

Lecture – Sequence analysis

Sequence analysis

Sequence alignment

Biological sequences evolved by evolution.

Why compare sequences?

- ✓ We often analyse a sequence **by aligning** to one or multiple sequences.
- ✓ This provides information about **homology**.
- ✓ Thus, we can **infer structure/function** using the similarities.



- ✓ Given a new sequence, infer its **function** based on similarity to **another sequence**
- ✓ Find **important molecular regions** – conserved across species
- ✓ Determine the **evolutionary constraints** at work
- ✓ Find **mutations** in a population or family of genes
- ✓ Find **similar looking sequence** in a database
- ✓ Find **secondary/tertiary structure** of a sequence of interest – molecular modeling using a template (homology modeling)

Sequence analysis

Sequence alignment

Biological sequences evolved by evolution.

Are two sequences related?

- ✓ Align sequences or parts of them
- ✓ Decide if alignment is by chance or evolutionarily linked?



Issues?

- ✓ What sorts of alignments to consider?
- ✓ How to **score an alignment** and hence **rank**?
- ✓ **Algorithm** to find **good alignments**
- ✓ Evaluate the significance of the alignment

Sequence analysis

Sequence alignment

How to align sequences?

✓ Using matrix

Sequence 1

AGGCTATCACCTGACCTCCAGGCCGATGCCC

Sequence 2

TAGCTATCACGACCGCGGGTCGATTGCCCCGAC

Aligned sequences:

Sequence 1

-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---

Sequence 2

TAG-CTATCAC--GACCGC--GGTCGATTGCCCCGAC

Gaps

Sequence analysis

Sequence alignment

- Sequence comparison lies at the heart of Bioinformatics analysis.
- It is an important first step toward **structural and functional analysis** of newly determined sequences.
- As **new biological sequences** are being generated at **exponential rates**, sequence **comparison** is becoming increasingly important to draw **functional and evolutionary inference** of a **new** protein with proteins **already existing** in the database.
- The most fundamental process in this type of comparison is **sequence alignment**.
- This is the process by which sequences are compared by searching for common character patterns and establishing residue–residue correspondence among related sequences.
- **Pairwise sequence alignment** is the process of aligning two sequences and is the basis of database similarity searching and **multiple sequence alignment**.

Sequence analysis

Sequence alignment - Basics

EVOLUTIONARY BASIS



- **DNA and proteins** are products of evolution.
- The **building blocks** of these biological macromolecules, nucleotide bases, and amino acids form **linear** sequences that determine the **primary structure** of the molecules.
- These molecules can be considered **molecular fossils** that **encode the history** of millions of years of evolution.
- During this time period, the **molecular sequences undergo random changes**, **some** of which are **selected** during the process of evolution.
- As the selected sequences **gradually accumulate mutations** and **diverge** over time, traces of evolution may still remain in certain portions of the sequences to allow **identification** of the **common ancestry**.
- The presence of evolutionary traces is because some of the residues that perform **key functional** and structural roles tend to be preserved by **natural selection**; other residues that may be less crucial for structure and function tend to mutate more frequently.

Sequence analysis

Sequence alignment - Basics

EVOLUTIONARY BASIS



Example

- For example, **active site residues of an enzyme family tend to be conserved** because they are responsible for catalytic functions.
- Therefore, **by comparing sequences** through alignment, **patterns of conservation and variation can be identified.**
- The degree of **sequence conservation in the alignment reveals evolutionary relatedness** of different sequences,
- whereas the **variation** between sequences reflects the **changes that have occurred during evolution** in the form of **substitutions, insertions, and deletions.**

Sequence analysis

Sequence alignment - Basics

EVOLUTIONARY BASIS



How to predict structure and function?

- Identifying the evolutionary relationships between sequences **helps to characterize the function** of unknown sequences.
- When a sequence alignment reveals **significant similarity among a group of sequences**, they can be considered as **belonging to the same family**.
- If **one member** within the family has a **known structure and function**, then that information can be transferred to those that have not yet been experimentally characterized.
- Therefore, sequence alignment can be used as **basis for prediction of structure and function of uncharacterized sequences**.

Sequence analysis

Sequence alignment - Basics

EVOLUTIONARY BASIS



How to predict common evolutionary origin?

- Sequence alignment provides inference for the **relatedness of two sequences** under study.
- If the two sequences share **significant similarity**, it is extremely unlikely that the extensive similarity between the two sequences has been acquired randomly, meaning that the **two sequences must have derived from a common evolutionary origin**.
- When a sequence alignment is generated correctly, it **reflects the evolutionary relationship** of the two sequences:
 - regions that are **aligned but not identical** represent **residue substitutions**;
 - regions where **residues from one sequence correspond to nothing in the other** represent **insertions or deletions** that have taken place on one of the sequences during evolution.
- It is also possible that **two sequences have derived from a common ancestor**, but **may have diverged to such an extent** that the common ancestral relationships are not recognizable at the sequence level.
- In that case, the distant evolutionary relationships have to be detected using other methods.

Sequence analysis

Sequence alignment

SEQUENCE HOMOLOGY VERSUS SEQUENCE SIMILARITY



Homology?

- An important concept in sequence analysis is **sequence homology**.
- When **two sequences are descended from a common evolutionary** origin, they are said to have a homologous relationship or **share homology**.

Similarity?

- A related but different term is **sequence similarity**, which is the **percentage of aligned residues that are similar** in physiochemical properties such as size, charge, and hydrophobicity.

Sequence analysis

Sequence alignment

SEQUENCE HOMOMOLOGY VERSUS SEQUENCE SIMILARITY



Homology versus similarity?

- It is important to distinguish sequence homology from the related term sequence similarity because the two terms are often confused by some researchers who use them interchangeably in scientific literature.
- To be clear, **sequence homology is an inference** or a conclusion about a common ancestral relationship drawn from sequence similarity comparison when the two sequences share a high enough degree of similarity.
- On the other hand, **similarity is a direct result of observation from the sequence alignment**.
- Sequence **similarity can be quantified** using percentages; **homology is a qualitative** statement.

Example



- ✓ For example, **one may say that two sequences share 40% similarity**.
- ✓ **It is incorrect to say that the two sequences share 40% homology.**
They are either homologous or nonhomologous.

Sequence analysis

SEQUENCE HOMOLOGY VERSUS SEQUENCE SIMILARITY



Sequence alignment

What we call as homologous relationship?

- Generally, if the sequence similarity level is high enough, a common evolutionary relationship can be inferred.
- In dealing with real research problems, the issue of at what similarity level can one infer homologous relationships is not always clear.
- The answer depends on the type of sequences being examined and sequence lengths.

Nucleotide



- ✓ Nucleotide sequences consist of only **four characters**, and therefore, **unrelated sequences have at least a 25% chance of being identical**.

Proteins



- ✓ For protein sequences, there are **twenty possible amino acid residues**, and so **two unrelated sequences can match up 5% of the residues by random chance**.



Sequence analysis

SEQUENCE HOMOMOLOGY VERSUS SEQUENCE SIMILARITY



Sequence alignment

What we call as homologous relationship?

Nucleotide



- ✓ Nucleotide sequences consist of only **four characters**, and therefore, **unrelated sequences have at least a 25% chance of being identical**.

Proteins



- ✓ For protein sequences, there are **twenty possible amino acid residues**, and so **two unrelated sequences can match up 5% of the residues by random chance**.



Sequence length



- ✓ If **gaps are allowed**, the percentage could increase to **10–20%**.
- ✓ Sequence **length** is also a **crucial** factor.
- ✓ The **shorter the sequence**, the **higher the chance** that some alignment is attributable to **random chance**.
- ✓ The **longer the sequence**, the **less likely** the **matching** at the same level of similarity is attributable to random chance.





Sequence analysis

SEQUENCE HOMOLOGY VERSUS SEQUENCE SIMILARITY



Sequence alignment

What we call as homologous relationship?

- Sequence length  ✓ This suggests that shorter sequences require higher cutoffs for inferring homologous relationships than longer sequences.
- ≥30% identity  ✓ For determining a homology relationship of two protein sequences, for example, if both sequences are aligned at full length, which is 100 residues long, an identity of 30% or higher can be safely regarded as having close homology.
- 20-30% identity  ✓ If their identity level falls between 20% and 30%, determination of homologous relationships in this range becomes less certain.
✓ In this area remote homologs mix with randomly related sequences.
- ≤20% identity  ✓ Below 20% identity, where high proportions of nonrelated sequences are present homologous relationships cannot be reliably determined.

Note: ✓ Note that the percentage identity values only provide a tentative guidance for homology identification.

Sequence analysis

SEQUENCE SIMILARITY VERSUS SEQUENCE IDENTITY



Sequence alignment

Similarity versus identity?

Another set of related terms for sequence comparison are sequence similarity and sequence identity.

Nucleotide



Sequence similarity and sequence identity are synonymous for nucleotide sequences.

Protein



- For protein sequences, however, the two concepts are very different.
- In a protein sequence alignment, **sequence identity refers to the percentage of matches of the same amino acid residues** between two aligned sequences.
- **Similarity refers to the percentage of aligned residues that have similar physicochemical characteristics** and can be more readily substituted for each other.

Sequence analysis

Sequence alignment

SEQUENCE SIMILARITY VERSUS SEQUENCE IDENTITY



How to calculate similarity and identity?

Two methods



- There are two ways to calculate the sequence similarity/identity.
- One involves the use of the overall sequence lengths of both sequences;
- the other normalizes by the size of the shorter sequence.

The first method uses the following formula:

Percentage sequence similarity (S)



$$S = [(L_s \times 2) / (L_a + L_b)] \times 100$$

- ✓ where **S** is the percentage sequence similarity,
- ✓ **L_s** is the number of aligned residues with similar characteristics,
- ✓ and **L_a** and **L_b** are the total lengths of each individual sequence.

Percentage sequence identity (I)



$$I = [(L_i \times 2) / (L_a + L_b)] \times 100$$

- ✓ where **L_i** is the number of aligned identical residues.

Sequence analysis

Sequence alignment

SEQUENCE SIMILARITY VERSUS SEQUENCE IDENTITY



How to calculate similarity and identity?

Two methods



- There are two ways to calculate the sequence similarity/identity.
- One involves the use of the overall sequence lengths of both sequences;
- the other normalizes by the size of the shorter sequence.

The **second method** of calculation is to derive the percentage of identical/similar residues over the full length of the smaller sequence using the formula:

Percentage sequence identity **I** (or similarity **S**)



$$I(S)\% = L_{i(s)} / L_a \%$$

✓ where **L_a** is the length of the shorter of the two sequences.

- ✓ where **S** is the percentage sequence similarity,
- ✓ **I** is the percentage sequence identity
- ✓ **L_s** is the number of aligned residues with similar characteristics,
- ✓ **L_i** is the number of aligned identical residues.