Computational Challenges for Systems Biology and Personalized Medicine

Nathan D. Price

Frontiers of Extreme Computing

Oct 23, 2007
“The Human Genome Project has catalyzed two paradigm changes in contemporary biology and medicine—systems biology and predictive, preventive and personalized medicine. “ – Lee Hood

Acknowledgement
Lee Hood
Institute for Systems Biology
Systems Biology

High-Throughput Data Generation (Genome Sequencing, DNA Arrays, Proteomics, Metabolomics)

Bioinformatics (Computation, Engineering, Systems Science, Modeling, Simulation)

Data-rich Data-poor

Systems biology challenge: Utilize high-throughput technologies to drive large scale biological and medical discovery
The Exponential Rise in Computation Power and Biological Data: The Fuel of the Biomedical Revolution

- The harnessing of the (exponentially increasing) biological data through the use of (exponentially increasing) computational capabilities will revolutionize biology and medicine.

- For given problems, one or the other may be limiting, and this can change in the future depending on how the problem constraints grow in each dimension, and the rate of growth of each.
Computational Challenges in Systems Biology

- How to fully decipher the (digital) information content of the genome
- How to do all-vs-all comparisons of 1000s of genomes (or more)
- How to extract protein and gene regulatory networks from 1 & 2
- How to integrate multiple high-throughput data types dependably
- How to visualize & explore large-scale, multi-dimensional data
- How to convert static network maps into dynamic mathematical models
- How to predict protein function \textit{ab initio}
- How to identify signatures for cellular states (e.g. healthy vs. diseased)
- How to build hierarchical models across multiple scales of time & space
- How to reduce complex multi-dimensional models to underlying principles
- Text searching to bring the literature and experimental data together

Scales in Biology with Computational Opportunities

- Protein folding
- Detailed molecular simulations
- \textit{Cell-scale simulation}
- Cell-cell interactions
- Microbial communities
- Ecosystem
Integrating Heterogeneous Biological Data through Cell-Scale Computational Models

Whole cell simulation represents an exascale computing challenge
Biochemical Reaction Networks

<table>
<thead>
<tr>
<th>Phylogenetic Data</th>
<th>Physiological Data</th>
</tr>
</thead>
</table>

Statistical Influence Networks

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Literature</th>
<th>Genome Annotation</th>
<th>Interactomics</th>
</tr>
</thead>
</table>

More detail (biochemistry, etc.)

Reaction Stoichiometry

<table>
<thead>
<tr>
<th>Reaction Stoichiometry Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enz</td>
</tr>
<tr>
<td>Rxn. 1</td>
</tr>
<tr>
<td>Rxn. 2</td>
</tr>
<tr>
<td>Rxn. 3</td>
</tr>
</tbody>
</table>

Application of Constraints

Constraint-Based Model

<table>
<thead>
<tr>
<th>Constraint-Based Model Equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>S · v = 0</td>
</tr>
<tr>
<td>v ≤ v̅max</td>
</tr>
</tbody>
</table>

Network Inference

Statistical Influence Network

<table>
<thead>
<tr>
<th>Statistical Influence Network Equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>C = f(A, B, D)</td>
</tr>
</tbody>
</table>

Two types of biological network models

**Biochemical Reaction Networks**

1) Directly mechanistic
2) Require significant knowledge of the system
3) Broadly applicable where biochemistry is known
4) Laws of physics and chemistry can be directly applied
5) Relate more closely to phenotype (i.e. fluxes)
6) Once reconstructed from biochemical data, network not likely to change (other than additional reactions)

**Statistical Influence Networks**

1) Not generally mechanistic
2) Can be applied without needing prior knowledge (although can be incorporated)
3) Broadly applicable without knowledge of biochemistry
4) Physico-chemical laws typically not applicable/applied
5) Relate more directly to high-throughput data (i.e. transcriptomes)
6) Additional data can lead to significant network rewiring

From Component to Systemic Annotation of Genomes

Palsson, *Nature Biotechnology*, 2004
Mathematical Representation of a Biochemical Network

With the network represented in matrix form, the tools of linear algebra, linear programming, and convex analysis are available.
The Systems Biology Process

The role of reconstruction

1. Components
   Plurality of omics

2. Reconstruction
   "Systemic annotation"
   one set of reactions arising from the genome

3. *in silico* modeling
   plurality of methods

4. Hypothesis generation and testing
   -CHiP-Chip
   -Fluxomics

---

Simulation
Experiment

phenotypic space
"practically infinite" for most organisms

gen- transcript- prote- metabol-...etc.
Computational challenges for biochemical network reconstruction

- Automated reconstructions of cellular networks in detail from genomes (e.g. Rick Steven’s SEED)
- Can it be extended to forming computable networks?
- Can experimental data (e.g. knockout lethality data) be used as input as well
- Are their limits to the algorithmic nature of the problem?
- Is automated text mining a necessary component? (probably)
- Given inputs of genome, biochemical characterizations, knockout lethality, probabilistic inference, can a predictive computable network be generated? (Seems the answer should be YES)
- Also, can probabilistic metabolic networks be generated and shown to be useful (similar in ideas to Shmulevich’s Probabilistic Boolean Networks)
- Reconstruction of dynamic models (would require sequence to kinetics computational capabilities)
Computational challenges for statistical inference networks

- Network inference at the genome-scale
  - Currently, this is done by bi-clustering before network inference can be done
  - Largest models currently work with hundreds of variables, but what is needed is thousands to tens of thousands
  - Is an interplay between computational and data limitations
- Enabling computation of genome-scale networks with feedback
  - For example, current models predict gene expression given TF expression
- Integrating heterogeneous data types
  - E.g. linking proteome and transcriptome

Beyond Genome-Scale Networks to Whole Cell Simulators

- Adding spatial component to models will require greatly increased computational power (and data generation)
- Simulation of physical properties of molecular machines

Image by David Goodsell
COMPUTATIONAL CHALLENGES FOR PERSONALIZED MEDICINE
Systems View of Disease

- Diseases are the result of one or more perturbed biomolecular networks.
- These perturbations lead to differences in the abundance of biomolecules (e.g. mRNA, proteins, metabolites).
- These changes can then be measured and used for molecular diagnostics of disease.
- Thus, reporters of the state of these networks are available, if we can learn to read the signals.
Foundational Data for Personalized and Predictive Medicine

- **Individual Genome Sequences**
  - Two people sequenced so far
    - James Watson: Co-discoverer of structure of DNA
    - J. Craig Venter - One of pioneers leading human genome sequencing
  - Will provide personalized information for each patient
  - Will provide probabilistic health futures

- **Blood Diagnostics**
  - 1000s of individual measurements
  - Patterns will be used to read out a current health state
  - Taken repeatedly over the lifetime of a patient
    - Thus, patient can act as their own 'control' when changes from previous norms emerge
  - Represents the interplay between the digital information of the genome and the interactions with the environment throughout the lifetime of the patient
Example Types of Biological Information

DNA → messenger RNA → PROTEIN → SYSTEMS → HUMAN
Computational Challenges of Systems Medicine

- The $1000 genome will mean that eventually we can have genome sequences for at least hundreds of thousands of individuals.
- Diseases, such as cancer, are multi-genic and influenced by a variety of environmental factors.
- Thus, we will need to mine this data to identify diagnostic patterns that will guide therapy choice.
- Computational challenges associated with the high-combinatorics of personalized medicine are enormous.
Computational Challenges for Personalized Medicine Diagnostics

Thus, we can imagine that to identify powerful diagnostics, we will need to:

- Link genome sequence data to patient records and blood diagnostic measurements (current and over time)
- Search for patterns requiring mutations in multiple points in the genome, requiring a number of environmental factors (hopefully linked through previous studies to the blood) etc.

There are 6 Billion bases in the human genome, each of which could, in theory, be involved in disease (the actual number may be much lower, though)

Most diseases are multi-geneic – current thought is that cancers results from ~6 mutations that are different in each case – even within what are considered the same cancer – and that symptoms can be non-existant from subsets of these mutations.

The combinatorics of these problems leads quickly into very large search space where full enumeration will not be possible – and thus advanced computation with efficient approximate search strategies will be needed.
Tracking perturbed networks through molecular signals in blood

By learning to track network perturbations through blood measurements, the blood will become a novel means for studying human biology and medicine, including:

- Normal physiology
- Virtually all diseases
- Development

However, we are not currently at the point we can do this in higher organisms.

To develop this vision in full will require

- Models of each cell type in body
- Models of secretion patterns in vivo
- Computational separation of signal from multiple sources in the blood
Moving Beyond Diagnostics to Therapy Discovery

- To truly understand and be able to control multi-geneic diseases, we must be able to model the effects of these mutations on systems.
- This requires personalized computational models for personalized medicine.
- Eventually, we would like to have human disease simulators that, given a genome and then supplemented over time with additional information (e.g. blood diagnostics) from interaction with the environment, could predict therapies and provide the possibility to actually fix disease states.
- There is a huge information barrier (in addition to experimental) that will require huge amounts of computation to realize.
Conclusions

- Systems Biology presents a myriad of computational challenges that must be met, including simulation at the level of molecules, pathways, cells, communities, and ecosystems.
- Today, I focused primarily on challenges for simulating genome-scale networks in cells.
- The future of systems medicine clearly will require large investments in increasing computational power in order to harness the amount of information that will be available for each patient—VASTLY greater than it is today.
Contact Information

University of Illinois

Price Lab

James Eddy
Swati Gupta
Caroline Milne
Ravali Raju
Jay Sung
Joe Dolivo
Luke Edelman
Seth Hanson
Azeem Husain

Lab Website

www.igb.uiuc.edu/labs/price/home.html

Email

ndprice@uiuc.edu

Further Reading

Challenges for genome-scale biochemical network simulations

- Constraint-based approaches only ones currently used at the genome-scale
- Optimization methods (e.g. flux balance analysis) computationally inexpensive
- Optimization-based re-engineering of networks is feasible for small numbers of permutations, but not large - so are areas of large-scale optimization of re-engineering problems that could benefit from ‘exascale’ resources

- Genome-scale dynamic models not yet feasible
- Genome-scale stochastic, single-cell models even further away
Moving into the Blood

- Differentiating multiple diseases
  - Organ specific molecular signatures
  - Combining multiple signatures
  - Broader studies for differentiation

- Stratification vs. Progression
  - Different diseases
  - Different stages

\[ \begin{array}{ccccccc}
A & > & F & \ldots & G & > & L & \ldots & C & > & H \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1
\end{array} \]

Readout: Cancer 1, 2, 3 ...
Stage 1, 2, 3, ...

Molecular State

Time
Acknowledgments

**University of Illinois**

**Price Lab**

James Eddy
Swati Gupta
Caroline Milne
Ravali Raju
Jay Sung
Joe Dolivo
Luke Edelman
Seth Hanson
Azeem Husain

**Institute for Systems Biology**

Lee Hood, M.D., Ph.D.
Ilya Shmulevich, Ph.D.

**M.D. Anderson Cancer Center**

Wei Zhang, Ph.D.
Jonathan Trent, M.D.
David Cogdell

**Lab Website**

[www.igb.uiuc.edu/labs/price/home.html](http://www.igb.uiuc.edu/labs/price/home.html)
Disease Arises from Disease Perturbed Networks

Non-Diseased

Diseased

dynamics of pathophysiology

diagnosis

therapy

prevention
Biology dictates what new technology should be developed; technology opens new frontiers in biology for exploration.
Computational Challenges: moving to the exascale

- Systems Biology
- Personalized Medicine

Individualized genome sequences for most people in U.S. and other developed nations (moving into rest of world eventually as well)

So, billions of sequences
  - How to read the signals

Blood diagnostic
Dynamics of a Prion Perturbed Network

2 wks

12 wks

20 wks

18 wks

Clinical Signs

No Clinical Signs

University of Illinois at Urbana-Champaign

INSTITUTE FOR GENOMIC BIOLOGY
University of Illinois at Urbana-Champaign
Innovations for the Next 20 Years in Biology and Medicine:

- Systems approaches to biology and medicine will dominate biological sciences in the 21st century—global science networks and strategic partnerships.

- Systems approaches will pioneer new opportunities in agriculture, bio-energy, biology, bio-remediation, health, nutrition, an understanding of human development, neurobiology and better educational strategies.

- Systems approaches will move medicine from its current reactive mode to predictive, preventive, personalized and participatory (P4 medicine) modes—with a focus on wellness.

- The digitalization of biology and medicine will constitute a far greater revolution than the digitalization of information technologies.

- P4 medicine and the digitalization of medicine will enable health care to become cheap and easily executed. Therefore exportable throughout the globe including to the developing world.

- Strategic partnerships and international networks in science will allow us to attack big scientific problems.

Enormous economic opportunities in biology
Outline

- Molecular signature classifiers
- Relative expression reversals
- Diagnosis similar cancers requiring very different treatments
  - Gastrointestinal stromal tumor
  - Leiomyosarcoma
- Survival prognosis marker
Molecular signature classifiers

- The goals of molecular classification of tumors:
  - Identify subpopulations of cancer
  - Inform choice of therapy
- Generally, a set of microarray experiments is used with
  - ~100 patient samples
  - ~ $10^4$ transcripts (genes)
- This very small number of samples relative to the number of transcripts is a key issue
  - Feature selection & model selection
- Also, the platform of microarray used can have a significant effect on results
Estimating Error on Future Cases

- **Methodology**
  - Best case: have an independent test set
- **Resampling techniques**
  - Use cross validation to estimate accuracy on future cases
  - Feature selection and model selection must be within loop to avoid overly optimistic estimates

Resampling: Shuffled repeatedly into training and test sets.

Training Set → Test Set

Data Set

Average performance on test set provides estimate for behavior on future cases

Can be MUCH different than behavior on training set
Relative Expression Reversal Classifiers: The Top Scoring Pair Approach

- Find a classification rule as follows:
  - IF gene A > gene B THEN class1, ELSE class2
- The classifier is chosen finding the most accurate and robust rule of this type from all possible pairs in the dataset
- If needed a set of classifiers of the above form can be used, with final classification resulting from a majority vote (k-TSP)

Tan et al., Bioinformatics, 21:3896-904, 2005
Rationale for k-TSP

- Based on concept of relative expression reversals

Advantages
- Does not require data normalization
- Does not require population-wide cutoffs or weighting functions
- Has reported accuracies in literature comparable to SVMs, PAM, other state-of-the-art classification methods
- Results in classifiers that are easy to implement
- Designed to avoid overfitting
  - \( n = \text{number of genes}, \ m = \text{number of samples} \)
  - For the example I will show, this equation yields:
  - \( \binom{n}{2} \ll 2^m \)
  - \( 10^9 \ll 10^{20} \)
Diagnosing Types of Sarcoma

GIST and LMS

- Gastrointestinal Stromal Tumor (GIST) and Leiomyosarcoma (LMS)
  - Morphologically very similar and thus hard to correctly distinguish using current methods
  - Have different treatments, so correct diagnoses is critical
    - Can be a life or death decision

- Key differences
  - Gleevec efficacy
    - GIST: 50-80% effectiveness in completely eradicating the cancer
    - LMS: negligible effects
  - Expression of c-kit protein
    - Almost never expressed in LMS
    - Mixed for GIST - but majority have c-kit expression
    - Heterogeneous within the tumor
C9orf65 expression

OBSCN expression

Clinicopathological Diagnosis

X – GIST
O - LMS

Accuracy on data = 99% Predicted accuracy on future data (LOOCV) = 98%

RT-PCR Classification Results

100% Accuracy
- 19 Independent Samples
- 20 samples from microarray study
- including previously indeterminate case

Comparative accuracies of c-kit expression and the 2-gene classifier

Kit Protein Staining of GIST-LMS

Top Row – GIST Positive Staining

Bottom Row – GIST negative staining

Blue arrows - GIST   Red arrows - LMS

Accuracy as a classifier ~ 87%.

TSP and survival of GIST patients

K-means Clusters

Survival Fraction

Time (years)

High OBSCN/PRUNE2
Low OBSCN/PRUNE2

0 2 4 6 8 10

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

DEPARTMENT OF CHEMICAL AND BIOMOLECULAR ENGINEERING
OBSCN Expression & Survival

K-means Clusters

Survival Fraction

Time (years)

High OBSCN2

Low OBSCN

No real predictive capacity
TSP methods perform comparably to best machine learning techniques on multiple cancers

### Table 3. LOOCV accuracy of classifiers for binary class expression datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>Leukemia</th>
<th>CNS</th>
<th>DLBCL</th>
<th>Colon</th>
<th>Prostate1</th>
<th>Prostate2</th>
<th>Prostate3</th>
<th>Lung</th>
<th>GCM</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSP</td>
<td>93.80</td>
<td>77.90</td>
<td><strong>98.10</strong></td>
<td>91.10</td>
<td><strong>95.10</strong></td>
<td>67.60</td>
<td>97.00</td>
<td>98.30</td>
<td>75.40</td>
<td>88.26</td>
</tr>
<tr>
<td>$k$-TSP</td>
<td>95.83</td>
<td><strong>97.10</strong></td>
<td>97.40</td>
<td>90.30</td>
<td>91.18</td>
<td>75.00</td>
<td>97.00</td>
<td>98.90</td>
<td>85.40</td>
<td><strong>92.01</strong></td>
</tr>
<tr>
<td>DT</td>
<td>73.61</td>
<td>67.65</td>
<td>80.52</td>
<td>80.65</td>
<td>87.25</td>
<td>64.77</td>
<td>84.85</td>
<td>96.13</td>
<td>77.86</td>
<td>79.25</td>
</tr>
<tr>
<td>NB</td>
<td><strong>100.00</strong></td>
<td>82.35</td>
<td>80.52</td>
<td>58.06</td>
<td>62.75</td>
<td>73.86</td>
<td>90.91</td>
<td>97.79</td>
<td>84.29</td>
<td>81.17</td>
</tr>
<tr>
<td>$k$-NN</td>
<td>84.72</td>
<td>76.47</td>
<td>84.42</td>
<td>74.19</td>
<td>76.47</td>
<td>69.32</td>
<td>87.88</td>
<td>98.34</td>
<td>82.86</td>
<td>81.63</td>
</tr>
<tr>
<td>SVM</td>
<td>98.61</td>
<td>82.35</td>
<td>97.40</td>
<td>82.26</td>
<td>91.18</td>
<td>76.14</td>
<td><strong>100.00</strong></td>
<td><strong>99.45</strong></td>
<td><strong>93.21</strong></td>
<td><strong>91.18</strong></td>
</tr>
<tr>
<td>PAM</td>
<td>97.22</td>
<td>82.35</td>
<td>85.71</td>
<td>85.48</td>
<td>91.18</td>
<td><strong>79.55</strong></td>
<td><strong>100.00</strong></td>
<td><strong>99.45</strong></td>
<td>79.29</td>
<td>88.91</td>
</tr>
</tbody>
</table>

The best prediction rate for each particular data set is highlighted in boldface.
A few general lessons

- Choosing markers based on relative expression reversals of gene pairs has proven to be very robust with high predictive accuracy in sets we have tested so far
  - Simple and independent of normalization
- Easy to implement clinical test ultimately
  - All that’s needed is RT-PCR on two genes
- M.D. Anderson Cancer Center is now further evaluating this test as part of its clinical biomarkers program - so hopefully this will benefit patients in the near term
Future Directions

- Evaluate and extend methods based on relative expression reversals to:
  - Differentiate many diseases simultaneously
  - Evaluate more complex phenotypes
    - Survival
    - Distant and local recurrence
    - Metastasis
  - Integrate with network information to inform types of comparisons to make