



H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics online course: IBT

Module: Sequence Alignment Theory and Applications

Session: Pair-wise Sequence Alignment



H3ABioNet

Pan African Bioinformatics Network for H3Africa



Introduction to Bioinformatics online course : IBT
Jonathan Kayondo

Learning Objectives

- Understand the difference between global and local pair-wise alignment algorithms
- Understand the basic principles of pair-wise alignments, scoring matrices and gap penalties
- Understand the concepts of the dynamic programming approach for pair-wise sequence alignment (global and local)

Learning Outcomes

- Align two sets of sequences using both a global and local alignment approach
- Examine the effect of changing parameters such as scoring matrices, gap penalties etc
- Understand the output of the pair-wise alignment and when the global or local alignment method is more appropriate to use

Types of Pair-wise Comparisons

Different Types of Pair-wise comparisons	
Method	Scenario
Dot plot/matrix	<p>Exploratory view of sequence for:</p> <ul style="list-style-type: none"> • Discovering repeats • Finding long insertions/deletions • Extracting portions for use in multiple alignments
Local alignments	<p>Comparison of sequences with partial homology:</p> <ul style="list-style-type: none"> • Ensures high quality sub alignments • Detailed residue-per-residue analysis
Global alignment	<p>Comparing two sequences over their entire length:</p> <ul style="list-style-type: none"> • Identifying long insertions/deletions • Checking quality of your data • Identifying every mutation in your sequences • For making multiple sequence alignments

Picking a Method

- Good practice to start with a dot plot:
 - Provides an overview of the possible relationships between your two sequences
 - Helpful tool in deciding next steps
- Dot plots not enough for detailed examination:
 - Most Don't really produce alignment (simply give generic indications)
- Alignments needed for in-depth pair-wise sequence comparisons

Picking a Method cont..

- There are two kinds of alignments:
 - Global alignments (sequences aligned along their entire lengths)
 - Local alignments (alignments of only the most similar portions of the two sequences)
- Global alignments chosen if:
 - Sequences related along their entire length
 - Sequences don't contain long insertions or deletions

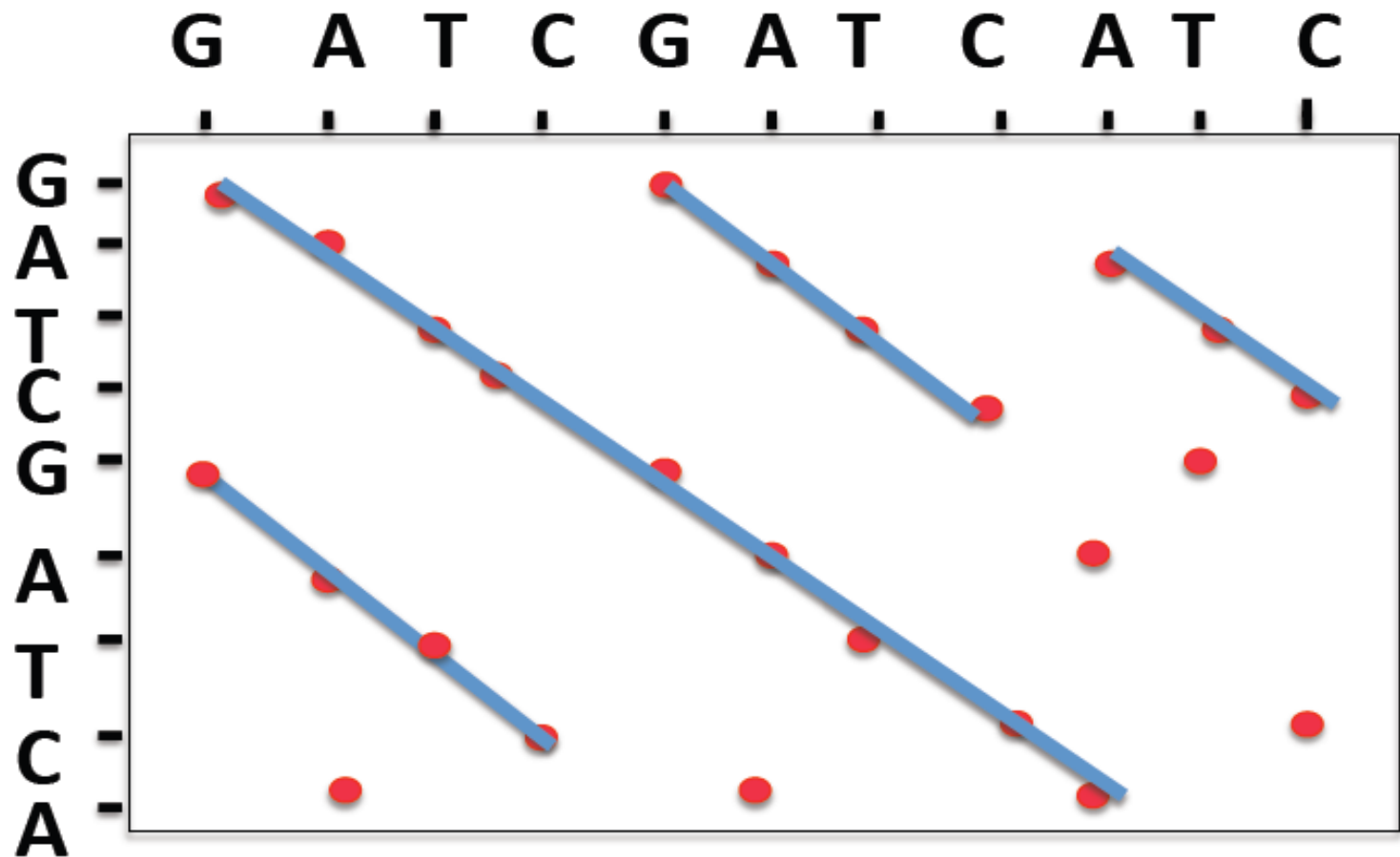
Making a Dot Plot

- Dot plot compares two sequences for possible alignments
- **Dot plot algorithm:**
 1. Draw a grid to write out the sequences
 2. One sequence (A) is listed across the top and the second sequence (B) listed down the left side
 3. Starting from the first character in B, one moves across the page keeping in the first row and placing a dot in any column where the character in A is the same
 4. The process is continued until all possible comparisons between A and B are made
 5. Any region of similarity is revealed by a diagonal row of dots
 6. Isolated dots not on diagonal represent random matches

Dot Plot/Matrix Considerations

- There are many dot plot flavors/programs to pick from
- If using computer program, default output might need to be refined by adjusting various settings e.g. sliding window size (comparisons of residues in immediate vicinity)
 - Long windows make clean dot plots (i.e. more stringent)
 - Shorter windows more sensitive but come with noise..can help in scenarios of distantly related proteins
 - Start large then progressively reduce until signal in question appears.

Dot Plot Example



Credit Pandam Salifu , IBT 2016



H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics online course: IBT

Sequence Alignment Theory and Applications | Jonathan Kayondo

Sequence Alignment Algorithms:

- **An algorithm** is a sequence of actions to be performed to arrive at a solution
- Rigorous algorithms (optimal alignments)- Dynamic programming
 - Needleman-Wunsch ('70) used for **global alignments**
 - Smith-Waterman ('81) for **local alignments**; provides one or more alignments of the sequences
- Heuristic algorithms (faster but only just approximate alignments...**see session 3**)
 - BLAST ('90)
 - FASTA ('85)

Dynamic Programming : Basic Idea

- Alignment of two sequences without allowing gaps requires making a large number of residue comparisons
- When gaps are allowed number of comparisons is even higher and is not achievable by direct comparison
- Dynamic programming is a method of sequence alignment that can take gaps into account but keeping manageable number of comparisons

Dynamic Programming : Basic Idea cont..

Designed to return best / optimal alignment that can be achieved between the two sequences....

Optimal not necessarily biologically meaningful

Approach/ How it works:

- Break up the problem into smaller sub-problems
- Solve the smaller problems optimally
- Use the sub-problem solutions to construct an optimal solution for the original problem

Dynamic Programming : Algorithm

- A computational method that considers all possible pairing of characters between the two sequences following a defined scoring scheme for matches, mismatches and gaps
- Produce a matrix of scores for all possible alignments between the sequences
- Highest set of sequential scores in the matrix defines an optimal alignment

Dynamic Programming : Algorithm cont..

- The method provides the highest scoring or optimal alignment between two sequences (nt, or A.a)
- Programs, including web hosted ones (e.g. LALIGN <http://www.ebi.ac.uk/Tools/psa/lalign/>, EMBOSS Water www.ebi.ac.uk/Tools/psa/emboss_water/), performing this type of analysis are readily available
- Method requires careful consideration for choice of variable settings in the programs:
 - Scoring matrix
 - Gap penalties
- Method is highly computationally demanding

Dynamic Programming: Algorithm procedure

- **Initialize matrix to guide alignment path:** computation of alignment quality depends on scoring system deriving from probabilities that:
 1. A particular amino acid pair is found in alignments of related proteins (p_{xy});
 2. The same amino acid pair is aligned by chance ($p_x p_y$) in the sequences given that some A.a are abundant in proteins, yet others rare;
 3. Insertion of a gap would be a better choice if it increases the score
- **Scores for aligning:** Ratio of probabilities 1:2 above, are usually provided as substitution matrix compilations, e.g. PAM, BLOSSUM (many sub-classes to pick from)

Dynamic Programming: Algorithm procedure cont..



- Each matrix table entry gives ratio of observed frequency of substitution between each possible amino acid pair in related proteins to that expected by chance, given the freq of amino acids in proteins
- The scores above are odd scores, usually transformed into log odds for easier manipulations

Calculating Scoring Matrices

- Log-odds matrix where each cell gives the probability of aligning those two residues
- Score of alignment = Sum of log-odds scores of residues
- Score for each residue given by:

$$s(a,b) = \frac{1}{\lambda} \log\left(\frac{p_{ab}}{f_a f_b}\right)$$

Scoring a sequence alignment

Sequence 1 V D S - C Y

Sequence 2 V E S L C Y

Score 4 2 4 -11 9 7

**Score = sum of amino acid pair scores (26)
minus single gap penalty (11) = 15**

1. Individual alignment scores are taken from an amino acid substitution matrix
2. Non-identical amino acids can be placed in corresponding positions.
3. Scores gained by each match are not always the same, for instance two rare amino acids will score more than two common.
4. Alignment gap(s) may be introduced for optimising the score. Gaps cause penalties.

Steps for dynamic programming algorithm

1. Score of new alignment = Score of previous alignment (A) + Score of new aligned pair

alignment		alignment (A)		aligned pair
V D S - C Y		V D S - C		Y
V E S L C Y		V E S L C		Y
15	=	8	+	7

2. Score of alignment (A) = Score of previous alignment (B) + Score of new aligned pair

alignment (A)		alignment (B)		aligned pair
V D S - C		V D S -		C
V E S L C		V E S L		C
8	=	-1	+	9

3. Repeat removing aligned pairs until end of alignments is reached

Dynamic Programming: Algorithm procedure- Gap scores



- Gaps in sequence alignments are given a big penalty to reflect fact that they are not expected to occur very often
- Gap Opening penalty- defines the cost for opening a gap in one of the sequences
 - No simple rule to predict optimal value for the gap penalty but if much higher than the default, local alignments containing gaps may split into shorter alignments
- Gap extension penalty – an extra penalty proportional to length of the gap. Usually <<< than Opening penalty.
- Gaps should be penalized more on their existence than their length

Global alignment: Needleman-Wunsch algorithm



- Dynamic programming method can be used to give global alignments as described by Needleman and Wunsch (1970)
- The optimal score at each matrix position is calculated by adding the current match score to previously scored positions and subtracting gap penalties if applicable
- Each matrix position may have a **+ve**, or **–ve** score or Zero (**0**)

Global alignment: Needleman-Wunsch algorithm cont..



- Needleman-Wunsch algorithm maximizes the number of matches between the sequences along the their entire length
- To produce a global sequence alignment from the scoring matrix, a second matrix called a **trace-back matrix** is produced
- Trace-back matrix keeps track of all the moves (residue alignment comparisons) in the scoring matrix
- Alignment produced by stringing together pairings from the optimal scores at each matrix position



Advantages and Disadvantages of Needleman-Wunsch algorithm



- Suitable for global alignment of two closely related sequences
- Can miss best local similarities
- Unsuitable for aligning:
 - very divergent sequences,
 - sequences with different domain structures

Local Alignment: Smith-Waterman Algorithm



- A modification of the dynamic programming algorithm for sequence alignment enables ability to create local sequence alignments (Smith and Waterman 1981a,b)
- Rules for calculating scoring matrix varies slightly from Needleman-Wunsch algorithm:
 - Scoring system includes –Ve scores for mismatches
 - When dynamic prog scoring matrix value becomes –ve, it is set to zero, which terminates alignment up to that point, and a new one can begin
- Trace-back matrix keeps track of all the trial moves
- Alignment produced by starting at the highest scoring positions in the matrix following a path up to Zero



Advantages and disadvantages of Smith-Waterman Algorithm

- Local alignments more meaningful than global matches because:
 - Can identify conserved local sequence domains present in both sequences
 - Can match two sequences with different lengths of overlap
- Smith-Waterman algorithm struggles finding highest scoring alignment when sequences include regions that align locally separated by other poorly aligning regions

Dynamic Programming: Algorithm procedure- Substitution Matrix

- Substitution matrix controls the cost of mutations:
 - Substitutions that are more likely should get a higher score
 - Substitutions that are less likely should get a lower score
- Appropriate substitution matrix can help determine likelihood of homology between two sequences

Dynamic Programming: Algorithm procedure- Types of Substitution Matrices

- Percent Identity:
 - Standard scoring matrix for aligning nucleotide (DNA) sequences
- Percent Accepted Mutation (PAM)
 - Protein sequence alignment
 - Estimates the rate at which each possible residue in a sequence changes to each other residue over time
- BLOSUM
 - For proteins
 - Derived from alignments found in BLOCKS database
 - BLOSUM-X: Identifies sequences that are X% similar to the query sequence

DNA Substitution Matrices

Sequence 1 ACTACCAGTTCATTTGATACTTCTCAAA

Sequence 2 TACCATTACCGTGTTAACTGAAAGGACTTAAAGACT

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

Match: 1

Mismatch: 0

Score: 6

Protein Scoring Systems

- Matrices reflect:
 - # of mutations to convert one A.a. residue to another
 - Chemical similarity
 - Observed mutation frequencies
 - Probability of occurrence of each A.a. residue
- Widely used scoring matrices:
 - PAM
 - BLOSUM

Percent Accepted Mutation (PAM)

- Family of matrices listing likelihood of change from one A.a. to another in homologous (i.e. Similar) protein sequences during evolution
 - Tracks evolutionary origin of proteins
 - Each gives changes expected for a given period of evolutionary time
- Predicted changes used to produce optimal alignments between two proteins and to score it
 - Assumption: A.a. substitutions observed over a shorter period of evolutionary history can be extrapolated to longer periods for higher PAMs

Percent Accepted Mutation (PAM) cont..

- Dayhoff, 1978 calculated, PAM1 matrix reflecting average change of 1% of all A.a. positions
- The common PAM250 matrix represents extrapolation to a level of 250% change expected in 2500 million years
 - NB: sequences at this level of divergence still show 20% similarity
- Higher PAMs yield better alignments than lower numbered PAMs for distantly related proteins and vice versa. For example could use:
 - PAM120: aligning sequences with 40% similarity
 - PAM80: ~50% similarity
 - PAM60: ~60% similarity

The PAM250 Matrix

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

the side group: (C) sulfhydryl (NDEQ) acid, acid amide and small hydrophobic, and (FYW) Log odds values ($\times 10^{-3}$) calculated (probability A.a pair will be found in alignment of two homologous proteins divided by the probability that the pair will be found in alignment of two unrelated proteins) means that ancestor probability are equal, -4 means (more by chance). Thus the probability of 10+10=20, whereas YY/TP is -1 between homologous sequences.

Amino acids are grouped according to the chemistry of the side group: (C) sulfhydryl, (STPAG)-small hydrophilic, (NDEQ) acid, acid amide and hydrophilic, (HRK) basic, (MILV) small hydrophobic, and (FYW) aromatic. Each matrix value is Log odds values ($\times 10$) calculated from odds scores: (probability A.a pair will be found in alignments of homologous proteins divided by probability that the pair will be found in alignment of unrelated proteins) : +10 means that ancestor probability is greater, 0 means that the probability are equal, -4 means that the change is random (more by chance). Thus the probability of alignment YY/YY is $10+10=20$, whereas YY/TP is $-3-5=-8$, a rare and unexpected between homologous sequences.

- ^[1] M.O. Dayhoff: *Survey of new data and computer methods of analysis* (1978), Atlas of protein sequence and structure, **5:3**

Blocks Amino Acid Substitution Matrices (BLOSUM)

- Use different strategy to estimate target frequencies
- Derived from alignments of domains of more diverse proteins (Henikoff & Henikoff, 1992)
- Occurrences of each amino acid pair in each column of each block alignment is counted
- Numbers derived from all blocks were used to compute the BLOSUM matrices

The BLOSUM Matrix

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

BLOSUM is based on local alignments. Scans for A.a. in very conserved regions of protein families (that do not have gaps in the sequence alignment) / blocks are made and the relative frequencies of amino acids and their substitution probabilities counted. Then, log-odds score for each of the 210 possible substitutions of the 20 standard amino acids are calculated. Zero (0) score means freq of A.a pair in the database is as expected by chance; +ve score means pair found more often than by chance; -ve means pair found less often than by chance

[2] S. Henikoff and J.G. Henikoff: *Amino acid substitution matrices from protein blocks* (1992), Proc. Natl. Acad. Sci., **89**:10915–10919s

Comparing PAM and BLOSUM


PAM Approach :

- Uses an evolutionary model
- Uses closely related sequences
- Extrapolates to greater distances

BLOSUM Approach:

- Looks at more distantly related sequences
- Observes actual mutations in motifs (not extrapolations)
- Uses sets of different overall identity

Comparing PAM and BLOSUM cont..



PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

More distant sequences

- PAM 120: for general use
- PAM 60: for close relations
- PAM 250: for distant relations
- BLOSUM 62: for general use
- BLOSUM 80: for close relations
- BLOSUM 45: for distant relations
- **BLOSUM** looks at mutations directly in motifs of more diverse sequences, whereas, **PAM** extrapolate evolutionary information based on a small set of related sequences

Statistical Significance of Alignments

- Alignment scores from substitution matrix represent quality of an alignment, but question remains whether score is high enough for evidence of homology:
 - To address this, important to know how high of a score can be expected due to chance alone
 - For global: one way is comparing observed alignment score with those of many alignments made from random sequences of the same length and composition
 - For local: Statistical models provide theory to test score against expected distributions of random local alignment scores in form of E-value (extreme value distribution) analysis.
 - Interpreted as # of alignments with scores at least equal to S (the observed) that would be expected by chance alone

Statistical Significance of Alignments cont.

- Note that:
 - Optimal methods always report the best alignment that can be achieved-
 - even if it has no biological meaning
 - When searching for local alignments, there could be several significant ones-
 - so it could be a mistake to look at only the optimal one
 - Looking at suboptimal alignments can be useful when comparing multi-module proteins (e.g. coagulation factors)
 - Bayesian statistics (*beyond scope for this module*) can allow examination of effect of prior information (e.g. chosen A.a. substitution matrix) to improve the call towards the most likely alignment

Conclusion

- Dynamic programming algorithm can provide alignment of DNA or protein sequences
 - Entire length of sequence (Needleman-Wunsch)
 - Localized regions (Smith-Waterman)
- Finding best alignment depends on appropriate choices for scoring matrix and gap penalties
- Important to validate alignment by showing that there are no alternative alignments almost as good (in terms of both score and probability)
- List of some Sequence Alignment resources can be found at:
http://en.wikipedia.org/wiki/List_of_sequence_alignment_software