(Aspekte der Thermodynamik in der Strukturbiologie)

Einführung in die Bioinformatik

Wintersemester 2012/13
16:00-16:45 Hörsaal N100 B3

Peter Güntert

Pairwise sequence alignment

Outline

- Definitions
- Reasons for comparing two sequences

- Principles of dot plot comparisons
- Using Dotlet

- Relation between dot plots and alignments
- Basic principles of alignment scoring

Definition: Pairwise Alignment

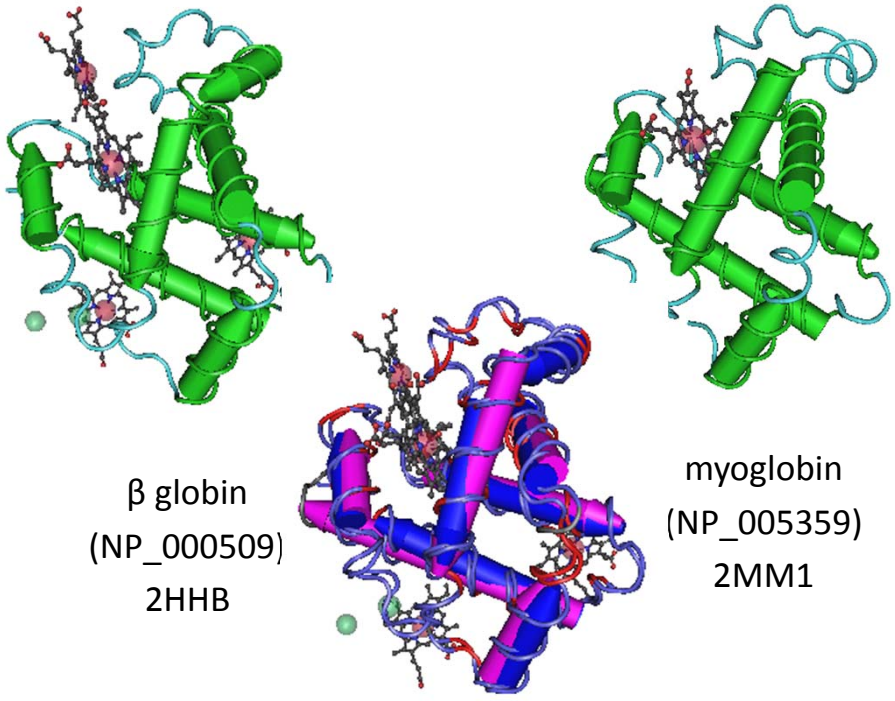
The process of lining up two sequences to achieve maximal levels of identity (and conservation, for amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

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.

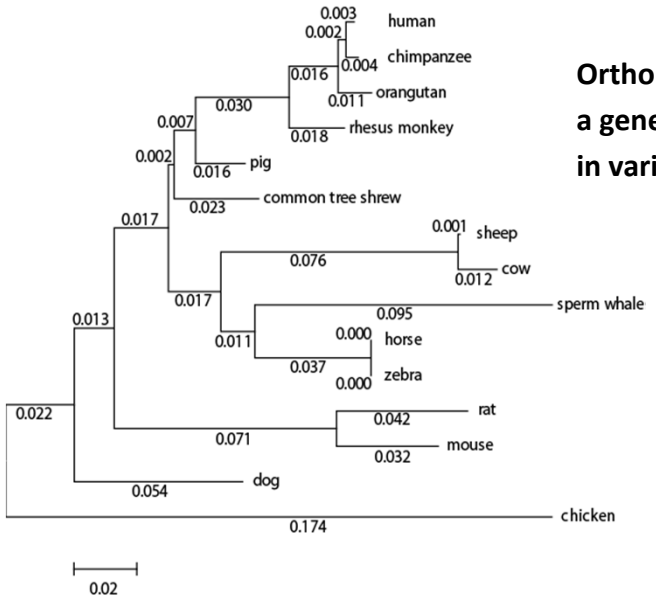
Conservation
Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.



Definitions: Homology, Orthologs, Paralogs

- **Homology:** Similarity attributed to descent from a common ancestor.
- **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.

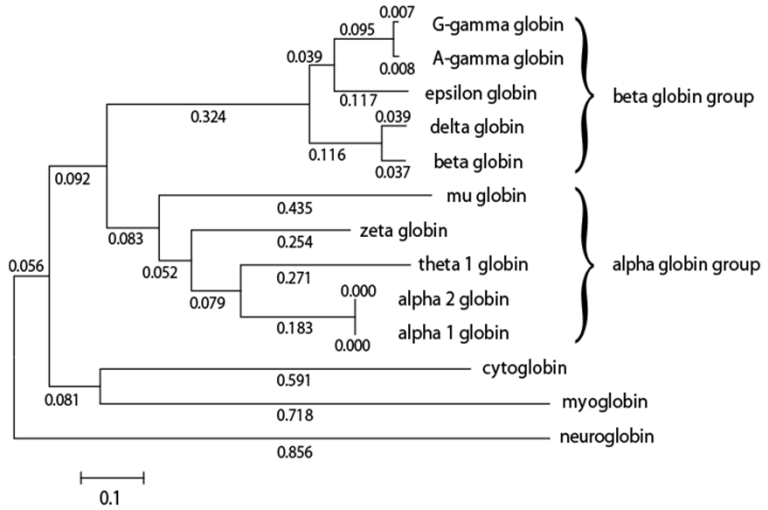
Globin Orthologs



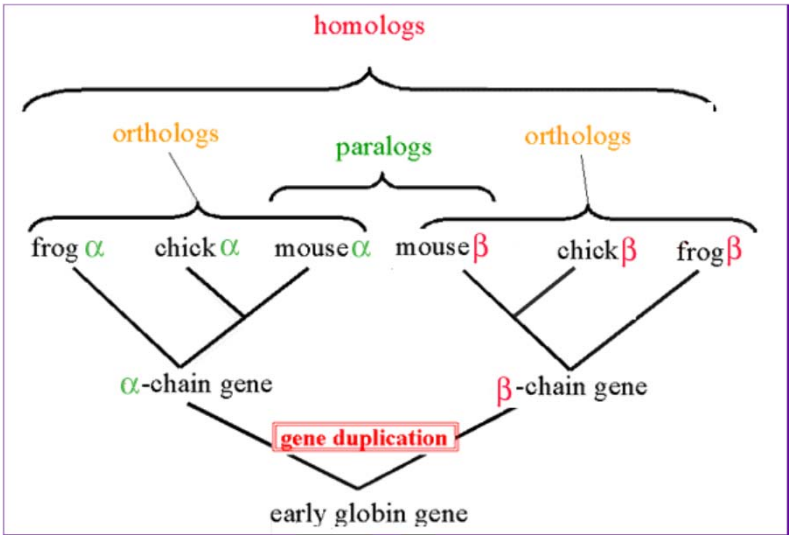
Orthologs: members of a gene (protein) family in various organisms.

Globin Paralogs

Paralogs: members of a gene (protein) family within a species.



Orthologs and paralogs are often viewed in a single tree



Why Compare Two Sequences?

- Database searches are useful for finding homologues
- Database searches don't provide precise comparisons
- More precise tools are needed to analyze the sequences in detail including
  - Dot plots for graphic analysis
  - Local or global alignments for residue/residue analysis
- The alignment of two sequences is called a *pairwise* alignment

Using The Right Tool

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

Some Applications of Pairwise Alignments

- Convince yourself two sequences are homologous
- Identify a shared domain
- Identify a duplicated region
- Locate important features such as
  - Catalytic domains
  - Disulphide bridges
- Compare a gene and its product

What is a Dot Plot ?

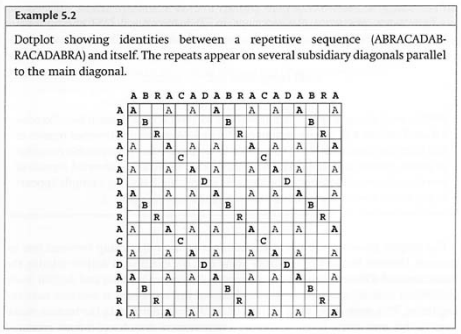
- A dot plot is a graphic representation of pairwise similarity
- The simplicity of dot plots prevents artifacts
- Ideal for looking for features that may come in different orders
- Reveal complex patterns
- Benefit from the most sophisticated statistical analysis tool . . . your brain

Example 5.1

Dotplot showing identities between short name (DOROTHYHODGKIN) and full name (DOROTHYCROWFOOTHODGKIN) of a famous protein crystallographer.

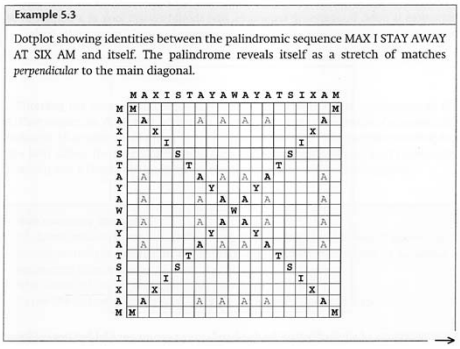
Letters corresponding to *isolated* matches are shown in non-bold type. The longest matching regions, shown in boldface, are the first and last names DOROTHY and HODGKIN. Shorter matching regions, such as the OTH of dorOTHy and crowfoOTHodgkin, or the RO of doROthy and cROwfoot, are noise.

Features in a dot plot



Example 5.2

Dotplot showing identities between a repetitive sequence (ABRACADABRACADABRA) and itself. The repeats appear on several subsidiary diagonals parallel to the main diagonal.



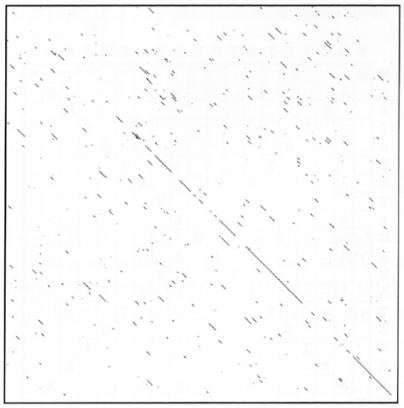
Example 5.3

Dotplot showing identities between the palindromic sequence MAX I STAY AWAY AT SIX AM and itself. The palindrome reveals itself as a stretch of matches *perpendicular* to the main diagonal.

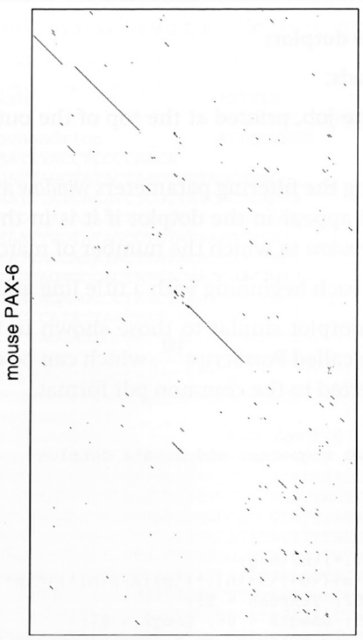
- A dot plot is a graphic representation of pairwise similarity
- The simplicity of dot plots prevents artifacts
- Ideal for looking for features that may come in different orders
- Reveal complex patterns
- Benefit from the most sophisticated statistical-analysis tool in the universe . . . your brain

Examples of dot plots

ATPases lamprey / dogfish shark

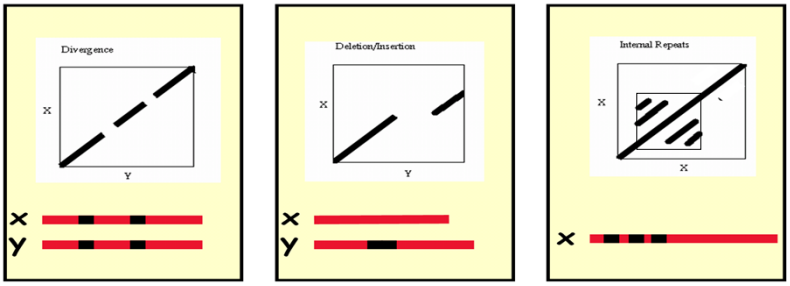


Drosophila eyeless



Some typical dot plot comparisons

- Divergent sequences where only a segment is homologous
- Long insertions and deletions
- Tandem repeats: The square shape of the pattern is characteristic of these repeats



Self-comparisons

- Start comparing your sequence with itself
- You can discover
  - Repeated domains
  - Motifs repeated many times (low complexity)
  - Mirror regions (palindromes) in nucleic acids

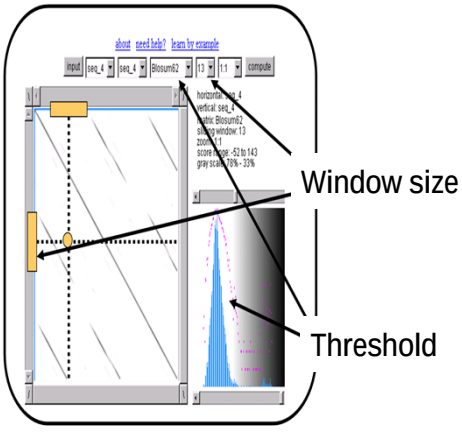
Using Dotlet to make a dot plot

- Dotlet is one of the handiest tools for making dot plots
- Dotlet is a Java applet
- Open and download the applet at the following site: www.isrec.isb-sib.ch/java/dotlet

- Alternative: Dotter http://sonnhammer.sbc.su.se/Dotter.html

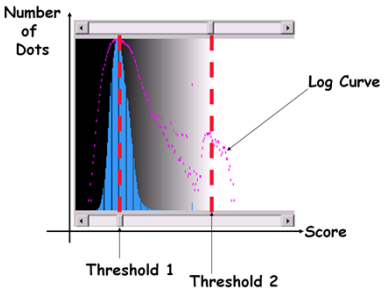
Set Dotlet parameters

- Dotlet slides a window along each sequence
- If the windows are more similar than the threshold, Dotlet prints a dot at their intersection
- You can control the similarity threshold with the little window on the left



The Dotlet threshold

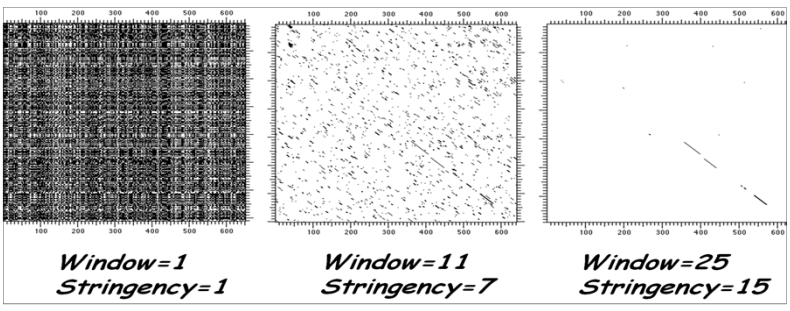
- Every dot has a score given by the window comparison
- When the score is
  - Below threshold 1 ⇔ black dot
  - Between thresholds 1 and 2 ⇔ grey dot
  - Above threshold 2 ⇔ white dot
- The blue curve is the distribution of scores in the sequences
- The peak ⇔ most common score,
  - Most common ⇔ less informative



Log curve

Getting your dot plot right

- Window size and the stringency control the aspect of your dot plot
  - Very stringent = clean dot plot, little signal
  - Not stringent enough = noisy dot plot, too much signal
- Play with the threshold until a usable signal appears

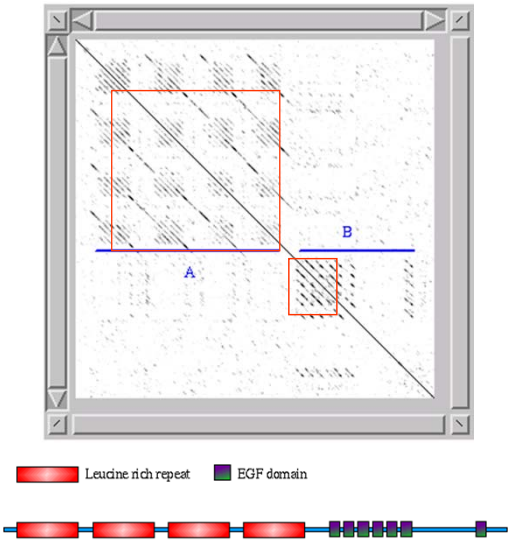


Which size for the window?

- Long window
  - Clean dot plots
  - Little sensitivity
- Short window
  - Noisy dot plots
  - Very sensitive
- The size of the window should be in the range of the elements you are looking for
  - Conserved domains: 50 amino acids
  - Transmembrane segments: 20 amino acids
- Shorten the window to compare distantly related sequences

Looking at repeated domains with Dotlet

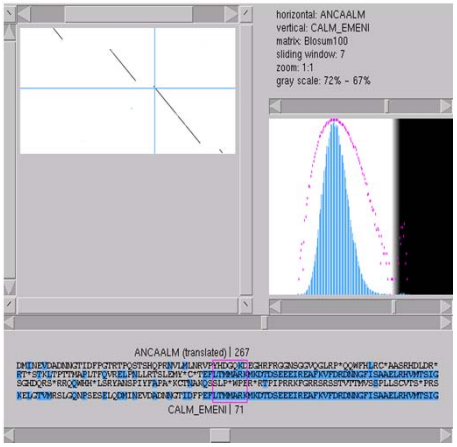
- The square shape is typical of tandem repeats.
- The repeats are not perfect because the sequences have diverged after their duplication.



Comparing a gene and its product

- Eukaryotic genes are transcribed into RNA
- The RNA is then spliced to remove the introns

- It may be necessary to compare the gene and its product
- Dotlet makes this comparative analysis easy



Relationship between dot plots and alignments

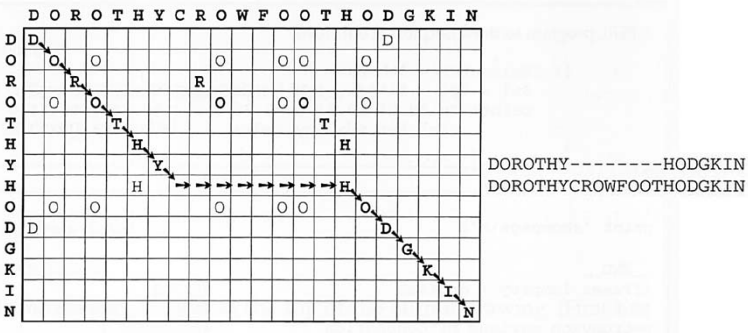


Fig. 5.1 Any path through the dotplot from upper left to lower right passes through a succession of cells, *each of which picks out a pair of positions, one from the row and one from the column, that are matched in the alignment that corresponds to that path; or that indicates a gap in one of the sequences.* The path need not pass through filled-in points only. However, the more filled-in points on the path, the more matching residues in the alignment.

Aligning sequences

- Dotlet dot plots are a good way to provide an overview
- Dot plots don't provide residue/residue analysis
- For this analysis you need an alignment
- The most convenient tool for making precise local alignments is Lalign

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

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

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

Measures of sequence similarity

To go beyond 'alignment by eyeball' via dotplots, we must define quantitative measures of sequence similarity and difference.

Given two character strings, two measures of the distance between them are:

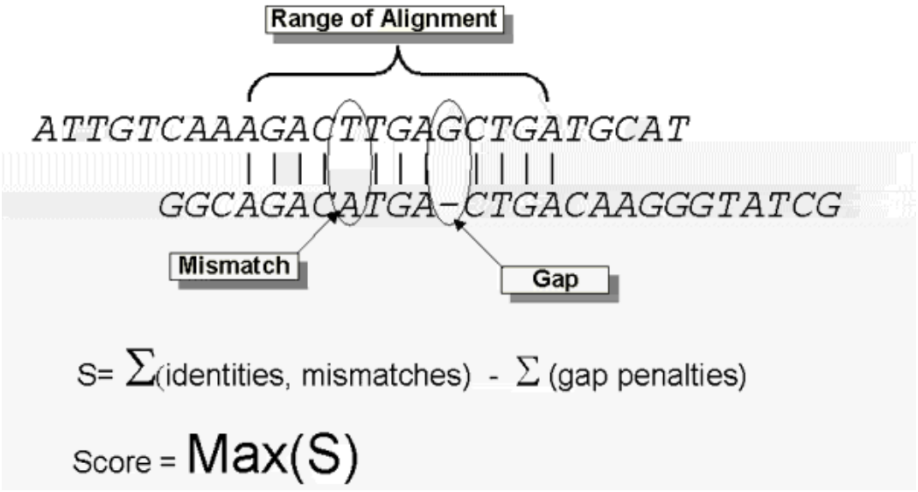
1. The **Hamming distance**, defined between two strings of equal length, is the number of positions with mismatching characters.
2. The **Levenshtein**, or **edit distance**, defined between two strings of not necessarily equal length, is the minimal number of 'edit operations' required to change one string into the other. An edit operation is a deletion, insertion or alteration of a single character in either sequence. A given sequence of edit operations induces a unique alignment, but not vice versa.

For example:

```
agtc
cgta
```
Hamming distance = 2

```
ag-tcc
cgctca
```
Levenshtein distance = 3

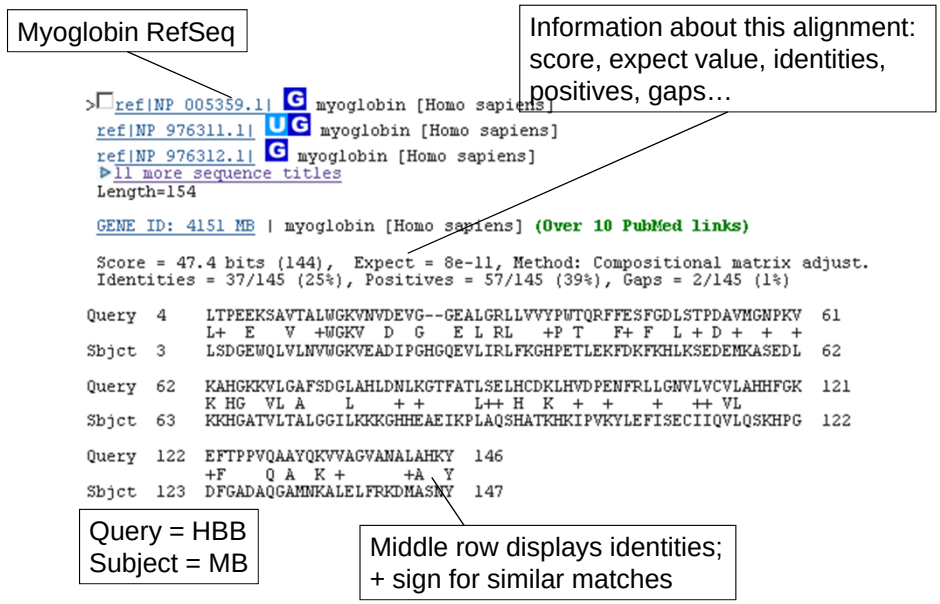
Calculation of an alignment score



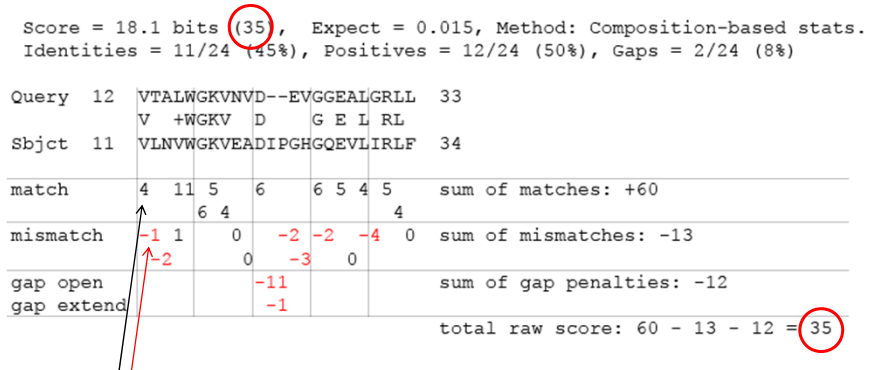
$S = \Sigma(\text{identities, mismatches}) - \Sigma(\text{gap penalties})$

$\text{Score} = \text{Max}(S)$

Pairwise alignment result of human β globin and myoglobin



The alignment 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".

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 6 4 4	sum of matches: +60
mismatch	-1 1 0 -2 -2 -4 0 -2 0 -3 0	sum of mismatches: -13
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 <u>creation</u> tends to have a large negative score.

Gap <u>extension</u> involves a small penalty.

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
```

Comparison alignment vs. dot plot

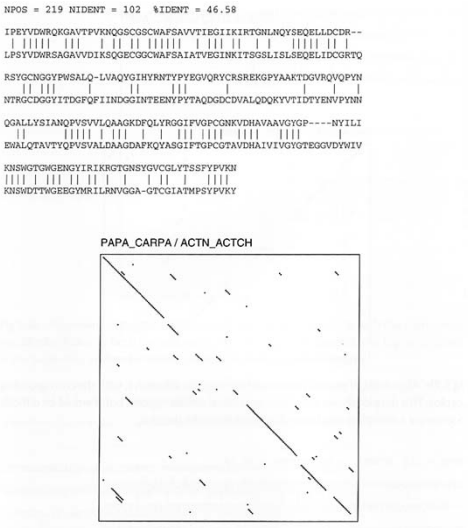


Fig 5.2a Alignment of papaya papain and kiwi fruit actinidin, with the corresponding dotplot.

Fig 5.2b Alignment of papaya papain and human procathepsin L, with the corresponding dotplot. This dotplot shows that there are several similar regions, but it would be difficult to generate a complete sequence alignment from the dotplot.

Comparison alignment vs. dot plot

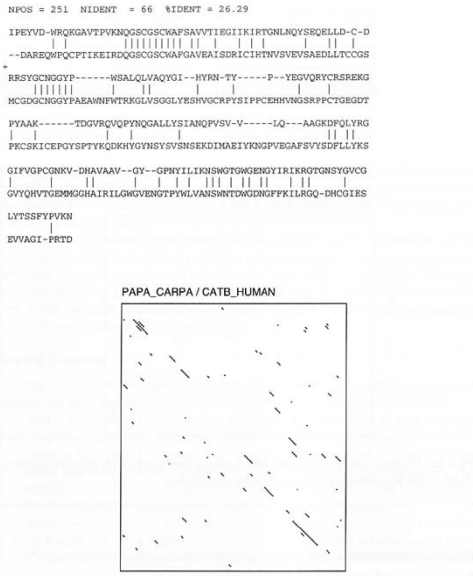


Fig 5.2c Alignment of papaya papain and human liver cathepsin B, with the corresponding dotplot. Note, in *both* the sequence alignment and the dotplot, the higher similarity at the beginning and end of the sequences than in the middle region.

Fig 5.2d Alignment of papaya papain and *S. aureus* staphopain, with the corresponding dotplot. The alignment is not derivable from this dotplot.

Substitution matrix for scoring

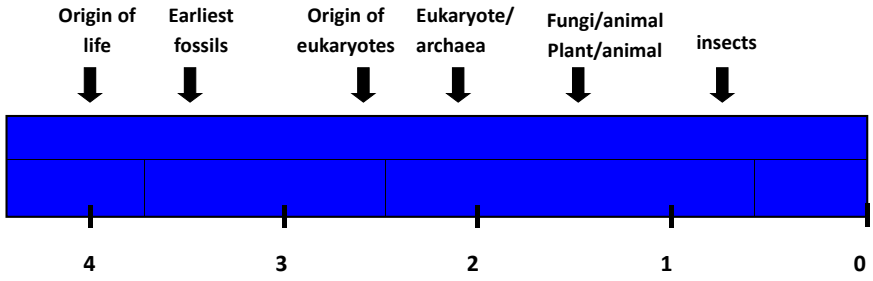
Substitution matrices used for scoring amino acid sequence similarity

The entries are in alphabetical order of the THREE-letter amino acid names. Only the lower triangles of the matrices are shown, as the substitution probabilities are taken as symmetric. (This is not because we are sure that the rate of any substitution is the same as the rate of its reverse, but because we cannot determine the differences between the two rates.)

The Dayhoff PAM250 matrix (MDM78):

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala (A)	2																			
Arg (R)	−2	6																		
Asn (N)	0	0	2																	
Asp (D)	0	−1	2	4																
Cys (C)	−2	−4	−4	−5	12															
Gln (Q)	0	1	1	2	−5	4														
Glu (E)	0	−1	1	3	−5	2	4													
Gly (G)	1	−3	0	1	−3	−1	0	5												
His (H)	−1	2	2	1	−3	3	1	−2	6											
Ile (I)	−1	−2	−2	−2	−2	−2	−2	−3	−2	−5										
Leu (L)	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	6									
Lys (K)	−1	3	1	0	−5	1	0	−2	0	−2	−3	5								
Met (M)	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6							
Phe (F)	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9						
Pro (P)	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6					
Ser (S)	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2				
Thr (T)	1	−2	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3			
Trp (W)	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17		
Tyr (Y)	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	
Val (V)	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4

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



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!

Unterlagen zur Vorlesung

http://www.bpc.uni-frankfurt.de/guentert/wiki/index.php/Teaching