COLUMBIA UNIVERSITY

E6893 Big Data Analytics:

Network Analysis on the Big Cancer Genome Data

Team Members*:* **Tai-Hsien Ou Yang and Kaiyi Zhu**

Dec 09, 2014

Motivation

Cancer as the King of Diseases

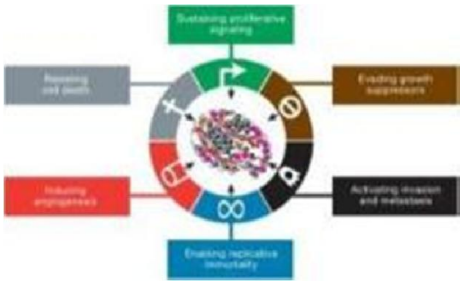
- Uncontrolled growth and spread of abnormal cells caused by genetic mutation and abnormal genomic interactions.
- In 2013, 580,350 people died of cancer in the U.S. (American Cancer Society)

Cancer is to complicated to cure

- >1,000 genes may involve in cancer development.
- Thousands of subtypes and treatment strategies.

But we still have a chance

- Different subtypes of cancer present different genomic profiles but share some common traits: *Hallmarks of Cancers*.
- To get the profile, common next-generation sequencing technology reveals a patient's genomic profile of 20,000 genes for $500 in a week.
- But the medical professionals do not have the tools for analysis because of the complexity.



D Hanahan, RA Weinberg, Cell, 2011

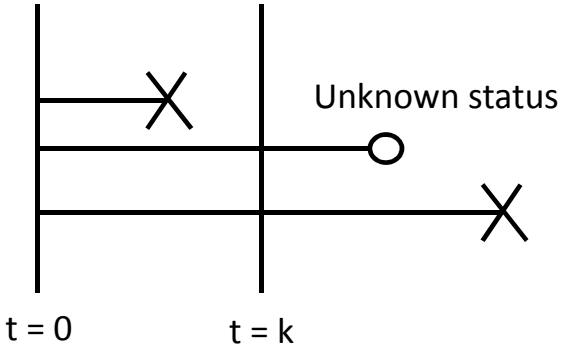
Our Proposal

Our Assumptions

- We assume the genomic features which are associated with the patient's outcome are related to the cancer hallmarks.
- Understanding their interrelationships may help us making treatment plan for complicated profiles.
- We assume the optimal treatment strategy can be identified by sorting the outcomes among the patients with similar genomic profiles.

Our Proposal

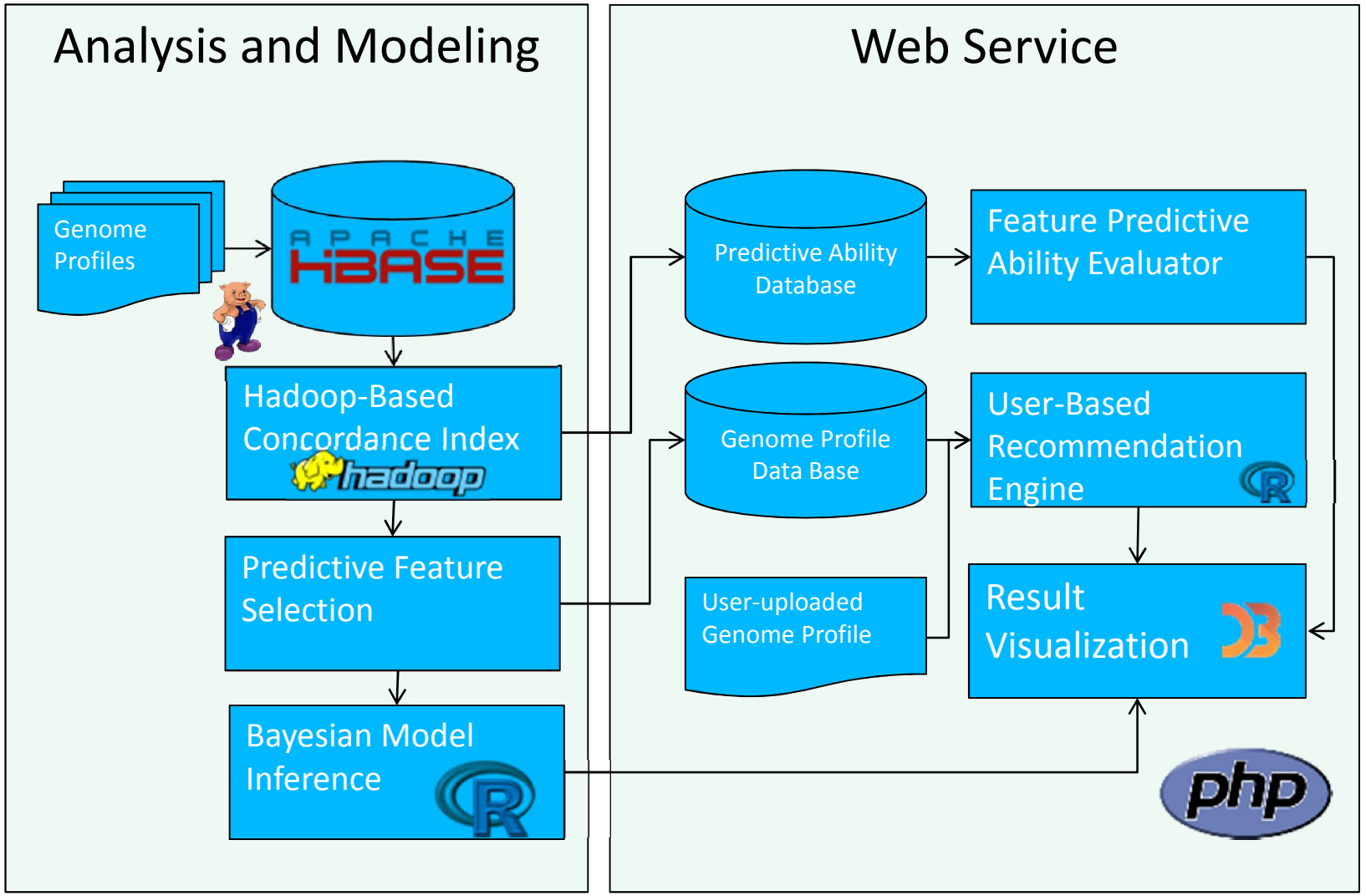
- To identify the genomic-clinical network using the "Big Data of Cancer"
  - The Cancer Genome Atlas: 3,316 patients with 20,530 features from 12 cancer types.
  - Build the toolkit for analyzing *censored biological data* using Hadoop.
- To build a patient-based predictive model for treatment suggestion.
- To build a friendly web service for cancer diagnosis and treatment suggestion.



√ Concordance Index

X Pearson Correlation

System Architecture



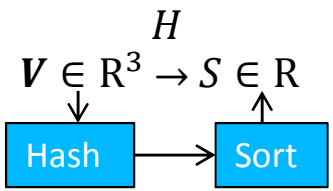
Contributions (1/3)

Hadoop-Based CI and Kendall's τ

- Non-parametric similarity metric for censored data.
- Check if a pair of numbers' order is concordant to the order of which in the response.

$$\tau = \frac{2(n_c - n_d)}{\sqrt{(n(n-1) - \sum_i t_i(t_i-1))(n(n-1) - \sum_j t_j(t_j-1))}}$$

- We hashed the input to reduce >66% data transmission.
- Hbase connection for memory efficiency.
- We identified 30 outcome-related genes.



Patient-Based Recommendation System

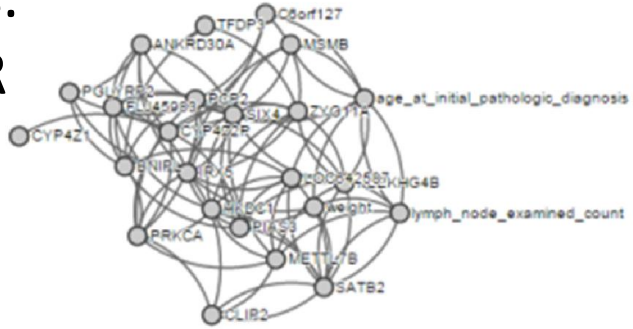
- Neo4j has Java heap problem and licensing problem.
- Hadoop-Mahout is too bulky for real-time analysis.
- We created our own one with web-API.
- LOOCV true positive rate (TPR) of recommendation=87.12%.

Index	Pearson Correlation	Kendall's τ
TPR	80.34%	87.12%

Contributions (2/3)

Bayesian Network

- On the top 30 outcome-associated genes.
- Inferred using the “bnlearn” package in R
- Visualized using D3.js.



PHP-Based Web Facility

- Diagnosis and treatment recommendation for the user-uploaded profile
- Fast response (1.0752 sec/profile)
- Interactive CI evaluation API to see if a gene is related to the outcome.

Contributions (3/3)

Network Enrichment Analysis

- The ontology network analysis shows the genes we identified involve in the networks related to cancer hallmarks:
  - Tissue development: Proliferation
  - Embryonic organ morphogenesis: Invasion
  - Interleukin-4 secretion : Evading immune system

Gene Ontology Consortium

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

Term	Background frequency	Sample frequency	Expected	+/-	P-value
positive regulation of ureteric bud formation (GO:0072107)	4	2	4.586e-03	+	3.826e-04
regulation of ureteric bud formation (GO:0072106)	4	2	4.586e-03	+	3.826e-04
interleukin-4 secretion (GO:0072602)	5	2	5.733e-03	+	5.974e-04
interleukin-4 production (GO:0032633)	7	2	8.026e-03	+	1.169e-03
interferon-gamma secretion (GO:0072643)	8	2	9.173e-03	+	1.526e-03
gonad development (GO:0008406)	199	4	2.282e-01	+	2.861e-03
development of primary sexual characteristics (GO:0045137)	202	4	2.316e-01	+	3.030e-03
uterus development (GO:0060065)	15	2	1.720e-02	+	5.339e-03
embryonic skeletal system morphogenesis (GO:0048704)	93	3	1.066e-01	+	6.321e-03
pharyngeal system development (GO:0060037)	17	2	1.949e-02	+	6.848e-03
tissue development (GO:0009888)	1496	8	1.715e+00	+	6.977e-03
sex differentiation (GO:0007548)	254	4	2.912e-01	+	7.278e-03
interferon-gamma production (GO:0032609)	18	2	2.064e-02	+	7.672e-03
embryonic organ morphogenesis (GO:0048562)	275	4	3.153e-01	+	9.839e-03

http://geneontology.org/

Demo

http://128.59.65.192/
http://128.59.65.250/html/

Conclusion

COLUMBIA UNIVERSITY

What we achieved

- A genomic-clinical regulatory network reconstructed from 3,316 patients. We also validated that the genes represent the Cancer Hallmarks.
- A web service for precise genome-based cancer diagnosis and treatment recommendation.
- User-based collaborative recommendation engine using R.
- A Hadoop-based toolkit for computational biomedical research.
- Performance evaluation for treatment suggestion engine.
- Efficient biomedical data storage using Hbase.

Further work

- Incorporate more genome profiles for analysis.
- Eliminate the PIG-based Hbase interface.
- Hadoop streaming/Apache Thrift.

Thank you!

Questions are welcome.