



BIOINF 525

http://bioboot.github.io/bioinf525_w16/

19-Jan-2016

MODULE OVERVIEW

Objective: Provide an introduction to the practice of bioinformatics as well as a practical guide to using common bioinformatics databases and algorithms

1.1. ▶ *Introduction to Bioinformatics*

1.2. ▶ **Sequence Alignment and Database Searching**

1.3. ▶ *Structural Bioinformatics*

1.4. ▶ *Genome Informatics: High Throughput Sequencing Applications and Analytical Methods*

WEEK ONE REVIEW

- Answers to last weeks homework (19/20):

[Answers week 1](#)

- Muddy Point Assessment (14/20):

[Responses](#)

- NCBI BLAST frustrations
- Need for FASTA header lines ">example1"
- More on protein structure viewing and finding
- "Nice Assignment".

THIS WEEK'S HOMEWORK

- Check out the “Background Reading” material online:

[Dynamic Programming](#)

[Database Searching](#)

- Complete the **lecture 1.2 homework questions**:

<http://tinyurl.com/bioinf525-quiz2>

TODAYS MENU

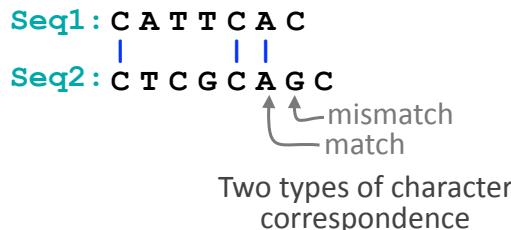
- Alignment basics
 - Why compare biological sequences?
- Homologue detection
 - Orthologs, paralogs, similarity and identity
 - Sequence changes during evolution
 - Alignment view: matches, mismatches and gaps
- Pairwise sequence alignment methods
 - Brute force alignment
 - Dot matrices
 - Dynamic programming (global vs local alignment)
- Rapid heuristic approaches
 - BLAST
- Practical database searching
 - PSI-BLAST and HMM approaches

Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.

Seq1: C A T T C A C

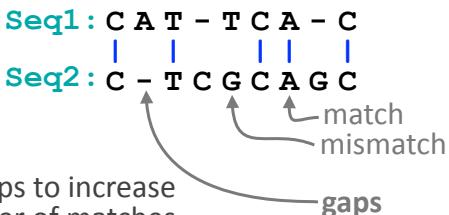
Seq2: C T C G C A G C

Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.



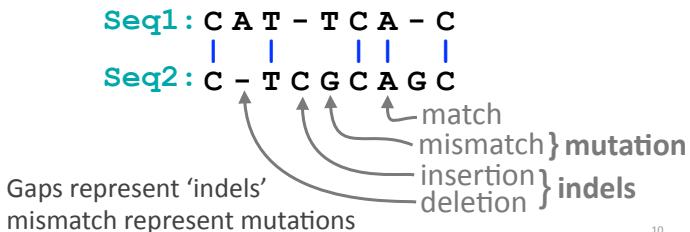
8

Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.



9

Basic Idea: Display one sequence above another with spaces (termed **gaps**) inserted in both to reveal **similarity** of nucleotides or amino acids.



10

Why compare biological sequences?

- To obtain **functional or mechanistic insight** about a sequence by inference from another potentially better characterized sequence
- To find whether two (or more) genes or proteins are **evolutionarily related**
- To find **structurally or functionally similar regions** within sequences (e.g. catalytic sites, binding sites for other molecules, etc.)
- Many practical bioinformatics applications...

11

Practical applications of sequence alignment include...

- **Similarity searching of databases**
 - Protein structure prediction, annotation, etc...
- **Assembly of sequence reads** into a longer construct such as a genomic sequence
- **Mapping sequencing reads to a known genome**
 - "Resequencing", looking for differences from reference genome - SNPs, indels (insertions or deletions)
 - Mapping transcription factor binding sites via ChIP-Seq (chromatin immuno-precipitation sequencing)
 - Pretty much all next-gen sequencing data analysis

Practical applications of sequence alignment include...

- **Similarity searching of databases**
 - Protein structure prediction
- **Assembly of sequence reads** into a longer construct such as a bacterial genome
- **Mapping sequencing reads to a known genome**
 - Looking for differences from reference genome - SNPs, indels (insertions or deletions)
 - Mapping transcription factor binding sites via ChIP-Seq (chromatin immuno-precipitation sequencing)
 - Pretty much all next-gen sequencing data analysis

N.B. Pairwise sequence alignment is arguably the most fundamental operation of bioinformatics!

Outline for today

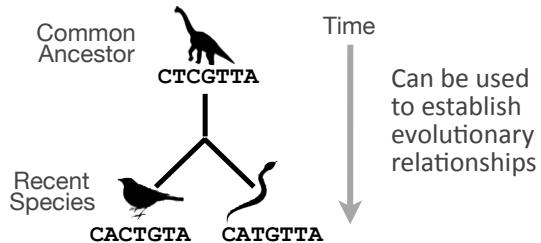
- Alignment basics
 - Why compare biological sequences?
- Homologue detection
 - Orthologs, paralogs, similarity and identity
 - Sequence changes during evolution
 - Alignment view: matches, mismatches and gaps
- Pairwise sequence alignment methods
 - Brute force alignment
 - Dot matrices
 - Dynamic programming (global vs local alignment)
- Rapid heuristic approaches
 - BLAST
- Practical database searching
 - PSI-BLAST and HMM approaches

14

Sequence comparison is most informative when it detects **homologs**

Homologs are sequences that have common origins i.e. they share a **common ancestor**

- They may or may not have common activity



15

Key terms

When we talk about related sequences we use specific terminology.

Homologous sequences may be either:

– **Orthologs or Paralogs**

(Note. these are all or nothing relationships!)

Any pair of sequences may share a certain level of:

– **Identity and/or Similarity**

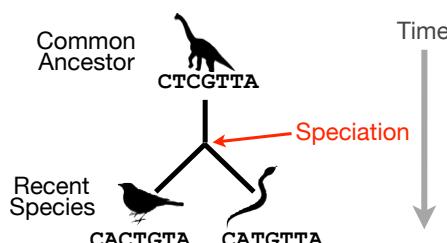
(Note. if these metrics are above a certain level we often infer homology)

16

Orthologs tend to have similar function

Orthologs: are homologs produced by speciation that have diverged due to divergence of the organisms they are associated with.

- Ortho = [greek: straight] ... implies direct descent

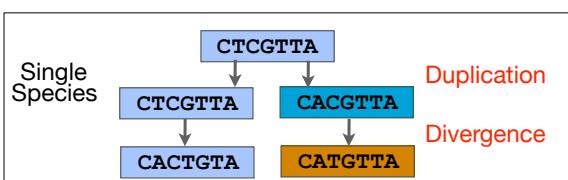


17

Paralogs tend to have slightly different functions

Paralogs: are homologs produced by **gene duplication**. They represent genes derived from a common ancestral gene that *duplicated within an organism* and then subsequently *diverged by accumulated mutation*.

– Para = [greek: along side of]



18

Orthologs vs Paralogs

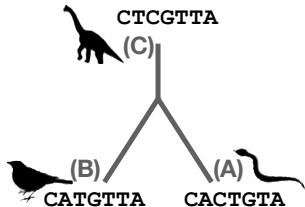
- In practice, determining ortholog vs paralog can be a complex problem:
 - gene loss after duplication,
 - lack of knowledge of evolutionary history,
 - weak similarity because of evolutionary distance
- Homology does not necessarily imply exact same function
 - may have similar function at very crude level but play a different physiological role

19

Sequence changes during evolution

There are three major types of sequence change that can occur during evolution.

- Mutations/Substitutions
- Deletions
- Insertions

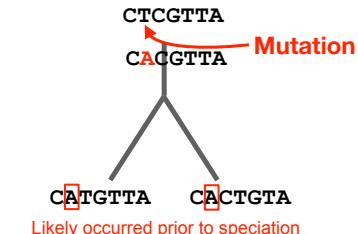


20

Mutations, deletions and insertions

There are three major types of sequence change that can occur during evolution.

- **Mutations/Substitutions** $\text{CTCGTTA} \rightarrow \text{CAGTTA}$
- Deletions
- Insertions

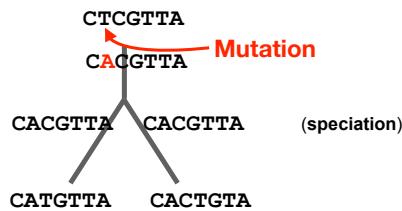


21

Mutations, deletions and insertions

There are three major types of sequence change that can occur during evolution.

- Mutations/Substitutions $\text{CTCGTTA} \rightarrow \text{CAGTTA}$
- Deletions
- Insertions

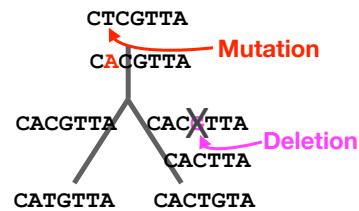


22

Mutations, deletions and insertions

There are three major types of sequence change that can occur during evolution.

- Mutations/Substitutions $\text{CTCGTTA} \rightarrow \text{CAGTTA}$
- Deletions $\text{CACGTTA} \rightarrow \text{CACTTA}$
- Insertions

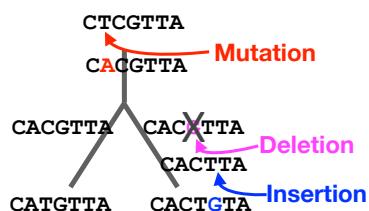


23

Mutations, deletions and insertions

There are three major types of sequence change that can occur during evolution.

- Mutations/Substitutions $\text{CTCGTTA} \rightarrow \text{CAGTTA}$
- Deletions $\text{CACGTTA} \rightarrow \text{CACTTA}$
- Insertions $\text{CACTTA} \rightarrow \text{CACTGTA}$

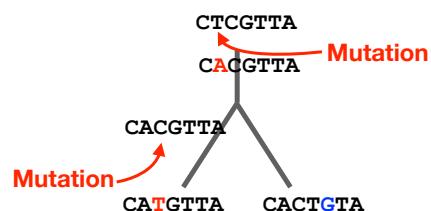


24

Mutations, deletions and insertions

There are three major types of sequence change that can occur during evolution.

- **Mutations/Substitutions** $\text{CTCGTTA} \rightarrow \text{CAGTTA}$
- Deletions $\text{CACGTTA} \rightarrow \text{CATGTTA}$
- Insertions

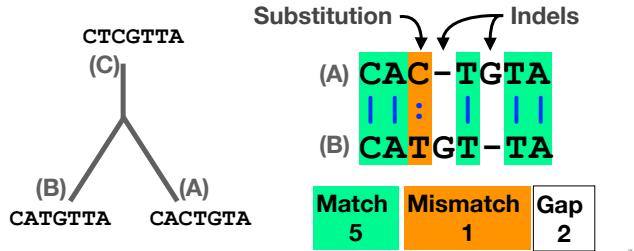


25

Alignment view

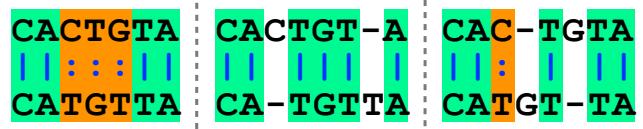
Alignments are great tools to visualize sequence similarity and evolutionary changes in homologous sequences.

- **Mismatches** represent mutations/substitutions
- **Gaps** represent insertions and deletions (indels)



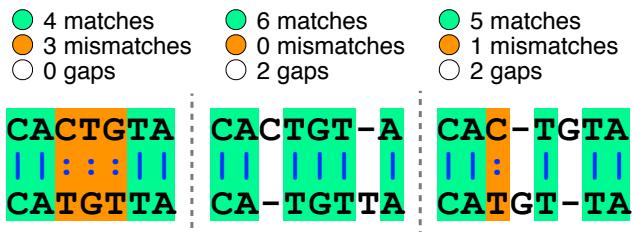
Alternative alignments

- Unfortunately, finding the correct alignment is difficult if we do not know the evolutionary history of the two sequences
 - There are many possible alignments
 - Which alignment is best?



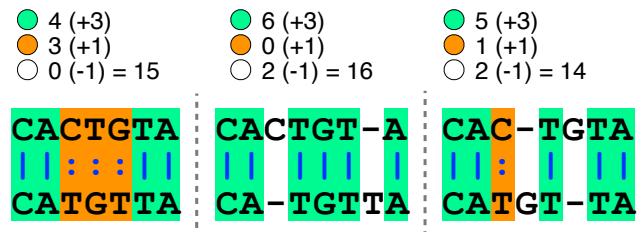
Alternative alignments

- One way to judge alignments is to compare their number of matches, insertions, deletions and mutations



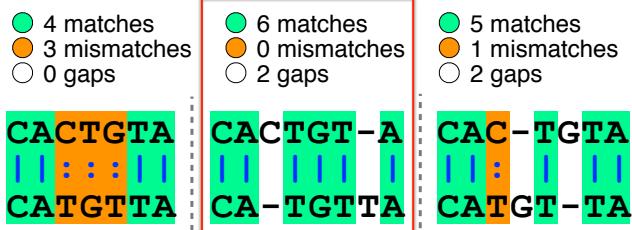
Scoring alignments

- We can assign a score for each match (+3), mismatch (+1) and indel (-1) to identify the **optimal alignment for this scoring scheme**



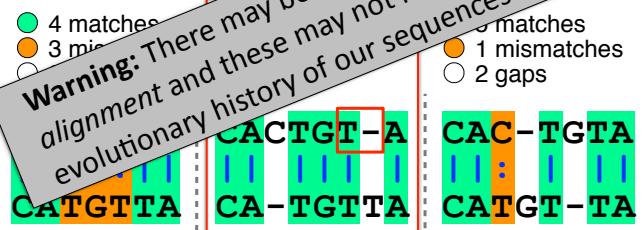
Optimal alignments

- Biologists often prefer **parsimonious alignments**, where the number of postulated sequence changes is minimized.



Optimal alignments

- Biologists often prefer **parsimonious alignments**, where the number of postulated sequence changes is minimized.



Side note: sequence *identity* and *similarity*

- Two commonly quoted metrics for pairs of aligned sequences.
 - Sequence identity:** typically quotes the percent of identical characters in the aligned region of two sequences
 - Sequence similarity:** typically the score resulting from optimal pair-wise alignment (**note dependence on parameters used: i.e. scoring scheme**)
- N.B. In contrast, **homology is an all or nothing relationship, you can not have a percent homology!**

32

Side note: sequence identity and similarity

- High sequence similarity is frequently used as an indicator of homology
 - Use to find genes and/or proteins with potentially similar or identical function
 - Can query a database of sequences by performing a series of pair-wise alignments
- Knowledge of the difference between sequences can also yield valuable functional and mechanistic insights
 - A gene from a normal and an affected subject – possible cause of a heritable disease
 - Similar proteins with different substrate specificities – what amino acid changes might be responsible for this?

33

Outline for today

- Alignment basics
 - Why compare biological sequences?
- Homologue detection
 - Orthologs, paralogs, similarity and identity
 - Sequence changes during evolution
 - Alignment view: matches, mismatches and gaps
- Pairwise sequence alignment methods
 - Brute force alignment
 - Dot matrices
 - Dynamic programming (global vs local alignment)
- Rapid heuristic approaches
 - BLAST
- Practical database searching
 - PSI-BLAST and HMM approaches

34

Outline for today

- Alignment basics
 - Why compare biological sequences?
- Homologue detection
 - Orthologs, paralogs, similarity and identity
 - Sequence changes during evolution
 - Alignment view: matches, mismatches and gaps
- Pairwise sequence alignment methods
 - How do we compute the optimal alignment between two sequences?
(global vs local alignment)
- Rapid heuristic approaches
 - BLAST
- Practical database searching
 - PSI-BLAST

35

Pair-wise Sequence Alignment

- Objective:** arrange two sequences in such a fashion that pairs of matching characters between the two sequences are maximized
 - Match does not have to be identity, can be defined by a function that ranks or scores the characters being compared (often termed a **substitution matrix**)
 - Ungapped alignment example – bars indicate matching characters

Seq1: GTAATCTG-
 | | | | |
Seq2: -TAAGCTGA

36

Simplest case – brute force alignments

- In the simplest case we can simply slide one sequence across the other and count matching characters for each possible alignment
 - Chose a scoring scheme and do not allow internal gaps within sequences
 - Algorithmic complexity is linear
 $N + M$ alignments to consider
(where N and M are the length of each sequence)

37

GTAATCTG
TTAAGCTGA

GTAATCTG
| |
TTAAGCTGA

GTAATCTG
|
TTAAGCTGA

GTAATCTG
|| | |
TTAAGCTGA

GTAATCTG
|
TTAAGCTGA

GTAATCTG
|| |
TTAAGCTGA

GTAATCTG
TTAAGCTGA

GTAATCTG
|
TTAAGCTGA

GTAATCTG
| |
TTAAGCTGA

GTAATCTG
| | |
TTAAGCTGA

GTAATCTG
TTAAGCTGA

GTAATCTG
|
TTAAGCTGA

GTAATCTG

TTAAGCTGA

Etc...

Slide from Jeffery de Wet

Gaps make the brute force method unusable for all but the shortest sequences

- Pairs of related sequences often have insertions or deletions relative to one-another, we therefore require **gapped pair-wise alignment**
 - Need to generate all the possible gap lengths and combinations of gaps at all possible positions in both sequences
 - For two sequences of equal length, the formula is:

$$\binom{2N}{N} = \frac{(2N)!}{(N!)^2} \cong \frac{2^{2N}}{\sqrt{\pi N}}$$

N = 10: 184756
N = 50: ~1.00E29
N = 250: ~1.17E149

Slide from Jeffery de Wet

Three general solutions to the alignment problem

- The **dot plot or dot matrix** approach
 - A simple graphical method for pair-wise alignment
 - No scoring, so difficult to compare alternative alignments
 - Can give visual clues to sequence structure but requires human interaction
- Dynamic programming** algorithms
 - Provides Optimal solutions (but not necessarily unique solutions)
- Heuristic **word** or **k-tuple** approaches
 - Much faster (e.g. **BLAST** and **FASTA**)
 - Widely used for database searches
 - May miss some pairs with low similarity

40

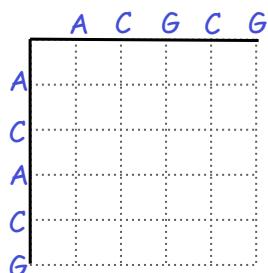
Three general solutions to the alignment problem

- The **dot plot or dot matrix** approach
 - A simple graphical method for pair-wise alignment
 - No scoring, so difficult to compare alternative alignments
 - Can give visual clues to sequence structure but requires human interaction
- Dynamic programming** algorithms
 - Provides Optimal solutions (but not necessarily unique solutions)
- Heuristic **word** or **k-tuple** approaches
 - Much faster (e.g. **BLAST** and **FASTA**)
 - Widely used for database searches
 - May miss some pairs with low similarity

41

Dot plots: simple graphical approach

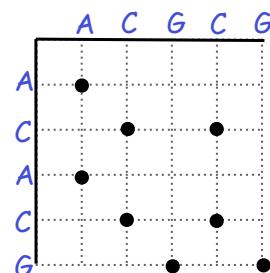
- Place one sequence on the vertical axis of a 2D grid (or matrix) and the other on the horizontal



42

Dot plots: simple graphical approach

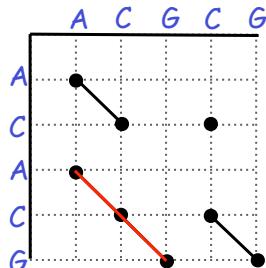
- Now simply put dots where the horizontal and vertical sequence values match



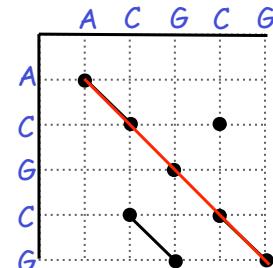
43

Dot plots: simple graphical approach

- Diagonal runs of dots indicate matched segments of sequence



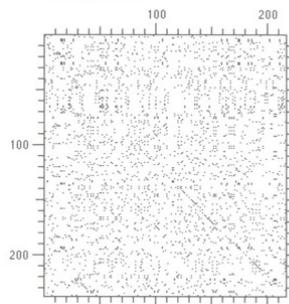
44



45

Dot plots: simple graphical approach

- Dot matrices for long sequences can be noisy



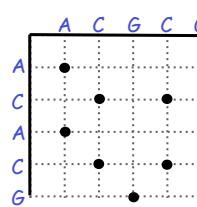
46

Dot plots: window size and match stringency

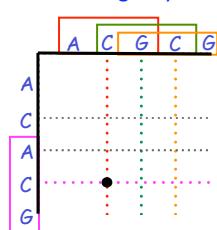
Solution: use a window and a threshold

- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.

- You have to choose window size and stringency



Filter
Window = 3
Stringency = 3



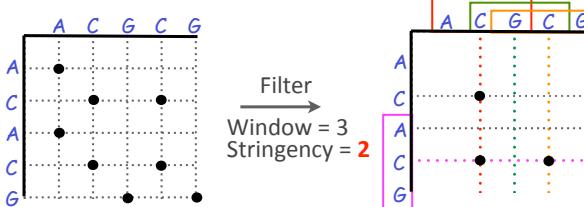
47

Dot plots: window size and match stringency

Solution: use a window and a threshold

- compare character by character within a window
- require certain fraction of matches within window in order to display it with a dot.

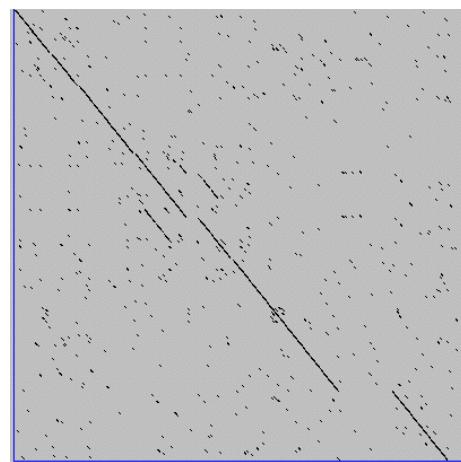
- You have to choose window size and stringency



Filter
Window = 3
Stringency = 2

48

Window size = 5 bases



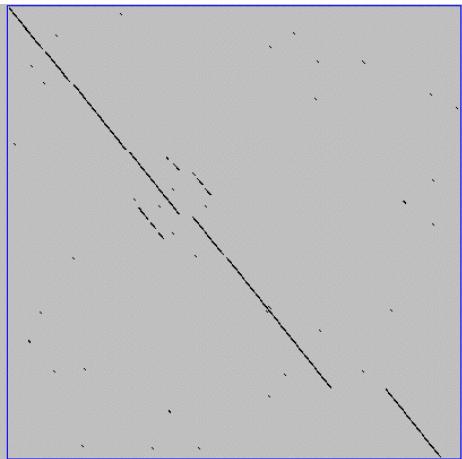
A dot plot simply puts a dot where two sequences match. In this example, dots are placed in the plot if 5 bases in a row match perfectly. Requiring a 5 base perfect match is a **heuristic** – only look at regions that have a certain degree of identity.

Do you expect evolutionarily related sequences to have more word matches (matches in a row over a certain length) than random or unrelated sequences?

Dot plots: simple graphical approach

- Q. What would the dot matrix of a two identical sequences look like?

Window size = 7 bases



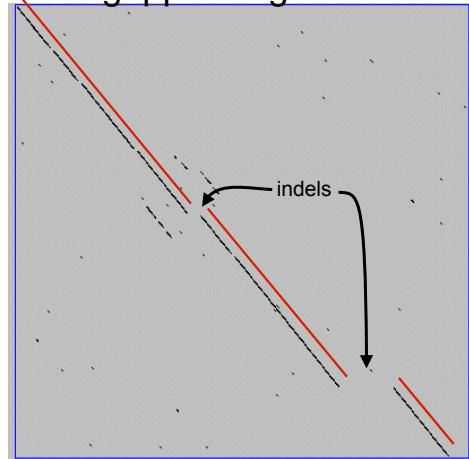
Web site used: <http://www.vivo.colostate.edu/molkit/dnadot/>

This is a dot plot of the same sequence pair. Now 7 bases in a row must match for a dot to be placed. Noise is reduced.

Using windows of a certain length is very similar to using words (kmers) of N characters in the heuristic alignment search tools

Bigger window (kmer) fewer matches to consider

Ungapped alignments

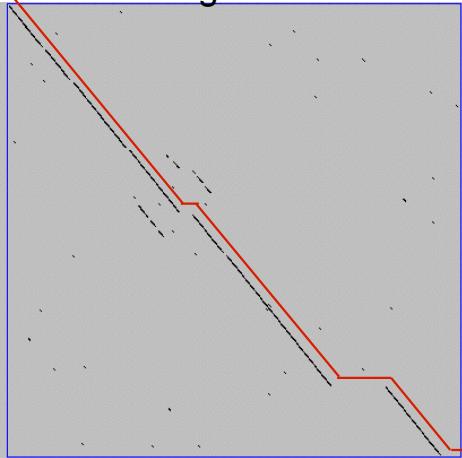


Web site used: <http://www.vivo.colostate.edu/molkit/dnadot/>

Only diagonals can be followed.

Downward or rightward paths represent **insertion or deletions** (gaps in one sequence or the other).

Global alignments



Web site used: <http://www.vivo.colostate.edu/molkit/dnadot/>

Global alignments go from end to end, i.e. from the upper left corner to the lower right corner.

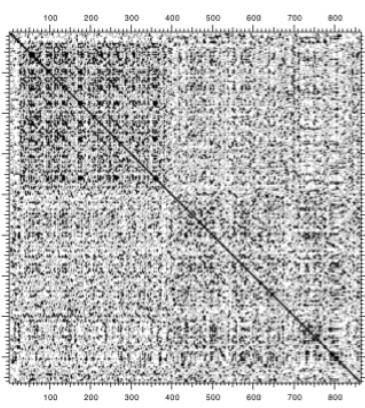
Global alignments do not have good statistical characterization and are **not used for database searches**.

Uses for dot matrices

- Visually assessing the similarity of two protein or two nucleic acid sequences
- Finding local repeat sequences within a larger sequence by comparing a sequence to itself
 - Repeats appear as a set of diagonal runs stacked vertically and/or horizontally

53

Repeats



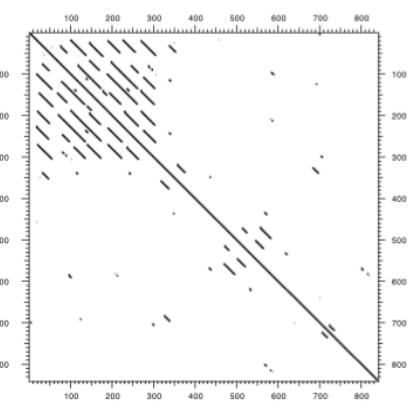
(Figure from Mount, "Bioinformatics sequence and genome analysis")

54

Human LDL receptor protein sequence (Genbank P01130)

W = 1
S = 1

Repeats



(Figure from Mount, "Bioinformatics sequence and genome analysis")

55

Human LDL receptor protein sequence (Genbank P01130)

W = 23
S = 7

Side note: dots can have “weights”

- Some matches can be rewarded more than others, depending on likelihood
- Use PAM or BLOSUM **substitution matrix**
– (more on these later)
- Put a dot only if a minimum total or average weight is achieved
– See chapter 3 in Mount, “*Bioinformatics sequence and genome analysis*”.

56

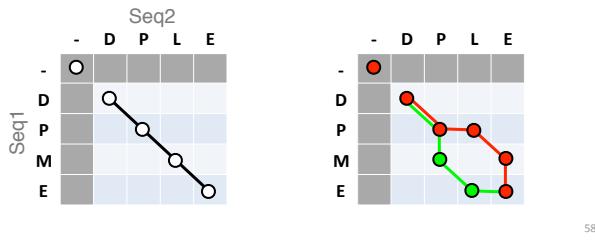
Three general solutions to the alignment problem

- The **dot plot or dot matrix** approach
 - A simple graphical method for pair-wise alignment
 - No scoring, so difficult to compare alternative alignments
 - Can give visual clues to sequence structure but requires human interaction
- Dynamic programming algorithms**
 - Provides Optimal solutions (but not necessarily unique solutions)
- Heuristic word or k-tuple approaches**
 - Much faster (e.g. **BLAST** and **FASTA**)
 - Widely used for database searches
 - May miss some pairs with low similarity

57

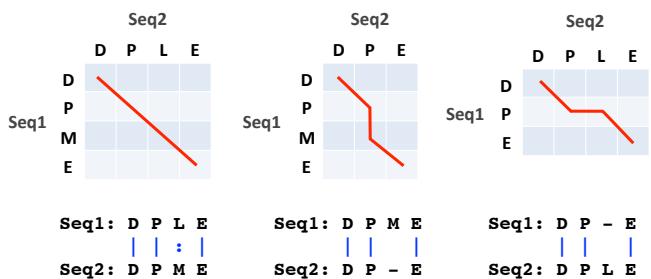
The Dynamic Programming Algorithm

- The dynamic programming algorithm can be thought of an extension to the dot plot approach
 - One sequence is placed down the side of a grid and another across the top
 - Instead of placing a dot in the grid, we **compute a score** for each position
 - Finding the optimal alignment corresponds to finding the path through the grid with the **highest possible score**



58

Different paths represent different alignments

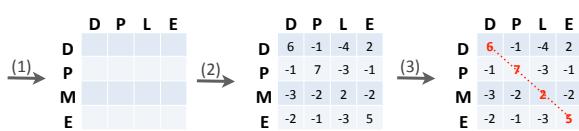


Matches are represented by diagonal paths and indels with horizontal or vertical path segments

59

Algorithm of Needleman and Wunsch

- The Needleman–Wunsch approach to global sequence alignment has three basic steps:
 - (1) setting up a 2D-grid (or **alignment matrix**),
 - (2) **scoring the matrix**, and
 - (3) identifying the **optimal path** through the matrix

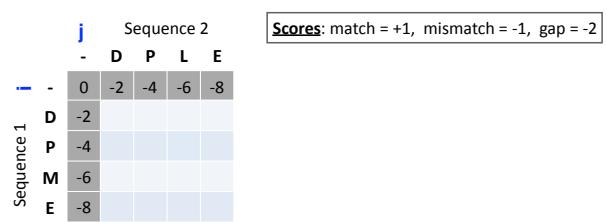


Needleman, S.B. & Wunsch, C.D. (1970) “A general method applicable to the search for similarities in the amino acid sequences of two proteins.” J. Mol. Biol. 48:443-453.

60

Scoring the alignment matrix

- Start by filling in the first row and column – these are all indels (gaps).
 - Each step you take you will add the **gap penalty** to the score ($S_{i,j}$) accumulated in the previous cell



61

Scoring the alignment matrix

- Start by filling in the first row and column – these are all indels (gaps).
 - Each step you take you will add the **gap penalty** to the score ($S_{i,j}$) accumulated in the previous cell

		Sequence 2				
		-	D	P	L	E
Sequence 1	-	0	-2	-4	-6	-8
	D	-2				
	P	-4				
	M	-6				
	E	-8				

Scores: match = +1, mismatch = -1, gap = -2

$S_{i+4} = (-2) + (-2) + (-2) + (-2)$

Seq1: DPME
Seq2: -----

62

Scoring the alignment matrix

- Then go to the empty corner cell (upper left). It has filled in values in up, left and diagonal directions
 - Now can ask which of the three directions gives the highest score?
 - keep track of this score and direction

		Sequence 2				
		-	D	P	L	E
Sequence 1	-	0	-2	-4	-6	-8
	D	-2	?			
	P	-4				
	M	-6				
	E	-8				

Scores: match = +1, mismatch = -1, gap = -2

$j-1$ j

$i-1$ i

$S(i-1, j-1)$ $S(i-1, j)$

$S(i, j-1)$ $S(i, j)$

63

Scoring the alignment matrix

- Then go to the empty corner cell (upper left). It has filled in values in up, left and diagonal directions
 - Now can ask which of the three directions gives the highest score?
 - keep track of this score and direction

		Sequence 2				
		-	D	P	L	E
Sequence 1	-	0	-2	-4	-6	-8
	D	-2	?			
	P	-4				
	M	-6				
	E	-8				

Scores: match = +1, mismatch = -1, gap = -2

$S(i, j) = \text{Max} \begin{cases} S(i-1, j-1) + (\text{mis})\text{match} \\ S(i-1, j) - \text{gap penalty} \\ S(i, j-1) - \text{gap penalty} \end{cases}$

① ② ③

64

Scoring the alignment matrix

- Then go to the empty corner cell (upper left). It has filled in values in up, left and diagonal directions
 - Now can ask which direction gives the highest score
 - keep track of direction and score

		Sequence 2				
		-	D	P	L	E
Sequence 1	-	0	-2	-4	-6	-8
	D	-2	1			
	P	-4				
	M	-6				
	E	-8				

Scores: match = +1, mismatch = -1, gap = -2

① (0)+(+1) = +1 <= (D-D) match!
Alignment D D

② (-2)+(-2) = -4

③ (-2)+(-2) = -4

65

Scoring the alignment matrix

- At each step, the score in the current cell is determined by the scores in the neighboring cells
 - The maximal score and the direction that gave that score is stored (we will use these later to determine the optimal alignment)

		Sequence 2				
		-	D	P	L	E
Sequence 1	-	0	-2	-4	-6	-8
	D	-2	1	1	-1	
	P	-4				
	M	-6				
	E	-8				

Scores: match = +1, mismatch = -1, gap = -2

① (-2)+(-1) = -3 <= (D-P) mismatch!
Alignment D P

② (-4)+(-2) = -6

③ (1)+(-2) = -1

66

Scoring the alignment matrix

- We will continue to store the alignment score ($S_{i,j}$) for all possible alignments in the alignment matrix.

		Sequence 2				
		-	D	P	L	E
Sequence 1	-	0	-2	-4	-6	-8
	D	-2	1	1	-1	3
	P	-4				
	M	-6				
	E	-8				

Scores: match = +1, mismatch = -1, gap = -2

① (-4)+(-1) = -5 <= (D-L) mismatch!
Alignment D-- DPL

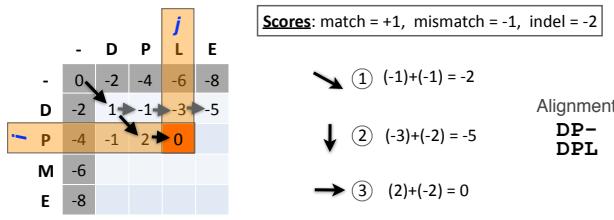
② (-6)+(-2) = -8

③ (-1)+(-2) = -3

67

Scoring the alignment matrix

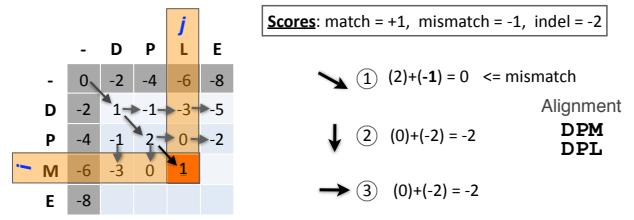
- For the highlighted cell, the corresponding score ($S_{i,j}$) refers to the score of the optimal alignment of the first i characters from sequence1, and the first j characters from sequence2.



68

Scoring the alignment matrix

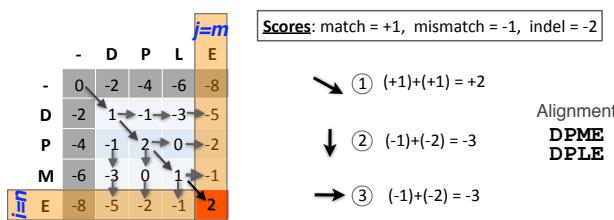
- At each step, the score in the current cell is determined by the scores in the neighboring cells
 - The maximal score and the direction that gave that score is stored



69

Scoring the alignment matrix

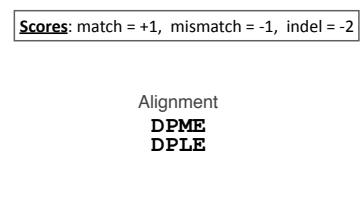
- The score of the best alignment of the entire sequences corresponds to $S_{n,m}$
 - (where n and m are the length of the sequences)



70

Scoring the alignment matrix

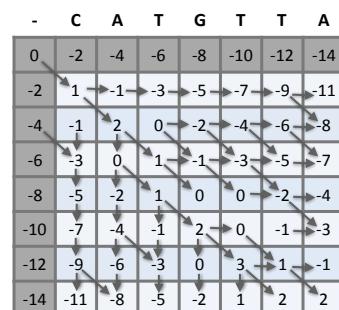
- To find the best alignment, we retrace the arrows starting from the bottom right cell
 - N.B. The optimal alignment score and alignment are dependent on the chosen scoring system



71

Questions:

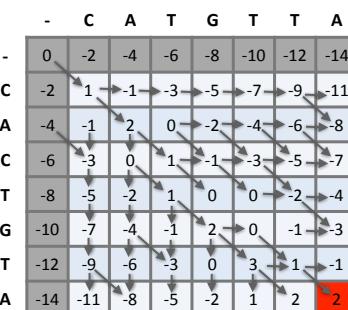
- What is the optimal score for the alignment of these sequences and how do we find the optimal alignment?



72

Questions:

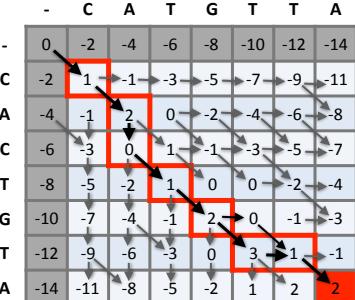
- What is the optimal score for the alignment of these sequences and how do we find the optimal alignment?



73

Questions:

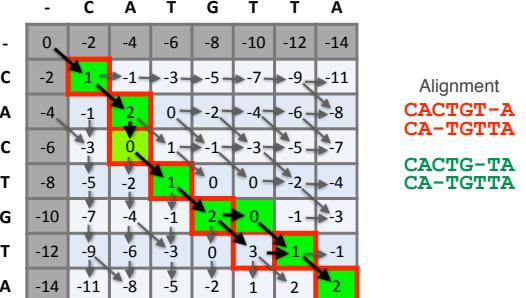
- To find the best alignment we retrace the arrows starting from the bottom right cell



74

More than one alignment possible

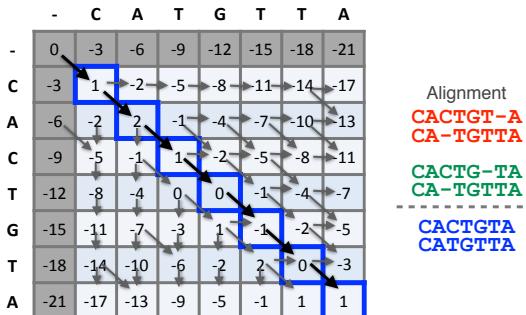
- Sometimes more than one alignment can result in the same optimal score



75

The alignment and score are dependent on the scoring system

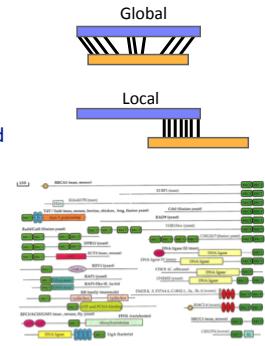
- Here we increase the gap penalty from -2 to -3



76

Global vs local alignments

- Needleman-Wunsch is a **global alignment algorithm**
 - Resulting alignment spans the complete sequences end to end
 - This is appropriate for closely related sequences that are similar in length
- For many practical applications we require **local alignments**
 - Local alignments highlight sub-regions (e.g. protein domains) in the two sequences that align well



77

Local alignment: Definition

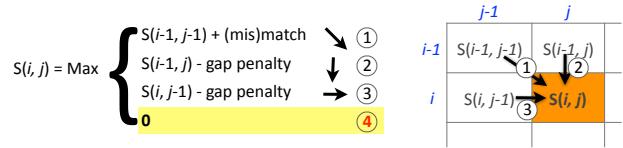
- Smith & Waterman proposed simply that a local alignment of two sequences allow arbitrary-length segments of each sequence to be aligned, with no penalty for the unaligned portions of the sequences. Otherwise, the score for a local alignment is calculated the same way as that for a global alignment

Smith, T.F. & Waterman, M.S. (1981) "Identification of common molecular subsequences." J. Mol. Biol. 147:195-197.

78

The Smith-Waterman algorithm

- Three main modifications to Needleman-Wunsch:
 - Allow a node to start at 0
 - The score for a particular cell cannot be negative
 - if all other score options produce a negative value, then a zero must be inserted in the cell
 - Record the highest-scoring node, and trace back from there



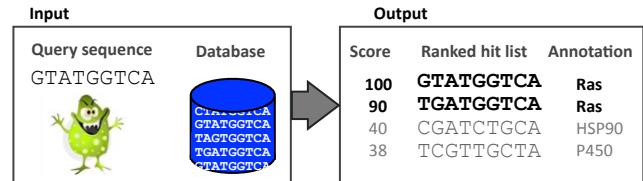
79

		Sequence 1													
		-	C	A	G	C	C	U	C	G	C	U	U	A	G
-		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A		0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A		0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U		0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
		0.0	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7
G		0.0	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C		0.0	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
C		0.0	1.0	0.7	0.0	1.0	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
A		0.0	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
U		0.0	0.0	0.7	1.7	0.3	1.3	1.3	2.3	1.0	0.7	1.7	2.0	1.0	1.0
U		0.0	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
G		0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.0	2.0	1.7	1.3	2.3	2.7
A		0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
C		0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
G		0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G		0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

80

Local alignments can be used for database searching

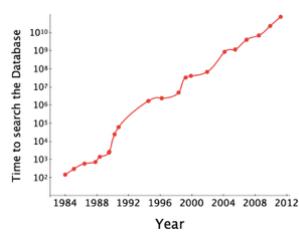
- **Goal:** Given a query sequence (Q) and a sequence database (D), find a list of sequences from D that are most similar to Q
- **Input:** Q, D and scoring scheme
- **Output:** Ranked list of hits



81

The database search problem

- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
 - Time to search with SW is proportional to $m \times n$ (m is length of query, n is length of database), **too slow for large databases!**

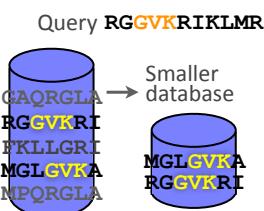


To reduce search time **heuristic algorithms**, such as BLAST, first remove database sequences without a strong local similarity to the query sequence in a quick initial scan.

82

The database search problem

- Due to the rapid growth of sequence databases, search algorithms have to be both efficient and sensitive
 - Time to search with SW is proportional to $m \times n$ (m is length of query, n is length of database), **too slow for large databases!**



To reduce search time **heuristic algorithms**, such as BLAST, first remove database sequences without a strong local similarity to the query sequence in a quick initial scan.

83

Outline for today

- Alignment basics
 - Why compare biological sequences?
- Homologue detection
 - Orthologs, paralogs, similarity and identity
 - Sequence changes during evolution
 - Alignment view: matches, mismatches and gaps
- Pairwise sequence alignment methods
 - Brute force alignment
 - Dot matrices
 - Dynamic programming (global vs local alignment)
- Rapid heuristic approaches
 - BLAST
- Practical database searching
 - PSI-BLAST and HMM approaches

84

Rapid, heuristic versions of Smith-Waterman: **BLAST**

- **BLAST** (Basic Local Alignment Search Tool) is a simplified form of Smith-Waterman (SW) alignment that is popular because it is **fast** and **easily accessible**
 - BLAST is a heuristic approximation to SW - It examines only part of the search space
 - BLAST saves time by restricting the search by scanning database sequences for likely matches before performing more rigorous alignments
 - Sacrifices some sensitivity in exchange for speed
 - In contrast to SW, BLAST is not guaranteed to find optimal alignments

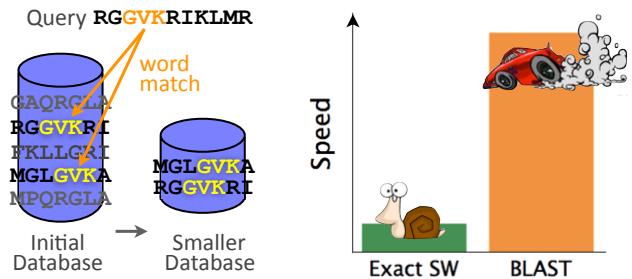
85

Rapid, heuristic versions of Smith–Waterman: BLAST

- BLAST (Basic Local Alignment Search Tool) is a rapid form of Smith-Waterman (SW) alignment
 - BLAST finds regions of local sequence similarity because it is **fast** and **easily parallelizable**
 - The central idea of the BLAST algorithm is to confine attention to sequence pairs that contain an initial **word pair match**
 - BLAST performs the search by scanning the database for sequence pairs that likely matches before performing alignments
 - Altschul et al. (1990) found some sensitivity in exchange for speed
 - In contrast to SW, BLAST is not guaranteed to find optimal alignments
- “The central idea of the BLAST algorithm is to confine attention to sequence pairs that contain an initial word pair match”*
Altschul et al. (1990)

86

- BLAST uses this pre-screening heuristic approximation resulting in an approach that is about 50 times faster than the Smith-Waterman algorithm



87

How BLAST works

- Four basic phases

- Phase 1: compile a list of query word pairs ($w=3$)

RGGVKRI Query sequence
 RGG
 GGV
 GVK
 VKR
 KRI

generate list of $w=3$ words for query

88

RGGVKRI Query sequence

RGG RAG RIG RLG ...
 GGV GAV GTV GCV ...
 GVK GAK GIK GGK ...
 VKR VRR VHR VER ...
 KRI KKI KHI KDI ...

extend list of words similar to query

89

Blast

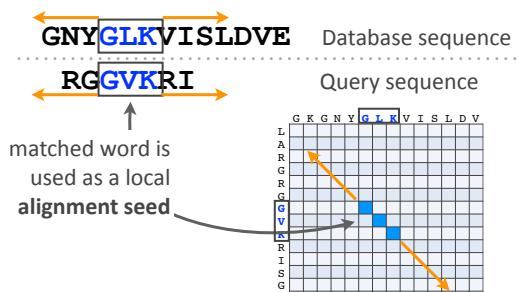
- Phase 3: a database is scanned to find sequence entries that match the compiled word list

GNY GLK VISLDVE Database sequence
 RGGVKRI Query sequence
 RGG RAG RIG RLG ...
 GGV GAV GTV GCV ...
 GVK GLK GIK GGK ...
 VKR VRR VHR VER ...
 KRI KKI KHI KDI ...

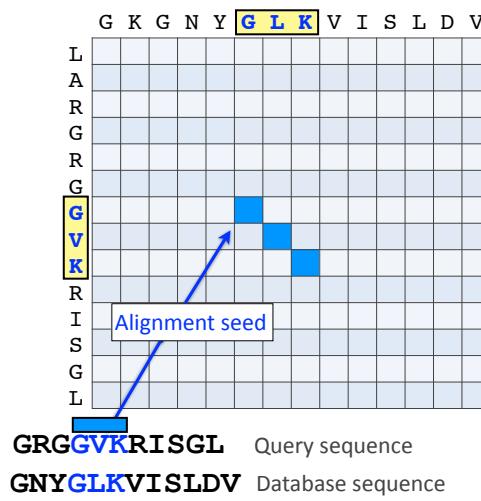
search for perfect matches in the database sequence

90

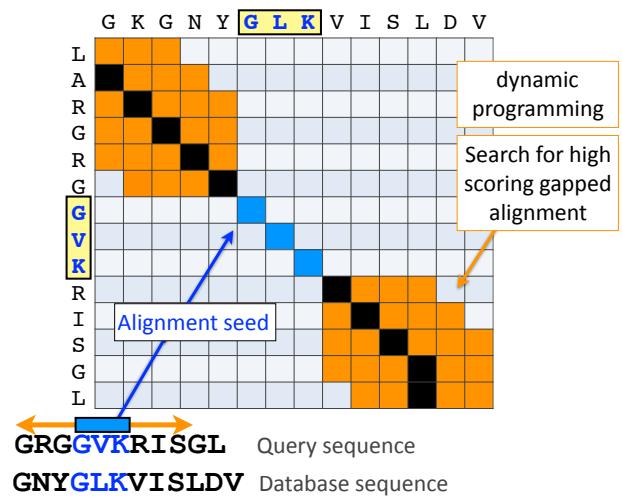
- Phase 4: the initial database hits are extended in both directions using dynamic programming



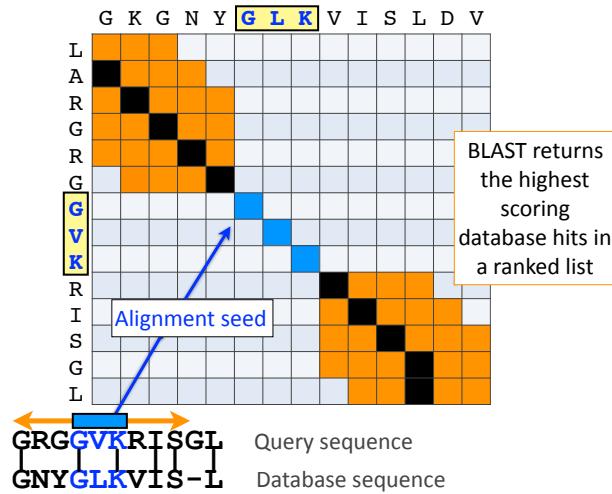
91



92



93



94

BLAST output

- BLAST returns the highest scoring database hits in a ranked list along with details about the target sequence and alignment statistics

Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo]	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1
Kinesin-14 heavy chain [Danio rerio]	595	595	88%	0	78%	XP_00320703
hypothetical protein EGK_18589	48.2	52	40%	0.03	32%	ELK35081.1
mKIAA4102 protein [Mus musculus]	42.7	42.7	38%	3.02	24%	EHH28205.1

95

Statistical significance of results

- An important feature of BLAST is the computation of statistical significance for each hit. This is described by the **E value** (expect value)

Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo]	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1
Kinesin-14 heavy chain [Danio rerio]	595	595	88%	0	78%	XP_00320703
hypothetical protein EGK_18589	48.2	52	40%	0.03	32%	ELK35081.1
mKIAA4102 protein [Mus musculus]	42.7	42.7	38%	3.02	24%	EHH28205.1

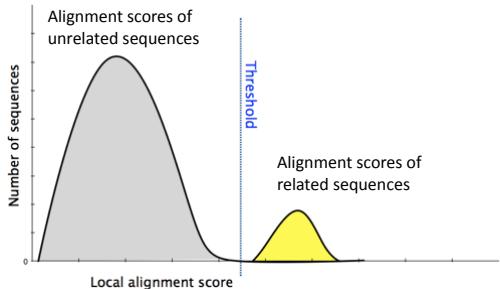
96

BLAST scores and E-values

- The **E value** is the **expected** number of hits that are as good or better than the observed local alignment score (with this score or better) if the query and database are **random** with respect to each other
 - i.e. the number of alignments expected to occur by chance with equivalent or better scores
- Typically, only hits with E value **below** a significance threshold are reported
 - This is equivalent to selecting alignments with score above a certain score threshold

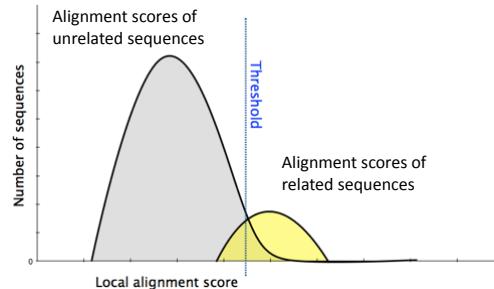
97

- Ideally, a threshold separates all query related sequences (yellow) from all unrelated sequences (gray)



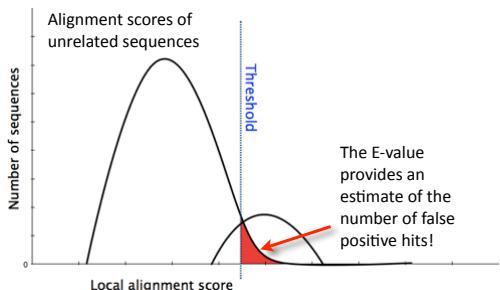
98

- Unfortunately, often both score distributions overlap
 - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



99

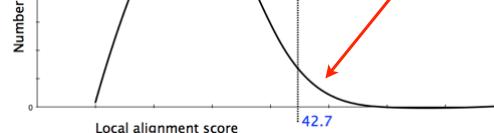
- Unfortunately, often both score distributions overlap
 - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



100

Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo]	677	677	100%	0	100%	NP_004512.1
Kif5b protein [Mus musculus]	676	676	100%	0	98%	AAA20133.1
Kinesin-14 heavy chain [Danio rerio]	595	595	88%	0	78%	XP_00320703
hypothetical protein EGK_18589	42.7	52	40%	0.03	32%	ELK35081.1

A score of 42.7 or better is expected to occur by chance 3 in 100 times (E-value = 0.03)



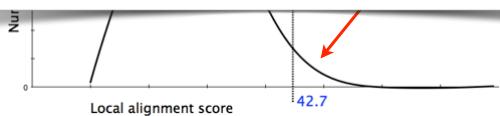
101

Description	Max score	Total score	Query cover	E value	Max ident	Accession
kinesin-1 heavy chain [Homo]	677	677	100%	0	100%	NP_004512.1

In general E values < 0.005 are usually significant.

To find out more about E values see: "The Statistics of Sequence Similarity Scores" available in the help section of the NCBI BLAST site:

<http://www.ncbi.nlm.nih.gov/blast/tutorial/Altschul-1.html>



102

Outline for today

- Alignment basics
 - Why compare biological sequences?
- Homologue detection
 - Orthologs, paralogs, similarity and identity
 - Sequence changes during evolution
 - Alignment view: matches, mismatches and gaps
- Pairwise sequence alignment methods
 - Brute force alignment
 - Dot matrices
 - Dynamic programming (global vs local alignment)
- Rapid heuristic approaches
 - BLAST
- Practical database searching
 - BLAST, PSI-BLAST and HMM approaches

103

Practical database searching with BLAST

The screenshot shows the NCBI BLAST Home Page. It includes a sidebar with species selection (Human, Mouse, Rat, Arabidopsis thaliana) and a main area with search buttons for nucleotide blast, protein blast, blastx, tblastn, and tblastx. A red box highlights the 'Basic BLAST' section.

104

Practical database searching with BLAST

- There are four basic components to a traditional BLAST search
 - (1) Choose the sequence (query)
 - (2) Select the BLAST program
 - (3) Choose the database to search
 - (4) Choose optional parameters
- Then click “BLAST”

105

Step 1: Choose your sequence

- Sequence can be input in FASTA format or as accession number

The screenshot shows the NCBI Protein search results for "hemoglobin subunit beta [Homo sapiens]". It displays the sequence in FASTA format: >gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens] MVLHLPEKSAVITALWGVNVEVGEALGRLLVVYPWTQPRFFESFGDLSTPDAVGNPKVARGKKVLG AFSDQLAHNLNLKGTFATLSELHCKDHLNVDPENFRLCNVLVCAHMFGKEFTPPVQAAAYQRUVVAGVAN ALAHKYN. A red circle highlights the "Display Settings" button, which is set to "FASTA".

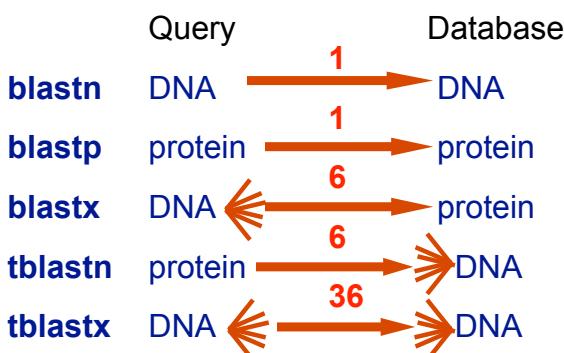
106

Step 2: Choose the BLAST program

The screenshot shows the NCBI Basic BLAST search interface. It lists several BLAST programs: nucleotide blast, protein blast, blastx, tblastn, and tblastx. Each has a brief description and an "Algorithms" link. A red box highlights the "Basic BLAST" section.

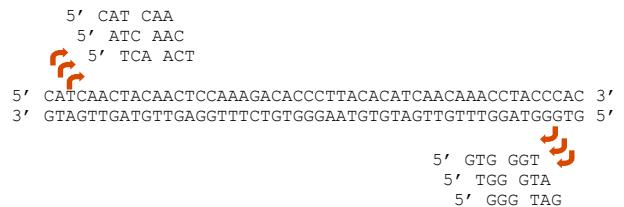
107

Step 2: Choose the BLAST program



108

DNA potentially encodes six proteins



109

Protein BLAST: search protein databases using a protein query

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
From _____ To _____
Query subrange _____
Or, upload file no file selected
Job Title _____
Enter a descriptive title for your BLAST search _____
Align two or more sequences

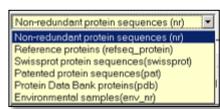
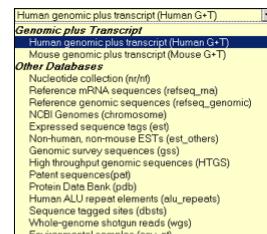
Choose Search Set
Database Non-redundant protein sequences (nr)
Organism Optional
Exclude Optional
Entrez Query Optional
Enter an Entrez query to limit search _____
Program Selection
Algorithm blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PH-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm _____

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
 Show results in a new window
Algorithm parameters

110

Step 3: Choose the database

nr = non-redundant (most general database)
dbest = database of expressed sequence tags
dbsts = database of sequence tag sites
gss = genomic survey sequences



protein databases

nucleotide databases

111

Protein BLAST: search protein databases using a protein query

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
From _____ To _____
Query subrange _____
Or, upload file no file selected
Job Title _____
Enter a descriptive title for your BLAST search _____
Align two or more sequences

Choose Search Set
Database Non-redundant protein sequences (nr)
Organism Optional
Exclude Optional
Entrez Query Optional
Enter an Entrez query to limit search _____
Program Selection
Algorithm blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PH-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm _____

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
 Show results in a new window
Algorithm parameters

112

Step 4a: Select optional search parameters

Algorithm parameters
General Parameters
Max target sequences Select the maximum number of aligned sequences to display _____
Short queries Automatically adjust parameters for short input sequences _____
Expect threshold _____
Word size _____
Max matches in a query range _____
Scoring Parameters
Matrix BLOSUM62 Scoring matrix
Gap Costs Existence: 11 Extension: 1 _____
Compositional adjustments Conditional compositional score matrix adjustment
Filters and Masking
Filter Low complexity regions _____
Mask Mask for lookup table only _____
 Mask lower case letters _____
BLAST Search database Non-redundant protein sequences (nr) using Blastp
 Show results in a new window

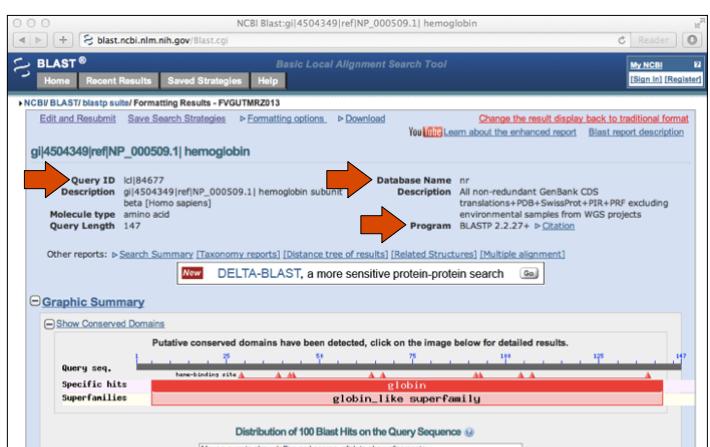
113

Step 4: Optional parameters

- You can...
 - choose the organism to search
 - change the substitution matrix
 - change the expect (E) value
 - change the word size
 - change the output format

114

Results page



Further down the results page...

NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments

Color key for alignment scores

Query	<40	40-60	60-80	80-200	>=200		
1	20	40	60	80	100	120	140

Further down the results page...

NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin

Sequences producing significant alignments:

Select: All None Selected: 0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total cover	E value	Max ident	Accession
hemoglobin beta [synthetic construct]	301	301	1e-103	100%	AAX37051.1
hemoglobin beta [synthetic construct]	301	301	1e-102	100%	AAX29557.1
hemoglobin subunit beta Homo sapiens >ref XP_508242.1 PREDICTED: hemoglobin_s	301	301	1e-102	100%	NP_000509.1
RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin_beta	300	300	1e-102	99%	P02024.2
beta globin chain variant Homo sapiens	299	299	1e-102	99%	AAN84548.1
beta globin Homo sapiens >gb AAZ39781.1 beta globin Homo sapiens >gb AAZ39782	299	299	1e-102	99%	AAZ39780.1
beta-globin Homo sapiens	299	299	1e-102	99%	ACU56984.1
hemoglobin beta chain Homo sapiens	299	299	1e-102	99%	AAD19696.1
Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At	298	298	9e-102	100%	1COH_B
hemoglobin beta subunit variant Homo sapiens >gb AAE88054.1 beta-globin Homo sa	298	298	1e-101	99%	AAE00489.1
Chain B, Human Hemoglobin D Los Angeles: Crystal Structure >PdbID:2YRSID Chain D_H	298	298	9e-101	99%	2YRS_B
Chain B, High-Resolution X-Ray Study Of Deoxy Recombinant Human Hemoglobins Syn	297	297	9e-101	99%	1DXU_B
Chain B, Analysis Of The Crystal Structure, Molecular Modeling And Infrared Spectroscop	297	297	9e-101	99%	1HDB_B

Further down the results page...

NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin

Download GenPept Graphics ▾ Next Match ▾ Previous Descriptions

hemoglobin subunit beta |Homo sapiens|

Sequence ID: ref|NP_000509.1| Length: 147 Number of Matches: 1

► See 84 more title(s)

Range 1 to 147 Download Graphics ▾ Next Match ▾ Previous Descriptions

Score Expect Method Identities Positives Gaps

301 bits(770) 1e-102 Compositional matrix adjust. 147/147(100%) 147/147(100%) 0/147(0%)

Query 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

Sbjct 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

Query 61 VRAAGKVKLGAfSOCLAHLNLNKGTTATSELHDXCLJUVDFNRLFQVYCVTAHHIFC 120

Sbjct 61 VRAAGKVKLGAfSOCLAHLNLNKGTTATSELHDXCLJUVDFNRLFQVYCVTAHHIFC 120

Query 121 KEETPPVQAAYQKVYAGVANALAHRYH 147

Sbjct 121 KEETPPVQAAYQKVYAGVANALAHRYH 147

Range 1 to 147 Download Graphics ▾ Next Match ▾ Previous Descriptions

RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain

Sequence ID: sp|P02024.2|HBB_GORG| Length: 147 Number of Matches: 1

Range 1 to 147 Download Graphics ▾ Next Match ▾ Previous Descriptions

Score Expect Method Identities Positives Gaps

300 bits(767) 4e-102 Compositional matrix adjust. 146/147(99%) 147/147(100%) 0/147(0%)

Different output formats are available

NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin

BLAST® Basic Local Alignment Search Tool My NCBI [Sign In] [Register]

Home Recent Results Saved Strategies Help

► NCBIBLAST /blast suite/Formatting Results - FVGUTMP2013

Edit and Resubmit Save Search Strategies ▾ Formatting options ▾ Download Change the result display back You Tube Learn about the enhanced report Bias Reform

Formatting options

Show Alignment as HTML Old View Reset form to defaults

Alignment View Query-anchored with letters for identities

Display Graphical Overview Sequence Retrieval NCBI-gi

Masking Character: Lower Case Color: Gray

Limit results Descriptions: 50 Graphical overview: 50 Alignments: 50

Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown. Enter organism name or id--completions will be suggested Exclude

Enter query:

Expect Min: Expect Max:

Percent Identity Min: Percent Identity Max:

Format for PSI-BLAST with inclusion threshold:

gl|4504349|ref|NP_000509.1| hemoglobin

E.g. Query anchored alignments

NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin

Download GenPept Graphics ▾ Next Match ▾ Previous Descriptions

Query 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

AAA37051 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

AKAZ3557 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

NP_000509 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

P02024 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

AAN84548 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

AAZ39780 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

ACU56984 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

AAE00489 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

I1COB_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

AAFO0489 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

ZYRS_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

IDXU_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

I1HDB_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

I1DPC_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

JKWC_C 2 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

AAE08978 2 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

INOP_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

I1KIK_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

AAE11320 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

AAE0822173 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

IY85_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

IY80_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

I1O10_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

CAA23759 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

IYE2_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

I1ZP_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

I1ABY_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

I1CHY_B 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 59

... and alignments with dots for identities

NCBI Blast:gi|4504349|ref|NP_000509.1| hemoglobin

Download GenPept Graphics ▾ Next Match ▾ Previous Descriptions

Query 1 MVHLTPEEKSAVTALGKVNVDVGGALGRLLVVYPWTQRFESFGDLSTPDAVGNPK 60

AAX37051 1 60

NP_000509 1 60

P02024 1 60

AAZ39780 1 60

ACU56984 1 60

AAE00489 1 60

I1COB_B 1 60

ZYRS_B 1 60

IDXU_B 1 59

I1HDB_B 1 59

JKWC_C 2 59

AAE08978 2 59

INOP_B 1 59

I1KIK_B 1 59

AAE11320 1 60

AAE0822173 1 60

IY85_B 1 59

IY80_B 1 59

I1O10_B 1 59

CAA23759 1 60

IYE2_B 1 59

I1ZP_B 1 59

I1ABY_B 1 59

I1CHY_B 1 59

Common problems

- Selecting the wrong version of BLAST
 - Selecting the wrong database
 - Too many hits returned
 - Too few hits returned
 - Unclear about the significance of a particular result - are these sequences homologous?

122

How to handle too many results

- Focus on the question you are trying to answer
 - select “refseq” database to eliminate redundant matches from “nr”
 - Limit hits by organism
 - Use just a portion of the query sequence, when appropriate
 - Adjust the expect value; lowering E will reduce the number of matches returned

123

How to handle too few results

- Many genes and proteins have no significant database matches
 - remove Entrez limits
 - raise E-value threshold
 - search different databases
 - try scoring matrices with lower BLOSUM values (or higher PAM values)
 - use a search algorithm that is more sensitive than BLAST (*e.g.* PSI-BLAST or HMMer)

124

Side note: Scoring matrices

- A substitution matrix contains values proportional to the probability that amino acid i mutates into amino acid j for all pairs of amino acids
 - Substitution matrices are constructed by assembling a large and diverse sample of verified pairwise alignments (or multiple sequence alignments) of amino acids.
 - Substitution matrices should reflect the probabilities of mutations occurring through a period of evolution
 - The two major types of substitution matrices are **PAM** and **BLOSUM**

125

BLOSUM62 is the default BLASTp scoring matrix

- BLOSUM matrices are based on short, ungapped blocks of conserved amino acid sequences from multiple alignments
 - members of a block that have a most X percent sequence identity to each other are used to generate a BLOSUMX matrix
 - For example, using a cutoff of 62% identity will generate the BLOSUM62 matrix
 - PAM matrices are similar but built from multiple alignments where amino acid substitutions are at rate of 1% (PAM 1)
 - Matrix multiplication is used generate higher PAM matrices
 - PAM3 = (PAM1 x PAM1 x PAM1) etc...

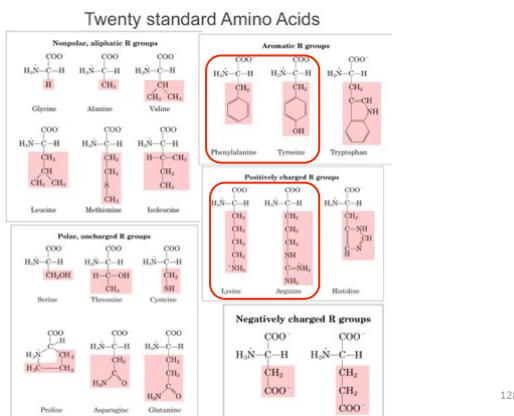
126

By default BLASTp Match scores come from the BLOSUM62 matrix

Note. Some amino acid mismatches have positive scores – highlighted in red

127

Protein scoring matrices reflect the properties of amino acids



Two problems standard BLAST cannot solve

- Use human beta globin as a query against human RefSeq proteins, and blastp does not “find” human myoglobin
 - This is because the two proteins are too distantly related
- **PSI-BLAST** at NCBI as well as hidden Markov models (HMMs) easily solve this problem
- How can we search using 10,000 base pairs as a query, or even millions of base pairs?
 - Many BLAST-like tools for genomic DNA are now available such as Megablast

129

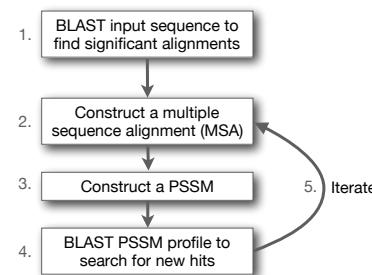
PSI-BLAST: Position specific iterated BLAST

- The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query
 - PSI-BLAST constructs a multiple sequence alignment from the results of a first round BLAST search and then creates a “profile” or specialized **position-specific scoring matrix (PSSM)** for subsequent search rounds

130

PSI-BLAST: Position-Specific Iterated BLAST

- Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



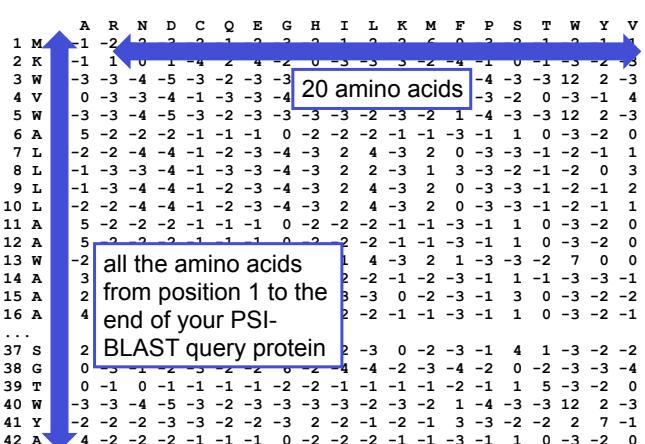
131

Inspect the blastp output to identify empirical “rules” regarding amino acids tolerated at each position

730496	66	FTVDENGQMSATAKGRVRVLFNNNDVCAHDIGSFTDTEPAKFHKYWGVASFLQKGNDHH	125
200679	63	FSVDEKGHMSATAKGRVRVLLSNWEVCADMWTDTEDPAFKHMKYWGVASFLQKGNDHH	122
206589	34	FSVDEKGHMSATAKGRVRVLLSNWEVCADMWTDTEDPAFKHMKYWGVASFLQKGNDHH	93
2136812	2	MSATAKGRVRVLLNNWDVCADMWTDTEDPAFKHMKYWGVASFLQKGNDHH	53
132408	65	FKIEDNGKTTATAKGRVRVILDKLELCKANWVGTFTEDTPDPAFKHMKYWGASFLQKGNDHH	124
267584	44	FSVDESGKVATATAHGRVIIILNNWEMCANMFGTFEDTPDPAFKHMKYWGASFLQKGNDHH	103
267585	44	FSVDGSGKVATATAQGRVIIILNNWEMCANMFGTFEDTPDPAFKHMKYWGAAAYLQSQNDHH	103
877608	63	FTIHEDGAMATAKGRVIIILNNWEMCADMHATFETTPDPAFKHMKYWGAAAYLQTQNDHH	122
6687453	60	FKVEEDGTMATAIAGRVIILNNWEMCANMFGTFEDTPDPAFKHMKYWGAAAYLQTQYDH	119
10697027	81	FKVQEDGTMATAATAGRVIILNNWEMCANMFGTFEDTEEPARFKHMKYWGAAAYLQTQYDH	140
13645517	1	MVGTFTDTEDPAFKHMKYWGVASFLQKGNDHH	32
13925316	38	FSVDESGKMTATAQGRVIIILNNWEMCANMFGTFEDTPDPAFKHMKYWGAAAYLQSQNDHH	97
131649	65	YTVVEDGTMASSKGRVKLFGWVICADMAAQYDPTTPAKHMYNTVQLGLASYLSSQGDNT	126

R,I,K C D,E,T K,R,T N,L,Y,G

132



133

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1 M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2 K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4 V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6 A	5	2	-2	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	
7 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	
8 L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	2	0	
9 L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	0	-3	-2	0	3	1	
10 L	-2	-2	-2	-4	-4					0	-3	-3	-1	-2	-1	1				
11 A	5	2	-2	-2	-2					0	-3	-3	-1	-2	-1	1				
12 A	5	2	-2	-2	-2					-3	-1	1	0	-3	-2	0				
13 W	-2	-2	-3	-4	-4					1	-3	-3	-2	7	0	0				
14 A	3	-2	-1	-2	-2					-3	-1	1	-1	-3	-3	-1				
15 A	2	1	0	-1	-2					-3	-1	3	0	-3	-2	-2				
16 A	4	-2	-2	-2	-2					-3	-1	1	0	-3	-2	-1				
...																				
37 S	2	-1	0	-1						-3	-1	4	1	-3	-2	-2				
38 G	0	-3	-1	-2						-4	-2	0	-2	-3	-3	-4				
39 T	0	-1	0	-1						-2	-1	1	5	-3	-2	0				
40 W	-3	-3	-3	-4	-5					1	-4	-3	-3	12	2	-3				
41 Y	-2	-2	-2	-2	-3					3	-3	-2	-2	2	7	-1				
42 A	4	-2	-2	-2	-2					-3	-1	1	0	-3	-2	0				

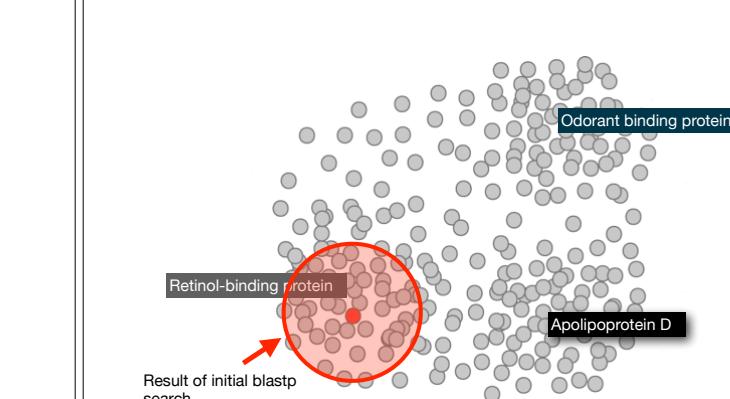
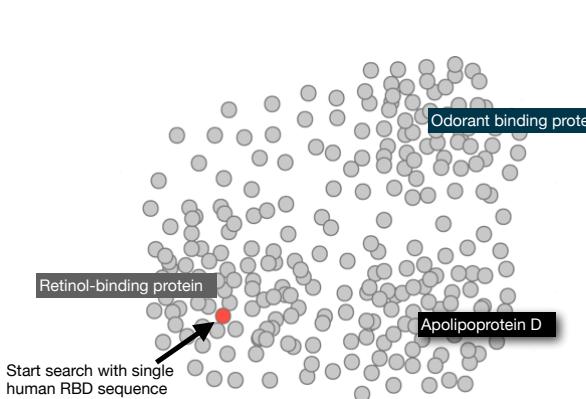
note that a given amino acid (such as alanine) in your query protein can receive different scores for matching alanine—depending on the position in the protein

134

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1 M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	
2 K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	
3 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	12	2	-3	
4 V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	
5 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	12	2	-3	
6 A	5	2	-2	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	
7 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	1	
8 L	-1	-3	-3	-4	-1	-3	-3	-4	-4	3	2	2	-3	1	3	-3	-2	-1	0	
9 L	-1	-3	-4	-4	-1	-3	-3	-4	-4	1	-3	-4	-4	0	-3	-3	-1	-2	1	
10 L	-2	-2	-4	-4	-1	-3	-3	-4	-4	1	-3	-4	-4	0	-3	-3	-1	-2	1	
11 A	5	2	-2	-2	-2					-3	-1	1	0	-3	-2	0				
12 A	5	2	-2	-2	-2					-3	-1	1	0	-3	-2	0				
13 W	-2	-3	-4	-4						1	-3	-3	-2	7	0	0				
14 A	3	-2	-1	-2						-3	-1	1	-1	-3	-3	-1				
15 A	2	1	0	-1						-3	-1	3	0	-3	-2	-2				
16 A	4	-2	-2	-2						-3	-1	1	0	-3	-2	-1				
...																				
37 S	2	-1	0	-1						-3	-1	4	1	-3	-2	-2				
38 G	0	-3	-1	-2						-4	-2	0	-2	-3	-3	-4				
39 T	0	-1	0	-1						-2	-1	1	5	-3	-2	0				
40 W	-3	-3	-3	-4	-5					1	-4	-3	-3	12	2	-3				
41 Y	-2	-2	-2	-2	-3					3	-3	-2	-2	2	7	-1				
42 A	4	-2	-2	-2	-2					-3	-1	1	0	-3	-2	0				

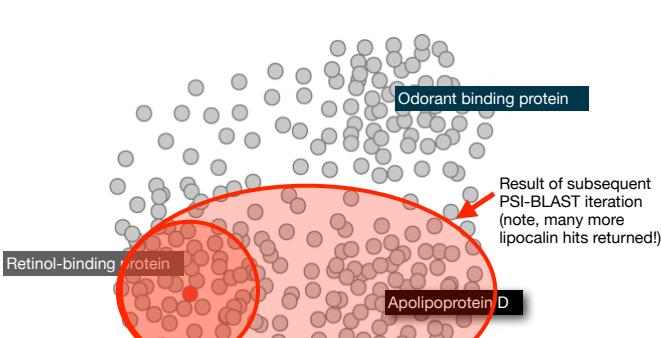
The PSI-BLAST PSSM is essentially a query customized scoring matrix that is more sensitive than PAM or BLOSUM.

135

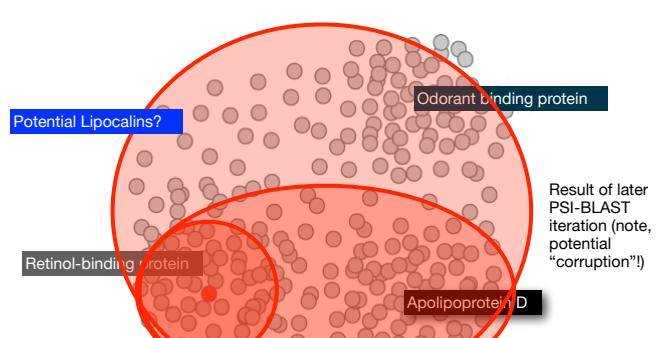


136

137



138



139

PSI-BLAST returns dramatically more hits

- The search process is continued iteratively, typically about five times, and at each step a new PSSM is built
 - You must decide how many iterations to perform and which sequences to include!
 - You can stop the search process at any point - typically whenever few new results are returned or when no new "sensible" results are found

Iteration	Hits with E < 0.005	Hits with E > 0.005
1	34	61
2	314	79
3	416	57
4	432	50
5	432	50

Human retinol-binding protein 4 (RBP4; P02753) was used as a query in a PSI-BLAST search of the RefSeq database.

140

The screenshot shows the HMMER3 website. At the top, there's a navigation bar with links for Home, Search, Results, Software, Help, and About. The main content area features a banner for "HMMER3: a new generation of sequence homology search software". Below the banner, there's a section for "Download HMMER" with a "Search" button. The central part of the page contains a search form with fields for "Query sequence" and "Database", along with "Score", "Taxonomy", "Domain", and "Download" buttons. A "Pfam Domains" section is also visible. At the bottom, there's a footer with links for "Comments or questions on the site? Send a mail to hmmr@janelia.hhmi.org" and "Howard Hughes Medical Institute".

This screenshot shows the HMMER search interface. The top navigation bar includes links for Home, Search, Results, Software, Help, and About. Below the navigation is a search form for "protein sequence vs protein sequence database". It has a "Paste in your sequence or use the example" input field containing a sequence example, and a "Submit" button. The bottom of the page includes a "Comments or questions on the site? Send a mail to hmmr@janelia.hhmi.org" link and a "Howard Hughes Medical Institute" footer.

This screenshot shows the HMMER search results for a Pfam Domains search. The top navigation bar is identical to the previous screenshot. The main content area displays a "Pfam Domains" search results page with a "Score" tab selected. It shows a "Distribution of Significant Hits" chart and a table of "Query Matches (5100)". The table includes columns for Target, Description, Species, E-value, and Alignments. The results list various kinesin family members from different species like Homo sapiens, Macaca mulatta, and Pan troglodytes, along with synthetic constructs and other sequences.

This screenshot shows the HMMER search results for a domain architecture search. The top navigation bar is identical. The main content area displays a "Query" search results page with a "Domain" tab selected. It shows a "Domain Architectures" section with several entries for Kinesin, FHA, KIF1B, DUF3694, PH, HHH_3, CH, and an exact match for Kinesin, HHH_3. Each entry includes a diagram of the domain architecture and a "View Scores" link.

This screenshot shows the HMMER search results for a domain architecture search with the "Download" tab selected. It displays a "Format" section with options for FASTA, Full length FASTA, Aligned FASTA, STOCKHOLM, Text, XML, and JSON. Each option has a brief description and a corresponding file icon. At the bottom, there are "Download" and "Reset" buttons.

Summary

- Alignment basics
 - Why compare biological sequences?
- Homologue detection
 - Orthologs, paralogs, similarity and identity
 - Sequence changes during evolution
 - Alignment view: matches, mismatches and gaps
- Pairwise sequence alignment methods
 - Brute force alignment
 - Dot matrices
 - Dynamic programming
(global vs local alignment)
- Rapid heuristic approaches
 - BLAST
- Practical database searching
 - BLAST, PSI-BLAST and HMM approaches