure the similarity of unaligned sequences.

Word search

- The original BLAST algorithm attempts to find *high-scoring segment pairs* (HSP).
- The HSP is the set of equal-length segments from 2 sequences that maximizes the total similarity score.
- BLAST constructs an *index* of all "words" of length k.
- These words are often called "k-mers".
- What are the frequencies of 2-mers in `TACCTAGGGG`?

INCA Q2

- Sequence: TACCTAGGGG

- 2-mer counts:

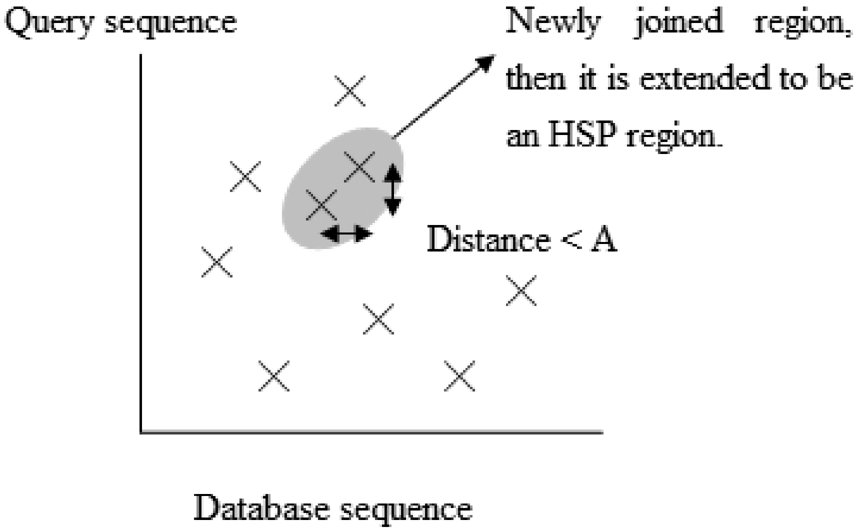
AC	AG	CC	CT	GG	TA

- How I handled this question in R:

```r
s <- paste(sample(c('A','C','G','T'), 10, replace=T), collapse='')
pieces <- sapply(1:(nchar(s)-1), function(i) substr(s, i, i+1))
table(pieces)
```

Building the HSP

- BLAST scans the database for high-scoring words (3-mers for proteins).
- From one pair of high-scoring words (*hit*), search left and right for a second hit within some maximum distance A.
- Require 2 hits to trigger a gap-free *extension* (incorporate flanking residues into candidate alignment).


Query sequence
Newly joined region, then it is extended to be an HSP region.
Distance < A
Database sequence

Finishing the HSP

- If the gap-free extension retains a high enough score, BLAST calculates a *gapped extension* (tolerate indels).
- Gapped extensions are very time consuming - two-hit method is designed to minimize the number carried out.
- Only high scoring gapped extensions are reported.
- Clearly, *scoring* plays an important role in BLAST searches.

What is a score?

- A measure of sequence homology (similarity that implies common ancestry).
- Sequences do not have to be exactly the same to be closely related.
- BUT this means that we have to know how some residues are more similar than others!
- *e.g.*, is glutamic acid (E) closer to cysteine (C) or aspartic acid (D)?
- A score is a rough estimate of how likely one type of substitution is over another.

Calculating scores

- Dayhoff pioneered the concept of quantifying amino acid substitution rates from the comparative analysis of protein sequences.

- Dayhoff *et al.* (1978) mapped 1,572 AA substitutions to trees relating protein sequences in the *Atlas* with <15% divergence.

	A	R	N	D	C	Q
A Ala						
R Arg	30					
N Asn	109	17				
D Asp	154	0	532			
C Cys	33	10	0	0		
Q Gln	93	120	50	76	0	
E Glu	266	0	94	831	0	422

PAM matrices

- accepted point mutations (abbreviated as PAM)
- calculate *mutation probability matrix* (M) from observed mutation counts (A):

  $$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}}$$

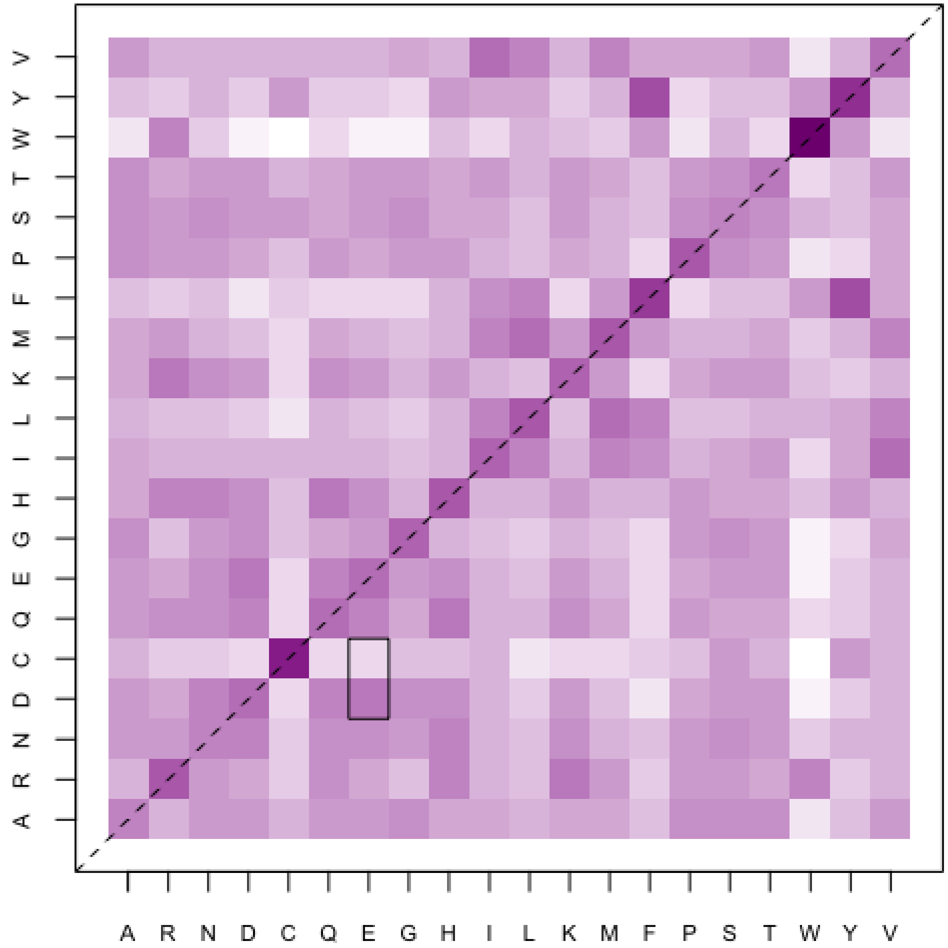
  where λ is a scaling constant, and

  $$m_j = \frac{\sum_{i \neq j} A_{ij}}{n_j}$$

  (the total number of mutations away from amino acid j, divided by total number of occurrences of this AA in the sequences).

PAM250 matrix

250 mutations per 100 amino acids (scaled from PAM1).



Is glutamic acid (E) closer to cysteine (C) or aspartic acid (D)?

BLOSUM62

- BLOcks SUbstitution Matrix
- Calculated from the (no longer maintained) BLOCKS database of local alignments of highly conserved regions of proteins.
- PAM is based on mutations mapped to a phylogeny.
- BLOSUM is based on odds ratios of AAs in an alignment.

Log-odds

- Consider an alignment of protein sequences.
- Frequency of amino acid a is p_a.
- If an aligned pair of AAs a and b are independent, then their probability is $p_a \times p_b$.
- Ratio of the *observed* probability ($q_{a,b}$) to this expectation is the *odds*.
- Taking the log of the odds gives us the log-odds:
  $$s(a,b) = \lambda \log \frac{q_{a,b}}{p_a p_b}$$
  where λ is used to round s to nice integers.

INCA Q3

- The observed frequency of aligned pairs of tryptophan (W) is $q_{W,W} = 0.0065$.
- The observed frequency of W alone is $p_W = 0.013$.
- What is s_{WW} if we set $\lambda = 2.88$? Round to one decimal place (*i.e.*, xy.z).
- Now do the same for leucine ($q_{L,L} = 0.0371, p_L = 0.099$)

Example stolen from SR Eddy (2004), Nature Biotechnol 22(8):1035.

BLOSUM62

- Like PAM, there are several BLOSUM matrices for different levels of evolutionary divergence.
- Unlike PAM, each BLOSUM matrix is derived from its own alignment, rather than being extrapolated from one data-derived matrix.
- BLOSUM62 derived from an alignment of protein segments of <62% identity
- Considered to be comparable to PAM250.
- BLAST generally uses BLOSUM62

Back to BLAST - evaluating significance

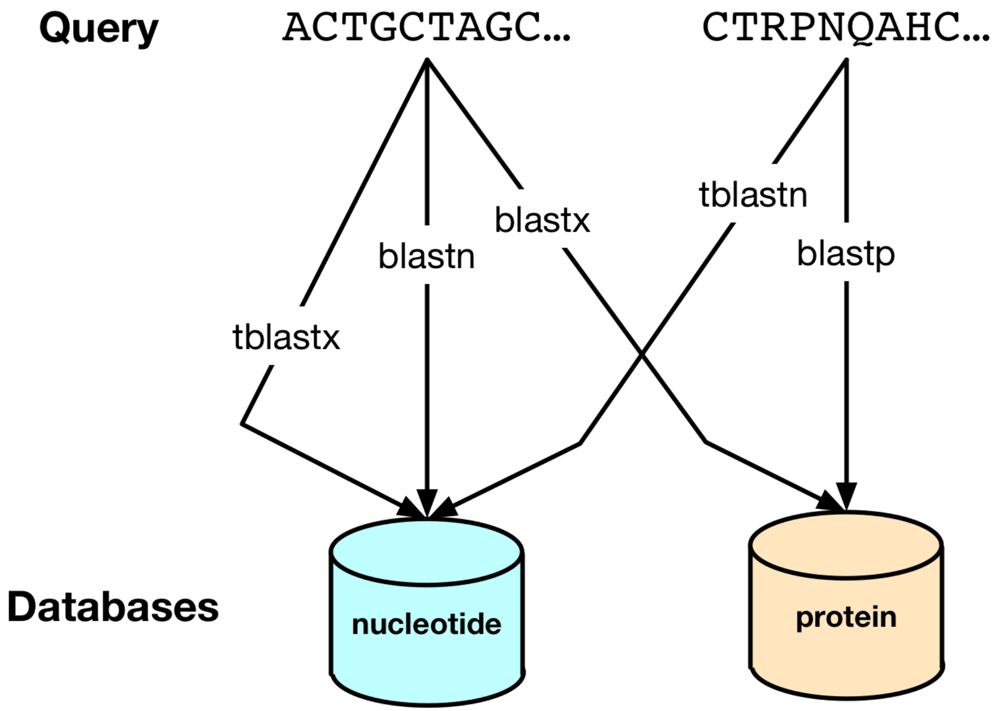
- Recall BLAST searches for high-scoring sequence pair (HSP).
- The expected number of HSPs with score $\geq S$:

  $$E = Kmne^{-\lambda S}$$

  where m and n are the sequence lengths.
- K and λ are pre-defined parameters that depend on the BLOSUM matrix.

Types of BLAST queries

- NCBI maintains both nucleotide and protein databases



BLAST databases

- **nr/nt** (non-redundant nucleotide collection) - identical sequences merged into same record
- 16S rRNA
- **est** (expressed sequence tags) - partial cDNA sequences
- **SRA** (sequence read archive) next-generation sequence data
- **VecScreen** - identify segments of vector origin
- **IgBLAST** - search immunoglobulin and T-cell receptor sequences

Further readings

- The Statistics of Sequence Similarity Scores
- Selecting the Right Similarity-Scoring Matrix
- Database resources of the National Center for Biotechnology Information