Pairwise alignment

Alignment between two or more protein or nucleic acid sequences represents an explicit hypothesis regarding the evolutionary history of those sequences. As a direct result, it facilitates many recent advances in understanding the information content and function of genetic sequences

>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPD
AVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR
LLGNVLVCVLAHHFG KEFTPPVQAAYQKVVAGVANALAHKYH

>sp|P02112|HBB_CHICK Hemoglobin subunit beta OS=Gallus gallus
VHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSP
TAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENF
RLLGDILIIVLAAHFS KDFTPECQAAWQKLVRVVAHALARKYH

Sequence comparison

Different types of pairwise comparisons	
Method name	***Situation***
Dot plot	**General exploration of your sequence** Discovering repeats Finding long insertions and deletions Extracting portions of sequences to make a multiple alignment
Local alignments	**Comparing sequences with partial homology** Making high-quality alignments Making residue-per-residue analysis
Global alignments	**Comparing two sequences over their entire length** Identifying long insertions and deletions Checking the quality of your data Identifying every mutation in your sequences

Dot Plots

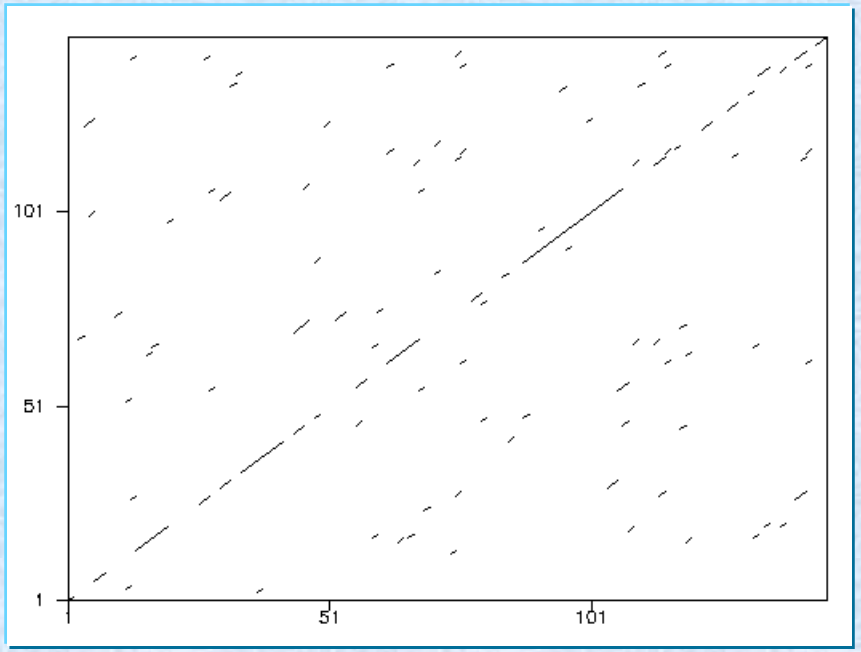
One of the simplest methods for evaluating similarity between two sequences is to visualize regions of similarity using dot plots.

To construct a simple dot plot, the first sequence to be compared is assigned to the horizontal axis of a plot space and the second is then assigned to the vertical axis.

Dots are then placed in the plot space at each position where both of the sequence elements are identical.

Adjacent regions of identity between two sequences give rise to diagonal lines of dots in the plot.

Human and chicken hemoglobin



VHLTPEEKSAVTALWGKVNV

VHWTAEEKQLITGLWGKVNV

Window size and cutoff (or threshold from substitution matrix)
E.g. 5 and 3

Simple alignment

An alignment between two sequences is simply a pairwise match between the characters in each sequence.

A true alignment of nucleotide or amino acid sequences is one that reflects evolutionary relationship between two or more homologues (sequences that share a common ancestor).

The three kinds of changes between sequences are

(i) a mutation that replaces one character with another

(ii) an insertion that adds one or more positions

(iii) a deletion that deletes one or more positions

Insertions and deletions occur less frequently than mutations. In these cases, gaps in alignments are commonly added.

Examples for mutation, insertion and deletion

Example

AATCTATA and AAGATA

Dot plots are useful for visual inspection of the regions of similarity between sequences.

A numeric system for evaluating sequence similarity has obvious advantages for objective determination of optimal alignments.

Scoring alignment without gap: credit for aligned residues and penalty for mismatches

AATCTATA
AAGATA

AATCTATA
AAGATA

AATCTATA
AAGATA

match score =1, if seq1(i) = seq2(i)

match score = 0, if seq1(i) ≠ seq2(i)

Identity matrix

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

The alignment scores are 4, 1, 3

Gaps

The insertions and deletions complicate sequence alignments by increasing the number of possible alignments between two or more sequences.

E.g. AATCTATA and AAGATA can be aligned in only three different ways.

These sequences can be aligned in 28 different ways using gaps.

AATCTATA	AATCTATA	AATCTATA
AAG-AT-A	AA-G-ATA	AAG--ATA

An alignment that includes gaps, an additional term, the gap penalty must be included in the scoring function.

Gap penalty = -1, if seq1(i) = "-" or seq2(i) = "-"

match score =1, if seq1(i) = seq2(i)

match score = 0, if seq1(i) ≠ seq2(i)

The scores are 1, 3, 3

Sequence alignment

Alignment 1

Position	1	2	3	4	5	6	7	8	9	10
Seq A:	V	–	E	I	T	G	E	I	S	T
Seq B:	P	R	E	–	T	E	R	I	–	T
Score:	0	-1	1	-1	1	0	0	1	-1	1
Total:	1									

Alignment 2

Seq A:	V	E	I	T	G	E	I	S	T
Seq B:	P	R	E	T	–	E	R	I	T
Score:	0	0	0	1	-1	1	0	0	1
Total:	2								

Alignment 3

Seq A:	–	V	E	I	T	G	E	–	I	S	T
Seq B:	P	R	E	–	T	–	E	R	I	–	T
Score:	-1	0	1	-1	1	-1	1	-1	1	-1	1
Total:	0										

VEITGEIST
PRETERIT

Origination and length penalties

One method to further distinguish between alignment is to differentiate between the alignments that contain many isolated gaps and those that contain fewer, but longer, sequence of gaps. E.g.

Consider two arbitrary sequences of lengths 12 and 9. Any alignment will have a shortage of 3 gaps in the shorter sequence.

Assuming the two sequences are homologous from the beginning to the end, then the differences are due to insertions in the longer or deletions in the shorter or both. Such events are called indel (insertion/deletion).

Considering this situation, we can bias our alignment scoring function to reward alignments that are more likely form an evolutionary perspective.

Gap penalty is divided into origination penalty (for starting new series of gaps on one of the sequences being aligned) and length penalty (number of sequential missing characters)

Origination and length penalties

AATCTATA	AATCTATA	AATCTATA
AAG-AT-A	AA-G-ATA	AAG--ATA

E.g. Origination penalty: -2; length penalty: -1, match score: +1 and mismatch score: 0

Case 1: 2x-2 + 2x-1 + 3x1 + 3x0 = -4 -2 +3 + 0 = -3

Case 2: -1

Case 3: 1x-2 + 2x-1 + 5x1 + 1x0 = -2 -2 + 5 + 0 = +1

Case 2 and case 3 are different; they were same in previous scoring method

Scoring matrices

In previous alignment, for non-gap positions scores are given as 1 for match and 0 for mismatch.

These scores can be further refined based on the substitutions.

For example Alanine is replaced with Valine or Lysine

Once the alignment score for each possible pair of nucleotides/amino acid residues are determined, the resulting scoring matrix is used to score each non-gap position in the alignment.

For nucleotides scoring is simple:

In BLAST, same nucleotides are given a score of +5 and different ones have -4

Case 2: matching nucleotides: mild reward (+1)

Transitions (purine to purine, A or G)/ pyramidine to pyramidine (C or T): mild penalty (-1)

Transversions (purine to pyramidine or vice versa); severe penalty (-5)

BLAST matrix

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

Transition Transversion matrix

	A	T	C	G
A	1	-5	-5	-1
T	-5	1	-1	-5
C	-5	-1	1	-5
G	-1	-5	-5	1

Scoring matrix: amino acids

Different criteria can be considered when devising a scoring matrix for amino acid sequence alignments

Most common ones are based on observed physical/chemical similarity and observed substitution frequencies

E.g. pairing two amino acids that both have aromatic functional groups might receive a good positive score,

pairing an amino acid that has a nonpolar functional group with one that has a charged functional group might result in a scoring penalty.

Scoring matrices have been derived based on residue hydrophobicity, charge and size

Another option is based on genetic code: minimum number of nucleotide substitutions are necessary to convert a codon from one residue to other

Scoring matrix: amino acids

A common method for deriving scoring matrices is to observe the actual substitution rates among various amino acid residues in nature.

If substitution between two amino acid residues is observed frequently, then positions in which these residues are aligned favorably.

Likewise, alignments between residues that are not observed to interchange frequently in natural evolution is penalized.

One commonly used scoring matrix based on observed substitution rates is the point accepted mutation (PAM) matrix.

The scores in a PAM matrix are computed by observing the substitutions that occur in alignments between similar sequences.

Development of PAM matrix

1. Alignment is constructed with very high sequence identity (usually >85%).

2.The relative mutatbility, m_j, for each amino acid is computed. It is the number of times the amino acid was substituted by any other amino acids. E.g. Ala to others

3. Pair of amino acids, A_{ij}, the number of times amino acid j was replaced by amino acid i, tallied for each amino acid pairs i and j. E.g. A_{cm} is the number of time Met is replaced with cysteine.

4. The substitution tallies are divided by relative mutability.

5. Normalize with the frequency of occurrence of each amino acid

6. Take log of each resulting entries in the PAM-1 matrix (PAM-1 means 1 substitution per 100 residues or 1 PAM unit) . This matrix is also called log odds matrix, since the entries are based on the log of the substitution probability for each amino acid.

PAM-1 matrix is appropriate to compare sequences are closely related. PAM-1000 matrix might be used to compare sequences with distant relationships. Usually PAM-250 is used for sequence alignment.

Calculation of a PAM matrix

PAM matrix is a 20x20 matrix for all pairs

Assumption:

Substitutions are equal is both directions (A to G and G to A)

E.g.: Element GA

Frequency of pairs, $F_{G,A} = 3$

Relative mutability, $m_A = 4$

Normalizing factor = number of mutations in the entire tree times 2, times relative frequency of A residues multiplied by 100 (1 substitution per 100 residues)

i.e., 6 x 2 x (10/63) x 100 = 190.4762

Hence, normalized relative mutability,

$m_A = 4/190.4762 = 0.021$

Consider a multiple sequence alignment

```
1.ACGCTAFKI
2.GCGCTAFKI  (1: A→G)
3.ACGCTAFKL  (1: I→L
4.GCGCTGFKI  (2: A→G)
5.GCGCTLFKI  (2: A→L)
6.ASGCTAFKL  (3: C→S)
7.ACACTAFKL  (3: G→A)
```

Construct tree

Calculation of a PAM matrix

$m_A = 4/190.4762 = 0.021$

Mutation probability, $M_{ij} = m_j F_{ij}/\Sigma f_{ij}$

$M_{G,A} = 0.021 \times 3 / 4$

$= 0.0157$

Σf_{ij}, total number of substitutions involving A

$R_{ij} = \log(M_{ij}/f_i) = \log(M_{GA}/f_G)$

$f_G = 10/63 = 0.1587$

$R_{GA} = \log(0.0157/0.1587) = \log(0.0989)$

$R_{GA} = -1.005$

Repeat for all off-diagonal elements.

For diagonal elements: $M_{jj} = 1 - m_j$

Calculate R_{jj}

Consider a multiple sequence alignment

```
1.ACGCTAFKI
2.GCGCTAFKI  (1: A→G)
3.ACGCTAFKL  (1: I→L
4.GCGCTGFKI  (2: A→G)
5.GCGCTLFKI  (2: A→L)
6.ASGCTAFKL  (3: C→S)
7.ACACTAFKL  (3: G→A)
```

Calculate the element R_{AA}

Calculate the element R_{IL}

PAM-120 mutation matrix

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

M. Michael Gromiha, BT3040

PAM 250 mutation matrix

	Cys	Gly	Pro	Ser	Ala	Thr	Asp	Glu	Asn	Gln	His	Lys	Arg	Val	Met	Ile	Leu	Phe	Tyr	Trp
Cys	12																			
Gly	-3	5																		
Pro	-3	-1	6																	
Ser	0	1	1	1																
Ala	-2	1	1	1	2															
Thr	-2	0	0	1	1	3														
Asp	-5	1	-1	0	0	0	4													
Glu	-5	0	-1	0	0	0	3	4												
Asn	-4	0	-1	1	0	0	2	1	2											
Gln	-5	-1	0	-1	0	-1	2	2	1	4										
His	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17