LSM2241
Fundamentals of Sequence Comparison I

Greg Tucker-Kellogg

19 August 2015

Outline

Introduction

Sequence homology

The issues and challenges in sequence analysis

Comparing sequences with dotplots

Setting up for pairwise alginment

Substitution matrices

Scoring matrices

Roundup and next week

Topic

Introduction

Sequence homology

The issues and challenges in sequence analysis

Comparing sequences with dotplots

Setting up for pairwise alginment

Substitution matrices

Scoring matrices

Roundup and next week

The most basic question in all of bioinformatics

Are these two sequences related?

- We address this problem by **aligning their sequences** with each other
- The workhorse of sequence alignment is **pairwise** sequence alignment

What is pairwise sequence alignment

Definition

The process of lining up two sequences to maximise conservation of sequence so that one can:

1. Assign the degree of *similarity*
2. Assess the likelihood of *homology*

Categories

1. Global alignment aligns sequences across their entire length
2. Local alignment finds stretches within sequences that align, even if they don't run the whole length

Note: One letter amino acid coding

Amino Acid	3-Letter	1-Letter	Side-chain polarity	Side-chain charge
Alanine	Ala	A	nonpolar	neutral
Arginine	Arg	R	polar	positive
Asparagine	Asn	N	polar	neutral
Aspartic acid	Asp	D	polar	negative
Cysteine	Cys	C	polar	neutral
Glutamic acid	Glu	E	polar	negative
Glutamine	Gln	Q	polar	neutral
Glycine	Gly	G	nonpolar	neutral

Histidine	His	H	polar	positive(10%)
				neutral(90%)

Isoleucine	Ile	I	nonpolar	neutral
Leucine	Leu	L	nonpolar	neutral
Lysine	Lys	K	polar	positive
Methionine	Met	M	nonpolar	neutral
Phenylalanine	Phe	F	nonpolar	neutral
Proline	Pro	P	nonpolar	neutral
Serine	Ser	S	polar	neutral
Threonine	Thr	T	polar	neutral
Tryptophan	Trp	W	nonpolar	neutral
Tyrosine	Tyr	Y	polar	neutral
Valine	Val	V	nonpolar	neutral

Some example alignments

A first example

```
HBA_HUMAN_FRA      1 GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL     41
                     |:.:||.|||||..|.::.:||:|::....:.||:||..||
HBB_HUMAN_FRA      1 GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL     41
```

A second example

```
HBA_HUMAN_FRA      1 GSAQVKGHGKKVADALTNA-----VAHVDDMPNALSALSDLHAHKL     41
                     .:.:::.|..||...:..|     |..|......|..|..:|..|.
LGB2_LUPLU_FR      1 NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG     46
```

What's the problem with this one?

```
HBA_HUMAN_FRA      1 GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL     41
                     ||..:.|......|.|  ...|..|:..|.:||.|    ..||:.
F11G11.2_FRAG      1 GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE     43
```

SMS to 77577 "gtk025 your answer" Link

Topic

Introduction

Sequence homology

The issues and challenges in sequence analysis

Comparing sequences with dotplots

Setting up for pairwise alginment

Substitution matrices

Scoring matrices

Roundup and next week

Homology and relatedness

- Two sequences are said to be *homologous* if they share common ancestry in evolution
- This is a yes or no question; there is no *degree* of homology
- Homology can arise from two processes in evolution
  - *Speciation* creates **orthologs** (orthologous sequences)
  - *Duplication* creates **paralogs**
- *Gene loss* sometimes leaves homologous fossils of ancestral genes
- *Gene conversion* sometimes makes it hard to infer orthologs

Paralogs and Orthologs

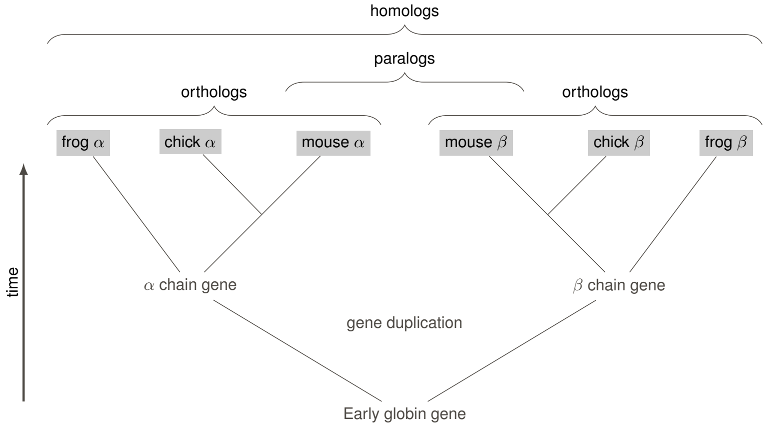
Definition (Paralogs)

Genes related by duplication within a genome

Definition (Orthologs)

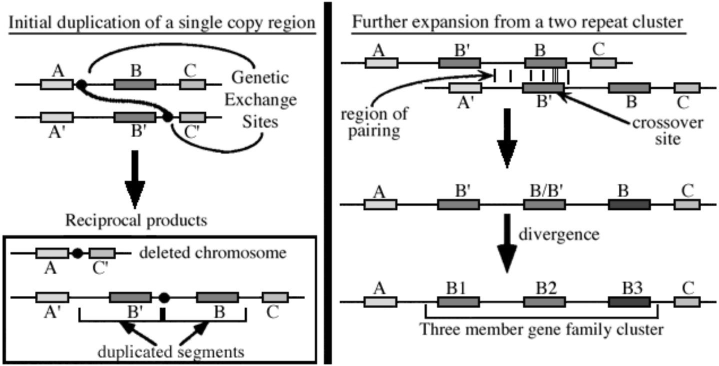
Genes related by speciation (the evolutionary process of creating new species)

Homologues through shared ancestry



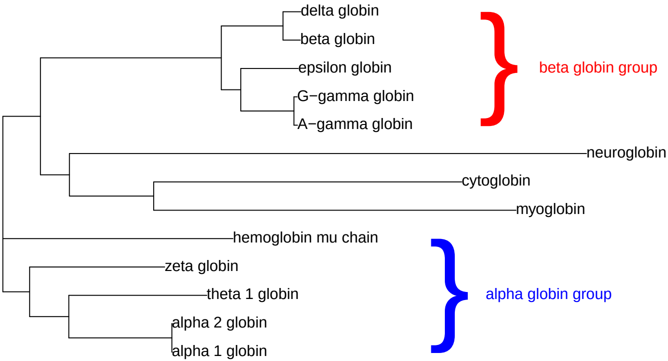
Relationships are often arranged on a *phylogenetic tree*, which we will cover in later weeks

How does gene duplication give rise to paralogs?



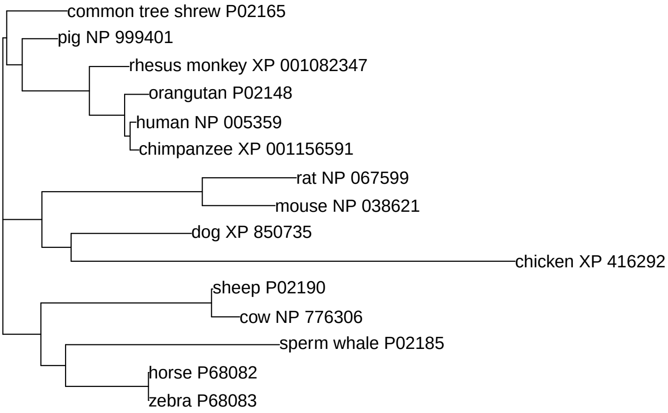
schematic of gene duplication by unequal cross-over (from Wikipedia). Gene duplication may also occur via retrotransposition

Paralogs: human globin protein sequences



human globin paralogs (members of a gene family *within* species) arranged in a phylogenetic tree (similar to Pevsner 2009 Fig 3.3)

Orthologs: the same gene shows relationships between species



myoglobin orthologs across species. Similar to Figure 3.2 of Pevsner.

Topic

Introduction

Sequence homology

The issues and challenges in sequence analysis

Comparing sequences with dotplots

Setting up for pairwise alginment

Substitution matrices

Scoring matrices

Roundup and next week

The issues in sequence alignment

The issues themselves

- What sorts of alignments to consider
- How to score alignments and rank them
- How to find the best (or at least good) scoring alignments
- How to evaluate — statistically — the significance of an alignment score.

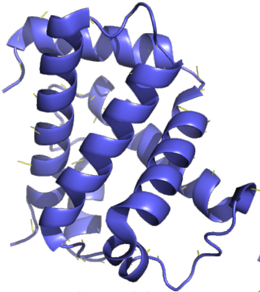
Some other things to keep in mind

- Some basic concepts of probability and statistics
- Finding optimal alignments can be too slow in some cases; sometimes speed is preferable to a guaranteed right answer.

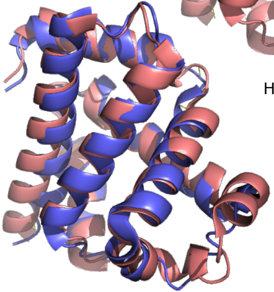
What sort of alignments to consider

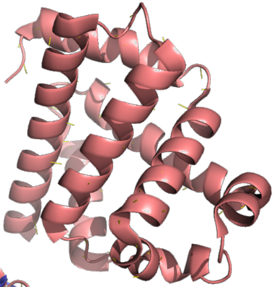
- Protein sequence alignment
- Nucleic acid sequence alignments
- Hybrid alignments
  - Aligning a protein with a translation of a nucleic acid sequence
- “Alignment” works with structure as well as sequence
  - Homologous sequences may show structural similarity even when sequence similarity is difficult to determine
  - Sequences were first compared *because* their structures were similar!
  - We will look at structure alignment

Aligning myoglobin and β-hemoglobin chains by their structures



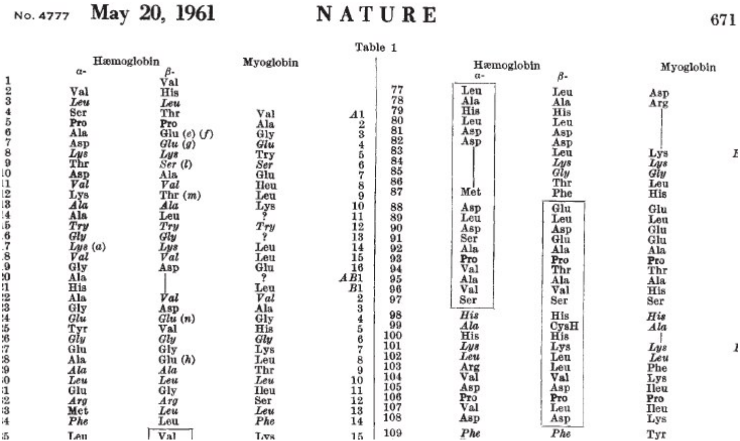
Human hemoglobin beta chain (2HHB)





Human myoglobin (3RGK)

An early example of sequence alignment



No. 4777 May 20, 1961 NATURE 671

Table 1

	Hæmoglobin		Myoglobin	
	α-	β-		
1		Val		
2	Val	His		
3	*Leu*	*Leu*		
4	Ser	Thr	Val	*A*1
5	**Pro**	**Pro**	Ala	2
6	Ala	Glu (*e*) (*f*)	Gly	3
7	Asp	*Glu* (*g*)	*Glu*	4
8	*Lys*	*Lys*	Try	5
9	Thr	*Ser* (*l*)	*Ser*	6
10	Asp	Ala	Glu	7
11	*Val*	*Val*	Ileu	8
12	Lys	Thr (*m*)	Leu	9
13	*Ala*	*Ala*	Lys	10
14	Ala	Leu	?	11
15	*Try*	*Try*	*Try*	12
16	*Gly*	*Gly*	?	13
17	*Lys* (*a*)	*Lys*	Leu	14
18	*Val*	*Val*	Leu	15
19	Gly	Asp	Glu	16
20	Ala		?	*AB*1
21	His		Leu	*B*1
22	Ala	*Val*	*Val*	2
23	Gly	Asp	Ala	3
24	*Glu*	*Glu* (*n*)	Gly	4
25	Tyr	Val	His	5
26	*Gly*	*Gly*	*Gly*	6
27	Glu	Gly	Lys	7
28	Ala	Glu (*h*)	Leu	8
29	*Ala*	*Ala*	Thr	9
30	*Leu*	*Leu*	*Leu*	10
31	Glu	Gly	Ileu	11
32	*Arg*	*Arg*	Ser	12
33	Met	*Leu*	*Leu*	13
34	*Phe*	Leu	*Phe*	14
35	Leu	Val	Lys	15

	Hæmoglobin		Myoglobin
	α-	β-	
77	Leu	Leu	Asp
78	Ala	Ala	Arg
79	His	His	
80	Leu	Leu	
81	Asp	Asp	
82	Asp	Asp	
83		Leu	Lys
84		*Lys*	*Lys*
85		*Gly*	*Gly*
86		Thr	Leu
87	Met	Phe	His
88	Asp	Glu	Glu
89	Leu	Leu	Leu
90	Asp	Asp	Glu
91	Ser	Glu	Glu
92	Ala	Ala	Ala
93	**Pro**	**Pro**	**Pro**
94	Val	Thr	Thr
95	Ala	Ala	Ala
96	Val	Val	His
97	Ser	Ser	Ser
98	*His*	His	*His*
99	*Ala*	CysH	*Ala*
100	His	His	
101	*Lys*	Lys	*Lys*
102	*Leu*	Leu	*Leu*
103	Arg	Leu	Phe
104	Val	Val	Lys
105	Asp	Asp	Ileu
106	**Pro**	**Pro**	**Pro**
107	Val	Leu	Ileu
108	Asp	Asp	Lys
109	*Phe*	*Phe*	Tyr

Myoglobin and hemoglobin chain sequences aligned in 1961 (Watson and Kendrew 1961; Perutz 1962)

Topic

Introduction

Sequence homology

The issues and challenges in sequence analysis

Comparing sequences with dotplots

Setting up for pairwise alginment

Substitution matrices

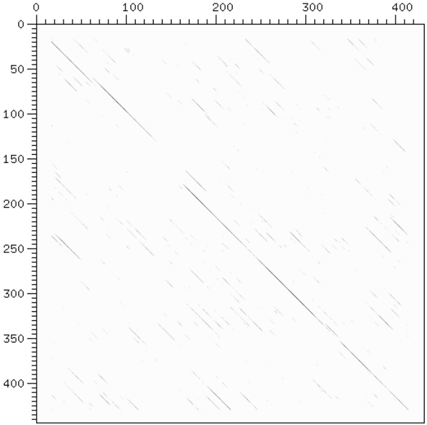
Scoring matrices

Roundup and next week

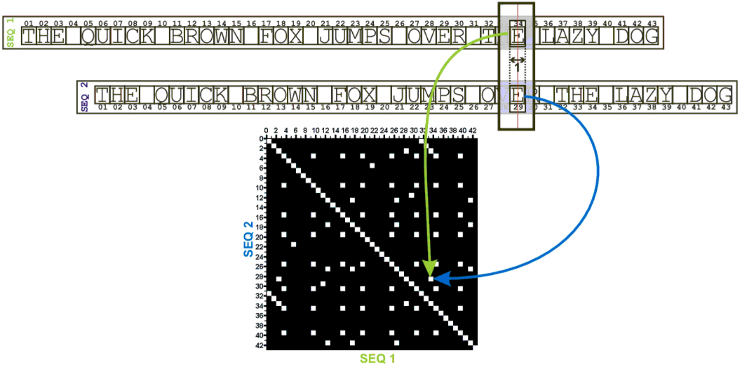
Dotplots provide a graphical overview of similarity

Human HbA vs HbB

- Similarity is plotted for a window across both sequences, without attempt at alignment.
- Overall similarity is evident from the main diagonal line
- Local similarity is evident from the diagonal lines elsewhere

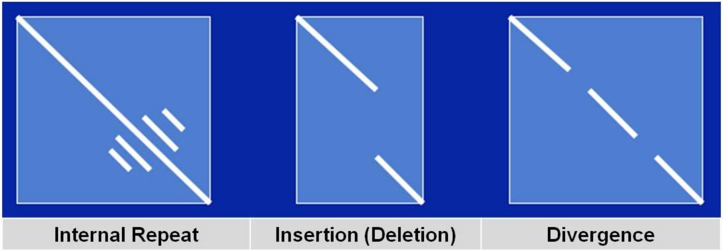


What dotplots are doing


SEQ 1
THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG
34
1
SEQ 2
THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG
29
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42
SEQ 2
SEQ 1

Each dot represents an identical letter at the corresponding position. Comparing against the self always gives a main diagonal of dots

Different dotplot patterns represent different things



- Dotplots give an informative picture of patterns of sequence similarity, even without an optimal alignment
- Even when you have an optimal alignment, dotplots can tell you what you have missed from it

Topic

Introduction

Sequence homology

The issues and challenges in sequence analysis

Comparing sequences with dotplots

Setting up for pairwise alginment

Substitution matrices

Scoring matrices

Roundup and next week

We can assess likely homology by sequence similarity

Assuming for the moment that we have a sequence alignment in hand. For that alignment, we can use the following terms, usually expressed in %:

Identity the extent to which two sequences in an alignment (protein or nucleotide) are, well, identical

Similarity the extent to which two sequences have similar residues at aligned positions[1]

Conservation Preservation of the physical and chemical properties in the sequence

The general approach to pairwise alignment

- Start with two sequences
- Pick a way to generate a score for differences at each point (point mutations).
- Pick a way to generate a score for gaps (insertions and deletions). This is always a penalty.
- The score S should reflect a measure of similarity

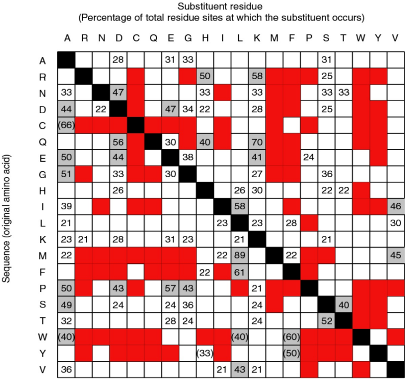
$$S = \sum(\text{identities,mismatches}) - \sum(\text{gap penalties})$$

- Estimate the probability that the alignment occurred by chance

Substitutions of amino acids for each other

Zuckerkandl and Pauling (1965)

- tabulated frequencies of substitutions in 18 globins
- some substitutions were forgiving, and occurred frequently
- others never occurred
- forgiving substitutions coincide with biochemical properties
- See Zuckerkandl and Pauling 1965



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

Sequence (original amino acid)

	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			🟥			🟥		■

from J. Pevsner, Bioinformatics and Functional Genomics

Topic

Introduction

Sequence homology

The issues and challenges in sequence analysis

Comparing sequences with dotplots

Setting up for pairwise alginment

Substitution matrices

Scoring matrices

Roundup and next week

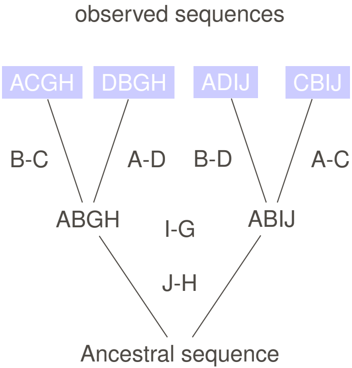
The Dayhoff Model

- Dayhoff, Schwartz, and Orcutt 1978 studied 34 protein superfamilies in 71 groups, mostly very conserved, some with more divergence
- Started with the frequencies of observed substitutions, came up with the frequencies of *acceptable* substitutions
- Alignments were checked for “Accepted Point Mutations” (APM), which were defined as changes (mutations) adopted by different species

What is an accepted point mutation?

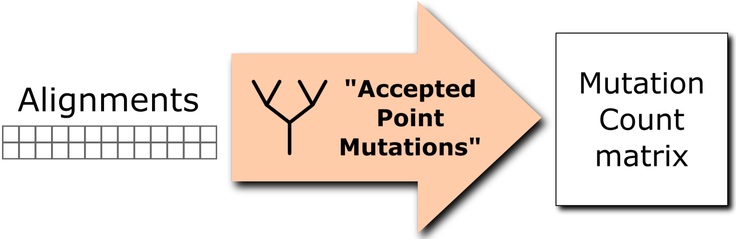
A model of protein sequence evolution

- Mutations compare with inferred ancestral sequences, not observed sequences
- The process of successive mutations form a *Markov chain*
- 1% accepted point mutations is a measure of evolutionary distance (1 PAM)



A matrix of mutation counts

From the Accepted Point mutations, a *matrix* counting mutations was developed

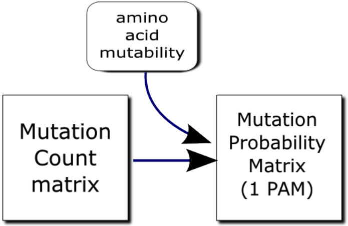


Derivation of Dayhoff mutation count matrix

The PAM1 mutation *probability* matrix

The process

- With counts of mutations
- Mutabilities of amino acids
- Derive relative probabilities of mutation
- Tune to 1% accepted mutation (1 PAM)

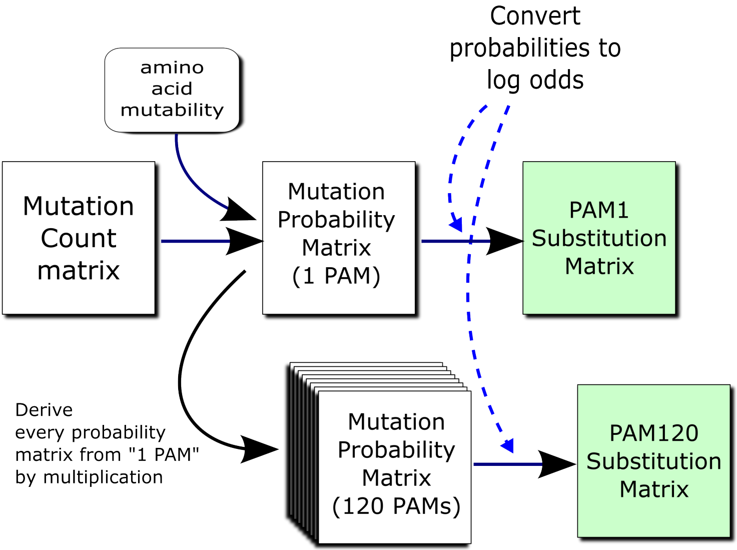


Derivation of the mutation probability matrix at 1 PAM

PAM Matrices

- APM was changed to PAM because it sounds better
- PAM1 represents 1 accepted point mutation per 100 amino acid residues (1 PAM unit of evolutionary divergence)
- Other PAM mutation probability matrices represent extrapolations from PAM1 for greater evolutionary distance
- PAM2 is generated by multiplying PAM1 with itself (matrix multiplication!)
- PAM250 is PAM1^{250}

Too many words!


Convert probabilities to log odds
amino acid mutability
Mutation Count matrix
Mutation Probability Matrix (1 PAM)
PAM1 Substitution Matrix
Derive every probability matrix from "1 PAM" by multiplication
Mutation Probability Matrix (120 PAMs)
PAM120 Substitution Matrix

Summary of the process used to construct PAM substitution matrices

Topic

Introduction

Sequence homology

The issues and challenges in sequence analysis

Comparing sequences with dotplots

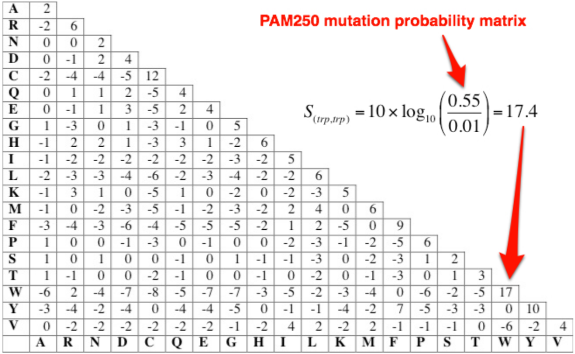
Setting up for pairwise alginment

Substitution matrices

Scoring matrices

Roundup and next week

PAM mutation probability matrices are converted to log-odds matrices for scoring



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

PAM250 mutation probability matrix

$$S_{(trp,trp)} = 10 \times \log_{10}\left(\frac{0.55}{0.01}\right) = 17.4$$

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

What is log-odds, and why do we do this?

- Remember
  - We want a scoring system that allows us to *add* the scores of individual residues
  - log odds lets us do this
- An *odds ratio* compares an observed event to a chance event:

$$\left(\frac{\text{odds the alignment is authentic}}{\text{odds the alignment is by chance}}\right)$$

- We take the $\log_{10}$ of this ratio to get the log-odds
- Taking the logarithm means that positive entries are more likely authentic than chance, negative entries are less likely authentic than chance

Later scoring matrices

A lot has changed since 1978!

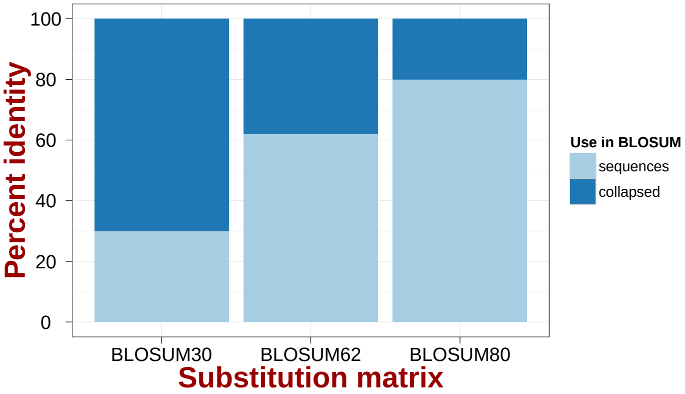
- Obviously, we have more sequences available for scoring
- But the PAM matrices are still quite widely used and very good.
- Newer versions based on the same principles

Alternatives to PAM: BLOSUM

- The BLOSUM family of matrices (S. Henikoff and J. G. Henikoff 1992) were formed from “block substitution alignments”, which were local alignments of related sequences. Different matrices emphasize different levels of identity
- Focused on *conserved* regions of *distantly related* proteins
- BLOSUM62 (the default for NCBI Blast) weights substitution frequencies towards less than 62% identity
- BLOSUM also uses log odds:

$$s_{i,j} = 2 \times \log_2 \left(\frac{q_{i,j}}{p_i} \right)$$

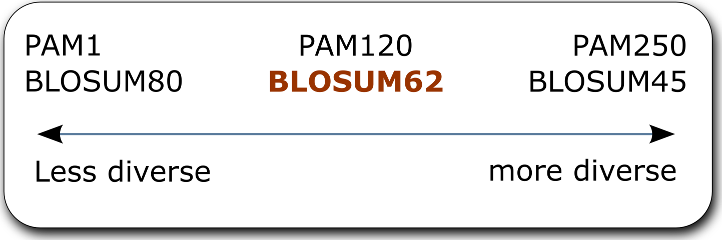
BLOSUM matrices are formed by collapsing similar sequences



What is "collapsing"?

- Every sequence initially compared with every other
- Define a cutoff percentage X
- Any set of sequences more than X percent identical are replaced by a single sequence

So which scoring matrix should you use?



PAM1 BLOSUM80

PAM120 **BLOSUM62**

PAM250 BLOSUM45

Less diverse ↔ more diverse

Matrix model a range of divergence of your sequences.
BLOSUM62 (in red) is the default scoring matrix for Blast protein sequence searches at the NCBI

Topic

Introduction

Sequence homology

The issues and challenges in sequence analysis

Comparing sequences with dotplots

Setting up for pairwise alginment

Substitution matrices

Scoring matrices

Roundup and next week

What have we learned?

- The meaning and importance of pairwise sequence alignment
  - This is the most basic bioinformatics question one can ask
- What sorts of alignments we might like to think about
  - Questions of relatedness (homology)
  - Questions of interpretation (similarity)
  - Questions of evolutionary history (orthologs, paralogs)
  - Local versus global alignments

What have we learned?

- How to score alignments, and rank them against each other
  - Scoring in general
  - Frequency substitution matrices
  - Probability substitution matrices
  - Scoring in terms of *log odds*
  - PAM and BLOSUM family matrices

Upcoming practical and next week's lecture

The practical practical

- deeper into working with databases and tools
- exploring orthologs and paralogs
- exploring the scores of sequence alignments

Next week

- The basic alignment algorithms
- Comparing more than two sequences

References I

Dayhoff, MO, RM Schwartz, and BC Orcutt (1978). “A Model of Evolutionary Change in Proteins”. In: *Atlas of protein sequence and structure* 5, pp. 345–352 (cit. on p. 29).

Henikoff, S and J G Henikoff (1992). “Amino acid substitution matrices from protein blocks.” In: *Proceedings of the National Academy of Sciences of the United States of America* 89.22, pp. 10915–9 (cit. on p. 39).

Perutz, M. F. (1962). “Relation between Structure and Sequence of Hæmoglobin”. en. In: *Nature* 194.4832, pp. 914–917. DOI: 10.1038/194914a0 (cit. on p. 19).

Pevsner, Jonathan (2009). *Bioinformatics and Functional Genomics*. Wiley-Blackwell, p. 992 (cit. on p. 13).

Watson, H C and J C Kendrew (1961). “The amino-acid sequence of sperm whale myoglobin. Comparison between the amino-acid sequences of sperm whale myoglobin and of human hemoglobin.” In: *Nature* 190, pp. 670–2 (cit. on p. 19).

Zuckerkandl, E and L Pauling (1965). “Molecules as documents of evolutionary history.” In: *Journal of Theoretical Biology* 8.2, pp. 357–66 (cit. on p. 27).