

Chapter 3: Pairwise Sequence Alignment

Learning objectives

- Upon completion of this material , you should be able to:
 - define homology as well as orthologs and paralog;
 - explain how PAM (accepted point mutation) matrices are derived;
 - contrast the utility of PAM and BLOSUM scoring matrices;
 - define dynamic programming and explain how global (Needleman–Wunsch) and local (Smith–Waterman) pairwise alignments are performed; and
 - perform pairwise alignment of protein or DNA sequences at the NCBI website.

Outline

Introduction

- Protein alignment: often more informative than DNA alignment

- Definitions: homology, similarity, identity

- Gaps

- Pairwise alignment, homology, and evolution of life

Scoring matrices

- Dayhoff model: 7 steps

- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local

- Global sequence alignment: algorithm of Needleman and

- Wunsch

- Local sequence alignment: Smith and Waterman algorithm

- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST

- Basic Local Alignment Search Tool (BLAST)

- Pairwise alignment with dotplots

The statistical significance of pairwise alignments

- Statistical significance of global alignments

- Percent identity and relative entropy

Perspective

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
- It is used in the analysis of genomes

Sequence 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

Example:

- searching for plant globins using human beta globin DNA yields no matches;
- searching for plant globins using human beta globin protein yields many matches

Pairwise alignment: DNA sequences can be more informative than protein

- Many times, DNA alignments are appropriate
 - to study noncoding regions of DNA
(e.g. introns or intergenic regions)
 - to study DNA polymorphisms
 - genome sequencing relies on DNA analysis

Definition: pairwise alignment

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: identity, similarity, conservation

◦ Homology

Similarity attributed to descent from a common ancestor.

Identity

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

Similarity

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

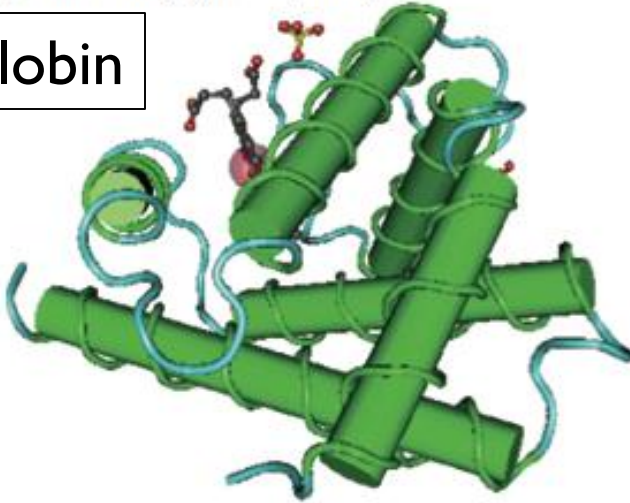
Conservation

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

Globin homologs

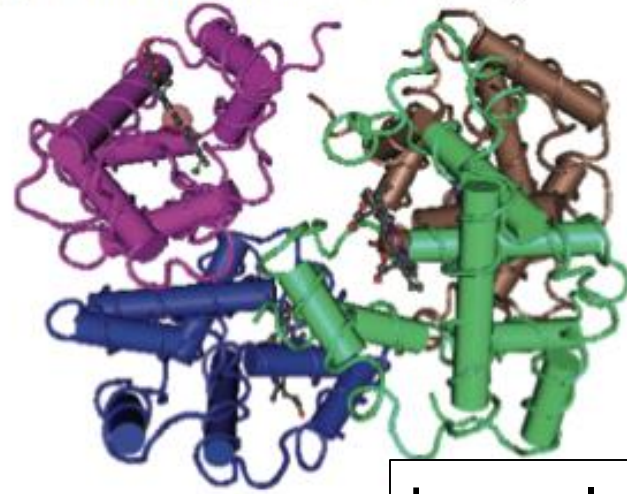
(a) Human myoglobin (3RGK)

myoglobin

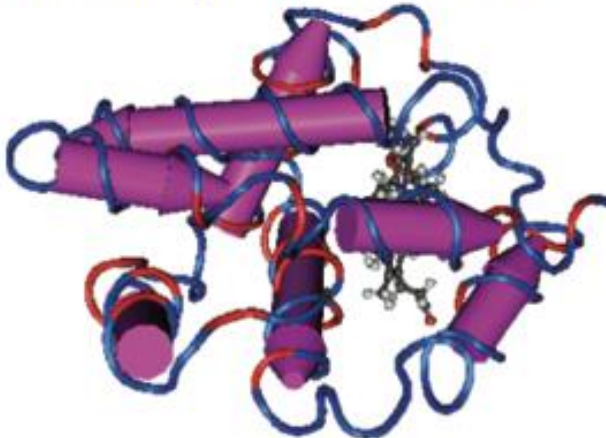


(b) Human hemoglobin tetramer (2H35)

hemoglobin

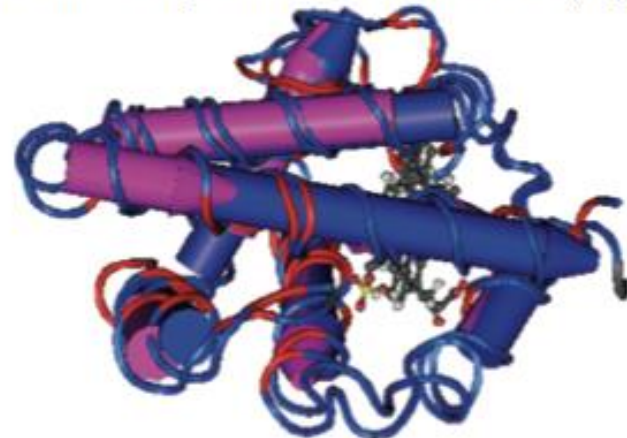


(c) Human beta globin (subunit of 2H35)



beta globin

(d) Pairwise alignment of beta globin and myoglobin



beta globin and myoglobin (aligned)

Outline

Introduction

- Protein alignment: often more informative than DNA alignment

- Definitions: homology, similarity, identity

- Gaps

- Pairwise alignment, homology, and evolution of life

Scoring matrices

- Dayhoff model: 7 steps

- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local

- Global sequence alignment: algorithm of Needleman and

- Wunsch

- Local sequence alignment: Smith and Waterman algorithm

- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST

- Basic Local Alignment Search Tool (BLAST)

- Pairwise alignment with dotplots

The statistical significance of pairwise alignments

- Statistical significance of global alignments

- Percent identity and relative entropy

Perspective

Definitions: two types of homology

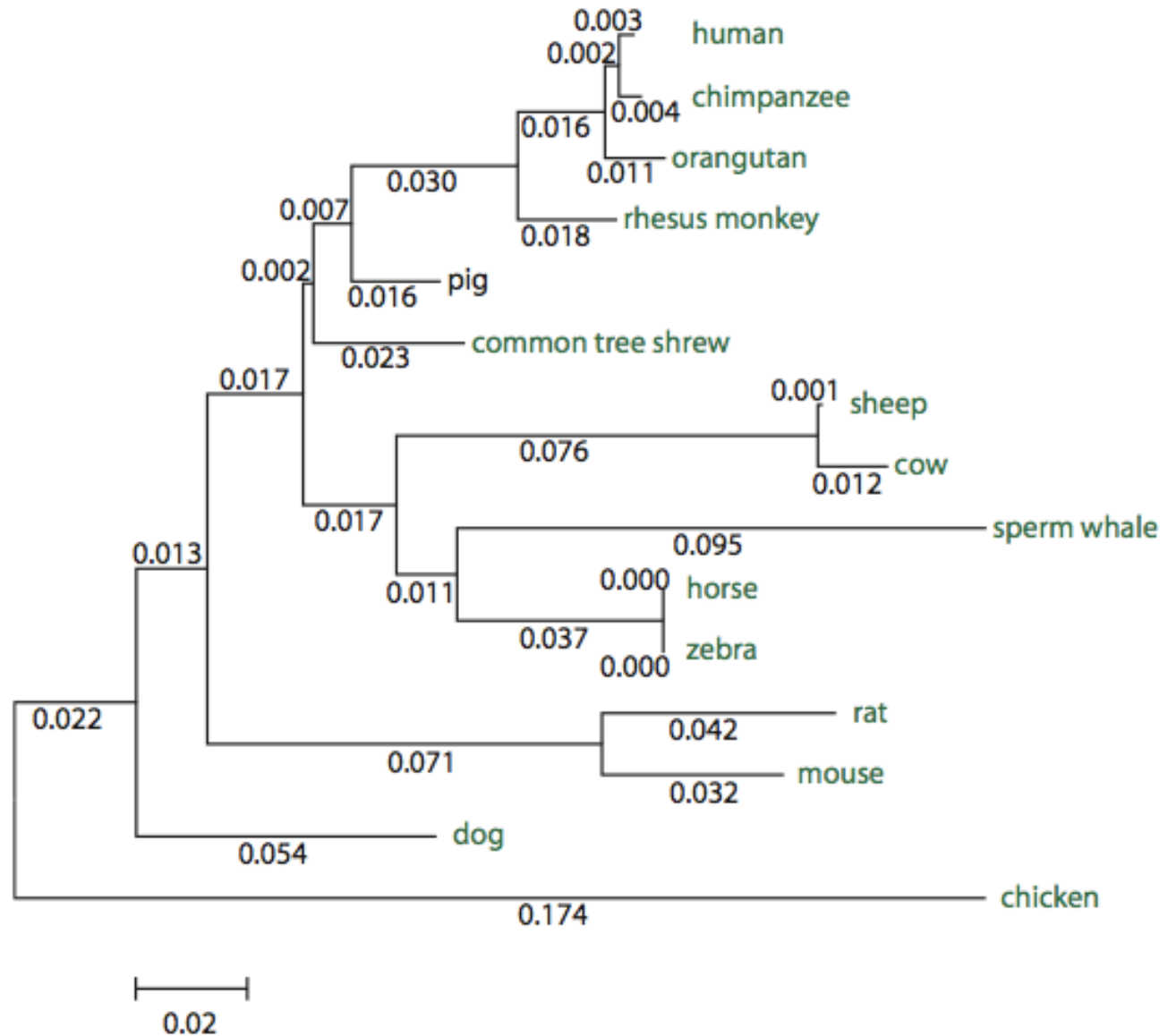
Orthologs

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

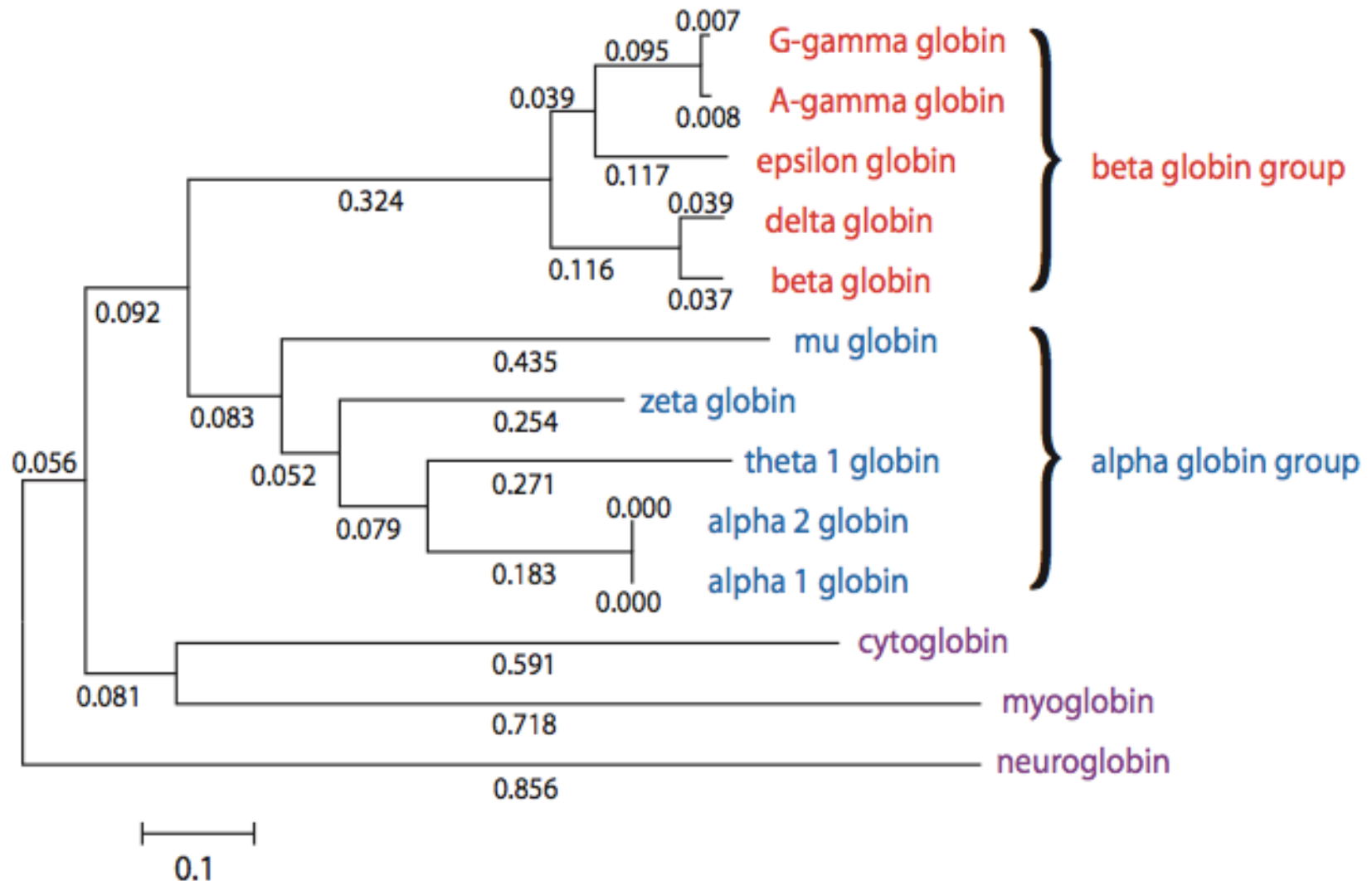
Paralogs

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

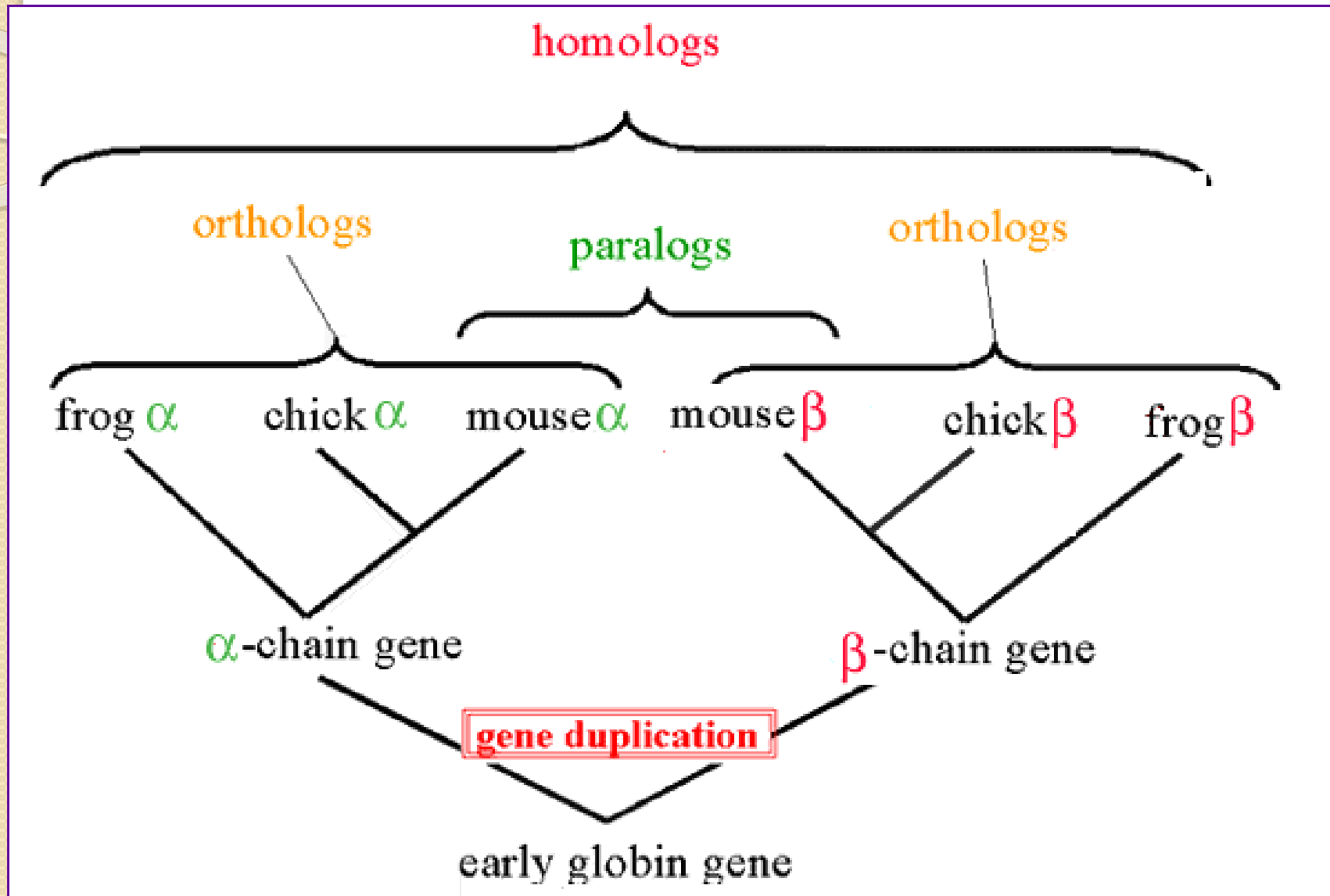
Myoglobin proteins: examples of orthologs



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



Orthologs and paralogs are often viewed in a single tree



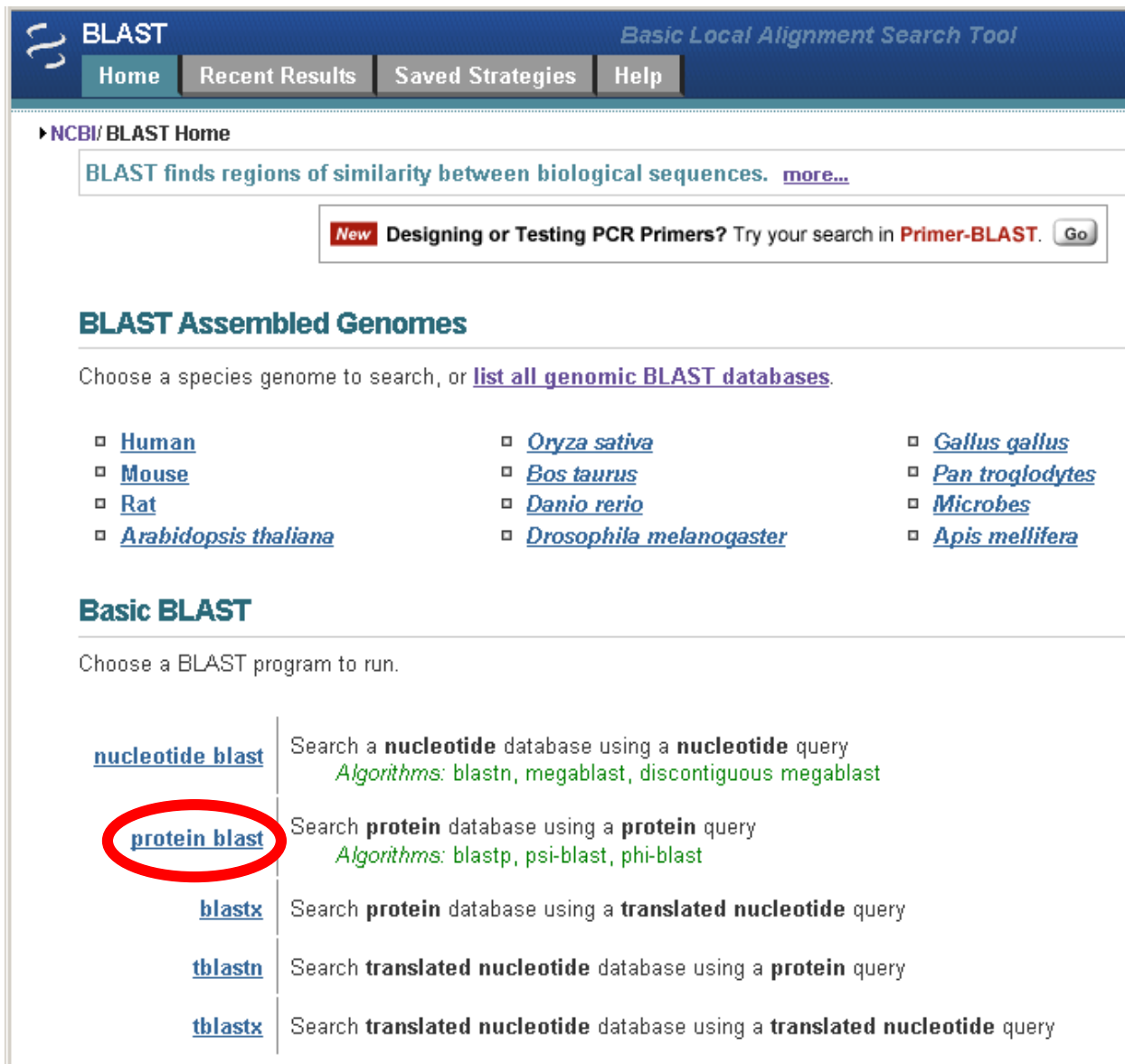
General approach to pairwise alignment

- Choose two sequences
- Select an algorithm that generates a score
- Allow gaps (insertions, deletions)
- Score reflects degree of similarity
- Alignments can be global or local
- Estimate probability that the alignment occurred by chance

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...



The screenshot shows the NCBI BLAST homepage. At the top, there's a navigation bar with 'BLAST' and 'Basic Local Alignment Search Tool'. Below this, there are tabs for 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. The main content area starts with 'NCBI/ BLAST Home' and a description: 'BLAST finds regions of similarity between biological sequences. [more...](#)'. There's a promotional banner for 'Primer-BLAST'. The 'BLAST Assembled Genomes' section lists various species genomes for selection, including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera. The 'Basic BLAST' section prompts the user to choose a BLAST program. A table lists several programs: 'nucleotide blast', 'protein blast' (circled in red), 'blastx', 'tblastn', and 'tblastx', each with a brief description and the algorithms used.

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

► NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New 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](#).

<input type="checkbox"/> Human	<input type="checkbox"/> Oryza sativa	<input type="checkbox"/> Gallus gallus
<input type="checkbox"/> Mouse	<input type="checkbox"/> Bos taurus	<input type="checkbox"/> Pan troglodytes
<input type="checkbox"/> Rat	<input type="checkbox"/> Danio rerio	<input type="checkbox"/> Microbes
<input type="checkbox"/> Arabidopsis thaliana	<input type="checkbox"/> Drosophila melanogaster	<input type="checkbox"/> Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
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 nucleotide query

BLAST *Basic Local Alignment Search Tool*



Home Recent Results Saved Strategies Help

► NCBI/BLAST/blastp suite

blastn **blastp** blastx tblastn tblastx


BLASTP programs search protein databases using a protein query

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 

Organism 
Optional

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...

https://www.ncbi.nlm.nih.gov/protein/NP_000509.1

hemoglobin subunit
beta [Homo sapiens]

And

https://www.ncbi.nlm.nih.gov/protein/np_005359

myoglobin [Homo
sapiens]

BLAST Basic Local Alignment

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite

blastn blastp blastx tblastn tblastx

BLASTP programs search protein s

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [Clear](#)

>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPREKSAVTALWCKVNVDEVGCEALGRLLVVPWTQRFESFGDLSTPDVAVMGNPKVKAH
AFSDCLAHLDNLKCTFATLSELHCDKLHVDPENFRLLGNVLCVLAHHFCKEFTPPVQAAYQKV
ALAHKYH

Or, upload file Browse...

Job Title
gi|4504349|ref|NP_000509.1| beta globin [Homo...
Enter a descriptive title for your BLAST search

☒ Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence [Clear](#)

np_005359

Or, upload file Browse...

Program Selection

Algorithm ☒ blastp (protein-protein BLAST)
Choose a BLAST algorithm

BLAST Search protein sequence using Blastp (protein-protein BLAST) ☐ 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 BLAST) ☐ Show results in a new window

Algorithm parameters [Note:](#)

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

Scoring Parameters

Matrix **BLOSUM45**

Gap Costs Existence: 13 Extension: 3

Compositional adjustments Conditional compositional score matrix adjustment

 **BLAST®**

Basic Local Alignment Search Tool

HomeRecent ResultsSaved StrategiesHelp

NCBI/BLAST/blastp suiteAlign Sequences Protein BLAST

blastblastblasttblasttblastx

BLASTP programs search protein subjects using a protein query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) 

Clear

Query subrange 

>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta (Homo sapiens)
MVHLTPEEKSAVTALVGRKVVVD EYGGKALGRLLVYVPWTORFFESFGDLSTPDAVHGKPPVCAH
GKQVVGAFSDGLAHLDNLGGTFATLSLHCDKLHVD PENFRLLGNVLVCVLAHDTGKEFTPPVQ
AAYQKVVAGVANALAHDCN

...

From

To

Or, upload file

Browse... 

Job Title

gi|4504349|ref|NP_000509.1| hemoglobin subunit...

Enter a descriptive title for your BLAST search 

☒ Align two or more sequences 

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence 

Clear

Subject subrange 

NP_005359

...

From

To

Or, upload file

Browse... 

Program Selection

Algorithm

☒ blastp (protein-protein BLAST)

Choose a BLAST algorithm 

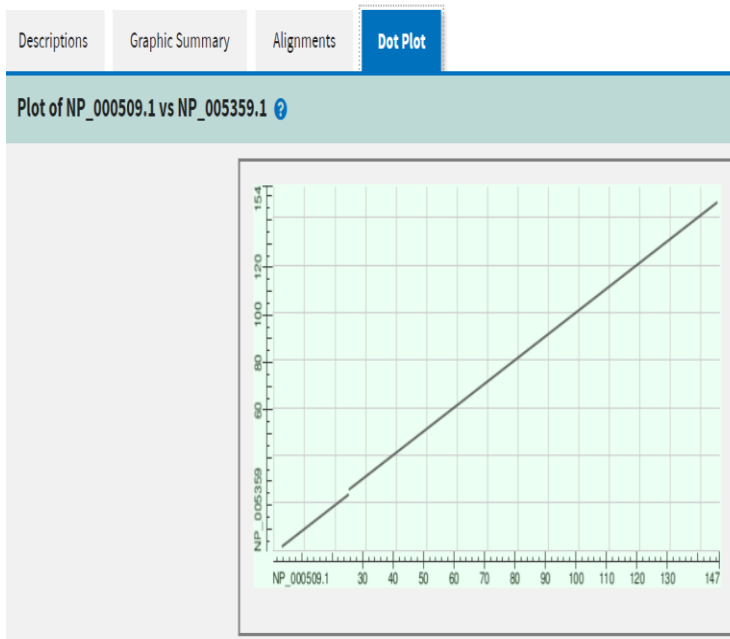
BLAST

Search protein sequence using Blastp (protein-protein BLAST)

☐ Show results in a new window

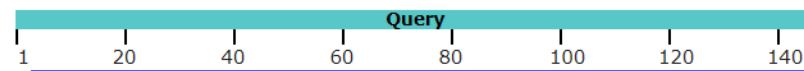
 [Algorithm parameters](#)

BLAST output



1 sequences selected ?

Distribution of the top 1 Blast Hits on 1 subject sequences



Pairwise alignment of human beta globin (the “query”) and myoglobin (the “subject”)

Score = 43.9 bits (102), Expect = 1e-09, Method: Composition-based stats.
Identities = 37/145 (25%), Positives = 57/145 (39%), Gaps = 2/145 (1%)

```

Query  4      LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
      → L+  E   V  +WGKV  D    G E L RL   +P T   F+ F  L + D +   +   +
Sbjct  3      LSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62     KAHGKKVLGAFSDGLAHLNLDNLKGTTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
      → K HG  VL A    L    + +    L++ H  K  +   +   +   ++ VL
Sbjct  63     KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

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

How raw scores are calculated: an example

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	VLNVWGKVEADIPGHGQEV LIRLF	34

match	4	11	5	6	6	5	4	5	sum of matches: +60 (round up to +61)
			6	4				4	
mismatch	-1	1	0	-2	-2	-4	0	0	sum of mismatches: -13
	-2		0	-3	0				
gap open			-11						sum of gap penalties: -13
gap extend			-2						

total raw score: 61 - 13 - 13 = 35

For a set of aligned residues we assign scores based on matches, mismatches, gap open penalties, and gap extension penalties. These scores add up to the total raw score.

		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
		GUU }	GCU }	GAU }	GGU }	U

Where do scores come from? We'll examine scoring matrices. These are related to the properties of the 20 common amino acids.

Outline

Introduction

- Protein alignment: often more informative than DNA alignment

- Definitions: homology, similarity, identity

Gaps

- Pairwise alignment, homology, and evolution of life

Scoring matrices

- Dayhoff model: 7 steps

- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local

- Global sequence alignment: algorithm of Needleman and

Wunsch

- Local sequence alignment: Smith and Waterman algorithm

- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST

- Basic Local Alignment Search Tool (BLAST)

- Pairwise alignment with dotplots

The statistical significance of pairwise alignments

- Statistical significance of global alignments

- Percent identity and relative entropy

Perspective

Gaps

- Positions at which a letter is paired with a null are called gaps.
- Gap scores are typically negative.
- Since a single mutational event may cause the insertion or deletion of more than one residue, the presence of a gap is ascribed more significance than the length of the gap. Thus there are separate penalties for gap creation and gap extension.
- In BLAST, it is rarely necessary to change gap values from the default.

Outline

Introduction

- Protein alignment: often more informative than DNA alignment

- Definitions: homology, similarity, identity

- Gaps

- Pairwise alignment, homology, and evolution of life

Scoring matrices

- Dayhoff model: 7 steps

- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local

- Global sequence alignment: algorithm of Needleman and

- Wunsch

- Local sequence alignment: Smith and Waterman algorithm

- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST

- Basic Local Alignment Search Tool (BLAST)

- Pairwise alignment with dotplots

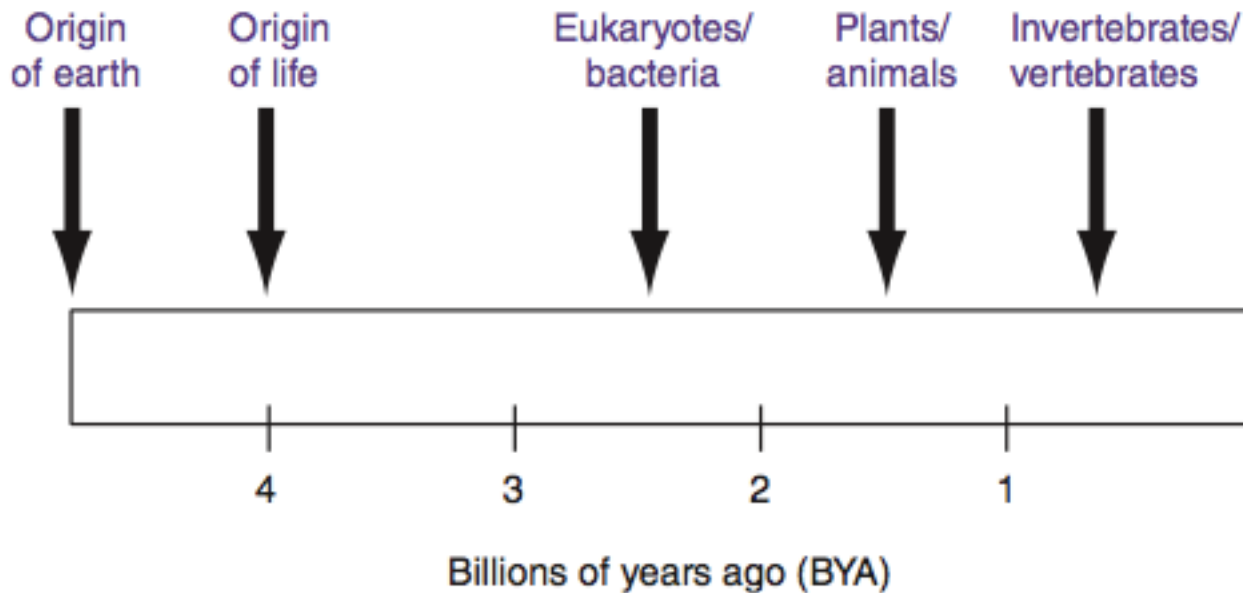
The statistical significance of pairwise alignments

- Statistical significance of global alignments

- Percent identity and relative entropy

Perspective

Pairwise alignment and the evolution of life



When two proteins (or DNA sequences) are homologous they share a common ancestor. We can infer the sequence of that ancestor. When we align globins from human and a plant we can imagine their common ancestor, a single celled organism that lived 1.5 billion years ago, and we can infer that ancient globin sequence. Through pairwise alignment we can look back in time at sequence evolution.

Outline

Introduction

- Protein alignment: often more informative than DNA alignment

- Definitions: homology, similarity, identity

- Gaps

- Pairwise alignment, homology, and evolution of life

Scoring matrices

- Dayhoff model: 7 steps

- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local

- Global sequence alignment: algorithm of Needleman and

- Wunsch

- Local sequence alignment: Smith and Waterman algorithm

- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST

- Basic Local Alignment Search Tool (BLAST)

- Pairwise alignment with dotplots

The statistical significance of pairwise alignments

- Statistical significance of global alignments

- Percent identity and relative entropy

Perspective

Step 1: Accepted point mutations (PAMs) in protein families

PROTEIN	PAMS PER 100 MILLION YEARS
Immunoglobulin (Ig) kappa chain C region	37
Kappa casein	33
Epidermal growth factor	26
Serum albumin	19
Hemoglobin alpha chain	12
Myoglobin	8.9
Nerve growth factor	8.5
Trypsin	5.9
Insulin	4.4
Cytochrome c	2.2
Glutamate dehydrogenase	0.9
Histone H3	0.14
Histone H4	0.10

Margaret Dayhoff and colleagues developed scoring matrices in the 1960s and 1970s. They defined PAMs as “accepted point mutations.” Some protein families evolve very slowly (e.g. histones change little over 100 million years); others (such as kappa casein) change very rapidly.

Dayhoff's 34 protein superfamilies

Protein PAMs per 100 million years

Ig kappa chain	37
Kappa casein	33
luteinizing hormone b	30
lactalbumin	27
complement component 3	27
epidermal growth factor	26
proopiomelanocortin	21
pancreatic ribonuclease	21
haptoglobin alpha	20
serum albumin	19
phospholipase A2, group IB	19
prolactin	17
carbonic anhydrase C	16
Hemoglobin a	12
Hemoglobin b	12

Dayhoff's 34 protein superfamilies

Protein

PAMs per 100 million years

Ig kappa chain

37

Kappa casein

33

luteinizing hormone b

30

lactalbumin

27

complement component 3

27

epidermal growth factor

26

proopiomelanocortin

21

pancreatic ribonuclease

21

haptoglobin alpha

20

human (NP_005203) versus mouse (NP_031812) kappa casein

Score = 57.8 bits (138), Expect = 3e-07

Identities = 39/118 (33%), Positives = 61/118 (51%), Gaps = 2/118 (1%)

```

Query 1  MKSFLLVVNALALTLPFLAVEVQNQKPACHENDERPFYQKTAPYVPMYYVPNSYPYYGT 60
          M++F++V+N LALTLPFLA E+QN          E ++ + ++ Y P+ V N + Y
Sbjct 2  MRNFIVVMNILALTLPFLAAEIQNPD SNCRGEKNDIVYDEQRVLYTPVRSVLN-FNQYEP 60

Query 61  NLYQRRPAI-AINNPYVPRTTYANPAVVRPHAQIPQRQYLPNSHPPTVVRLPNLHPSF 117
          N Y RP++ A +PY+          ++R A I + Q +PN          V +PSF
Sbjct 61  NYYHYRPSLPATASPYMYYP LVVRL LLLRSPAPISKWQSM PNFPQSAGVPYAIPNPSF 118
    
```

Dayhoff's 34 protein superfamilies

Protein

PAMs per 100 million years

apolipoprotein A-II	10
lysozyme	9.8
gastrin	9.8
myoglobin	8.9
nerve growth factor	8.5
myelin basic protein	7.4
thyroid stimulating hormone b	7.4
parathyroid hormone	7.3
parvalbumin	7.0
trypsin	5.9
insulin	4.4
calcitonin	4.3
arginine vasopressin	3.6
adenylate kinase I	3.2

Dayhoff's 34 protein superfamilies

<u>Protein</u>	<u>PAMs per 100 million years</u>
triosephosphate isomerase I	2.8
vasoactive intestinal peptide	2.6
glyceraldehyde phosph. dehydrogease	2.2
cytochrome c	2.2
collagen	1.7
troponin C, skeletal muscle	1.5
alpha crystallin B chain	1.5
glucagon	1.2
glutamate dehydrogenase	0.9
histone H2B, member Q	0.9
ubiquitin	0

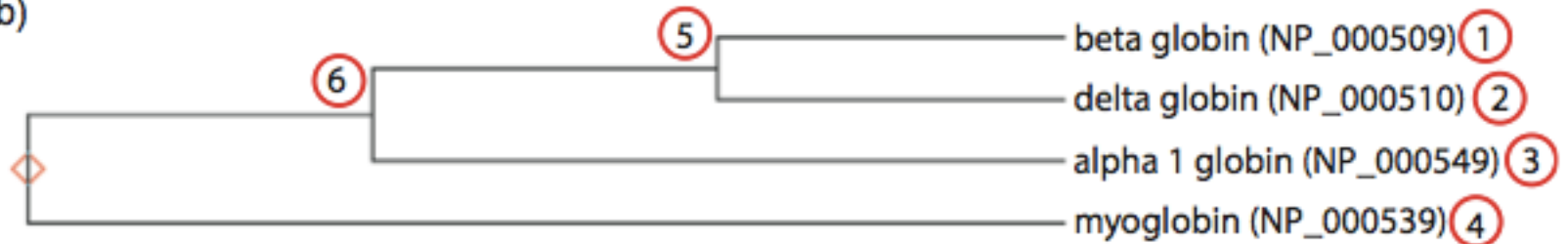
Step 1: accepted point mutations are defined not by the pairwise alignment but with respect to the common ancestor

(a)

beta globin	MVHLTPEEKSAVTALWGKV
delta globin	MVHLTPEEKTAVNALWGKV
alpha 1 globin	MV.LSPADKTNVKA AW GKV
myoglobin	.MGLSDGEWQLVLNVWGKV
5	MVHLSP EE KTAVNALWGKV
6	MVHLTP EE KTAVNALWGKV



(b)



Dayhoff et al. evaluated amino acid changes. They applied an evolutionary model to compare changes such as 1 versus 2 not to each other but to an inferred common ancestor at position 5.

Dayhoff model step 2 (of 7): Frequency of amino acids

TABLE 3.1 Normalized frequencies of amino acid. These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence.

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

If 20 amino acids occurred in nature at equal frequencies, each would be observed 5% of the time. However some are more common (G,A, L, K) and some rare (C,Y, M,W).

https://molbiol-tools.ca/Amino_acid_abbreviations.htm

Normalized frequencies of amino acids:

Gly	8.9%	Arg	4.1%
Ala	8.7%	Asn	4.0%
Leu	8.5%	Phe	4.0%
Lys	8.1%	Gln	3.8%
Ser	7.0%	Ile	3.7%
Val	6.5%	His	3.4%
Thr	5.8%	Cys	3.3%
Pro	5.1%	Tyr	3.0%
Glu	5.0%	Met	1.5%
Asp	4.7%	Trp	1.0%

Dayhoff model step 3: amino acid substitutions

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	y	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

From a survey of 1572 observed substitutions, the original amino acid (columns) are compared to the changes (rows).

Dayhoff model step 3: amino acid substitutions

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
A								
R	30							
N	109	17						
D	154	0	532					
C	33	10	0	0				
Q	93	120	50	76	0			
E	266	0	94	831	0	422		
G	579	10	156	162	10	30	112	
H	21	103	226	43	10	243	23	10

Zooming in on the previous table, note that substitutions are very common (e.g. $D \rightarrow E$, $A \rightarrow G$) while others are rare (e.g. $C \rightarrow Q$, $C \rightarrow E$). The scoring system we use for pairwise alignments should reflect these trends.

Dayhoff step 4 (of 7): Mutation probability matrix for the evolutionary distance of 1 PAM

		Original amino acid																			
		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
Replacement amino acid	A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.4	0.3	0.0	0.0	0.2
	R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
	N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0
	D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
	E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	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
	G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1
	H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.7	0.1	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.3
	L	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	99.5	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.2
	K	0.0	0.4	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	99.3	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
	M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	99.5	0.0	0.0	0.0	0.0	0.3	0.0
	P	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	99.3	0.1	0.0	0.0	0.0	0.0
	S	0.3	0.1	0.3	0.1	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.2	98.4	0.4	0.1	0.0	0.0
	T	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.3	98.7	0.0	0.0	0.1
	W	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	0.0	0.0	0.0	99.8	0.0	0.0
	Y	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	0.2	0.0	0.0	0.0	0.0	99.5	0.0
	V	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	99.0

This mutation probability matrix includes original amino acids (columns) and replacements (rows). The diagonals show that at a distance of 1 PAM most residues remain the same about 99% of the time (see shaded entries). Note how cysteine (C) and tryptophan (W) undergo few substitutions, and asparagine (N) many.

Substitution Matrix

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 true probabilities of mutations occurring through a period of evolution.

The two major types of substitution matrices are PAM and BLOSUM.



PAM matrices: Point-accepted mutations

PAM matrices are based on global alignments of closely related proteins.

The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. At an evolutionary interval of PAM1, one change has occurred over a length of 100 amino acids.

Other PAM matrices are extrapolated from PAM1. For PAM250, 250 changes have occurred for two proteins over a length of 100 amino acids.

All the PAM data come from closely related proteins (>85% amino acid identity).

Dayhoff step 4 (of 7): Mutation probability matrix for the evolutionary distance of 1 PAM

		A	R	N	D	C	Q	E	G	H
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His
amino acid	A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0
	R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1
	N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2
	D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0
	C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0
	Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2
	E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0
	G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0
	H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1
	.									

At this evolutionary distance of 1 PAM, 1% of the amino acids have diverged between each pair of sequences. The columns are percentages that sum to 100%.

Dayhoff step 5 (of 7): PAM250 and other PAM matrices

<u>NP 002037.2</u>	164	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSGKLWRDGRGALQNII	207
<u>XP 001162057.1</u>	164	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSGKLWRDGRGALQNII	207
<u>NP 001003142.1</u>	162	IHDHFGIVEGLMTTVHAIITATQKTVDGPGSGKMWRDGRGAAQNII	205
<u>XP 893121.1</u>	168	IHDNFGIMEGLMTTVHAIITATQKTVDGPGSGKLWRDGRGAAQNII	211
<u>XP 576394.1</u>	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSGKLWRDGRGAAQNII	205
<u>NP 058704.1</u>	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSGKLWRDGRGAAQNII	205
<u>XP 001070653.1</u>	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSGKLWRDGRGAAQNII	205
<u>XP 001062726.1</u>	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSGKLWRDGRGAAQNII	205
<u>NP 989636.1</u>	162	IHDNFGIVEGLMTTVHAIITATQKTVDGPGSGKLWRDGRGAAQNII	205
<u>NP 525091.1</u>	161	INDNFEIVEGLMTTVHATTATQKTVDGPGSGKLWRDGRGAAQNII	204
<u>XP 318655.2</u>	161	INDNFGILEGLMTTVHATTATQKTVDGPGSGKLWRDGRGAAQNII	204
<u>NP 508535.1</u>	170	INDNFGIIEGLMTTVHAVTATQKTVDGPGSGKLWRDGRGAGQNII	213
<u>NP 595236.1</u>	164	INDTFGIEEGLMTTVHATTATQKTVDGPSSKDWRGGRGASANII	207
<u>NP 011708.1</u>	162	INDAFGIEEGLMTTVHSLTATQKTVDGPSSHKDWRGGRTASGNII	205
<u>XP 456022.1</u>	161	INDEFGIDEALMTTVHSITATQKTVDGPSSHKDWRGGRTASGNII	204
<u>NP 001060897.1</u>	166	IHDNFGIIEGLMTTVHAIITATQKTVDGPSSSKDWRGGRAASFNII	209

Consider a multiple alignment of glyceraldehyde 3-phosphate protein sequences. Some substitutions are observed in columns (arrowheads). These give us insight into changes tolerated by natural selection.

Dayhoff step 5 (of 7): PAM250 and other PAM matrices

mouse AIPNPSFLAMPTNENQDNTAIP TIDPITPIVST--PVPTM-----ESIVNTVANPEAST
 rabbit S--HPFFMAILPNKMQDKAVTPTTNTIAAVEPT--PIPTT-----EPVVSTEVIAEASP
 sheep PHPHLSFMAIPPKKDQDKTEIPAINTIASAEPTVHSTPTT-----EAVVNAVDNPEASS
 cattle PHPHLSFMAIPPKKNQDKTEIPTINTIASGEPT--STPTT-----EAVESTVATLEDSP
 pig PRPHASFIAIPPKKNQDKTAIPAINSIATVEPT--IVPATEPIVNAEPIVNAVVTPEASS
 human PNLHPSFIAIPPKKIQDKIIIPTINTIATVEPT--PAPAT-----EPTVDSVVTPEAFS
 horse PCPHPSFIAIPPKKLQEITVIPKINTIATVEPT--PIPTP-----EPTVNNAVIPDASS
 . : *:* : : : * : * : : * : : * : : * : : * : : * : : : :

Now consider the alignment of distantly related kappa caseins. There are few conserved column positions, and many some columns (double arrowheads) have five different residues among the 7 proteins. We want to design a scoring system that is tolerant of distantly related proteins: if the scoring system is too strict then the divergent sequences may be penalized so heavily that authentic homologs are not identified or aligned.

Dayhoff step 5 (of 7): PAM250 and other PAM matrices

replacement amino acid

original amino acid

PAM0	A	R	N	D	C	Q	E	G
A	100	0	0	0	0	0	0	0
R	0	100	0	0	0	0	0	0
N	0	0	100	0	0	0	0	0
D	0	0	0	100	0	0	0	0
C	0	0	0	0	100	0	0	0
Q	0	0	0	0	0	100	0	0
E	0	0	0	0	0	0	100	0
G	0	0	0	0	0	0	0	100

original amino acid

PAM ∞	A	R	N	D	C	Q	E	G
A	8.7	8.7	8.7	8.7	8.7	8.7	8.7	8.7
R	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1
N	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
D	4.7	4.7	4.7	4.7	4.7	4.7	4.7	4.7
C	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
Q	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8
E	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
G	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9

At the extreme of perfectly conserved proteins (PAM0) there are no amino acid replacements. At the extreme of completely diverged proteins (PAM ∞) the matrix converges on the background frequencies of the amino acids.

PAM250 matrix: for proteins that share ~20% identity

		Original amino acid																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Replacement amino acid	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

Compare this to a PAM1 matrix, and note the diagonal still has high scores but much information content is lost.

Dayhoff step 6 (of 7): from a mutation probability matrix to a relatedness odds matrix

$$R_{ij} = \frac{M_{ij}}{f_i}$$

A relatedness odds matrix reports the probability that amino acid j will change to i in a homologous sequence.

The numerator models the observed change. The denominator f_i is the probability of amino acid residue i occurring in the second sequence by chance.

A positive value indicates a replacement happens more often than expected by chance. A negative value indicates the replacement is not favored.

Why do we go from a mutation probability matrix to a log odds matrix?

- We want a scoring matrix so that when we do a pairwise alignment (or a BLAST search) we know what score to assign to two aligned amino acid residues.
- Logarithms are easier to use for a scoring system. They allow us to sum the scores of aligned residues (rather than having to multiply them).

Log-odds matrix for PAM250

6									
-3	5								
4	0	6							
2	-5	0	9						
-3	-1	-2	-5	6					
-3	0	-2	-3	1	2				
-2	0	-1	-3	0	1	3			
-2	-3	-4	0	-6	-2	-5	17		
-1	-4	-2	7	-5	-3	-3	0	10	
2	-2	2	-1	-1	-1	0	-6	-2	4
L	K	M	F	P	S	T	W	Y	V

This is a useful matrix for comparing distantly related proteins. Note that an alignment of two tryptophan (W) residues earns +17 and a W to T mismatch is -5.

This is an example of a scoring matrix with “severe” penalties. A match of W to W earns +13, but a mismatch (e.g. W aligned to T) has a score of -19, far lower than in PAM250.

BLOSUM62 scoring matrix

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

BL62 is the default scoring matrix at the NCBI BLAST site.

BLOSUM Matrices

BLOSUM matrices are based on local alignments.

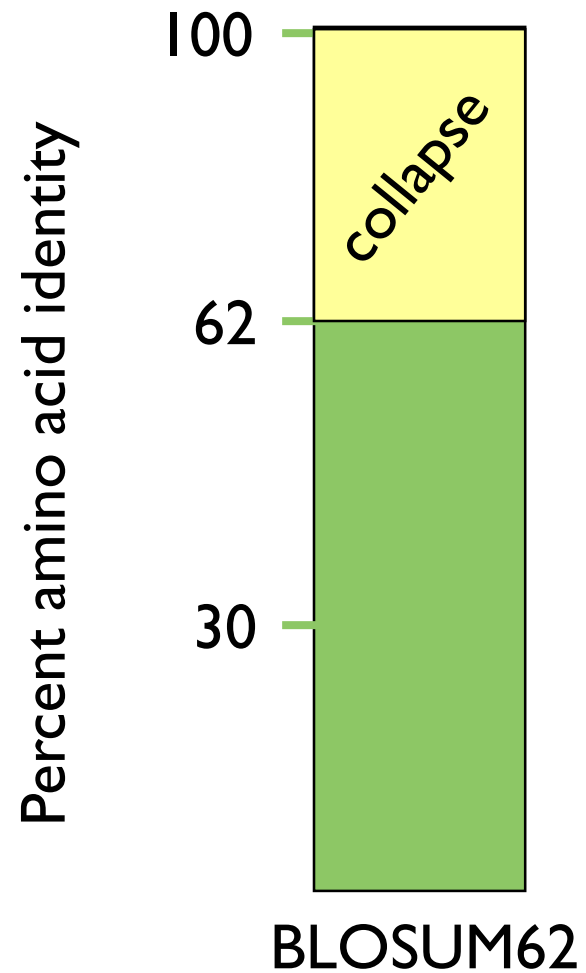
All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins.

BLOSUM stands for blocks substitution matrix.

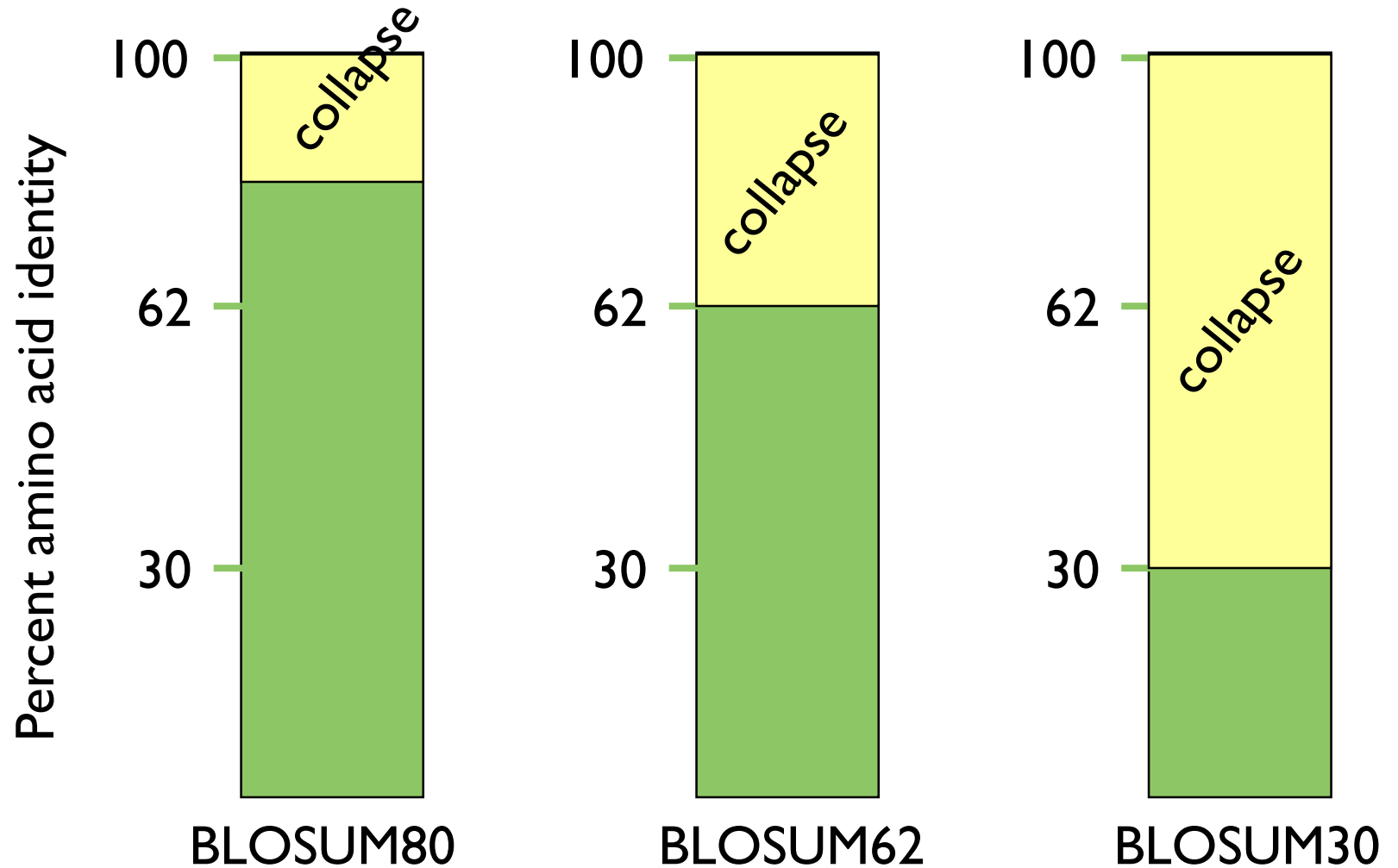
BLOSUM62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.

BLOSUM62 is the default matrix in BLAST 2.0.

BLOSUM Matrices



BLOSUM Matrices



Summary of PAM and BLOSUM matrices

BLOSUM90

BLOSUM62

BLOSUM45

PAM30

PAM120

PAM250

Less divergent



More divergent

Human versus
chimpanzee beta globin

Human versus
bacterial globins

A higher PAM number, and a lower BLOSUM number, tends to correspond to a matrix tuned to more divergent proteins.

Outline

Introduction

- Protein alignment: often more informative than DNA alignment

- Definitions: homology, similarity, identity

- Gaps

- Pairwise alignment, homology, and evolution of life

Scoring matrices

- Dayhoff model: 7 steps

- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local

- Global sequence alignment: algorithm of Needleman and

Wunsch

- Local sequence alignment: Smith and Waterman algorithm

- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST

- Basic Local Alignment Search Tool (BLAST)

- Pairwise alignment with dotplots

The statistical significance of pairwise alignments

- Statistical significance of global alignments

- Percent identity and relative entropy

Perspective

Two kinds of sequence alignment: global and local

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

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

BLAST, a heuristic version of Smith-Waterman.

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)

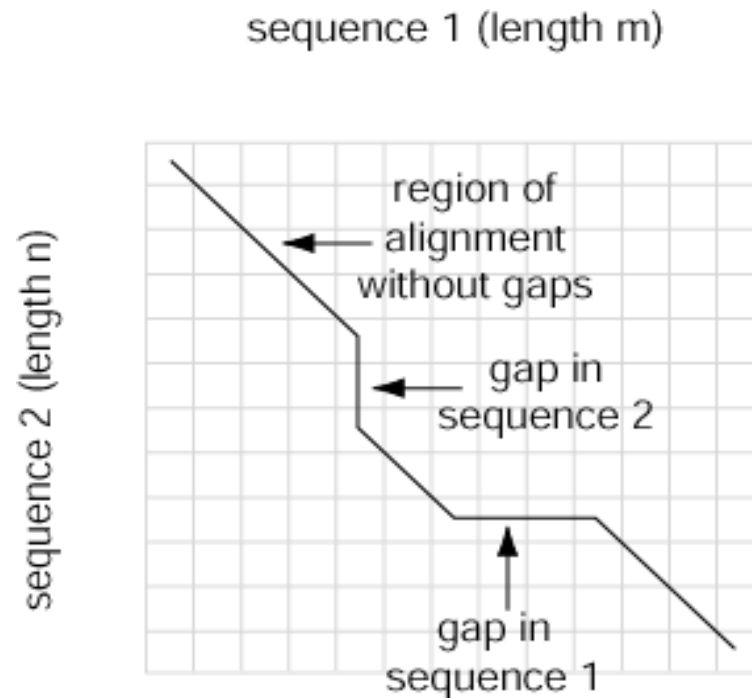
Three steps to global alignment with the Needleman-Wunsch algorithm

[1] set up a matrix

[2] score the matrix

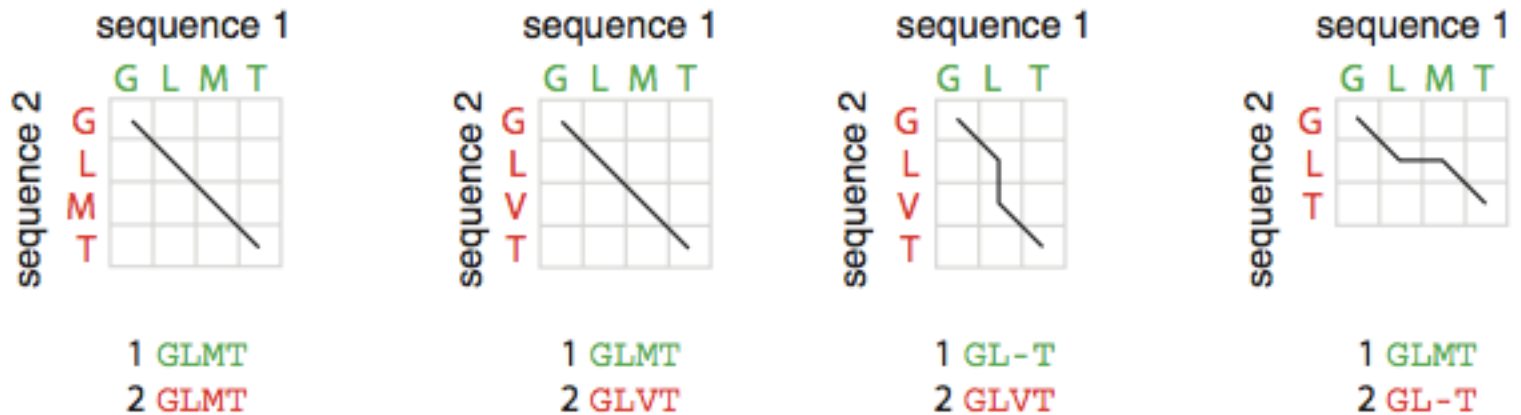
[3] identify the optimal alignment(s)

Four possible outcomes in aligning two sequences



- [1] identity (stay along a diagonal)
- [2] mismatch (stay along a diagonal)
- [3] gap in one sequence (move vertically!)
- [4] gap in the other sequence (move horizontally!)

Four possible outcomes in aligning two sequences



Global pairwise alignment using Needleman-Wunsch

(a)

c)

		Sequence 2								
		F	M	D	T	P	L	N	E	
Sequence 1		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								

Identify positions of identity (shaded gray).

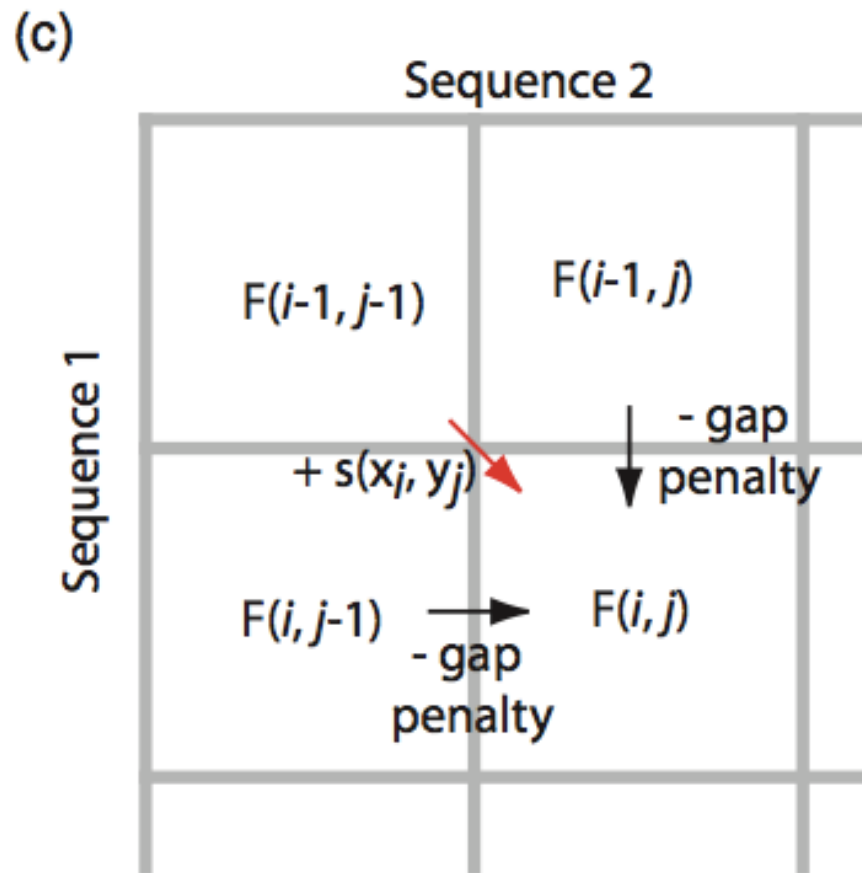
Global pairwise alignment using Needleman-Wunsch

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

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

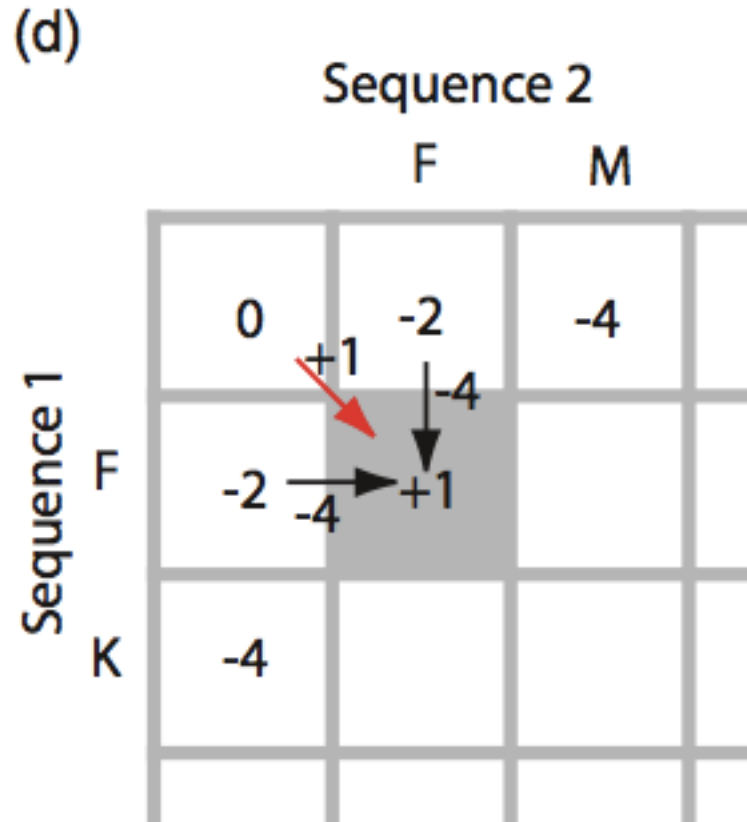
Define an overall score that maximizes cumulative scores at each position of the pairwise alignment, allowing for substitutions and gaps in either sequence.

Global pairwise alignment using Needleman-Wunsch



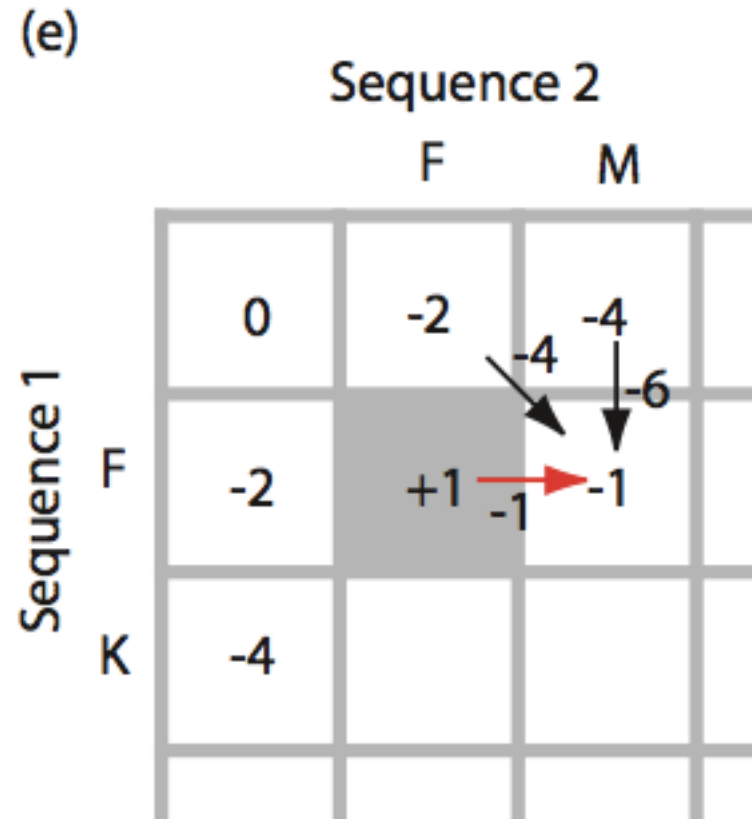
To decide how to align sequences 1 and 2 in the box at lower right, decide what the scores are beginning at upper left (not requiring a gap), or beginning from the left or top (each requiring a gap penalty).

Global pairwise alignment using Needleman-Wunsch



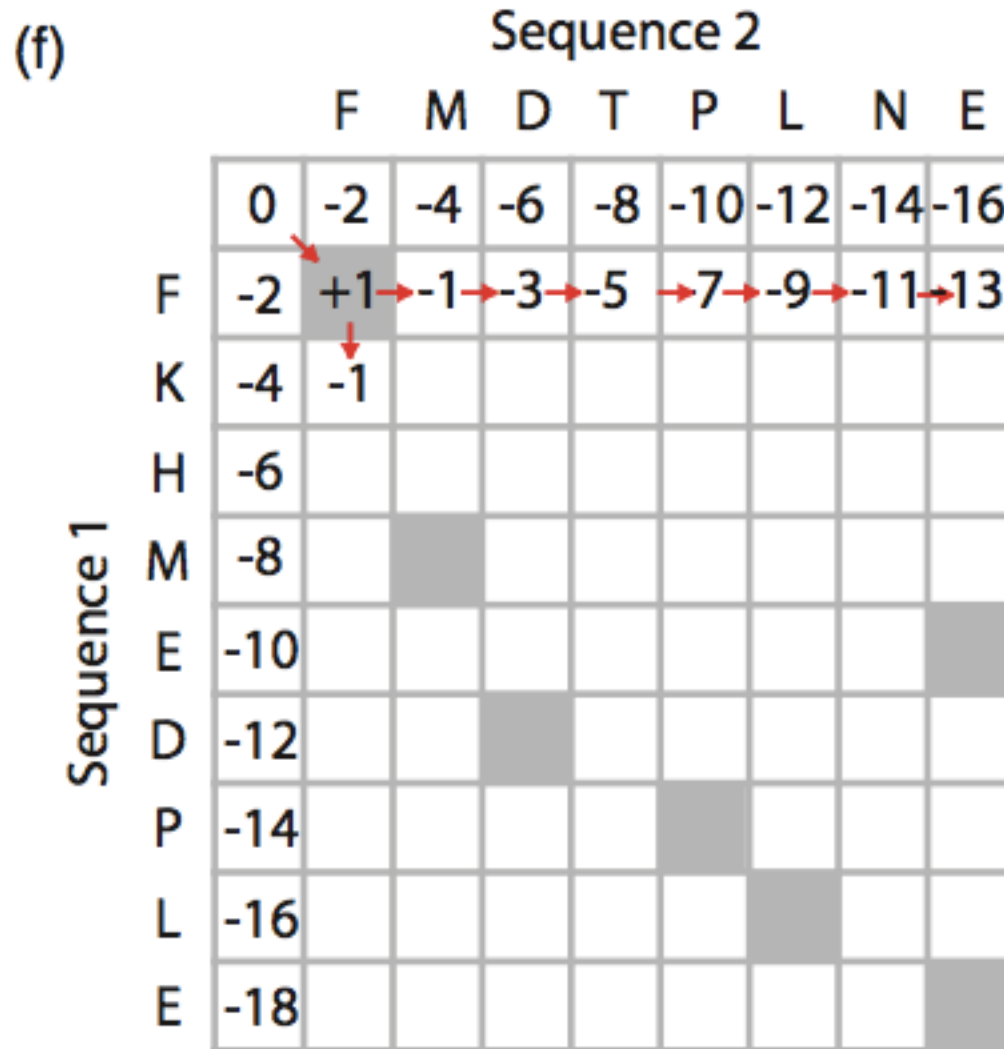
Here the best score involves +1 (proceed from upper left to gray, lower right square). If we instead select an alignment involving a gap the score would be worse (-4).

Global pairwise alignment using Needleman-Wunsch



Proceed to calculate the optimal score for the next position.

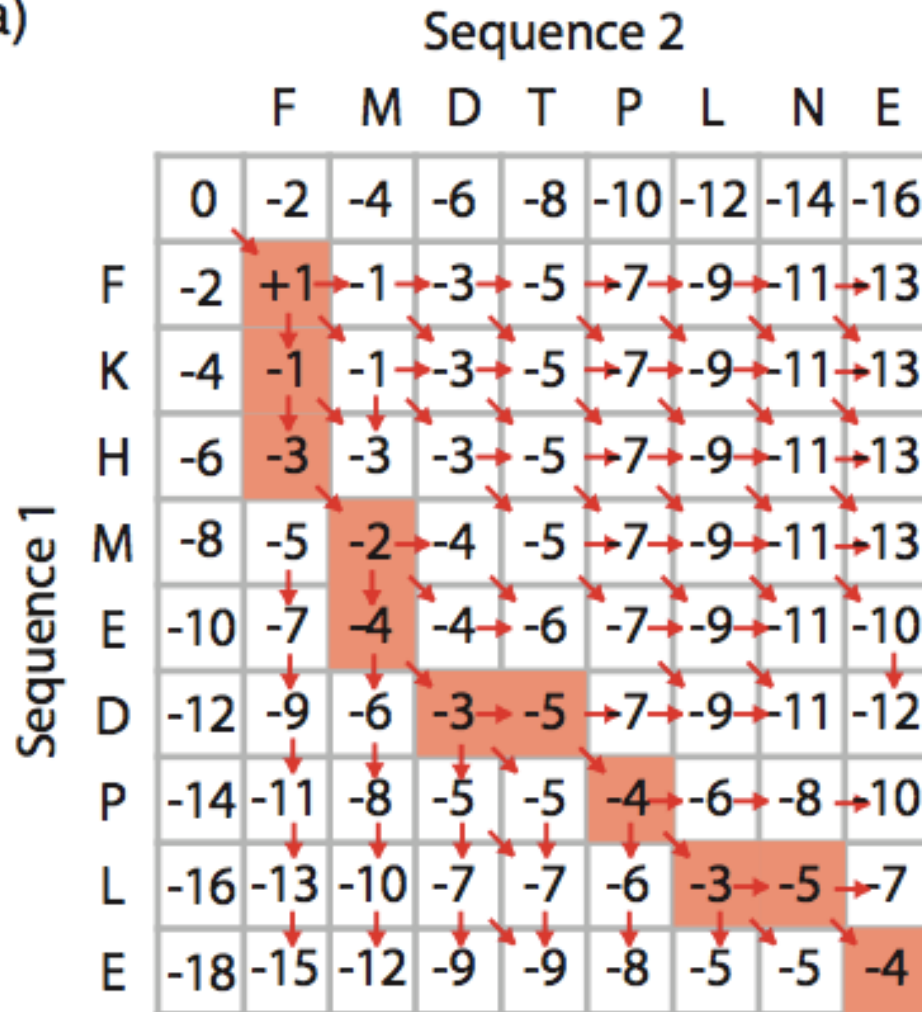
Global pairwise alignment using Needleman-Wunsch



Continue filling in the matrix.

Global pairwise alignment using Needleman-Wunsch

(a)



Highlighted cells indicate the optimal path (best scores), indicating how the two sequences should be aligned.

Global pairwise alignment using Needleman-Wunsch

(b)

		Sequence 2								
		F	M	D	T	P	L	N	E	
Sequence 1		0	-2	-4	-6	-8	-10	-12	-14	-16
	F	-2	+1	-1	-3	-5	-7	-9	-11	-13
	K	-4	-1	-1	-3	-5	-7	-9	-11	-13
	H	-6	-3	-3	-3	-5	-7	-9	-11	-13
	M	-8	-5	-2	-4	-5	-7	-9	-11	-13
	E	-10	-7	-4	-4	-6	-7	-9	-11	-10
	D	-12	-9	-6	-3	-5	-7	-9	-11	-12
	P	-14	-11	-8	-5	-5	-4	-6	-8	-10
	L	-16	-13	-10	-7	-7	-6	-3	-5	-7
	E	-18	-15	-12	-9	-9	-8	-5	-5	-4

Equivalent representation, showing the traceback procedure: begin at the lower right cell and proceed back to the start.

Global pairwise alignment using Needleman-Wunsch

(b)

		Sequence 2								
		F	M	D	T	P	L	N	E	
Sequence 1		0	-2	-4	-6	-8	-10	-12	-14	-16
	F	-2	+1	-1	-3	-5	-7	-9	-11	-13
	K	-4	-1	-1	-3	-5	-7	-9	-11	-13
	H	-6	-3	-3	-5	-7	-9	-11	-13	-13
	M	-8	-5	-2	-4	-5	-7	-9	-11	-13
	E	-10	-7	-4	-6	-7	-9	-11	-10	-10
	D	-12	-9	-6	-3	-5	-7	-9	-11	-12
	P	-14	-11	-8	-5	-5	-4	-6	-8	-10
	L	-16	-13	-10	-7	-7	-6	-3	-5	-7
	E	-18	-15	-12	-9	-9	-8	-5	-5	-4

	↖	↑	↑	↖	↑	↖	←	↖	↖	←	↖
	+1	-1	-3	-2	-4	-3	-5	-4	-3	-5	-4
Sequence 1	F	K	H	M	E	D	-	P	L	-	E
Sequence 2	F	-	-	M	-	D	T	P	L	N	E

Equivalent representation, showing the traceback procedure: begin at the lower right cell and proceed back to the start.

Needleman-Wunsch: dynamic programming

N-W is guaranteed to find optimal alignments, although the algorithm does not search all possible alignments.

It is an example of a dynamic programming algorithm: an optimal path (alignment) is identified by incrementally extending optimal subpaths.

Thus, a series of decisions is made at each step of the alignment to find the pair of residues with the best score.

Global alignment versus local alignment

Global alignment (Needleman-Wunsch) extends from one end of each sequence to the other.

Local alignment finds optimally matching regions within two sequences (“subsequences”).

Local alignment is almost always used for database searches such as BLAST. It is useful to find domains (or limited regions of homology) within sequences.

Smith and Waterman (1981) solved the problem of performing optimal local sequence alignment. Other methods (BLAST, FASTA) are faster but less thorough.

Global alignment (top) includes matches ignored by local alignment (bottom)

(a)

NP_824492.1	1	MCGDMTVHTVEYIRYRIPEQQSAEFLAAYTRAAQLAAAPQCV DYELARC	50
NP_337032.1	1		0
NP_824492.1	51	EEDFEHFVLRITWTSTEDHIEGFRKSELPDFLAEIRPYISSIEEMRHYK	100
NP_337032.1	1		0
NP_824492.1	101	PTTVRGTTGAAPVPTLYAWAGGAEAFARL TEVFYEKVLKDDVLAPVFEGMAP	150
NP_337032.1	1	MEGMDQMPKSFYDAVGGAKTFDAIVSRFYAQVAEDEV LRRVY----P	43
NP_824492.1	151	EH-----AAHVALWLGEVFGGPAAYSETQGGHGHMVAKHLGKNITEVQRR	195
NP_337032.1	44	EDDLAGAEERLRMFLEQYWGGRPTYSE-QRGHPRLMRHAPFRISLIERD	92
NP_824492.1	196	RWVNLLQDAADDAGLPT-DAEFRSAFLAYAEWGTRLAVYFSGPD AVPPAE	244
NP_337032.1	93	AWLRCMHTAVASIDSETLDDEHRRELLDYLEMAAHS LV--NSPF	134
NP_824492.1	245	QPVPQWSWGAMPPYQP	260
NP_337032.1	135		134

Global:
15% identity

(b)

NP_824492.1	113	TLYAWAGGAEAFARL TEVFYEKVLKDDVLAPVFEGMAPEH-----AAHVA	157
NP_337032.1	10	SFYDAVGGAKTFDAIVSRFYAQVAEDEV LRRVY----PEDDLAGAEERLR	55
NP_824492.1	158	LWLGEVFGGPAAYSETQGGHGHMVAKHLGKNITEVQRRRWVNLLQDAADD	207
NP_337032.1	56	MFLEQYWGGRPTYSE-QRGHPRLMRHAPFRISLIERDAWLRCMHTAVAS	104
NP_824492.1	208	AGLPT-DAEFRSAFLAYAE	225
NP_337032.1	105	IDSETLDDEHRRELLDYLE	123

Local:
30% identity

NP_824492, NP_337032

Outline

Introduction

- Protein alignment: often more informative than DNA alignment

- Definitions: homology, similarity, identity

- Gaps

- Pairwise alignment, homology, and evolution of life

Scoring matrices

- Dayhoff model: 7 steps

- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local

- Global sequence alignment: algorithm of Needleman and

- Wunsch

- Local sequence alignment: Smith and Waterman algorithm

- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST

- Basic Local Alignment Search Tool (BLAST)

- Pairwise alignment with dotplots

The statistical significance of pairwise alignments

- Statistical significance of global alignments

- Percent identity and relative entropy

Perspective

How the Smith-Waterman algorithm works

Set up a matrix between two proteins (size $m+1, n+1$)

No values in the scoring matrix can be negative! $S \geq 0$

The score in each cell is the maximum of four values:

[1] $s(i-1, j-1) + \text{the new score at } [i,j]$ (a match or mismatch)

[2] $s(i, j-1) - \text{gap penalty}$

[3] $s(i-1, j) - \text{gap penalty}$

[4] zero ← this is not in Needleman-Wunsch

Where to use the Smith-Waterman algorithm

[1] Galaxy offers “needle” and “water” EMBOSS programs.

[2] EBI offers needle and water.
<http://www.ebi.ac.uk/Tools/psa/>

[3] Try using SSEARCH to perform a rigorous Smith-Waterman local alignment:
<http://fasta.bioch.virginia.edu/>

[4] Next-generation sequence aligners incorporate Smith-Waterman in some specialized steps.

Rapid, heuristic versions of Smith-Waterman: FASTA and 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).

Gotoh (1982) and Myers and Miller (1988) improved the algorithms so both global and local alignment require less time and space.

FASTA and BLAST provide rapid alternatives to S-W.

Outline

Introduction

- Protein alignment: often more informative than DNA alignment

- Definitions: homology, similarity, identity

- Gaps

- Pairwise alignment, homology, and evolution of life

Scoring matrices

- Dayhoff model: 7 steps

- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local

- Global sequence alignment: algorithm of Needleman and

- Wunsch

- Local sequence alignment: Smith and Waterman algorithm

- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST

- Basic Local Alignment Search Tool (BLAST)

- Pairwise alignment with dotplots

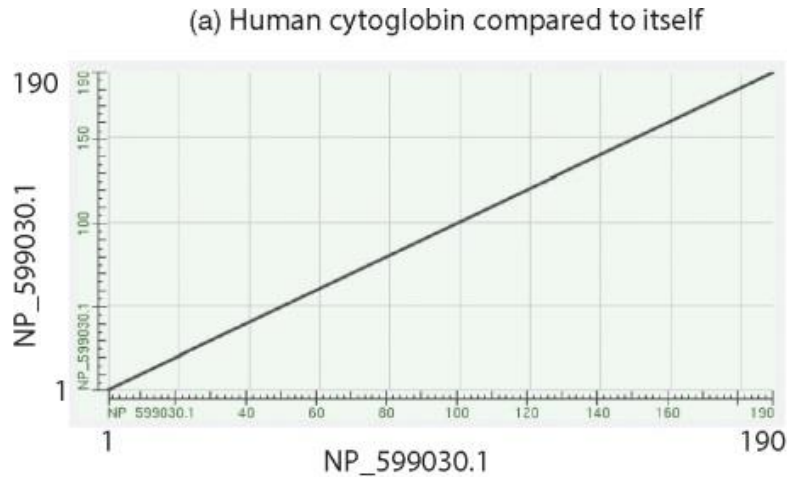
The statistical significance of pairwise alignments

- Statistical significance of global alignments

- Percent identity and relative entropy

Perspective

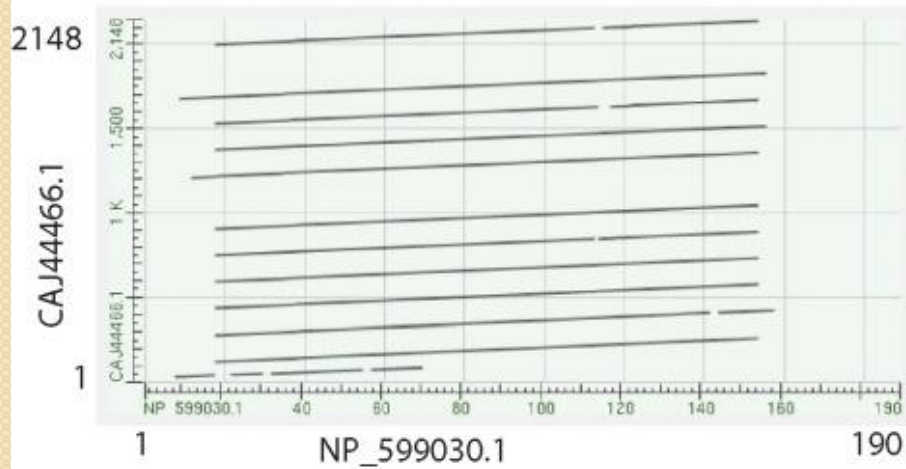
Pairwise alignment with dotplots



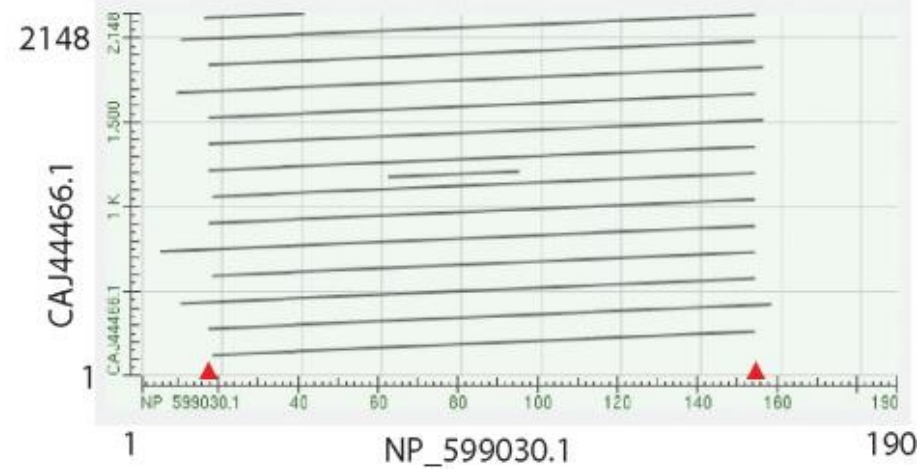
A human globin searched against itself produces a unit diagonal on a dot plot (NCBI BLASTP, aligning 2 sequences).

Pairwise alignment with dotplots

(b) Cytoglobin compared to a snail globin (BLOSUM62)



(c) Cytoglobin compared to a snail globin (PAM250)



Search human cytoglobin against a large snail globin (having many globin repeats). More repeats are observed using PAM250 than BLOSUM62.

To “read” this plot note that cytoglobin (x-axis) matches the snail globin (y-axis) at about a dozen locations across the snail protein. Red arrows indicate that the first few and last few amino acids of cytoglobin do not participate in this repeat structure.

Pairwise alignment with dotplots

haemoglobin type 1 [Biomphalaria glabrata]

Sequence ID: [emb|CAJ44466.1|](#) Length: 2148 Number of Matches: 15

Range 1: 1529 to 1669 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
55.0 bits(189)	4e-13	Composition-based stats.	36/141(26%)	83/141(58%)	4/141(2%)
Query 18	ELSEAERKAVQAMNARLYANCEDV---GVAILVRFFVNFPSAKQYFSQFKHMEDPLEMER	74	LSE++R+A+++ W RL A ++V GV ++++FF N+P+ ++ F++F + +		
Sbjct 1529	GLSETDRRALDSSWKRLTAGENGVOQKAGVNLVLWFFNNIPNMRERFTKFDANQADDALRA	1588			
Query 75	SPQLRKHACRVMGALNTVVENLHDPDKVSSVLALVGKAH-ALKHKVEPVYFKILSGVILE	133	P+++K+ ++G+L++ +++++DP + + + V+ AH ++ V YF LS I		
Sbjct 1589	DPEFQKQVNVIVGGLKSFLDVNDPIALQANMDRVAEHLSDPVGVPYFSALSQNIHR	1648			
Query 134	VVAEEFASDFPPETQRANAKL	154	+ ++ ++ +AW+ L		
Sbjct 1649	FIEISLGVTADSDESQANTDL	1669			

BLASTP output includes the various sequence alignments. One is shown here: human cytoglobin (residues 18-154) aligns to the snail globin (at residues 1529-1669). The expect value is convincing ($4e-13$), and this is one of a dozen sequence alignments.

Conclusion: the dotplot is an excellent way to visualize complex repeats.

Outline

Introduction

- Protein alignment: often more informative than DNA alignment

- Definitions: homology, similarity, identity

- Gaps

- Pairwise alignment, homology, and evolution of life

Scoring matrices

- Dayhoff model: 7 steps

- Pairwise alignment and limits of detection: the “twilight zone”

Alignment algorithms: global and local

- Global sequence alignment: algorithm of Needleman and

- Wunsch

- Local sequence alignment: Smith and Waterman algorithm

- Rapid, heuristic versions of Smith–Waterman: FASTA and BLAST

- Basic Local Alignment Search Tool (BLAST)

- Pairwise alignment with dotplots

The statistical significance of pairwise alignments

- Statistical significance of global alignments

- Percent identity and relative entropy

Perspective

Statistical significance of pairwise alignments

Information based on a "gold standard" (e.g. 3D structure)

	sequences are homologous	sequences are not homologous	
alignment result: sequences reported as related	True positives (TP)	False positives (FP)	All positives
alignment result: sequences reported as not related (or, sequences not reported)	False negative (FN)	True negative (TN)	All negatives

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

Statistical significance of pairwise alignments

The statistical significance of global alignments is not well described. We can apply a z-score.

$$Z = \frac{x - \mu}{s}$$

For local alignment the statistical significance is thoroughly understood.

Perspective

Pairwise alignment is a fundamental problem in bioinformatics. We discussed concepts of homology, and global versus local alignment (e.g. Needleman-Wunsch versus Smith-Waterman algorithms).

Substituent residue
(Percentage of total residue sites at which the substituent occurs)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A				28			31	33								31				
R								50			58				25					
N	33			47				33			33				33	33				
D	44		22			47	34	22			28				25					
C	(66)																			
Q				56		30		40			70									
E	50			44			38				41			24						
G	51			33		30					27					36				
H				26							26	30				22	22			
I	39										58									46
L	21									23	23		28							30
K	23	21		28			31	23			21				21					
M	22									22	89			22						45
F								22		61										
P	50			43			57	43			21									
S	49			24			24	36			24					40				
T	32						28	24			24					52				
W	(40)									(40)		(60)								
Y							(33)				(50)									
V	36									21	43	21								

Sequence (original amino acid)

We end with a remarkable scoring matrix reported by Zuckerkandl and Pauling in 1965, soon after the very first protein sequences were identified. While the data set was very sparse, these authors already found patterns of amino acid substitutions that occur in nature.