Exercise: Multiple Sequence Alignment and Analysis

Multiple Sequence Alignment (MSA) is a very useful tool to identify conserved patterns, predict structure, pinpoint protein-protein interaction sites, as well as to reveal functional sites in proteins (or in genomic sequences in general). Extracting useful information from an MSA is therefore an important task in bioinformatics studies and also in biology research.

In this exercise, you will play with a few ways to obtain information from an MSA:

1. Start from a single protein identifier and find the sequences using BLAST and PSI-BLAST
2. Build the MSA through an alignment tool and evaluate the quality of the alignment.
3. Identify different conservation patterns between two sub-families derived from MSA to identify protein functinal sites.
4. Optionally, and time allowing, you may observe the dependance of the results on the alignment tools and options used.

At the end of this manual, you may find a list of tools, websites and web-servers used, and in addition a few other useful ones.

Back ground knowledge:

Multiple sequence alignments are often used to reveal functionally important residues within a protein family. Many protein families contain sub-families with functional specialization, such as binding different ligands or being involved in different protein– protein interactions; typically, these proteins will be paralogs. A small, number of amino acids generally determine functional specificity. The identification of these residues can aid the understanding of protein function and help finding targets for experimental analysis. For more back ground, please refer to:

- Pirovano, Feenstra & Heringa. "Sequence Comparison by Sequence Harmony Identifies Subtype Specific Sites", *Nucleic Acids Res.*, 34, 6540-6548 (2006), which gives a concise introduction into the topic and on some of the available tools to detect this specialization from alignments; and
- Brandt, Feenstra, & Heringa "Multi-Harmony: detecting functional specificity from sequence alignment", *Nucl. Acids Res.*, doi:10.1093/nar/gkq415 (2010), which gives an extension of the method, and an updated benchmarking and comparison with several other specificity determining sites.

In this assignment, we will try to identify the specific functional sites in the MIP ('Major Intrinsic Protein' – Integral membrane transporters) family. MIP family members are mainly involved in facilitating the transport of both water and small neutral solutes through the cellular membrane in all domains of life. There are about six MIP subfamilies, the two major being the aquaporins (AQPs) and the glycerol-uptake facilitators (GLPs).

We first fetch the sequences for this MIP protein family using BLAST, then align them and evaluate the alignment and its quality. To pinpoint the functional specific residues, Multi-Harmony will be used to find subfamily specific sites between two sub-families from the input alignment.

Part 1: BLAST and PSI-Blast in the MIP family

Perform a BLAST search using NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) on the Swissprot/Uniprot db using the following query sequence ID: P0AER2.1

While BLAST runs, it already starts showing you the domain lay-out of your query sequence; have a look at this. When BLAST is finished, have a look at the names of the domain hits and families BLAST returns; are they all from the AQPs and GLPs sub-family? If not, what do you think is happening here?

From the Blast output, select all AQPs 1(Aquaporin-1), AQPs 2(Aquaporin-2) and GLPs (glycerol facilitators) proteins and download them as FASTA.

Part 2: Alignment and quality scoring

1 Multiply Aligning:

Generate four alignments of all sequences using:

- Clustal Omega: www.ebi.ac.uk/Tools/msa/clustalo/
  Select Pearson/FASTA as Output format. Use default settings otherwise.
- PRALINE: www.ibi.vu.nl/programs/pralinewww/
  Select Standard progressive strategy – *be sure to turn off SS and TM prediction!* (be aware that PRALINE is a sophisticated but rather slow tool – SS and TM prediction makes it much slower still.)
- Optionally, you may experiment with alignment settings (e.g., using PRALINE's global or PSI-BLAST pre-profile processing options, and/or SS prediction), or other alignment tools (e.g., Muscle, ProbCons; there are many others).

Save the web output pages – these contain useful statistics on the alignment produced.

2 Alignment Visualisation:

Study your alignments using JalView. It is convenient to turn on a color-coded view in the Colour menu (Zappo is recommended); this will colour amino acids according to their properties (e.g., red for positive charge, blue for negative charge, green for hydrophobic). This allows you to quickly review conservation properties in different areas of your alignment.

In your alignment, you may have seen that some MIP subfamilies (AQPs) have larger differences compared to others (GLPs). Please see below at Part 3 for why these particular subfamilies may be different (Pirovano *et al.,* 2006 also provides some more detail and references about the MIP family and its sub groups).

Also, you may notice some sequences are (clearly) outliers. You could decide to remove them now, or you can re-evaluate this decision later when you have calculated a phylogenetic tree.

3 Optional – Assessing Alignment Quality:

Measuring alignment quality is a tricky problem. Often-used methods are the SP score (also known as 'Alignment score', using a substitution matrix), the total alignment length (more compact alignments tend to be better defined and are more likely to correspond to evolutionary patterns), and visual inspection. Alignment/SP score is generally reported by the alignment tool (check your web

or command line output), but scores are usually not compatible between alignment tools (often not even for the same tool with different settings).

Surprisingly, none of the available tools can easily calculate an overall alignment score. With the following bit of code, we can let Jalview do it for us (thanks to Jim Proctor). It will calculate the score for each alignment currently open in Jalview (sadly, it does not seem to work in all versions of JalView):

```
// Open the Tools->Groovy Console
// paste the following lines in and
// press 'CTRL+R' or 'CMD+R' to execute
// output is sum of quality scores for each alignment

def alf = Jalview.getAlignframes();
print "Title\tQuality sum\n"
def qual;
for (ala in alf)
{
    // ala is an jalview.gui.AlignFrame object
    print ala.getTitle()+"\t";
    // get the conservation object and sum the quality scores
    def alcons = ala.viewport.getAlignmentQualityAnnot();
    qual=0;
    for (q in alcons.annotations)
    {
        if (q!=null) { qual+=q.value; }
    }
    print "\t"+qual+"\n";
}
// end of groovy script
```

(P.S., for some help with groovy - see here: www.jalview.org/help/html/features/groovy.html)

An alternative way to compare two alignments (it only works pairwise), is VerAlign www.ibi.vu.nl/programs/veralignwww/. However, it requires identifiers between the alignments to be identical, so you will probably need to fix them first using a text editor (if you know how to use a regexp search/replace, that will save you quite some time). PRALINE, particularly, mangles the sequence identifiers badly.

Part 3: Phylogeny and Specificity Scoring

1 Phylogeny vs. Functional annotations:

As explained in the introduction, functionally important sites in a protein sequence may be found by comparing homologous groups of proteins with (slightly) different functions (typically, you expect these to be paralogs). To find these so-called specificity sites in a protein family, one needs to define the subgroups of interest.

This can be done in two different ways:

1. using (functional) annotations.
2. using phylogeny.

For phylogeny, several methods are available. Here, you will use two simple methods, Neighbour Joining (NJ) and average distance (UPGMA) which can be run from Jalview.

Calculate NJ and UPGMA trees for the alignments. Compare the trees you get. Pay attention to how sequences of the 3 different MIP subfamilies are distributed along the phylogenetic tree. Are all members of each subfamily grouped together in a single clade? Are subfamily clades side-by-side in

the tree, or are subfamilies in a sub-clade of another subfamily? Note, that in Jalview you can set a cut-off in the tree window (by simply clicking), and observe the corresponding grouping of sequences highlighted in the alignment window. You can also select a tree node, and get the corresponding sequences selected in the alignment window. Conversely, you can select a set of sequences in the alignment, and see them highlighted in the tree as well.

Coming back to the alignment quality, quite often (certainly when starting from a set of Blast hits), one finds a few sequences in the alignment are clearly outliers. Usually this is reflected in the phylogeny as well, as a sequence that end up as a separate clade separated from the rest. As a practical solution, to streamline further analysis, such sequences may be removed. This can be done in Jalview (select the sequence name, and press Delete). One may then clean up the alignment by removing gapped columns (Edit → Remove Empty Columns), but, strictly, one should re-align the remaining sequences.

2 Calculating Specificity:

Use Multi-Harmony to find subfamily specific sites from an input alignment. Use groupings based on the functional annotation (in the sequence labels) to make groups. You need to compare GLPs with AQPs , in terms of sequence annotations this means AQPs 1&2 versus GLPs. Pirovano *et al.* (2006) gives some more background on this protein family and its subfamilies.

You can run Multi-Harmony from the website. Specify '1FX8' (sequence P0AER2.1) as 'Reference Structure (optional): PDB Id:' on the Multi-Harmony webserver, so you can easily map your results to the known functional sites.

There are several ways to indicate groups in the Multi-Harmony webserver.

Easiest one, it is possible to run the Multi-Harmony analysis directly from Jalview (Web Service → Analysis → Multi Harmony). Advantage is that you will also visually see the scores displayed below your alignment. All sequences in your input must be part of one group, and you need at least two groups. You can define two (or more) groups in Jalview by defining a cut in the tree. (Manually selecting sequences into groups is also possible, though that is much more involved; see the Jalview manual if you need this.) Optionally, you can specify a reference sequence or structure for analysis.

Outside of Jalview, the most straightforward way to define groups is by specifying the number of sequences in each group. That assumes groups are consecutive in your alignment. You can do that manually (using a text editor), or for example using the option 'Sort Alignment by Tree' in the Jalview tree view window. If you prepend a group name to your sequence labels, you can also sort them by name in Jalview (Calculate → Sort → By ID).

The thirdmethod is more generic and also allows sequences in your alignment not to belong to any of the groups you want to analyse. This requires sequences to be labelled in its sequence name; this should look something like in the FASTA file:

```
>GROUP:AQP|P0CD91.1
----PVLVPRHSEYNPQLSLLAKFRSASLHSEPLMPHNATYPDSFQQSLC-PAPPSSPGH---VFPQSPCPT-----
SYPHSPGSPSESDSL…
>GROUP:GLP|P23900.2
----LEKAITTQNCNTKCVTIPSTCSEIWGLSTANTVDQWDTTGLYSFSEQTRSLD
GRLQVSHRKGLPHVIYCRLWRWPDLHSHHELKAIENCEYAFNLKKDEVCVNPYHYQRVET
PVLPPVLVPRHTEILTELPPLDDYTH-------SIPENTNFPAGIEPQSN--YIPETPPP…
```

In this example, all sequences labelled '`GROUP:AQP`' will constitute one group; and similarly for each unique label in between '`GROUP:`' and '`|`'. Doing this manually tends to be tedious, but it is very suitable for scripting.

3 Evaluating Specificity Detection:

When you have run Sequence Harmony, sort your output table on increasing Sequence Harmony score. Alternatively, on the output page you can set a cutoff for the SH score, and sort by alignment position. For comparison, the specific functional sites can be seen in the below, as defined in the paper based on proximity to the glycerol ligand in the crystal structure (numbering according to PDB 1FX8 and sequence P0AER2.1) :

L	W	V	I	A	H	L	V	T	Y	P	N	P	L	I	I	G	P	L	G	F	A	M
21	48	52	56	65	66	67	71	137	138	139	140	141	159	163	187	195	196	197	199	200	201	202

You may use this to assess the quality of your predictions based on the SH scores in terms of True Positives (TP) and False Positives, e.g. by creating a ROC plot (sort on ascending SH score), which you can compare to the one Pirovano *et al.* (2006), Fig. 4. It also makes sense to calculate the area under the curve in the ROC plot (AUC-ROC). Alternatively, you may simply choose a single cutoff for the SH score and calculate TP and FP rate for the different alignments.

Repeat this SH analysis for another alignment. You may find your results (SH scores, selected positions, and benchmark statistics) depend on the alignment used, and that some positions might be more sensitive to these changes.

Tools needed

References for the tools mentioned above. You are not strictly limited to these tools, but please be specific in your report if you chose differently and provide a justification for your choice.

- Blast (blast.ncbi.nlm.nih.gov/Blast.cgi)
- GPCR-DB (www.gpcr.org/7tm)
- Multi-Harmony webserver www.ibi.vu.nl/programs/shmrwww/
- PRALINE (www.ibi.vu.nl/programs/pralinewww): pre-processing, iteration, secondary and transmembrane structure
- CLUSTAL Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/): Several integrated heuristic strategies
- MUSCLE (www.ebi.ac.uk/Tools/muscle): Iterative tree optimisation
  - if installed, can also be used from the command line:
    `muscle -in <input> -out <output>`
- optionally: T-COFFEE (www.ebi.ac.uk/Tools/t-coffee): Search matrix extension strategy
- JalView (www.jalview.org): alignment viewing, editing and analysis
- Weblogo (weblogo.threeplusone.com/create.cgi): visualisation of sequence conservation patterns
- Phylogeny (www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=phyml): Several strategies for phylogenetic tree reconstruction, including Maximum Likelihood
- MEGA (www.megasoftware.net): A nice and versatile tool for alignment analysis, including phylogeny (including maximum likelihood), but also alignment editing. It is MS windows-based but it runs fine on Linux using Wine (which unfortunately currently doesn't run fine on our Linux computers here).