Introduction to Bioinformatics
3. Sequence Alignment #1

Department of
Computer Science and
Electronics
Gadjah Mada University

Recap - Why Align Sequences?

- DNA sequences (4 letters in alphabet)
  - GTAAACTGGTACT…
- Amino acid (protein) sequences (20 letters)
  - SSHLDKLMNEFF…
- Align them so we can search databases
  - To help predict structure/function of new genes
    - In particular, look for homologues (evolutionary relatives)
- 3D-pssm (Imperial College - Structure Prediction)
  - http://www.sbg.bio.ic.ac.uk/servers/3dpssm
  - Give it a gene sequence
    - It predicts the protein structure

Recap - Example matches

```
1. gattcagacctagct   (no indels)
     gtcagatcct
2. gattcaga-cctagct (with indels)
   g-t-cagatcct
3. gattcagacctagc-t
   g-t-----cagatcct
```

- Need to come up with algorithms producing:
  - Ways of scoring alignments
  - Ways to search for high scoring alignments
- Concentrate today on alignments <u>without</u> indels

Word of Warning

- These algorithms are still very much in flux
  - Both the techniques and the ways of assessing them (the statistics) change all the time
- Various parameters in algorithm have defaults
  - But you can change these defaults
  - So you need to know exactly how the algorithm works
- Always read the manual

Hamming Distances

- Suppose we have
  - Query sequence Q and database sequence D
- Hamming distance:
  - Number of places where Q and D are <u>different</u> (distance)
- Example (stars mark differences)
  - `SSHLDKLMNEFF`
  - `*  **  *`
  - `HSHLKLLMKEFFHDMN`
  - Scores 4 for Hamming distance (sometimes worry about ends)
- Simple alignment algorithm: slide Q along D
  - Remember where the Hamming distance was minimised

Scoring Schemes (Amino Acids)

- Hamming distance doesn't take into account
  - Likelihood of one amino acid changing to another
  - Some amino acid substitutions are disastrous
    - So they don't survive evolution
  - Some substitutions barely change anything
    - Because the two amino acids are chemically quite similar
- Scoring schemes address this problem
  - Give scores to the chances of each substitution
- 2 possibilities:
  - Use empirical evidence
    - Of actual substitutions in known homologues (families)
  - Use theory from chemistry (hydrophobicity, etc.)

BLOSUM62 Scheme

- Blocks Amino Acid Substitution Matrices
- Empirical method
  - Based on roughly 2000 amino acid patterns (blocks)
  - Found in more than 500 families of related proteins
- Calculate the **Log-odds** scores for each pair (R_1, R_2)
  - Let O = observed frequency R_1 <=> R_2
  - Let E = expected frequency R_1 <=> R_2 [happening by chance]
  - I.e., Score = round($2 * \log_2(O/E)$)
- To calculate the score for an alignment of two sequences
  - <u>Add up</u> the pairwise scores for residues
    - We've calculated <u>log</u> odds

BLOSUM62 Substitution Matrix

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

Zero: by chance

- + more than chance
- - less than chance

Arranged by

- Sidegroups
- So, high scoring in the end boxes

Example

- M,I,L,V
- Interchangeable

Example Calculation

Query =	S	S	H	L	D	K	L	M	R
Dbase =	H	S	H	L	K	L	L	M	G
Score =	-1	4	8	4	-1	-2	4	5	0

- Total score = -1+4+8+4+-1+-2+4+5+-2
  = 21
- Write Blosum(Query,Dbase) = 21
  - Not standard to do this

BLAST Algorithm
Basic Local Alignment Search Tool

- Fast alignment technique(s)
  - Similar to FASTA algorithms (not used much now)
  - There are more accurate ones, but they’re slower
  - BLAST makes a big use of lookup tables
- Idea: statistically significant alignments (hits)
  - Will have regions of at least 3 letters same
    - Or at least high scoring with respect to BLOSUM matrix

```
CCNDHRKMTCSPNDNNRK
TTNDHRMTACSPDNNNKH
```

more likely than

```
CCNDHRKMTCSPNDNNRK
YTNHHMMTTYSLDNNNKK
```

- Based on small local alignments

BLAST Overview

- Given a query sequence Q
- Seven main stages
  - Remove (filter) low complexity regions from Q
  - Harvest k-tuples (triples) from Q
  - Expand each triple into ~50 high scoring words
  - Seed a set of possible alignments
  - Generate high scoring pairs (HSPs) from the seeds
  - Test significance of matches from HSPs
  - Report the alignments found from the HSPs

BLAST Algorithm Part 1
Removing Low-complexity Segments

- Imagine matching
  - HHHHHHHHKMAY and HHHHHHHHURHD
  - The KMAY and URHD are the interesting parts
  - But this pair score highly using BLOSUM
- It’s a good idea to remove the HHHHHHHs
  - From the query sequence (low complexity)
- SEG program does this kind of thing
  - Comes with most BLAST implementations
  - Often doesn’t do much, and it can be turned off

Removing Low-complexity Segments

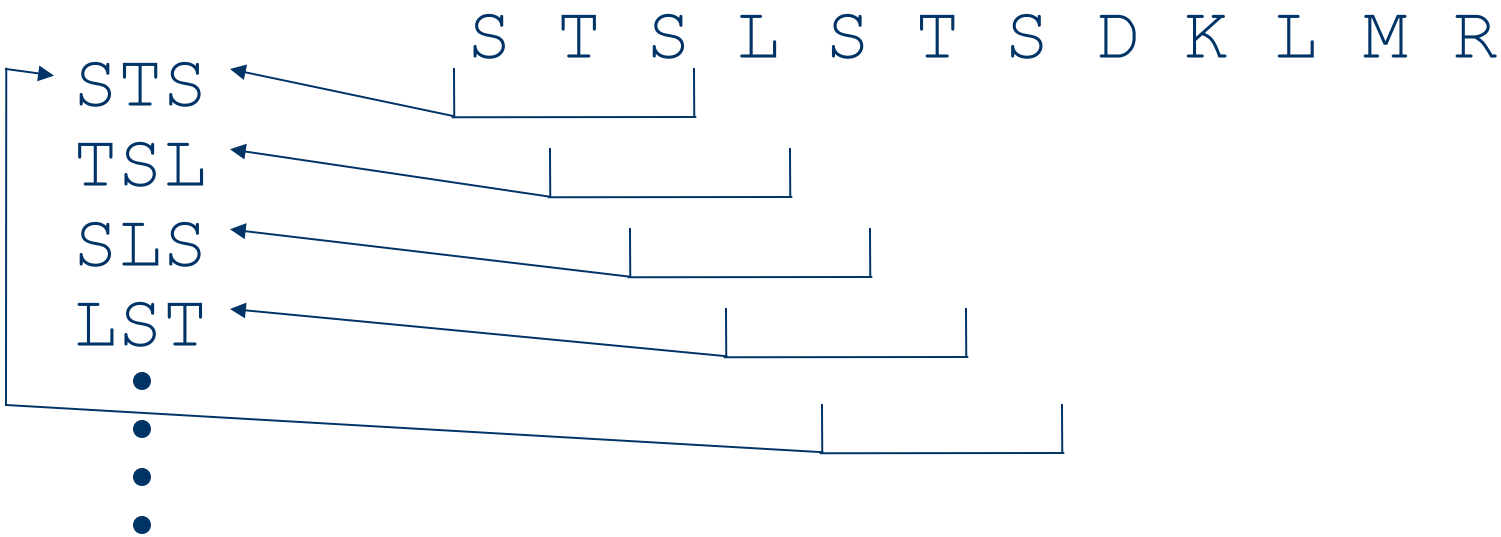
- Given a segment of length L
  - With each amino acid occurring n_1 n_2 … n_{20} times
- Use the following measure for "compositional complexity":

$$K = \frac{1}{L} log_{20} \left(\frac{L!}{\prod_{i=1}^{20} n_i!} \right)$$

- To use this measure
  - Slide a "window" of ~12 residues along Query Sequence Q
  - Use a threshold to determine low complexity windows
  - Use a minimise routine to replace the segment
    - With an optimal minimised segment (or just an X)
- Will do an example calculation in tutorial

BLAST Algorithm Part 2
Harvesting k-tuples

- Collect all the k-tuples of elements in Q
  - k set to 3 for residues and 11 for DNA (can vary)
  - Triples are called ‘words’. Call this set W



BLAST Algorithm Part 3 Finding High Scoring Triples

- Given a word w from W
  - Find all other words w' of same length (3), which:
    - Appear in some database sequence
    - Blosum(w,w') > a threshold T
- Choose T to limit number to around 50
  - Call these the high scoring triples (words) for w
- Example: letting w=PQG, set T to be 13
  - Suppose that PQG, PEG, PSG, PQA are found in database
  - Blosum(PQG,PQG) = 18, Blosum(PQG,PEG) = 15
  - Blosum(PQG,PSG) = 13, Blosum(PQG,PQA) = 12
  - Hence, PQG and PEG *only* are kept

Finding High Scoring Triples

- For each w in W, find all the high scoring words
  - Organise these sets of words
    - Remembering all the places where w was found in Q
- Each high scoring triple is going to be a seed
  - In order to generate possible alignment(s)
    - One seed can generate more than one alignment
- End of the first half of the algorithm
  - Going to find alignments now

BLAST Algorithm Part 4 Seeding Possible Alignments

- Look at first triple V in query sequence Q
  - Actually from Q (not from W - which has omissions)
  - Retrieve the set of ~50 high scoring words
    - Call this set H_V
  - Retrieve the list of places in Q where V occurs
    - Call this set P_V
- For every pair (*word, pos*)
  - Where *word* is from H_V and *pos* is from P_V
    - Find all the database sequences D
      - Which have an <u>exact</u> match with *word* at position *pos*'
    - Store an alignment between Q and D
      - With V matched at *pos* in Q and *pos*' in D
- Repeat this for the second triple in Q, and so on

Seeding Possible Alignments Example

- Suppose Q = QQGPHUIQEGQQG
- Suppose V = QQG, H_V = {QQG, QEG}
  - Then P_V = {1, 11}
- Suppose we are looking in the database at:
  - D = PKLMMQQGKQEG
- Then the alignments seeded are:

```
     QQGPHUIQEGQQG   word=QQG
PKLMMQQGKQEG         pos=1
```

```
QQGPHUIQEGQQG        word=QQG
     PKLMMQQGKQEG    pos=11
```

```
        QQGPHUIQEGQQG   word=QEG
```

```
QQGPHUIQEGQQG    word=QEG
```

BLAST Algorithm Part 5 Generating High Scoring Pairs (HSPs)

- For each alignment A
  - Where sequences Q and D are matched
  - Original region matching was M
- Extend M to the left
  - Until the Blosum score begins to decrease
- Extend M to the right
  - Until the Blosum score begins to decrease
- Larger stretch of sequence now matches
  - May have higher score than the original triple
  - Call these high scoring pairs
- Throw away any alignments for which the score S of the extended region M is lower than some cutoff score

Extending Alignment Regions Example

```
QQGPHUIQEGQQGKEEDPP        Blosum(QQG,QQG) = 16
     PKLMMQQGKQEGM
QQGPHUIQEGQQGKEEDPP        Blosum(QQGK,QQGK) = 21
     PKLMMQQGKQEGM
QQGPHUIQEGQQGKEEDPP        Blosum(QQGKE,QQGKQ) = 23
     PKLMMQQGKQEGM
QQGPHUIQEGQQGKEEDPP        Blosum(QQGKEE,QQGKQE) = 28
     PKLMMQQGKQEGM
QQGPHUIQEGQQGKEEDPP        Blosum(QQGKEED,QQGKQEG) = 27
     PKLMMQQGKQEGM
```

So, the extension to the right stops here

HSP (before left extension) is QQGKEE, scoring 28

BLAST Algorithm Part 6 Checking Statistical Significance

- Reason we extended alignment regions
  - Give a more accurate picture of the probability of that BLOSUM score occurring by chance
- Question: is a HSP significant?
- Suppose we have a HSP such that
  - It scores S for a region of length L in sequences Q & D
- Then the probability of two random sequences Q' and D' scoring S in a region of length L is calculated
  - Where Q' is same length as Q and D' is same length as D
- This probability needs to be low for significance
- We cover the statistics (briefly) later

BLAST Algorithm Part 7 Reporting the Alignments

- For each statistically significant HSP
  - The alignment is reported
- If a sequence D has two HSPs with Query Q
  - Two different alignments are reported
- Later versions of BLAST
  - Try and unify the two alignments

NCBI BLAST Server (protein-protein)

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help | Links CSG Webpages DoC Homepage Windows TinOfBeans

Address =Semiauto&PAGE=Proteins&NCBI_GI=yes&HITLIST_SIZE=100&COMPOSITION_BASED_STATISTICS=yes&SHOW_OVERVIEW=yes&AUTO_FORMAT=yes&CDD_SEARCH=yes&FILTER=L&SHOW_LINKOUT=yes Go

NCBI protein–protein BLAST

Nucleotide Protein Translations Retrieve results for an RID

Search

Set subsequence From: To:

Choose database nr

Do CD-Search ☑

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: (none)

Composition-based statistics ☑

Choose filter ☑ Low complexity ☐ Mask for lookup table only ☐ Mask lower case

Expect 10

Word Size 3

http://www.ncbi.nlm.nih.gov/blast Local intranet

- http://www.ncbi.nlm.nih.gov/BLAST/

Real Example

- MRPQAPGSLVDPNEDELRMAPWYWGRISREEAKSILHGKPDGSFLVRDAL
  SMKGEYTLTLMKDGCEKLIKICHMDRKYGFIETDLFNSVVEMINYYKENS
  LSMYNKTLDITLSNPIVRAREDEESQPHGDLCLLSNEFIRTCQLLQNLEQ
  NLENKRNSFNAIREELQEKKLHQSVFGNTEKIFRNQIKLNESFMKAPADA
  PSTEAGGAGDGANAAASAAANANARRSLQEHKQTLLNLLDALQAKGQVLN
  HYMENKKKEELLLERQINALKPELQILQLRKDKYIERLKGFNLKDDDLKM
  ILQMGFDKWQQLYETVSNQPHSNEALWLLKDAKRRNAEEMLKGAPSGTFL
  IRARDAGHYALSIACKNIVQHCLIYETSTGFGFAAPYNIYATLKSLVEHY
  ANNSLEEHNDTLTTTLRWPVLYWKNNPLQVQMIQLQEEMDLEYEQAATLR
  PPPMMGSSAPIPTSRSREHDVVDGTGSLEAEAAPASISPSNFSTSQ

- A gene taken from a fruit fly (Drosophila Melanogaster)
  - We'll alter this a little
  - And see if the NCBI BLAST server can find it for us