# Short Intro to Sequence Alignment and Analysis for Biologists

Sebastian Schmeier

20th May 2014

## Short Intro to Sequence Alignment and Analysis for Biologists

In *bioinformatics*, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences [1].

Sequence alignment falls into three classes: pairwise, multiple, and database searches:

- Pairwise is easy to do (computationally). Often used to align large sequences (comparative genomics).
- Multiple is more complex, but is more widely used (e.g. phylogenetics, comparative genomics).
- Database searches – very widespread and useful. Usually in the form of BLAST. We will deal with this first!

### 1. Database sequence search

We will use the BLAST tool to find a sequence of interest, e.g. to identify what the sequence possibly does.

**BLAST** stands for **B**asic **L**ocal **A**lignment **S**equence **T**ool [2]. It is available online through NCBI (National Center for Biotechnology Information). The basic idea behind BLAST is that it searches your input sequence against a *database of DNA sequences* to find matches. Matches are returned with *e-values* that reflect the likelihood that the match is real (as opposed to a chance match). It works in association with GenBank – the NCBI database for DNA sequences.

Attention! It is very important if you are using BLAST in your research, that you know how it works, what its limitations are, and that you use the correct

BLAST options and interpret the meaning of the results correctly. If usage of BLAST is presented in your thesis, this knowledge will be expected, and you will also be expected to interpret your results correctly.

### 1.1. Get a sequence

Go to http://www.ncbi.nlm.nih.gov/ and search for a gene ( e.g. BRCA2). Now we see plenty of different databases where our search term was found. All of these provide different information about our search term.

We want the sequence of the gene. Thus, we need to go to the nucleotide database (e.g. BRCA2). Choose for example the human sequence of your gene in FASTA format (more information on data-formats can be found here). Copy a part of the sequence (e.g. BRCA2).

### 1.2. Find the right BLAST tool

Several different flavours of BLAST are available.

- nucleotide blast - Search a nucleotide database using a nucleotide query
- protein blast - Search protein database using a protein query
- 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
- Align - Align two (or more) sequences using BLAST (bl2seq)

### 1.3. Choosing the right options/parameters

Some of these tools mentioned above have options/parameters for the type of BLAST you can perform. It is important to know what they are and what situations to use the different options. Plus there are various other tools – have a look around. There are tutorials on the website.

The parameters section of th BLAST tool is very important. Here we talk about the most important ones:

- **Max target sequences** - Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).
- **Short queries** - Automatically adjust word size and other parameters to improve results for short queries.
- **Expect threshold** - Expected number of chance matches in a random model. The default value (10) means that 10 such matches are expected to be found merely by chance.

Note! The Expect value (*e-value*) is critical – simply put, it is the number of such matches we expect by random chance. However, you should look into exactly what the *e-value* is. The lower the e-value, the greater the chance the match is real. More information is available here and a video tutorial is available here.

- **The word size** - The length (number of nucleotides) of the seed that initiates an alignment (More information here).
- **Match/Mismatch Scores** - Reward and penalty for matching and mismatching bases, which determine whether to align residues or not – this is the basic engine of the search (More information here).
- **Gap Costs** - Cost to create and extend a gap in an alignment (More information here).

Finally we have various filters for filtering out low-complexity sequences. You can just filter these out from just the initial alignment seeding, but include them the full alignment search, as well as customize what regions get masked.

### 1.4. Interpreting the results

> ***Example 1*** 1. Use the nucleotide sequence of your gene and play with BLAST. 2. Copy a fairly short chunk of around 20-30 bp and "BLAST" it. What is the result? 3. Copy a bigger chunk, maybe 100-200bp of the gene and "BLAST" it. What is the result? 4. Delete some stretches of the bigger chunk and "BLAST" again. What is the result? 5. Insert some random DNA into the sequence and "BLAST" again. What is the result?

When looking at the results from BLAST, you should know what the various results mean. Although the *e-value* is very important, there are other result options that are also important, such as the *% coverage* of the sequence. Familiarize yourselves with these. You can check the sequences to see what they are, by clicking on them, and the alignments are also presented. Remember that BLAST is a local alignment tool. It will find as many matches between the query and the subject as fulfill the search criteria. This means that it is important to look at all the results, as often there are many highly similar results, so to say that the top one is the closest match isn't always true, and its also important to distinguish between what is a real match versus what is a spurious match.

### 1.5. Where to go from here?

There is a lot of stuff in the NCBI website, so it is worth exploring. Another very important sequence resource is the database of individual genome projects. These can have the whole genome and/or the individual sequence reads that

were used to construct the genome. Very often they have an online BLAST engine that you can search that database with. There are also public databases that people have downloaded genome sequence data onto that can be searched.

## 2. Alignments

### 2.1. Pairwise alignments

Align aligns two sequences. In other words, this is a pairwise alignment tool. This is one good way of aligning two sequences if they have a reasonable match over their entire lengths. It has advantages and disadvantages. It can align in both orientations, but it often struggles if the starts of the sequence are not similar, or if the sequences don't match over their entire lengths. It will also often produce very many short matches of bits of the two sequences that are similar.

> ***Example 2*** 1. Copy the identifier for BRCA2 from human (U43746.1) and mouse (U65594.1). 2. Paste them into the align program. 3. Run it!

Pipmaker can be used to very quickly find matches between two sequences is. Here you can align sequences of very different lengths, and it also aligns in both orientations, and is very fast. It produces a dot-plot type output of the level of similarity. The disadvantage is that it doesn't give you the actual sequence alignment, but it shows you what parts you can align. You can also see repetitive elements in the dotplots.

Stretcher is a useful tool for aligning long sequences that are quite similar. Its usually OK if the sequences are not the same lengths, but not always. You also have to make sure the sequences are in the same orientation. It produces a nice aligned sequence output. It is part of the Mobyle website that uses the EMBOSS suite of programs (run by the Pasteur). This has quite a number of other, useful programs on there, such as SQUIZZ for sequence/alignment conversions (see below).

Vista is another alignment tool that produces a visual display of the level of similarity across the alignment, allowing you to see conserved and diverged regions easily. The aligner is surprisingly accurate and fast, and you can include multiple sequences, although they all get pairwise aligned, not multiply aligned as we will cover next .

### 2.2. Multiple alignments

CLUSTALW2 is the most commonly-used tool [3]. You can find CLUSTALW all over the web. You must choose whether you do fast or slow (depending on

how accurate you want the results to be), and also DNA or protein. There are also a few parameters that can be adjusted. CLUSTALW2 often has trouble if the starts of the sequences are not very similar, and can also have trouble with repeats and large indels. Changing some of the parameters can help (often the gap settings).

CLUSTAL OMEGA is a newer version of CLUSTALW (and will eventually replace it), which is not yet as commonly used, but seems promising (Paper reference here [4]).

> ***Example 3*** 1. Use this file and paste it into the interface (it contains the BRCA protein sequence from 7 species). 2. Run it with default parameters. 3. Look at the output. 4. Highlight the alignments with colors. 5. Send the results to the EBI Phylogeny. 6. Look at the produced phylogentic tree. 7. What does it tell us? 8. Do the lengths of the branches have a meaning?

There are numerous other multiple alignment tools available, and multiple alignments is a field of research in its own right. Many are available as web-tools. Try out some and find one you like. But the most important thing is that multiple alignments must be checked by eye for accuracy. Therefore, you need to import the completed alignment into an alignment editor, and then go through to check that nothing strange has happened.

## 3. Additional tools for sequence analysis

There are a huge number of sequence analysis tools online, that do useful things like make reverse complements of your DNA, translate it, look for genes, look for repeat elements, make restriction maps, design primers, etc. Try a few programs, and find something that works.

There are many different primer design tools out there, and I imagine most do a good job. NCBI has one on their website called Primer-BLAST. Primer3Plus is one, which gives a large number of options, although is a little hard to use at first.

For analysis of sequence data, small scale analyses are often done with software that individual lab groups have. If not, BioEdit is a free program for Windows that is a little painful but works OK. Geneious is a (comparatively) cheap DNA analysis package that is useful for a wide range of DNA analyses, such as alignments, phylogenetic tree-building and cloning design that is powerful, easy to use, and also has the ability to handle next-generation sequencing data (below). However Geneious is not free software and as such costs money.

If you want to move into analysis of genome scale data, then you will probably need to know some *command-line* (below). For assembly of Sanger sequences

the Phred/Phrap/Consed package is a very powerful free assembly tool. I will deal with next-generation sequence analysis below.

Attention! Often sequence/alignment formats are an issue when working with different programs. Sometimes it is necessary to convert between different sequence/alignment formats, and usually there are sites online that can do this (e.g. GALAXY see below). Also be aware that Macs/*Unix*/Windows use different paragraph/space/newline notations, so sometimes these can be incompatible between each other and need to be converted. Many cases where you get errors for no apparent reason will be different line return formats (in my experience!)

There are also increasing numbers of databases for a huge variety of biological information. One good source for information about these is the journal, Nucleic Acids Research. Periodically (once a year) they put out a special issue in which each paper is a description of an online database of some sort. It is a large issue! The DNA and protein databases are obviously the main ones, and you are probably familiar with these. However, there are also specialized database for model organisms (e.g. Yeast SGD database), types of genes, and genomic comparisons.

Then there are more specialized ones, such as TFSEARCH that searches the JASPAR database public version of the TRANSFAC database for known binding sites in your DNA (i.e. a database for finding cis-elements). If you want to discover new binding sites MEME might help.

You also might want to find gene ontology terms or pathways that are associated to your gene-set (DAVID) or your genomic regions (GREAT).

You might want to identify protein-domains in our DNA or protein sequences. Common databases and tools for such a task are CD-SEARCH, PROSITE or the PFAM and Superfamily database.

Attention! Many of the specialized tools that are available with a web-front-end impose limitations, e.g. regarding sequence length, number of motifs searched, etc. Generally speaking, the downloaded version of these tools are the ones to circumvent such limitations. But then again, we need to use the *command-line* (see below).

> ***Example 4*** 1. Copy the proteins from this file. 2. Paste the list to DAVID. 3. Look at the functional annotation of the genes/proteins. 4. What are those genes? 5. Look for Pathway enrichment (KEGG) of the gene set and select one pathway. 6. Explore the KEGG database. 7. Convert the list using DAVID to OFFICIAL_GENE_SYMBOLS. 8. Take the same gene and paste it into the STRING search field. 9. Look at the association network. What is the difference to KEGG? 10. Take the first gene and search it in the UCSC Genome Browser.

> ***Example 5*** 1. From the file, take the human one and the chicken

protein one and search for protein-domains with PFAM. 2. What do you expect to find? 3. What do you find??

## 4. Genome browsers

Now we are in the age of sequencing, people often want to look at and represent larger chunks of DNA, and also to integrate multiple pieces of information into one view of your DNA of interest. If this is something you are starting to do, then it might be worthwhile looking at using a genome browser. Genome browsers are software tools that allow you to view, in a scalable way, large pieces of DNA and to simultaneously view many different features that are present in that sequence.

Usually you start by either putting your DNA sequence of interest into the program , and then load various features onto this. For example, you might want to load on where all the genes are, and then overlay this with where histone variants map (from ChIP mapping) and which genes from a related species are present. But you can basically put anything on it (e.g. I put primers in). In order to load features, you have to have a special type of file, called a GFF3 file. This is actually not so special: it is just a tab-delimited text file (you can make these with Excel) that contains the various information you want to load on (gene name, position, orientation, etc). There are a number of freely-available genome browsers you can use, such as the Broad's IGV.

The UCSC Genome Browser is also of this type, but here lots of data is already integrated and you can upload specail annotation for the sequences in form of bed-files. This browser is actually a piece of software that one can use to create one's own genome browser. Other projects use it as well, e.g. the UCSC Microbial Genome browser, etc.

## 5. Next-generation sequence analysis

The costs of sequencing are dropping at a phenomenal rate, due to the development of so-called next-generation sequencing technologies. Many of these technologies produce a very large number of sequences of short length that need to be assembled to get the original DNA sequence, or mapped back to the genome (e.g. in the case of mRNA sequencing). The sheer numbers of sequences generated by these technologies means that in many cases traditional tools for sequence analysis don't work. More-and-more biologists have to learn to deal with high volumes of sequences, and often this requires using the *command-line* and programming. If you are interested in continuing with biological research, then gaining experience in dealing with this sort of data will become an advantage for you. The biggest problem currently in biology is the lack of people who can analyse large data sets, and can do so in a quantitative manner. In most cases this requires the use of *command-line* driven software. Luckily, the majority are pretty simple to use once you understand the basics. The great thing about

*command-line* software is that almost all of it is free! It takes a bit more time to get comfortable using it, and the support depends on the software, but at least you don't have to harass your supervisor for money!

This is an area that is developing incredibly rapidly, with a bewildering range of tools that all seem to do the same or similar things. Probably the best thing to do is to talk to someone who has done these kinds of analyses before to help you get started. There are many papers being published on a monthly basis providing new and varied ways to a whole variety of analyses, but you don't want to spend your whole time testing out new versions unless you have to. If in doubt, let published studies doing the similar things to you guide you (if possible).

There is also intensive development of downstream analyses of data generated by these technologies. For transcriptome data, you may want to know what set of genes has changed significantly between two different samples (e.g. a mutant and wild-type), and what classes these genes fall into. Alternatively, you may want a list of where all the polymorphisms from a genome sequence fall; what genes they fall into, etc. These kinds of analyses are now often done using *command-line* driven software. An important feature to note is that next generation sequence datasets are often publically available (usually a criterion for publishing), therefore you can analyse these datasets yourself, which means that you don't even have to do the experiments and then pay for the sequencing yourself. This is important to keep in mind, as often the people who originally produced the dataset did so for a completely different reason, but the information that is important for you is in there, and just needs to be analysed to get it out.

Another key aspect in the analysis of biological data is statistical analysis. For basic statistics, there are commercial packages available that will do statistics on data you enter. However, you should be familiar with the statistical analyses that are being performed: how they work and what they mean. As more-and-more data are produced, there is a need to do more sophisticated analyses, and to be able to plot these data. For this, more powerful statistical packages may be required. Many of these also require use of the *command-line*. A particularly good (and free!) one is called R. Once again it takes a bit to learn, but will definitely be beneficial in the long-term!

Finally, the best way to learn computational biology is to talk to people who have experience in the area and who can point you in the right direction. Very often, the hardest part is knowing where to start. Therefore ask around: there are a number of people in the Institute who have experience with bioinformatics and computational biology who will be able to help.

## 6. Getting data

There is lots of data already out there, and depending of what kind of data you are looking for there are different databases and tools available to get this data,

e.g.

- UCSC Table Browser
- BioMart
- Gene Expression Omnibus
- Short read archive
- dbSNP

## 7. A word on the command-line

Many of the analysis tools that are available on the web, plus a whole lot that are not, are also available as downloadable programs to run locally on your computer. One of the most common is the set of BLAST scripts – you can download them all. Many of these require (e.g. the BLAST series) require *command-line* usage, although increasing numbers are doing it through graphical user interfaces. The advantage with running these programs off your own computer is they are often faster, and you can often put in bigger/more sequences that on the web (they often limit this so people don't take over the whole server with a huge job). I highly recommend trying out *command-line* – you can do a lot with just a little knowledge, and it opens up a whole lot of stuff you couldn't otherwise use (in particular free stuff!).

This often requires use of a *Unix*/*Linux* computers (luckily, Macs are basically *Unix*, but Windows is not). An easy way to try out a *Linux* working environment is through an *Linux* LiveCD. After you download the *Linux* system and burn it onto a DVD, you insert it into your computer drive and restart the computer. It will boot your computer into a *Linux* system that is running from the CD, without any installation. The next time you restart your computer without the CD inserted, your old system starts as normal without any change.

There are good little tutorials on the web for the *command-line* stuff (a short one can be found here), but it is also a good idea to ask someone who already knows some to teach: a quick run through the key points will be enough to get you started.

## 8. Beyond

If you really get into sequence analysis, then you will need to know some programming! Amazingly, this is not as scary as it sounds. There are a variety of languages you can use, but Python is an easy one to start with. On of the main advantages of Python is the presence of a library written for bioinformatics tasks BioPython. This is collection of what are called "modules" that people have already written to do all sorts of sequence analyses and manipulations. This means you don't have to write the program yourself, but you just have to figure

out how to use the provided module (which admittedly can take some time as well). These modules are not really programs in their own right: you have to write a program to use the modules to do things that you want. If you learn how to use Python and write scripts, it can be incredibly powerful! Another language that is widely used is Perl. For statistical analyses, the language R is very widely used and there are also a large number of modules that can be utilized for biological analyses, known as Bioconductor.

Hint! The good news for people wary of *command-line* is a lot of these tools are slowly being incorporated into web-based packages, such as GALAXY or GenePattern, that have a graphical user interface. These are useful, however, understanding the underlying tools is important to make sure that your results are as expected.

## 9. Links

- BLAST - http://blast.ncbi.nlm.nih.gov/
- NCBI - http://www.ncbi.nlm.nih.gov/
- NCBI Nucleotide - http://www.ncbi.nlm.nih.gov/nuccore
- Stretcher - http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=stretcher
- Vista - http://genome.lbl.gov/vista/index.shtml
- CLUSTALW2 - https://www.ebi.ac.uk/Tools/msa/clustalw2/
- CLUSTAL OMEGA - https://www.ebi.ac.uk/Tools/msa/clustalo/
- Primer-BLAST - http://www.ncbi.nlm.nih.gov/tools/primer-blast/
- Phred/Phrap/Consed - http://www.phrap.org/phredphrapconsed.html
- TFSEARCH - http://www.cbrc.jp/research/db/TFSEARCH.html
- JASPAR database - http://jaspar.genereg.net/
- TRANSFAC database - http://www.gene-regulation.com/pub/databases.html
- MEME - http://meme.nbcr.net/meme/
- DAVID - http://david.abcc.ncifcrf.gov/
- GREAT - http://bejerano.stanford.edu/great/public/html/
- STRING - http://string-db.org/
- KEGG - http://www.genome.jp/kegg/
- CD-SEARCH - http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi?
- PROSITE - http://prosite.expasy.org/
- PFAM - http://pfam.xfam.org/
- Superfamily - http://supfam.org/SUPERFAMILY/hmm.html
- UCSC Genome Browser - http://genome.ucsc.edu/
- IGV - https://www.broadinstitute.org/igv/home
- UCSC Microbial Genome Browser - http://microbes.ucsc.edu/
- UCSC Table Browser - http://genome.ucsc.edu/cgi-bin/hgTables?

command=start

- BioMart - http://www.biomart.org/biomart/martview/
- Gene Expression Omnibus - http://www.ncbi.nlm.nih.gov/geo/
- Short read archive - http://www.ncbi.nlm.nih.gov/sra
- dbSNP - http://www.ncbi.nlm.nih.gov/snp/
- Python - https://www.python.org/
- BioPython - http://www.biopython.org/
- Perl - http://www.perl.org/
- R - http://www.r-project.org/
- Bioconductor - http://www.bioconductor.org/
- GALAXY - https://usegalaxy.org/
- GenePattern - https://www.broadinstitute.org/cancer/software/genepattern/
- Ubuntu LiveCD - https://help.ubuntu.com/community/LiveCD#How-To_LiveCD_Ubuntu

## 10. References


1. Mount DM. Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Laboratory Press (2004): Cold Spring Harbor, NY. ISBN 0-87969-608-7.
2. Altschul SF et al. Gaped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res (1997). 25:3389-3402.
3. Larkin MA et al. Clustal W and Clustal X version 2.0. Bioinformatics (2007), 23, 2947-2948.
4. Sievers F et al. Fast, scalable generation of highquality protein multiple sequence alignments using Clustal Omega.Mol Syst Biol. (2011) 7: 539. DOI: 10.1038/msb.2011.75


***FILE:*** ***bioinf*****seqintro.md - Download as PDF - Sebastian Schmeier - Last update: 2014/05/2014_**