

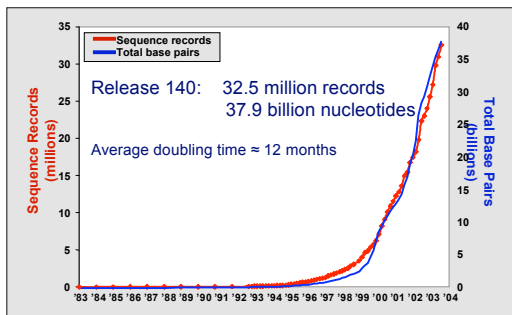
Sequence Alignment and Approaches to Database Searching

Jessica Kissinger
WHO-TDR Delhi 2005



Ian Korf, and M. Yandell O'Reilly Publishing

The Growth of GenBank



NCBI - BLAST

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

<ul style="list-style-type: none"> Getting started News FAQs 	Nucleotide <ul style="list-style-type: none"> Quickly search for highly similar sequences (megablast) Quickly search for divergent sequences (discontiguous megablast) Nucleotide-nucleotide BLAST (blastn) Search for short, nearly exact matches (tblastx) Search trace archives with megablast or discontiguous megablast 	Protein <ul style="list-style-type: none"> Protein-protein BLAST (blastp) Position-specific iterated and pattern-hit enhanced BLAST (PSI-BLAST) Search for short, nearly exact matches (tblastx) Search the conserved domain database (cdart) Protein homology by domain architecture (odart)
Software <ul style="list-style-type: none"> Downloads Developer info 	Translated <ul style="list-style-type: none"> Translated query vs. protein database (blastx) Protein query vs. translated database (blastp) Translated query vs. translated database (tblastx) 	Genomes <ul style="list-style-type: none"> Human, mouse, rat, chimp, dog, sheep, cat Chicken, puffer fish, zebrafish Environmental samples Metagen Insects, nematodes, plants, fungi, microbial genomes, other eukaryotic genomes
Other resources <ul style="list-style-type: none"> References NCBI Contributors Mailing list Contact Us 	Special <ul style="list-style-type: none"> Search for gene expression data (GEO BLAST) Align two sequences (tblast) Screen for vector contamination (VecScreen) ImmunoSpot BLAST (igblast) SNP BLAST 	Meta <ul style="list-style-type: none"> Retrieve results

<http://www.ncbi.nlm.nih.gov/BLAST/>

NCBI BLAST

Nucleotide Protein translation Retrieve results for an hit

Search:

Set resequence from: To:

Choose database:

Do CD Search:

Now:

Options for advanced blasting

Limit by matrix: or select from:

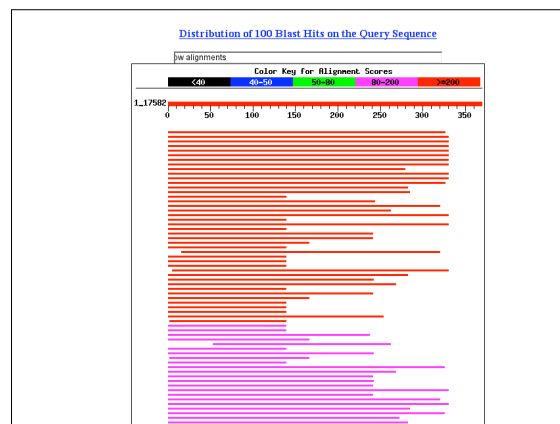
Composition-based statistics:

Choose filter: Low complexity Mask for lookup table only Mask lower case

Expect:

Word size:

Matrix: Gap Costs: Existence: 11 Extension: 1



Algorithms: definition

Webster's definition:

“a procedure for solving a mathematical problem in a finite number of steps that frequently involves a repetition of an operation; or *broadly*: a step-by-step procedure for solving a problem or accomplishing some end”

Alignments

- Alignment types:
 - global/local
 - gapped/ungapped
 - pairwise/multiple
- In what follows we will focus on *pairwise alignments*.

Pairwise Alignment

- There are two types of pairwise alignments
 - **Global (Needleman-Wunsch)**
 - Compare two sequences in their entirety
 - Insert gaps as necessary to make the sequences the same lengths
 - **Local (Smith-Waterman)**
 - Compare a portion of one sequence to a portion of another
 - Look for the “best” possible alignment of sub-regions

Global vs Local Alignment

- **Global**

```

L G P S S K Q T G K G S - S R I W D N
|   |   |   |   |
L N - I T K S A G K G A I M R L G D -
            
```
- **Local**

```

- - - - - G K G - - - - -
            | | |
- - - - - G K G - - - - -
            
```

Substitutions, Insertions, Deletions

- *Mutation*: one of
 - switch from one nucleotide to another
 - *insertion*
 - *deletion*
- *Substitution*: a switch in nucleotides which spreads throughout most of a species.
- Substitutions, insertions and deletions passed along two independent lines of descent cause a divergence of the two sequences from the original (and from each other):



Example

- For the previous example $cggatgcca \rightarrow cgggatccaa, ccctaggtecca$, the two descendent sequences align as follows

```

c g g g t a - - t - c c a a
c c c - t a g g t c c c - a
            
```

- “-” (*indel*) represents an insertion or deletion.

Alignments (*cont.*)

- Given two sequences, find an “optimal” alignment between them and use it to answer the questions stated above.
- What is an “optimal” alignment?
- Need a way to compare alignments.
 - Attach a score to each alignment.
 - This should reflect the likelihood that this alignment was produced as a consequence of divergence from a common ancestor.

Scoring schemes

- Given a scoring scheme,
 - an optimal alignment between two sequences is one with the *best* score (there might be more than one optimal alignment).
 - the *score of the sequence pair* is such a best score.
- Using the scores of sequence pairs one can:
 - investigate the hypothesis that two sequences diverged from a common ancestor
 - use the alignment of a pair of sequences that are judged to be related in order to discover common patterns.
 - by comparing scores among different species, get information to help reconstruct the phylogenetic tree that relates them all.

Types of scores

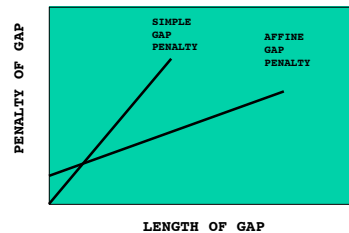
- *Similarity* Scores: the higher the score, the more closely related are the two aligned sequences.
- *Distance* scores (or distance measures): the lower the score, the more closely related the sequences.

In what follows we will use similarity score.

GAP PENALTIES

Linear = #gaps x penalty

Affine = Opening penalty + #gaps x extension penalty



Substitution Matrices

- A 4x4 (NA) or a 20x20 (AA) symmetric matrix.
- Example:
 1. $s(X,Y)=1$ if $X=Y$, -1 otherwise.
- In what follows we will assume that a scoring scheme, consisting of a substitution matrix and a gap penalty function, is given.

Example

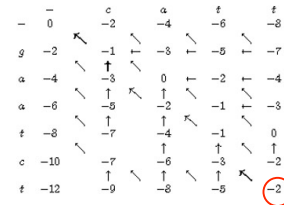
- Let $s(X,Y)=1$ if $X=Y$, -1 otherwise and use a linear gap score with $d=-2$. Then the score of the alignment

$$\begin{array}{cccccccc}
 & c & t & t & a & g & - & g & - & - \\
 & c & a & t & - & g & a & g & a & a \\
 \text{is} & 1 & -1 & +1 & -2 & +1 & -2 & +1 & -2 & -2 = -5
 \end{array}$$

Example

- $x=gaatct, y=catt$ ($m=6$ and $n=4$)
- $s(X,Y)=1$ if $X=Y, -1$ otherwise
- $d=2$

Example (cont.)



- 3 optimal alignments:

$gaatct$ $gaatct$ $gaatct$
 $c-at-t$ $ca-t-t$ $-cat-t$

Database Searching

- Database Searching \neq Sequence alignment
- Similarity \neq Homology
- Similarity is a measure of "sameness". It is expressed as a percentage, and it does not imply any reasons for the observed sameness, it is simply a measure of the observed likeness.
- Homology is an evolutionary term used to describe relationship via descent from a common ancestor. Homologous things are often similar, but not always, for example the flipper of a whale and your arm, or the DNA sequence for Actin in humans and chickens. Homology is NEVER expressed as a percent, either you are related or you aren't.

Similarity and Homology

- Sequence homology can be reliably inferred from **statistically significant** similarity over a majority of the sequence length.
- Non-homology CANNOT be inferred from non-similarity because non-similar things can still share a common ancestor.
- Homologous proteins share common structures, but not necessarily common sequence or function.

Origins of similarity NOT based on common ancestry

- Similarity is often observed in regions of low sequence complexity, I.e. SSSSSS or ATATATATATAT, such similarity is also almost always local and will not span the length of the sequences being compared.
- Similarity can also be caused by underlying biases in nucleotide or amino acid usage
- Similarity can be caused by shared motifs that have been acquired.

Similarity Assessment

- Our assumption is that unrelated sequences will behave like random sequences
- Biological sequences are not random, so the statistics of **extreme value distributions** apply.
- **Scores** for matches are influenced by the scoring matrix used
- **Sensitivity** and **Selectivity** are affected by choice of matrix and choice of database (redundancy and size).
- Choice of search molecule (query)

Sensitivity and Selectivity

- Sensitivity is a measure of your ability to find all the true matches
- Selectivity is a measure of your ability to not erroneously include false matches
- Database searching is a balancing act between sensitivity and selectivity. The factors that affect searches most are:
 - Scoring matrix
 - Gap model and Gap penalties
 - Filtering of low complexity regions (or not)
 - Size and redundancy of database

A quick talk about probability

- What fraction of a nucleotide database will contain a hit to the letter “A”?
- To “AT”
- To “ATCG”
- In a protein database, what fraction will hit “W” Tryptophan?
- Are biological sequences random?

Scoring Matrices

- Scoring Matrices are designed to detect signal above background, to detect similarities beyond what would be observed by chance alone
- The simplest scoring mechanism is match = 1, mismatch = -1, but these values don't work well for biological data.
- Because amino acids affect structure and reactivity, not all of the 400 aa pairs can be treated via a unitary match/mis-match matrix

	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	-
G	0	-3	-1	-2	-3		
H	-2	-3	-1	0			

BLOSUM 62

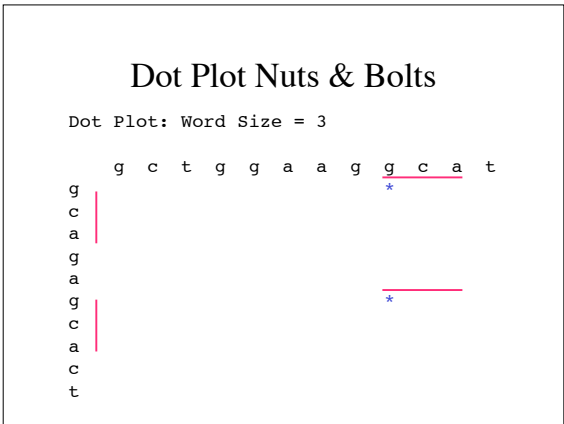
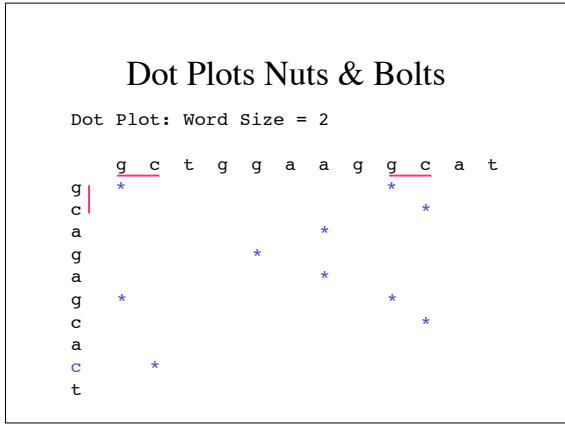
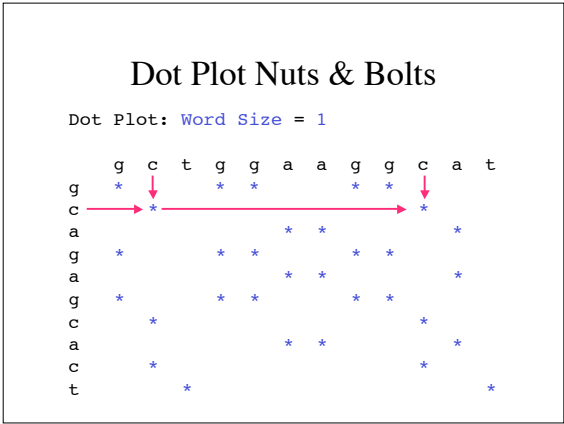
Matrix differences

- PAM matrices are based on an explicit evolutionary model, BLOSUM on an implicit model
- PAM matrices based on mutations observed throughout a global alignment, BLOSUM is based on conserved regions (blocks) which contain no gaps
- In BLOSUM, not all mutations are counted equally (similar sequences are clustered and together)
- PAM matrices were the first matrices, BLOSUM matrices came later. For most applications, BLOSUM 62 is the default scoring matrix.

Matrix rules of thumb

- Need different levels of sensitivity
 - Close relationships (Low PAM, high Blosum)
 - Distant relationships (High PAM, low Blosum)

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent	←	→ More divergent

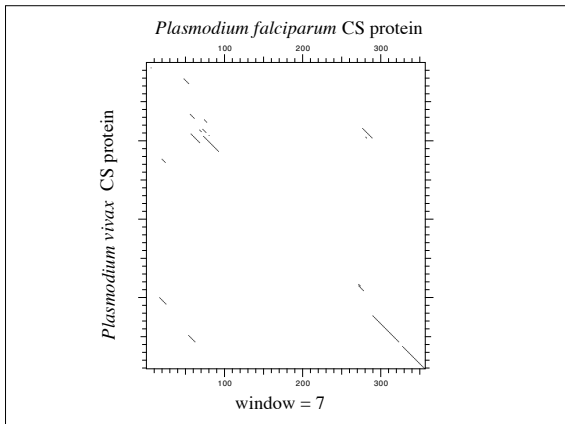
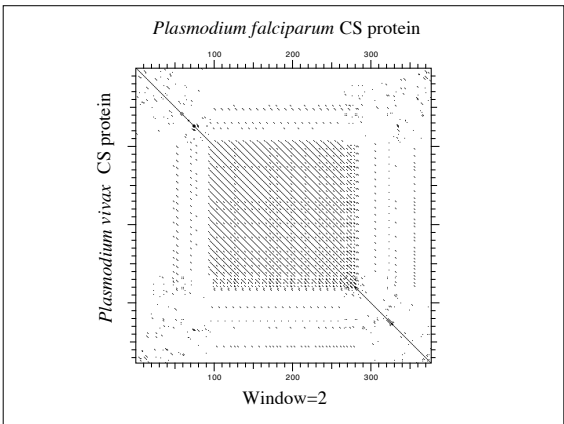


Plasmodium falciparum circumsporozoite protein

MMBRKLAISVSSFLVEALFQFYQCYSSSNTRVLNELNYDNAGTNLVNLEMLNYGQENWYSYLLKKNRSLGENDGGNN
 NNGDNGREGKDEKDRGDNEDNEKLRPKHKLKLQPGDGNPDNANPNVDPNANPNVDPNANPNVDPNANPNANPN
 ANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPN
 ANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPN
 ANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPNANPN
 NNNEEPSDKHEQLKIKNSISTEWSPCSVCNGIQVRIKFGSANKPKDELDEYDIEKKIKMEKCSSFNVNSSI
 GLIMVLSFLFLN

Plasmodium vivax circumsporozoite protein

MKNFLLAVSSILLVDLFTTHCGHVDLSKAINLGVNFNVDASSLGAHVQGSASRGRGLGENPDDEEIDAKKKKDGK
 KAEFKVPEKELKQAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGD
 RADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGD
 RAAGQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAGDRADQQPAG
 GVGVRVRIRVNAANKKPEDLILNDELDVCTMDKAGIFNVSSLSGLVILLVLAFLN



Database Searching

- **Applied Considerations**
 - The choice of search algorithm influences the sensitivity and selectivity of the search
 - The choice of matrix determines both the pattern and the extent of substitution in the sequences the database search is most likely to discover

Protein vs Nucleotide

- Which molecules should you search with?
- Which databases should you search, nucleotide or protein?

	T	C	A	G
T	TTT Phe (F) TTC * TTA Leu (L) TTG *	TCT Ser (S) TCC * TCA * TCG *	TAT Tyr (Y) TAC TAA Ter TAG Ter	TGT Cys (C) TGC TGA Ter TGG Trp (W)
C	CTT Leu (L) CTC * CTA * CTG *	CCT Pro (P) CCC * CCA * CCG *	CAT His (H) CAC * CAA Gln (Q) CAG *	CGT Arg (R) CGC * CGA * CGG *
A	ATT Ile (I) ATC * ATA * ATG Met (M)	ACT Thr (T) ACC * ACA * ACG *	AAT Asn (N) AAC * AAA Lys (K) AAG *	AGT Ser (S) AGC * AGA Arg (R) AGG *
G	GTT Val (V) GTC * GTA * GTG *	GCT Ala (A) GCC * GCA * GCG *	GAT Asp (D) GAC * GAA Gln (E) GAG *	GGT Gly (G) GGC * GGA * GGG *

The “Universal” genetic code. WARNING: There are others!

Remember your translation frames

```

1/1
G L S D A L E N V T H S V C T L
W P V R C V G E N C D P L R V H T
H A C Q H R W R M * P T P C A H
ATG GCC TGT CAG ATG CGT TGG AGR ATG TGA CCC ACT CCG TGT GCA CAC TT
TAC CGG ACR GTC TAC GCA ACC TCT TAC ACT GGG TGA GGC ACR CGT GTG AA
H G T L H T P S H S G S A T C V
P R D S A N S F T V W E T H V S
R Q * I R Q L I H G V G H R C K
    
```

Each strand has three reading frames. Frames 1-3 indicate the top or “+” strand and frames 4-6 indicate the bottom or “-” strand. Sometimes the notation is +1, +2, +3 and -1, -2, -3

Why can't we just look at the DNA sequence for the protein?

- It was one thought that we might be able to calculate a minimum mutation matrix, i.e. one in which the minimum number of steps needed to change from one aa to another we counted. The problem is, because of the degeneracy of the genetic code, often likely and unlikely mutations would receive the same score

Database search algorithms need to impose some sort of heuristic

- Because we cannot realistically search large databases for optimal alignments
- So, we use heuristic approaches to simplify the search
- These approaches are good, but they are not guaranteed to find “the optimal alignment”

The FASTA approach

William Pearson

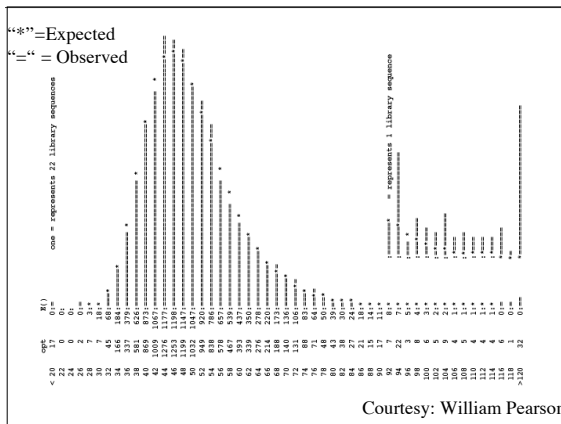
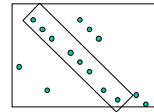
- Apply a dot-plot like approach to rapidly find regions of similarity.
- Instead of comparing each residue in two sequences, FASTA looks for patterns of k -tuples (words).
- When “ k ” number of consecutive “hits” are found (4-6 for DNA, 1-2 for protein k -tups) a scoring matrix is applied to identify the highest scoring segments
- Apply Smith Waterman algorithm to find the optimal alignment within the area searched

FASTA Nuts & Bolts

Create a list of words for the query sequences (W=2 for AA W=6 for nt):

```
g c t g g a a g g c a t
g c t g g a
c t g g a a
t g g a a g
```

Compare the words from the two sequences for identical words using dot plots. Only attempt SW Alignment on “hits” that are on the same diagonal within some distance of each other



The best scores are:

	s-w	z-score	E(12805)	%	len
PWH06 H--trans. ATP synth.-human mito.	1400	1767.8	10-92	100.0	226
PWB06 H--trans. ATP synth.-bovine mito.	1157	1460.9	10-75	77.9	226
PWMS6 H--trans. ATP synth.-mouse mito.	1118	1411.6	10-72	75.7	226
PWKL6 H--trans. ATP synth.-frog mito.	745	940.6	10-46	53.3	226
PWFF6Y H--trans. ATP synth.-fruit fly mito.	473	597.1	10-27	37.8	222
PWFF6 H--trans. ATP synth.-fruit fly mito.	471	594.6	10-26	37.5	224
PWRY3 H--trans. ATP synth.-yeast mito.	438	551.7	10-25	36.2	232
PWAG6H H--trans. ATP synth.-aspergillus mito.	365	459.6	10-19	30.4	230
PWQ66 H--trans. ATP synth.-Cochliobolus mito.	353	444.4	10-18	31.3	214
PWNT6 H--trans. ATP synth.-wheat mito.	309	385.4	10-15	28.9	235
PWV6M H--trans. ATP synth.-tobacco mito.	309	385.2	10-15	28.3	233
PWZ6M H--trans. ATP synth.-corn mito.	283	355.0	10-15	31.1	291
LMEC6 H--trans. ATP synth.-E. coli	178	223.0	10-6	23.3	236
LRRE6 H--trans. ATP synth.-rice chloro.	144	180.8	0.00037	24.2	231
PWMA6 H--trans. ATP synth.-pea chloro.	143	179.5	0.00044	25.0	232
PWYBA H--trans. ATP synth.-Synechocystis	142	177.3	0.00058	26.5	170
PWSP6 H--trans. ATP synth.-spinach chloro.	138	173.2	0.00098	24.2	231
PWYCA6 H--trans. ATP synth.-cyanobacteria	127	158.9	0.0062	26.3	167
LMNT6 H--trans. ATP synth.-tobacco chloro.	126	158.1	0.0069	22.1	231
LMLV6 H--trans. ATP synth.-Marchantia chloro.	126	158.0	0.0069	24.0	167
PWEG6C H--trans. ATP synth.-Mugilans chloro.	123	154.1	0.011	25.7	214
S17420 ubiquinol-cytochrome-c reductase	113	138.0	0.09	23.4	158
S17418 ubiquinol-cytochrome-c reductase	108	131.7	0.20	24.5	208
QK028M NADH dehydrogenase (ubiquinone)	107	131.2	0.22	26.1	211
S17415 ubiquinol-cytochrome-c reductase	105	127.9	0.33	27.7	137
DMH022 NADH dehydrogenase (ubiquinone)	103	126.1	0.41	20.1	149
QRECA amino acid trans. protein-E. coli	104	125.1	0.47	23.4	111
CBU ubiquinol-cytochrome-c reductase	102	124.1	0.53	26.8	205
S17419 ubiquinol-cytochrome-c reductase	101	122.9	0.63	23.4	158
S17407 ubiquinol-cytochrome-c reductase	98	120.3	0.87	23.6	140
QBE85 integral membrane protein- herp	98	119.4	0.99	20.8	202

Courtesy: William Pearson

BLAST

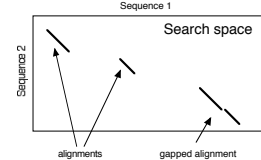
BLAST = Basic Local Alignment Search Tool
 BLAST uses a word based heuristic similar to that of FASTA to approximate a simplification of the SW algorithm known as the “maximal segment pairs” algorithm

MSP alignments are valuable because their statistics (Karlin, Altschul) are well understood

Basic BLAST does not allow gaps, thus, the evolutionary model requires that there be a long region of sequence that has evolved without insertions or deletions (indels) that would disrupt the alignment

Alignment Overview

Sequence alignment takes place in a 2-dimensional space where diagonal lines represent regions of similarity. Gaps in an alignment appear as broken diagonals. The search space is sometimes considered as 2 sequences and sometimes as query x database.



- Global alignment vs. local alignment
 - BLAST is local
- Maximum scoring pair (MSP) vs. High-scoring pair (HSP)
 - BLAST finds HSPs (usually the MSP too)
- Gapped vs. ungapped
 - BLAST can do both

Basic BLAST Algorithms

- **BLASTN** - compares a nucleotide query to a nucleotide database
- **BLASTP** - compares a protein query to a protein database
- **BLASTX** - compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- **TBLASTN** - compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
- **TBLASTX** - compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that tblastx program cannot be used with the nr database on the BLAST Web page.

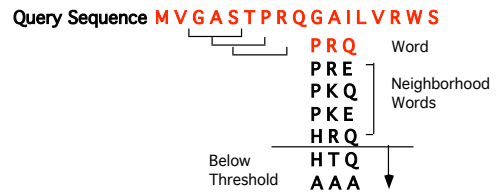
The 5 Standard BLAST Programs

Program	Database	Query	Typical Uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, amplicers, ESTs, and repeats to a genome. Identifying related transcripts.
BLASTP	Protein	Protein	Identifying common regions between proteins. Collecting related proteins for phylogenetic analysis.
BLASTX	Protein	Nucleotide	Finding protein-coding genes in genomic DNA.
TBLASTN	Nucleotide	Protein	Identifying transcripts similar to a known protein (finding proteins not yet in GenBank). Mapping a protein to genomic DNA.
TBLASTX	Nucleotide	Nucleotide	Cross-species gene prediction. Searching for genes missed by traditional methods.

BLAST

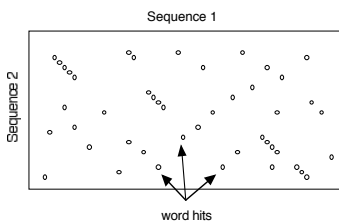
- BLAST is less sensitive than Smith-Waterman
- Basic BLAST uses a word size of 3 for proteins and is more sensitive than FASTA (even though FASTA uses a word of size 2)
- Basic BLAST uses a word size of 11 or 12 for nucleic acid sequences
- The Heuristic is applied to the words in BLAST via a "threshold value, T" for alignments of words.

Blast in a Nutshell



The BLAST Algorithm: Seeding (W and T)

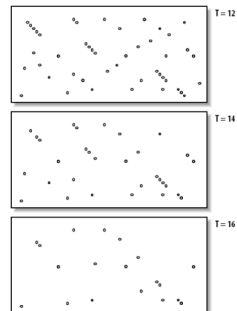
- Speed gained by minimizing search space
- Alignments require word hits
- Neighborhood words
- W and T modulate speed and sensitivity

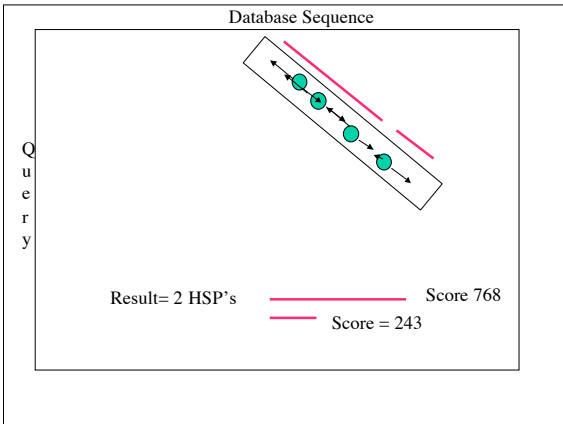
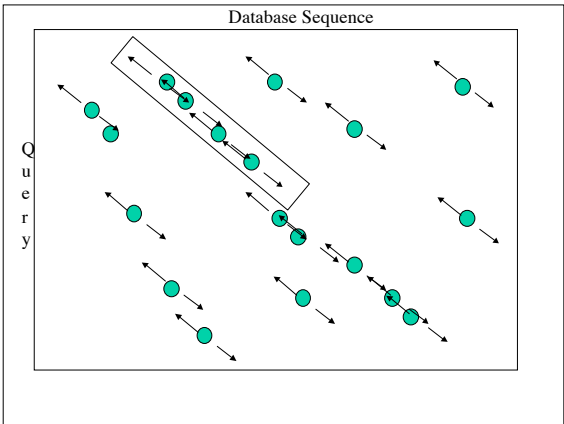
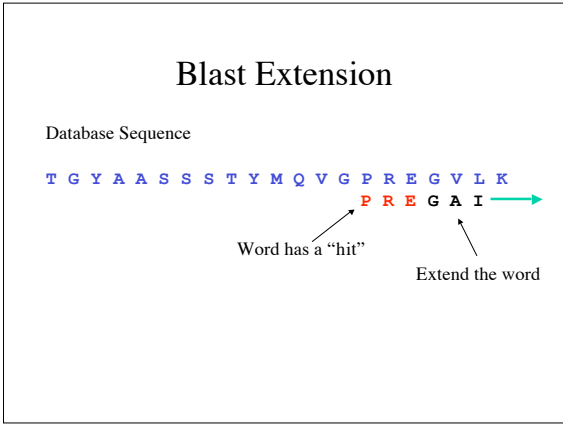
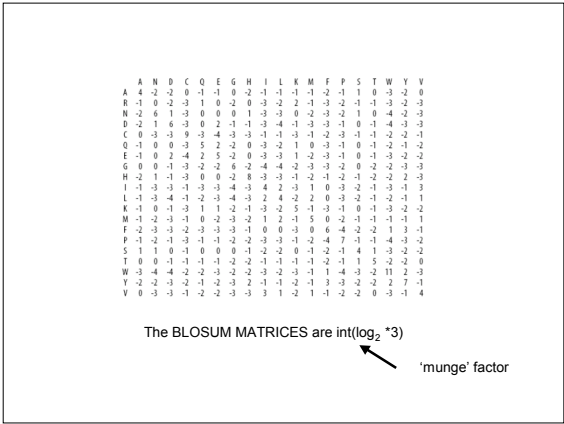


BLOSUM62 neighborhood of RGD

RGD	17
KGD	14
QGD	13
RGE	13
EGD	12
HGD	12
NGD	12
RGN	12
AGD	11
MGD	11
RAD	11
RGQ	11
RGS	11
RND	11
RSD	11
SGD	11
TGD	11

T=12





When does extension stop?

- When you hit the end of the sequence
- Or more likely when the "score" drops off by some number "X" from its optimal score
- The extension has no hope of achieving some minimal cut off score (~55-70, for BLOSUM 62)
- Note: in older versions of blast (prior to 2.0), there is no gapping. If there are multiple hits to a given gene that are not continuous, they are reported as "HSP"s. These HSP's need to be manually assembled into an alignment.

The Statistics

- The score is literally the score of your alignment according to the chosen substitution matrix and gap penalty (Sum based on each pair of residues).
- Since different matrices will give different scores for the same sequence, a normalized "bit" score is provided that removes the effects of scoring matrix upon the score. The bigger the bit score, the better.
- The E value is the probability of observing the null hypothesis. The null hypothesis is that the observed database hit occurred by chance (for this given query, matrix and database [size]).

Some common parameter values

- Normal word sizes for proteins are $W=3$ with $T = 14$ or $W=4$ with $T=16$.
- Normal word sizes for nucleic acids are $W=11$ or $W=12$
- The default scoring matrix for nucleic acid sequences is $(+1, -3)$ for NCBI BLAST and $(+5, -4)$ for WUBLAST

```

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PFF
Posted date: Oct 19, 2002 5:06 AM
Number of letters in database: 386,243,062
Number of sequences in database: 1,212,440

Lambda      K      H
0.316      0.133  0.373

Gapped
Lambda      K      H
0.267      0.0410 0.140

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 251,807,139
Number of Sequences: 1212440
Number of extensions: 1012625
Number of successful extensions: 46557
Number of sequences better than 10.0: 339
Number of HSP's better than 10.0 without gapping: 150
Number of HSP's successfully gapped in prelim test: 151
Number of HSP's that attempted gapping in prelim test: 47307
Number of HSP's gapped (non-prelim): 505
length of query: 369
length of database: 386,243,062
effective HSP length: 124
effective length of query: 245
effective length of database: 235,900,502
effective search space: 57795622990
effective search space used: 57795622990
T: 11
A: 40
X1: 16 ( 7.3 bits)
X2: 39 (14.6 bits)
X3: 24 (24.7 bits)
S1: 41 (21.6 bits)
S2: 73 (32.7 bits)

```

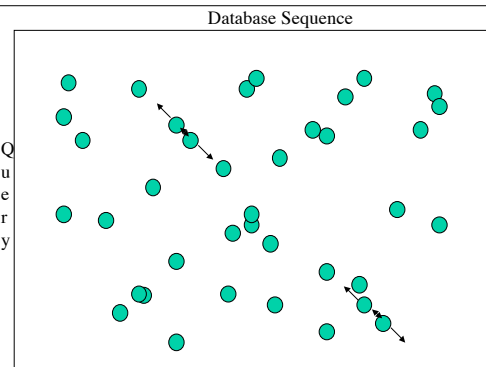
Gapped BLAST & PSI BLAST

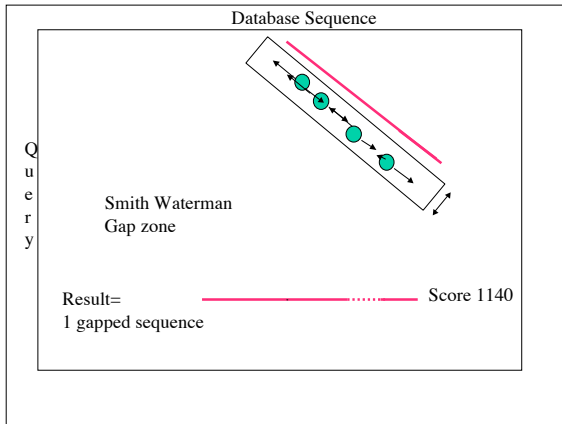
Gapped BLAST(Blast2.0) 3 Changes to the Algorithm

- Threshold for neighborhood word generation was decreased.
- Criterion for extending word pairs modified, there must be two hits on the same diagonal within some distance X, (this gives an increase in speed)
- Smith-Waterman calculations are used to produce the final alignment on successful extensions (thus, it will contain gaps)

Word Extension

- In the older versions of BLAST, if a word pair with a score above T was encountered when screening the DB, it was extended.
- In the newer version, two non-overlapping words located at some distance X (the “hitdist”) from each other must hit the same sequence in the DB before an extension is performed.
- To maintain sensitivity, must lower the value of T. This yields more hits, but few are extended.





The BLAST Algorithm: 2-hit Seeding

- Alignments tend to have multiple word hits.
- Isolated word hits are frequently false leads.
- Most alignments have large ungapped regions.
- Requiring 2 word hits on the same diagonal (of 40 aa for example), greatly increases speed at a slight cost in sensitivity.

Gapped Alignment

- Original BLAST found many HSP's and used all to generate a SUM statistic
- If you gap then you only need to find only one rather than all ungapped alignments.
- T is lowered to achieve more hits on initial scan
- Only pairs of hits on the same diagonal within some distance "H" are extended
- Gapped alignments are achieved via dynamic programming to extend the pairs of aligned residues in both directions within some window of gap tolerance.

PSI-BLAST

- Distant relationships are often best detected by motif or profile searches rather than pairwise comparisons
- BLAST uses a generalized matrix
- PSI-BLAST automatically generates a new matrix based on the output from the previous BLAST search.
- May not be as sensitive as motif search but is very general and easy to use.

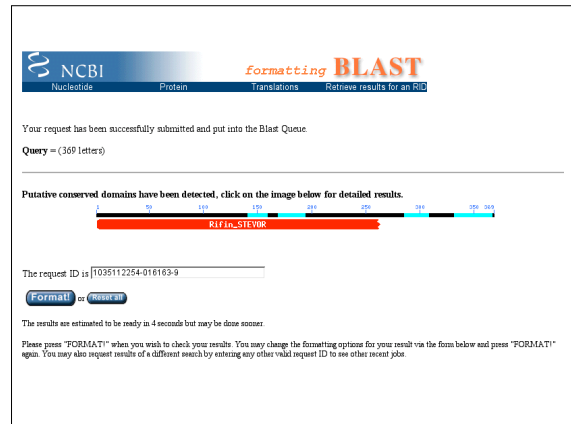
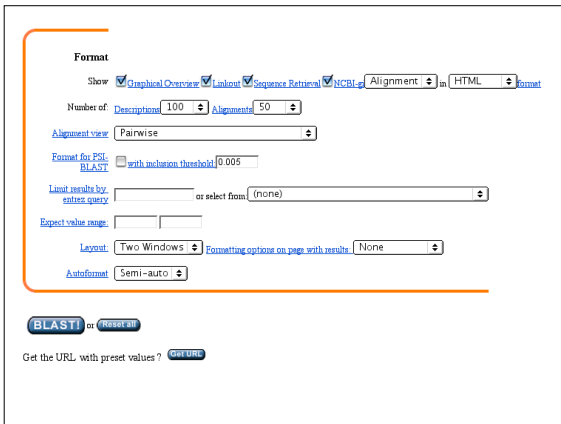
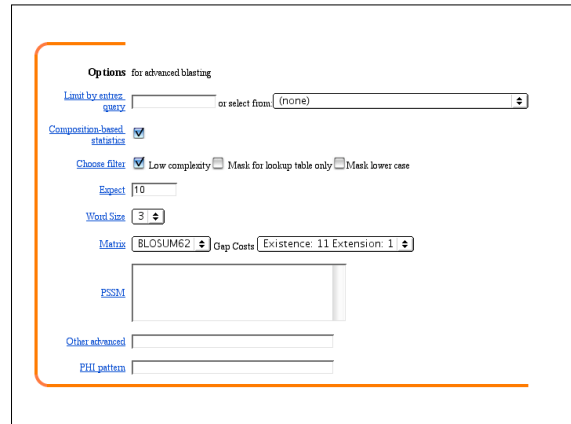
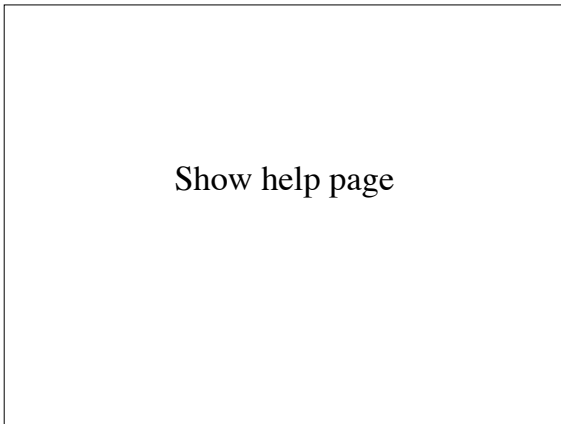
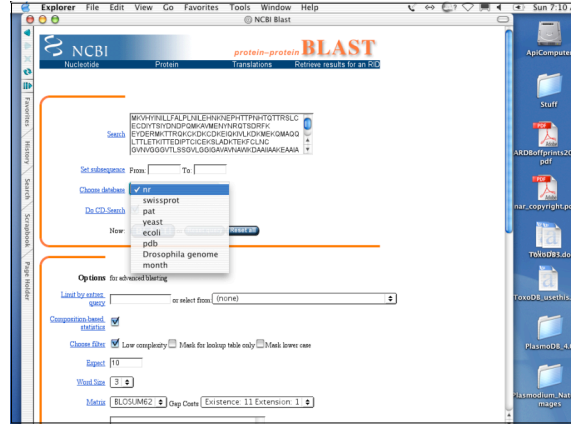
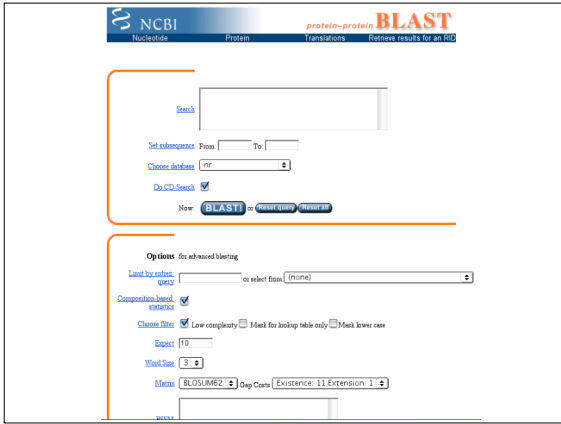
3 Changes to Algorithm

- Criterion for extending word pairs modified, this gives an increase in speed
- Ability to create gapped alignments added
- BLAST searches may be iterated, with a position-specific matrix generated from significant alignments found in round i used in round $i + 1$.

A PSSM (position specific scoring matrix) for PSI-BLAST

The 20 Amino acids

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
20 N	0	0	3	-2	-4	2	0	0	-2	0	0	2	-2	-4	-3	2	0	-5	-3	-3
21 S	-2	0	3	0	-4	0	0	0	-2	-4	-4	1	-3	-4	-3	2	2	4	-3	-3
22 G	1	0	2	-2	-3	0	-2	1	2	-2	0	1	-2	-3	-3	1	-2	-4	-3	0
23 W	-2	2	1	1	-4	0	1	0	2	-1	-3	0	-3	2	-3	1	-2	3	-2	-3
24 D	-3	0	0	4	-4	-1	3	-3	1	-2	0	0	-2	-4	0	-2	0	-5	-3	-1
25 Q	-2	0	1	0	-4	2	3	0	-2	-1	-4	-1	-3	-3	-3	1	2	-4	0	-3



Command line BLAST

Format: algorithm db query options

Example: `blastp nr myprot.txt -
matrix=pam70 V=10 B=10`

Example: `blastn nt mynuc.txt M=5 N=-4
E=1.0e-5`

Example: `blastn nt mynuc.txt M=5 N=-4
E=1.0e-5 > blast.out`

Making your own BLAST DB

- Any sequence file of fasta formatted sequences can be turned into a BLAST DB.
- How you do this depends on which BLAST variant you are using.
- **NCBI BLAST-protein DB:** `formatdb -p T myseqfile`
- **NCBI BLAST-nucleotide DB:** `formatdb -p F myseqfile`
- **WUBLAST - proteinDB:** `xdformat -p myseqfile`
- **WUBLAST-nucleotideDB:** `xdformat -n myseqfile`

Practical Exercises

- Install the WU-BLAST program in linux
- Make your own custom BLAST-searchable database
- Run a command-line BLAST search in Linux
- Run a PSI-BLAST search at NCBI
- Download BLAST results from NCBI