Ranklust - A bioinformatics solution to identify network biomarkers in cancer

Henning Lund-Hanssen

12th May 2015

Contents

Part I
Introduction

To this day we still struggle with cancer. Even with all our modern equipment and knowledge we have still not been able to tame this horrible disease. My thesis is about making a tool which gives cancer researchers an easier way of identifying network biomarkers in cancer.

1 State of today

1.1 Biomarkers

Biomarkers are at the centre of this thesis. They are what Ranklust should be able to detect and rank in order to identify network biomarkers, and not just single molecules of them. A biomarker is a "biological measure of a biological state" [10]. It can be represented by the levels of a specific protein in our blood, a specific gene, or a combination the two.

Biomarkers can be used for different purposes. They can be used to measure the effect of cancer drug treatment. That the drug does what it is supposed to do. It can be used to predict disease development or the current stage of the disease.

Characteristics for a good biomarker:

- Safe and easy to measure
- Cost efficient to follow up
- Modifiable with treatment
- Consistent across gender and ethnic groups

1.2 Prostate specific antigen

An example of a single molecule biomarker is the prostate specific antigen (PSA). This is a protein produced by the prostate gland in male humans. The identification of cancer with PSA is simple, the higher the level of PSA, measured in ng/mL (nanograms per milliliter), the higher is the chance of the patient having prostate cancer [1].

Today it is used for both identifying and evaluating the current stage of prostate cancer. This biomarker can be found by analyzing blood examples from patients, thus fulfilling some of the demands for a good biomarker, but not all of them. It is easy to measure and easy to acquire, but not reliable enough to be used as the only tool to identify and determine the stage or remission of prostate cancer. The low grade of reliability comes from the fact that even though higher levels of PSA shows higher chance of having prostate cancer, prostate cancer is not the only reason to have elevated levels of PSA [1]. Namely inflammation and enlargement of the prostate. Though, a man with both these cases may or may not develop prostate cancer.

The conclusion for the PSA biomarker is that it is not reliable enough and are causing faulty treatment of prostate cancer that may not even exist. Because even if a patient has prostate cancer, it may not be developing any further, promoting the case of not taking any action against it at all. So there is need for a new biomarker, or at least a better way of identifying them.

1.3 Next-generation sequencing

Today, rapid analyzing of genes and proteins are made available through Next-Generation Sequencing (NGS)[2]. This opens up the possibility of looking at a bigger picture when trying to diagnose cancer patients. Through acquiring more data, faster than before, there now exists databases with much information that is easy to access. This makes room for building huge networks of proteins and genes, allowing for more extensively and thorough assays to be done. For example, what if something that is classified as a prostate cancer biomarker only is viable, when proteins that has not been classified as a biomarker, also is present? Together they could represent a more appropriate *network biomarker*.

2 Single gene prioritization

3 Network generation

4 Clustering

5 Cytoscape

Cytoscape is the open-source software platform the Ranklust app will be developed on. Its main purpose is to visualize molecular interaction networks and biological pathways. It has an easy way to integrate ***Apps***, which may be combined by other apps again to build big and complex applications which may solve problems in a bigger picture. The goal of Ranklust is to cluster the networks we get from single gene prioritization and rank them in order to identify network biomarkers. Apps taking care of making the networks already exists, but there still has to be made a decision about whether or not they could be modified in order to better support the clustering.

Which databases to use has to be considered. The reason to use databases is because they have information on how protein and genes form a network based on how they interact with eachother. The initial database candidates in Ranklust are iRefIndex [3], GeneMania [4] and STRING [5]. These databases all have in common that it exists Cytoscape apps made to use these databases. STRING however, does not have any repository available through the Cytoscape app store, so interacting with the database through a new app in Cytoscape without making new plugins may be difficult. On the other hand, both iRefIndex and GeneMania have their repositories easily available to the public together with decent documentation. However, the difference between them is what they contain information about. iRefIndex contains data about protein-protein interaction (PPI), while GeneMania contains data about genes. Since proteins come from genes, GeneMania can

also give us some information about proteins. Differences between the two databases will be discussed in greater depth at a later stage.

The open-source plugins in Cytoscape to communicate with the databases are iRefScape [6] for iRefIndex and GeneMANIA [7] for GeneMania.

Cytoscape is based on the Java programming language, which is a little bit untraditional for a software platform used to develop bioinformatics tools. The reason to choose Java above Python, Perl or other popular programming languages is simply because I am more versed in Java programming than any other language. The Cytoscape community also promotes the idea of developing apps that follow the *OSGi* standard [8].

Developing OSGi software should promote modularization of the code and increase the probability of the app being launched as an official Cytoscape app, in addition to provide other developers with the possibility of reusing my modules in their own apps. There exists several design patterns that could prove to be useful in the development of Ranklust. Another strategy to follow may not be a direct design patter, but more of a collection of them, is the clean code principles. More thorough examination of these strategies will follow.

Prototyping of different parts of the app will be done in Python and the Galaxy environment [9]. Python is an easy language to prototype with, and Galaxy is an easy environment to test small scripts with. Galaxy also has great logging of previous experiments combined with settings, so recreating simulations and comparing results is easy to do and reliable. However, in my experience, Java comes up to par with Python in development speed once all of the boilerplate code is written and a good deployment tool is used. Therefore the Cytoscape platform is used for the final deployment of Ranklust.

Part II
Background

Introduction

We have our body, and inside our body we have our organs. These organs are made up of tissue, and tissue is made up of cells. Our cells performs two type of functions, to execute chemical reactions needed to stay alive, and to pass information for maintaining life onto the next generation.

Background

DNA, RNA and protein in general

The name of the process when DNA goes from DNA to RNA to protein is called the **Central Dogma**. I will focus on the cell, how we perceive it and the interaction inside it. The cell can be seen as a network community of interacting protein molecules. The protein comes from our DNA.

In our DNA, there are areas that contain codes for making protein. These areas are called genes. Also, all of the possible interactions between the cells are specified by the proteins in complex social networks. The reason for calling these networks for "social networks" is because it is the interactions that we look at as the connections between the cells. It is our protein that performs chemical reactions in our body. DNA on the other hand stores and passes information about how our body is built up. RNA is the intermediate stage between DNA and protein.

Protein

The protein in our body is made up from amino acids. The amount of amino acids in a protein may vary from 20 to 5000. But on average there is about 350 amino acids in a protein in our body. The way we identify amino acids is through the alphabet. We use almost every letter in the alphabet, and it is organized in a chronological order when we show them in a list, but we miss some characters. They can also be identified by three letters, or the whole name of the amino acid.

Amino acids have some very basic attributes like volume and mass. But beeing acids we also have information about their polarity and their basicity/acidity. So the amino acids are either polar or non-polar, combined with beeing neutral, acidic or basic. And of course we are able to see to what degree they are acidic/basic.

Amino acids consist of an amino group, carboxyl group and a R group. Very often there is also a central carbon that combines all of these groups together. We have about 20 different amino acids and they can be classified into 4 types. The positively charged, the negatively, the polar and the non-polar. The positive amino acids are basic and the negative ones are acidic. The four types just mentioned are not totally exclusively to eachother, though an amino acid cannot be both basic and acidic at the same time. At the same time an amino acid cannot be both polar and non-polar. But any combination of acidity/basicity together with polarity can occur.

Genome

When the genome changes suddenly and unexpected, we have a mutation. A mutation may happen when several things occur:

- Insertion
  - A part of a chromosome gets inserted into another one
- Deletion
  - A part of a chromosome gets deleted
- Duplication
  - A part of a chromosome gets duplicated
- Inversion
  - A part of a chromosome gets inverted, but yet it stays in place
- Translocation
  - Two chromosomes exchanges parts and they become what is called a derivative

References

[1] URL: http://www.cancer.gov/cancertopics/types/prostate/psa-fact-sheet (visited on 11/05/2015).

[2] URL: http://www.illumina.com/technology/next-generation-sequencing.html (visited on 11/05/2015).

[3] URL: http://irefindex.org/wiki/index.php?title=iRefIndex (visited on 07/05/2015).

[4] URL: http://www.genemania.org/ (visited on 07/05/2015).

[5] URL: http://string-db.org/ (visited on 07/05/2015).

[6] URL: http://apps.cytoscape.org/apps/irefscape (visited on 07/05/2015).

[7] URL: http://apps.cytoscape.org/apps/genemania (visited on 07/05/2015).

[8] URL: http://wiki.cytoscape.org/Cytoscape_3/AppDeveloper (visited on 08/05/2015).

[9] URL: https://usegalaxy.org/ (visited on 08/05/2015).

[10] MD Dr Ananya Mandal. URL: http://www.news-medical.net/health/What-is-a-Biomarker.aspx (visited on 12/05/2015).