Here are some responses to **Questions** asked during the **Introductory Session** of this and previous runs of the course you now attend. Some answers are quite long and "***in depth***". Where this is so, we have provided both a "***Short***" and simple answer **AND** a "***Longer***" extended answer. If you find the "***Short*** " answer sufficient, that is fine. Read the "***Longer***" version only if it amuses you.

We have tried to answer all the **Questions** that were asked of us as fully as we can. Do remember though, that many of the **Questions** you asked will be considered again in more specialised sessions of the course still ahead of you. If you find our answers insufficient, why not ask the presenter(s) of the more targeted session when it arrives.

In hope you find these **Answers** useful,

Dave and Pedro

PDF generated ***2020-06-25 01:10:01***

QUESTION INDEX

Broadly Informatics Questions

- Are the **Linux** and **Windows Operating Systems** compatible? ***Answer***
- Can I install both **Linux *AND* Windows** on the same machine? ***Answer***
- What is **Cygwin**? ***Answer***

Return to Main Index

Biological Queries

Homology, Parology, Othology:

- Explain the difference between **Similarity** and **Homology** in **MSA**? *Answer*
- … and how can we deduce, with a precision, that two or more sequences are **Homologous**? *Answer*
- Explain briefly the term **Homologues**, **Orthologues** and **Paralogues**? *Answer*
- What is the difference between a **Protein Domain** and a **Motif**? *Answer*

Sequence Alignment:

- Explain the differences between **Pairwise Alignmen**t, **Global Alignment** and **Local Alignment**? *Answer*
- When would a **Local** approach be preferable to **Global** approach for a **Pairwise Alignment**? *Answer*
- Which is the best software tool to compute **Pairwise Sequence Alignments**? *Answer*
- Which is the best software tool to compute **Multiple Sequence Alignments** (**MSAs**)? *Answer*

General:

- Is it possible for one to create a **Personal Profile** at **NCBI**, and/or any other site providing similar services, enabling search results (and similar analytical computations) for future references? *Answer*
- What is the difference between **Bioinformatics** and **Computational Biology**? *Answer*

Return to Main Index

Query: Are the **Linux** and **Windows Operating Systems** compatible?

Reply: These two **Operating Systems** are not really compatible, as they manage and explore the hardware differently and use incompatible mechanisms to host applications. In other words you cannot expect to move around seamlessly. Specifically, software developed and installed in one system will not necessarily just run it in the other. A variety of workarounds do exist to reduce the impact of this incompatibility, however, generally at the expense of resources such as memory and disk space, resulting in significantly reduced program execution speeds.

Return To Index

Query: Can I install both **Linux *AND* Windows** on the same machine?

Reply: Yes, just not one on top of the other. And in general terms it only requires free software and some labour.

- If you have a comfortable amount of memory and disk space, from either **Linux** or **Windows**, you can use a technique called **Virtualisation**, which involves installing a **Virtual Machine (VM)** that runs the alternative **Operating system**. **Virtual machines (VMs)** are managed by specialised software such as **Oracle VM Virtualbox**. It is a good idea to have a shared disk area, such as the **Downloads** folder. This solution may slow down things, and you may be encouraged to add memory to your computer. The two operating systems will be sharing hardware resources all the time.

- A technique called **Containerisation** also virtualises a machine with a different **Operating System**. A container carries the **Operating System**, installed applications and data, and can be moved around seamlessly in heterogeneous systems. The most commonly used software to do this is **Docker**. The containers do not run autonomously as **VMs**, they share software resources with the native **Operating System** that hosts them.

- Alternatively, you can create a separate **Disk Partition** and install the second **Operating System** there. In this case they will not be running simultaneously, and as you will be starting one **OS** or the other (so, to change **Operating Systems** for always involve a reboot of the entire machine), this technique is called "**Dual Boot**". Sharing files between the systems is a bit harder to set-up.

Return To Index

Query: What is **Cygwin**?

Reply: Cygwin is just a collection of open source tools that operate in a **Linux-like Terminal** under **MS Windows**. You should not expect native **Linux** applications to run under **Cygwin**. Its usage is limited to trivial **Command Line** file manipulations. **Cygwin** is not suitable for most serious **Bioinformatics** work.

Return To Index

Query: 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?

Return To Index

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 deduce, 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.

Return To Index

Query: Explain briefly the term **Homologues**, **Orthologues** and **Paralogues**?

Reply: “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.

Return To Index

Query: What is the difference between a **Protein Domain** and a **Motif**?

Briefly: ***From a purely Biological perspective:***

*A **Motif** is an **Sequence pattern** or a **Structural Feature** that is to be found repeated, with significant accuracy, in many **Proteins**, **RNAs** or regions of a **Genome** (typically within **Genes**). The conservation of a **Motif** should have, or at least should be conjectured to have, some biological significance.*

*A **Protein Domain** is a conserved part of a **Protein** that can evolve, function and exist independently of the **Proteins** of which it forms part. A **Protein Domain** may occur in any number of different **Proteins** that are, outside of the extent of the **Domain**, not **Homologous**.*

*From the limited perspective of their representation in **Domain/Motif Databases**, it is tempting (if inaccurate) to consider somewhat simpler definitions. This is elaborated below.*

In full:

Motifs

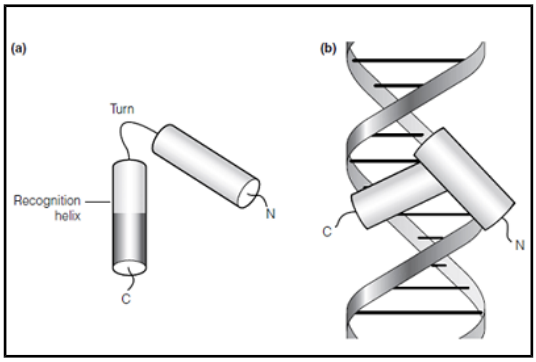
Accepting the **Biologically** complete definitions above.

Motifs can be either **Sequence Motifs** or **Structural Motifs**.

As the names suggest, a **Sequence Motif** is identified by a region of significantly conserved **Sequence** (**DNA/RNA/Protein**). A **Structural Motif** is define by a region in which **Structure** is conserved (**DNA/RNA/Protein**).

A **Sequence Motif**, might well also be a **Structural Motif**, but the definition does not insist on such.

A **Structural Motif** will often also involve high levels of **Sequence** conservation, but not necessarily. Many **Motifs** idenfied by **Structural** conservation are **NOT** well conserved at the **Sequence Level**.



Specifically, some **Structural Motifs** can be found in **non-Homologous Proteins**. In such cases, the **Structure** and biological significance of the **Motif** will be conserved, but there should be no great expectation of high **Sequence Conservation**. And example of such a **Motif** is the **Helix-Turn-Helix** (**HTH**) **Motif**, an arrangement of **2 Helices** separated by a **Turn**, that binds **DNA**.

Programs exist to try and identify the presence of **HTH motifs** in **DNA** Sequence. These programs look for **Sequence Patterns** in the **DNA** that are typical of **HTH motifs**. Due to relatively poor general **Sequence Conservation**, these programs are not very effective. To work at all, they use search models very specific to particular, narrowly defined families of **HTHs**.

All **Protein Domains** whose function includes **DNA Binding** will include one, or more, **HTH Motif** (or something very similar).

In **Sequence Motif Databases** (e.g. **Prosite**) hightly conserved **Motifs** are sometimes represented (crudely!) by consensus **Sequence Patterns** computed from **Multiple Sequence Alignments (MSAs)** of known examples of the **Motif**.

All **Motifs** represented in commonly used **Databases** are **Sequence Motifs**. It is not easy to see how conserved structure could be represented in a **Computer** usefully.

Protein Domains

Again, accepting the **Biologically** complete definitions above.

Protein Domains are generally **Homologous**, even if the entire **Proteins** in which they exist are not. **Sequence Consevation** should therefore, geneally, be expected at some level through the span of a **Protein Domain**.

Protein Domains often include various **Motifs** that, commonly define the function of the **Domain** (see example of **HTH Motifs** above).

The sequence conservation of a **Protein Domain** is, generally, not consistent over its length. There will be regions of **Low Conservation** (typically at each end). Accordingly, representing a **Protein Domain** in a database using a simple **Sequence Pattern** will almost never be sufficient. More sophisticated techniques are required.

Currently, the most popular way to represent **Domains** is to construct a probablistic model (a **Profile Hidden Markov Model**) from an alignment of known examples of the **Protein Domain**. As their name suggests, these **Domain** "***Profiles***" are constructed using the principles behind "**Hidden Markov Models**". These profiles can represent successfully regions of high or low (but not ***no***) conservation effectively.

Profile Hidden Markov Models are also used (in addition to **Sequence Patterns**) to represent **Motifs** (for example, in **Prosite**). They are far more flexible, complete and effective than **Sequence Patterns**.

Motif and Domain Databases

The main role of both **Motifs** and **Domains**, from a purely **Computational** perspective, is to represent conserved features that can be searched for in **Proteins**. This being so, **Motifs** and **Domains** need to take some form that can be matched against a **Protein(s)** to determine anywhere they might match sufficiently well to imply their presence.

The computer programs that implement these searches need no concept of "*Biological Significance*". They just seek **Similarity** between a Database entry and a Protein that is significantly beyond that expected to occur randomly.

Motif searches, generally at least, rely solely on **Sequence** matching. It is not easy to see how **Structure** might be usefully/simply represented in a computer in a way that would enable comparison with a **Protein Sequence**. It is thus tempting (if wrong) for those whose interest is focus only on the computational aspects of **Bioinformatics** to think soley of **Sequence Motifs** and conveniently forget **Structural Motifs**!

Representing **Sequence** conservation is relatively simple. A simple **Consensus Pattern** would work for very very highly conserved **Motifs**, slightly more sophistication (e.g. **Profile Hidden Markov Models**, **HMMs**) are available when conservation is less obvious.

Usefully representing **Domains** in a computer is, generally, done more honestly. The entire **Domain** (as defined properly above) is typically, represented. Almost always, there will be parts of a **Domain** that are not particularly well conserved, so any **Consensus Pattern** approach is unlikely to succeed. **HMMs** are commonly used to represent the elements of **Domain Databases**.

An exception is the **Domain Database** **PRINTS**. Here, regions of **Proteins** that correspond to **Protein Domains** by specified by arrangement of "*motifs*" referred to as "*fingerprints*".

But, the "motifs" employed by **PRINTS** are not required to be true **Biological Motifs** (as defined earlier). **PRINTS** has no need for them to have any **Biological significance**, after all. **PRINTS** only requires that its "*motifs*" be well conserved, that is, likely to match in the right place. Perhaps using the term **Motif** is not

strictly appropriate? Maybe "*signature*" or "*conserved region*" might be more exact?

Also, the "*Domains*" represented in **PRINTS** do not always correspond exactly to the **Biological** "*best specification*" of the position of a **Domain**. It does not really have to, after all. It is generally sufficient for a user to be informed that some specified **Domain** is predicted in a position including the region (the "*fingerprint*") identified by **PRINTS**.

Return To Index

Query: Explain the differences between **Pairwise Alignment**, Global Alignment and Local Alignment?

Briefly: ***Pairwise Alignment*** *is the* ***Alignment*** *of exactly* ***two*** *(a* ***pair****) of* ***DNA\RNA*** *or* ***Protein*** *sequences.*

Global Alignment *is the* ***Alignment*** *of* ***two or more DNA\RNA*** *or* ***Protein*** *sequences in which the* ***entire length*** *of* ***all*** *sequences must be included.*

Local Alignment *is the* ***Alignment*** *of* ***two or more*** *of* ***DNA\RNA*** *or* ***Protein*** *sequences in which only convincingly similar regions of* ***some or all*** *sequences need be involved.*

In full:

Pairwise Alignment is the Alignment of exactly **two** (a **pair**) of **DNA/RNA** or **Protein** sequences between which Homology is assumed for all or part of the length of the biological entities represented by those sequences.

The programs that are used to compute **Pairwise** (or **Multiple**, come to that) **Sequence Alignments** analyse only the **Sequences**, not the **DNA/RNA/Proteins** themselves. They can only consider the strings of characters that ***represent*** the biological entities. All the software has to use as a guide are some rather basic and (typically) very general assumptions concerning how the various sequence elements (**Bases/Amino Acids**) relate to each other in an evolutionary context.

For **DNA/RNA**, this amounts to almost nothing! Essentially, to the **Alignment** software, corresponding **Bases** in **Pairwise Alignments** (usually) are either seen as simply identical or different. In addition, with only a **four** letter alphabet (**A**, **C**, **G** , **T/U**), the chances of random **Alignments** looking better than they are must be high. This is especially so as **DNA/RNA** sequences are not randomly (evenly) composed (e.g. Base Bias) which most **Alignment** software does not take into consideration when evaluating the quality of an **Alignment** possibility. Thus there is a significant risk of False Positives (**Alignments** that look good, but are not).

Conversely, for coding **DNA/RNA**, due to the redundancy of the Genetic Code, quite diverse sequences can code for very similar proteins. True similarity (at the **Protein** level) can be masked at the **DNA/RNA** level. So there is also a significant chance of False Negatives (**Alignments** that look random at the **DNA/RNA** level, but not at the **Protein** level). This last observation should loudly encourage protein coding **DNA/RNA** always to be translated into **Protein** sequences before any alignment is attempted. There is much more information in a **Protein** sequence than in a **DNA/RNA** sequence.

Finally, ***ALL*** alignment programs assume that the sequences they seek to align represent **Homologous** biological entities. If this is not true, then trivially, the **Alignments** generated are worthless! The software will always produce an alignment that is arithmetically "*the best*" according to some set of simple rules. However, there is no meaningful **Alignment** to find if there is no **Homology**! It is up to the user to start with sensible data.

Global Alignment is the **Alignment** of **two or more DNA/RNA** or **Protein** sequences ***all*** assumed to represent biological entities that are **Homology** over their full length (**Globally**). Obviously, if a **Global Alignment Program** is asked to align things that are ***NOT*** **Globally Homologous**, an **Alignment** will be generated but it will be nonsense. It is up to the user to pose a sensible question before a sensible answer can be expected.

A **Global Alignment** must include the entire length of all the sequences it is given to align. Clearly, if all the sequences represents entities that are **Homologous** end to end, then it makes no sense to "*shave off*" any part of any sequence, however poorly it matches the rest of the **Alignment**.

Multiple Sequence Alignments (MSAs) computed for the determination of Phylogenetic Trees are typically **Global**. They are usually **Alignments** between the same **Gene** (or similar) from each of the organisms under study. Logically, they should match over their entire length. However, they are not always consistently impressively similar over their entire length!

All **Global Sequence Alignments** that involve only ***two*** sequences are **Pairwise Sequence Alignments**.

Local Alignment is the **Alignment** of **two or more** of **DNA** or **Protein** sequences in which **Homology** is assumed to exist between **DNA/RNA** or **Protein** regions (e.g. Domains, Sequence Motifs) represented by

the **Aligned Sequences**, but possible not over their entire length.

Local Alignment is appropriate when the purpose of the **Alignment** is to determine the position of **Domains** (or other regions conserved due to **Homology**) between two **DNA**/**RNA** regions or **Proteins**. The software is written to seek out highly Similar portions of the sequences it is considering and then to extend and align those regions only until the degree of similarity falls below some specified level. Surrounding sequence that does not match significantly can be ignored (as is not allowed for **Global Alignment**).

All **Local Sequence Alignments** that involve only ***two*** sequences are **Pairwise Sequence Alignments**.

Return To Index

Query: When would a **Local** approach be preferable to **Global** approach for a **Pairwise Alignment**?

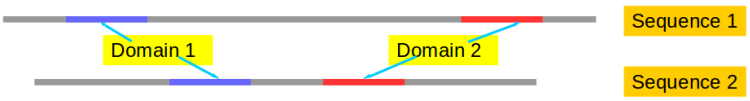
Briefly: A **Local** approach is the correct choice if the sequences to be aligned are assumed to to represent **DNA/RNA/Proteins** that share **Homologous** regions but are not necessarily **Homologous** over their entire lengths.

A **Global** approach is the correct choice only if the sequences to be aligned are assumed to represent **DNA/RNA/Proteins** that are **Homologous** over their entire lengths.

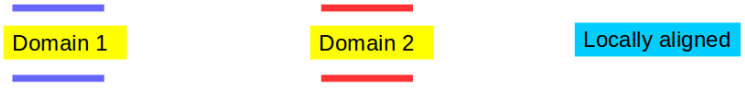
In full: I think the best way to expand upon the explanation above is with an example or two.

Example 1:

Imagine that you wish to align two **Proteins**, each of which include two domains (a **Blue** one, and a **Red** one).

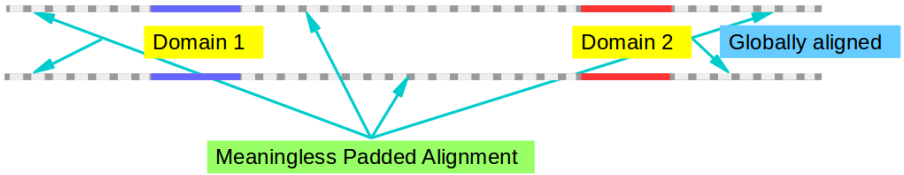


Clearly, the correct choice would be to use a **Local** strategy. Assuming each **Domain** is sufficiently similar (well conserved) in both proteins, this should result in **two** separate **Alignments**, one for each **Domain**.

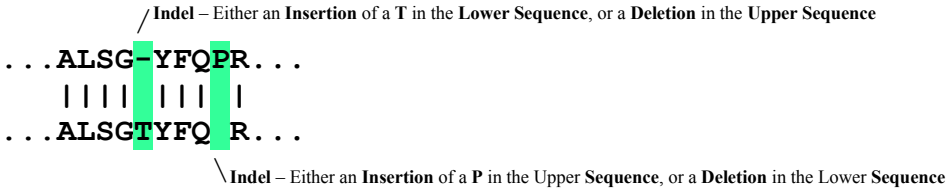


Assuming there are no serendipitously **Similar** regions in the **Non-Homologous** parts of the **Proteins**, only the **Domain** regions should be considered. The worst that could happen is that the **Domain Boundaries** are not determined with **100%** accuracy, as it is likely that **Sequence Similarity** will decline towards the extremities.

Now consider the likely result of using a **Global** approach. The software must assume the **Proteins** to be **Homologous** throughout their length (***not true!***). Assuming the **Similarity** between the **Domains** is sufficiently strong, a broadly correct alignment of the **Domains** might be achieved (as illustrated). However, the software will insist on padding out the two sequences it has been given to align, until they are the same length. The **Linker Regions** between the two **Domains** must also be padded (typically with '**-**' characters) so that they are also the same length.



The insertion of **Padding Characters** ('**-**'s) into one or both **Sequences** is the way the way the software mimics **Evolution** causing **Insertions** and/or **Deletions** (commonly referred to as **Indels**) in **Proteins**.

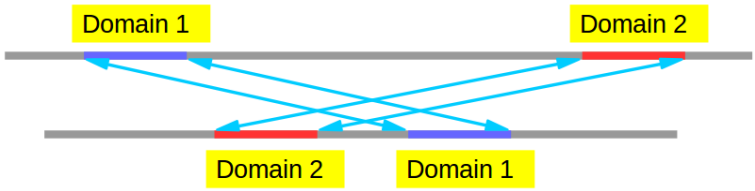


So, if you are lucky, the **Global Alignment** might tell you much that you wanted to know. However, the alignment at each end is between **Non-Homologous** stretches of protein and is thus, very probably, complete nonsense. Similarly, the **Alignment** of the **Linker Region** is without meaning.

Remember also, if the **Non-Homologous** regions are of greatly dissimilar length and/or if the **Domain Similarity** is weak, then the software might be reluctant to predict the required number of **Indels** for a tidy answer (as illustrated). Instead, it is very possible the entire **Global** answer could be nonsense!

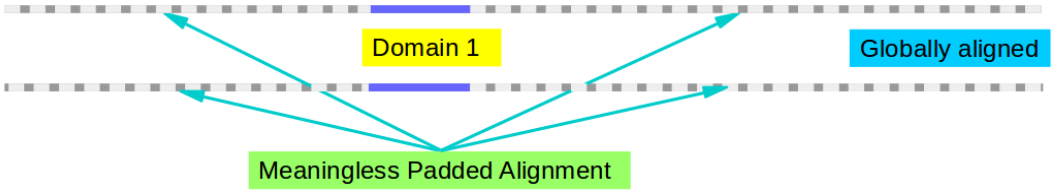
Example 2:

Imagine again that you wish to align two **Proteins**, each of which include two domains (a **Blue** one, and a **Red** one). The difference now is that the two **Domains** no longer occur in the same order!



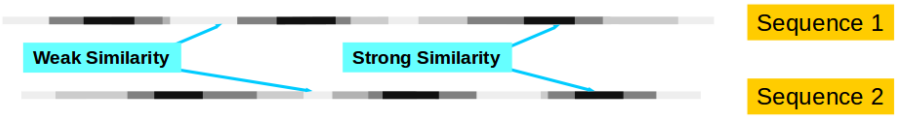
Again, and equally clearly, the correct choice would be to use a **Local** strategy. Assuming each **Domain** is sufficiently **Similar** (well conserved) in both **Proteins**, this should result in **two** separate **Alignments**, one for each **Domain**. The **Local Alignment** software has no requirements for regions to be in any particular order. The **Local Alignment** software will seek out **Similar** regions from any part of the **Sequences** it is examining and … **Align** them!

The **Global** approach is doomed to failure! It is (typically) written to find just one solution (the arithmetically "*best*" on according to rules either built into the programs or supplied by the user). It can either align the **Blue Domain**, or the **Red Domain** (or neither if things are not sufficiently obvious!). It cannot align both in a **Global Alignment**! It will choose the most **Similar Domain** (the **Blue** one in my illustration) and ignore the weaker (Red) one altogether. The entire **Alignment** either side of the "*correctly*" aligned **Domain**, is meaningless!



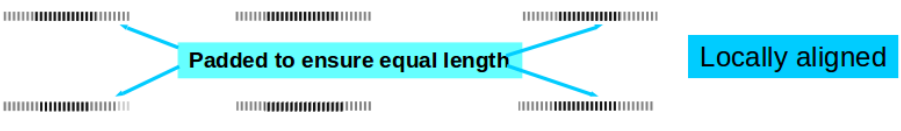
Example 3:

Imagine this time that you have **two Proteins**. The confident assumption is that they are **Homologous** over their entire lengths. However, **Sequence Similarity** is not likely to be consistent. In some regions, the **Protein Sequences** will look very **Similar** (to both the user and the software). In other regions, the **Similarity** will be far less obvious. In my illustration, the darker a region is coloured in one **Sequence**, the greater **Similarity** it has with its corresponding region in the other **Sequence**. I have arbitrarily decided to have **three** well matched regions between my **two Sequences**.



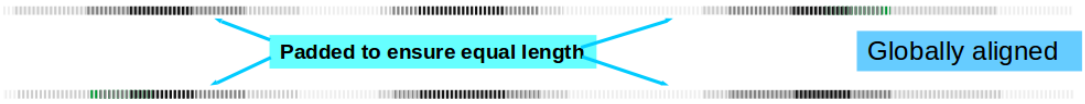
If a **Local Alignment** is computed between these two **Sequences**, it is most likely that the **three** most **Similar** regions will be detected, isolated and aligned reasonably satisfactorily. **Three** answers would generally be expected. However, although these regions would probably be aligned accurately enough, it cannot be satisfactory that so much **Sequence** has been eliminated because **Similarity** is insufficiently obvious. The assumption here is that ***ALL*** of both proteins are **Homologous**. It is therefore it is logical that ***ALL*** of both sequences be represented in a ***single*** **Alignment**.

Even the partial solutions of the **Local Alignments** must be between **Sequence** regions of equal length. This requires the software to insert padding characters, that is, to predict a number of Indel evolutionary events. **Indels** may be predicted anywhere in the **Local Alignments**, however, it should to be expected that the poorer the **Similarity**, the more willing the software would be to predict **Indels**.



Now consider the application of a **Global** approach to the **Aligning** of the two **Sequences**.

A single **Alignment** including the full length of both sequences will be generated. It is reasonable to expect that the strongly **Similarly Sequence** regions will be matched correctly. Padding (i.e. the assumption of **Indels**) will normally be required (particularly in the poorly matching regions) in order to stretch out the **Sequences** to allow **Alignment** of ***ALL*** corresponding regions.



The Global Alignment is clearly the one that best fits the assumptions of the problem.

The message of these thee examples is that it is very important to think carefully before choosing between **Global** and **Local Alignment** for you data.

Return To Index

Query: Which is the best software tool to compute **Pairwise Sequence Alignments**?

Briefly: For **Pairwise Alignment**, the major choice to be made is whether to compute a **Local Alignment** or a **Global Alignment**. This is discussed quite fully in the answer to another **FAQ**.

Most of the more popular **Local Alignment** programs use the same basic approach (**Algorithm**). Specifically, something very close to the **Smith-Waterman Algorithm**. The **Alignments** you compute will vary slightly with the implementation of the **Algorithm** you choose, but not (usually) significantly. There is room for poetry, even in the creation of computer programs. Most of the major alternatives produce fine **Alignments**.

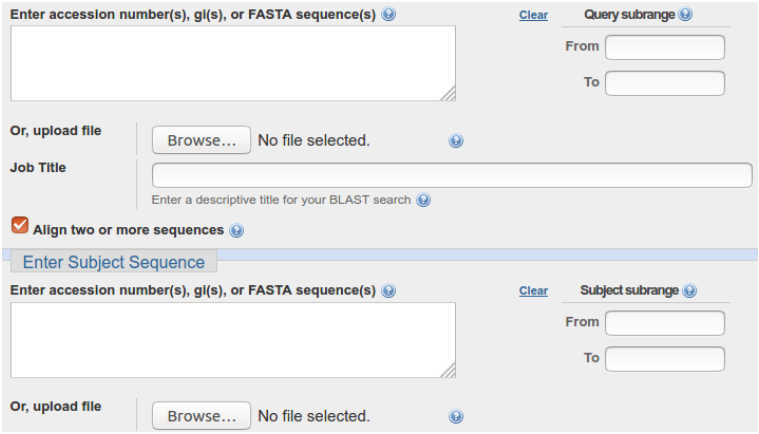
Most of the more popular **Global Alignment** programs are based upon the same **Algorithm**. Specifically, something very close to the **Needleman–Wunsch Algorithm**. The **Alignments** you compute will vary slightly with the implementation of the **Algorithm** you choose, but not (usually) significantly. Most of the major alternatives produce fine **Alignments**.

In full: **Wikipedia** (a wonderful resource!) provide an extensive list of software for **Pairwise Alignment**. Here, I will just mention a couple of options. Unsurprisingly, the ones I tend to rely on most.

The easiest way to compute a **Pairwise Alignment**, is to do so using one of the major **Bioinformatics Services** via a **Web Browser**.

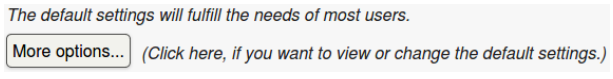
For example, at the **NCBI**, amongst the **Specialisations** of **Blast** (very loosely!), **Global Pairwise Alignment** (based upon the **Needleman-Wunsch Algorithm**) is offered. This option gives very good results.

Local Pairwise Alignment is also offered as apart of the **Blast** service at the **NCBI**. This does not employ the **Smith-Waterman Algorithm**. Instead a version of the standard **Blast Algorithm** is used, with **Parameters** set to compare **two Sequences** with each other rather than **one Sequence** with a **Sequence Database**. In theory, this should not really be as sensitive as using a **Smith-Waterman Algorithm** method. Nevertheless, the results are generally satisfactory in my experience. If you want to try this method, go first to the normal **Blast page**, then chose either the **Proteins** or **DNA** option. Next click on the "**Align two or more sequences**" button, and finally, follow your nose.



At the **EBI** (and many many other sites) a selection of programs from the **EMBOSS Package of Bioinformatics Software** are offered. The **Local Pairwise Alignment** option (an implementation of **Smith-Waterman**) is called **matcher**. The **Global Pairwise Alignment** option (an implementation of **Needleman-Wunsch**) is called **needle**. Both are quite easy to use and give good results.

A downside of running programs implemented on the **INTERNET**, is that all to often you will be denied full access to all the program's parameters. This is particularly the case for the **Pairwise Alignments** options at the **EBI**. The **EBI** declare:



Even if you insist, and click on "**More options**", you are not given access to all the options that the program allows! Well, OK, it is true that the default setting are usually the right settings, but I still resent being denied access to the programs full range of capabilities.

For some programs demanding a lot of compute time, you may also be restricted as to the volume of data you can process. Specifically, some **Multiple Sequence Alignment** (**MSA**) Programs limit the number of **Sequences** you may align!

The solution is to download the programs to your own machine. Read the Manual! And run the software locally. This is not as hard as it sounds. **Blast** is easy to **download and install** on a **LINUX** laptop, as are the programs of the **EMBOSS Package**. Why not try it sometime? Once the programs are running on your hardware, you have access to all the parameters and the data volume limits are determined by the capacities of your computer.

I conclude with a cynical note. Only my opinion, feel free to disagree.

Aligning pairs of sequences is a crude process. It relies on incomplete information (just the sequences, NOT the **DNA**/**RNA** or **Proteins** themselves). Hugely general assumptions are made concerning the way substitutions occur during **Evolution**, but surely the way substitutions happen varies widely according to circumstances (e.g. substitutions in a **Linker Region** must be far more likely than in the middle of a highly conserved **Protein Domain**), also very simple rules governing where **Indels** might occur are applied evenly over the length of the sequences being aligned. This cannot be right? The likelihood of an **Indel** in the middle of a **Helix** (for example) just has to be far less than in a region where there is no important secondary structure.

So, I wonder, how much effort should one take fine tuning the parameters of such a sloppy analysis? Of course, if a program is to be written to align ***ANY*** two sequences, there is little option but to implement such improbably general assumptions. The programs (I suggest), even when run as carefully as is possible, will only align **Sequences** that you could align manually (with a bit of knowledge about the relationships between amino acids). The programs will just produce the "*best*" **Alignments** a lot faster than any human can.

Return To Index

Query: Which is the best software tool to compute **Multiple Sequence Alignments (MSAs)**?

Briefly: The first **MSA** program commonly used is **Clustal**. It was born around **1988** and was just about the only real option for many years. It is still around in several forms, most still in common use and generating believable **MSAs**.

Over the years, a wide variety of alternatives to **Clustal** have emerged (starting with **T-Coffee** in 2000) ... most claiming to be improvements on **Clustal**. The list is intimidatingly long! I will mention some of the more popular examples here, but will resist giving too much detail until I have read and understood more about the various "new" approaches.

In full:

The most popular **Pairwise Alignments Algorithms** rely on a technique called **Dynamic Programming**. The **Dynamic Programming Pairwise Alignment Algorithms** (see previous Answer) can, in theory at least, also be used for **Multiple Sequence Alignments**. Sadly, however, to do so is not practical. The requirement for computer memory (in particular) increases exponentially as the number of **Sequences** to be aligned increases. Even in a world of cheap and copious computer resources, direct application of **Dynamic Programming** strategies to **Multiple Sequence Alignment** leads to impossible memory requirement to align even relatively modest numbers of **Sequences**.

Clustal, and many other programs following the lead of **Clustal**, use a compromise approach. In overview, they build **Multiple Sequence Alignments** by combining **Sequences** in a series of **Pairwise Alignment** steps. This is, theoretically, not as good as using the **Dynamic Programming Algorithms** directly, ***BUT*** it allows the **Multiple Alignment** of very large numbers of **Sequences** using only the resources expected to be available on a modern laptop.

So the original question was asking which was the best program to use. As usual, it all depends what you want to do, how many **Sequences** you wish to align and the length of those **Sequences**.

The **EBI** offer a range of **MSA** programs and advice on which one to use for what purpose. I cannot really improve on this advice (mostly concerning only the size of the **Alignment**?). Note however, when it is suggested that a particular program is suitable for "small alignments", that often means only that the program (e.g. **T-Coffee**) is slow and resource hungry, so if you use it for a large **MSA**, you will use an unacceptable amount of **EBI** computing power. **T-Coffee**, at the **EBI**, imposes a limit on the number of **Sequences** it will accept, for example. However, that limit is **500 Sequences**, which really seems quite generous … at first glance. Do not miss the further restriction implying a maximum total **Sequence** file size of **1MB**!? So … **500 Sequences** of mean length … **2 Residues**? I hope I have something wrong here, I will investigate further soon. I hope the **1MB** limit only applies when **Sequence Data** is supplied in an up loaded file rather than pasted into the box provided on the **Web Page**.

Some of the **EBI** choices claim to be particularly appropriate for specific purposes. For example, **webPRANK** is intended for people interested in **Phylogeny**. **T_Coffee** will combine alignments produced by other **MSA** methods.

Muscle (offered at the **EBI**) is an option that works in a similar, but more sophisticated way to **Clustal**. It should produce better **Alignments** more swiftly (generally) than **Clustal** and is quite "light" on computer resources.

The **NCBI** offer an **MSA** program called **COBALT**. This aligns **Protein Sequences** only. Interesting features of COBALT include the use of **Domain**/**Motif Databases** to detect regions of the **Sequences** being processed that really should align well. Also, **COBALT** uses models for the way evolution works computed from the **Sequences** under investigation, rather than models assumed to apply to the whole of **Nature**. Clearly, this has to be a good idea.

So, **COBALT** works on a more informative set of data than most of its competitors. I suggest that generating better **Alignments** by simply designing better computer **Algorithms** is limited by the usual paucity of the raw data. Only by providing ***MORE*** information, in particular ***MORE*** specific information, can real

improvements be achieved. That, it would appear, is what **COBALT** attempts to do.

At this time, my experimentations with **COBALT** are at an early stage, I hope to have more to say after I look more closely. I suggest it is worth a try however, given you have a suitable set of **Protein Sequences** to align.

As with all software, if you wish to get the best out of the programs, download them and run them on your on computer (after reading the manual, of course). This is possible for most (if not all) of the **MSA** programs mentioned above.

Return To Index

Query: Is it possible for one to create a **Personal Profile** at **NCBI**, and/or any other site providing similar services, enabling search results (and similar analytical computations) for future references?

If all you wish to do is to save your results to your laptop for further consideration, this is generally possible at most relevant sites without having a **Personal Profile**. You might lose some of the pretty ways to view your results offered by the web, but you will have the essential ingredients of your analysis stored locally.

It is certainly the case that **Personal Profiles** are available at the NCBI. The EBI also offer something of the sort (that I have never used). There is no reason to believe such **Profiles** are not also available at other large **Bioinformatics Service Sites** but I have no personal experience. I investigate and will expand upon this answer if I learn anything interesting.

Ensembl (a **Genome Database** site, not a general service site like the **NCBI** or **EBI**), also offers a **Personal Profile** option for more serious users of their facilities.

NCBI: To create a **Personal Profile** at the **NCBI**, go to the NCBI Home Page, click on the link "**Sign in to NCBI**" (top right hand corner of the page), click then on the link "**Register for an NCBI account**" (bottom of the page. And follow your nose.

Or … take the following Shortcut and then follow your nose.

EBI: I feel sure it is possible to create a useful **Personal Profile** at the **EBI**. However, I have never serious made and used such a **Profile**. It does not look particularly straight forward. Instructions are few (it is claimed "*More details to come soon here...*" in the place that claims to offer to explain "**How to use the AAP?**".

If you wish to explore, I suggest you start Here. I will expand these notes if/when I learn more.

Ensembl: To create a **Personal Profile** at **Ensembl**, go to the Ensembl Home Page, click on the link "**Login/Register**" (top right hand corner of the page), elect to **Register** and follow your nose.

Return To Index

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 instance:

- 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**!

Return To Index