DREAM: a Dialogue on Reverse Engineering Assessment and Methods

Andrea Califano:

**MAGNet**: Center for the Multiscale Analysis of Genetic and Cellular Networks
**C2B2**: Center for Computational Biology and Bioinformatics
**ICRC**: Irving Cancer Research Center
**Columbia University**
Reverse Engineering

- Inference of a predictive (generative) model from data.

  E.g. argmax[\(P(Data|Model)\)]

- Assumptions:
  - Model variables (E.g., DNA, mRNA, Proteins, cellular sub-structures)
  - Model variable space: At equilibrium, temporal dynamics, spatio-temporal dynamics, etc.
  - Model variable interactions: probabilistics (linear, non-linear), explicit kinetics, etc.
  - Model topology: known a-priori, inferred.

- Question:
  - Model \(\sim\) Reality?
Reverse Engineering

Biological System

High-throughput Biology

Biochemical Validation

Specific Prediction

Data

Expression

Proteomics

Structure

Sequence

Model

Control X-Y- Control X+Y+

Control X+Y- Control X-Y+

NKAT
ATGATGGATG
CTCGCATG
CGACGATCAG
GTGTAGCCTG
GGCTGGA
Some Reverse Engineering Methods

• Optimization: High-Dimensional objective function max corresponds to best topology
  – Liang S, Fuhrman S, Somogyi (REVEAL)
  – Gat-Viks and R. Shamir (Chain Functions)
  – Segal E, Shapira M, Regev A, Pe’er D, Botstein D, Koller D, and Friedman N (Prob. Graphical Models)
  – Jing Yu, V. Anne Smith, Paul P. Wang, Alexander J. Hartemink, Erich D. Jarvis (Dynamic Bayesian Networks)
  – …

• Regression: Create a general model of biochemical interactions and fit the parameters
  – Gardner TS, di Bernardo D, Lorentz D, and Collins JJ (NIR)
  – Alberto de la Fuente, Paul Brazhnik, Pedro Mendes
  – Roven C and Bussemaker H (REDUCE)
  – …

• Probabilistic and Information Theoretic: Compute probability of interaction and filter with statistical criteria
  – Atul Butte et al. (Relevance Networks)
  – Gustavo Stolovitzky et al. (Co-Expression Networks)
  – Andrea Califano et al. (ARACNE, MINDY)
  – …

• Evidence Integration: Use databases to provide scaffolding
  – P. Shannon, A. Markiel, O. Ozier, NS Baliga, JT Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker (Cytoscape)
  – …
Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling

Timothy S. Gardner, Diego di Bernardo, David Lorez, James J. Collins

The complexity of cellular gene, protein, and metabolite networks can hinder attempts to elucidate their structure and function. To address this problem, we used systematic transcriptional perturbations to construct a first-order model of regulatory interactions in a nine-gene subnetwork of the SOS pathway in *Escherichia coli*. The model correctly identified the major regulatory genes and the transcriptional targets of mitomycin C activity in the subnetwork. This approach, which is experimentally and computationally scalable, provides a framework for elucidating the functional properties of genetic networks and identifying molecular targets of pharmacological compounds.

Fig. 1. Diagram of interactions in the SOS network. DNA lesions caused by mitomycin C (MMC) are converted to single-stranded DNA during chromosomal replication. Upon binding to ssDNA, the RecA protein is activated (RecA*) and serves as a coprotease for the LexA protein. The LexA protein is degraded, thereby diminishing the repression of genes that mediate multiple protective responses. Boxes denote genes, ellipses denote proteins, hexagons indicate metabolites, arrows denote positive regulation, filled circles denote negative regulation. Red emphasis denotes the primary pathway by which the network is activated after DNA damage.
Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data

Eran Segal\textsuperscript{1,6}, Michael Shapira\textsuperscript{2}, Aviv Regev\textsuperscript{3,5,6}, Dana Pe’er\textsuperscript{4,6}, David Botstein\textsuperscript{2}, Daphne Koller\textsuperscript{1} & Nir Friedman\textsuperscript{4}

\textbf{Figure 3} The respiration and carbon regulation module (55 genes). (a) Regulation tree/program. Each node in the tree represents a regulator (for example, Hap4) and a query of its qualitative value (for example, red upward arrow next to Hap4 for “is Hap4 upregulated?”). The expression of the regulators themselves is shown below their respective nodes. (b) Gene expression profiles. Genes, rows, arrays, columns. Arrays are arranged according to the regulation tree. For example, the rightmost leaf includes the arrays in which both Hap4 and HML Alpha2 are upregulated. Content that consist primarily of one or two types of experimental conditions are labeled. (c) Significant annotations. Colored boxes indicate genes with the respective annotation. The most significantly enriched annotations for this module were selected for display (the number of annotated genes and the calculated \( P \) value for the enrichment of each annotation are shown in parentheses). Note the enrichment of three annotations representing a biochemical process, cellular compartment and physiological process, respectively, all relating to cellular respiration. (d) Promoter analysis. Lines represent 500 bp of genomic sequence located upstream to the start codon of each of the genes; colored boxes represent the presence of cis-regulatory motifs located in these regions. Note the enrichment of both the HAP4 motif (purple) and the stress response element (STRE, green), recognized by Hap4 and Mox4, respectively, supporting their inclusion in the module’s regulation program.
Reverse engineering of regulatory networks in human B cells

Katia Basso⁴, Adam A Margolin⁴, Gustavo Stolovitzky⁵, Ulf Klein¹, Riccardo Dalla-Favera¹,⁴ & Andrea Califano²

Does it generalize?

Does it work only for MYC?
Towards a proteome-scale map of the human protein-protein interaction network

Jean-François Rual*, Kavitha Venkatesan*, Tong Hao, Tomaž Hirozare-Kishikawa, Amélie Drucet, Ning Li, Gabriel F. Berriz, Francis D. Gibbons, Matja Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S. Goldberg, Lan V. Zhang, Shanyi L. Wong, Giovanni Franklin, Siming Li†, Joanna S. Abala†, Janghoo Lim†, Carlene Fraughton†, Estelle Llamas†, Sebha Cevik†, Camille Bex†, Philippe Lamesch†, Robert S. Sikorski†, Jean Vandenhaute†, Hucia Y. Zoghi†, Alex Smialow†, Stephanie Bosak†, Reynaldo Sequerra†, Lynn Doucette-Steere†, Michael E. Cusick†, David E. Hill†, Frederick P. Roth† & Marc Vidal

Validation by co-affinity purification:

LCI: 62%
Y2H: 78%
Both: 81%

And which part?
Gene Regulatory Networks and the Evolution of Animal Body Plans

Eric H. Davidson* and Douglas H. Erwin

Development of the animal body plan is controlled by large gene regulatory networks (GRNs), and hence evolution of body plans must depend upon change in the architecture of developmental GRNs. However, these networks are composed of diverse components that evolve at different rates and in different ways. Because of the hierarchical organization of developmental GRNs, some kinds of change affect terminal properties of the body plan such as occur in speciation, whereas others affect major aspects of body plan morphologies. A notable feature of the palaeontological record of animal evolution is the establishment by the early Cambrian of virtually all phylum-level body plans. We identify a class of GRN component, the "kernels" of the network, which, because of their developmental role and their particular internal structure, are most impervious to change. Conservation of phyletic body plans may have been due to the retention since pre-Cambrian times of GRN kernels, which underlie development of major body parts.
CASP-style: CASP (Critical Assessment of Structure Predictions) is a bi-yearly workshop where blind protein structure predictions (obtained by various methods) are compared to structures assessed by experimental methods. The latter are kept secret until the deadline for submission of the computational structure predictions. Algorithms compete within specific categories (i.e. ab-initio, homology-based methods, etc.)

CASP-style workshop and database to assess the quality of methods and data for the reverse engineering of cellular networks
- Assess: against which standard?
- Assess: what type of predictions?
  - Transcriptional
  - Signaling
  - Metabolic
  - Protein-Protein
- Assess: which cellular context?
- Assess: which methods?

Planning meeting: May 9-10 NYAS
DREAM Working Group (SC)

- Gary Bader (MSK)
- Joel Bader (JHU)
- Diego Di Bernardo (TIGEM)
- Hamid Bolouri (ISB)
- Harmen Bussemaker (CU)
- Andrea Califano (CU)
- Jim Collins (BU)
- Eric Davidson (Caltech)
- Tim Gardner (BU)
- Mark Gerstein (Yale)
- Alexander Hartemink (Duke)
- Trey Ideker (UCSD)

- Andre Levchenko (JHU)
- Pedro Mendes (VP)
- John Moult (U.Maryland)
- Andrey Rzhetsky (CU)
- Benno Schwikowski (Pasteur)
- Eran Segal (Weitzman)
- Ron Shamir (TAU)
- Mike Snyder (Yale)
- Gustavo Stolovitzky (IBM)
- Marc Vidal (Harvard)
- Mike Yaffe (MIT)
What is DREAM

• An attempt to assemble:
  – Useful metrics for the evaluation of reverse engineering methods
  – Plausible/reasonable predictive/generative models
    • Synthetic (Biologically motivated)
    • Biological
    • Bioengineered
  – “Gold Standard” data that would be useful to assess method’s performance
  – Common threads to engage the reverse-engineering community in a dialogue to help further structure and consolidate the field
  – A Database with Data, Methods, Publications, and Predictions.

• Similar Efforts in complementary fields
  – CASP (Critical Assessment of Structure Prediction)
  – GAW (Genetic Analysis Workshop)
  – CAMDA (Critical Assessment of Microarray Data Analysis)
  – CoEPrA (Comparative Evaluation of Predictive Algorithms)
  – CAPRI (Comparative Assessment of Protein Interactions)
1st DREAM Workshop

Sept. 7-8, 2006. Wave Hill NY

- NIH National Centers for Biomedical Computing (NCBC)
- IBM
- New York Academy of Science
- Center for Discrete Mathematics and Theoretic Computer Science

Participants 120
Invited Presentations 2+8
Submitted papers 30
Accepted papers 12
Accepted posters 13
NYAS Annals Volume
Session I: Experimental Gold Standards

- **Invited:**
  - Dana Pe'er – Inferring Regulatory Pathways: Data and experimental design
  - Joel Bader – Estimating the size of the Human Interactome

- **Submitted:**
  - I. Cantone, D. di Bernardo, and M.P. Cosma – Benchmarking reverse-engineering strategies via a synthetic gene network in Saccharomyces cerevisiae
  - Theodore J. Perkins – The gap gene system of Drosophila melanogaster: Model-fitting and validation
Interventions on a known map...

Activators
1. α-CD3
2. α-CD28
3. ICAM-2
4. PMA
5. β2cAMP

Inhibitors
6. G06976
7. AKT inh
8. Psitect
9. U0126
10. LY294002

D. Pe'er
Inferred T cell signaling map


Phospho-Proteins
Phospho-Lipids
Perturbed in data

<table>
<thead>
<tr>
<th>T Cells</th>
<th>15/17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported</td>
<td>17/17</td>
</tr>
<tr>
<td>Reversed</td>
<td>1</td>
</tr>
<tr>
<td>Missed</td>
<td>3</td>
</tr>
</tbody>
</table>
Engineered Network topology

H0 Locus

She2 Locus

Cbf1 Locus

Ash1

Swi5

Gal80

D. di Bernardo
Session II: Synthetic Gold Standard

- Invited:
  - Pedro Mendes – In Silico Models for Reverse Engineering: Complexity and Realism versus Well-Defined Metrics
  - Leslie Loew – In Silico Gold Standards from Virtual Cell

- Submitted:
  - B. Stigler, M. Stillman, A. Jarrah, P. Mendes, and R. Laubenbacher – Reverse Engineering of Network Topology
  - Winfried Just – Data requirements of reverse-engineering algorithms
Session II: Synthetic Gold Standard

\[ \frac{dG_t}{dt} = s(G_t, K, G_n) - b(G_t), \]

\[ \delta_t(G_1, K, G_n) = V_t \cdot \prod_j \left( \frac{K_i}{I_j} + (1 + K_i) \right) \]

BIOINFORMATICIANS

Pedro Mendes

P. Mendes
Synthetic Benchmarks:

- Erdos-Renyi
- Scale-free

(a) Erdos-Renyi network
(b) Scale-free network

(a) Precision vs. Recall for different models
(b) Precision vs. Recall for different models
Session III: Data Generation and Validation

• Invited:
  – Riccardo Dalla-Favera – Validating Pathways in Human B Cells
  – Eric Schadt – Simulations and Multifactorial Gene Perturbation Experiments as a Way to Validate Reverse Engineered Gene Networks Reconstructed via the Integration of Genetic and Gene Expression Data

• Submitted:
The germinal center

B-Cell Subpopulations

Pre-GC
Naïve B → Germinal Center (GC) → Centroblast → Centrocyte → Post-GC → Plasma Cell

Memory B

Germinal Center (GC)

B-Cell Derived Malignancies

Mantle Cell Lymphoma (CD5+)
B-CLL (CD5+)
Follicular Lymphoma
Hodgkin Disease
Burkitt Lymphoma
Diffuse Large Cell Lymphoma
Multiple Myeloma
• Changes occurred in 11,814 non-unique edges

Edge Phenotype Map (sig. level = 6.2813e-11)

- MI (+) when phen. removed
- MI (-) when phen. removed

Gene 1 vs. Gene 2
Session IV: RE Algorithms and Metrics

• Invited:
  – James J. Collins – Reverse Engineering Gene-Protein Networks
  – Mark Gerstein – Understanding Biological Function through Evaluation of Genome-scale Networks

• Submitted:
  – M. Socolovsky, M. Murrell, Y. Liu, R. Pop, E. Porpiglia, and A. Levchenko – Computational Modeling of Fetal Erythroblasts Predicts Negative Autoregulatory Interactions Mediated by Fas and its ligand
  – L. David and C. Wiggins – Quantifying Reliability of Dynamic Bayesian Networks
  – A. Bernard and A.J. Hartemink – Evaluating Algorithms for Learning Biological Networks
How do we use the Networks?

Treat cells with drug compound

Obtain expression profile

Filter profile using identified network

ID direct genetic targets of drug

Solved Using NIR

Profiling Drug Targets

J.J. Collins
Validation: recA/lexA Double Perturbation

Expression changes Following recA/lexA double perturbation

Cannot distinguish affected genes using just expression data

Correct mediators of expression profile identified using NIR approach

Predicted mediators: lexA and recA identified as perturbed genes by network model

J.J. Collins
Individual Features and their Integration for Yeast Membrane Protein Interaction Prediction

Combination of All Features

- MIPS
- GO
- Genetic Interactions
- mRNA co-expr.

Mark Gerstein
DREAM Metrics
Based on your evaluation of the workshop goals and the program, would you recommend that this project is continued in future years

Yes 31
No 0
Total 31

What is your research background

Computational 11
Experimental 1
Both 4
Total 16
Q2: Based on the current session topics did you find that
The sessions accurately represent and address the workshop goals

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>27</td>
<td>1</td>
</tr>
</tbody>
</table>

- [b] Some sessions are redundant and should be eliminated

<table>
<thead>
<tr>
<th>Session</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>0</td>
</tr>
<tr>
<td>Session 2</td>
<td>1</td>
</tr>
<tr>
<td>Session 3</td>
<td>0</td>
</tr>
<tr>
<td>Session 4</td>
<td>1</td>
</tr>
</tbody>
</table>

- [c] Some sessions are critically missing

1. Reverse Engineering vs. model inversion problems.
2. Validation
3. Estimating Dynamic Parameters
4. Temporal simulations of RE nets.
5. Metabolic modeling.
6. Validation of Sequence motifs (DNA, protein)
7. The format should probably evolve as the field matures.
8. Definition of Reverse Engineering.
9. Some more on cellular phenotypes and applications.
10. Inverse methods for real parameter inference.
11. The session topics were good, but the talks didn’t always fit.
12. Statistical rigorous evaluation
14. RE methods comparisons
15. Short poster presentations before poster sessions.
Should the DREAM Workshop directly address the issue of comparing the performance of reverse engineering algorithm using blind data (similar to CASP)?

<table>
<thead>
<tr>
<th>Option</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>22</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
</tr>
<tr>
<td>Maybe</td>
<td>1</td>
</tr>
<tr>
<td>Premature</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>29</strong></td>
</tr>
</tbody>
</table>
On a scale from 1 to 5 (5 showing the highest interest), would you use (or find useful) a DREAM curated database with the following information:

<table>
<thead>
<tr>
<th>Category</th>
<th>Very Low</th>
<th>Low</th>
<th>Med</th>
<th>High</th>
<th>Very High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. RE Algorithms/Software</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>19</td>
<td>30</td>
</tr>
<tr>
<td>b. Experimental Data</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>c. Blind Experimental Data</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>d. Synthetic Data</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>11</td>
<td>29</td>
</tr>
<tr>
<td>e. Biological Network Questions</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>f. Predictions for bio-validation</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>13</td>
<td>29</td>
<td>29</td>
</tr>
</tbody>
</table>
**Questionnaire (Suggestions)**

### Longer Event and more sessions/discussion

- More time for discussions.
- 3 day workshop + ½ hour submitted papers.

### Comments on Topic Focus

- Protein-level was under-represented in comparison to genetic nets
- ODE models were under-represented in comparison to genetic nets
- Discussion on Network Analysis.
- Metabolic analysis is premature.
- Separate regulatory validation from interaction validations, since data + results are different.
- Interference between different levels of modeling: Protein interaction nets interact with metabolic and transcription nets dynamically.
- Better organization with respect to the wide range of questions we are dealing with.

### Logistics

- Have the workshop closer to the hotel.
- Organize the conference closer to hotel.
- Common dinner.
- A social mixer so that people can get to know each other better.
- Integrate with GEO?

### Comments on Biological Focus

- Once a net is validated, there are Q to be answered which have a biological/physiological impact. Speakers should answer which questions were predicted with networks.
- More talks on experimental validation techniques: TF-DNA, signaling, protein interaction.
- Invite more experimentalists.
- Discussions of applications, biological context and phenotypic context.

### CASP-like Competition

- Weaker version of CASP would be good (written by someone who said no to Q3): start easy: topology. Separate supervised and unsupervised learning.
- Perhaps a CASP like contest could take place in the future for defined projects. But much more discussion is needed.
- If a CASP like contest is implemented, avoid the trap of having talk after talk saying: “My algo is better than yours”. Concentrate on why and methodology of assessment.

### DREAM Website

- Launch a challenge to infer some nets from Data in the DREAM website.
- Items b and f in Q4 would be useful but impractical, and a logical nightmare.
Recommendations

• Gold Standards
  – Synthetic GS provide complete data and are relatively realistic. Experimental Biologists do not trust them yet
  – Experimental GS provide very incomplete data and contain both false positives (few) and negatives (many)
  – Engineered GS address both issues but are artificial and small scale

• Biochemical Validation
  – Several criteria/recommendations for validation should emerge as a result of the “Dialogue”
    • Literature, Binding Site Analysis, ChIP, PP interaction assays, reporter assays, co-IP, etc.
  – Define an acceptable Precision. (30%-50%?)

• Methods
  – Repository of methods (with publications/data) is needed

• Models
  – Current Subdivisions (metabolic/signaling/transcriptional) are artificial. The cellular interaction networks are composite
  – Context Specificity is very important

• Use of the Networks
  – Emerging as a key justification for the field
Issues and Roadblocks

• Field is relatively new and rapidly evolving. DREAM Goals may become a moving target

• All similar efforts have a well-defined methodology to produce “Gold-Standard” data

• Reverse Engineering feeds on many data types and infers diverse cellular network types
  – Sequence, Expression profiles, Structure, ChIP-Chip, Synthetic Lethality, etc.
  – Regulatory, Signaling, Metabolic, etc.

• Welding the computational and experimental communities will be hard
  – Can we expect experimentalists to hold off on publication of new biological circuits?
Acknowledgments

- **Co-organizers**
  - Gustavo Stolovitzky
  - Jim Collins

- **Funding Agencies**
  - NIH Roadmap
  - IBM
  - NYAS
  - DIMACS

- **Speakers**
  - See individual Sessions

- **Steering Committee**
  - Gary Bader (MSK)
  - Joel Bader (JHU)
  - Diego Di Bernardo (TIGEM)
  - Hamid Bolouri (ISB)
  - Harmen Bussemaker (CU)
  - Andrea Califano (CU)
  - Jim Collins (BU)
  - Eric Davidson (Caltech)
  - Tim Gardner (BU)
  - Mark Gerstein (Yale)
  - Alexander Hartemink (Duke)
  - Trey Ideker (UCSD)
  - Andre Levchenko (JHU)
  - Pedro Mendes (VP)
  - John Moult (U.Maryland)
  - Andrey Rzhetsky (CU)
  - Benno Schwikowski (Pasteur)
  - Eran Segal (Weitzman)
  - Ron Shamir (TAU)
  - Mike Snyder (Yale)
  - Gustavo Stolovitzky (IBM)
  - Marc Vidal (Harvard)
  - Mike Yaffe (MIT)