

Algorithms in Bioinformatics FOUR Pairwise Sequence Alignment

0	-	A	T	C	A
1	C	-4	-2	-4	-4
2	A	-4	-2	-4	-4
3	T	-4	-2	-4	-4
4	C	-4	-2	-4	-4
5	A	-10	-7	-4	-4
6	G	-12	-9	-4	-4

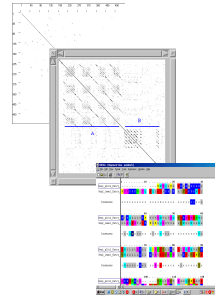
Sami Khuri
Department of Computer Science
San José State University

sami.khuri@sjsu.edu

0	-	A	T	C	A
1	C	-4	-2	-4	-4
2	A	-4	-2	-4	-4
3	T	-4	-2	-4	-4
4	C	-4	-2	-4	-4
5	A	-10	-7	-4	-4
6	G	-12	-9	-4	-4

©2018 Sami Khuri

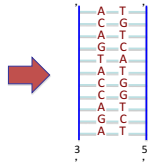
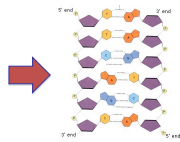
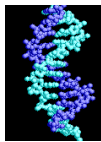
Pairwise Sequence Alignment



- Homology
- Similarity
- Global string alignment
- Local string alignment
- Dot matrices
- Dynamic programming
- Scoring matrices
- BLAST

©2018 Sami Khuri

Convention: DNA Sequences



5' ACAGTACCAGACCAGACCATACATACCATC 3'
3' TGTCATGGTCTGGTCTGGTATGTATGGTAG 5'

ACAGTACCAGACCAGACCATACATACCATC

©2018 Sami Khuri

Sequence Alignment

- **Sequence alignment** is the procedure of comparing sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences.
 - Comparing two sequences gives us a **pairwise sequence alignment**.
 - Comparing more than two sequences gives us **multiple sequence alignment**.

©2018 Sami Khuri

Pairwise vs Multiple Alignment

- **Pairwise Sequence Alignment**
 - Infer biological relationships from the sequence similarity
- **Multiple Sequence Alignment**
 - Infer sequence similarity from biological relationships

Starting point: sequences that are biologically related. Use the MSA to infer phylogenetic relationships. They can help elucidate biological facts about proteins since most conserved regions are biologically significant. MSA's can help formulate and test hypotheses about protein 3-D structure and function.

©2018 Sami Khuri

Why Do We Align Sequences?

- Phylogenetic trees and molecular evolution
- Identifying genes in a genome
- Predicting function of unknown genes
- Predicting protein structure
- Assembling genome sequences

©2018 Sami Khuri

Why Do We Align Sequences?

- The basic idea of aligning sequences is that **similar DNA sequences** generally produce **similar proteins**.
- To be able to predict the characteristics of a protein using only its sequence data, the **structure** or **function** information of known proteins with similar sequences can be used.
- To be able to check and see whether two (or more) genes or proteins are evolutionarily related to each other.

©2018 Sami Khuri

Importance of Alignments

- Alignment methods are at the core of many of the software tools used to search the databases.
- Alignment is the task of locating equivalent regions of two or more sequences to maximize their similarity.
- In order to assess the similarity of two sequences it is necessary to have a quantitative measure of their alignment, which includes the degree of similarity of two aligned residues as well as accounting for insertions and deletions.

Understanding Bioinformatics by Zvelebil and Baum

©2018 Sami Khuri

Query Sequence

If a query sequence is found to be significantly similar to an already annotated sequence (DNA or protein), we can use the information from the annotated sequence to possibly infer **gene structure** or **function** of the query sequence.

©2018 Sami Khuri

Similarity and Difference

- The **similarity** of two DNA sequences taken from different organisms can be explained by the theory that all contemporary genetic material has one common ancestral DNA.
- **Differences** between families of contemporary species resulted from mutations during the course of evolution.
 - Most of these changes are due to local mutations between nucleotide sequences.

©2018 Sami Khuri

When To Do The Pairwise Comparison?

- You have a strong suspicion that two sequences are homologous.
 - Two sequences are homologous, when they share a common ancestor.

©2018 Sami Khuri

Homology and Similarity

Homology

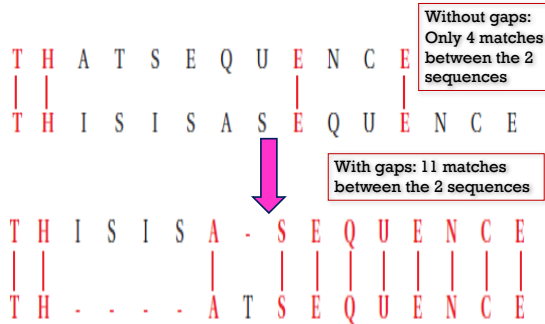
- Evolutionary related sequence.
- A common ancestral molecular sequence.

Similarity

- Sequences that share certain sequence patterns.
- Directly observable from alignment.

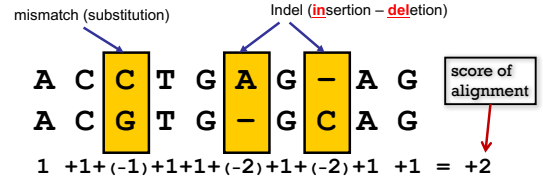
©2018 Sami Khuri

Need for Gaps: An Example



Scoring a Pairwise Alignment

- The two sequences are 70% identical



- Score of the alignment where:
Match → +1 Mismatch → -1 Indel → -2

Problem Definition

Given:

- Two sequences.
- A scoring system for evaluating match or mismatch of two characters.
- A penalty function for gaps in sequences.

Find:

- An **optimal pairing** of sequences that retains the order of characters in each sequence, perhaps introducing gaps, such that the total score is optimal.
- ©2018 Sami Khuri

Local and Global Alignments

- Global alignment**

- find alignment in which the **total score** is highest, perhaps at the expense of areas of great local similarity.

- Local alignment**

- find alignment in which the **highest scoring subsequences** are identified, at the expense of the overall score.
 - Local alignment can be obtained by performing minor modifications to the global alignment algorithm.
- ©2018 Sami Khuri

Shall we perform: Global or Local Alignment?

- Global Alignment:**

- Are these two sequences generally the same?

- Local Alignment:**

- Do these two sequences contain high scoring subsequences?

- Local similarities** may occur in sequences with different structure or function that share common substructure or subfunction.
- ©2018 Sami Khuri

Local Sequence Alignment

- The **optimal local alignment** of two sequences is the one that finds the longest segment of high sequence similarity between the two sequences.
- ©2018 Sami Khuri

Example: Local and Global Alignments

(A) **Local alignment**

```

P13-kinase  RENSSEIEMVKKDGGDFHIDFG
CAMP PK    RLSSEFELIDGGSEVYVIRK
    
```

(B) **Global alignment**

```

P13-kinase  HGLGNLE--LEEERE--MSSAKRPELWLNENFDI RSELEFENNEEIKNGDDERDRET
CAMP PK    GNAAAAGKGEKESVKEF LAKAKEDFLKKNENFAGNTAHEDDFEREKYLGTSGERVRE--
    
```

Understanding Bioinformatics by Zvelebil and Baum ©2018 Sami Khuri

Methods of Alignment

- A) By hand - slide sequences on two lines of a word processor
- B) Dot plot (also known as dot matrix)
 - with windows
- C) Rigorous mathematical approach
 - Dynamic programming (optimal but slow)
- D) Heuristic methods (fast but approximate)
 - BLAST and FASTA
 - Word matching and hash tables

A) Pairwise Sequence Alignment by Hand

- Write sequences across the page in two rows.
- Place identical or similar characters in the same column.
- Place non-identical characters either in the same column as a mismatch or opposite a gap in the other sequence.

B) Dot Matrix Method (I)

- Dot matrices are the simplest means of comparing two sequences.
- Dot matrices are designed to answer the following questions:
 - Where are all sites of similarity between my sequence and a second sequence?
 - Where are all sites of internal similarity in my sequence?
- Dot plots are not quantitative, they are qualitative.

The Dot Matrix Method (II)

- Dot plots place one sequence on the X axis, the other on the Y axis and compare the sequence on one axis with that on the other:
 - If the sequences match according to some criteria, a dot is placed at the XY intercept.
 - The dots populate a 2-dimensional space representing similarity between the sequences along the X and the Y axes.
- Dot plots present a visual representation of the similarity between two sequences, but do not give a numerical value to this similarity.

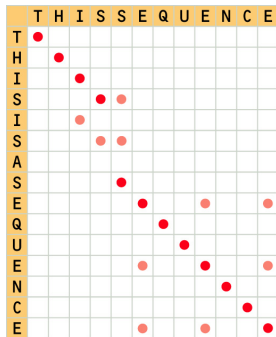
Dot Matrices

	A	T	G	C	C	T	A	G
A	*						*	
T		*				*		
G			*					*
C				*	*			
C				*	*			
T		*				*		
A	*						*	
G			*					*

Window Size = 1

The diagonal line always appears when a sequence is compared to itself.

Background Noise



Red dots represent identities that are meaningful – they are true matchings of identical residue-pairs.

Pink dots represent identities that are due to noise – they are matchings of random identical residues-pairs

Understanding Bioinformatics by Zvelebil and Baum

©2018 Sami Khuri

Improving Dot Matrices

- In a dot matrix, detection of matching regions may be improved by **filtering** out random matches.
- **Filtering** is achieved by using a sliding window to compare the two sequences.

©2018 Sami Khuri

Sliding Window

4
GAA CTCA TACGAATTCACATTAGAC

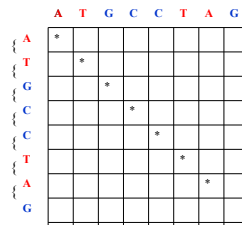
Window Size: Number of characters to compare

Stringency: Number of characters that have to match exactly

There are some default values for the window size and for the stringency, but one has to play around with the numbers to see what gives the best result.

©2018 Sami Khuri

Dot Matrices with Windows



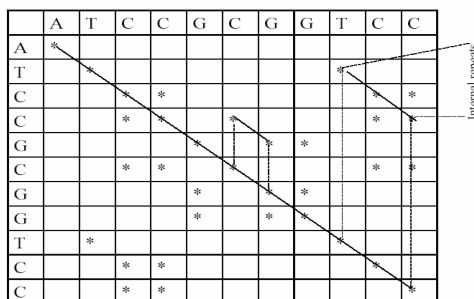
Window Size = 2

Compare two nucleotides at a time.

Windows **filter** out the noise.

©2018 Sami Khuri

Internal Repeats



Number of letters to compare (window)=1

©2018 Sami Khuri

Determining Window Size

- | | |
|---|--|
| <p>DNA Sequences</p> <ul style="list-style-type: none"> • A typical window size is 15. • A suitable match (stringency) requirement in this window is 10. | <p>Protein Sequence</p> <ul style="list-style-type: none"> • Often the matrix is not filtered, but a window size requirement is 2 or 3. • A match requirement of 2 will highlight matching regions. |
|---|--|

©2018 Sami Khuri

Advantages of Dot Matrix

- All possible matches of residues between two sequences are found
 - The investigator now has the choice of identifying the most significant ones.
- The sequences of the actual regions that align can be detected by using one of two other methods for performing sequence alignments.
- The presence of **insertions/deletions** and direct and inverted **repeats** can be revealed.

©2018 Sami Khuri

Dot Matrices Shortcomings

- Most dot matrix computer programs do not show an actual alignment.
- Dot matrices rely on visual analysis.
- It is difficult to find optimal alignments.
 - We need scoring schemes more sophisticated than identical match.
- It is difficult to estimate the significance of alignments.
- Dot matrices do not allow gaps in the sequence alignments.

©2018 Sami Khuri

Other Applications of Dot Matrix

- Finding direct, inverted or tandem repeats in protein and DNA sequences.
- Predicting regions in RNA that are self-complementary and that have the potential of forming secondary structure.

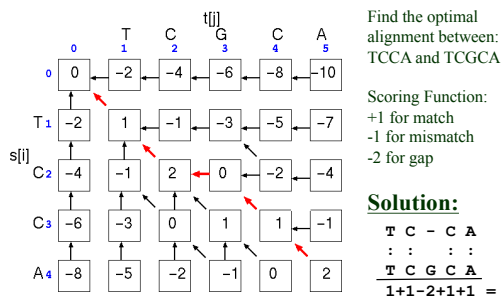
©2018 Sami Khuri

C) Dynamic Programming

- Dynamic programming provides a reliable and optimal computational method for aligning DNA and protein sequences.
- The **optimal alignments** provide useful information to researchers, who make **functional**, **structural**, and **evolutionary predictions** of the sequences.

©2018 Sami Khuri

Needleman Wunsch: Example



©2018 Sami Khuri

Local Sequence Alignment

- Their dynamic algorithm gives a **global alignment** of sequences.
- A modification of the dynamic programming algorithm for sequence alignment provides a **local sequence alignment** giving the highest-scoring local match between two sequences (Smith and Waterman 1981).
- **Local alignments** are usually more meaningful than global matches because they include patterns that are conserved in the sequences.

©2018 Sami Khuri

Local Alignments

	G	A	A	C	G	T	A	G	G	C	G	T	A	T		
o	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
A	0	0	1	1	0	0	0	1	0	0	0	0	0	1	0	
T	0	0	0	0	0	0	1	0	0	0	0	0	1	0	2	
A	0	0	1	1	0	0	0	2	0	0	0	0	0	0	2	0
C	0	0	0	0	2	0	0	0	1	0	1	0	0	0	1	0
T	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0
A	0	0	1	1	0	0	0	2	0	0	0	0	0	0	2	0
C	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0
G	0	1	0	0	0	3	1	0	1	1	0	2	0	0	0	0
G	0	1	0	0	0	1	2	0	0	2	0	1	1	0	0	0
A	0	0	2	1	0	0	0	3	1	0	0	0	0	0	2	0
G	0	1	0	0	0	1	0	0	4	2	0	1	0	0	0	0
G	0	1	0	0	0	1	0	0	1	5	3	1	0	0	0	0
G	0	1	0	0	0	1	0	0	1	2	4	4	2	0	0	0

Thus, the best local alignment achieved from the above Dynamic Programming is:

A C G T A G G

©2018 Sami Khuri

Scoring Systems

- Use of the **dynamic programming** method requires a scoring system for
 - the comparison of symbol pairs (**nucleotides** for DNA sequences & **amino acids** for protein sequences),
 - a scheme for insertion/deletion (gap) penalties.
- The most commonly used scoring systems for protein sequence alignments are the log odds form
 - of the **PAM250** matrix and
 - the **BLOSUM62** matrix.
- A number of other choices are available.

©2018 Sami Khuri

Successful Substitution Matrices

- The most successful **substitution matrices** are the ones that use actual evidence of what has happened during evolution, and are based on the analysis of alignments of numerous homologs of well-suited proteins from many different species.

Understanding Bioinformatics by Zvelebil and Baum

©2018 Sami Khuri

Scoring Matrices (I)

- Upon evaluating a sequence alignment, we are really interested in knowing whether the alignment is random or meaningful.
- A **scoring matrix** (table) or a **substitute matrix** (table) is a table of values that describe the probability of a residue (amino acid or base) pair occurring in an alignment.

©2018 Sami Khuri

Scoring Matrices II

- The alignment algorithm needs to know if it is more likely that a given amino acid pair has occurred **randomly** or that it has occurred as a result of an **evolutionary** event.
- Similar amino acids are defined by high-scoring matches between the amino acid pairs in the substitution matrix.

©2018 Sami Khuri

Amino Acid Substitution Matrices (I)

- For proteins, an **amino acid substitution matrix**, such as the Dayhoff Percent Accepted Mutation matrix 250 (**PAM250**) or **Block Substitution Matrix 62** (**BLOSUM62**) is used to score matches and mismatches.
- Similar matrices are available for aligning DNA sequences.

©2018 Sami Khuri

Amino Acid Substitution Matrices (II)

- In the **amino acid substitution matrices**, amino acids are listed both across the top of a matrix and down the side, and each matrix position is filled with a score that reflects how often one amino acid would have been paired with the other in an alignment of related protein sequences.

©2018 Sami Khuri

PAM Matrices

Point Accepted Mutation

- An **accepted mutation** is any mutation that doesn't kill the protein or organism; that is, amino acid changes "accepted" by natural selection.

One PAM (PAM1) = 1% of the amino acids have been changed.

©2018 Sami Khuri

Dayhoff Amino Acid Substitution Matrices

- **PAM Matrices** are Dayhoff amino acid substitution or percent accepted mutation matrices.
- This family of matrices lists the likelihood of change from one amino acid to another in homologous protein sequences during evolution.
- These predicted changes are used to produce **optimal alignments** between two protein sequences and to score the alignment.

©2018 Sami Khuri

Extrapolating PAM1

The assumption in this evolutionary model is that the amino acid substitutions observed over short periods of evolutionary history can be extrapolated to longer distances.

©2018 Sami Khuri

Constructing More PAM Matrices

- The **PAM1** Matrix is best used for comparing sequences where 1% or less of the amino acids have changed.
- What do you do with sequences that are more divergent?
- You multiply the PAM1 matrix by itself N times to get a new matrix that works best with sequences that have PAM2, PAM20, PAM100, PAM200, etc.
- For example $PAM20 = (PAM1)^{20}$

©2018 Sami Khuri

PAM Matrices for Low Level of Similarities

- As seen, **PAM1** matrix could be multiplied by itself N times, to give transition matrices for comparing sequences with lower and lower levels of similarity due to separation of longer periods of evolutionary history.
- The PAM120, PAM80, and PAM60 matrices should be used for aligning sequences that are 40%, 50%, and 60% similar, respectively.

©2018 Sami Khuri

PAM250 Matrix

- The PAM250 matrix provides a better-scoring alignment than lower-numbered PAM matrices for distantly related proteins of 14-27% similarity.
- Scoring matrices are also used in database searches for similar sequences.

©2018 Sami Khuri

BLOSUM vs PAM

- The BLOSUM matrix was constructed from actual substitutions.
- The BLOSUM matrix was derived much more recently than the Dayhoff matrices, in the early 1990's, using local multiple alignments rather than global alignments.

Understanding Bioinformatics by Zvelebil and Baum

©2018 Sami Khuri

BLOSUM Matrices

- The **BLOSUM** scoring matrices (especially BLOSUM62) appear to capture more of the distant types of variations found in protein families.
- Another criticism: PAM scoring matrices are not much more useful for sequence alignment than simpler matrices, such as the ones based on chemical grouping of amino acid side chains.

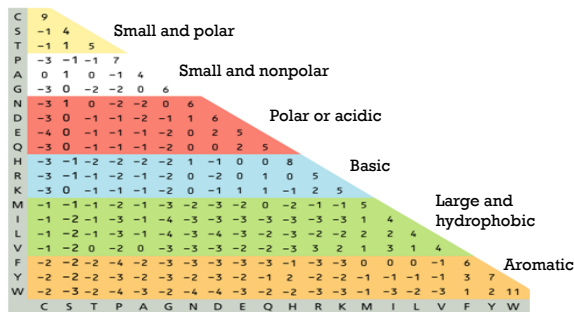
©2018 Sami Khuri

BLOSUM

- **Block** Substitution **M**atrix
 - created from BLOCKS database.
- Currently the most widely used comparison matrix.
- More sensitive than PAM or other matrices
- Finds more sequences that are related
- The BLOSUM matrices are based on an entirely different type of sequence analysis and a much larger data set than the Dayhoff PAM Matrices.

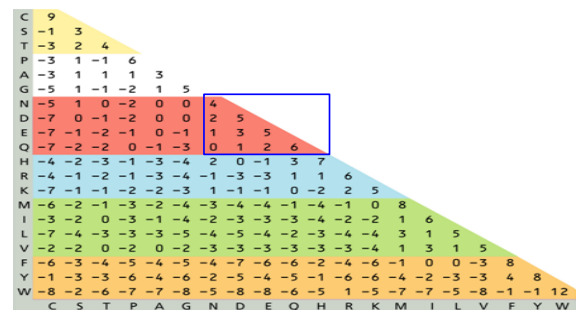
©2018 Sami Khuri

BLOSUM62 Substitution Matrix



©2018 Sami Khuri

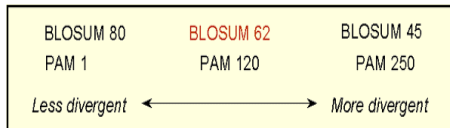
PAM120 Substitution Matrix



©2018 Sami Khuri

Comparison: PAM and BLOSUM Matrices

The **PAM** model is designed to track the evolutionary origins of proteins, whereas the **BLOSUM** model is designed to find their conserved domains.



©2018 Sami Khuri

D) Approximate Methods BLAST

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
– Altschul et al. 1990, 1994, 1997
- Heuristic method for local alignment
- Designed specifically for database searches
- Idea: Good alignments contain short lengths of exact matches.

©2018 Sami Khuri

The BLAST Family

- **blastp**: compares an amino acid query sequence against a protein sequence database.
- **blastn**: compares a nucleotide query sequence against a nucleotide sequence database.
- **blastx**: compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

©2018 Sami Khuri

Chance or Homology?

- In all methods of sequence comparison, the fundamental question is whether the similarities perceived between two sequences are due to chance, and are thus of little biological significance, or whether they are due to the derivation of the sequences from a common ancestral sequence, and are thus homologous.

Understanding Bioinformatics by Zvelebil and Baum

©2018 Sami Khuri

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query Algorithms: blastn , megablast, discontinuous megablast
protein blast	Search protein database using a protein query Algorithms: blastp, psi-blast, phi-blast
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Blastn: general purpose nucleotide search and alignment program to align tRNA or rRNA sequences as well as mRNA or genomic DNA sequences containing a mix of coding and noncoding regions.

©2018 Sami Khuri

The Expected Value

SW:P11A	BOVIN	P32871	PHOSPHATIDYLINOSITOL 3-KINAS	(1068)	2228	493	1.2e-138
SW:P11A	HUMAN	P42336	PHOSPHATIDYLINOSITOL 3-KINAS	(1068)	2216	490	7.4e-138
SW:P11A	MOUSE	P42337	PHOSPHATIDYLINOSITOL 3-KINAS	(1068)	2204	488	4.5e-137
SW:P11B	HUMAN	P42338	PHOSPHATIDYLINOSITOL 3-KINAS	(1070)	1126	254	1.1e-66

The **e-value** tells us how likely it is that the similarity between the query sequence and the database sequence is due to chance.

The lower the **e-value**, the more likely it is that the two sequences are truly similar and not just chance matches

©2018 Sami Khuri