Homology

Bio5488
Ting Wang
1/27/19, 1/29/19
How to align them?

Why we can align them?

Why +1 for match, and -1 for mismatch?

What does the score mean?

Is 16 a good score?
Outline

- Nobel-prize-worthy work on homology
- What is homology?
  - How to detect homology?
  - How to quantify homology?
- How to use homology?
- Homology beyond sequence analysis
- Next-gen sequencing alignment
Russell Doolittle  
(Bishop and Varmus)

Simian Sarcoma Virus one Gene, v-sis, is Derived from the Gene (or Genes) Encoding a Platelet-Derived Growth Factor  

Author(s): Russell F. Doolittle, Michael W. Hunkapiller, Leroy E. Hood, Sushilkumar G. Devare, Keith C. Robbins, Stuart A. Aaronson, Harry N. Antoniades  

Bishop and Varmus strategy (Nobel prize 1989)

Doolittle strategy (could be the first Nobel prize for computational biology)

What is the significance?
A few Definitions

**Homologs**: genes/sequences sharing a common origin

**Orthologs**: genes originating from a single ancestral gene in the last common ancestor of the compared genomes; genes related via speciation

**Paralogs**: genes related via duplication

**Xenolog**: sequences that have arisen out of horizontal transfer events (symbiosis, viruses, etc)

**Co-orthologs**: two or more genes in one lineage that are, collectively, orthologous to one or more genes in another lineage due to a lineage-specific duplication(s)

**Outparalogs**: paralogous genes resulting from a duplication(s) preceding a given speciation event

**Inparalogs**: paralogous genes resulting from a lineage-specific duplication(s) subsequent to a given speciation event
Need ancestral sequences to distinguish orthologs and paralogs
Similarity versus Homology

• Similarity refers to the likeness or % identity between 2 sequences
• Similarity means sharing a statistically significant number of bases or amino acids
• Similarity does not imply homology
• Similarity can be quantified
• It is ok to say that two sequences are X% identical
• It is ok to say that two sequences have a similarity score of Z
• It is generally incorrect to say that two sequences are X% similar

• Homology refers to shared ancestry
• Two sequences are homologous if they are derived from a common ancestral sequence
• Homology usually implies similarity
• Low complexity regions can be highly similar without being homologous
• Homologous sequences are not always highly similar
• A sequence is either homologous or not.
• Never say two things are X% homologous
Why Compare Sequences?

• Sequence comparisons lie at the heart of all bioinformatics
• Identify sequences
  • What is this thing I just found?
• Compare new genes to known ones
• Compare genes from different species
  • information about evolution
• Guess functions for entire genomes full of new gene sequences
  • Metagenomics
• What does it matter if two sequences are similar or not?
  • Globally similar sequences are likely to have the same biological function or role
  • Locally similar sequences are likely to have some physical shape or property with similar biochemical roles
  • If we can figure out what one does, we may be able to figure out what they all do
Sequence alignment

• How to optimally align two sequences
  • Dot plots
  • Dynamic programming
    • Global alignment
    • Local alignment

• How to score an alignment

• Fast similar sequence search
  • BLAST
  • BLAT
  • More recent development: short read alignment

• Determine statistical significance

• Using information in multiple sequence alignment to improve sensitivity
Visual Alignments (Dot Plots)

• Build a comparison matrix
  • Rows: Sequence #1
  • Columns: Sequence #2

• Filling
  • For each coordinate, if the character in the row matches the one in the column, fill in the cell
  • Continue until all coordinates have been examined
Noise in Dot Plots

• Nucleic Acids (DNA, RNA)
  • 1 out of 4 bases matches at random

• Windowing helps reduce noise
  • Can require >X bp match before plotting
  • Percentage of bases matching in the window is set as threshold
Met14 vs Met2
“DotPlot”

Match = 1
Mismatch = -1
Gray: 1
α chain of human hemoglobin

β chain of human hemoglobin
MAZ: Myc associated zinc finger isoform 1 self alignment
Human vs Chimp Y chromosome comparison
Dot plots of DNA sequence identity between chimpanzee and human Y chromosomes and chromosomes 21

Aligning sequences by residue

- Match: award
- Mismatch (substitution or mutation): penalize
- Insertion/Deletion (INDELS – gaps): penalize (gap open, gap extension)
More than one solution is possible

• Which alignment is best?
Alignment Scoring Scheme

• Possible scoring scheme:
  match: +2
  mismatch: -1
  indel –2

• Alignment 1: 5*2 + 1*-1 + 4*-2 = 10 – 1 – 8 = 1
• Alignment 2: 6*2 + 1*-1 + 2*-2 = 12 – 1 – 4 = 7
Dynamic Programming

• **Global Alignments:**

• **Local Alignments:**
  • Smith T.F. and Waterman M.S. (1981) *J. Mol. Biol.* 147, 195-197
  • One simple modification of Needleman/Wunsch: when a value in the score matrix becomes negative, reset it to zero (begin of new alignment)

• Guaranteed to be mathematically optimal:
  • Given two sequences (and a scoring system) these algorithms are guaranteed to find the very best alignment between the two sequences!

• **Slow $N^2$ algorithm**

• Performed in 2 stages
  • Prepare a scoring matrix using recursive function
  • Scan matrix diagonally using traceback protocol
$S(i,j) = \max \left\{ S(i-1,j-1) + \text{score}(a_i, b_j), S(i-1,j) + \gamma, S(i,j-1) + \gamma \right\}$
Dynamic Programming

![Dynamic Programming Diagram]

The diagram illustrates a dynamic programming problem, likely involving sequence comparison or alignment. The matrices show the distances between corresponding characters of two sequences, "GENETICS" and "GENTICS", with the goal of finding the minimum cost of transformation. The asterisk (*) indicates the optimal alignment path.
DP (demo)

- Match=5, mismatch=-3, gap=-4

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DP (demo)

\[ S_{1,1} = \text{MAX}\{S_{0,0} + 5, \ S_{1,0} - 4, \ S_{0,1} - 4, 0\} = \text{MAX}\{5, -4, -4, 0\} = 5 \]
DP (demo)

\[ S_{1,2} = \text{MAX}\{S_{0,1}-3, S_{1,1}-4, S_{0,2}-4, 0\} = \text{MAX}\{0-3, 5-4, 0-4, 0\} = \text{MAX}\{-3, 1, -4, 0\} = 1 \]
$$S_{1,3} = \text{MAX}\{S_{0,2} - 3, S_{1,2} - 4, S_{0,3} - 4, 0\} = \text{MAX}\{0 - 3, 1 - 4, 0 - 4, 0\} = \text{MAX}\{-3, -3, -4, 0\} = 0$$
Trace Back (Local Alignment)

• Maximum local alignment score is the highest score anywhere in the matrix (14 in this example)

• 14 is found in two separate cells, indicating two possible multiple alignments producing the maximal local alignment score
Trace Back (Local Alignment)

• Trace-back begins in the position with the highest value.

• At each cell, we look to see where we move next according to the pointers.

• When a cell is reached where there is not a pointer to a previous cell, we have reached the beginning of the alignment.
Trace Back Demo
Trace Back Demo

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>7</td>
<td>12</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>9</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Trace Back Demo
Maximum Local Alignment

G A A T T C - A
|   |   |   |
G G A T - C G A
+ - + + - + - +
5 3 5 5 4 5 4 5
=14

G A A T T C - A
|   |   |   |
G G A T - C G A
+ - + + - + - +
5 3 5 5 4 5 4 5
=14
Linear vs. Affine Gaps

• So far, gaps have been modeled as linear

• More likely contiguous block of residues inserted or deleted
  • 1 gap of length k rather than k gaps of length 1

• Can create scoring scheme to penalize big gaps relatively less
  • Biggest cost is to open new gap, but extending is not so costly
Affine Gap Penalty

\[ w_x = g + r(x-1) \]

- \( w_x \): total gap penalty
- \( g \): gap open penalty
- \( r \): gap extend penalty
- \( x \): gap length

- gap penalty chosen relative to score matrix
Scoring Alignments

• Pick a scoring matrix
  • BLOSUM62
  • PAM250
  • Match=5, mismatch=-4

• Decide on gap penalties
  • -gap opening penalty (-8)
  • -gap extension penalty (-1)

• Assume every position is independent

• Sum scores at each position
  • \[\log(x*y) = \log x + \log y\]
Scoring Matrices

$$S_{ij} = \frac{\log\left(\frac{q_{ij}}{p_i p_j}\right)}{\lambda}$$

• An empirical model of evolution, biology and chemistry all wrapped up in a 20 X 20 (or 4 X 4) table of numbers

• Structurally or chemically similar residues should ideally have high diagonal or off-diagonal numbers

• Structurally or chemically dissimilar residues should ideally have low diagonal or off-diagonal numbers

• What does the score mean: The likelihood of seeing two residues align (preserved) than random expected.
# Scoring Alignments

**Blosum62 Scoring Matrix**

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

Developed for distantly related proteins

Substitutions only from multiple alignments of conserved regions of protein families, hand curated, constitute the known homologous blocks

Identity threshold to define conserved blocks can be varied, e.g. 62% identity gives BLOSUM62

Scores calculated from frequency of amino acids in aligned pairs compared to what would be expected due to abundance alone, given all sequences

Blosum Matricies

What score should we give to a ser residue aligned with a thr residue?

\[
\text{score}(S:T) \propto \log_2 \frac{P(S:T \mid \text{homology})}{P(S:T \mid \text{random})}
\]
Example of deriving Blosum scores for S:S, S:T, and T:T

Database of known alignments

<table>
<thead>
<tr>
<th>SDHIP</th>
<th>HKSA</th>
<th>WMFET</th>
<th>RTQC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDHLP</td>
<td>HRTA</td>
<td>WMFDT</td>
<td>RTNC</td>
</tr>
<tr>
<td>SDHIP</td>
<td>HKS</td>
<td>WLFDT</td>
<td>KTQC</td>
</tr>
<tr>
<td>SEHLP</td>
<td></td>
<td>K</td>
<td>SQC</td>
</tr>
<tr>
<td>SEHLP</td>
<td></td>
<td>K</td>
<td>TQC</td>
</tr>
</tbody>
</table>

**Homology Model (consider each pair of sequences separately)**

S:S pairs in alignments = 11  
P(S:S|homology) = 11/117 = .094
S:T pairs in alignments = 6  
P(S:T|homology) = 6/117 = .051
T:T pairs in alignments = 9  
P(T:T|homology) = 9/117 = .078

Total pairs in alignments = 117
Example of deriving Blosum scores for S:S, S:T, and T:T

Database of known alignments

```
| SDHIP | HKSA | WMFET | RTQC |
| SDHLP | HRTA | WMFDT | RTNC |
| SDHIP | HKSG | WLFDT | KTQC |
| SEHLPE | KSQC |
| SEHLPE | KTQC |
```

Random Model

- Number of S residues = 8  \[ P(S:S|\text{random})=P(S)P(S)=(8/72)^2=0.012 \]
- Number of T residues = 8  \[ P(S:T|\text{random})=2*P(S)P(T)=2*(8/72)^2=0.024 \]
- Total residues = 72  \[ P(T:T|\text{random})=P(T)P(T)=(8/72)^2=0.012 \]
Example of deriving Blosum scores for S:S, S:T, and T:T

\[
\text{score}(S : S) = \log_2 \frac{P(S : S \mid \text{homology})}{P(S : S \mid \text{random})} = \log_2 \frac{0.094}{0.012} = 2.96
\]

\[
\text{score}(S : T) = \log_2 \frac{P(S : T \mid \text{homology})}{P(S : T \mid \text{random})} = \log_2 \frac{0.051}{0.024} = 1.09
\]

\[
\text{score}(T : T) = \log_2 \frac{P(T : T \mid \text{homology})}{P(T : T \mid \text{random})} = \log_2 \frac{0.078}{0.012} = 2.70
\]
BLOSUM and PAM

<table>
<thead>
<tr>
<th>BLOSUM 45</th>
<th>BLOSUM 62</th>
<th>BLOSUM 90</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM 250</td>
<td>PAM 160</td>
<td>PAM 100</td>
</tr>
</tbody>
</table>

*More Divergent*  *Less Divergent*

- BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.
- PAM matrices: point accepted mutation
Scoring Matrices Take Home Points

• Based on log odds scores
  • Ratios > 1 give positive scores, ratios < 1 give negative scores
  • Because $\log(x \times y) = \log x + \log y$ the score of an alignment is the sum of the scores for each pair of aligned residues

• Assume independence of adjacent residues when scoring

• Introduced the concept that the frequency of a residue in a multiple alignment is informative
Fast Similar Sequence Search

• Can we run Smith-Waterman between query and every DB sequence?
• Yes, but too slow!
• General approach
  • Break query and DB sequence to match subsequences
  • Extend the matched subsequences, filter hopeless sequences
  • Use dynamic programming to get optimal alignment
BLAST

• Basic Local Alignment Search Tool
• Altschul et al. *J Mol Biol.* 1990
• One of the most widely used bioinformatics applications
  • Alignment quality not as good as Smith-Waterman
  • But much faster, supported at NCBI with big computer cluster
• For tutorials or information: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html
BLAST Algorithm Steps

• Query and DB sequences are optionally filtered to remove low-complexity regions
  • E.g. ACACACACACAA, TTTTTTTTTT
BLAST Algorithm Steps

• Query and DB sequences are optionally filtered to remove low-complexity regions

• Break DB sequences into k-mer words and hash their locations to speed later searches
  • \( k \) is usually 11 for DNA/RNA and 3 for protein

```
LPPQGLL
LPP
PPQ
PQG
QGL
GLL
```
BLAST Algorithm Steps

- Query and DB sequences are optionally filtered to remove low-complexity regions
- Break DB sequences into k-mer words and hash their locations to speed later searches
- Each k-mer in query find possible k-mers that matches well with it
  - “well” is evaluated by substitution matrices
BLAST Algorithm Steps

• Only words with $\geq T$ cutoff score is kept
  • $T$ is usually 11-13, ~ 50 words make $T$ cutoff
  • Note: this is 50 words at every query position

• For each DB sequence with a high scoring word, try to extend it in both ends

  Query: LP PQG LL
  DB seq: MP PEG LL
  HSP score $9 + 15 + 8 = 32$

  • Form HSP (High-scoring Segment Pairs)
  • Use BLOSUM to score the extended alignment
  • No gaps allowed
The BLAST Search Algorithm

Query Word

Neighbourhood Words

Score Threshold (13)

High-scoring Segment Pair
BLAST Algorithm Steps

• Keep only statistically significant HSPs
  • Based on the scores of aligning 2 random seqs
• Use Smith-Waterman algorithm to join the HSPs and get optimal alignment
  • Gaps are allowed default (-11, -1)
BLAST algorithm summary

“query”

“seeds”
(111111)
(111000111)

“neighborhood words”
(branch and bound algorithm)

Indexing all seeds

“subjects” (database)

Scan the index and find all word hits

DP extension to recover the high scoring pairs

Extending high scoring pairs

Evaluate Significance of HSPs by
Karlin-Altschul Statistic: \( E = KN \exp(-\lambda \cdot S) \)
Different BLAST Programs

BLAST DB:
- nr (non-redundant):
  - GenBank, RefSeq, EMBL...
- est:
  - expressed sequences (cDNA), redundant
- Swissprot and pdb:
  - protein databases

If query is DNA, but known to be coding (e.g. cDNA):
- Translate cDNA into protein
- Zero gap-extension penalty
# Different BLAST Programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastp</td>
<td>Compares an amino acid query sequence against a protein sequence database.</td>
</tr>
<tr>
<td>blastn</td>
<td>Compares a nucleotide query sequence against a nucleotide sequence database.</td>
</tr>
<tr>
<td>blastx</td>
<td>Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.</td>
</tr>
<tr>
<td>tblastn</td>
<td>Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.</td>
</tr>
<tr>
<td>tblastx</td>
<td>Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is too computationally intensive.</td>
</tr>
</tbody>
</table>
PSI-BLAST

- **Position Specific Iterative BLAST**
  - Align high scoring hits in initial BLAST to construct a profile for the hits
  - Use profile (PSSM) for next iteration BLAST

- Find remote homologs or protein families
- FP sequences can degrade search quickly
PSI-BLAST

Query
GKATFGKLFAAHPEYQQMFRFF

Initial Matches
GKATFGKLFAAHPEYQQMFRFF
GKDCLIKFLSAHPQMAAVFGFS
GLELWKILREHEIKAFFSRV
SLHFWEKFLHDDPDLVSLFKRV
GFDILISVLDKPVLDQALAHY

PSSM Profile

<table>
<thead>
<tr>
<th>Pos</th>
<th>Ala</th>
<th>Cys</th>
<th>Glu</th>
<th>Asp</th>
<th>Phe</th>
<th>Gly</th>
<th>His</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-20</td>
<td>-30</td>
<td>-30</td>
<td>-40</td>
<td>20</td>
<td>-30</td>
<td>-20</td>
</tr>
<tr>
<td>2</td>
<td>-10</td>
<td>-50</td>
<td>-20</td>
<td>-30</td>
<td>-20</td>
<td>-30</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>-30</td>
<td>0</td>
<td>0</td>
<td>-50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>-30</td>
<td>-80</td>
<td>-50</td>
<td>-40</td>
<td>0</td>
<td>-60</td>
<td>-20</td>
</tr>
<tr>
<td>5</td>
<td>-50</td>
<td>-30</td>
<td>-60</td>
<td>-60</td>
<td>100</td>
<td>-70</td>
<td>-10</td>
</tr>
</tbody>
</table>
...

Refined Matches
TSTMYKMFQTYPEVRSLFYVNMT
GKATFGKLFAAHPEYQQMFRFF
SGIAMKRQALVFAGILQEFVAN
GKDCLIKFLSAHPQMAAVFGFS
WAKASAAWGTAGEFMMALFDA
GLELWKILREHEIKAFFSRV
SLHFWEKFLHDDPDLVSLFKRV
GVALMFTLFDNQETIGYFKRL
GFDILISVLDKPVLDQALAHY
VDPHLMVSEWPEKLWMSEFWPI

Search
Search again

Reciprocal Blast

• Search for orthologous sequences between two species

  • GeneA in Species1 BLAST Species2 □ GeneB
  • GeneB in Species2 BLAST Species1 □ GeneA

  • GeneA orthologous GeneB

• Also called bi-directional best hit
BLAT

• **BLAST-Like Alignment Tool**
  • Compare to BLAST, BLAT can align much longer regions (MB) really fast with little resources
  • E.g. can map a sequence to the genome in seconds on one Linux computer
  • Allow big gaps (mRNA to genome)
  • Need higher similarity (> 95% for DNA and 80% for proteins) for aligned sequences

• **Basic approach**
  • Break long sequence into blocks
  • Index k-mers, typically 8-13
  • Stitch blocks together for final alignment
BLAT: Indexing

Genome: cacaattatcagaccgc

3-mers: cac aat tat cac gac cgc

Index: aat 3, gac 12
       cac 0,9,9 tat 6
       cgc 15

cDNA (mRNA -> DNA): aattctcac

3-mers: aat att ttc tct ctc tca cac
        0 1 2 3 4 5 6

hits: aat 0,3 -3
      cac 6,0 6
      cac 6,9 -3

clump: cac**AAT**tat**CAC**gaccgc
        |||   |||
aattctcac
BLAT Example

• Enter sequence and parameters

BLAT Search Genome

Human BLAT Search

Genome: Human
Assembly: May 2004
Query type: DNA
Sort output: query.score
Output type: hyperlink

Paste a query sequence to find its location in the genome. Multiple sequences may be searched at once if separated by a line starting with > followed by the sequence name.

Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

Upload sequence: Browse... Submit File
## BLAT Example

- Get result instantly!!

### Human BLAT Results

<table>
<thead>
<tr>
<th>ACTIONS</th>
<th>QUERY</th>
<th>ACTION</th>
<th>ACCESS</th>
<th>CHROM</th>
<th>STRAND</th>
<th>START</th>
<th>END</th>
<th>QSIZE</th>
<th>IDENTITY</th>
<th>CHRO</th>
<th>STRAND</th>
<th>START</th>
<th>END</th>
<th>SEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>1734</td>
<td>1</td>
<td>1750</td>
<td>1750</td>
<td>99.9%</td>
<td>12</td>
<td>-</td>
<td>101735652</td>
<td>10163548</td>
<td>77167</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>634</td>
<td>1</td>
<td>641</td>
<td>1750</td>
<td>99.7%</td>
<td>12_random</td>
<td>-</td>
<td>281957</td>
<td>286767</td>
<td>1011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>186</td>
<td>639</td>
<td>824</td>
<td>1750</td>
<td>100.0%</td>
<td>12_random</td>
<td>+</td>
<td>116309</td>
<td>116494</td>
<td>186</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>32</td>
<td>1456</td>
<td>1567</td>
<td>1750</td>
<td>97.1%</td>
<td>1</td>
<td>-</td>
<td>21612213</td>
<td>21612486</td>
<td>274</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>24</td>
<td>210</td>
<td>233</td>
<td>1750</td>
<td>100.0%</td>
<td>22</td>
<td>-</td>
<td>20374107</td>
<td>20374210</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>24</td>
<td>1377</td>
<td>1407</td>
<td>1750</td>
<td>88.5%</td>
<td>9</td>
<td>+</td>
<td>30738626</td>
<td>30738655</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>24</td>
<td>753</td>
<td>789</td>
<td>1750</td>
<td>96.2%</td>
<td>16</td>
<td>+</td>
<td>3494390</td>
<td>3494428</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>22</td>
<td>1009</td>
<td>1030</td>
<td>1750</td>
<td>100.0%</td>
<td>7</td>
<td>-</td>
<td>30429927</td>
<td>30429948</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>22</td>
<td>827</td>
<td>860</td>
<td>1750</td>
<td>82.4%</td>
<td>11</td>
<td>-</td>
<td>2145950</td>
<td>2145983</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>22</td>
<td>209</td>
<td>229</td>
<td>1750</td>
<td>100.0%</td>
<td>2</td>
<td>+</td>
<td>95471146</td>
<td>95471167</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>22</td>
<td>1564</td>
<td>1585</td>
<td>1750</td>
<td>100.0%</td>
<td>16</td>
<td>+</td>
<td>56691362</td>
<td>56691393</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>22</td>
<td>153</td>
<td>175</td>
<td>1750</td>
<td>100.0%</td>
<td>1</td>
<td>+</td>
<td>37402859</td>
<td>37402882</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>21</td>
<td>1095</td>
<td>1115</td>
<td>1750</td>
<td>100.0%</td>
<td>X</td>
<td>+</td>
<td>5847999</td>
<td>5848019</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>browser details</td>
<td>NM_000277.1</td>
<td>17</td>
<td>253</td>
<td>283</td>
<td>1750</td>
<td>77.5%</td>
<td>17</td>
<td>+</td>
<td>61370186</td>
<td>61370216</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary of Fast Search

• Fast sequence similarity search
  • Break seq, hash DB sub-seq, match sub-seq and extend, use DP for optimal alignment
  • *BLAST, most widely used, many applications with sound statistical foundations
  • *BLAT, align sequence to genome, fast yet need higher similarity
BLAST score and significance

• Report DB sequences above a threshold
  • E value: Number (instead of probability \ pvalue) of matches expected merely by chance

\[ E = Kmn e^{-\lambda S} \]

\[ p(s \geq x) \approx 1 - \exp[-e^{-x}] \]

• m, n are query and DB length
• K, \ are constants
• Smaller E, more stringent
Are these proteins homologs?

SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
       L P    W L   Y N    Y C     L

SEQ 2: QFFPLMPPAPYWILATDYENLPLVYSCTFFWLF

SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
       L P    W LDATYKNYA  Y C     L

SEQ 2: QFFPLMPPAPYWILDATYKNYALVYSCDVTYKL

SEQ 1: RVVNLVPS--FWVLDATYKNYAGAINYNCDVTYKLY
       RVV L PS    W LDATYKNYA  Y CDVTYKL

SEQ 2: RVVPLMPSAPYWILDATYKNYALVYSCDVTYKL

Most likely (score = 24)

MAYBE (score = 15)

Probably not (score = 9)
Significance of scores

Homology detection algorithm

HPDKKAHSIHAWILSKSKVLEGNTKEVVVDNLKT
LENENQGKCTIAEYKDGGKASVYNFSNVNGVE

Low score = unrelated
High score = homologs

How high is high enough?
Other significance questions

• Pairwise sequence comparison scores
• Microarray expression measurements
• Sequence motif scores
• Functional assignments of genes
• Call peaks from ChIP-seq data
The null hypothesis

- We are interested in characterizing the distribution of scores from sequence comparison algorithms.

- We would like to measure how surprising a given score is, assuming that the two sequences are not related.

- The assumption is called the null hypothesis.

- The purpose of most statistical tests is to determine whether the observed results provide a reason to reject the hypothesis that they are merely a product of chance factors.
Gaussian vs. Extreme Value Distribution (EVD)
What is the chance of picking a person at least 75 in. tall $P(X \geq 75)$?

$Z_{score}(x) = \frac{x - \mu}{\sigma} = \frac{75 - 68}{3} = 2.33$

From Table:
$z=2.33 \quad P=0.01$
Each alignment/score that BLAST returns is the very best alignment/score among a large number of alignments/scores for those two sequences (i.e., the EVD problem).

Computing a p-value

• The probability of observing a score >4 is the area under the curve to the right of 4.

• This probability is called a p-value.

• p-value = Pr(data|null)
Scaling the EVD

- An extreme value distribution derived from, e.g., the Smith-Waterman algorithm will have a characteristic mode $\mu$ and scale parameter $\lambda$.

\[ P(S \geq x) = 1 - \exp\left[-e^{-\lambda(x-\mu)}\right] \]

- These parameters depend upon the size of the query, the size of the target database, the substitution matrix and the gap penalties.
An example

You run BLAST and get a score of 45. You then run BLAST on a shuffled version of the database, and fit an extreme value distribution to the resulting empirical distribution. The parameters of the EVD are $\mu = 25$ and $\lambda = 0.693$. What is the p-value associated with 45?

\[
P(S \geq x) = 1 - \exp[-e^{-\lambda(x-\mu)}]
\]
\[
P(S \geq 45) = 1 - \exp[-e^{-0.693(45-25)}]
\]
\[
= 1 - \exp[-e^{-13.86}]
\]
\[
= 1 - \exp[-9.565 \times 10^{-7}]
\]
\[
= 1 - 0.999999043
\]
\[
= 9.565 \times 10^{-7}
\]
Summary of statistical significance

• A distribution plots the frequency of a given type of observation.

• The area under the distribution is 1.

• Most statistical tests compare observed data to the expected result according to the null hypothesis.

• Sequence similarity scores follow an extreme value distribution, which is characterized by a larger tail.

• The p-value associated with a score is the area under the curve to the right of that score.
• How to align them?

• Why we can align them?
• Why +1 for match, and -1 for mismatch?
• What does the score mean?
• Is 16 a good score?
Applying homology: concept and technology

• Genome evolution
  • Mammalian genome evolution
  • Human genome variation
  • Cancer genome evolution

• Gene finding
  • Comparative approaches
  • Ab initio approaches
    • Hidden Markov Model

• Protein structure
  • Threading

• Regulatory motif finding
  • Profile comparison

• Pathway/Network comparison
  • PathBLAST

• Conservation
  • Ultra conserved elements
  • Human accelerated regions
Gene prediction

• Comparing to a known gene from a different species

• Using EST evidence (aligning transcript to genome)

• Predicting from sequence (HHM)

• Using conservation
  • Signature of coding potential
  • What about RNA gene?

• Using other genomics signals
  • Specific epigenetic marks of promoters and gene bodies
Genscan (Burge and Karlin, 1998)

- Dramatic improvement over previous methods
- Generalised HMM
- Different parameter sets for different GC content regions (intron length distribution and exon stats)
Predicting non-coding RNA?

- From sequence?
  - Not clear which properties can be exploited
  - Sequence features such as promoters are too weak
- Histone modifications + conservation worked
Figure 1 | Intergenic K4–K36 domains produce multi-exonic RNAs.
So far, only linear sequence comparison
Expanding the idea of a sequence
Central theme of the new algorithm – compare profiles

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T</th>
<th>G</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T</th>
<th>G</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>
Met14 vs Met2
“DotPlot”

MET14 (1000nt)
MET2 (895nt)

Match = 1
Mismatch = -1
Gray: 1
Met14 vs Met2
PhyloNet

HSPs:
E < 0.1

TTTCACGTGA
P=1.75E-5

P=0.002
P=0.003
P=0.03
P=0.02

MET2 (895nt)

MET14 (1000nt)
PathBlast, NetworkBlast

Trey Ideker
Aligning Short Reads

(material for yourself)
0 and 1st generation sequencing

Pre-1992
“old fashioned way”
- S35 ddNTPs
- Gels
- Manual loading
- Manual base calling

1992-1999
- ABI 373/377
- Fluorescent ddNTPs*
- Gels
- Manual loading
- Automated base calling*

1999
- ABI 3700
- Fluorescent ddNTPs
- Capillaries*
- Robotic loading*
- Automated base calling
- Breaks down frequently

2003
- ABI 3730XL
- Fluorescent ddNTPs
- Capillaries
- Robotic loading
- Automated base calling
- Reliable*
Next or 2\textsuperscript{nd}-generation sequencing

454/Roche GS-20/FLX  
(Oct 2005)

ABI SOLiD  
(Oct 2007)

Illumina/Solexa  
1G Genetic Analyser  (Feb 2007)
Cluster generation

FLOWCELL STATION AREA

REAGENT COMPARTMENT AREA

8 channels (lanes)
IGA without cover
Flow cell imaging
A flow cell contains eight lanes. Each lane/channel contains three columns of tiles. Each column contains 100 tiles. Each tile is imaged four times per cycle – one image per base. 345,600 images for a 36-cycle run.
Data analysis pipeline

- Firecrest
  - tiff image files (345,600)
  - intensity files

- Bustard
  - Sequence files

- Eland

Additional Data Analysis

Alignment to Genome
Primary tools and analysis tasks

• Image processing
  • (unique to each manufacturer)

• Basecalling
  • (unique to each manufacturer)

• Align sequence reads to reference genome

• Assemble contigs and whole genomes using quality scores and/or paired-end information

• Peak finding for Chip-Seq applications
  • (and statistics to validate, map to regulated genes, etc)

• SNP calling/genotyping

• Transcript profiling
  • measure gene expression, identifying alternative splicing, etc.
NGS: Sequence alignment

- Map the large numbers of short reads to a reference genome
  - In a broader sense: Identify similar sequences (DNA, RNA, or protein) in consequence of functional, structural, or evolutionary relationships between them
  - Applications: Genome assembly, SNP detection, homology search, etc

- \textit{large} $\Rightarrow$ faster search speed
- \textit{short} $\Rightarrow$ greater search sensitivity.
Mapping Reads Back

• Hash Table (Lookup table)
  • FAST, but requires perfect matches

• Array Scanning
  • Can handle mismatches, but not gaps

• Dynamic Programming (Smith Waterman, Forward, Viterbi)
  • Indels
  • Mathematically optimal solution
  • Slow (most programs use Hash Mapping as a prefilter)

• Burrows-Wheeler Transform (BW Transform)
  • FAST (memory efficient)
  • But for gaps/mismatches, it lacks sensitivity
Many short read aligners

- Bfast
- BioScope
- Bowtie
- BWA
- CLC bio
- CloudBurst
- Eland/Eland2
- GenomeMapper
- GnuMap
- Karma
- MAQ
- MOM
- Mosaik
- MrFAST/MrsFAST
- NovoAlign
- PASS
- PerM
- RazerS
- RMAP
- SSAHA2
- Segemehl
- SeqMap
- SHRiMP
- Slider/SliderII
- SOAP/SOAP2
- Srprism
- Stampy
- vmatch
- ZOOM
- ......
Short read mapping

• Input:
  • A reference genome
  • A collection of many 25-100bp tags (reads)
  • User-specified parameters

• Output:
  • One or more genomic coordinates for each tag

• In practice, only 70-75% of tags successfully map to the reference genome. Why?
Multiple mapping

- A single tag may occur more than once in the reference genome.
- The user may choose to ignore tags that appear more than $n$ times.
- As $n$ gets large, you get more data, but also more noise in the data.
Inexact matching

• An observed tag may not exactly match any position in the reference genome.
• Sometimes, the tag *almost* matches one or more positions.
• Such mismatches may represent a SNP or a bad read-out.
• The user can specify the maximum number of mismatches, or a phred-style quality score threshold.
• As the number of allowed mismatches goes up, the number of mapped tags increases, but so does the number of incorrectly mapped tags.
Using base qualities to evaluate

READ: AGGTCCGGGATAACCGGGGAC

CHR1: CCGTCCGGGATAACCGGGGAC

CHR2: AGGTCCCGGATACCGGGGAC

Q: 30

BETTER R!

Q: 10+10

BETTER R!
Hash table (Eland, SOAP)

• Main idea: preprocess genome to speed up queries
  • Hash every substring of length k
  • K is a tiny constant

• For each query p, can easily retrieve all suffixes of the genome that start with p1, p2, … pk.

• Easy to implement.

• Significant speed up in practice.

• Large memory consumption.

• Inexact match is difficult.
  • Need multiple hash tables
  • More memory
Spaced seed alignment (MAQ)

- Tags and tag-sized pieces of reference are cut into small “seeds.”
- Pairs of spaced seeds are stored in an index.
- Look up spaced seeds for each tag.
- For each “hit,” confirm the remaining positions.
- Report results to the user.
Index the reference genome: Suffix Tree

- Each suffix corresponds to exactly one path from the root to a leaf.
- Edges spell non-empty strings.
- Construction: linear time and space.
- Check if a string of length m is a substring.
- Each substring is a prefix of a suffix!
Burrows-Wheeler (Bowtie, BWA)

- Store entire reference genome.
- Align tag base by base from the end.
- When tag is traversed, all active locations are reported.
- If no match is found, then back up and try a substitution.
Why Burrows-Wheeler?

BWT very compact:
- Approximately $\frac{1}{2}$ byte per base
- As large as the original text, plus a few “extras”
- Can fit onto a standard computer with 2GB of memory

- Linear-time search algorithm
  proportional to length of query for exact matches
Burrows-Wheeler Transform (BWT)

Burrows-Wheeler Matrix (BWM)

acaacg$

$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$aca

BWT

gc$aaac

g$acaac

acaacg$
Key observation

The $i$-th occurrence of character X in the last column corresponds to the same text character as the $i$-th occurrence of X in the first column.
Burrows-Wheeler Matrix

$\text{acaacg}$

$\begin{array}{c|c}
3 & \text{aacg}$\text{ac} \\
1 & \text{acaacg}$ \\
4 & \text{acg}$\text{aca} \\
2 & \text{caacg}$\text{a} \\
5 & \text{cg}$\text{acaa} \\
6 & \text{g}$\text{acaac}$ \\
\end{array}$

See the suffix array?
Burrows-Wheeler Transform

- Originally designed for data compression for large text
- Burrows-Wheeler matrix: sort lexicographically all cyclic rotations of S$
- BWT(S): the last column of Burrows-Wheeler matrix
- Compression: runs of repeated characters are easy to compress using move-to-front transform and run-length encoding, etc.
- BWT(S) is a reversible permutation of S
• BW Matrix Property: Last-First (LF) Mapping
• The ith occurrence of character X in the last column correspond to the same text character as the ith occurrence of X in the first column
Searching BWT
Searching BWT

BWT(aggcagcagact) = tggcc$ggaac

Search for pattern: gca
Human genome memory footprint