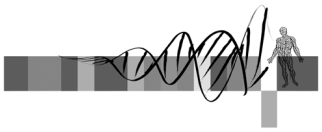


Bioinformatics for Biologists

Sequence Analysis: Part I. Pairwise alignment and database searching

Fran Lewitter, Ph.D.
Director
Bionformatics & Research Computing
Whitehead Institute

Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
- Significance of alignments
- Database searching methods

Topics to Cover

- Introduction
  - Why do alignments?
  - A bit of history
  - Definitions
- Scoring alignments
- Alignment methods
- Significance of alignments
- Database searching methods

Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor.

Doolittle RF, Hunkapiller MW, Hood LE, Devare SG, Robbins KC, Aaronson SA, Antoniades HN. *Science* 221:275-277, 1983.

Evolutionary Basis of Sequence Alignment

- ***Similarity*** - observable quantity, such as per cent identity
- ***Homology*** - conclusion drawn from data that two genes share a common evolutionary history; no metric is associated with this

Some Definitions

- An ***alignment*** is a mutual arrangement of two sequences, which exhibits where the two sequences are similar, and where they differ.
- An ***optimal alignment*** is one that exhibits the most correspondences and the least differences. It is the alignment with the highest score. May or may not be biologically meaningful.

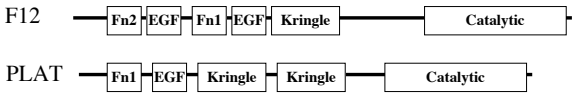
Alignment Methods

- ***Global alignment*** - Needleman-Wunsch (1970) maximizes the number of matches between the sequences along the entire length of the sequences.
- ***Local alignment*** - Smith-Waterman (1981) is a modification of the dynamic programming algorithm gives the highest scoring local match between two sequences.

Alignment Methods

Global vs Local

Modular proteins



Possible Alignments

```
A:   T C A G A C G A G T G
B:   T C G G A G C T G

I.   T C A G A C G A G T G
     T C G G A - - G C T G
II.  T C A G A C G A G T G
     T C G G A - G C - T G
III. T C A G A C G A G T G
     T C G G A - G - C T G
```

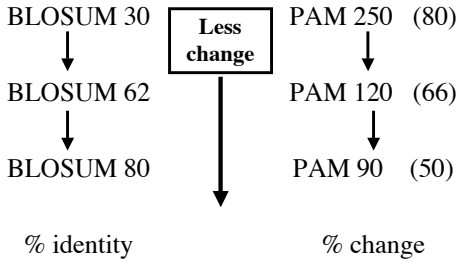

Topics to Cover

- Introduction
- Scoring alignments
  - Nucleotide vs Proteins
  - Definitions
- Alignment methods
- Significance of alignments
- Database searching methods

Amino Acid Substitution Matrices

- ***PAM -*** point accepted mutation based on *global* alignment [evolutionary model]
- ***BLOSUM*** - block substitutions based on *local* alignments [similarity among conserved sequences]

Substitution Matrices



Part of BLOSUM 62 Matrix

	C	S	T	P	A	G	N
C	9						
S	-1	4					
T	-1	1	5				
P	-3	-1	-1	7			
A	0	1	0	-1	4		
G	-3	0	-2	-2	0	6	
N	-3	1	0	-2	-2	0	

Log-odds = obs freq of aa substitutions / freq expected by chance

Part of PAM 250 Matrix

	C	S	T	P	A	G	N
C	1	2					
S	0	2					
T	-2	1	3				
P	-3	1	0	6			
A	-2	1	1	1	2		
G	-3	1	0	-1	1	5	
N	-4	1	0	-1	0	0	

Log-odds = pair in homologous proteins / pair in unrelated proteins by chance

Gap Penalties

- ***Insertion and Deletions*** (indels)
- ***Affine gap costs*** - a scoring system for gaps within alignments that charges a penalty for the existence of a gap and an additional per-residue penalty proportional to the gap's length

Example of simple scoring system for nucleic acids

- Match = +1 (ex. A-A, T-T, C-C, G-G)
- Mismatch = -1 (ex. A-T, A-C, etc)
- Gap opening = - 5
- Gap extension = -2

```
  T  C  A  G  A  C  G  A  G  T  G
  T  C  G  G  A  -  -  G  C  T  G
---------------------------------
+1 +1 -1 +1 +1 -5 -2 -1 -1 +1 +1 = -4
```


Scoring for BLAST 2 Sequences

```
Score = 94.0 bits (230), Expect = 6e-19
Identities = 45/101 (44%), Positives = 54/101 (52%), Gaps = 7/101 (6%)

Query: 204 YTGPFCDV----DTKASCYDGRGLSYRGLARTTLSGAPCQPWASEATYRNVTAEQ---AR 256
           Y+  FC      +  + CY G G +YRG    T SGA C PW S      V   Q   A+
Sbjct: 198 YSSEFCSTPACSEGNSDCYFGNGSAYRGTHSLTESGASCLPWNSMILIGKVYTAQNPSAQ 257

Query: 257 NWGLGGHAFCRNPDNDIRPWCFVLNRDRLSWEYCDLAQCQT 297
            GLG H +CRNPD D +PWC VL   RL+WEYCD+  C T
Sbjct: 258 ALGLGKHNYCRNPDGDAKPWCHVLKNRRLTWEYCDVPSCST 298
```

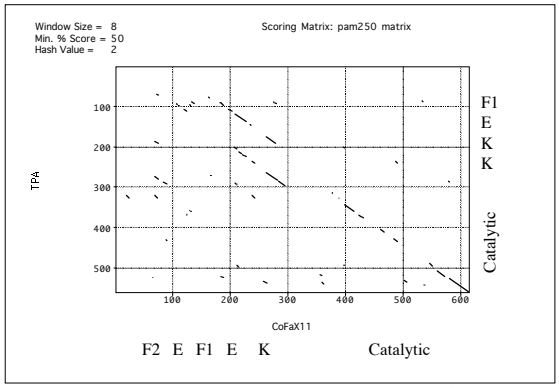
Based on BLOSUM62

```
Position  1: Y - Y =   7
Position  2: T - S =   1
Position  3: G - S =   0
Position  4: P - E =  -1
         . . .
Position  9: - - P = -11
Position 10: - - A =  -1
         . . .
               Sum     230
```

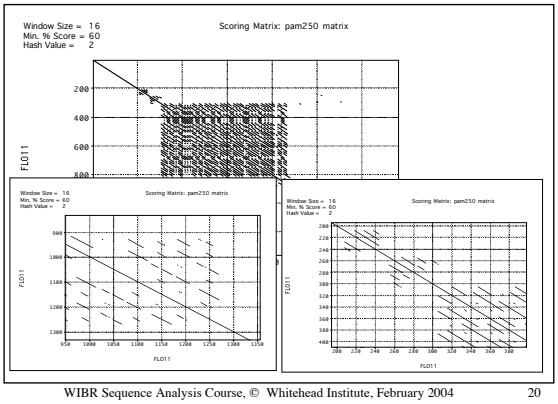

Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
  - Dot matrix analysis
  - Exhaustive methods; Dynamic programming algorithm (Smith-Waterman (Local), Needleman-Wunsch (Global))
  - Heuristic methods; Approximate methods; word or k-tuple (FASTA, BLAST, BLAT)
- Significance of alignments
- Database searching methods

Dot Matrix Comparison



Dot Matrix Comparison



Dynamic Programming

- Provides very best or optimal alignment
- Compares every pair of characters (e.g. bases or amino acids) in the two sequences
- Puts in gaps and mismatches
- Maximum number of matches between identical or related characters
- Generates a score and statistical assessment
- Nice example of global alignment using N-W:
  http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

Global vs Local Alignment

(example from Mount 2001)

	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6	-6	-10	-14	-18	-22	-26	-30
G	-16	-6	6	-5	-10	-13	-17	-22	-26
S	-20	-10	-5	7	-5	-8	-13	-17	-21
D	-24	-14	-8	-5	3	-5	-4	-14	-17
R	-28	-18	-14	-10	-8	3	-6	2	-10
T	-32	-22	-18	-13	-12	-7	3	-7	5
T	-36	-26	-22	-17	-15	-11	-7	2	-4
E	-40	-30	-25	-22	-20	-15	-8	-8	0
T	-44	-34	-30	-24	-24	-21	-15	-9	-5

	GAP	M	N	A	L	S	D	R	T
GAP	0	0	0	0	0	0	0	0	0
M	0	6	0	0	4	0	0	0	0
G	0	0	6	1	0	5	1	0	0
S	0	0	1	7	0	2	5	1	1
D	0	0	0	2	1	3	0	6	4	1
R	0	0	0	0	0	0	3	0	12	3
T	0	0	0	0	1	0	1	3	0	15
T	0	0	0	0	1	0	1	1	2	3
E	0	0	0	1	0	0	0	4	0	2
T	0	0	0	0	2	0	1	0	3	3

```
sequence 1   M - N A L S D R T
sequence 2   M G S D R T T E T
score        6 -12 1 0 -3 1 0 -1 3 = -5
```

```
sequence 1   S D R T
sequence 2   S D R T
score        2 4 6 3 = 15
```


Original "Ungapped" BLAST Algorithm (1990)

- To improve speed, use a word based hashing scheme to index database
- Limit search for similarities to only the region near matching words
- Use **T**hreshold parameter to rate neighbor words
- Extend match left and right to search for high scoring alignments

Original BLAST Algorithm (1990)

Query word (W=3)

Query: GSVEDTTGSQSLAALLNKCKT*PQG*QRLVNQWIKQPLM

Neighborhood words

PQG	18	PHG	13
PEG	15	PMG	13
PNG	13	PTG	12
PDG	13	Etc.	

Neighborhood Score threshold (T=13)

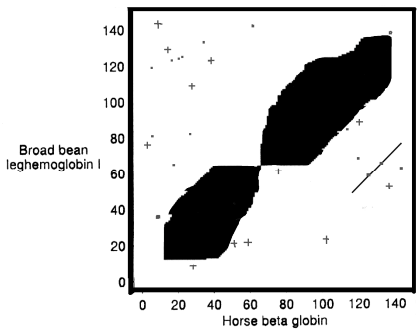
```
Query: 325   SLAALLNKCKT*PQG*QRLVNQWIKQPLMDKNRIEERLNLVEA
             +LA++L+   TP  G R++ +W+  P+ D   + ER   I A
Sbjct: 290   TLASVLDCTVT*PMG*SRMLKRWLHMPVRDTRVLLERQQTIGA
```

BLAST Refinements (1997)

- "two-hit" method for extending word pairs
- Gapped alignments
- Additional algorithms
  - Iterate with position-specific matrix (PSI-BLAST)
  - Pattern-hit initiated BLAST (PHI-BLAST)

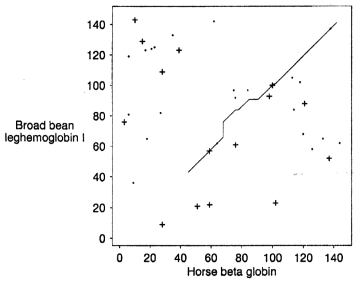
Gapped BLAST

15(+) > 13
22(•) > 11



(Altschul et al 1997)

Gapped BLAST



(Altschul et al 1997)

Programs to Compare two sequences - Unix or Web

NCBI

BLAST 2 Sequences

EMBOSS

water - Smith-Waterman
needle - Needleman -Wunsch
dotmatch (dot plot)
einverted or palindrome (inverted repeats)
equicktandem or etandem (tandem repeats)

Other

lalign (multiple matching subsegments in two sequences)

Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
- Significance of alignments
  - How strong can alignment be by chance alone?
- Database searching methods

Statistical Significance

- ***Raw Scores*** - score of an alignment equal to the sum of substitution and gap scores.
- ***Bit scores*** - scaled version of an alignment's raw score that accounts for the statistical properties of the scoring system used.
- ***E-value*** - expected number of distinct alignments that would achieve a given score by chance. Lower E-value => more significant.

Some formulas

$$\mathbf{E} = Kmn\, e^{-\lambda S}$$

This is the Expected number of high-scoring segment pairs (HSPs) with score at least S for sequences of length m and n.

This is the E value for the score S.

Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
- Significance of alignments
- Database searching methods
  - BLAST
  - BLAST vs. FASTA
  - BLAT

Questions

- Why do a database search?
- What database should be searched?
- What alignment algorithm to use?
- What do the results mean?
- What parameters can be changed?
  - Substitution matrices
  - Statistical significance
  - Filtering for low complexity

BLASTP Results

Sequences producing significant alignments:		Score (bits)	E Value	
gi|34862150|ref|XP_345634.1|	similar to mismatch repair pro...	209	5e-54	L
gi|36949366|ref|NP_002431.2|	mutS homolog 4; mutS (E. coli)...	162	9e-40	L
gi|34481396|emb|CAC79990.1|	sperm protein [Homo sapiens]	152	1e-36	L
gi|34861090|ref|XP_227831.2|	similar to MutS homolog 4 [Rat...	147	3e-35	L
gi|34872785|ref|XP_213395.2|	similar to hypothetical protei...	33	0.62	L
gi|34853116|ref|XP_345138.1|	similar to hypothetical protei...	32	1.3	L
gi|34783109|gb|AAH01726.2|	Unknown (protein for IMAGE:35345...	32	1.6	
gi|16307283|gb|AAH09731.1|AAH09731	Similar to hypothetical ...	31	3.1	L
gi|34868124|ref|XP_221530.2|	similar to mKIAA0719 protein [...	31	3.4	L
gi|34853816|ref|XP_344817.1|	similar to FGFR1 oncogene part...	30	7.8	L

Alignments

Get selected sequences | Select all | Deselect all

>gi|34862150|ref|XP_345634.1| L similar to mismatch repair protein MSH6 [Rattus norvegicus]
Length = 1541

Score = 209 bits (533), Expect = 5e-54
Identities = 174/617 (28%), Positives = 283/617 (45%), Gaps = 78/617 (12%)

WU-BLAST vs NCBI BLAST

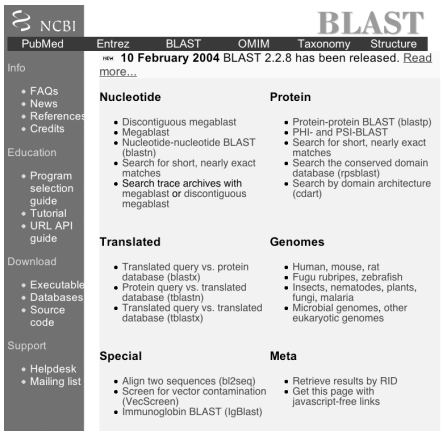
- WU-BLAST first for gapped alignments
- Use different scoring system for gaps
- Report different statistics
- WU-BLAST does not filter low-complexity by default
- WU-BLAST looks for and reports multiple regions of similarity
- Results will be different!

BLAT

- ***B***last-***L***ike ***A***lignment ***T***ool
- Developed by Jim Kent at UCSC
- For DNA it is designed to quickly find sequences of >= 95% similarity of length 40 bases or more.
- For proteins it finds sequences of >= 80% similarity of length 20 amino acids or more.
- DNA BLAT works by keeping an index of the entire genome in memory - non-overlapping 11-mers (< 1 GB of RAM)
- Protein BLAT uses 4-mers (~ 2 GB)

FASTA

- Index "words" and locate identities
- Rescore best 10 regions
- Find optimal subset of initial regions that can be joined to form single alignment
- Align highest scoring sequences using Smith-Waterman



Basic Searching Strategies

- Search early and often
- Use specialized databases
- Use multiple matrices
- Use filters
- Consider Biology