Here are some responses to Questions asked during the Introductory Session of previous runs of the course you now attend.

In hope you find them useful,

Dave and Pedro

Query: Can you please explain the difference between **Similarity** and **Homology** in **MSA**? ….

Briefly: *In all contexts of sequence alignment, **Similarity** is a measure of how well the aligned sequences match. **Homology** is an assertion that the **Proteins/DNA** represented by the aligned sequences are related by evolution.*

In full: In this context, **Similarity** is a quantitative evaluation of the degree to which a set of **Sequences** are similar. This measure typically becomes interesting when it is significantly greater than that which might be expected by chance.

Homology is the assertion that a given set of **Proteins**/**DNA**/**RNAs** are Biologically related by evolution[1]. That is, all the elements of the set evolved directly from a single origin. **Homology** is a binary judgement (i.e. not **quantitative**!). A set of **Proteins**/**DNA**/**RNAs** are either Biologically related by evolution (i.e. **Homologous**) or they are not.

Where **Proteins**/**DNA**/**RNAs** are judged to be **Homologous**, ***ALL*** differences between their aligned **Sequences** should be exclusively due to evolutionary processes (e.g. **Substitutions**, **Insertions**, **Deletions** …).

In all contexts (not just **M**ultiple **S**equence **A**lignment, **MSA**), **Similarity**, typically represented as alignments between **Sequences** (simple, incomplete ***representations*** of **Proteins**, **DNA**, **RNA**) is an association computed, most commonly, by computer programs.

An alignment of Sequences (**Pairwise** or **Multiple**) is but an attractively intuitive way to display **Sequence** data in a fashion that highlights regions that are **Similar** beyond random expectation, *maybe* for some reason of Biological significance (e.g. evolutionary **Conservation**, the sequences represent different instances of a **repeated region**) … or just by chance.

Whether **Homology** exists, or not, between a given set of **Proteins**/**DNA**/**RNA** is a judgement that can only be made by an informed human(s) after careful reference to all the evidence. Such evidence will usually include the degree of **Similarity** indicated by carefully computed alignments of the relevant **Sequences**.

A computer program has no way to directly infer **Homology** between **Proteins**/**DNA**/**RNA** from **Similarity**, however complete/perfect that **Similarity** might be.

Only informed people, assisted by real Biological research, can meaningfully speculate upon **Homology**. **Similarity** between **Sequences** can but be a part of the evidence for claiming **Homology** between **Proteins**/**DNA**/**RNA**.

Aligning the **Sequences** of biological entities that are not **Homologous** is not usually sensible. In particular, **Alignment Programs** are written to assume the sequences they process represent **Homologous** entities. What can the **Alignment** mean if this is not true?

---

1 We both agree that:

Sequences can be **Similar** (or even **Identical**) but they **CANNOT** be **Homologous**!!!!

Proteins, for example, can be **Homologous**. That is, two proteins ***CAN*** evolve from a common ancestor, but **Protein Sequences** are merely ***representations*** of **Proteins** and so cannot evolve In any sense whatsoever!

Many of the links we have provided refer to "**Sequence Homology**". We think this is inaccurate, but this we overlook as otherwise, the links are useful.

Pedantic? Possibly, but we are convinced we are right to be pedantic!

Query: …. and how can we deduct with a precision that two or more sequences are homologous?

Briefly: *Just from the alignment, you cannot, although a highly **similar** alignment can be taken as a pretty strong indication of **homology**. To be confident that a promising alignment truly indicates **homology**, you must know something about the **Proteins**/**DNA** represented by the aligned sequences. **Bioinformatics** rarely completely replaces **Biology**.*

In full: Well … there is certainly no universal formula for inferring **Homology** from **Similarity**, especially if you desire "*precision*". Specifically, there is no objective level of **Similarity** between **Sequences** that assures **Homology** between the **Proteins**/**DNA**/**RNAs** that those sequences represent.

Convincingly high **Similarity** evident in **Sequence Alignments** is commonly indicative of **Protein**/**DNA**/**RNA Homology**. However, **Sequence Alignments** for **Proteins**/**DNA**/**RNAs** that are not **Homologous** can appear temptingly **Similar**. Also, **Sequence Alignments** for **Proteins**/**DNA**/**RNAs** that ***are*** **Homologous** can appear surprisingly dissimilar. Reasons include:

- **DNA**/**RNA** only uses a four letter alphabet. Relatively small shared compositional bias between a set of **Sequences** is going to start looking "interesting" unreasonable early.
- **Compositional Bias** even in **Protein Sequences** can look a bit more like **Homology** than perhaps it deserves.
- **DNA**/**RNA** sequences can be quite different and yet code for very **Similar**, genuinely **Homologous**, proteins due to **Redundancy in the Genetics code**. It is quite possible to ***MISS*** more distant protein **Homology** when trying to make a judgement based on **DNA**/**RNA Sequence** alignments. This is why it is always so much better to align ***PROTEIN*** **Sequences** rather than **DNA**/**RNA Sequences** wherever it is possible. **Protein Sequences** encapsulate much more information than do **DNA**/**RNA Sequences**.
- **Convergent Evolution** can occur to produce apparent similarity between proteins that are evolutionarily unrelated but perform similar functions and have similar structures.

Query: Can you explain briefly the notions **Homologues**, **Orthologues** and **Paralogues** (genes)?

Resply: "briefly"? … afraid this half of Pedro and Dave is not known for his close acquaintance with "briefly", but one will do one's very best :-)

To say two **Genes**/**Proteins** are **Homologues** is to say that they are related by **Evolution**. That is, multiple copies emerged from a single common ancestor **Gene**/**Protein** and then evolved apart (generally) from each other whilst (normally) retaining essential original elements of function and/or structure.

Homologues are also **Paralogues** if they exist separately exclusively because of duplication event(s). Specifically, no **Speciation** event(s) were involved in the creation of multiple instances of the original single **Gene**/**Protein**. This being so, **Paralogues** are **Homologues** that occur always in the ***same*** species.

Homologues are also **Orthologues** if they exist separately because of duplication event(s) plus at least one **Speciation** event. This being so, **Orthologues** are **Homologues** that occur always in ***different*** species.

Query: What is the difference between **Bioinformatics** and **Computational Biology**?

Briefly: ***Bioinformatics*** *is the application of* ***Mathematics***, ***Statistics*** *and* ***Computing Science*** *to achieve a better understanding of the information encoded in biological molecules (****Protein****/****DNA****/****RNA****).*

Computational Biology *includes all aspects of* ***Bioinformatics***, *emphasising the development of algorithms and software tools.*

In full: A number of different terms to define the meeting of Computing Science/Informatics and Biology have emerged in recent years. Terms including **Bioinformatics**, **Computational Biology**, **Biological Computation** …

The distinctions between these terms typically require a bit of thought. The first step of finding clear definitions is not always trivial. Here your query concerns specifically the distinction between **Bioinformatics** and **Computational Biology**.

For **Bioinformatics**, there exist many definitions exhibiting variation of emphasis. Generally though, **Bioinformatics** is the application of **Mathematics**, **Statistics** and **Computing Science** to the achievement of a better understanding of the information encoded in biological molecules (**Protein**/**DNA**/**RNA**). A broad definition that allows **Bioinformatics** to encompass a very wide range of activities in very many forms of Biological research.

Computational Biology can be thought of as including all aspects of **Bioinformatics**, emphasising the development of algorithms and software tools.

Computational Biology also encompasses the construction of mathematical models and the use of simulations to study biological, behavioural and social systems. These, not being concerned with the information derived from biological molecules, are generally considered outside the definition of **Bioinformatics**. This implies that all of **Bioinformatics** forms a substantial subset of **Computational Biology**.

When attempting to determine whether a particular activity is **Bioinformatics**, **Computational Biology** or **both**, it is often more useful to consider that which is being analysed rather than the analysis itself. Studying Biology using **Bioinformatics** always involves the analysis of biological information encoded in the molecules of life (**DNA**, **RNA**, **Proteins**) and is also always **Computational Biology**. Using Computational methods to study any other form of Biological data/information is not **Bioinformatics** but could be **Computational Biology**.

Examples abound. For example:

- When studying **Ecology**, information encoded in specific organisms is not considered directly. However, computation and modelling using observed quantities such as areas/weight of biomass, temperatures, pH, etc will be central. **Computational Biology** but not **Bioinformatics**!
- Studying **Biological** (e.g. **Membrane**) **Transport** using numerical models that work with concentrations, fluxes, electrical charge, etc. and ***not*** the amino acids of the proteins that form pores in membranes. **Computational Biology** but not **Bioinformatics**!