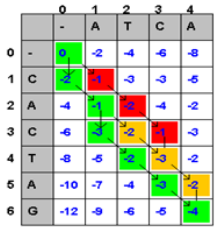
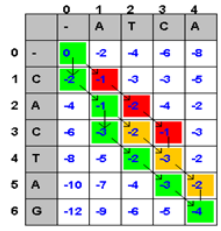
Algorithms in Bioinformatics
FOUR
Pairwise Sequence Alignment

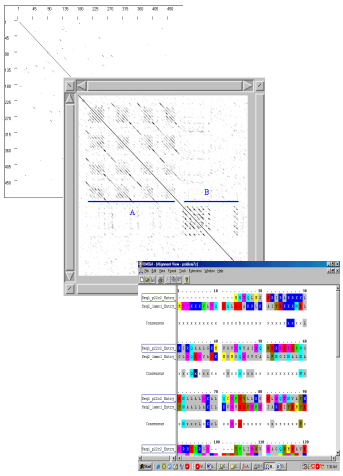


Sami Khuri
Department of Computer Science
San José State University

sami.khuri@sjsu.edu



Pairwise Sequence Alignment



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

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**.

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.

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

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.

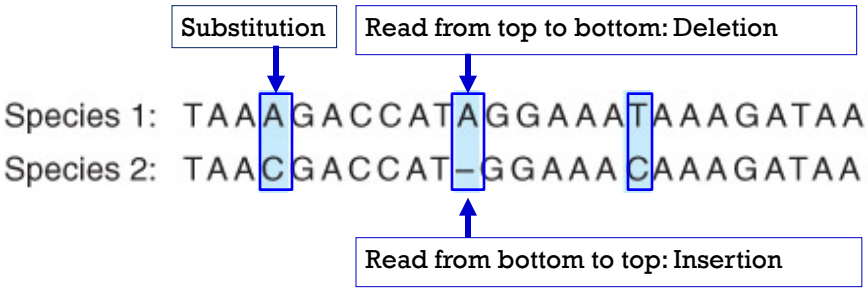
Similarity and Difference

- 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.

Evolution and Alignments

- Alignments reflect the probable evolutionary history of two sequences.
- Residues that align and that are not identical represent substitutions.
- Sequences without correspondence in aligned sequences are interpreted as indels and in an alignment are gaps.

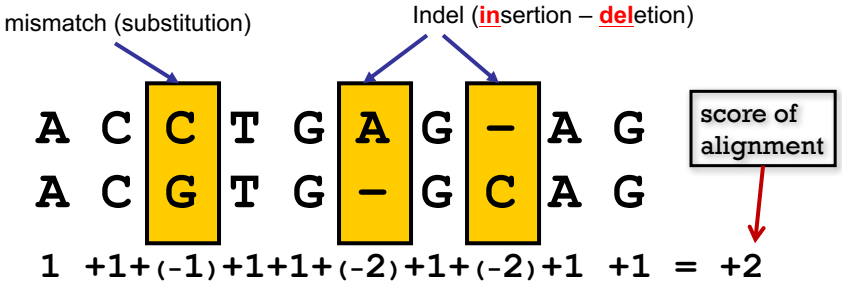
How do we Compare Sequences?



Determining the similarity of two genes by aligning their nucleotide sequences as well as possible; the differences due to mutation are shown in boxes.

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.

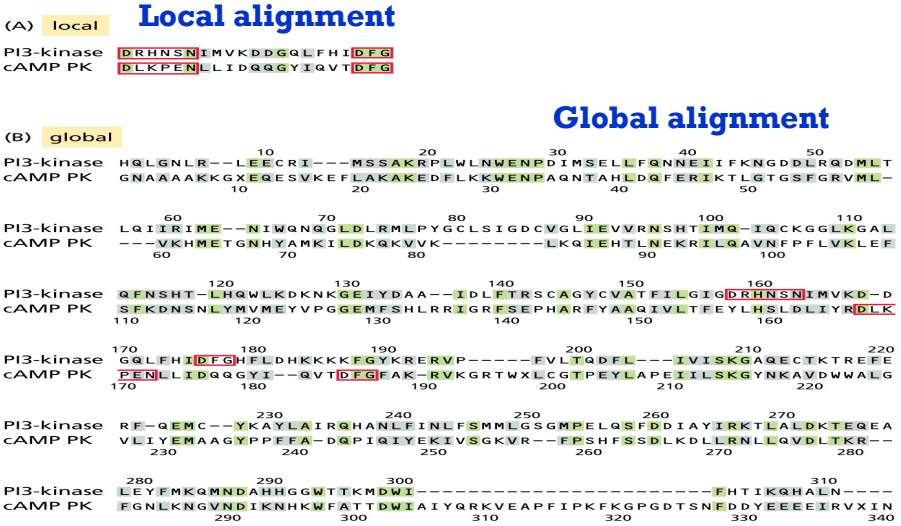
Local and Global Alignments

- 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.

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.

Example: Local and Global Alignments

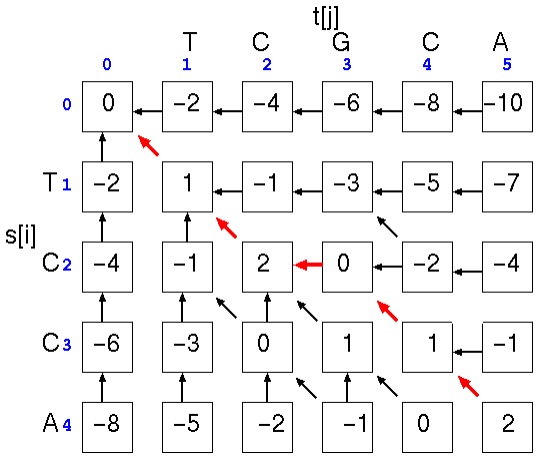


Understanding Bioinformatics by Zvelebil and Baum

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.

Needleman Wunsch: Example



Find the optimal alignment between: TCCA and TCGCA

Scoring Function:
+1 for match
-1 for mismatch
-2 for gap

Solution:

```
T C - C A
: :   : :
T C G C A
1+1-2+1+1 = 2
```

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.

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.

Scoring Matrices

- 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.

Amino Acid Substitution Matrices (I)

- For proteins, an **amino acid substitution matrix**, such as the Dayhoff **P**ercent **A**ccepted **M**utation matrix 250 (**PAM250**) or **Blo**ck **Su**bstitution **M**atrix 62 (**BLOSUM62**) is used to score matches and mismatches.
- Similar matrices are available for aligning DNA sequences.

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.

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.

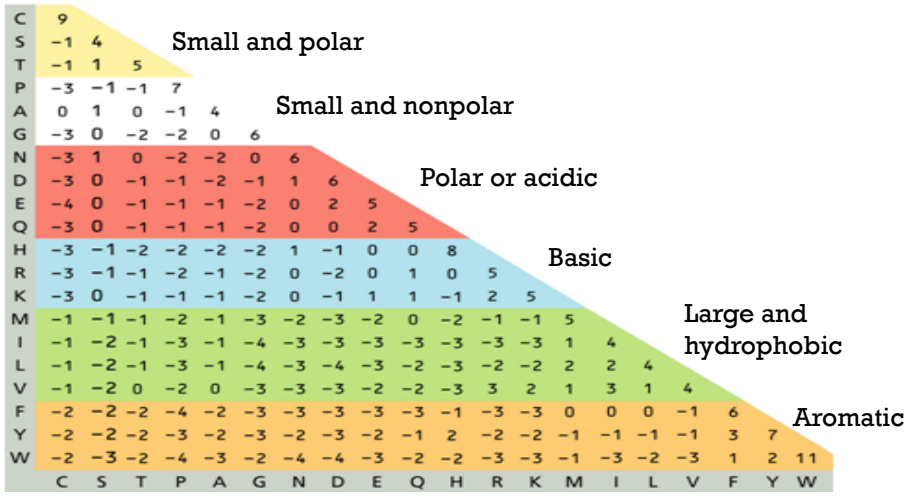
Constructing More PAM Matrices

- 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 = $(PAM1)^{20}$

BLOSUM

- **Blo**ck **S**ubstitution **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.

BLOSUM62 Substitution Matrix



Colors indicate different physicochemical properties of the residues.

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.

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent	⟷	*More divergent*

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.

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.

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

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