Ranklust

A contribution to the Cytoscape plugin clusterMaker2 and its applicaton to find prostate cancer candidate biomarker genes

Henning Lund-Hanssen

Thesis submitted for the degree of
Master in Programming and Networks

60 credits

Department of Informatics

Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Autumn 2016

Ranklust

A contribution to the Cytoscape plugin clusterMaker2 and its applicaton to find prostate cancer candidate biomarker genes

Henning Lund-Hanssen

Ranklust

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

New candidates for prostate cancer gene biomarkers are needed, and through the development of the Ranklust contribution to the Cytoscape app, clusterMaker2. We show how it can be used to identify genes that could be prostate cancer candidate biomarkers, through ranking clusters in Protein-Protein Networks (PPI) and filtering the data afterwards.

Contents

List of Figures

List of Tables

Listings

Part I

Intro

Chapter 1

Introduction

1.1 Oversight

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. This thesis is about implementing and using a tool named Ranklust. It is not a standalone tool, but rather a contribution to a Cytoscape plugin named clusterMaker2[1][2]. The goal of this tool is to rank clusters created from Protein-Protein Interaction (PPI) networks in Cytoscape. The ranks will be based on different node and edge attributes in the network. The resulting ranks will also indicate which clusters can be seen as cluster biomarkers, and which genes could be considered to be single gene candidate biomarkers.

Part II

Background

Chapter 2

TODO: Find a good chapter here!

2.1 Biomarkers

A biomarker is a "biological measure of a biological state" [22]. Among other things, 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: [34]

- 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

The National Institutes of Health Biomarkers Definitions Working Group has defined biomarkers as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic process, or pharmacologic responses to a therapeutic intervention."[24, 36]. The World Health Organization (WHO) has created a guideline for defining biomarkers: "almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical or biological. The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction". So a biomarker can be anything from the pulse of the heart or blood pressure. In this thesis, I will see if we can get better and more information from ranking gene clusters based on biomarker information. The result could in the best case be a new biomarker in itself, and can be a step in the direction of hierarchical biomarkers (biomarkers creating new biomarkers).

2.1.1 Biomarkers and Clinical Endpoints

Some of the important characteristics of biomarkers is that they are objective and quantifiable as biological processes[36]. They are used as a state indicator of the biological object that is under scrutiny. When the biological object is a human being screened for cancer, biomarkers may help.

Biomarkers can be an indicator of how far cancer has progressed in the human body, even though the human subject may feel no difference at all. It can also be the opposite, that the human subject feels huge differences during several weeks that the cancer might have developed at a grand scale, but the biomarkers show no objective change. This proves that the human subject's experience and sense of state that it is in does not necessarily have to correlate.

Clinical endpoints is the opposite[36]. They describe how the subjects feel or describe how they function. It is not as objective as biomarkers and it demands more resources to gather the information, as the subjects has to be interviewed in some form, rather than interpreting pure data. But one thing has to be noted; patients does not seek treatment for cancer due to their biomarkers being "off the charts". They seek treatment because they feel that they do not feel ok or do not function sufficiently. So biomarkers are not by any means far superior to clinical endpoints in all aspects.

Biomarkers can also be ruled out as a reliable predictor when population differences are too big[36]. As with clinical endpoints, people describe their feelings differently and might hold information back. As pointed out before, people's feelings are subjective, and different ways of interpreting their own body's state might skew statistical results with erronous feedback. Erronous feedback in this case can be examplified as two persons who are at the exact same state of cancer, with the same prerequisites, but they report how they feel differently. One person might be ignoring certain pains or lose hair after radiation treatment, but not the other. This can be mitigated to some degree with careful and accurate screening of patients admitted to treatment, but a totally unified group in terms of how they describe pain and other attributes that are interesting might be seen as borderline impossible. We are after all humans and very much fallible.

2.1.2 Prostate cancer statistics

Statistics show that "In 2013, there were 1.4 million incident cases of prostate cancer and 293 000 deaths"[31]. Taking population into concideration together with the increasing incidence rates, we have had a "3-fold increase in prostate cancer cases since 1990"[31].

2.1.3 Specific biomarkers

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 [3].

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 [3]. 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.

Figure 2.1: Cancers Ranked by Number of Incident Cases in Both Sexes, Globally, by Development Status, and in the 50 Most Populous Countries, 2013

Region	Country	Breast cancer	Tracheal, bronchus, and lung cancer	Colon and rectum cancer	Prostate cancer	Stomach cancer	Liver cancer	Cervical cancer	Non-Hodgkin lymphoma	Esophageal cancer	Leukemia	Lip and oral cavity cancer	Bladder cancer	Uterine cancer	Pancreatic cancer	Brain and nervous system cancer	Kidney cancer	Malignant skin melanoma	Ovarian cancer	Thyroid cancer	Gallbladder and biliary tract cancer	Larynx cancer	Other pharynx cancer	Multiple myeloma	Hodgkin lymphoma	Nasopharynx cancer	Testicular cancer	Mesothelioma
Global		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Developed		3	4	2	1	5	11	17	7	20	12	14	6	13	8	16	10	9	15	18	19	22	23	21	24	27	25	26
Developing		1	2	4	6	3	5	7	11	8	10	9	16	13	14	12	20	23	18	15	19	17	21	25	24	22	26	27
High-income Asia Pacific	Japan	4	2	1	6	3	5	17	10	11	13	14	9	18	7	22	12	23	15	16	8	21	20	19	27	26	25	24
	South Korea	5	2	1	8	3	4	13	10	15	14	18	11	19	9	16	12	23	17	6	7	21	22	20	25	24	27	26
High-income North America	Canada	4	3	2	1	12	18	20	5	19	10	14	6	8	11	13	9	7	15	17	21	22	23	16	24	27	25	26
	United States	2	3	4	1	13	18	20	7	19	11	12	6	9	10	17	8	5	15	14	23	21	22	16	24	27	25	26
Southern Latin America	Argentina[a]	1	4	3	2	5	18	6	8	14	12	16	10	7	9	19	11	15	17	21	13	20	25	23	24	26	22	27
Western Europe	France	3	4	2	1	10	14	19	6	18	13	12	5	7	11	16	9	8	17	22	23	21	15	20	24	27	25	26
	Germany	3	4	2	1	6	16	21	9	18	11	15	5	12	8	14	7	10	13	22	17	23	19	20	26	27	24	25
	Italy	2	4	1	3	6	8	21	7	22	11	19	5	12	9	14	10	13	15	18	16	20	24	17	25	27	26	23
	Spain	4	3	1	2	6	11	20	7	23	12	13	5	8	9	15	10	14	16	22	19	17	21	18	24	27	25	26
	United Kingdom	4	3	2	1	7	18	19	6	12	11	17	5	14	9	15	10	8	13	24	23	20	22	16	26	27	25	21
Central Asia	Uzbekistan[a]	1	3	8	11	2	9	4	12	5	6	10	20	13	16	7	17	18	19	24	21	14	23	25	15	26	22	27
Central Europe	Poland	3	1	2	4	6	19	11	14	20	12	16	5	9	8	13	7	17	10	21	18	15	23	22	25	26	24	27
Eastern Europe	Russia	3	2	1	5	4	17	9	14	19	12	11	8	10	7	18	6	16	13	15	21	20	22	24	23	26	25	27
	Ukraine	2	3	1	4	5	21	9	15	20	11	8	10	17	7	14	6	13	12	16	19	18	22	24	23	27	26	25
Andean Latin America	Peru[a]	3	6	4	1	2	9	5	7	21	8	18	20	10	12	15	14	17	16	11	13	24	25	19	23	27	22	26
Central Latin America	Colombia[a]	2	6	4	1	3	10	5	9	17	7	16	23	13	12	11	20	18	14	8	15	19	25	22	21	26	24	27
	Mexico[a]	2	7	3	1	5	10	4	9	22	8	18	24	16	12	14	11	20	13	6	17	19	25	23	21	27	15	26
	Venezuela[a]	2	4	5	1	6	13	3	7	19	8	16	22	10	11	17	12	20	15	9	18	14	25	21	23	26	24	27
Tropical Latin America	Brazil[a]	2	4	3	1	5	17	6	11	14	8	10	21	16	12	7	18	13	20	9	22	15	19	23	24	26	25	27
East Asia	China[a]	5	1	4	9	2	3	12	11	6	8	19	14	7	13	10	20	26	21	17	18	16	25	24	22	15	27	23
	North Korea[a]	4	1	5	11	3	2	7	12	6	9	14	17	8	13	10	22	23	21	20	19	16	25	24	18	15	27	26
Southeast Asia	Indonesia[a]	1	2	4	7	5	10	3	8	21	11	6	18	13	15	12	19	23	14	9	16	20	22	25	24	17	26	27
	Malaysia[a]	1	2	3	4	7	6	10	5	19	8	14	15	12	17	18	16	22	11	9	23	20	21	25	24	13	26	27
	Myanmar[a]	1	2	5	11	12	4	3	9	17	7	6	22	8	15	14	23	24	10	13	16	19	20	25	21	18	26	27
	Philippines[a]	1	2	4	3	12	7	5	11	23	6	10	20	9	15	14	18	21	13	8	19	17	22	26	25	16	24	27
	Thailand[a]	3	1	4	5	7	2	6	10	19	11	8	13	17	16	14	18	23	15	12	9	20	21	26	25	22	24	27
	Vietnam[a]	4	2	5	12	3	1	8	6	10	11	7	18	13	17	9	21	25	20	14	19	15	16	26	24	22	23	27
South Asia	Afghanistan[a]	3	2	8	10	1	9	4	12	17	5	14	11	13	18	6	20	23	22	16	21	15	25	24	7	19	26	27
	Bangladesh[a]	3	4	7	11	6	2	8	5	10	9	1	19	17	22	14	23	24	12	26	18	15	13	25	20	16	21	27
	India[a]	1	6	4	15	5	8	3	11	7	10	2	19	24	22	12	21	16	14	17	20	13	9	25	18	23	26	27
	Nepal[a]	1	4	6	13	7	9	3	12	5	8	2	21	22	17	14	23	25	11	20	16	15	10	24	18	19	26	27
	Pakistan[a]	1	3	6	8	13	12	15	5	4	7	2	9	17	24	14	22	25	11	18	20	10	16	23	19	21	26	27
North Africa and Middle East	Algeria[a]	1	3	2	10	4	15	7	6	23	5	21	14	20	16	8	22	24	17	12	9	18	25	19	13	11	26	27
	Egypt[a]	1	6	7	3	9	2	11	13	20	4	14	8	15	10	5	16	23	19	12	18	17	24	25	22	26	21	27
	Iran[a]	2	5	6	3	1	9	15	10	4	7	14	12	24	17	8	18	21	19	13	16	11	26	23	20	25	22	27
	Iraq[a]	1	2	5	8	6	7	10	11	19	3	15	14	9	12	4	16	26	17	13	20	18	24	23	21	25	22	27
	Morocco[a]	1	2	6	3	5	9	4	10	20	8	13	15	11	12	7	21	23	17	16	14	19	25	26	18	22	24	27
	Saudi Arabia[a]	1	4	2	6	9	3	16	5	17	8	13	14	22	11	7	15	26	18	10	12	19	21	23	24	25	20	27
	Sudan[a]	1	2	4	5	3	8	11	9	15	6	14	10	18	17	7	19	22	20	12	21	16	26	24	13	23	25	27
	Turkey[a]	2	1	3	4	5	13	18	7	24	6	20	10	11	9	8	16	22	14	12	19	15	27	21	23	25	17	26
	Yemen[a]	1	2	5	8	3	9	6	10	19	4	15	11	12	18	7	22	23	20	13	16	17	26	24	14	21	25	27
Central sub-Saharan Africa	DRC[a,b]	1	7	5	3	4	6	2	9	8	11	10	16	17	15	12	19	14	20	24	18	21	23	22	13	25	26	27
Eastern sub-Saharan Africa	Ethiopia[a]	1	8	4	3	7	6	2	9	5	18	11	21	15	16	10	12	13	19	22	17	25	24	20	14	26	23	27
	Kenya[a]	1	10	5	4	6	9	3	8	2	14	7	15	23	13	12	20	18	11	17	24	16	25	19	21	22	26	27
	Mozambique[a]	2	7	4	1	5	6	3	8	10	14	11	19	18	16	9	13	15	20	25	17	24	23	21	12	26	22	27
	Tanzania[a]	1	10	4	2	8	5	3	7	6	14	11	21	18	17	9	13	15	19	23	16	25	24	20	12	26	22	27
	Uganda[a]	3	9	6	1	8	7	2	4	5	12	11	16	15	19	21	18	20	10	13	25	24	22	23	14	17	26	27
Southern sub-Saharan Africa	South Africa[a]	2	4	3	1	10	11	6	7	5	13	8	15	14	9	21	17	12	16	20	23	18	22	19	24	27	25	26
Western sub-Saharan Africa	Ghana[a]	3	10	5	1	6	4	2	8	13	12	17	15	7	9	11	20	19	14	21	18	25	26	22	16	24	23	27
	Nigeria[a]	3	9	5	4	6	1	2	7	16	11	17	13	8	12	14	18	20	10	21	19	23	25	22	15	26	24	27

Figure 2.2: Cancers Ranked by Number of Deaths in Both Sexes, Globally, by Development Status, and in the 50 Most Populous Countries, 2013

Region	Country	Tracheal, bronchus, and lung cancer	Stomach cancer	Liver cancer	Colon and rectum cancer	Breast cancer	Esophageal cancer	Pancreatic cancer	Prostate cancer	Leukemia	Cervical cancer	Non-Hodgkin lymphoma	Brain and nervous system cancer	Bladder cancer	Ovarian cancer	Gallbladder and biliary tract cancer	Lip and oral cavity cancer	Kidney cancer	Larynx cancer	Multiple myeloma	Other pharynx cancer	Uterine cancer	Nasopharynx cancer	Malignant skin melanoma	Mesothelioma	Thyroid cancer	Hodgkin lymphoma
Global		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Developed		1	3	7	2	4	11	5	6	8	17	9	14	10	13	15	18	12	21	16	22	20	26	19	23	24	25
Developing		1	3	2	4	6	5	9	12	8	7	11	10	14	15	16	13	18	17	22	19	21	20	24	26	23	25
High-income Asia Pacific	Japan	1	2	4	3	9	7	5	8	11	16	10	18	12	14	6	17	13	23	15	19	20	24	25	22	21	26
	South Korea	1	3	2	4	8	7	5	9	11	13	10	14	12	16	6	20	15	18	17	21	22	24	23	25	19	26
High-income North America	Canada	1	7	13	2	3	12	5	4	8	20	6	9	10	11	17	18	14	22	15	23	19	25	16	21	24	26
	United States	1	9	8	2	3	13	4	5	7	18	6	12	14	10	20	19	11	21	15	22	17	26	16	23	24	25
Southern Latin America	Argentina[a]	1	4	8	2	3	7	6	5	11	10	13	16	14	15	12	19	9	17	20	23	18	26	21	24	22	25
Western Europe	France	1	7	6	2	4	11	5	3	8	20	10	14	9	13	17	16	12	21	15	18	22	25	19	23	24	26
	Germany	1	6	10	2	3	14	4	5	7	19	11	13	9	12	15	18	8	23	16	20	21	26	17	22	24	25
	Italy	1	3	6	2	4	16	5	7	9	22	10	12	8	13	14	18	11	20	15	23	21	26	17	19	24	25
	Spain	1	3	7	2	5	14	6	4	9	21	10	11	8	12	16	18	13	17	15	22	19	26	20	23	24	25
	United Kingdom	1	7	14	2	3	6	5	4	10	21	8	13	11	9	20	19	12	22	15	23	18	26	16	17	24	25
Central Asia	Uzbekistan[a]	2	1	5	7	4	3	10	18	6	9	11	8	16	15	17	12	14	13	24	20	19	23	21	27	26	22
Central Europe	Poland	1	3	12	2	4	16	5	6	11	14	15	8	7	9	13	18	10	17	21	22	19	25	20	26	23	24
Eastern Europe	Russia	1	3	7	2	4	13	5	6	10	12	16	11	14	9	18	15	8	17	22	21	19	25	20	26	23	24
	Ukraine	1	3	14	2	4	16	5	8	11	10	15	12	9	7	18	13	6	17	23	20	21	25	19	24	22	26
Andean Latin America	Peru[a]	2	1	6	3	7	17	8	4	9	5	10	12	16	14	11	20	13	22	18	23	15	27	19	24	21	25
Central Latin America	Colombia[a]	2	1	7	3	5	12	9	4	8	6	10	11	15	14	13	18	17	16	19	24	22	25	21	27	20	23
	Mexico[a]	1	2	3	5	6	15	8	4	9	7	10	13	17	12	14	19	11	16	18	26	22	27	23	25	20	21
	Venezuela[a]	1	2	7	4	5	13	8	3	9	6	10	14	17	11	16	19	12	15	18	23	20	25	21	27	22	24
Tropical Latin America	Brazil[a]	1	3	8	2	5	7	6	4	11	9	12	10	13	14	18	15	17	16	20	19	21	25	22	26	23	24
East Asia	China[a]	1	3	2	5	8	4	6	15	7	12	11	9	13	18	14	20	17	19	22	23	16	10	26	21	24	25
	North Korea[a]	1	3	2	4	7	5	10	20	6	9	12	8	11	16	14	19	18	17	22	24	15	13	26	25	23	21
Southeast Asia	Indonesia[a]	1	3	6	2	5	14	8	19	7	4	10	9	16	12	13	11	17	20	24	18	21	15	23	25	22	26
	Malaysia[a]	1	5	3	2	4	13	7	12	6	10	8	17	15	11	18	16	14	21	20	19	23	9	24	27	22	25
	Myanmar[a]	1	7	2	4	5	13	11	20	6	3	9	14	15	8	12	10	17	21	25	19	18	16	24	26	22	23
	Philippines[a]	1	6	2	4	3	16	8	11	5	7	10	12	19	9	15	13	17	22	23	21	20	14	24	27	18	25
	Thailand[a]	2	4	1	3	6	12	9	16	8	5	13	11	15	14	7	10	17	20	24	19	22	18	23	26	21	25
	Vietnam[a]	2	3	1	4	11	5	12	14	9	8	6	7	16	22	19	10	21	15	24	13	18	20	26	27	17	25
South Asia	Afghanistan[a]	2	1	5	7	3	12	13	14	4	6	10	9	8	18	17	20	16	11	24	23	21	19	26	25	22	15
	Bangladesh[a]	2	3	1	8	9	7	19	18	5	10	4	13	14	11	16	6	20	15	24	12	21	17	25	26	27	22
	India[a]	4	1	2	7	5	3	15	20	10	6	11	13	16	12	17	8	19	14	21	9	25	18	22	27	23	24
	Nepal[a]	2	3	6	5	4	1	16	20	8	7	11	15	17	10	14	9	19	13	21	12	23	18	25	27	22	24
	Pakistan[a]	1	10	5	9	2	3	18	19	8	16	4	13	7	11	15	6	20	12	21	14	22	17	26	27	23	24
North Africa and Middle East	Algeria[a]	1	3	9	4	2	18	13	15	5	12	6	8	10	14	7	22	19	16	17	23	24	11	25	26	20	21
	Egypt[a]	2	7	1	8	3	11	9	10	5	14	12	4	6	17	16	18	13	15	23	21	20	27	24	25	19	22
	Iran[a]	3	1	9	5	8	2	12	7	4	17	14	6	10	16	13	18	15	11	20	24	26	19	22	27	21	23
	Iraq[a]	2	5	4	7	1	15	8	11	3	12	10	6	9	13	17	19	14	16	21	23	18	22	25	27	20	24
	Morocco[a]	1	3	5	4	2	15	9	6	10	7	11	8	12	14	13	19	17	16	24	23	18	20	25	26	21	22
	Saudi Arabia[a]	2	6	1	3	4	12	8	10	7	17	9	5	13	15	11	16	14	18	22	21	25	20	24	27	19	23
	Sudan[a]	1	2	6	4	3	12	11	10	5	13	9	7	8	16	15	18	17	14	22	24	23	19	25	27	20	21
	Turkey[a]	1	2	10	3	4	16	5	6	7	18	11	8	9	12	15	22	13	14	17	27	19	23	24	21	20	25
	Yemen[a]	1	2	6	5	3	13	12	11	4	10	9	7	8	16	14	18	17	15	23	24	20	19	25	26	21	22
Central sub-Saharan Africa	DRC[a,b]	6	2	5	4	3	7	11	9	10	1	8	14	12	17	13	15	16	18	21	22	20	24	23	27	25	19
Eastern sub-Saharan Africa	Ethiopia[a]	7	6	4	3	5	2	10	9	14	1	8	12	16	17	13	15	11	22	18	21	20	26	23	27	25	19
	Kenya[a]	8	4	5	6	3	1	9	11	14	2	7	15	13	10	19	12	18	16	17	22	25	20	24	27	23	21
	Mozambique[a]	7	3	4	1	5	9	12	6	13	2	8	11	15	18	14	16	10	21	19	20	22	26	23	25	27	17
	Tanzania[a]	8	6	2	3	4	5	13	9	12	1	7	10	16	17	14	15	11	22	19	20	21	25	23	26	27	18
	Uganda[a]	9	8	4	6	7	2	13	5	11	1	3	20	12	10	24	16	15	21	18	17	23	14	25	27	19	22
Southern sub-Saharan Africa	South Africa[a]	1	9	8	3	5	2	7	4	11	6	10	17	14	12	21	13	16	20	15	22	19	25	18	23	24	26
Western sub-Saharan Africa	Ghana[a]	9	3	1	6	5	10	7	4	14	2	8	15	12	11	16	19	17	25	18	24	13	23	21	26	22	20
	Nigeria[a]	7	3	1	5	4	10	9	8	13	2	6	15	11	12	14	18	17	21	20	23	16	25	22	26	24	19

Chapter 3

Big data in bioinformatics and networks as a tool in cancer research

3.1 Existing biomarkers

The PSA biomarker is over 20 years old [23]. 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 [25]. 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 [34]. 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.

In the cases of pancreatic, lung, breast, brain and overian cancer, the somatic distribution of single-nucleotide polymorphisms (SNP) has a few altered genes that occur with a frequency higher than 10%, and many other genes that are mutated occur with a frequency of 5% or lower[20]. On the other hand, prostate cancer has relatively few SNPs and copy-number alterations (CNA), so that kind of cancer is more likely to be driven by another somatic variation, namely DNA methylation. Cancer driver genes can be detected by positive-selection signals in the mutation pattern of genes in tumors, but there is a drawback with this method. Genes that are less frequently mutated within tumor samples might not be identified by a statistical analysis, even if they might be functionally important. An

alternative method is to use prior knowledge of cellular mechanisms, and add this prior knowledge as attributes to genes represented as nodes in a network. Then we can use graph algorithms to gain novel information about cancer driver genes. Genes that is not listed as cancer driver genes can then be assigned a status "guilt-by-association" if they have a network-relationship with proven cancer driver genes.

But all of this is based on single genes or proteins. What if we looked at whole networks as biomarkers? In this case, we will look at clusters of networks.

3.2 High-throughput sequencing

Today, rapid analysis of genes and proteins are made available through Next-Generation Sequencing (NGS)[4]. Networks offers us an informatic, algorithmic, visual and mathematical tool to study this bigger picture. This project will integrate the opportunity networks offer to discover new biomarkers in cancer. Through acquiring more data, faster than before, there now exists databases with large amounts of 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 cluster 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 [30].

3.3 Networks as a tool in cancer research

Viewing the cell as a network of proteins and genes presents us with several assumptions to make. A way of defining edges between nodes has to be established. It exists several ways of doing this, but it depends on the context.

Bayesian probability networks are one way of defining edges [26]. 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). 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 in this thesis has yet to come.

3.3.1 Pathways and networks

Pathways are small networks of well studied processes where many areas are well documented. Networks on the other hand are bigger and less explored, but when properly analyzed, it might provide new information unknown to pathway systems.

Analyzing pathways and networks have an advantage over individual genes. They may reveal information that comes from molecular events that covers multiple genes or pathways. Aggregating this data can increase the chance of detecting driver genes through statistical analysis. Also, information gathered through pathways and networks may be enriched through genomic, transcriptomic and proteomic data and create a more unified view of the tumor biology.

Chapter 4

Network clustering: Toward network based biomarker discovery

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. 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 [19]. There is several algorithms to create clusters, and all need to be researched in order to find the best suited one. The major differences on how these algorithms work is centered around how they handle nodes 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 [1]. But there is faster cluster-creating algorithms than those implemented in ClusterMaker2, an example is the SPICi [29] 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. They should be ranked in the order of being a cluster biomarker, more specificly described, a cluster biomarker for prostate cancer. The idea of clustering is to identify relationships between cancer driver genes and genes not related to cancer that reside within the same cluster. Since a cancer driver gene and a gene that has not been proved to be a cancer driver gene has a relation, the latter might play a role in cancer. These genes will be identified as candidate biomarkeres for prostate cancer. This was mentioned earlier as the "guilt-by-association"-principle.

Chapter 5

Ranklust: An implementation to rank clusters

5.1 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 rank the clusters in the network. Apps taking care of making the networks and creating clusters already exists, so Ranklust will only concentrate on the part that ranks the clusters and visualizes the results to the user of Cytoscape.

5.2 Programming language

The reason to choose Java above Python, Perl or other popular programming languages in bioinformatics is simply because of development experience. Java 8 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* [5] and sklearn[6], 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 Cyto-

scape 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 [7].

Formatting and cleaning the data that Ranklust will provide and export out of Cytoscape is done in Python 2.7[35]. As mentioned over, it has many small libraries that can help with the bioinformatics part, and the scripts needed to format the data will rarely exceed over 100 lines of code. Third-party libraries like pandas[32] and numpy[21] will support the formatting and cleaning of data.

5.3 OSGi and design patterns

Developing OSGi software promotes modularization [8] 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[33], 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 [9], but more of a collection of them, and it is the clean code principles by Robert C. Martin [10]. Going against the coding principles promoted by the Cytoscape app community and exclude the use of OSGi modules will not be done, but third party Java libraries that is not OSGi-ready will be used, namely the JUNG library[11].

Chapter 6

Databases

6.1 Databases for network information

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 each other. The initial database candidates in Ranklust are iRefIndex [12], GeneMania [13] and STRING [14]. 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. The open-source plugins in Cytoscape to communicate with the databases are iRefScape [15] for iRefIndex and GeneMANIA [16] for GeneMania.

6.2 Neo4J

A problem we had with the Neo4J[28] database is the dump it creates. The dump creates a single Cypher[27] commit for the whole database. This way, if anything goes wrong while importing the data, it will get rolled back. It hinders faulty data and relationship between several parts of the data to be imported into the database. The problem with a single commit to

import the data is the required memory. The personal computer used in this thesis and the computers at University of Oslo that we have access to did not have enough memory to import the database as a single commit in Cypher. Splitting up the dump to several commits does not work without quirks. It conquers the memory problem, but the way the relationships are created from the Neo4J dump requires the import process to be done in a single commit. A way to fix this is either not to use Cypher queries as a way of importing the data, but instead use GraphML[17]. Another way is to refactor the Cypher dump so that the creation of relationships does not need to be done in a single commit. We will use both ways, because not being able to use the Cypher queries because of too little memory will most likely be a problem for several people and not just me. Using GraphML is good for exporting data from Cytoscape and from the database, but interaction between the two of them while performing the algorithms is not feasible at the time. The last drawback is where Cypher queries are good, because it exists plugins that aid in using data from the database directly in Cytoscape. So initializing a database for testing could easily be done with GraphML, and then use Cypher queries to instruct a database to perform computations on the data.

Creating a script to refactor the Cypher queries does not only help in regards of using the data with Cytoscape, but also for everyone using a Neo4J Cypher dump to create a database and does not have much memory. A problem that might come up is performance. Because the current way of creating relationships use node labels in Neo4J and it takes only a single line to create a relationship between two nodes. If we refactor the relationship-part of the Neo4J dump, it will increase to a two instruction query. Firstly, two nodes has to be found based on their indexed node name, which will relate to the protein/gene name. Secondly, we do the same thing that was done before the refactor, we create the relationship between the two nodes. The difference between the old and the new way of creating the relationship is how the nodes are found. The Neo4J dump creates temporary node variables prepended with an underscore and an integer ID. These ID's only exist within the commit the nodes are created. So splitting creation and relationship-making leaves us with worthless IDs that does not refer to any known node. That is when Neo4J tries to be smart and then creates new nodes based on the underscore ID and then creates a relationship between them. The end result, two new nodes that only contains the undescore IDs. So the refactored version aims at matching the underscore IDs in the relationship creation to the node creation, and to replace them with the name of the nodes instead of the temporary underscore ID. But as mentioned, matching the nodes with the name takes up more time and performance might be an issue. For comparison, with a Cypher dump of 4,500 lines, it takes about 20 seconds before the query engine gets a stack overflow error from too little memory (4GB memory). With GraphML, an XML file with 250,000 lines takes under 1 second to import and it gets relationships right without any form of refactoring.

6.3 Other database technologies

Other databse technologies like SQL and NoSQL has not been researched in this thesis to be used with Cytoscape or graph algorithms regarding

Part III

Results

Chapter 7

Graph analysis

PPI network from irefweb. downloaded with these settings: irefweb said 109276 interactions after hgnc it was 43706 (perturbation: 60% from iref) after clustering it was 13183 (perturbation: 87,9% from iref - 69,8% from hgnc) only uniprot and refseq uniprot said

7.1 Creating a network

Converting from proteins to genes through HGNC data resulten in a 60% data network edge perturbation. From 100k ish links to 40k ish.

7.1.1 Creating connections

7.1.2 Adding weights

7.2 Ranking results

Talking about MCL results

Inflation	Clusters	Avg. cluster size	Max. cluster size	Min. cluster size	Modularity
1.6	1068	8.88	968	2	0.367
1.8	1400	6.60	660	2	0.307
2.0	1599	5.68	405	2	0.269
2.5	2053	4.20	179	2	0.223
3.0	2210	3.75	122	2	0.199

Table 7.1: MCL clustering parameter and statistic results

The modularity of the clustered networks gives an indicator of how well the process of creating the clusters went. Modularity is given as a score from 0 to 1. A score closer to 1 is more preferrable, as this indicates that the clusters created have a good degree of separation to the other clusters in the network. The preferred score to end up with would be around 0.8, but in this network there has been a good amount of perturbation through the protein-to-gene process. Modularity is not the only indicator of how well a network was clustered, hence the choice of not setting the inflation value in MCL to 1.6, but rather 1.8. When a lower inflation value is set, MCL does not separate edges between nodes as vigorously and as a direct cause, inflation will go up. Taking the other attributes in the table 7.1 into consideration, 1.8 seemed like the best inflation value. An inflation value of 1.8 has also been proved to be good for large high-throughput constructed protein-protein networks with a large amount of alterations[18].

7.3 Cross-validation

Executing a cross-validation on the iRefWeb with PRWP and MAA rankings had two purposes. The first being to prove the fact that every gene that had its prior score removed by the cross-validation, should be found in the results of the cluster ranking and identified as candidate biomarkers.

The second purpose was to analyze the distribution of the genes with prior scores removed by the cross-validation in terms of how they placed in the cluster ranking. The analysis of this distribution was done by dividing the amount of genes removed by cross-validation in a cluster by the total amount of genes i the same cluster. This operation was repeated for every cluster in the ranking, resulting in a distribution of the average amount of genes removed by cross-validation, that was detected by Ranklust together with post-processing of data, as cancer candidate biomarkers. A distinct descending distribution from high to low ranked clusters would indicate that most of the genes, with a prior score of their relevance to prostate cancer, was ranked in a way that achieved the goal of Ranklust; ranking clusters in biological network according to network structure and prior knowledge of relevance to diseases.

7.3.1 Cross-validation in PRWP

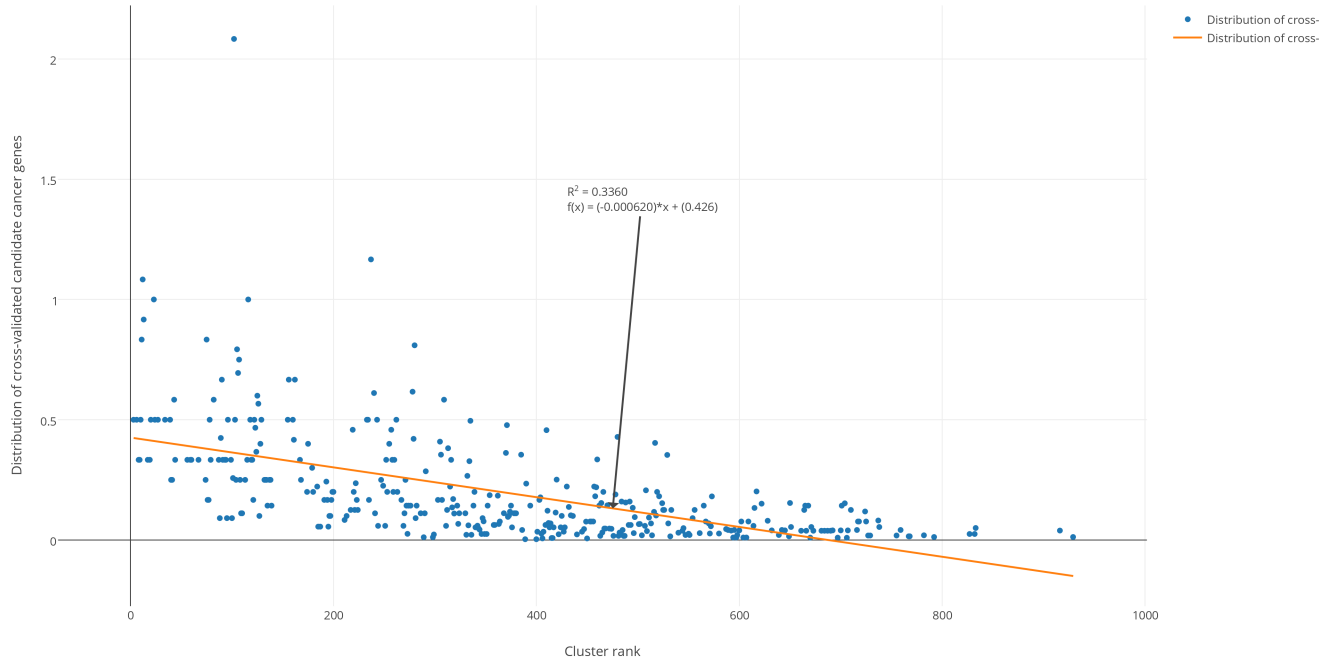

Figure 7.1: Cross-validation distribution in clusters (PRWP)

This plot is developed from the 10 random cross-validation runs ranked with PRWP. The results from this cross-validation shows that the higher the rank of the cluster, the more candidate cancer genes the cluster had.

7.3.2 Cross-validation in MAA

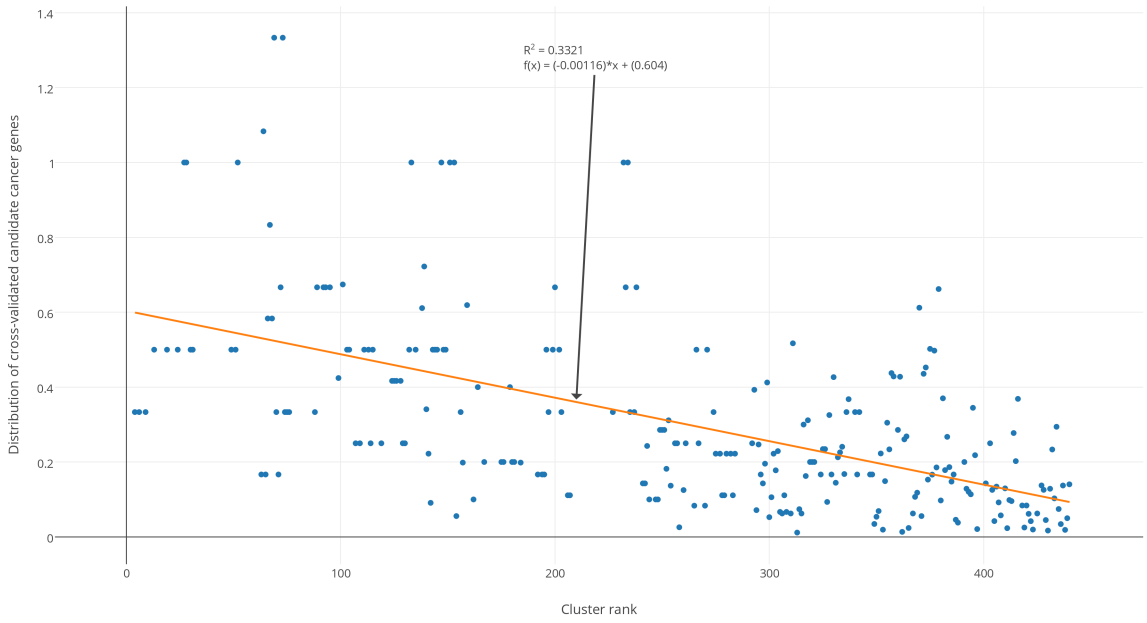

Figure 7.2: Cross-validation distribution in clusters (MAA)

This plot is developed from the 10 random cross-validatioon runs ranked with MAA. The results from this cross-validation shows that the higher the rank of the cluster, the more candidate cancer genes the cluster had.

7.4 Benchmarks

7.4.1 Ranked with PRWP

7.4.2 Ranked with MAA

7.5 Comparison to known biomarkers

7.6 Identification of possible biomarkers

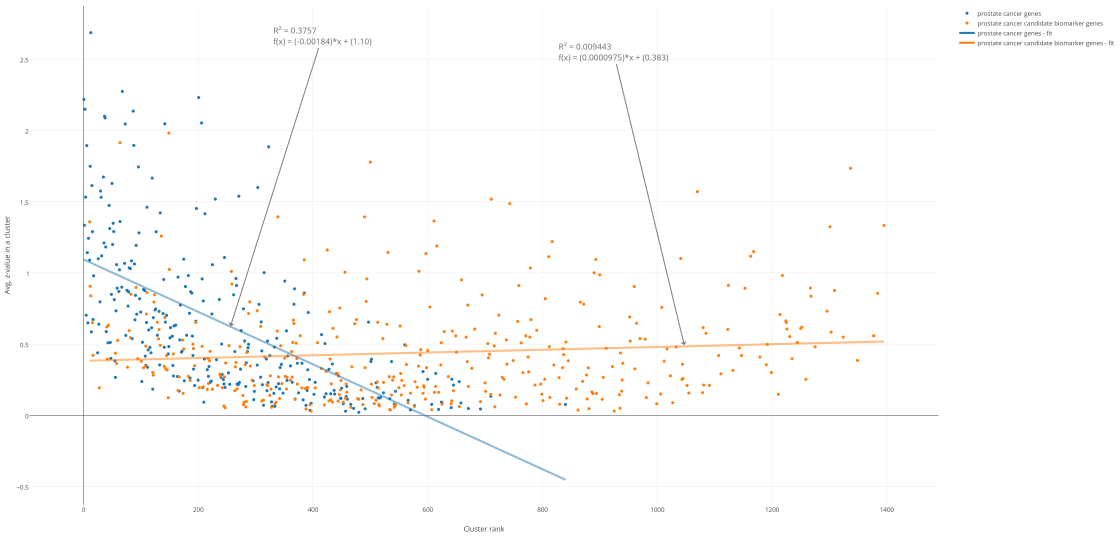

Figure 7.3: Average distribution of z-scores in clusters ranked by PRWP.

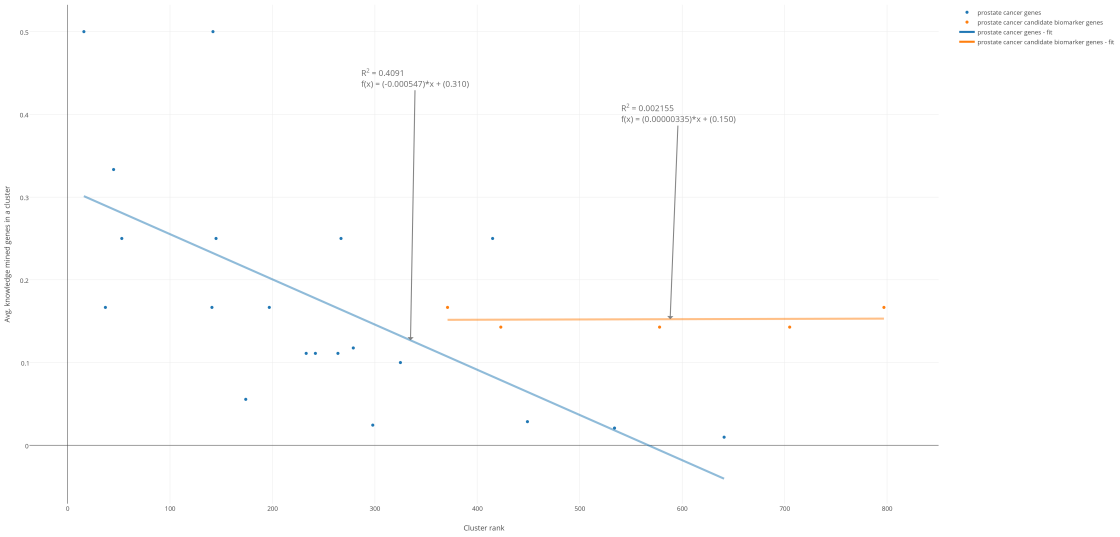

Figure 7.4: Average distribution of curated knowledge mined genes in clusters ranked by PRWP.

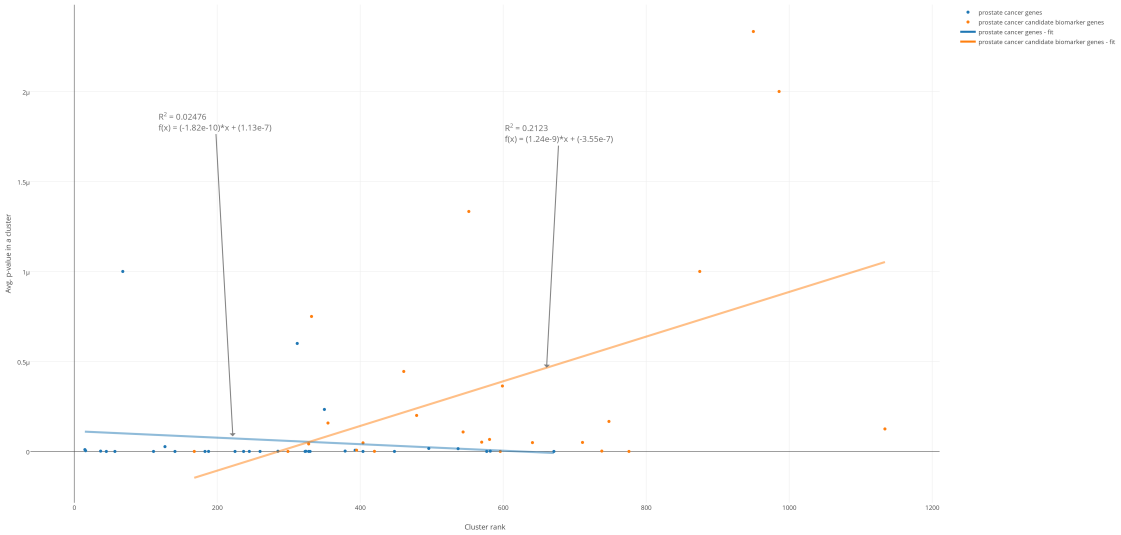

Figure 7.5: Average distribution of p-values in clusters ranked by PRWP.

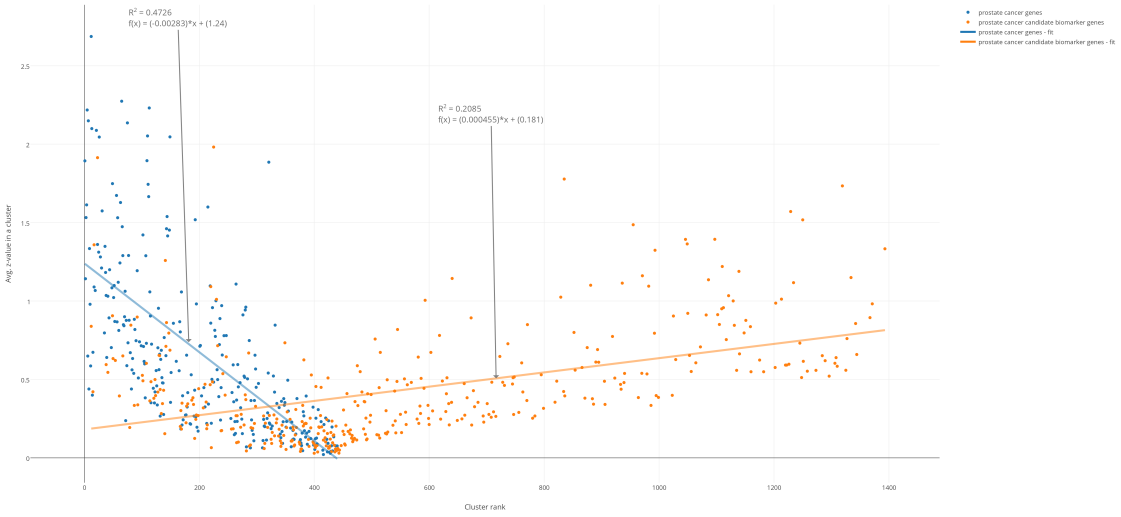

Figure 7.6: Average distribution of z-scores in clusters ranked by MAA.

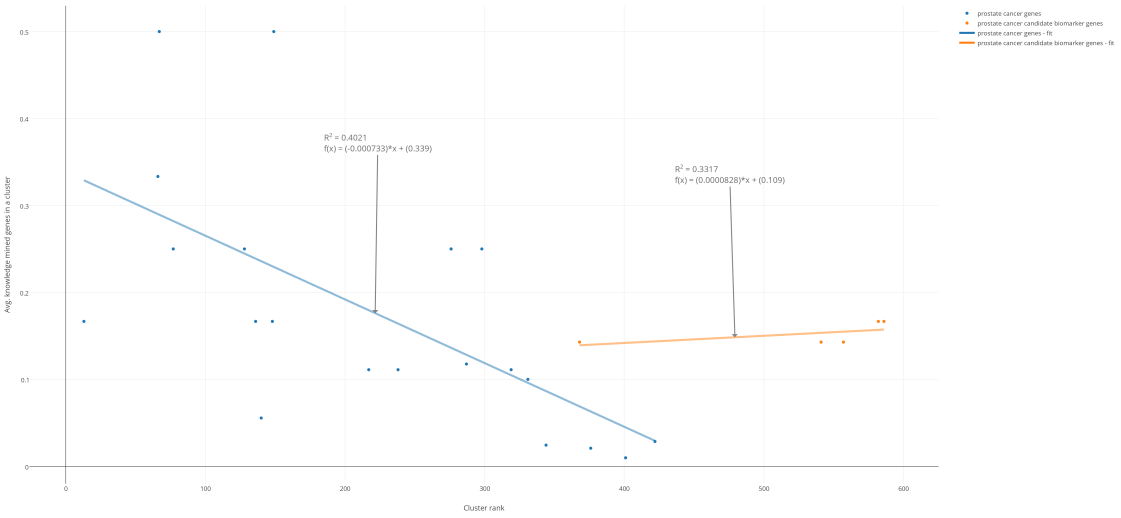

Figure 7.7: Average distribution of curated knowledge mined genes in clusters ranked by MAA.

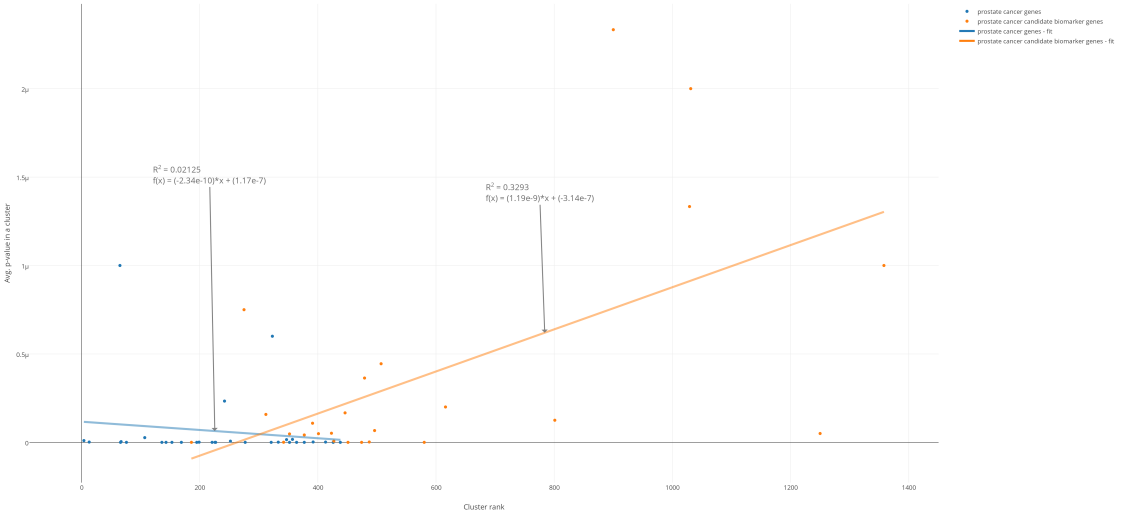

Figure 7.8: Average distribution of p-values in clusters ranked by MAA.

Glossary

MCL Markov Cluster algorithm used in Cytoscape to cluster the networks. 27, 28

Bibliography

[1] URL: http://apps.cytoscape.org/apps/clustermaker2 (visited on 22/05/2015).

[2] URL: https://github.com/RBVI/clusterMaker2 (visited on 08/07/2016).

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

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

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

[6] URL: http://scikit-learn.org/stable/ (visited on 08/07/2016).

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

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

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

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

[11] URL: http://jung.sourceforge.net/ (visited on 07/07/2016).

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

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

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

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

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

[17] URL: http://graphml.graphdrawing.org/ (visited on 19/02/2016).

[18] Sylvain Brohée and Jacques van Helden. 'Evaluation of cluster algorithms for protein-protein interaction networks'. In: *BMC Informatics* (2006). DOI: 10.1186/1471-2106-7-488.

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

[20] Pau Creixell, Jüri Reimand, Syed Haider, Guanming Wu, Tatsuhiro Shibata, Miguel Vazquez, Ville Mustonen, Abel Gonzalez-Perez, John Pearson, Chris Sander, Benjamin J Raphael, Debora S Marks, B F Francis Oulette, Alfonso Valencia, Gary D Bader, Paul C Boutros, Joshua M Stuart, Rune Linding, Nuria Lopez-Bigas and Lincoln D Stein. 'Pathway and network analysis of cancer genomes'. In: *Nature Methods* 12.7 (July 2015).

[21] Numpy Developers. *Numerical Python*. URL: http://www.numpy.org/ (visited on 24/07/2016).

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

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

[24] Biomarkers Definitions Working Group. 'Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework'. In: *Clinical Pharmacology & Therapeutics* 69.3 (18th Mar. 2001), pp. 89–95.

[25] 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).

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

[27] Neo Technology Inc. *Intro to Cypher*. URL: https://neo4j.com/developer/cypher-query-language/ (visited on 24/07/2016).

[28] Neo Technology Inc. *Neo4J*. URL: https://neo4j.com/ (visited on 24/07/2016).

[29] 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).

[30] 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).

[31] Mohsen Naghavi. 'The Global Burden of Cancer 2013'. In: *JAMA Oncology* (July 2015).

[32] NumFOCUS. *Python Data Analysis Libarary*. URL: http://pandas.pydata.org/ (visited on 24/07/2016).

[33] OpenJDK. *Project Jigsaw*. URL: http://openjdk.java.net/projects/jigsaw/ (visited on 24/07/2016).

[34] 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).

[35] Guido van Rossum. *Python*. URL: https://www.python.org/ (visited on 24/07/2016).

[36] Kyle Strimbu and Jorge A. Tavel. 'What are Biomarkers?' In: *Current Opinion in HIV and AIDS* (Nov. 2010).