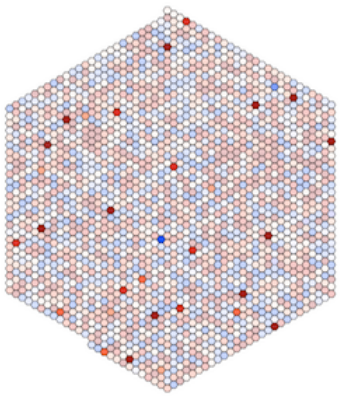
Master Thesis Project:
Dataset layout optimization

Project outline

Considering the complicated relationship between proteins and pathways (see Background section below), visual display of datasets pose several challenges. One such challenge is to visualise *hotspots* in the dataset, where a group of proteins that have some known (or speculative) common function. The goal of this project is to investigate different possibilities for optimising dataset layout for visualisation purposes. More explicitly put:

> Given a dataset with **n** proteins (**n** > 1000) that have associated quantitative information, is it possible to optimise the visualisation of the entire dataset so that the hotspots where proteins that frequently appear together in a pathway are displayed closer to one another?



To the left is an example visualisation where the dataset is visualised in hexagonal grid layout, with a single pathway highlighted. The cells that are greyed out are proteins not a part of the selected pathway.

In this case the layout is not particularly good as the proteins are spread over the entire dataset, instead of being clustered.

Using the same analogy described in the Background section, we can reformulate this question as follows; if we have a large group of individuals that are members of various groups, and a quantifiable characteristic (e.g. BMI or income etc) among these individuals have changed due to a factor (e.g. time) can we identify which groups have been effected the most and display this information in an intuitive manner?

The major questions that this project should aim to address are:

1. Which optimisation or machine learning methods can be used to solve this problem? How do they rate against one another?
2. Can we be sure of distinct optima for an arbitrary dataset; mathematically? biologically? A discussion of the model(s) used in this respect.

Since this is an active field of research, the project can easily be extended, or combined with another project, focusing on the following questions:

- How can a dataset be visualised? Hypergraphs and alternatives
- Practical factors in implementation of this optimisation; runtime? memory constraints? online/local machine? strategies to deal with cross-domain data transfer?
- Exploration of different programming paradigms/languages in this context; functional vs imperative programming?
- …

Requirements

The majority of the work will be *in silico*, thus programming skills and experience is a decisive factor. There is some existing code written in Java for optimisation, meanwhile the visualisation is designed in JavaScript using D3.js library. It is not necessary to use the existing code, although it would likely facilitate getting started with the project.

Besides being interested in systems biology, applicants should meet the following criteria:

- proficient in reading and writing Java code
- capable of working on Linux OS and Eclipse IDE
- familiar with JavaScript language and JSON formatted data

The following are not requirements but would merit extra priority:

- practical or theoretical experience with optimisation using Machine Learning and/or Artificial Intelligence methods (particularly genetic algorithms)
- familiarity with D3.js library and/or previous experience with HTML & CSS
- previous experience with hypergraphs

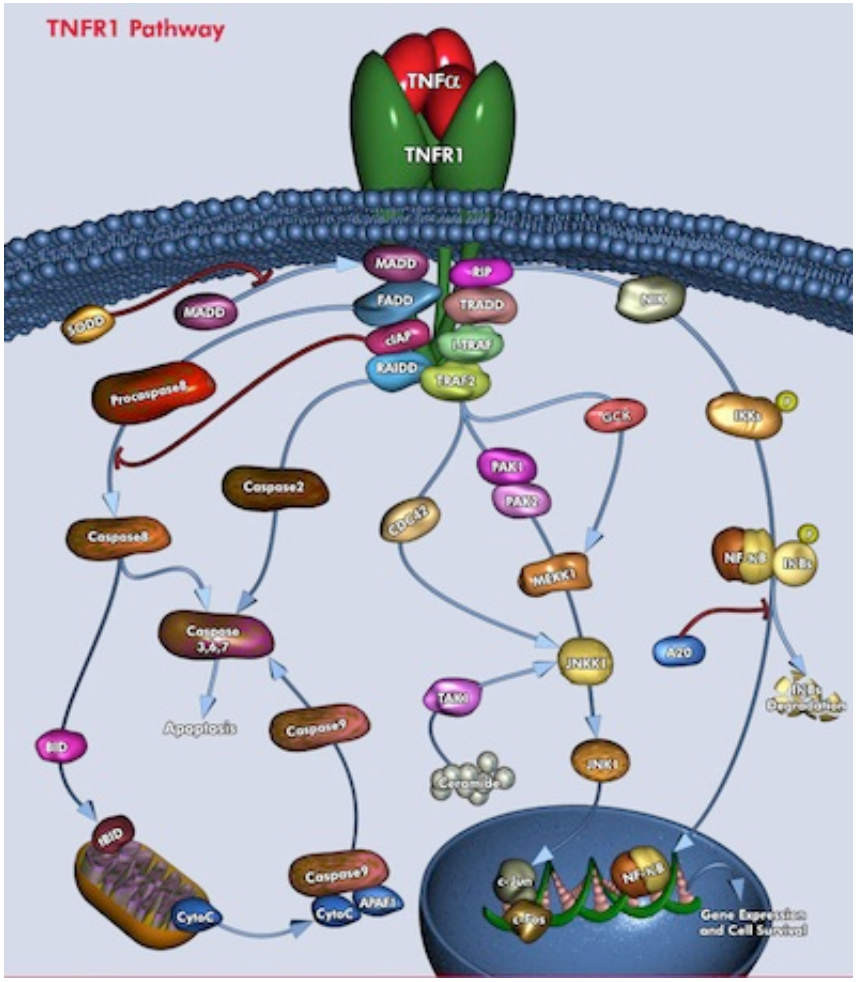
For questions regarding the project contact Ufuk Kirik via email:
`ufuk.kirik@immun.lth.se`

Background

Proteins and pathways in cell biology

Every living cell is dependent on the sustenance of hundreds of biochemical reactions in order to stay alive and fully function. These reactions vary to great extent including, but not limited to, energy metabolism, to maintain cell structure or transport of different molecules in and out of the cell.

Proteins are the macromolecules that are responsible for carrying out all of these reactions, and they typically do not act alone but rather interact with one another, as well as other macromolecules, to carry out their function. In eukaryotic cells, such as human cells, these interactions tend to be very complex and exhibit some degree of dependency to one another. A series of interactions that have result in a distinct functional result are usually expressed in terms of *pathways*, see example below:



Pathways are typically defined as sets of biochemical reactions, sometimes given in a particular order, to describe how a particular function is carried out in a cell or amongst a group of cells. Since there often is a degree of dependence between different interactions and a majority of the vital processes in a cell are composed of a large number of processes pathways typically have a hierarchical structure, such that larger pathways that describe more general functions (sometimes referred to as *superpathways*) include a variety of

subpathways that describe more specific parts of the process.

For instance “DNA Repair” pathway includes six subpathways, one of which is the “Double-strand break repair” pathway, which then in turn contains “Homologous Recombination Repair” and “Non-homologous End-joining” pathways, and so forth. Pathways that share a common parent could either be different steps of a larger process or multiple different ways a particular function might be carried out, or a combination of both. Finally it is worth noting that proteins typically have more than one function and they commonly participate in multiple pathways, that may or may not have a parent/child relationship.

An analogy to the relationship between proteins and pathways can be seen in group membership of people. A person can be a member of various groups of people, like residents of a country, residents of a particular city, member of various interest clubs etc. Some of these memberships are dependent on each other (i.e. being a resident of Lund implies being a resident of Sweden) whereas other memberships do not imply any dependency (i.e. being a member of Barcelona FC fan club does not imply being a resident of Sweden).

Proteomics

Analogous to its older sibling field of genomics which is the study of genome of an cell or organism, *proteomics* is the study of the proteome, or in other words the set of proteins and their activity within a cell, tissue or whole organism. The fundamental difference between the two fields is essentially the temporal and local differences across an organism. The genome of an organism is pretty much conserved across different types of cells and throughout time, whereas the proteome is very dynamical depending both on when and where you look. The dynamic nature of the proteome, together with the fundamental role of proteins in biological systems make proteomics a very active and relevant field.

A typical dataset acquired from a typical experiment contains several thousands of proteins which are quantified usually over multiple samples. It is common practice to compare these values across the samples to try and explain biologically interesting phenomenon, such as why a particular subtype of cancer is more resistant to therapy than a very similar but distinct subtype, or why a particular treatment is effective for a particular disease whereas another is not. Various strategies and software tools exist for this type of analysis.