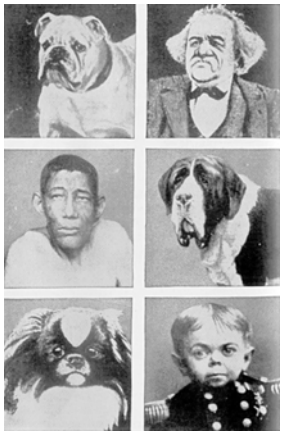
November 22, 2010

Pairwise sequence alignment



Jonathan Pevsner, Ph.D.
Bioinformatics
Johns Hopkins School Med.

Copyright notice

Announcements

The moodle quiz from lecture 1 is due one week later—by today at noon. After then the quiz "closes" and won't be available to you.

The quiz from today's lecture ("opens" at 10:30 am) is due in one week later at noon. Because of the Thanksgiving break, I'm extending the deadline a day to Tuesday November 30 (5:00 pm).

Outline: pairwise alignment

- Overview and examples
- Definitions: homologs, paralogs, orthologs
- Assigning scores to aligned amino acids: Dayhoff's PAM matrices
- Alignment algorithms: Needleman-Wunsch, Smith-Waterman

Learning objectives

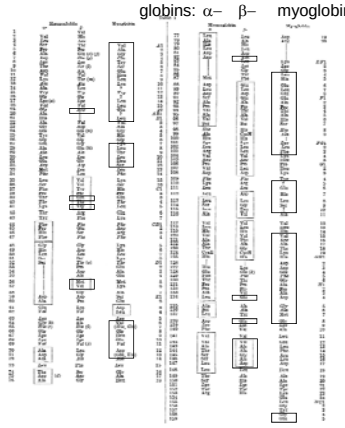
- Define homologs, paralogs, orthologs
- Perform pairwise alignments (NCBI BLAST)
- Understand how scores are assigned to aligned amino acids using Dayhoff's PAM matrices
- Explain how the Needleman-Wunsch algorithm performs global pairwise alignments

Pairwise alignments in the 1950s

β-corticotropin (sheep)	`ala gly glu asp asp glu`
Corticotropin A (pig)	`asp gly ala glu asp glu`
Oxytocin	`CYIQNCPLG`
Vasopressin	`CYFQNCPRG`

Page 46

globins: α– β– myoglobin



Early example of sequence alignment: globins (1961)

H.C. Watson and J.C. Kendrew, "Comparison Between the Amino-Acid Sequences of Sperm Whale Myoglobin and of Human Hæmoglobin." *Nature* 190:670-672, 1961.

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 week)
- It is used in the analysis of genomes

Page 47

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

Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Page 54

Pairwise alignment: protein sequences can be more informative than DNA

- Many times, DNA alignments are appropriate
  - --to confirm the identity of a cDNA
  - --to study noncoding regions of DNA
  - --to study DNA polymorphisms
  - --example: Neanderthal vs modern human DNA

```
Query: 181 catcaactacaactccaaagacacccttacacccactaggatatcaacaaacctacccac 240
           |||||||| |||| |||||| ||||| | ||||||||||||||||||||||||||||||
Sbjct: 189 catcaactgcaaccccaaagccaccct-cacccactaggatatcaacaaacctacccac 247
```

Outline: pairwise alignment

- Overview and examples
- Definitions: homologs, paralogs, orthologs
- Assigning scores to aligned amino acids: Dayhoff's PAM matrices
- Alignment algorithms: Needleman-Wunsch, Smith-Waterman

Definition: pairwise alignment

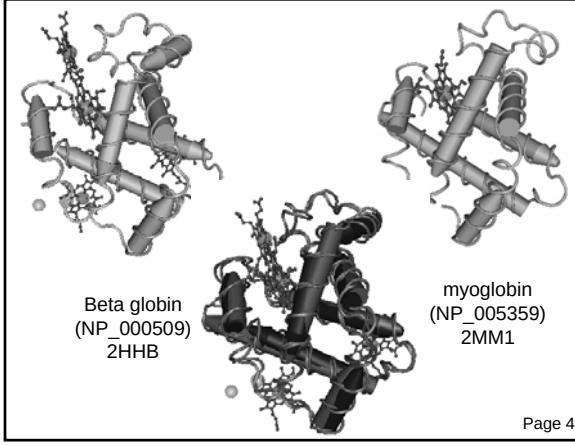
Pairwise alignment
The process of lining up two sequences
to achieve maximal levels of identity
(and conservation, in the case of amino acid sequences)
for the purpose of assessing the degree of similarity
and the possibility of homology.

Page 53

Definition: homology

Homology
Similarity attributed to descent from a common ancestor.

Page 49



Beta globin
(NP_000509)
2HHB

myoglobin
(NP_005359)
2MM1

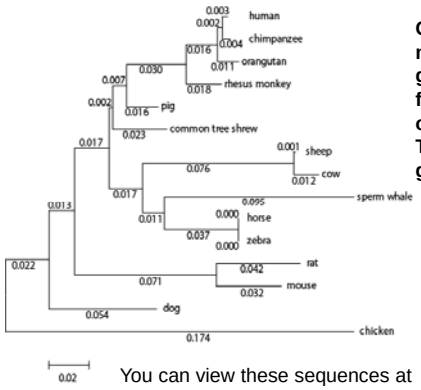
Page 49

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 49

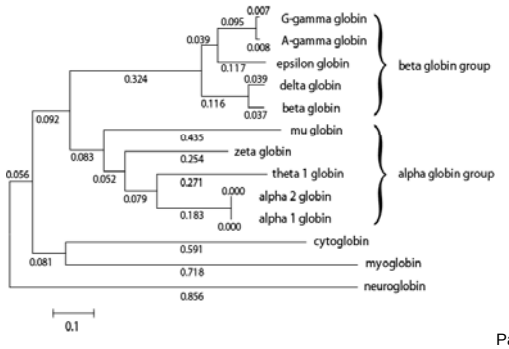


Orthologs: members of a gene (protein) family in various organisms. This tree shows globin orthologs.

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

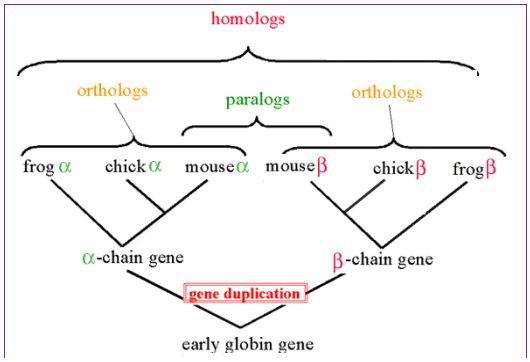
Page 51

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

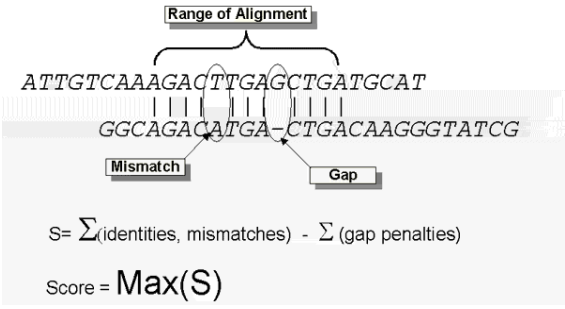


Source: NCBI

General approach to pairwise alignment

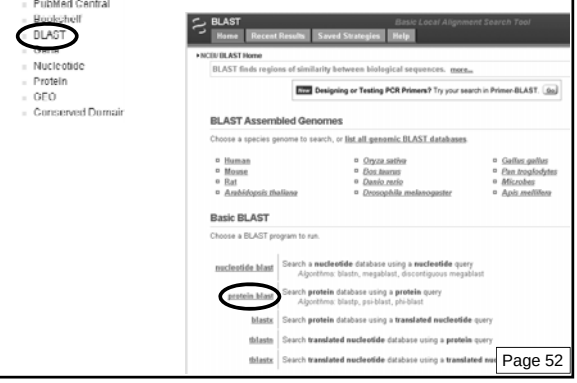
- Choose two sequences
- Select an algorithm that generates a score
- Allow gaps (insertions, deletions)
- Score reflects degree of similarity
- Alignments can be global or local
- Estimate probability that the alignment occurred by chance

Calculation of an alignment score

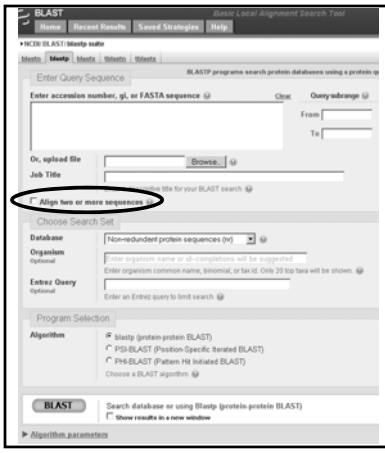


Source: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Alignment_Scores2.html

Find BLAST from the home page of NCBI and select protein BLAST…

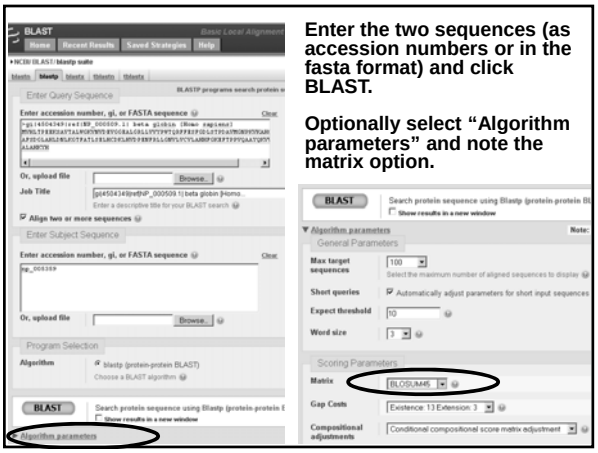


Page 52



Choose align two or more sequences…

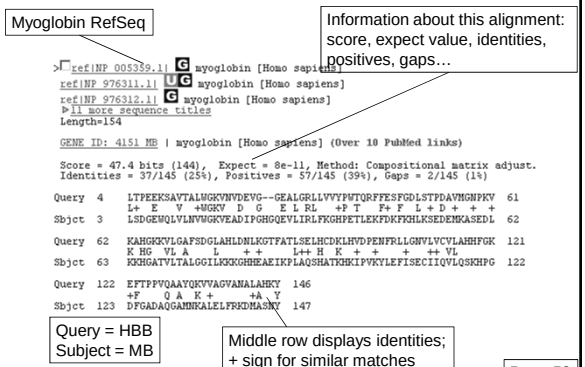
Page 52



Enter the two sequences (as accession numbers or in the fasta format) and click BLAST.

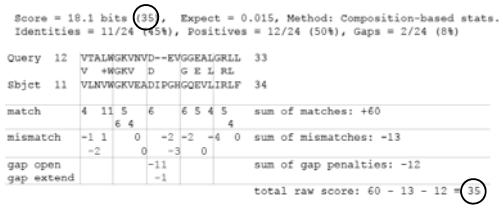
Optionally select "Algorithm parameters" and note the matrix option.

Pairwise alignment result of human beta globin and myoglobin



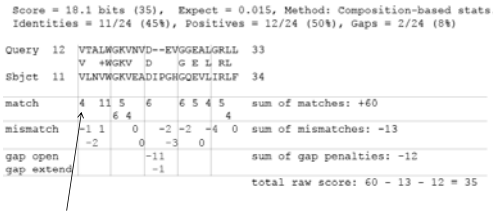
Page 53

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



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



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

These scores come from a "scoring matrix"!

Page 53

Definitions: homology

Homology
Similarity attributed to descent from a common ancestor.

Page 50

Definitions: identity, similarity, conservation

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.

Page 51

Definition: pairwise alignment

Pairwise alignment

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

Page 53

Mind the 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

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.

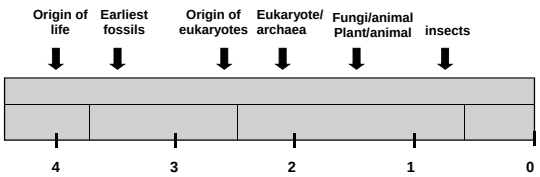
Pairwise alignment of retinol-binding protein and β-lactoglobulin: Example of an alignment with internal, terminal gaps

```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

 51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | |   |   |    ::  | .| .  ||  |:   ||        |.
 45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV...........QYSC 136 RBP
     || ||.        |        :.||||  | .             .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI....... 178 lactoglobulin
```

Pairwise alignment of retinol-binding protein from human (top) and rainbow trout (*O. mykiss*): Example of an alignment with few gaps

```
  1 .MKWVWALLLLA.AWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 48
     ::    ||  ||  ||   .||.||. .| :|||:.|:.| |||.|||||
  1 MLRICVALCALATCWA...QDCQVSNIQVMQNFDRSRYTGRWYAVAKKDP 47
            .         .         .         .         .
 49 EGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTED 98
     |||| ||:|:|||||.|.|.|.||| ||| :||||:.||.| ||| || |
 48 VGLFLLDNVVAQFSVDESGKMTATAHGRVIILNNWEMCANMFGTFEDTPD 97
            .         .         .         .         .
 99 PAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADS 148
     ||||||:||| ||:|| ||||||::||||| ||: |||| ..|||| |
 98 PAKFKMRYWGAASYLQTGNDDHWVIDTDYDNYAIHYSCREVDLDGTCLDG 147
            .         .         .         .         .
149 YSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL 199
    |||:||| | || || |||| :...|:|  ..|| : | |:|:
148 YSFIFSRHPTGLRPEDQKIVTDKKKEICFLGKYRRVGHTGFCESS..... 192
```

Pairwise sequence alignment allows us to look back billions of years ago (BYA)



When you do a pairwise alignment of homologous human and plant proteins, you are studying sequences that last shared a common ancestor 1.5 billion years ago!

Page 56

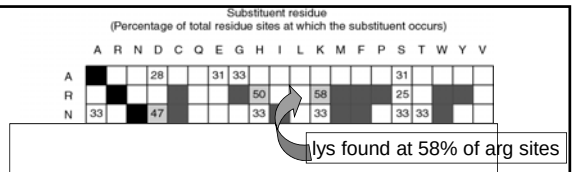
Multiple sequence alignment of glyceraldehyde 3-phosphate dehydrogenases: example of extremely high conservation

```
fly       GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human     GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant     GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast     GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon  GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly       KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human     KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant     KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast     KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon  KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly       GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human     GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant     GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast     GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon  GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```

Page 57

Outline: pairwise alignment

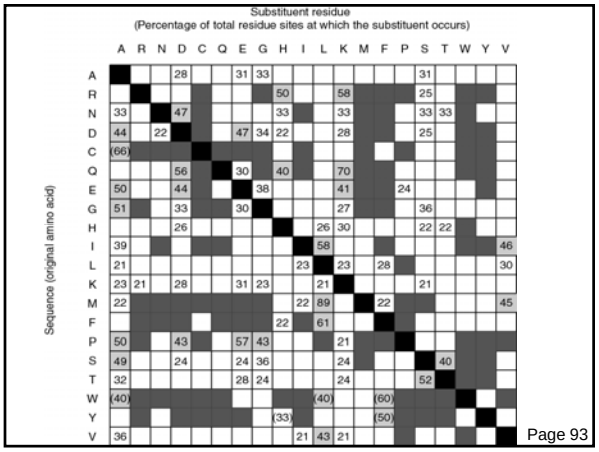
- Overview and examples
- Definitions: homologs, paralogs, orthologs
- Assigning scores to aligned amino acids: Dayhoff's PAM matrices
- Alignment algorithms: Needleman-Wunsch, Smith-Waterman



Emile Zuckerkandl and Linus Pauling (1965) considered substitution frequencies in 18 globins (myoglobins and hemoglobins from human to lamprey).

Black: identity
Gray: very conservative substitutions (>40% occurrence)
White: fairly conservative substitutions (>21% occurrence)
Red: no substitutions observed

Page 93



Page 93

Where we're heading: to a PAM250 log odds scoring matrix that assigns scores and is forgiving of mismatches… (such as +17 for W to W or -5 for W to T)

	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	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	

Page 69

…and to a whole series of scoring matrices such as PAM10 that are strict and do not tolerate mismatches (such as +13 for W to W or -19 for W to T)

	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	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-9	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8

Page 69

Dayhoff's 34 protein superfamilies

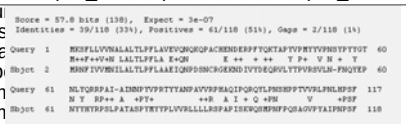
Protein	PAMs per 100 million years
Ig kappa chain	37
Kappa casein	33
luteinizing hormone b	30
lactalbumin	27
complement component 3	27
epidermal growth factor	26
proopiomelanocortin	21
pancreatic ribonuclease	21
haptoglobin alpha	20
serum albumin	19
phospholipase A2, group IB	19
prolactin	17
carbonic anhydrase C	16
Hemoglobin α	12
Hemoglobin β	12

Page 59

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
h	
seru	
phos	
prola	
carb	
Hem	
Hem	

human (NP_005203) versus mouse (NP_031812)



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 1	3.2

Page 59

Dayhoff's 34 protein superfamilies

Protein	PAMs per 100 million years
triosephosphate isomerase 1	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

Page 59

Pairwise alignment of human (NP_005203) versus mouse (NP_031812) ubiquitin

Dayhoff's approach to assigning scores for any two aligned amino acid residues

Dayhoff et al. defined the score of two aligned residues i,j as 10 times the log of how likely it is to observe these two residues (based on the empirical observation of how often they are aligned in nature) divided by the background probability of finding these amino acids by chance. This provides a score for each pair of residues.

$$s_{i,j} = 10 \times \log\left(\frac{q_{i,j}}{p_i}\right)$$

Page 58

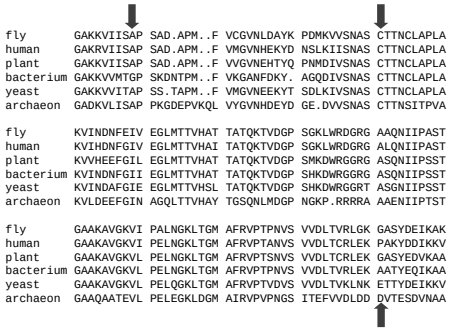
Dayhoff's numbers of "accepted point mutations": what amino acid substitutions occur in proteins?

	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

Dayhoff (1978) p.346.

Page 61

Multiple sequence alignment of glyceraldehyde 3-phosphate dehydrogenases: columns of residues may have high or low conservation



```
fly       GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human     GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant     GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast     GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon  GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly       KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human     KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant     KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast     KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon  KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly       GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human     GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant     GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast     GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon  GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```


The relative mutability of amino acids

Asn	**134**	**His**	**66**
Ser	**120**	**Arg**	**65**
Asp	**106**	**Lys**	**56**
Glu	**102**	**Pro**	**56**
Ala	**100**	**Gly**	**49**
Thr	**97**	**Tyr**	**41**
Ile	**96**	**Phe**	**41**
Met	**94**	**Leu**	**40**
Gln	**93**	**Cys**	**20**
Val	**74**	**Trp**	**18**

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%**

- **blue=6 codons; red=1 codon**
- **These frequencies f_i sum to 1**

Dayhoff's numbers of "accepted point mutations": what amino acid substitutions occur in proteins?

	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

Dayhoff's PAM1 mutation probability matrix

Original amino acid

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His
A	9867	2	9	10	3	8	17	21	2
R	1	9913	1	0	1	10	0	0	10
N	4	1	9822	36	0	4	6	6	21
D	6	0	42	9859	0	6	53	6	4
C	1	1	0	0	9973	0	0	0	1
Q	3	9	4	5	0	9876	27	1	23
E	10	0	7	56	0	35	9865	4	2
G	21	1	12	11	1	3	7	9935	1
H	1	8	18	3	1	20	1	0	9912
I	2	2	3	1	2	1	2	0	0

Dayhoff's PAM1 mutation probability matrix

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His
A	9867	2	9	10	3	8	17	21	2
R	1	9913	1	0	1	10	0	0	10
N	4	1	9822	36	0	4	6	6	21
D	6	0	42	9859	0	6	53	6	4
C	1	1	0	0	9973	0	0	0	1
Q	3	9	4	5	0	9876	27	1	23
E	10	0	7	56	0	35	9865	4	2
G	21	1	12	11	1	3	7	9935	1
H	1	8	18	3	1	20	1	0	9912
I	2	2	3	1	2	1	2	0	0

Each element of the matrix shows the probability that an original amino acid (top) will be replaced by another amino acid (side)

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).

Page 63

Dayhoff's PAM1 mutation probability matrix

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His
A	9867	2	9	10	3	8	17	21	2
R	1	9913	1	0	1	10	0	0	10
N	4	1	9822	36	0	4	6	6	21
D	6	0	42	9859	0	6	53	6	4
C	1	1	0	0	9973	0	0	0	1
Q	3	9	4	5	0	9876	27	1	23
E	10	0	7	56	0	35	9865	4	2
G	21	1	12	11	1	3	7	9935	1
H	1	8	18	3	1	20	1	0	9912
I	2	2	3	1	2	1	2	0	0

Page 66

Dayhoff's PAM0 mutation probability matrix: the rules for extremely slowly evolving proteins

PAM0	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu
A	100%	0%	0%	0%	0%	0%	0%
R	0%	100%	0%	0%	0%	0%	0%
N	0%	0%	100%	0%	0%	0%	0%
D	0%	0%	0%	100%	0%	0%	0%
C	0%	0%	0%	0%	100%	0%	0%
Q	0%	0%	0%	0%	0%	100%	0%
E	0%	0%	0%	0%	0%	0%	100%
G	0%	0%	0%	0%	0%	0%	0%

Top: original amino acid
Side: replacement amino acid

Page 68

Dayhoff's PAM2000 mutation probability matrix: the rules for very distantly related proteins

PAM∞	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
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%

Top: original amino acid
Side: replacement amino acid

Page 68

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

Page 68

	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

PAM250 log odds scoring matrix

Page 69

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).

Page 69

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

- The cells in a log odds matrix consist of an "odds ratio":

$$\frac{\text{the probability that an alignment is authentic}}{\text{the probability that the alignment was random}}$$

The score S for an alignment of residues a,b is given by:

$S(a,b) = 10 \log_{10} (M_{ab}/p_b)$

As an example, for tryptophan,

$S(trp,trp) = 10 \log_{10} (0.55/0.010) = 17.4$

Page 69

What do the numbers mean in a log odds matrix?

A score of +2 indicates that the amino acid replacement occurs 1.6 times as frequently as expected by chance.

A score of 0 is neutral.

A score of –10 indicates that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is one tenth as frequent as the chance alignment of these amino acids.

Page 58

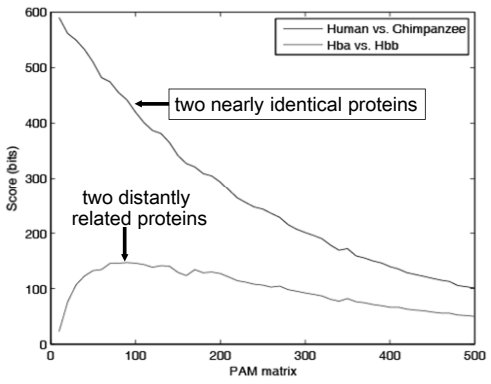
BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent	← →	*More divergent*

More conserved

Rat versus mouse globin

Less conserved

Rat versus bacterial globin



page 72

BLOSUM Matrices

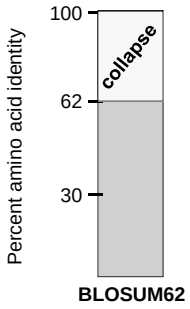
BLOSUM matrices are based on local alignments.

BLOSUM stands for blocks substitution matrix.

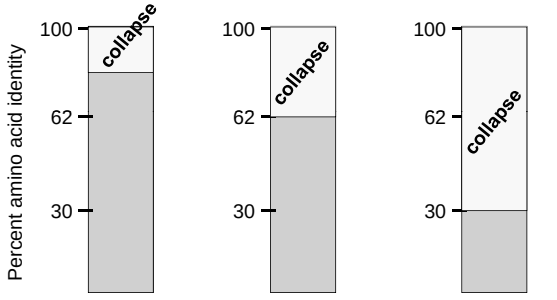
BLOSUM62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.

Page 70

BLOSUM Matrices



BLOSUM Matrices



BLOSUM Matrices

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

The BLOCKS database contains thousands of groups of multiple sequence alignments.

BLOSUM62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

Page 72

Blosum62 scoring matrix

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

Page 73

PAM matrices: Point-accepted mutations

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

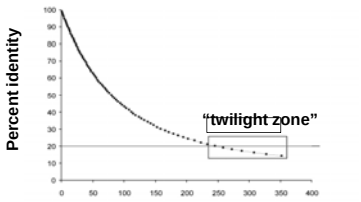
The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. At an evolutionary interval of PAM1, one change has occurred over a length of 100 amino acids.

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).

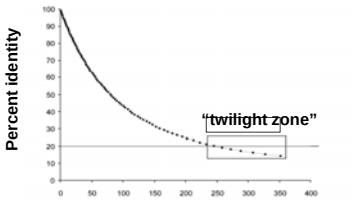
Page 74

Two randomly diverging protein sequences change in a negatively exponential fashion



Page 74

At PAM1, two proteins are 99% identical
At PAM10.7, there are 10 differences per 100 residues
At PAM80, there are 50 differences per 100 residues
At PAM250, there are 80 differences per 100 residues



Page 75

PAM: "Accepted point mutation"

• Two proteins with 50% identity may have 80 changes per 100 residues. (Why? Because any residue can be subject to back mutations.)

• Proteins with 20% to 25% identity are in the "twilight zone" and may be statistically significantly related.

• PAM or "accepted point mutation" refers to the "hits" or matches between two sequences (Dayhoff & Eck, 1968)

Page 75

Outline: pairwise alignment

- Overview and examples
- Definitions: homologs, paralogs, orthologs
- Assigning scores to aligned amino acids: Dayhoff's PAM matrices
- Alignment algorithms: Needleman-Wunsch, Smith-Waterman

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).

Finally, we will consider BLAST, a heuristic version of Smith-Waterman. We will cover BLAST in detail on Monday.

Page 76

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 76

Three steps to global alignment with the Needleman-Wunsch algorithm

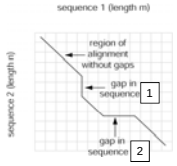
[1] set up a matrix

[2] score the matrix

[3] identify the optimal alignment(s)

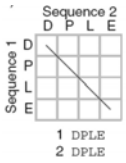
Page 76

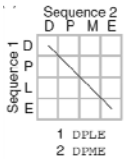
Four possible outcomes in aligning two sequences

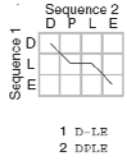


[1] identity (stay along a diagonal)
[2] mismatch (stay along a diagonal)
[3] gap in one sequence (move vertically!)
[4] gap in the other sequence (move horizontally!)

Page 77

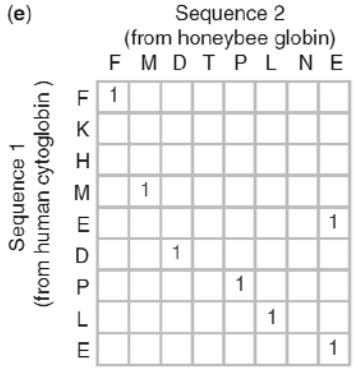






Page 77

Start Needleman-Wunsch with an identity matrix



Page 77

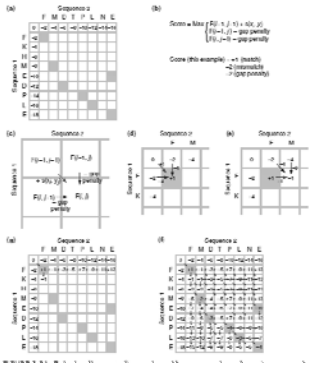
Start Needleman-Wunsch with an identity matrix

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

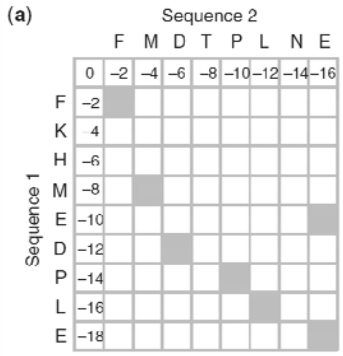
Fill in the matrix using “dynamic programming”



Page 78

Fill in the matrix using “dynamic programming”

(a)



Page 78

Fill in the matrix using “dynamic programming”

(b)

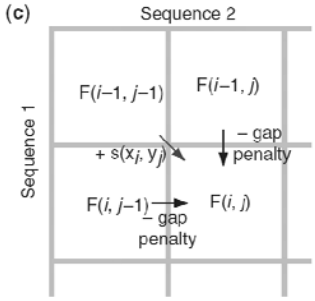
$$\text{Score} = \text{Max}\begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - \text{gap penalty} \\ F(i, j-1) - \text{gap penalty} \end{cases}$$

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

Page 78

Fill in the matrix using “dynamic programming”

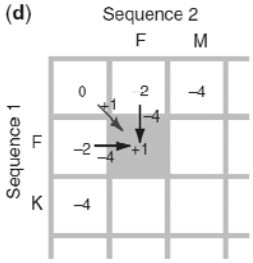
(c)



Page 78

Fill in the matrix using “dynamic programming”

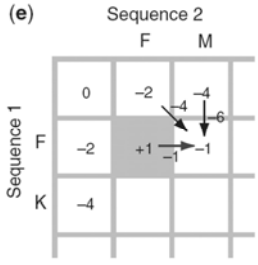
(d)



Page 78

Fill in the matrix using “dynamic programming”

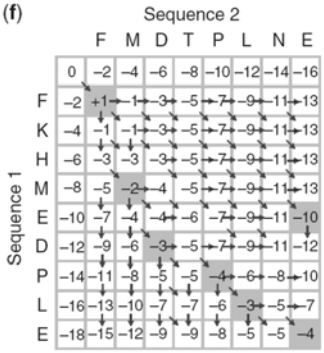
(e)



Page 78

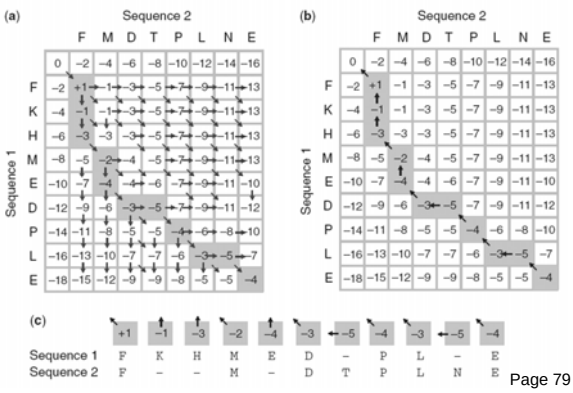
Fill in the matrix using “dynamic programming”

(f)



Page 78

Traceback to find the optimal (best) pairwise alignment



Page 79

Needleman-Wunsch: dynamic programming

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

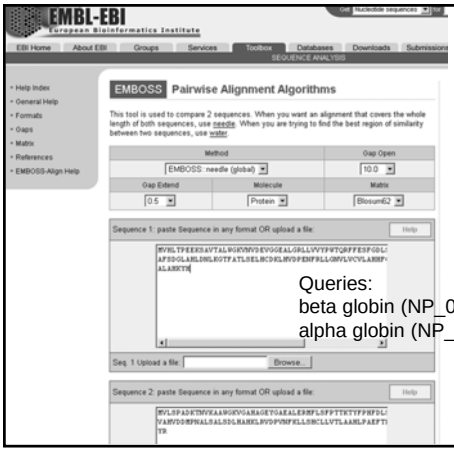
It is an example of a dynamic programming algorithm: an optimal path (alignment) is identified by incrementally extending optimal subpaths.
Thus, a series of decisions is made at each step of the alignment to find the pair of residues with the best score.

Page 80

Try using needle to implement a Needleman-Wunsch global alignment algorithm to find the optimum alignment (including gaps):

http://www.ebi.ac.uk/emboss/align/

Page 81



Queries:
beta globin (NP_000509)
alpha globin (NP_000549)

```
########################################
# Program:  needle
# Rundate:  Tue Aug 22 16:29:58 2006
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/needle-20060822-162957430033B5.output
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:          9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

EMBOSS_001         1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48
                     || |:|.:|:.|.|.|||| :..|.|.|||.|:.:..:|.|:.:|..| |
EMBOSS_001         1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48

EMBOSS_001        49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98
                     ||      |:..|.|.|||| .|.:..:||.|:.:..:.:||.|..||.
EMBOSS_001        49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93

EMBOSS_001        99 VDPENFRLLGHVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
                     |||.||:||.:.|:..||.|..||||.|.|:..|.:|.|:..|..||.
EMBOSS_001        94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
```

Global alignment versus local alignment

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

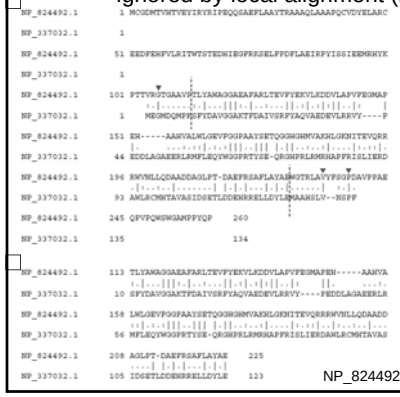
Local alignment finds optimally matching regions within two sequences (“subsequences”).

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.

Page 82

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



NP_824492, NP_337032

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 ≥ 0

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

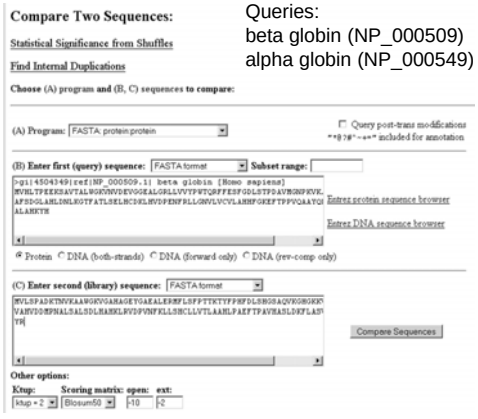
[1] s(i-1, j-1) + the new score at [i,j] (a match or mismatch)
[2] s(i,j-1) – gap penalty
[3] s(i-1,j) – gap penalty
[4] zero

Smith-Waterman algorithm allows the alignment of subsets of sequences

Sequence 1 (length m)

Sequence 2 (length n)	C	A	G	C	C	U	C	G	C	U	U	A	G
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
U	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
G	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
C	0.0	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
C	0.0	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
A	0.0	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
U	0.0	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
U	0.0	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
G	0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7
A	0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
C	0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
G	0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
G	0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

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



Queries:
beta globin (NP_000509)
alpha globin (NP_000549)

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.

Statistical significance of pairwise alignment

We will discuss the statistical significance of alignment scores in the next lecture (BLAST). A basic question is how to determine whether a particular alignment score is likely to have occurred by chance. According to the null hypothesis, two aligned sequences are not homologous (evolutionarily related). Can we reject the null hypothesis at a particular significance level alpha?

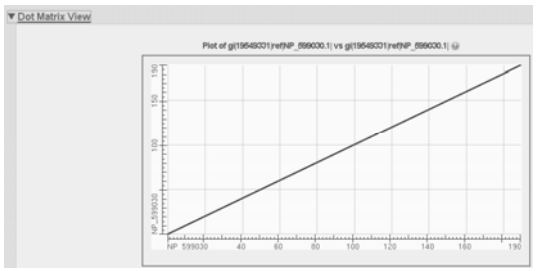
Pairwise alignments with dot plots: graphical displays of relatedness with NCBI's BLAST

[1] Compare human cytoglobin (NP_599030, length 190 amino acids) with itself. The output includes a dot plot. The data points showing amino acid identities appear as a diagonal line.

[2] Compare cytoglobin with a globin from the snail *Biomphalaria glabrata* (accession CAJ44466, length 2,148 amino acids. See lots of repeated regions!

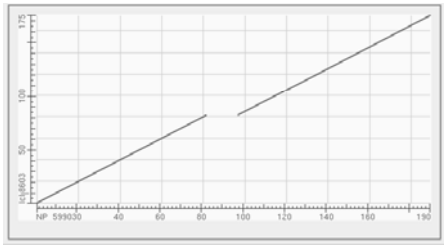
Page 85

Pairwise alignments with dot plots: cytoglobin versus itself yields a straight line



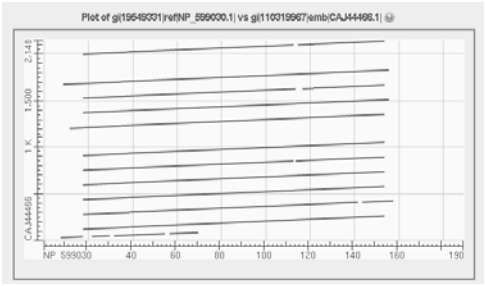
Page 85

Pairwise alignments with dot plots: cytoglobin versus itself (but with 15 amino acids deleted from one copy)



Page 85

Pairwise alignments with dot plots: cytoglobin versus a snail globin



Page 85

Next in the course...

Take the quiz (on pairwise alignment), due in a week (because of the Thanksgiving break, it's due TUESDAY at 5 pm).

Next Monday: BLAST