Network Analysis on the Big Cancer Genome Data Set

Tai-Hsien Ou Yang and Kaiyi Zhu
Department of Electrical Engineering
Columbia University
New York, NY, USA
e-mail: {to2232, kz2232}@columbia.edu

***Abstract*—The genome profile of the cancerous tissue reveals the underlying pathological mechanisms as well as implies the possible treatments. To create a model for precise diagnosis and treatment recommendation for cancer patients, we analyzed the interrelationships of the outcome-associated genes using the Hadoop-based concordance index tool kit which we implemented for this project. With the identified genes, we implemented an treatment recommendation engine and created a web service for medical professionals. The resulting treatment suggestion model yielded an 87.12% true positive rate.**

Keywords- MapReduce;Cancer; Treatment; Concordance Index; Network Analysis; Web Service

I. INTRODUCTION

Cancer is one of the most lethal diseases. It is caused by genetic mutation and abnormal genomic interactions in the cells. [1] The abnormalities lead to the uncontrolled proliferation and invasivess. The general traits are proposed by Hanahan and Weinberg as Hallmarks of Cancer [2].

With the "Big Data of cancer", we are able to scrutinize the relationships between the gene expressions and the response to the treatment. Understanding their interrelationships may help us making treatment plan for complicated profiles. Therefore, to create a model for genomic precision treatment strategy suggestion, we assume the genomic features which are associated with the patient's outcome are related to the cancer hallmarks.

But one of the most common tools for biomedical data analysis was not implemented in the Apache Hadoop ecosystem, which is the family the concordance index. The concordance index is a non-parametric similarity metric for the censored data. It requires massive computation for counting the number of concordant pairs among all possible combinations between the two numbers in the input.

In this project, we identified the genes which are associated with the outcome of cancer patients by creating a toolkit based on the Apache Hadoop ecosystem. We further created the Bayesian network of the genes as well as validated that the genes involves in the cancer hallmarks and the performance of the model. Finally, we implemented a publicly available web service for cancer profile analysis and treatment suggestion based on the assumption that assume the optimal treatment strategy can be identified by sorting the outcomes among the patients with similar genomic profiles.

II. RELATED WORKS

Due to the inherent complexity of the cause, which is abnormal interactions between genes, currently cancers are classified into multiple subtypes [3] and numerous treatment strategies have been developed [2]. Though the products for specific cancer types were approved [4, 5], how to connect the cancer subtypes to the optimal treatment remains challenging [6].

However, the breakthrough of high-throughput genomic profiling technology makes massive cancer genome data profiling possible. One of the biggest Pan-Cancer data set is curated and maintained by The Cancer Genome Atlas (TCGA) [7]. The data set contains more than 18,000 genome profiles with comprehensive clinical records including survival information, treatment record, and clinical phenotypes.

III. SYTEM OVERVIEW

In the following sections, we elaborate on the data set we used and the methods we implemented for this project.

A. Materials

We used the genomic and the clinical profiles from the 3,316 patients from 12 cancer types in TCGA data set. We extracted the RNA-seq-based mRNA expression data (20,530 features), age at diagnosis, weight, lymph node number, and the survival information (follow-up time and status), and 347 drugs from each patient for our analysis. The profiling platform for the mRNA expression is the Illumina HiSeq 2000 sequencing system. The source code for preprocessing can be found in *Supplementary Information S1*. We also investigated the association between the outcome of 2,635 patients and the 23,371 methylation sites, which were profiled using the Illumina Infiniumn Human Methylation 27k array.

Since we transformed the data set into the common CSV format, our programs can take any RNA-seq or microarrary-based genomic data if needed.

B. Methods

The architecture of the project is shown in Figure 1. We performed analysis and modeling on the TCGA data set using the programs we created for Hbase, Hadoop, and R. With the selected features, we created the web service using PHP and R and visualized the results for the queries from users. The source code files and instructions of our final project are available in Appendix.

We implemented the algorithms and modules using Java, PHP, R, Pig, and JavaScript. The methods are described as the following. We evaluated the time performance on a virtual machine with Intel 32-bit CPU and 2GB memory. The operation system is Ubuntu 14.04 LTS.

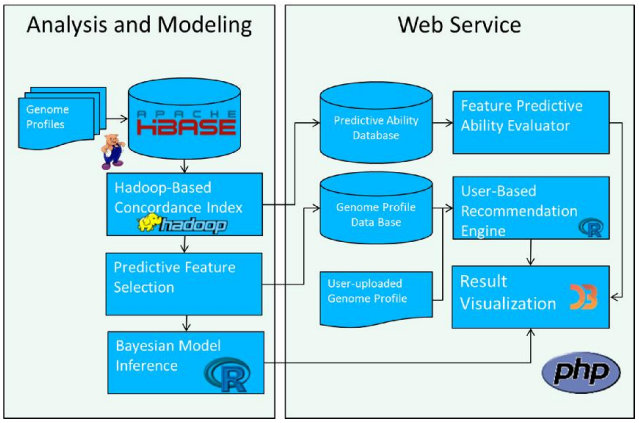


Figure 1. Archiecteure of the project.

IV. ALGORITHM

A. Map-Reduce Concordance Index Algorithm

We created the map-reduce concordance index function C by mapping all possible combination of the predictions into pairs and check if they are valid pairs and concordant with the response: If one pair's order is concordant to the response when patient with the shorter survival time is deceased, add 1 to n_c. Otherwise if the order is not concordant to the response, add 1 to n_d. In Kendall's τ, statuses of all subjects are always definitive. Then in the reducing part, the two groups of pairs are collected and the concordance index is computed as formula (1).

$$C = \frac{n_c}{n_c + n_d} \qquad (1)$$

Because the survival time is always integer and the status is binary (deceased and living/missing), we hashed the survival information into a floating number, then we sorted the survival information with regarded to the value of the feature-of-interest. Therefore, for each pair of numbers of the sorted input, the concordance checking subroutine requires only one iteration, which reduces not only the complexity on space but on time.

TABLE I. THE TWO CONCORDANCE INDEX ALGORITHMS

Algorithms	Interface of Mapper and Reducer	
	Key	Value
Algorithm A	Pair of two survivals	Score of a pair
Algorithm B	Gene Symbol	Concordance Index

We implemented two types of MapReduce–based concordance index program for validation. The difference is shown in Table I. The algorithm we described above is Algorithm A. We implemented Algorithm B for validation. In Algorithm B, the concordance index evaluation was done by the mapper instead of the reducer. The reducer only collects the final concordance index values. The customized writables and the sequence file conversion tool are imported from the pacakage available at http://vangjee.wordpress.com/2012/02/29/computing-pearson-correlation-using-hadoops-mapreduce-mr-paradigm/.

Using the concordance index algorithm, we identified 30 most survival-associated genes out of the 20,530 features on the TCGA data set.

B. Patient-Based Treatment Recommendation Engine

The treatment recommendation engine was implemented in R as the pseudo-code in Figure 2, in which u is the uploaded profile, V is the set of the profiles in the data set, R_v is the similarity between u and $v \in V$, and T is the covariate of treatments received by the patients. The most similar patients against the uploaded profile are found by sorting the similarity between the uploaded profile and the each patient in the data set. From the most similar patients, we identified the most applicable treatment strategies by sorting their survival time decreasingly.

The performance of the recommendation engine is evaluated using the Leave-One-Out cross validation strategy. The metric is the true positive rate of the predicted treatment against the treatment received by the held-out patient. If the predicted drug is among the treatment received by the patient, the prediction is true positive. We compared the performance of the engine using the Pearson correlation and Kendall's τ as the similarity in Step 3 of Algorithm C. Because of the superior performance, we used Kendall's τ similarity as the default similarity metric in the engine.

Algorithm C Treatment Recommendation Engine

1: procedure E(u)
2: for each $v \in V$ do
3: $R_v \leftarrow$ SIMILARITY(v,u)
4: end for
5: $R_{top} \leftarrow$ SORT(R) [1:20]
6: $T_{top} \leftarrow T$[INDEX (R_{top})]
7: $T_{sorted} \leftarrow$ SORT(T_{top}) BY SURVIAL(V[INDEX (R_{top})])
8: return T_{sorted}
9: end procedure

Figure 2. Pseudo-code for the treatment recommendation engine.

C. *Inference, Validation and Visualization of the Networks*

We inferred the Bayesian network of the 30 genes against the 3 using the *bnlearn* package in R. We used the default parameters. We visualize the inferred Bayesian Network using D3.js. The subnetworks of which the 30 genes enriched were validated using the publicly available gene ontology analysis tool Gene Ontology Consortium Enrichment Analysis (http://geneontology.org/).

D. *PHP-Based Interactive Web Service*

We created an interactive REST API for the pre-computed concordance index and the recommendation engine using PHP and R. The concordance index API returns a web page with the concordance index of the queried gene and the histogram of its expression in the data set. The recommendation engine returns a report presenting the applicable treatments with the longest survival. It also returns a line plot for verifying the associations between the profiles. The facilities are accessible at http://128.59.65.192/

E. *HBase-Hadoop Interface*

Because PIG and the Hbase shell do not provide the function for creating the schema for large matrix, we implemented a Java class to import the genomic data with 20,530 columns. For exporting the records, we utilized the PIG function to access the Hbase storage and creating the formatted output.

V. SOFTWARE PACKAGE DESCRIPTION

The package we provided contains the toolkit for performing two types of MapReduce-based concordance index evaluation in Hadoop environment and the HBase tools for data importing and exporting, as well as the R script files for preprocessing, feature analysis and Bayesian inference.

For the treatment engine, we include the source code of the engine and the implementation of the entire web service (Fig. 3). The package is available at Github repository https://github.com/Sapphirine/Network-Analysis-on-the-Big-Cancer-Genome-Data The detailed instruction and screenshots of those five parts of our package is available in *Appendix 1*.

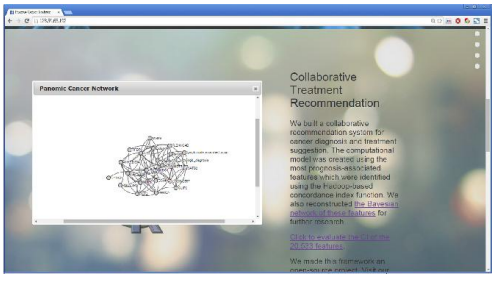


Figure 3. The screenshot of the web service

VI. EXPERIMENTAL RESULTS

A. *The Toolkit for Biomedical Survial Data Analysis*

We implemented a tool kit consists of the concordance index function and the corresponding preprocessing, importing and exporting tools. As described in the Materials and Methods, by hashing and sorting the input, we reduced the input size by 66.67% since the input of the algorithm is a floating array instead of a three-column matrix. However, this method does not allow ties. The scores used for identifying the features were evaluated using the traditional method.

We identified 30 outcome-related genes (Table II). Algorithm A and B returned the same results. It takes 72 seconds to evaluate one concordance index for 3,316 patients for Algorithm A and 3.06 seconds for algorithm B. The reason why Algorithm A is much slower than Algorithm B is that we ran Hadoop in pseudo-distributed mode. If the performance is gauged on a real distributed Hadoop system, Algorithm A should be faster than Algorithm B.

We performed the concordance index evaluation on the methylation site profiles to explore the possible epigenetic mechanisms. But the most outcome-associated methylation site, cg16414852 (CI=0.6569) did not outperformed the 30th mRNA gene expression. Therefore, we decided not to include it in further analysis.

TABLE II. **TABLE OF THE MOST OUTCOME-ASSOCIATED GENES**

Gene Symbols
SCGB2A2, CYP4Z2P, C6orf127, ANKRD30A, VTCN1, PIP, CYP4Z1, BNIPL, IRX5, ESR1, TFDP3, CRABP2, CALML5, LOC642587, ZYG11A, PCP2, PIAS3, GATA3 CLIP2, KRT7, FLJ45983, C20orf114, SIX4, MSMB, PRKCA, PLEKHG4B, SATB2, HKDC1, PGLYRP2, METTL7B

B. *Patient-Based Treatment Recommendation Engine*

The true positive rate of the recommendation engine is over 87% when Kendall's τ is used as the similarity metric for the patient similarity (Table III). Therefore we implemented it in the web service. We tried to implement the recommendation engine in Neo4j, but the shell returned java heap error matter how large the heap size is. Since Mahout recommendation engine does not provide Kendall's τ similarity, we did not compare the result with the result generated by Mahout.

TABLE III. PERFORMANCE OF THE RECOMMENDATION ENGINE

Performance	**Similarity Metric**	
	Pearson	***Kendall's τ***
True Positive Rate	80.34%	87.12%

C. *Bayesian Inference, Validation and Visualization of the networks*

We reconstructed the Bayesian network for the 30 genes and the three phenotypes as presented in Figure 4, which was visualized using D3.js.

The network enrichment analysis shows the genes we identified involve in the networks related to three cancer hallmarks (Fig. 5), which are tissue development, embryonic morphogenesis, and interleukin-4 secretion. This result validates the assumption that the outcome-associated genes involve in the pathways of the cancer hallmarks.

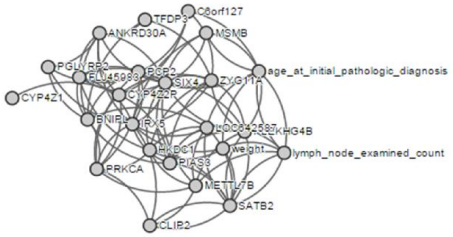


Figure 4. The inferred Bayesian network of the 30 genes and the three clinical phenotypes.

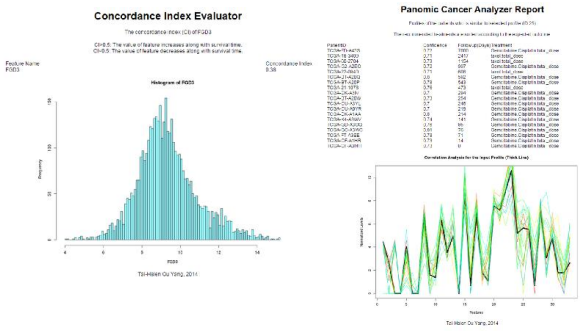
Term	Background frequency	Sample frequency	Expected	+/-	P-value
positive regulation of cytokine biosynthesis (GO:0001817)	4	2	4.596e-03	+	3.636e-04
regulation of cytokine biosynthesis (GO:0001817)	4	2	4.596e-03	+	3.636e-04
interleukin-4 secretion (GO:0072582)	5	2	5.732e-03	+	5.974e-04
interleukin-4 production (GO:0032633)	7	2	8.029e-03	+	1.189e-03
interferon-gamma secretion (GO:0072610)	8	2	9.172e-03	+	1.529e-03
gonad development (GO:0008406)	169	4	2.202e-01	+	2.811e-03
development of primary sexual characteristics (GO:0045137)	202	4	2.376e-01	+	4.224e-03
uterus development (GO:0060065)	16	2	1.720e-02	+	5.339e-04
embryonic skeletal system morphogenesis (GO:0048706)	93	3	1.066e-01	+	8.321e-03
pharyngeal system development (GO:0060037)	17	2	1.949e-02	+	6.346e-04
tissue development (GO:0009888)	1496	8	1.715e+00	+	6.977e-03
sex differentiation (GO:0007548)	214	4	2.512e-01	+	7.278e-03
interferon-gamma production (GO:0032609)	18	2	2.064e-02	+	7.672e-04
embryonic organ morphogenesis (GO:0048562)	275	4	3.152e-01	+	9.839e-03

Figure 5. The subnetworks in which the 30 genes are enriched.

D. *Interactive Web Service*

The web service provides the access to the interactive Bayesian network of the outcome-associated genes, the concordance index evaluator and the recommendation engine (Fig. 6). The two services can be queried as REST APIs. Since the concordance indices of the 20,530 genes were pre-computed and stored as a database, the CI of all features can be retrieved via the web services. The histogram of the expression of the gene is also visualized in the returned page.

The treatment recommendation engine allows users to upload the profile which complies with the format given on the web page. Even the recommendation engine requires scanning across the entire data set, the service still responds fast (1.08 seconds per profile). The returned report presents the top 20 suggested treatments. The line plot presents the profiles of the patients who received the treatments against the uploaded profile for further verification. The web service also provides all patients in the data set as the demonstrative queries.



(a) (b)

Figure 6. The screenshots of the returned results of (a) the concordance index evaluator and (b) the recommendation engine.

E. *HBase-Hadoop Interface*

We implemented the HBase-Hadoop interface using PIG, the importing time is 0.06 seconds per gene and the exporting time is 11.10 seconds per gene.

In the future, we aim to incorporate more genome profiles and platforms into our analysis pipeline to improve the accuracy. On the technical part, we attempt to incorporate Apache Thrift to connect Hadoop, HBase, and the web service for streaming and dynamic model training.

VII. CONCLUSION

By implementing the Hadoop-based concordance index tool kit, we identified a set of outcome-associated genes using a data set collected from multiple cancer types. We validated that the set of genes involve in multiple subnetworks representing the hallmarks of cancer as well as inferring and visualizing their Bayesian regulation network.

Based on the outcome-associated gene set, we designed a user-based collaborative recommendation engine and made it a web service for precise genome-based cancer diagnosis and treatment recommendation as well as evaluating the performance of it.

Finally, we implemented a publicly available web service of concordance index evaluation and treatment recommendation. Since there was no diagnosis and treatment planning tools were created on the Pan-Cancer basis. The research starting from this project may not only shed a light on the molecular pathologies of the complicated malignancies and the potential therapeutic regimens but provide more powerful tool for medical professionals to save lives.

VIII. CONTRIBUTIONS

T.-H Ou Yang contributed the idea of the algorithms and implemented the toolkit, the web service and performed the analysis, as well as the presentation and the report. K. Zhu created the interface between Hadoop and HBase and the movie clip.

APPENDIX

Please find the appendix file *Appendix 1* in the package for the detailed instructions and the descriptions about the package.

ACKNOWLEDGMENT

We acknowledge Professor Ching-Yung Lin at Columbia University for the opportunity for us to accomplish this project and to present it in the workshop. We acknowledge TCGA for the data set which we used in this project and Dr. Gee Vang for the example and the package of the customized Hadoop writables.

REFERENCES

[1] American Cancer Society, "Cancer Facts & Figures 2014," Atlanta: American Cancer Society, 2014, pp. 1-2.

[2] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," Cell, Mar. 2011, Vol. 144, pp. 646-674, doi:10.1016/j.cell.2011.02.013.

[3] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, and J. P. Mesirov, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, Vol. 286, Oct 1999, pp. 531-537, doi: 10.1126/science.286.5439.531.

[4] L. J. van't Veer, H. Dai, M. J. Van De Vijver , Y. D. He, A. A. Hart, M. Mao, "Gene expression profiling predicts clinical outcome of breast cancer," Nature, Jan. 2002, Vol. 415, pp. 530-536, doi:10.1038/415530a.

[5] M. J. Van De Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, "A gene-expression signature as a predictor of survival in breast cancer," New Eng. J. Med., Dec. 2002, Vol. 347, pp. 1999-2009, doi: 10.1056/NEJMoa021967.

[6] N. I. Simonds, M. J. Khoury, S. D. Schully, K. Armstrong, W. F. Cohn, D. A. Fenstermacher, "Comparative effectiveness research in cancer genomics and precision medicine: current landscape and future prospects," J. Nat. Cancer Insti., Mar. 2013, djt108, doi: 10.1093/jnci/djt108

[7] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, "The cancer genome atlas pan-cancer analysis project," Nature genetics, Sept. 2013, Vol. 45, pp. 1113-1120, doi:10.1038/ng.2764.