Exercise 1 – Functional Association Networks Online

Objectives:

- to get to know the STRING database, and its various features / resources.
- to explore a protein group/pathway that may be of particular interest for you.

Please note: depending on the number of participants in the course, the performance of the website may be slower than usual. Should that be the case, please be patient and do not click the reload button too frequently.

1) Follow a well-characterized example protein through interaction space

- open your browser, and point it to the STRING website: https://string-db.org/
- after clicking 'Search' on the STRING welcome page, click on example #1 in the category 'Protein by name' on the input page (this is the 'trpA' protein in Escherichia coli), and click 'Search'.
- this will bring you to a network-overview of the trpA protein together with some of its functional partners. The various lines indicate 'functional links' between two proteins (importantly, these do not necessarily signify direct binding connections; they could also just mean sharing a common pathway, proteins regulating each other, or having a connection in genetic terms). The various colors of the lines indicate different types of evidence. We will look at each of these in turn, in a minute.
- But first, let's spend some more time with the network itself … click on one of the proteins. You should get a pop-up window with information about the protein: its annotation, its domain structure, and in case there is a 3D structure known this will be shown as well.
- within the pop-up window of 'trpA' (the red protein), proceed to the link labeled 'homologs among STRING organisms'. It should show a list of homologs in other organisms. Locate the homolog in 'Citrobacter koseri', and click on its score – this brings up the Smith-Waterman alignment of the two proteins (the one from E.coli *vs.* the one from C.koseri). Note that STRING has all (or almost all) of those possible alignments already pre-computed, literally billions of alignments that are all accessible with a simple click.
- now, go back to the network view (simply close the popup window and click "back" in the current browser tab).
- now, click on the '+ / More'-button below the network … this will grow the network. Again, all links between proteins are pre-computed, so you should not have to wait long. Use 'Plus' a few times, then use 'Minus' again to bring the network back to its original size (i.e.: one query protein, and ten partners).
- now, below the network, there is a Tab labeled "Viewers", select this now to show a selection of "evidence viewers" … these allow to inspect some of the evidence supporting your current network. We'll look at the various viewers in turn.
- Neighborhood view: this view should show the genomic arrangement of the genes encoding the proteins in our network. Obviously, the genomes are in flux – changes have happened throughout evolutionary history. Can you infer some rearrangement events that brought about the current situation? Despite this

ongoing ‘genome shuffling’, the cell apparently chooses to keep this operon largely intact – this is evidence for a functional relation between its genes, and STRING has used it to predict an interaction between the proteins. Which pair of genes apparently has the largest selection pressures to always stay together in a operon? Does this make sense from what these genes encode?

- Fusion view: now switch over to the fusion viewer. It will show instances where some of the genes in our set have ‘fused’, i.e. merged to encode a single open reading frame. Usually, only functionally coupled genes tend to do this, again providing a method to discover functional connections from simple genome information. In our case, which pair of genes has apparently fused at least twice, in two different orientations?

- Cooccurrence viewer: This is a very powerful view, and we will spend a little bit of time there. In the Cooccurrence viewer, use the little ‘+’ button next to ‘Terrabacteria group’, then ‘Firmicutes’, then ‘Clostridia’, ‘Clostridiales’, ‘Clostridiaceae’, and finally ‘Clostridium’ to look at some organisms in detail.
  The occurrence view provides a quick overview on which are the genomes that contain homologs of our proteins. The darker the color, the better conserved (i.e. the more similar) the protein in the other genome. Again, clicking on one of the squares brings up the Smith-Waterman alignment of the respective E.coli protein with the most similar protein in the species you chose. Notice how in some of the Clostridia, many proteins named ‘trp[..]’ are apparently missing … what might have happened there?
  In this view, effectively each query protein is represented by a ‘profile’ of presences and absences. These profiles capture a summary of evolutionary history. Proteins with similar profiles can be predicted to be functional partners (why?), which is a very powerful prediction method. In general, will this also work well for human proteins? Why / why not?

- Coexpression viewer: this one is conceptually very simple. It shows which genes show a consistent, similar expression response to a variety of external stimuli/conditions. Notice how in this case all the ‘trp[..]’ proteins show a clear coherence in E.coli, but not so strongly elsewhere (probably has not been tested elsewhere).

- Experiments viewer: here, any known experimental evidence supporting a link between two proteins is listed. What is the evidence supporting a link between ‘trpA’ and ‘trpB’? To find out, click on the corresponding row in the table…
  Go back to the experiments view: There is also some evidence in organisms other than E.coli … check, for example, one of the records linking TRP2 and TRP3 in yeast (click ‘info’ again, there). Why is TRP3 shown with two colored dots? Click on the dots to find out (proceed to “homologs among STRING organisms”, and then restrict the view to E.coli K12) … you’ll find that, apparently, TRP3 is homologous to several proteins in E.coli – it may be the result of a gene fusion.

- Database viewer: again, a very simple view. Here we explore what is annotated in curated databases about the proteins we’re currently looking at. Follow the

KEGG pathway labeled “Phenylalanine, tyrosine and tryptophan biosynthesis”, until you leave STRING and reach the KEGG database. Can you recognize the various ‘trp[..]’ E.coli proteins there? If not, this illustrates a common problem with database cross-connections: identical items are often named/displayed in completely different ways. Hint: in this case, for example the ‘trpA’ protein is ‘4.2.1.20’ (E.C. nomenclature), as becomes evident when you click on the 4.2.1.20 box.

- Textmining viewer: this is one of the most powerful views. It shows bodies of text (mostly Pubmed abstracts) that mention at least two of the proteins currently of interest. This is valuable for a quick overview about the literature, but it also provides for yet another ‘prediction’ technique: Proteins that are very often mentioned together (more often than expected by chance alone) are probably functionally linked. Can you think of some problems of this simple assumption?

In summary, STRING provides a very quick way to find about any protein of interest (as long as it is one of the nearly 70 Mio proteins annotated in STRING). It provides domain, annotation, and structural information, and places the protein into a network of its functional partners (the latter derived by a number of direct and indirect prediction techniques). In some organisms, STRING works better than in others, but it is always worth a try … ☺

2). Explore a set of proteins of your interest.

- STRING also offers the possibility to enter several proteins at once. This enables a quick overview of the ‘interaction evidence landscape’ for any protein group of interest
- Think of a set of proteins that are of interest to you, possibly from your research project or of interest to your lab. Try to enter those into STRING: go back to the start page, and choose the tab ‘Multiple Proteins’. There, identify the proteins by name, and select the organism you are interested in. Then click ‘GO’.
- If STRING does not recognize your proteins by name, try to use a different name space (UNIPROT, ENSEMBL, HUGO, SGD, …). If this fails, use the protein sequences as query (on the tab ‘Multiple Sequences’). By the way: this is often the safest way to identity proteins, not only in the STRING database.
- Once you have entered the proteins successfully, proceed to the network view. Do you see any connections between your proteins? If not, lower the ‘confidence cutoff’ … to do this, go towards the bottom of the page: there is a selection box labeled ‘required confidence’, set that to ‘low confidence’ and click the button ‘Update Parameters’.
- now, take a few minutes to explore the evidence landscape. Does it agree with your expectations? Do you discover something that is new to you?
- Important: Feedback to the STRING developer team (via Dr. von Mering). Which interactions that you knew about were missing? Which interactions that STRING showed appeared to be wrong? What can be improved with respect to the user interface?