Ranklust - A bioinformatics solution to identify network biomarkers in cancer

Henning Lund-Hanssen

22nd May 2015

Contents

1 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, named Ranklust, which gives cancer researchers an easier way of identifying network biomarkers in cancer.

2 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" [16]. 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. Here is a list of what biomarkers currently are being used for:

Usages for biomarkers: [22]

- Disease disposition
  - What is a patient's risk of developing cancer in the future?
- Screening
  - Does earlier detection of patients with cancer decrease mortality?
- Diagnostic
  - Who has cancer? What is the grade of the cancer?
- Prognostic
  - What clinical outcome is most likely if therapy is not administered?
- Predictive
  - Which therapy is most appropriate?
- Monitoring
  - Was therapy effective? Did the patient's disease recur?
- Pharmacogenomic

– What is the risk for adverse reaction to the prescribed therapeutic dose?

Characteristics for a good biomarker:

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

3 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 PSA 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 marker 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 of these cases may or may not develop prostate cancer.

So 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 harmful, 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 to diagnose and predict the right treatment.

4 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*. The amount of data that can be analyzed also opens up for another more personalized approach to each cancer patient. Finding patient-specific biomarkers could make a huge impact on the quality of treatment [21].

5 Next generation of biomarkers

The PSA biomarker is over 20 years old [17]. Through those years, it has been discovered other and better biomarkers for prostate cancer than PSA. Among those, PCA3, which is detectable through urine samples from patients. It also has the benefit of not being affected by the size of the prostate gland [18]. But still the results could be better. Therefore, it has been tried to combine these two biomarkers in order to see if it is beneficial to see the results from each biomarker in light of each other [22]. The results from these tests is that they complement each other to a level of significance that makes it compelling to analyze them both to diagnose prostate cancer. It is important to point out that even if these biomarkers are not the best at indicating if a patient has cancer or not, these biomarkers are good at indicating progression and recurrence of prostate cancer.

But all of this is based on single genes or proteins. What if we looked at whole networks as biomarkers?

6 Network

Viewing the cell as a network of proteins and genes presents us with several assumptions to make. Firstly, there is no physical connection between the proteins and genes, which represents the nodes in the network. So a way of defining edges between nodes has to be established. It exists several ways of doing this, but there has yet to be discovered a silver bullet of defining the edges.

Bayesian probability networks are one way of defining edges [19]. It is based on the probability that if a node A exists, then node B exists. That way it is possible to create networks based on the assumption that if node A exists, then node B is 80% likely to exist. Maybe node C and D have a 100% chance of existing if node B exists. These percentages represents how strong the edges between the nodes are, and makes it able to construct a directed acyclic graph, DAG for short. It is important to note that a true bayesian network can not be achieved through random observations. Rather should some constant value(s) be introduced to be able to really measure the effect of the other variables.

The way of defining edges in a network also determines what kind of information that is possible to get from it. The different bio-databases has different ways of calculating edges. Should the database alone define the edges? If the database supply weights in the edges and/or nodes, should it be used, changed or ignored? The final decision of how and in what way the edges should be defined has yet to come.

7 Clustering

Networks is the pre-processing of single gene prioritization that is necessary in order to come to the next step, clustering of the network. Clustering is about making a hierarchical view of a network to be able to look at the bigger picture of the cell. The reason to group a network into clusters and rank them is that the edges we create represents different connections. They can represent function, probability of existence, interactions or contents of the cell [15]. The function of the cell is what Ranklust is after. There is several algorithms to create clusters, and all need to be heavily researched in order to find the best suiting one. The major differences on how these algorithms work is centered around how they handle vertices and edges. For example how the cluster is expanded, what density level it is aiming for, how robust the algorithm is towards incomplete networks. It is feasible if the cluster-making algorithm is easily, or even already implemented as an app in Cytoscape. ClusterMaker2 is such an app and will be considered [3]. But there is faster cluster-creating algorithms than those implemented in ClusterMaker2, namely the SPICi [20] algorithm. So a possible solution for Ranklust could be to implement the SPICi algorithm into the ClusterMaker2 app, in order to reuse most of the code that represents the GUI, single gene prioritization and network creation.

In addition to the cluster-creating algorithms, there is a need for cluster-ranking algorithms.

8 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 is easy to integrate ***Apps*** and even combine multiple apps to solve new problems, given that the source code of the apps is available or that it exists an easy way of piping results from one app to another.

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 about how protein and genes form a network based on how they interact with eachother. The initial database candidates in Ranklust are iRefIndex [4], GeneMania [5] and STRING [6]. 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 [7] for iRefIndex and GeneMANIA [8] 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. Java is known for being this big bloated enterprise programming language, and Python as the fast and easy mockup tool to develop good programs fast. Python also has big biological computation libraries like *Biopython* [9], that makes it easy to build your own standalone apps. Though when used in a bigger environment as Cytoscape, Java shines, having sturdy packaging and modelling standards. The Cytoscape community is big, alive and has standards for how the architecture of apps should look like. The community promotes this through the use of the *OSGi* standard [10].

Developing OSGi software should promote modularization [11] 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. Also, it seems like Java 9 is aimed at making it easier to modularize apps along the lines of the OSGi standard, so it might be easier to refactor an application in the future from Java 8 to Java 9 when the architecture already is in place. 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 pattern [12], but more of a collection of them, is the clean code principles by Robert C. Martin [13]. More thorough examination of these strategies will follow.

Prototyping of different parts of the app will be done in Python and the

Galaxy environment [14]. Python is an easy language to prototype with, and Galaxy is an easy environment to test small scripts in. 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.

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://apps.cytoscape.org/apps/clustermaker2 (visited on 22/05/2015).

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

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

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

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

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

[9] URL: http://biopython.org/wiki/Main_Page (visited on 20/05/2015).

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

[11] URL: http://www.javaworld.com/article/2878952/java-platform/modularity-in-java-9.html (visited on 20/05/2015).

[12] URL: http://www.techopedia.com/definition/18822/design-pattern (visited on 20/05/2015).

[13] URL: http://www.amazon.com/Clean-Code-Handbook-Software-Craftsmanship/dp/0132350882 (visited on 20/05/2015).

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

[15] Anne-Ruxandra Carvunis and Trey Ideker. 'Siri of the Cell: What biology Could Learn from the iPhone'. In: *CellPress* 157 (Apr. 2014).

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

[17] Michael F.Berger et al. 'The genomic complexity of primary human prostate cancer'. In: *Nature* 1 (2011).

[18] Alexander Haesea, Alexandre de la Tailleb, Hendrik van Poppelc, Michael Marbergerd, Arnulf Stenzle, Peter F.A. Muldersf, Hartwig Hulandg, Clément-Claude Abboub, Mesut Remzid, Martina Tinzld, Susan Feyerabende, Alexander B. Stillebroerf, Martijn P.M.Q. van Gilsf and Jack A. Schalkenf. 'Clinical Utility of the PCA3 Urine Assay in European Men Scheduled for Repeat Biopsy'. In: *European Urology* 54 (2008).

[19] David Heckerman. *A Tutorial on Learning With Bayesian Networks*. MSR-TR-95-06. Microsoft Research, Mar. 1995, p. 57. URL: http://research.microsoft.com/apps/pubs/default.aspx?id=69588.

[20] P. Jiang and M. Singh. 'SPICi: a fast clustering algorithm for large biological networks'. In: *Bioinformatics* 26.8 (15th Apr. 2010), pp. 1105–1111. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btq078. URL: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq078 (visited on 22/05/2015).

[21] Rebecca J. Leary, Isaac Kinde, Frank Diehl, Kerstin Schmidt, Chris Clouser, Cisilya Duncan, Alena Antipova, Clarence Lee, Kevin McKernan, Francisco M. De La Vega, Kenneth W. Kinzler, Bert Vogelstein, Luis A. Diaz Jr. and Victor E. Velculesc. 'Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing'. In: *Science Translational Medicine* 2 (2010).

[22] John R. Prensner, Mark A. Ruin, John T. Wei and Arul M. Chinnayiyan. 'Beyond PSA: The next generation of prostate cancer biomarkers'. In: *Sci Transl Med* 1 (2012).