

Cancer systems biology using high-performance computing

Tampere University of Technology
Department of Signal Processing

Matti Nykter
Antti Ylipää
Virpi Kivinen
Timo Erkkilä

December 2010

Laboratory of Complex Biosystems Modeling

- Principal investigator: Matti Nykter, D.Sc
- 4 post-docs, 4 graduate students and 4 undergraduates
- We study various complex biosystems using systems biology based methods and modeling approaches in combination with high-throughput genomics data
- In practice, we strive to develop generic mathematical models and computational tools, and then apply them to answer specific biological questions
- Our main application area is currently cancer systems biology, but we also study immunology and cell differentiation



Additional information: <http://csb.cs.tut.fi/>



TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Cancer systems biology

- Cancer systems biology combines empirical, mathematical and computational techniques to gain understanding of different aspects of the disease
- In our research, we take advantage of the data from the Cancer Genome Atlas (TCGA) project through our collaborators, Prof. Wei Zhang at M. D. Anderson Cancer Center and Prof. Ilya Shmulevich at the Institute for Systems Biology
- Profs. Shmulevich and Zhang have a joint Genome Data Analysis Center grant for conducting high-level analyses of the TCGA data
- Prof. Zhang is also a Professor in our group (funded by Finnish Funding Agency for Technology and Innovation Finland Distinguished Professor program – Tekes FiDiPro)
- In Prof. Zhang's project we develop new computational methods to gain understanding of prostate cancer, in particular



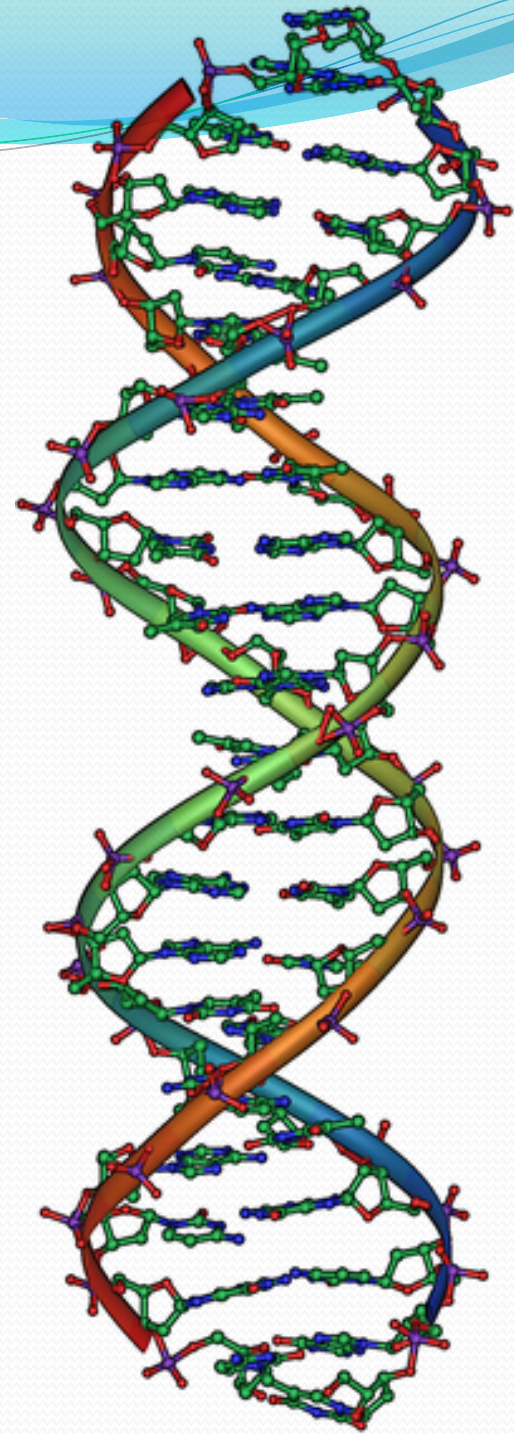
Prof. Wei Zhang



Prof. Ilya Shmulevich

The genomic challenge

- Human genome contains over 20,000 genes and for the most part we still do not know their role in cells
- The genes produce hundreds of thousands of different molecules
 - The grand challenge of "the post-genomic era" is to find out what these different genetic components actually do
- Given the vast amount of genes, this is an astounding challenge, both experimentally and computationally
- New experimental projects that produce terabytes of data, such as TCGA project, are undertaken constantly
- The main problem now is not the lack of data but the inefficiency of the current computational means to analyze them

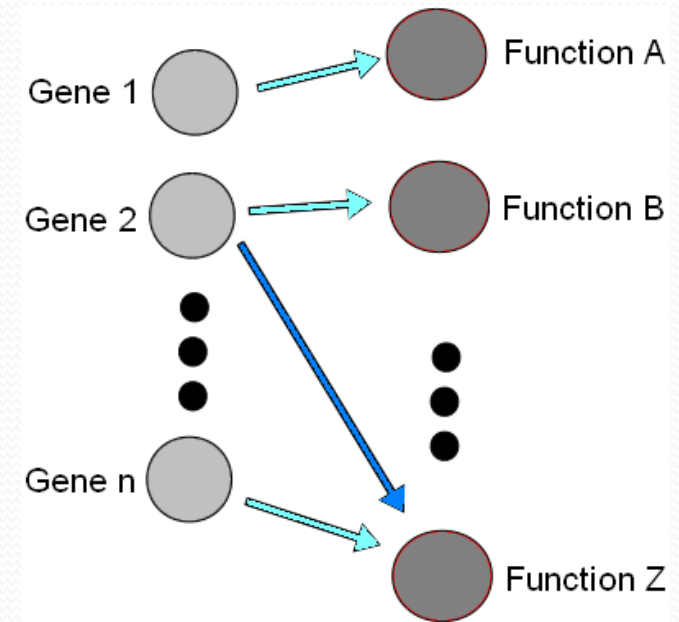


Gene-function association

- Objective: Combine gene expression measurement data from tumors with statistical models to infer the associations of genes with biological functions
- What we need:
 - 1) massive amounts of high-throughput data from TCGA project and others
 - 2) gene sets describing biological processes
 - 3) enough computing resources to identify the associations
- What we do:
 - 1) For each tumor sample and each gene set, we compute a gene set enrichment score
 - Statistical significance analysis of enrichment scores requires computationally intensive permutation testing
 - Fortunately, each test is independent, and thus parallelizable
 - 2) Compute the association between genes and gene sets
 - A parallelizable problem, too

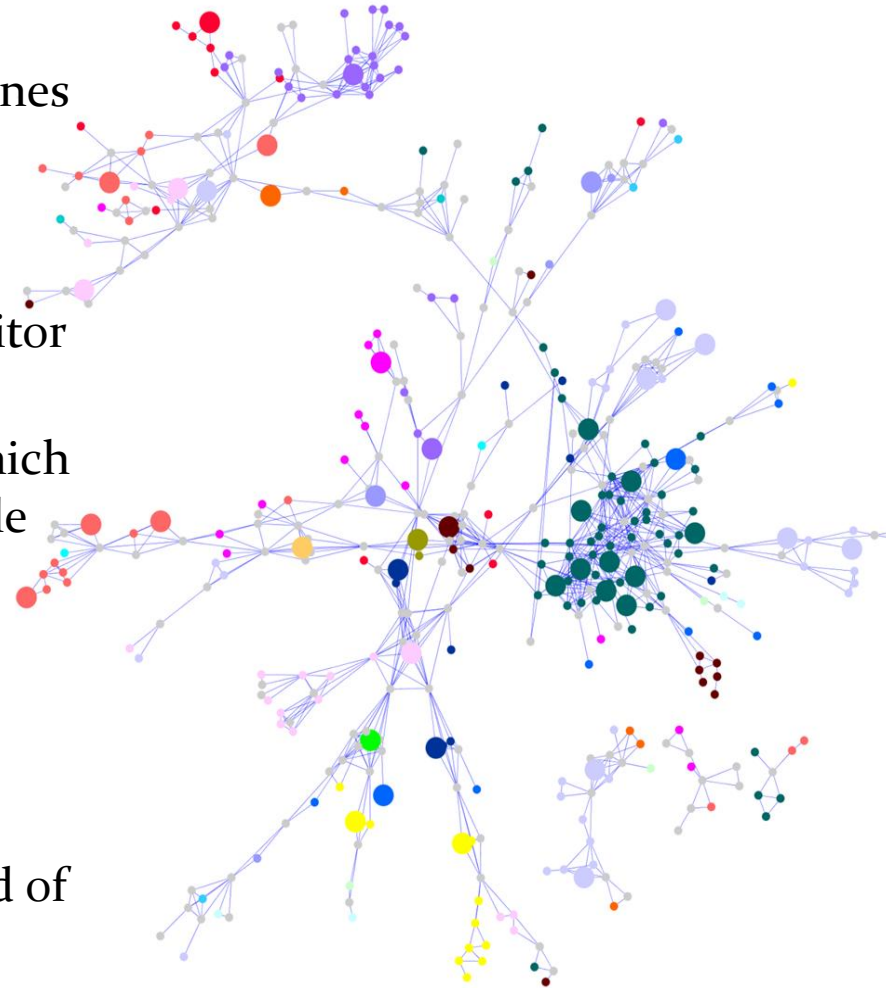
Gene-function association

- Currently, computing using ca. 300 tumor samples and 200 gene sets takes years of CPU time
- Parallelization to TUT's idle desktop computers via Techila's system reduces computing times to few days
- Future: We will take thousands of tumor measurements and gene sets and predict functions for all the molecules and mutations
 - It takes tens of years of CPU time, but we will only have to compute it once
- This is possible only using a high level of computation parallelization which we have implemented already



Indicator protein selection

- Gene-function associations tell us which genes work in which cellular process
- Measuring cellular processes through quantifying certain proteins from the bloodstream helps us to diagnose and monitor diseases
- Our aim is to find a small set of proteins which are indicative of as many diseases as possible
- We apply genetic algorithms to the gene-function associations to search for the best proteins
- Parallelizing the computationally intensive optimization algorithm enables us to find a profitable solution within few hours instead of months



User views

- Highly distributed computing enables us to try out completely new kinds of strategies when we are no longer limited by computing time as much as before
- Many standard methods that we use profit from the increase in accuracy, e.g. significance estimation and optimization
- Many of our new algorithms could not be developed or run at all without Techila's systems
- Algorithm development speed increases compared to other computing platforms as the Techila system is easy to use and fully integrated in the development environments we commonly use, mainly Matlab



Antti Ylipää, M.Sc.



Matti Nykter, D.Sc.



Virpi Kivinen, M.Sc.



Timo Erkkilä, M.Sc.