Automating Biomedical Research Through Machine Learning

Robert F Murphy
Ray & Stephanie Lane Professor of Computational Biology and Professor of Biological Sciences, Biomedical Engineering and Machine Learning
Affiliated Senior Fellow, Freiburg Institute for Advanced Studies
Honorary Professor, Faculty of Biology, University of Freiburg, Germany

RAY AND STEPHANIE LANE Center for Computational Biology
Carnegie Mellon
Why Comp Bio?

Personal Genomics Revolution

Personal Genome Company 23andMe Receives Google Investment
by Dr. Hsien-Hsien Lei
Posted May 22, 2007 in DNA Products, DNA in General

Email this • Stumble It! • Digg This! (10 Diggs, 1 comment) • Discuss on NewsVine

Google has invested $3.9 million in 23andMe along with Genentech and two venture capital firms, New Enterprise Associates and Mohr Davidow Ventures. 23andMe is a personal genome company that wants to help consumers get in tune with their genes. Check out the blurb on the newly revamped front page:

Even though your body contains

http://www.eyeondna.com/2007/05/22/personal-genome-company-23andme-receives-google-investment

Wednesday, January 11, 2012

Ion Torrent claims to be first with $1K genome sequencer

By Lori Valigra

Ion Torrent Systems Inc. of Guilford, Conn., showed off its new DNA sequencing machine and chip that it claims can map a human genome in 24 hours for a cost of $1,000 at the J.P. Morgan Healthcare Conference in San Francisco Tuesday, according a news release from parent company Life Technologies Corp.

The same day, Illumina Inc. of San Diego also announced a 24-day sequencer, but at a much higher initial machine price tag of $740,000. The price to sequence a genome was not disclosed.

Ion Torrent last summer wrote in a Nature paper that it had developed a DNA sequencing technique to target the $1,000 genome industry goal. At the time the paper was published, Ion Torrent CEO and founder Jonathan Rothberg, told Mass High Tech, "Sequencing on an ion semiconductor chip makes the $1,000 genome both inevitable and predictable. Extrapolating from our current progress we will break the $1,000 genome barrier in 2013." The company appears to be ahead of schedule.

http://www.bizjournals.com/boston/blog/mass-high-tech/2012/01/ion-torrent-claims-to-be-first-with-1k-genome.html
Why Comp Bio?
Cancer Diagnosis and Treatment

Agendia Gets FDA Clearance for MammaPrint on Agilent Systems
February 23, 2011

NEW YORK (GenomeWeb News) – Agendia has received another clearance for the firm's MammaPrint breast cancer recurrence assay.

The new clearance is for running the test on two additional Agilent microarray scanners and two Agilent Bioanalyzers. Agendia said the additional clearance will expand laboratory capacity to handle the increasing number of MammaPrint, TargetPrint, and BluePrint test results.

Previously, the MammaPrint test was only allowed to be performed at Agendia's lab in Irvine, Calif. The new clearance enables the test to be run in a central laboratory and allows Agendia to also perform the test at its lab in Amsterdam, where the firm is based.

MammaPrint was the first in vitro diagnostic multivariate index assay device approved by the FDA in early 2007.

Why Comp Bio?

Develop better drugs

http://www.projectsmagazine.eu.com/opinion/ahead_in_the_cloud_computational_drug_discovery_reaches_for_the_next_level
Why Comp Bio?

Solve the energy crisis with microorganisms

Environmental Capital

Weekly Journal


JULY 14, 2009, 11:14 AM ET

Biofuels Bonanza: Exxon, Venter to Team up on Algae

By Keith Johnson

Exxon’s $600 million deal with Synthetic Genomics to brew fuel from algae could mark a coming of age for alternative fuels.

As human genome mapper Craig Venter, now the chief executive of SGI said, turning algae into biofuel simply won’t happen without Big Oil’s deep pockets. He said it was the biggest program he knows of worldwide to produce biofuels.

At the same time, both Exxon and SGI were at pains to stress that the “collaborative research project” is in its early phases. “There’s no guarantee of success,” said Exxon’s vice-president for R&D, Emil Jacobs. “We’re giving a new reality to the timetables” of algae biofuel development, Mr. Venter said.
Why Comp Bio?

Get a great job:

**Future Forecast: 10 Hot Job Markets in 2012**

In our information-rich society there is an ever-increasing demand for workers in the fields of computers, health care, science and space technology—much of it driven by the demands of the retiring baby boomers. If you like to plan ahead, here is sampling of some of the jobs that will be hot in the next several years and beyond.

1) **Organic Food Industry**

By 2010, organic food and beverage will represent about 10 percent of the total market—a tenfold increase from 1998. Bob Scowcroft, executive director of the Organic Farming Research Foundation says the industry will soon need more organic food producers, certification experts, retailers and scientists as organic becomes mainstream.

Qualifications: Organic food expertise in farming, business or science.

Salary range: $50,000 to $80,000

2) **Computational Biology**

There is a growing need to combine computer science, biology and math to make sense of research data in massive quantities, says Leroy Hood, co-founder of the Institute for Systems Biology. This field may eventually allow physicians to test for a patient's unique genetic markers and tailor the best treatments and medicine for that patient.

Qualifications: A bachelor's degree or higher in bioinformatics, computer science, mathematics.


**Jobs in Computational Biology: Career Options and Requirements**

The study of biology using computer software and mathematics is called computational biology. Jobs in the field typically require a high level of integrated education in biology, mathematics, and computer science. A variety of organizations employ people in this application-oriented field.

**Career Options**

This multidisciplinary specialty has a number of practical and research applications. Computational biologists analyze large volumes of information and devise computer modeling simulations for academic research and health applications for businesses such as biotechnology and pharmaceutical companies, as well as for government health and research institutions.

**Career Options in Academia**

Computational biologists in academia can work with other theoretical and applied researchers to devise models, simulations, and predictions for molecular biological systems and interactions. They can also analyze large quantities of data related to genetics and genomics. Academic researchers typically teach courses in their field and supervise student research projects in addition to conducting their own research.

**Academic Career Requirements**

CAREERS in academia commonly require a graduate degree in computational biology. To obtain a faculty position at a university, a Ph.D. is recommended, though a master’s degree is typically sufficient to teach at a 2-year college. In addition, working in a postdoctoral position for several years after graduation is usually necessary before a permanent faculty or research position can be obtained.

**Career Options in Commercial Industries**

Pharmaceutical companies, scientific software companies and biotechnology companies all employ computational biologists in research and development. Modeling drug interactions, developing analytical software for use by biologists and analyzing large data sets for biotech companies that provide genetic

http://education-portal.com/articles/Jobs_in_Computational_Biology_Career_Options_and_Requirements.html
Examples of Computational Biology Problems

• Drug Development
• Cell Modeling
Observation, Modeling and Computation

• Traditional Approach (Scientific Method): Make observations, construct model/theory, make new predictions, test them
  – Good news
    • Computers can help build models
  – Bad news
    • Difficult/impossible to prove biological theories directly
    • Lots of possible theories and predictions to test!
Problem – Cell/Developmental Biology

- Tens of thousands of molecules within cells can change on time scales from below seconds to months during differentiation or disease processes
- Can undergo changes in expression, interaction, localization
- Too many combinations to measure all…
Problem – Drug Development

• Drugs fail late in development because of unanticipated side effects – only real solution is to choose drugs with desired effect on target and no undesired effect

• This requires determining the effects of millions of potential drugs on tens of thousands of potential targets

• Exhaustive experimentation too costly…
Solution?

- Represent problems as a matrix of possible experiments
- Build model to predict full matrix from whatever data we have
- Use active learning to choose new experiments and iteratively improve the model
Traditional drug development

A controlled experiment that measures the effect of a compound on a target’s activity

Establish Assay for a Target

Typically, a protein whose activity is believed to cause a disease
Traditional drug development

Chemicals that are potential drugs or might be modified to become drugs

Compounds

Screen \sim 10^7\text{ Compounds}

Negative  Positive
Two problems:
(1) We don’t know effects on other targets
Two problems:
(1) We don’t know effects on other targets

Comprehensive screening for one target does not reveal side effects
Two problems:
(1) We don’t know effects on other targets
(2) We have learned nothing for the next target
but we cannot afford to exhaustively perform every experiment

Much better if we could learn the entire matrix…

<table>
<thead>
<tr>
<th>Compounds/perturbagens</th>
<th>Targets/phenotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Intermediate</td>
<td></td>
</tr>
</tbody>
</table>
Solution: just do some experiments…

Compounds/perturbagens

Targets/phenotypes

- Negative
- Positive
- Intermediate
Solution: just do some experiments... and predict the rest
Active Learning Pipeline

- Efficiently learn accurate predictive model

Model Generation

Experiment Execution

Experiment Selection

Pick experiments most likely to improve model
Active Learning Assessment

- **Accuracy or Hits (%)**
- **Experimental Space Explored (%)**

Lines:
- Red: Active Learning
- Green: Random

Graph demonstrates the comparison between Active Learning and Random methods in terms of accuracy or hits against the experimental space explored.
Active Learning Assessment

Accuracy or Hits (%) vs. Experimental Space Explored (%)

- **Active Learning**
- **Random**
Active Learning Assessment

Model Improvement using Active Learning with constrained budget

Accuracy or Hits (%)

Experimental Space Explored (%)
Savings from using active learning rather than some other method.
Test with PubChem Data

- Assays: 177
  - 108 *in vitro*
  - 69 *in vivo*
  - Sign of score modified to reflect type of assay (inhibition or activation)
- Unique Protein Targets: 133
- Compounds: 20,000
- Experiments: ~1,000,000 (30% coverage)
- Compare discovery rate across different methods
  - Discovery: a drug-protein pair whose |rank score| > 80
Active Learning
Optimized $\mu$-QSAR
Randomized Search
Active Learning
Optimized $\mu$-QSAR
Randomized Search

With only 2.5% of the matrix covered, we can identify 57% of the active compounds!
BUILDING CELL MODELS FROM IMAGES
Cell Organization

• How do we learn and represent
  – the sizes and shapes of different cell types
  – the number, sizes, shapes, positions of subcellular structures
  – the distribution of proteins across those structures?
  – how these change in the presence of perturbagens?
Fluorescence microscopy

- Primary method used to determine the subcellular location of a protein is to "tag" it with a fluorescent probe
Subcellular Location
Subcellular Location
Subcellular Location
Subcellular Location

- mitochondrion
- centriole
- vesicle
- cytosol
- rough endoplasmic reticulum
- ribosomes
- smooth endoplasmic reticulum
- microtubules
- chromatin
- nuclear envelope
- nuclear pore
- nucleolus
- Golgi complex
- lysosome
- flagellum
- plasma membrane
Subcellular Location
Subcellular Location
Can recognition of subcellular location be automated?

- Problem is hard because different cells have different shapes, sizes, orientations
- Organelles not found in fixed locations
Supervised Machine Learning Approach

• Design features to describe subcellular patterns
• Use examples of images of different subcellular patterns to train classifier
Subcellular Pattern Classification: Computer vs. Human

Boland et al 1997; Murphy et al 2000; Boland & Murphy 2001; Murphy et al 2003; Huang & Murphy 2004
Subcellular Pattern Classification: Computer vs. Human

2D Images of HeLa cells

Chebira et al 2007; Nanni et al 2010
However...

• Assigning words not sufficient
  – Knowing that apples and oranges can be distinguished by their color does not allow you to understand how either are formed
Solution: Generative Modeling

• Human cognition

  Learn

  Mental model

  Write

  Generated examples

• Generative model

  Training images

  Statistical generative model

  Generated image
http://CellOrganizer.org

May 17, 2013: Version 1.9.0 released!

New: Now allows synthesis of cell and nuclear shape instances for Hela cells using a diffeomorphic model.

The CellOrganizer project provides tools for

- learning generative models of cell organization directly from images
- storing and retrieving those models in XML files
- synthesizing cell images (or other representations) from one or more models

Model learning captures variation among cells in a collection of images. Images used for model learning and instances synthesized from models can be two- or three-dimensional static images or movies.

CellOrganizer can learn models of

- cell shape
- nuclear shape
- chromatin texture
- vesicular organelle size, shape and position
- microtubule distribution.

These models can be conditional upon each other. For example, for a given synthesized cell instance, organelle position is dependent upon the cell and nuclear shape of that instance.

Cell types for which generative models for at least some organelles have been built include human HeLa cells, mouse NIH 3T3 cells, and Arabidopsis protoplasts. Planned projects include mouse T lymphocytes and rat PC12 cells.

Support for CellOrganizer has been provided by grants GM075205 and GM090033 from the National Institute of General Medical Sciences, grants MCB1121919 and MCB1121793 from the U.S. National Science Foundation, by a Forschungspreis from the Alexander von Humboldt Foundation, and by the School of Life Sciences of the Freiburg Institute for Advanced Studies.
What does CellOrganizer Do?

• Given a population of cells we can model
  – Distribution of nuclear and cell shapes
  – Distribution of vesicular patterns
    • Number, shape, intensity, location of objects
  – Distribution of microtubules
    • Number, stiffness, location of centrosome
  – Relationships between these models

• Given models we can
  – Synthesize new images
  – Model populations that change over time/condition
Conclusions

- Active learning provides solution to infeasibility of performing all possible perturbation and differentiation experiments
- Tools beginning to be available to build image-derived generative models
  - Learn the underlying cell “model” from which individual cell images are drawn
Active Learning

Josh Kangas  Armaghan Naik

Chris Langmead
CellOrganizer

**Project Leaders**
- Robert F. Murphy
- Gustavo Rohde

**Major Collaborators**
- Klaus Palme
- Hauke Busch
- Christoph Wülfing
- Melanie Boerries
- Jörn Dengiel
- Ivo Sbalzarini

**Current Team Members**
- Devin Sullivan
- Gregory R. Johnson
- Taraz Buck
- Ivan Cao-Berg

**Past Contributors**
- Ting Zhao
- Tao Peng
- Wei Wang
- Aabid Sharif
- Joshua Kangas
- Jianwei Zhang
- Alexander Dovzhenko
- Rüdiger Trojok
- Jieyue Li
- Baek Hwan Cho

**Support:**
- NIH GM075205, GM090033, GM103712
- NSF MCB1121919