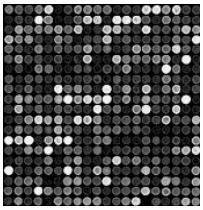
Protein Bioinformatics



Johns Hopkins Bloomberg School of Public Health
260.655
Thursday, April 1, 2010
Jonathan Pevsner

Outline for today

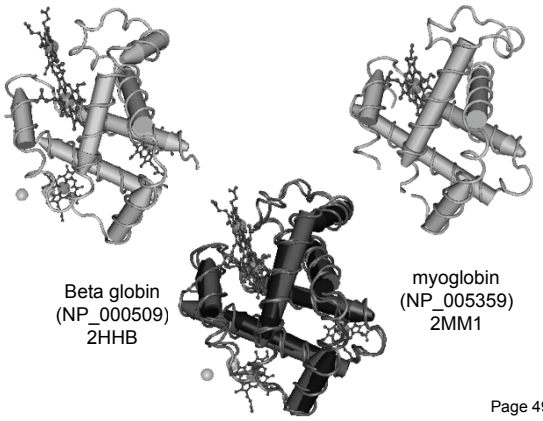
1. Homology and pairwise alignment
2. BLAST
3. Multiple sequence alignment
4. Phylogeny and evolution

Learning objectives: homology & alignment

1. You should know the definitions of homologs, orthologs, and paralogs
2. You should know how to determine whether two genes (or proteins) are homologous
3. You should know what a scoring matrix is
4. You should know how alignments are performed
5. You should know how to align two sequences using the BLAST tool at NCBI

Pairwise sequence alignment is the most fundamental operation of bioinformatics

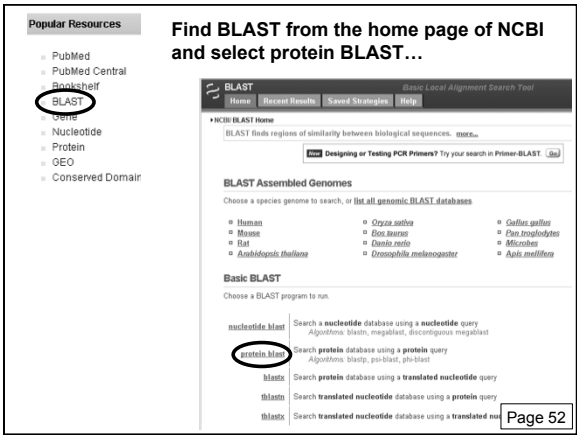
- It is used to decide if two proteins (or genes) are related structurally or functionally
- It is used to identify domains or motifs that are shared between proteins
- It is the basis of BLAST searching (next topic)
- It is used in the analysis of genomes

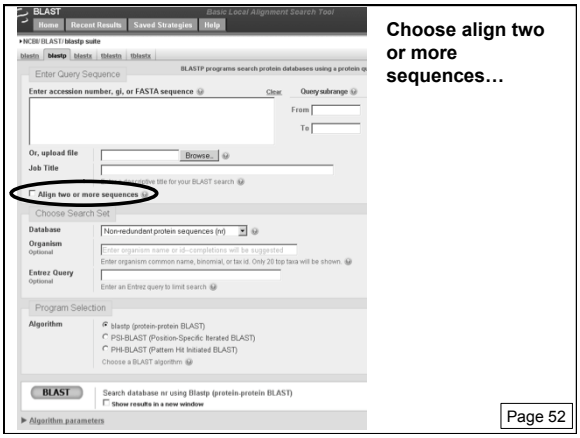


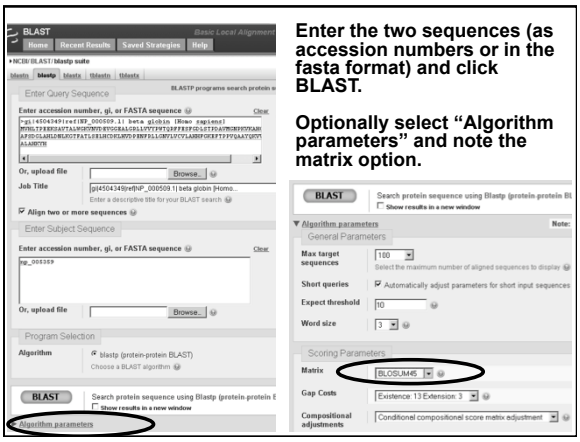
Page 49

Pairwise alignment: protein sequences can be more informative than DNA

- protein is more informative (20 vs 4 characters); many amino acids share related biophysical properties
- codons are degenerate: changes in the third position often do not alter the amino acid that is specified
- protein sequences offer a longer “look-back” time
- DNA sequences can be translated into protein, and then used in pairwise alignments


Popular Resources
PubMed
PubMed Central
Bookshelf
BLAST
Gene
Nucleotide
Protein
GEO
Conserved Domain
Find BLAST from the home page of NCBI and select protein BLAST…
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
Designing or Testing PCR Primers? Try your search in Primer-BLAST.
BLAST Assembled Genomes
Choose a species genome to search, or list all genomic BLAST databases
Human
Mouse
Rat
Arabidopsis thaliana
Oryza sativa
Bos taurus
Danio rerio
Drosophila melanogaster
Gallus gallus
Pan troglodytes
Microbes
Apis mellifera
Basic BLAST
Choose a BLAST program to run.
nucleotide blast
Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast
protein blast
Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast
blastx
Search protein database using a translated nucleotide query
tblastn
Search translated nucleotide database using a protein query
tblastx
Search translated nucleotide database using a translated nu
Page 52


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
Query subrange
From
To
Or, upload file
Browse...
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
Enter organism name or id--completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
Choose a BLAST algorithm
BLAST
Search database nr using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Choose align two or more sequences…
Page 52


BLAST
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
Or, upload file
Browse...
Job Title
Align two or more sequences
Enter Subject Sequence
Enter accession number, gi, or FASTA sequence
Clear
Or, upload file
Browse...
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein
Show results in a new window
Algorithm parameters
Enter the two sequences (as accession numbers or in the fasta format) and click BLAST.
Optionally select "Algorithm parameters" and note the matrix option.
BLAST
Search protein sequence using Blastp (protein-protein
Show results in a new window
Algorithm parameters
Note:
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Scoring Parameters
Matrix
BLOSUM45
Gap Costs
Existence: 13 Extension: 3
Compositional adjustments
Conditional compositional score matrix adjustment

Pairwise alignment result of human beta globin and myoglobin

Myoglobin RefSeq

Information about this alignment: score, expect value, identities, positives, gaps…

```
>ref|NP_005359.1| G myoglobin [Homo sapiens]
 ref|NP_976311.1| UG myoglobin [Homo sapiens]
 ref|NP_976312.1| G myoglobin [Homo sapiens]
 ▶11 more sequence titles
 Length=154

 GENE ID: 4151 MB | myoglobin [Homo sapiens] (Over 10 PubMed links)

 Score = 47.4 bits (144),  Expect = 8e-11, Method: Compositional matrix adjust.
 Identities = 37/145 (25%), Positives = 57/145 (39%), Gaps = 2/145 (1%)

Query  4    LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
            L+  E   V  +WGKV  D  G   E L RL   +P T   F+ F  L + D +  +  +
Sbjct  3    LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
            K HG  VL A    L    + +    L++ H  K  +  +    +   ++ VL
Sbjct  63   KGHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  122  EFTPPVQAAYQKVVAGVANALAHKY  146
            +F    Q A  K +      +A  Y
Sbjct  123  DFGADAQGAMNKALELFRKDMASKY  147
```

Query = HBB
Subject = MB

Middle row displays identities; + sign for similar matches

Page 53

Pairwise alignment result of human beta globin and myoglobin: the score is a sum of match, mismatch, gap creation, and gap extension scores

```
Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12  VTALWGKVNVD--EVGGEALGRLL  33
           V  +WGKV  D    G E L RL
Sbjct  11  VLNVWGKVEADIPGHGQEVLIRLF  34

match      4  11 5   6    6 5 4 5    sum of matches: +60
                6 4                 4
mismatch   -1 1    0   -2 -2  -4  0  sum of mismatches: -13
            -2      0   -3   0
gap open              -11            sum of gap penalties: -12
gap extend             -1
                                     total raw score: 60 - 13 - 12 = 35
```

Page 53

Pairwise alignment result of human beta globin and myoglobin: the score is a sum of match, mismatch, gap creation, and gap extension scores

```
Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12  VTALWGKVNVD--EVGGEALGRLL  33
           V  +WGKV  D    G E L RL
Sbjct  11  VLNVWGKVEADIPGHGQEVLIRLF  34

match      4  11 5   6    6 5 4 5    sum of matches: +60
                6 4                 4
mismatch   -1 1    0   -2 -2  -4  0  sum of mismatches: -13
            -2      0   -3   0
gap open              -11            sum of gap penalties: -12
gap extend             -1
                                     total raw score: 60 - 13 - 12 = 35
```

V matching V earns +4
T matching L earns -1

These scores come from a "scoring matrix"!

Page 53

Gaps

```
Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12  VTALWGKVNVD--EVGGEALGRLL  33
           V  +WGKV  D    G E L RL
Sbjct  11  VLNVWGKVEADIPGHGQEVLIRLF  34

match      4  11 5   6    6 5 4 5    sum of matches: +60
             6 4              4
mismatch  -1 1    0  -2 -2  -4  0    sum of mismatches: -13
   -2       0  -3   0
gap open          -11                 sum of gap penalties: -12
gap extend         -1
                                      total raw score: 60 - 13 - 12 = 35
```

First gap position scores -11
Second gap position scores -1
Gap creation tends to have a large negative score;
Gap extension involves a small penalty

Page 55

Definitions

Pairwise alignment
The process of lining up two sequences
to achieve maximal levels of identity
(and conservation, in the case of amino acid sequences)
for the purpose of assessing the degree of similarity
and the possibility of homology.

Definitions

Homology
Similarity attributed to descent from a common ancestor.

Identity

The extent to which two (nucleotide or amino acid)
sequences are invariant.

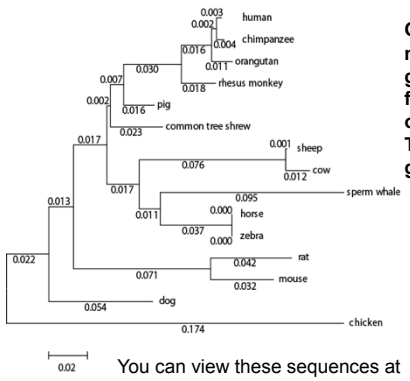
Page 44

Definitions: two types of homology

Orthologs
Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.

Paralogs
Homologous sequences within a single species that arose by gene duplication.

Page 43

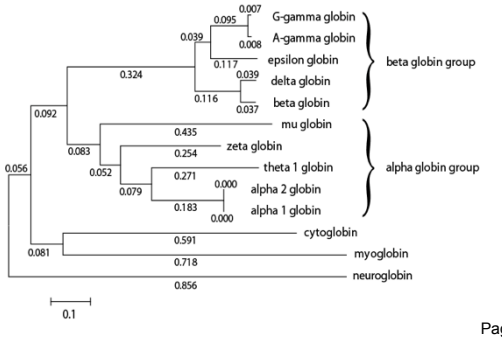


Orthologs: members of a gene (protein) family in various organisms. This tree shows globin orthologs.

You can view these sequences at www.bioinfbook.org (document 3.1)

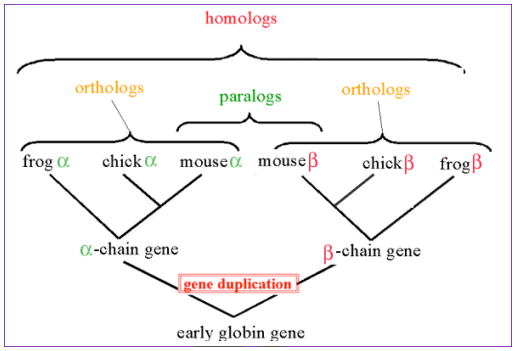
Page 51

Paralogs: members of a gene (protein) family within a species. This tree shows human globin paralogs.



Page 52

Orthologs and paralogs are often viewed in a single tree



Source: NCBI

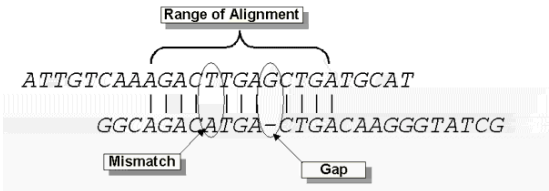
Definitions

Similarity
The extent to which nucleotide or protein sequences are related. It is based upon identity plus conservation.

Identity
The extent to which two sequences are invariant.

Conservation
Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

Calculation of an alignment score



S= Σ(identities, mismatches) - Σ (gap penalties)

Score = Max(S)

Source: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Alignment_Scores2.html

PAM250 mutation probability matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Top: original amino acid
Side: replacement amino acid

PAM250 log odds scoring matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

PAM10 log odds scoring matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-20	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-17	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-8	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

Two kinds of sequence alignment: global and local

We will first consider the global alignment algorithm of Needleman and Wunsch (1970).

We will then discuss the local alignment algorithm of Smith and Waterman (1981).

Finally, we will consider BLAST, a heuristic version of Smith-Waterman. BLAST is faster but less rigorous.

Page 63

Global alignment with the algorithm of Needleman and Wunsch (1970)

- Two sequences can be compared in a matrix along x- and y-axes.
- If they are identical, a path along a diagonal can be drawn
- Find the optimal subpaths, and add them up to achieve the best score. This involves
  - --adding gaps when needed
  - --allowing for conservative substitutions
  - --choosing a scoring system (simple or complicated)
- N-W is guaranteed to find optimal alignment(s)

Page 63

Three steps to global alignment with the Needleman-Wunsch algorithm

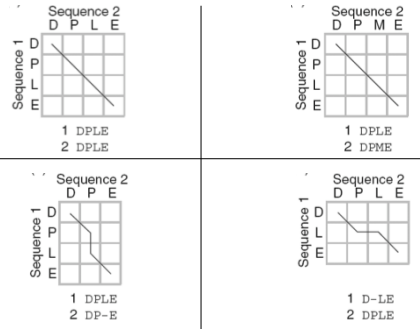
[1] set up a matrix

[2] score the matrix

[3] identify the optimal alignment(s)

Page 63

Four possible outcomes in aligning two sequences



Page 77

Start Needleman-Wunsch with an identity matrix

(e)

Sequence 2 (from honeybee globin)

Sequence 1 (from human cytoglobin)

	F	M	D	T	P	L	N	E
F	1							
K								
H								
M		1						
E								1
D			1					
P					1			
L						1		
E								1

Page 77

Start Needleman-Wunsch with an identity matrix
(or, as here, use values from a scoring matrix)

Sequence 2

Sequence 1

	F	M	D	T	P	L	N	E
F	6	0	–3	–2	–4	0	–3	–3
K	–3	–1	–1	–1	–1	–2	0	1
H	–1	–2	–1	–2	–2	–3	1	0
M	0	5	–3	–1	–2	2	–2	–2
E	–3	–2	2	–1	–1	–3	0	5
D	–3	–3	6	–1	–1	–4	1	2
P	–4	–2	–1	–1	7	–3	–2	–1
L	0	2	–4	–1	–3	4	–3	–3
E	–3	–2	2	–1	–1	–3	0	5

Page 77

Fill in the matrix using “dynamic programming”

(**a**)

Sequence 1 \ Sequence 2		F	M	D	T	P	L	N	E
	0	–2	–4	–6	–8	–10	–12	–14	–16
F	–2								
K	–4								
H	–6								
M	–8								
E	–10								
D	–12								
P	–14								
L	–16								
E	–18								

Page 78

Fill in the matrix using “dynamic programming”

(**b**)

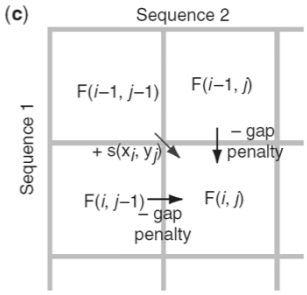
$$\text{Score} = \text{Max}\begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - \text{gap penalty} \\ F(i, j-1) - \text{gap penalty} \end{cases}$$

Score (this example) = +1 (match)
–2 (mismatch)
–2 (gap penalty)

Page 78

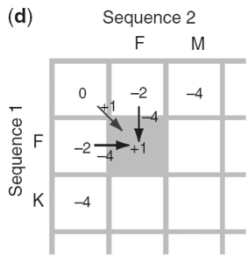
Fill in the matrix using “dynamic programming”

(**c**)



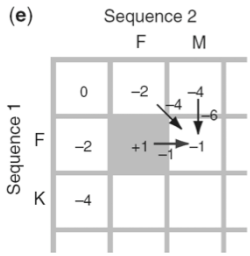
Page 78

Fill in the matrix using “dynamic programming”



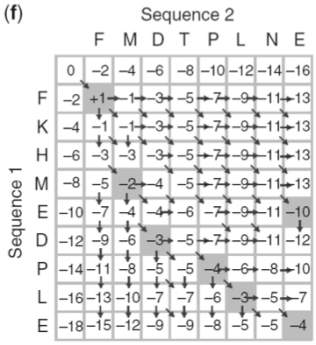
Page 78

Fill in the matrix using “dynamic programming”



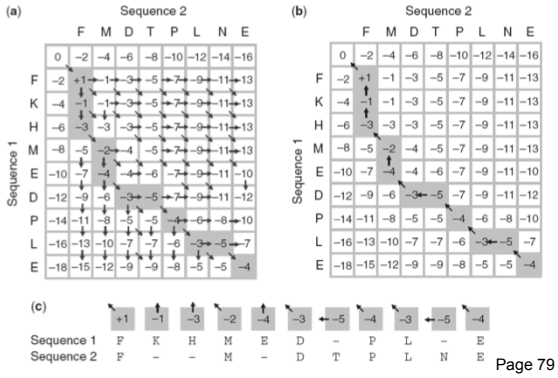
Page 78

Fill in the matrix using “dynamic programming”



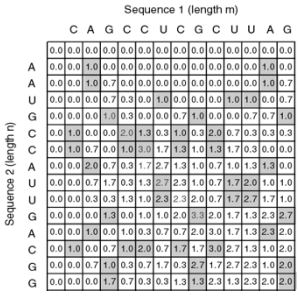
Page 78

Traceback to find the optimal (best) pairwise alignment



Page 79

Smith-Waterman allows local alignment (internal terminations)



Page 70

Rapid, heuristic versions of Smith-Waterman: BLAST

Smith-Waterman is very rigorous and it is guaranteed to find an optimal alignment.

But Smith-Waterman is slow. It requires computer space and time proportional to the product of the two sequences being aligned (or the product of a query against an entire database).

BLAST provides a rapid alternative to S-W, although it's not as accurate.

Page 71

Outline for today

1. Homology and pairwise alignment
2. BLAST
3. Multiple sequence alignment
4. Phylogeny and evolution

Learning objectives: BLAST

1. You should know what the five basic BLAST programs are
2. You should be able to perform a BLAST search
3. You should be able to interpret the results of a BLAST search

BLAST

BLAST (Basic Local Alignment Search Tool) allows rapid sequence comparison of a query sequence against a database.

The BLAST algorithm is fast, accurate, and web-accessible.

page 87

Why use BLAST?

BLAST searching is fundamental to understanding the relatedness of any favorite query sequence to other known proteins or DNA sequences.

Applications include
- identifying orthologs and paralogs
- discovering new genes or proteins
- discovering variants of genes or proteins
- investigating expressed sequence tags (ESTs)
- exploring protein structure and function

page 88

Four components to a BLAST search

(1) Choose the sequence (query)

(2) Select the BLAST program

(3) Choose the database to search

(4) Choose optional parameters

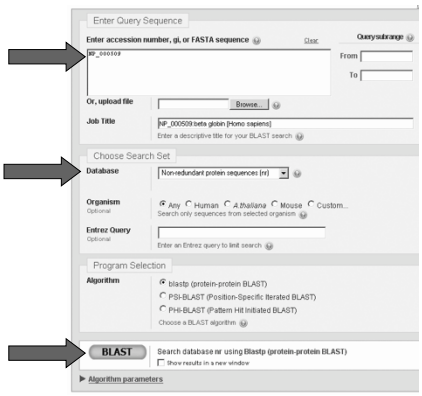
Then click “BLAST”

page 88

Step 1: Choose your sequence

Sequence can be input in FASTA format or as accession number

page 89



Step 2: Choose the BLAST program

Choose the BLAST program

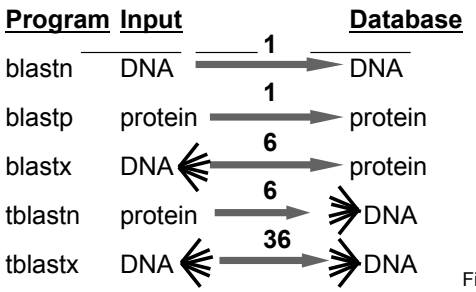


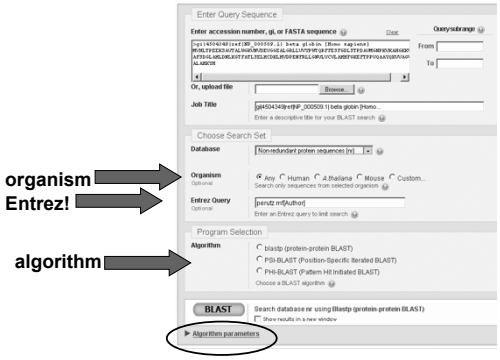
Fig. 4.3
page 91

Step 3: choose the database

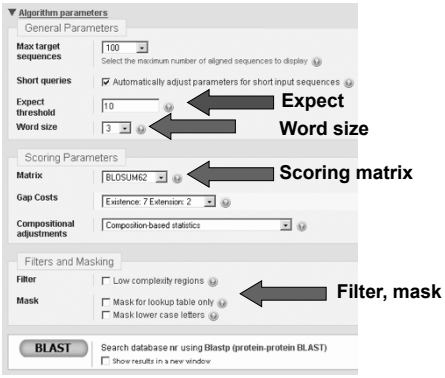
nr = non-redundant (most general database)
dbest = database of expressed sequence tags
dbsts = database of sequence tag sites
gss = genomic survey sequences
htgs = high throughput genomic sequence

page 92-93

Step 4a: Select optional search parameters



Step 4a: optional blastp search parameters



How a BLAST search works

"The central idea of the BLAST algorithm is to confine attention to segment pairs that contain a word pair of length w with a score of at least T."

Altschul et al. (1990)

(page 101, 102)

How the original BLAST algorithm works: three phases

Phase 1: compile a list of word pairs (w=3) above threshold T

Example: for a human RBP query …FSGTWYA… (query word is in yellow)

A list of words (w=3) is:

```
FSG SGT GTW TWY WYA
YSG TGT ATW SWY WFA
FTG SVT GSW TWF WYS
```

Fig. 4.13
page 101

Phase 1: compile a list of words (w=3)

	GTW	6,5,11	22
neighborhood	GSW	6,1,11	18
word hits	ATW	0,5,11	16
> threshold	NTW	0,5,11	16
(T=11)	GTY	6,5,2	13
	GNW		10
neighborhood	GAW		9
word hits			
< below threshold			

Fig. 4.13
page 101

How a BLAST search works: 3 phases

Phase 2:

Scan the database for entries that match the compiled list.

This is fast and relatively easy.

Fig. 4.13
page 101

How a BLAST search works: 3 phases

Phase 3: when you manage to find a hit (i.e. a match between a "word" and a database entry), extend the hit in either direction.

Keep track of the score (use a scoring matrix)

Stop when the score drops below some cutoff.

```
KENFDKARFSGTWYAMAKKDPEG 50  RBP (query)
MKGLDIQKVAGTWYSLAMAASD. 44  lactoglobulin (hit)
```

extend **Hit!** **extend**

page 101

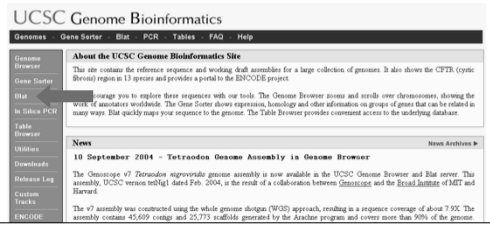
BLAST-related tools for genomic DNA

Recently developed tools include:

- MegaBLAST at NCBI.
- BLAT (BLAST-like alignment tool). BLAT parses an entire genomic DNA database into words (11mers), then searches them against a query. Thus it is a mirror image of the BLAST strategy. See http://genome.ucsc.edu
- SSAHA at Ensembl uses a similar strategy as BLAT. See http://www.ensembl.org

Page 136

To access BLAT, visit http://genome.ucsc.edu



"BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 20 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates." --BLAT website

BLAT output includes browser and other formats. Try a beta globin protein search to view homologs.

How to interpret a BLAST search: expect value

The expect value *E* is the number of alignments with scores greater than or equal to score *S* that are expected to occur by chance in a database search.

An *E* value is related to a probability value *p*.

The key equation describing an *E* value is:

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

page 105

How to interpret BLAST: *E* values and *p* values

Very small *E* values are very similar to *p* values.
E values of about 1 to 10 are far easier to interpret than corresponding *p* values.

E	*p*
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258 (about 0.1)
0.05	0.04877058 (about 0.05)
0.001	0.00099950 (about 0.001)
0.0001	0.0001000

Table 4.4
page 107

Sometimes a real match has an E value > 1

```
                                                                  Score     E
Sequences producing significant alignments:                       (bits)  Value

gi|5803139|ref|NP_006735.1|  retinol-binding protein 4, inte...   378  e-105
gi|230284|pdb|1RBP|     Retinol Binding Protein >gi|4938971|p...   371  e-103
gi|88364|pir||A27786  plasma retinol-binding protein - human       370  e-103
gi|4558179|pdb|1QAB|E  Chain E, The Structure Of Human Retin...    363  e-100
gi|7770173|gb|AAF69622.1|AF119917 30  (AF119868) PRO2222 [Ho...    324  5e-89
gi|13645517|ref|XP_005907.2|  retinol-binding protein 4, int...    233  9e-62
gi|296672|emb|CAA26553.1|  (X02775) RBP [Homo sapiens]             207  8e-54
gi|5419892|emb|CAB46489.1|  (X02824) RBP (aa 101-172) [Homo ...    149  2e-36
gi|2895204|gb|AAC02945.1|  (AF025334) mutant retinol binding...     90  2e-18
gi|2895206|gb|AAC02946.1|  (AF025335) mutant retinol binding...     73  2e-13
gi|4502163|ref|NP_001638.1|  apolipoprotein D precursor [Hom...     55  4e-08
gi|619183|gb|AAB32200.1|  apolipoprotein D, apoD [human, pla...     55  5e-08
gi|1246096|gb|AAB35919.1|  (S80440) apolipoprotein D, apoD (...     43  3e-04
gi|223373|prf||0801163A  complex-forming glycoprotein HC [Ho...     37  0.011
gi|4884164|emb|CAB43305.1|  (AL050169) hypothetical protein ...     35  0.043
gi|13639329|ref|XP_005360.3|  61620 [Homo sapiens] >gi|13639...     35  0.043
gi|4502067|ref|NP_001624.1|  alpha-1-microglobulin/bikunin p...     35  0.068
gi|14735821|ref|XP_029964.1|  progestagen-associated endomet...     35  0.070
gi|4557799|ref|NP_000597.1|  complement component 8, gamma p...     34  0.14
gi|4505583|ref|NP_002562.1|  progestagen-associated endometr...     32  0.49
gi|13639651|ref|XP_005430.2|  complement component 8, gamma ...     31  1.1
```

…try a reciprocal BLAST to confirm

Fig. 4.18
page 110

Outline for today

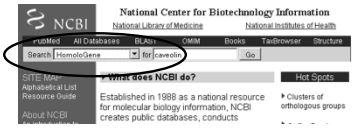
1. Homology and pairwise alignment
2. BLAST
3. Multiple sequence alignment
4. Phylogeny and evolution

Multiple sequence alignment: definition

- a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned
- homologous residues are aligned in columns across the length of the sequences
- residues are homologous in an evolutionary sense
- residues are homologous in a structural sense

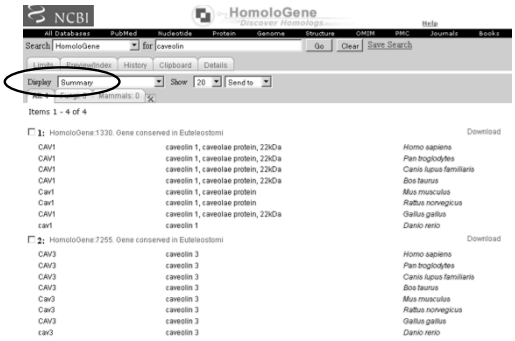
Page 180

Example: someone is interested in caveolin



Step 1: at NCBI change the pulldown menu to HomoloGene and enter caveolin in the search box

Step 2: inspect the results. We'll take the first set of caveolins. Change the Display to Multiple alignment.



Step 3: inspect the multiple alignment. Note that these eight proteins align nicely, although gaps must be included.

1: HomoloGene:1330. Gene conserved in Euteleostomi — Downlo

Multiple Sequence Alignment

Generated by MUSCLE [see reference] version 3.6 (using options: -maxiters 2).

```
NP_001744.2          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADELSEKQVYDAHT   46
XP_519325.2          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADELSEKQVYDAHT   46
NP_001003296.1       1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-AEEMSEKQVYDAHT   46
NP_776429.1          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-AEEMNEKQVYDAHT   46
NP_031642.1          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADEVTEKQVYDAHT   46
NP_113744.1          1   MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADEVNEKQVYDAHT   46
XP_001234148.1       1   ---MEYFQ----EAFLYAAPVREQGNIYKPNNKMM-ADELSEKAVHDVHT   42
NP_997816.1          1   MTSG-YKDGTPEEEYAHSPFIREQGNIYKPNNKEMDNDSINEKTLQDVHT   49

NP_001744.2         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
XP_519325.2         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
NP_001003296.1      47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
NP_776429.1         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
NP_031642.1         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
NP_113744.1         47   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   96
XP_001234148.1      43   KEIDLVNRDPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK   92
NP_997816.1         50   KEIDLVNRDPKHLNDDVVKVDFEDVIAEPAGTYSFDGVWKASFTTFTVTK   99
```

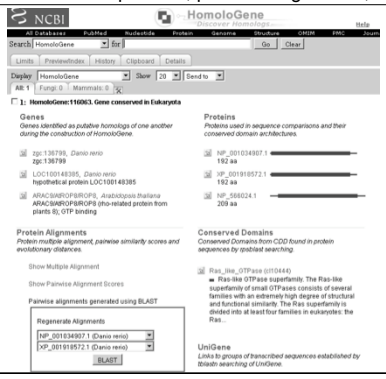
Here's another multiple alignment, Rac:

```
NP_061485.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
XP_855587.1          1   -----MQAIKCVVVEDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_776588.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_033033.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_599193.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_990348.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_956065.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
NP_648121.1          1   -----MQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM   45
XP_366655.1          1   MAAPGVQSLKCVVTGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSASVM   50
XP_329350.1          1   MLTGEMLTLDFLLL-------TCLLISYTTNAFPGEYIPTVFDNYSASVM   43
NP_195320.1          1   --MSASRFIKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV   48
NP_179371.1          1   --MSASRFIKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV   48
NP_190698.1          1   --MSASRFVKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV   48
NP_195228.1          1   --MSASRFIKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVI   48
NP_001048639.1       1   --MSASRFIKCVTVGDGAVGKTCMLISYTSNTFPTDYVPTVFDNFSANVV   48

NP_061485.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQTVGETYGKDITSRGKDKPIAD   95
XP_855587.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT----------------D   76
NP_776588.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT----------------D   76
NP_033033.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT----------------D   76
NP_599193.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT----------------D   76
NP_990348.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT----------------D   76
NP_956065.1         46   VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT----------------D   76
NP_648121.1         46   VDAKPINLGLWDTA
XP_366655.1         51   VDGKPISLGLWDTA
XP_329350.1         44   VDGKPVSLGLWDTA
NP_195320.1         49   VNGATVNLGLWDTA
NP_179371.1         49   VNGATVNLGLWDTA
NP_190698.1         49   VNGSTVNLGLWDTAGQEDYNRLRPLSYRGA------------------D   79
NP_195228.1         49   VDGNTINLGLWDTAGQEDYNRLRPLSYRGA------------------D   79
NP_001048639.1      49   VDGSTVNLGLWDTA--EDYNRLRPLSYRGA------------------D   77
```

This insertion could be due to alternative splicing

HomoloGene includes groups of eukaryotic proteins. The site includes links to the proteins, pairwise alignments, and more



Example: globins

Let's look at a multiple sequence alignment (MSA) of five globins proteins. We'll use five prominent MSA programs: ClustalW, Praline, MUSCLE (used at HomoloGene), ProbCons, and TCoffee. Each program offers unique strengths.

We'll focus on a histidine (H) residue that has a critical role in binding oxygen in globins, and should be aligned. But often it's not aligned, and all five programs give different answers.

Our conclusion will be that there is no single best approach to MSA. Dozens of new programs have been introduced in recent years.

ClustalW

```
CLUSTAL W (1.83) multiple sequence alignment

                                                          ▼
beta globin  ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin    -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin  -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean      ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
                         :   :   :   :.. .    .      ::   *     *.
                            ▽
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin    HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS---LGRKHRAVGVKLS 104
soybean      --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice         --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
               .      . ..  *  .::       :           :           :

beta globin  NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
myoglobin    YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin  SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean      QFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA------- 144
rice         HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
              :  :  ::  :       :     * .    .  :
```

Note how the region of a conserved histidine (▼) varies depending on which of five prominent algorithms is used

Praline

```
(a) Praline multiple sequence alignment

                                                                ▼
beta globin      ..........MVHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG
myoglobin        ...........MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK.FK
neuroglobin      ............MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean          ..........MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS..FL
rice             MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS..FL
Consistency      0000000000142654382579345734633643436244536864 33*35344*50063
                                  ▽                   ▼
beta globin      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSEL..HCDKLH....VDP
myoglobin        HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQS..HATKHK....IPV
neuroglobin      QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG....VKL
soybean          A.NGVDP..TNPKLTGHAEKLFALVRDSAGQL.KASGTVVADAA....LGSVHAQKAVTD
rice             R.NSDVPLEKNPKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRDTTLKRLGATHLKYGVGD
Consistency      3166354224776653*43686354244544513356343335420033354400000922

beta globin      ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKVVAGVANALAHKYH......
myoglobin        KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin      SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..
soybean          PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKKA........
rice             AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...
Consistency      437448444982585423053365544 54*5546542644675432 2001000
```

MUSCLE

```
(b)
MUSCLE (3.6) multiple sequence alignment

beta globin  ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin    -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin  -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean      ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
                            :  :  :  :  .. .   .     ::  *    *.

beta globin  DLSTPDAVMGNPKVKAHGKKVLGAF---SDGLAHLDNLKGTFATLSELHCDKLH--VDPE
myoglobin    HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAEIKPLAQSHATKHK--IPVK
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVI---DAAVTNVEDLSSLEEYLASLGRKHRAVGVKLS
soybean      NGVDF----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLRKGVGDA
                        . ..  *  .::         :                  :   :

beta globin  NFRLLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH------
myoglobin    YLEFISECIIQVLQSKHPGD-FGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin  SFSTVGESLLYMLEKCLGPA-FTPATRAAWSQLYGAVVQAMSRGWDGE----
soybean      QFVVVKEALLKTIKAAVGDK-WSDELSRAWEVAYDELAAAIKKA-------
rice         HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE---
              :  :   ::  :        :     * .    .   :
```

Page 194

Probcons

```
(c)
PROBCONS

beta globin  M----------VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin    M-----------GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin  M-------------ERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean      M----------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
             *          * :  :  :  :  .. .   .     ::  *    *.

beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP
myoglobin    HLKSEDEMKASEDLKKHGATVLTALGGI---LKKKGHHE---AEIKPLAQSHATKHKIPV
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLE---EYLASLGRKHRAV-GVKL
soybean      NGVDF----TNPKLTGHAEKLFALVRDSAGQLKASGTVV----ADAALGSVHAQK-AVTD
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKY-GVGD
             .    :   . ..  *  .::        ::   .          *.  *     :

beta globin  ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHK------YH
myoglobin    KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin  SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE
soybean      PQFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIK--------KA
rice         AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
               :  :   ::  :        :     * .    .   :
```

Page 195

TCoffee

```
(d)
CLUSTAL FORMAT for T-COFFEE Version_5.13

beta globin  ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
myoglobin    -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin  -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean      ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-FLA
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
                            :  :  :  :  .. .   .     ::  *    *.

beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL---KGTF---ATLSELHCDKLHVDP
myoglobin    HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAE---IKPLAQSHATKHKIPV
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDL---SSLEEYLASLGRKH-RAVGVKL
soybean      NGVDF----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
             .          . ..  *  .::        ::   .          *.  *

beta globin  ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------
myoglobin    KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin  SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDG----E
soybean      Q-FVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA--------
rice         H-FEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
               :  :   ::  :        :     * .    .   :
```

Conclusions: ClustalW (the most popular program) gives different answers than a set of recent, improved alternatives. No one method is ideal.

Page 195

Multiple sequence alignment: properties

- not necessarily one “correct” alignment of a protein family
- protein sequences evolve...
- ...the corresponding three-dimensional structures of proteins also evolve
- may be impossible to identify amino acid residues that align properly (structurally) throughout a multiple sequence alignment
- for two proteins sharing 30% amino acid identity, about 50% of the individual amino acids are superposable in the two structures

Page 180

Multiple sequence alignment: features

- some aligned residues, such as cysteines that form disulfide bridges, may be highly conserved
- there may be conserved motifs such as a transmembrane domain
- there may be conserved secondary structure features
- there may be regions with consistent patterns of insertions or deletions (indels)

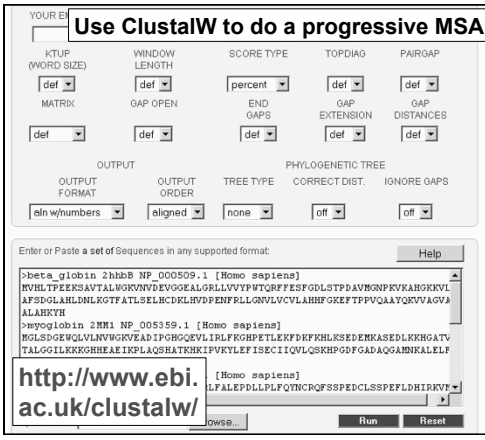
Page 181

Multiple sequence alignment: uses

- MSA is more sensitive than pairwise alignment to detect homologs
- BLAST output can take the form of a MSA, and can reveal conserved residues or motifs
- Population data can be analyzed in a MSA (PopSet)
- A single query can be searched against a database of MSAs (e.g. PFAM)
- Regulatory regions of genes may have consensus sequences identifiable by MSA

Page 181

Use ClustalW to do a progressive MSA



http://www.ebi.ac.uk/clustalw/

Page 186

Feng-Doolittle MSA (implemented in ClustalW and other programs) occurs in 3 stages

[1] Do a set of global pairwise alignments (Needleman and Wunsch's dynamic programming algorithm)

[2] Create a guide tree

[3] Progressively align the sequences

Page 185

Progressive MSA stage 1 of 3: generate global pairwise alignments

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	beta_globin	147	2	myoglobin	154	25
1	beta_globin	147	3	neuroglobin	151	15
1	beta_globin	147	4	soybean	144	13
1	beta_globin	147	5	rice	166	21
2	myoglobin	154	3	neuroglobin	151	16
2	myoglobin	154	4	soybean	144	8
2	myoglobin	154	5	rice	166	12
3	neuroglobin	151	4	soybean	144	17
3	neuroglobin	151	5	rice	166	18
4	soybean	144	5	rice	166	43

best score

Page 186

Number of pairwise alignments needed

For *n* sequences, (*n*-1)(*n*) / 2

For 5 sequences, (4)(5) / 2 = 10

For 200 sequences, (199)(200) / 2 = 19,900

Page 185

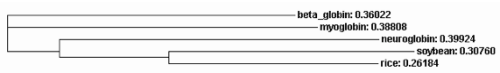
Feng-Doolittle stage 2: guide tree

- Convert similarity scores to distance scores
- A tree shows the distance between objects
- Use UPGMA (defined in the phylogeny lecture)
- ClustalW provides a syntax to describe the tree
- A guide tree is not a phylogenetic tree

Page 187

Progressive MSA stage 2 of 3: generate a guide tree calculated from the distance matrix (5 distantly related globins)

```
(
beta_globin:0.36022,
myoglobin:0.38808,
(
neuroglobin:0.39924,
(
soybean:0.30760,
rice:0.26184)
:0.13652)
:0.06560);
```

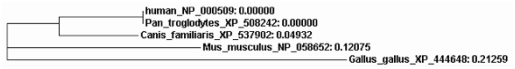


Page 186

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	human_NP_000509	147	2	Pan_troglodytes_XP_508242	147	100
1	human_NP_000509	147	3	Canis_familiaris_XP_537902	147	89
1	human_NP_000509	147	4	Mus_musculus_NP_058652	147	80
1	human_NP_000509	147	5	Gallus_gallus_XP_444648	147	69
2	Pan_troglodytes_XP_508242	147	3	Canis_familiaris_XP_537902	147	89
2	Pan_troglodytes_XP_508242	147	4	Mus_musculus_NP_058652	147	80
2	Pan_troglodytes_XP_508242	147	5	Gallus_gallus_XP_444648	147	69
3	Canis_familiaris_XP_537902	147	4	Mus_musculus_NP_058652	147	78
3	Canis_familiaris_XP_537902	147	5	Gallus_gallus_XP_444648	147	71
4	Mus_musculus_NP_058652	147	5	Gallus_gallus_XP_444648	147	66

```
(
(
(
human_NP_000509:0.00000,
Pan_troglodytes_XP_508242:0.00000)
:0.05272,
Canis_familiaris_XP_537902:0.04932)
:0.03231,
Mus_musculus_NP_058652:0.12075,
Gallus_gallus_XP_444648:0.21259);
```

5 closely related globins



Page 188

Feng-Doolittle stage 3: progressive alignment

- Make a MSA based on the order in the guide tree
- Start with the two most closely related sequences
- Then add the next closest sequence
- Continue until all sequences are added to the MSA
- Rule: “once a gap, always a gap.”

Page 188

Clustal W alignment of 5 distantly related globins

```
CLUSTAL W (1.83) multiple sequence alignment


beta_globin       ---------------MVHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFF 43
cytoglobin        MEKVPGEMEIERRERSEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFPSAKQYF 60
myoglobin         ----------------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKF 44
neuroglobin       -----------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLF 42
leghemoglobin     ----------------MGFTEKQEALVNSSWELFKQNPS-YSVLFYTIILKKAPAAKGMF 43
                                     :   :   :   *     .    .      ::   *     *

beta_globin       ES-FGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLA---HLDNLKGTFATLSELHCDKLHV 99
cytoglobin        SQ-FKGNEDPLEMERSPQLRKHACRVMGALNTVVENLHDPDKVSSVLALVGKAHALKHKV 119
myoglobin         DK-FKGLKSEDEMKASEDLKKHGATVLTALGGILK---KKGHHEAEIKPLAQSHATKHKI 100
neuroglobin       QYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG-V 101
leghemoglobin     S----FLKDSAEVVDSPKLQAHAEKVFGMVHDSAIQLRASGEVVLGDATLGAIHIQKGVV 99
                  .     :.       . ..  *   *:  .          .        :.  *     :

beta_globin       DPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------------ 147
cytoglobin        EPVYFKILSGVILEVVAEEFASDFPPETQRAWAKLRGLITSHVTAAYKEVGWVQQVPNAT 179
myoglobin         PVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRDMAANYKELGFQG------ 154
neuroglobin       KLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWDGE---------- 151
leghemoglobin     DP-HFVVVKEALLETIKEASGEKWSEELSTAWEVAYEGLASAIKKAMN------------ 146
                      :  :   ::  :       :      *       . . :    .

beta_globin       -----------
cytoglobin        TPPATLFSSGF 190
myoglobin         -----------
neuroglobin       -----------
leghemoglobin     -----------
```

Fig. 6.3
Page 187

Clustal W alignment of 5 closely related globins

```
CLUSTAL W (1.83) multiple sequence alignment


human_NP_000509                 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Pan_troglodytes_XP_508242       MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Canis_familiaris_XP_537902      MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLS 50
Mus_musculus_NP_058652          MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLS 50
Gallus_gallus_XP_444648         MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLS 50
                                *** *  **. :: **.***  * *.***.***:*******:* ***:**

human_NP_000509                 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Pan_troglodytes_XP_508242       TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Canis_familiaris_XP_537902      TPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVD 100
Mus_musculus_NP_058652          SASAIMGNPKVKAHGKKVITAFNEGLKNLDNLKGTFASLSELHCDKLHVD 100
Gallus_gallus_XP_444648         SPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVD 100
                                :. *::.*. *:******: :*.:.: :***:*.**: ************

human_NP_000509                 PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Pan_troglodytes_XP_508242       PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Canis_familiaris_XP_537902      PENFRLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH 147
Mus_musculus_NP_058652          PENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH 147
Gallus_gallus_XP_444648         PENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH 147
                                ****:***: :: **. *:.*:***  ***:**:*  ** ***:***
```

* asterisks indicate identity in a column

Fig. 6.5
Page 189

Why “once a gap, always a gap”?

- There are many possible ways to make a MSA
- Where gaps are added is a critical question
- Gaps are often added to the first two (closest) sequences
- To change the initial gap choices later on would be to give more weight to distantly related sequences
- To maintain the initial gap choices is to trust that those gaps are most believable

Page 189

Outline for today

1. Homology and pairwise alignment
2. BLAST
3. Multiple sequence alignment
4. Phylogeny and evolution

Learning objectives: phylogeny

1. You should know how to create a phylogenetic tree from a multiple sequence alignment

2. You should know the parts of a tree

3. You should know how to interpret the biological (historical) meaning of a tree

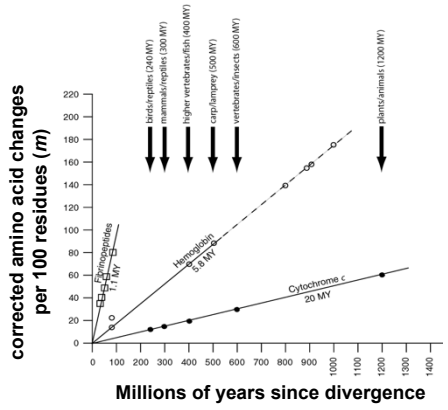
Molecular clock hypothesis

In the 1960s, sequence data were accumulated for small, abundant proteins such as globins, cytochromes *c*, and fibrinopeptides. Some proteins appeared to evolve slowly, while others evolved rapidly.

Linus Pauling, Emanuel Margoliash and others proposed the hypothesis of a molecular clock:

For every given protein, the rate of molecular evolution is approximately constant in all evolutionary lineages

Page 221



Dickerson (1971)

Fig. 7.5
Page 222

Molecular clock hypothesis: conclusions

Dickerson drew the following conclusions:

- For each protein, the data lie on a straight line. Thus, the rate of amino acid substitution has remained constant for each protein.
- The average rate of change differs for each protein. The time for a 1% change to occur between two lines of evolution is 20 MY (cytochrome c), 5.8 MY (hemoglobin), and 1.1 MY (fibrinopeptides).
- The observed variations in rate of change reflect functional constraints imposed by natural selection.

Page 223

Five stages of phylogenetic analysis

[1] Selection of sequences for analysis

[2] Multiple sequence alignment

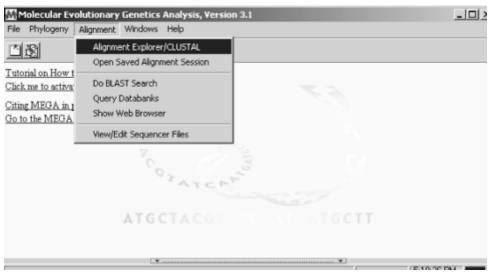
[3] Selection of a substitution model

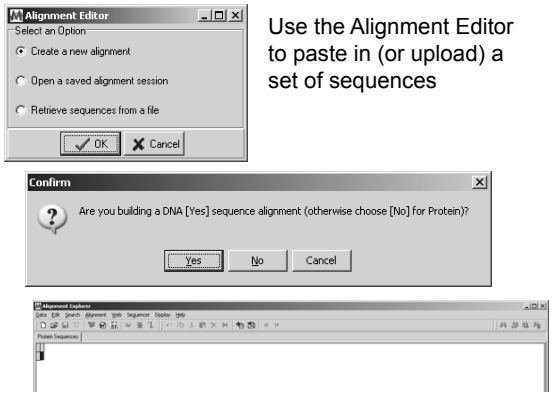
[4] Tree building

[5] Tree evaluation

Page 243

MEGA software for phylogeny:
http://www.megasoftware.net/

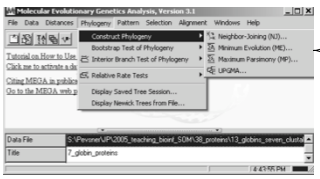




Use the Alignment Editor to paste in (or upload) a set of sequences

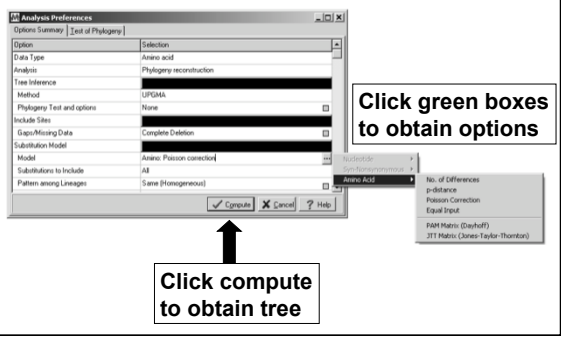
How to use MEGA to make a tree

[1] Enter a multiple sequence alignment (.meg) file
[2] Under the phylogeny menu, select one of these four methods…



Neighbor-Joining (NJ)
Minimum Evolution (ME)
Maximum Parsimony (MP)
UPGMA

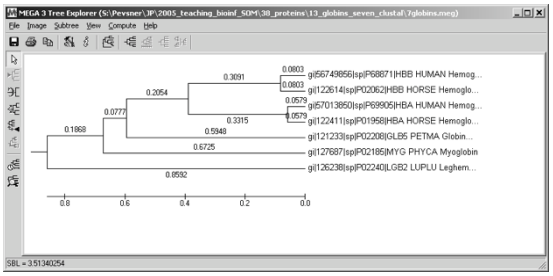
Use of MEGA for a distance-based tree: UPGMA



Click green boxes to obtain options

Click compute to obtain tree

Use of MEGA for a distance-based tree: UPGMA



Tree-building methods: UPGMA

UPGMA is
unweighted pair group method
using arithmetic mean

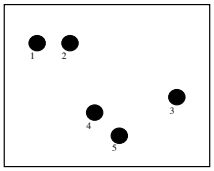


Fig. 7.26
Page 257

Tree-building methods: UPGMA

Step 1: compute the pairwise distances of all the proteins. Get ready to put the numbers 1-5 at the bottom of your new tree.

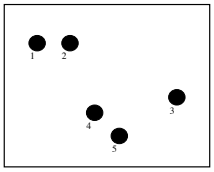


Fig. 7.26
Page 257

Tree-building methods: UPGMA

Step 2: Find the two proteins with the smallest pairwise distance. Cluster them.

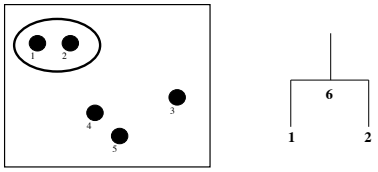


Fig. 7.26
Page 257

Tree-building methods: UPGMA

Step 3: Do it again. Find the next two proteins with the smallest pairwise distance. Cluster them.

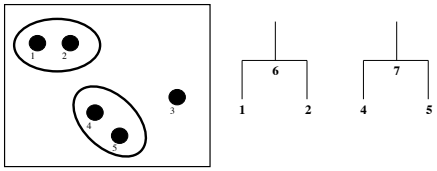


Fig. 7.26
Page 257

Tree-building methods: UPGMA

Step 4: Keep going. Cluster.

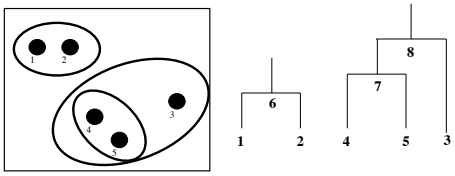


Fig. 7.26
Page 257

Tree-building methods: UPGMA

Step 4: Last cluster! This is your tree.

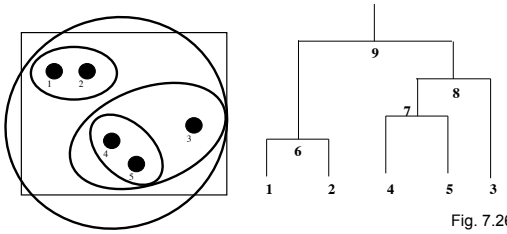


Fig. 7.26
Page 257