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.

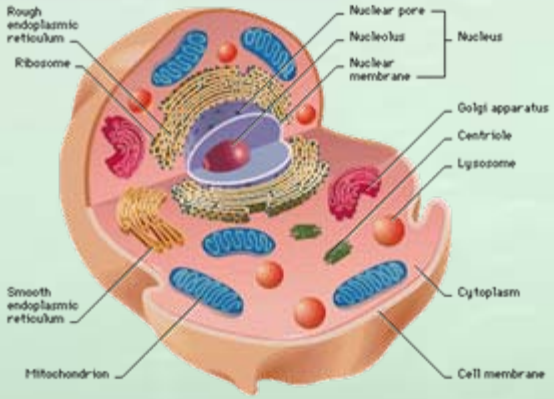
E.g. argmax[$P(Data|Model)$]

- 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 ~= Reality?

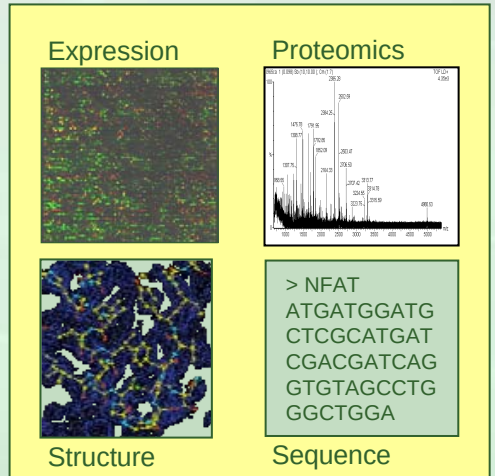
Reverse Engineering

Biological System

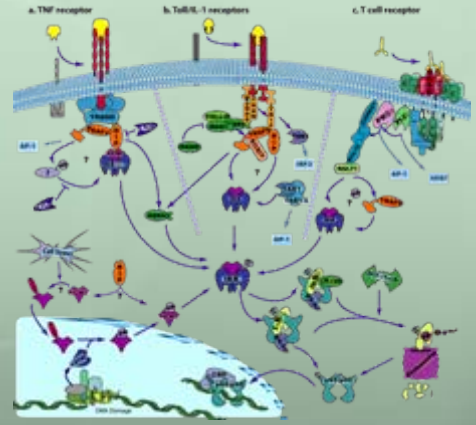


High-throughput Biology

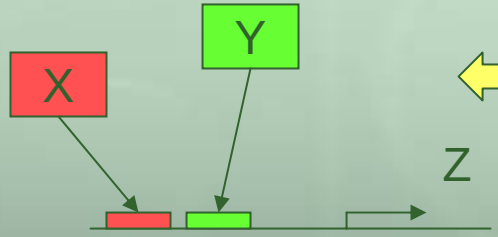
Data



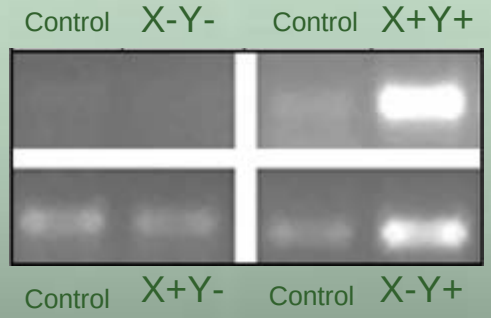
Model



Specific Prediction



Biochemical Validation



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)
  - Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. (Naïve Bayes + Bayesian Net Evidence Integration)
  - Brown CT, Rust AG, Clarke PJC, Panb Z, Schilstrab MJ, De Buysscher T, Griffin G, Wold B, Cameron RA, Davidson EH, Bolouri H (Net Builder)
  - Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W, Wilbur WJ, Hatzivassiloglou V, Friedman C. (GeneWays)
  - …

Is it the problem or the method?

Does it scale?

Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling

Timothy S. Gardner,[1]* Diego di Bernardo,[1,2]* David Lorenz,[1] James J. Collins[1]†

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.

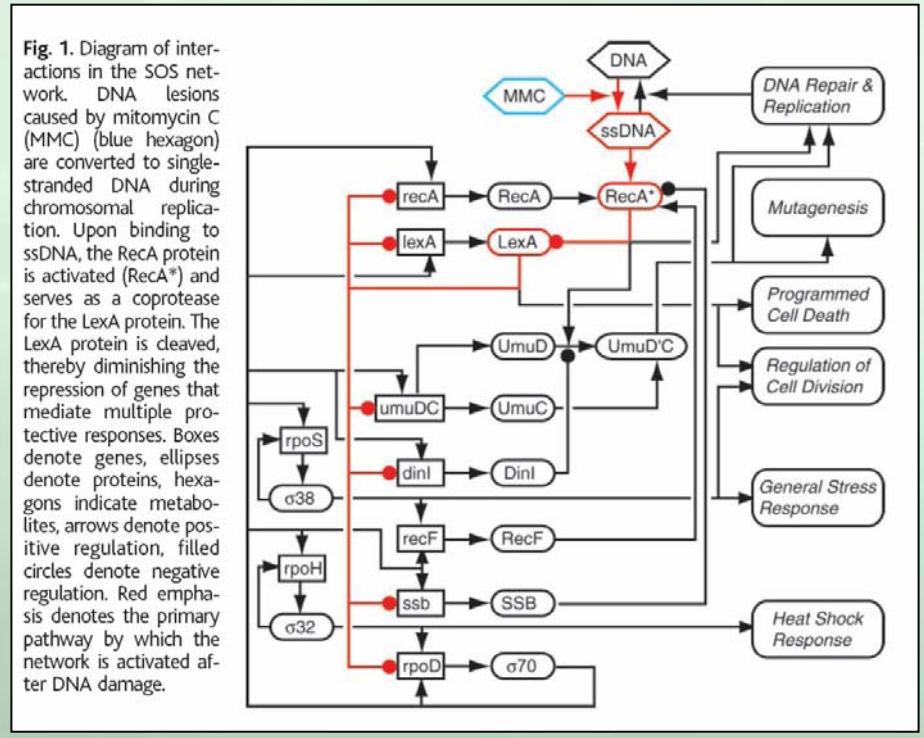


Fig. 1. Diagram of interactions in the SOS network. DNA lesions caused by mitomycin C (MMC) (blue hexagon) are converted to single-stranded DNA during chromosomal replication. Upon binding to ssDNA, the RecA protein is activated (RecA*) and serves as a coprotease for the LexA protein. The LexA protein is cleaved, 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.

Science

Robustness and Validation

Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data

Eran Segal[1,6], Michael Shapira[2], Aviv Regev[3,5,6], Dana Pe'er[4,6], David Botstein[2], Daphne Koller[1] & Nir Friedman[4]

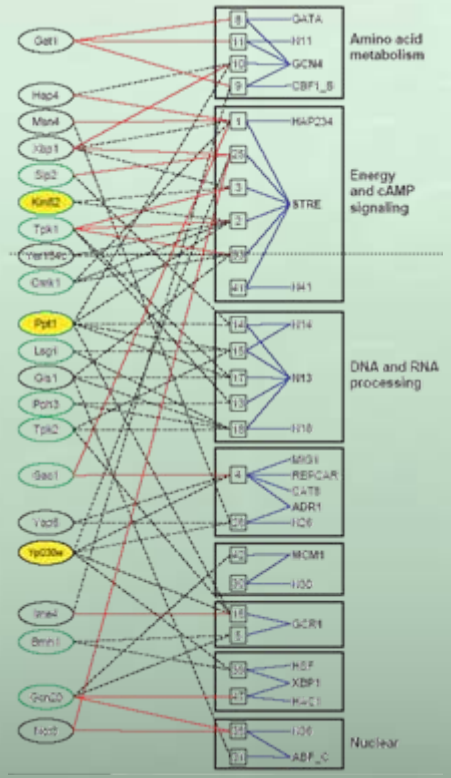
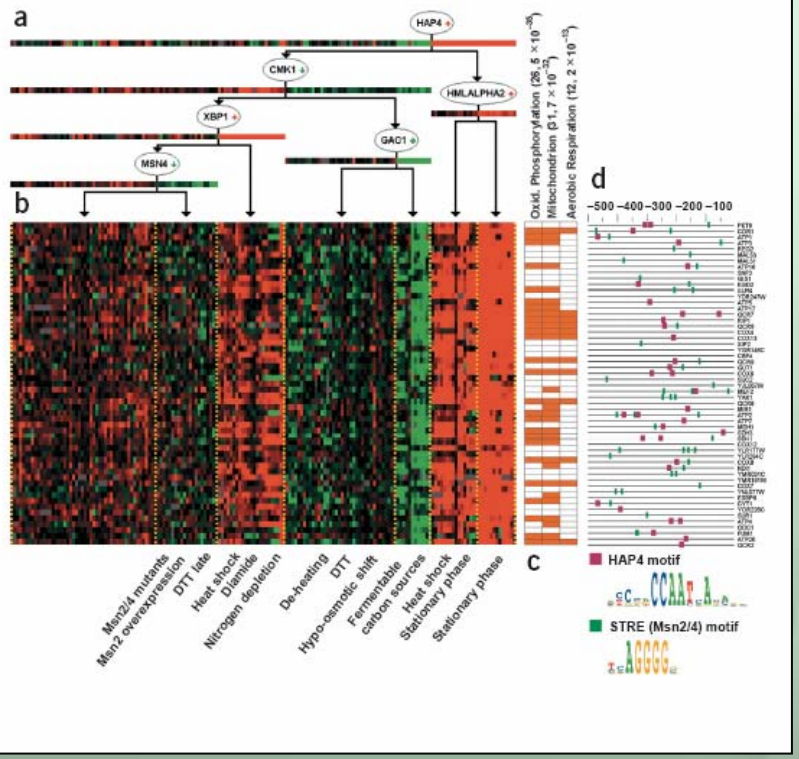


Figure 3 The respiration and carbon regulation module (55 genes). (**a**) Regulation tree/program. Each node in the tree represents a regulator (for example, Hap4) and a query of its qualitative value (for example, red upward arrow next to Hap4 for "is Hap4 upregulated?"). The expression of the regulators themselves is shown below their respective node. (**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 HMLAlpha2 are upregulated. Contexts that consist primarily of one or two types of experimental conditions are labeled. (**c**) Significant annotations. Colored entries 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 Msn4, respectively, supporting their inclusion in the module's regulation program.

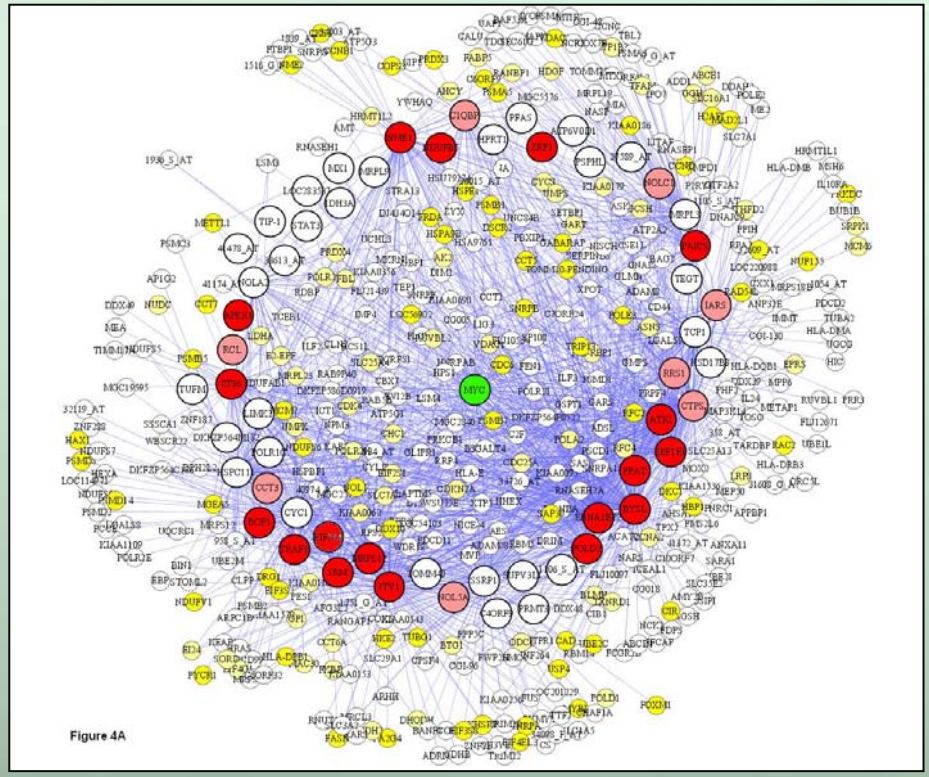


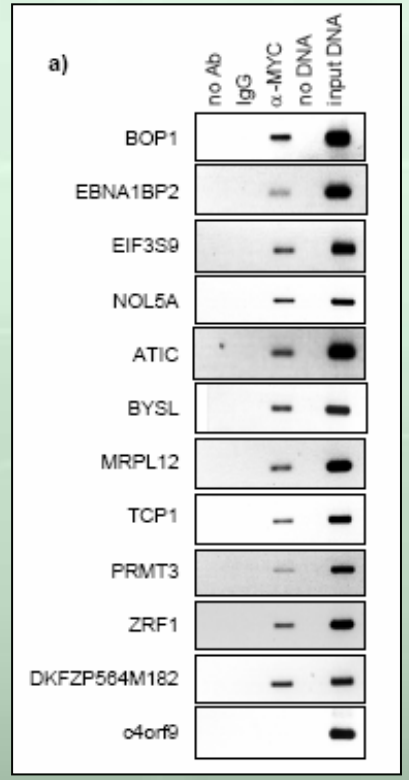
Nature Genetics

Does it generalize?

Reverse engineering of regulatory networks in human B cells

Katia Basso[1], Adam A Margolin[2], Gustavo Stolovitzky[3], Ulf Klein[1], Riccardo Dalla-Favera[1,4] & Andrea Califano[2]





Nature Genetics

Does it work only for MYC?

How much is real ? How much is missing ?

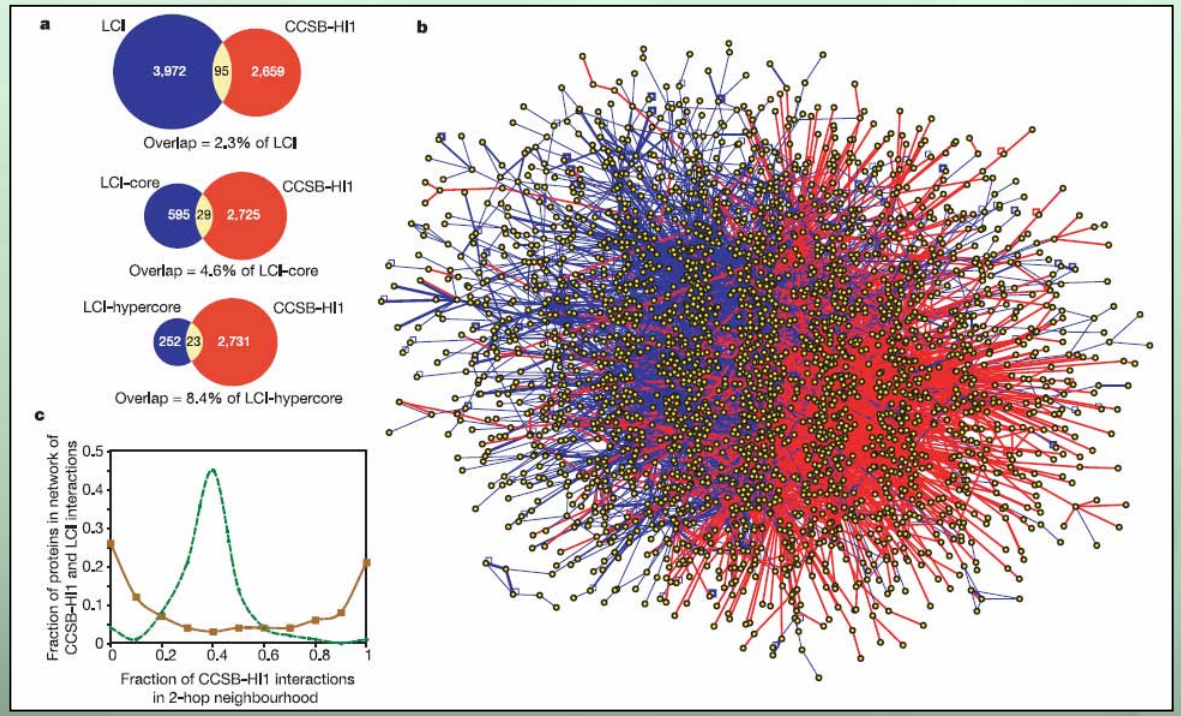
Towards a proteome-scale map of the human protein-protein interaction network

Jean-François Rual[1*], Kavitha Venkatesan[1*], Tong Hao[1], Tomoko Hirozane-Kishikawa[1], Amélie Dricot[1], Ning Li[1], Gabriel F. Berriz[2], Francis D. Gibbons[2], Matija Dreze[1,3], Nono Ayivi-Guedehoussou[1], Niels Klitgord[1], Christophe Simon[1], Mike Boxem[1], Stuart Milstein[1], Jennifer Rosenberg[1], Debra S. Goldberg[2], Lan V. Zhang[2], Sharyl L. Wong[2], Giovanni Franklin[2], Siming Li[1]†, Joanna S. Albala[1]†, Janghoo Lim[4], Carlene Fraughton[1], Estelle Llamosas[1], Sebiha Cevik[1], Camille Bex[1], Philippe Lamesch[1,3], Robert S. Sikorski[5], Jean Vandenhaute[3], Huda Y. Zoghbi[4], Alex Smolyar[1], Stephanie Bosak[6], Reynaldo Sequerra[6], Lynn Doucette-Stamm[6], Michael E. Cusick[1], David E. Hill[1], Frederick P. Roth[2] & Marc Vidal[1]

And which part ?

Validation by co-affinity purification:

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



Nature

The ultimate goal

Gene Regulatory Networks and the Evolution of Animal Body Plans

Eric H. Davidson[1*] and Douglas H. Erwin[2]

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 morphology. A notable feature of the paleontological 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 time of GRN kernels, which underlie development of major body parts.

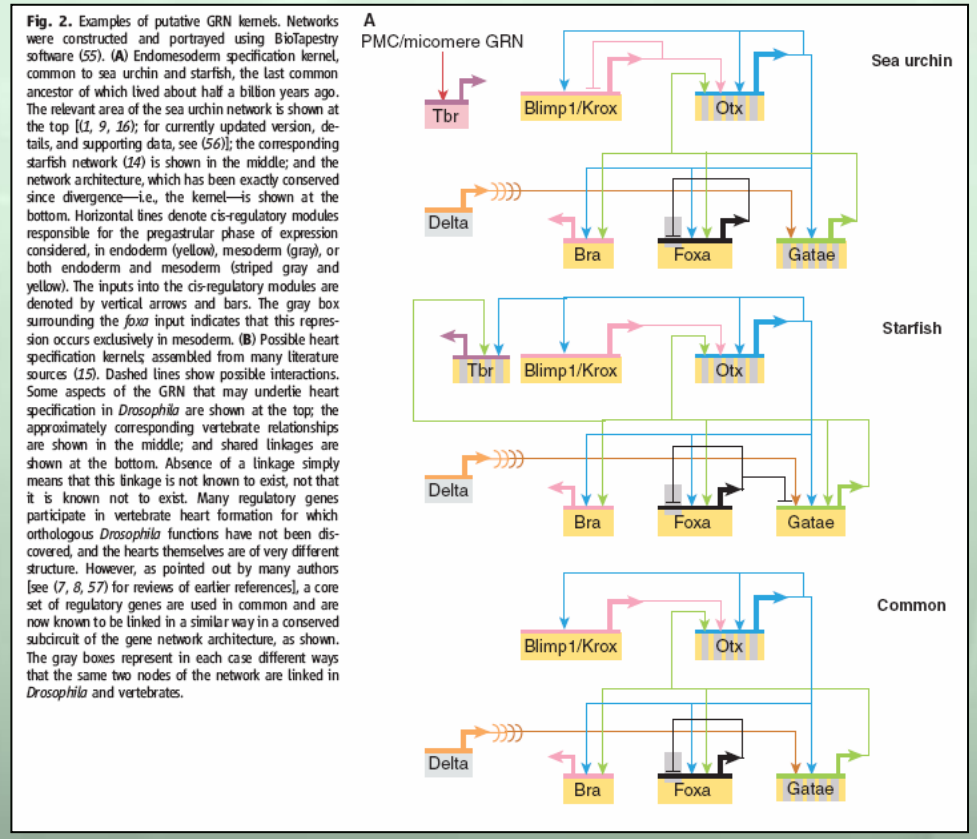


Fig. 2. Examples of putative GRN kernels. Networks were constructed and portrayed using BioTapestry software (*55*). (**A**) Endomesoderm specification kernel, common to sea urchin and starfish, the last common ancestor of which lived about half a billion years ago. The relevant area of the sea urchin network is shown at the top [(*1*, *9*, *16*); for currently updated version, details, and supporting data, see (*56*)]; the corresponding starfish network (*14*) is shown in the middle; and the network architecture, which has been exactly conserved since divergence—i.e., the kernel—is shown at the bottom. Horizontal lines denote cis-regulatory modules responsible for the pregastrular phase of expression considered, in endoderm (yellow), mesoderm (gray), or both endoderm and mesoderm (striped gray and yellow). The inputs into the cis-regulatory modules are denoted by vertical arrows and bars. The gray box surrounding the *foxa* input indicates that this repression occurs exclusively in mesoderm. (**B**) Possible heart specification kernels; assembled from many literature sources (*15*). Dashed lines show possible interactions. Some aspects of the GRN that may underlie heart specification in *Drosophila* are shown at the top; the approximately corresponding vertebrate relationships are shown in the middle; and shared linkages are shown at the bottom. Absence of a linkage simply means that this linkage is not known to exist, not that it is known not to exist. Many regulatory genes participate in vertebrate heart formation for which orthologous *Drosophila* functions have not been discovered, and the hearts themselves are of very different structure. However, as pointed out by many authors [see (*7*, *8*, *57*) for reviews of earlier references], a core set of regulatory genes are used in common and are now known to be linked in a similar way in a conserved subcircuit of the gene network architecture, as shown. The gray boxes represent in each case different ways that the same two nodes of the network are linked in *Drosophila* and vertebrates.

Science

DREAM: a Dialogue

- 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.)

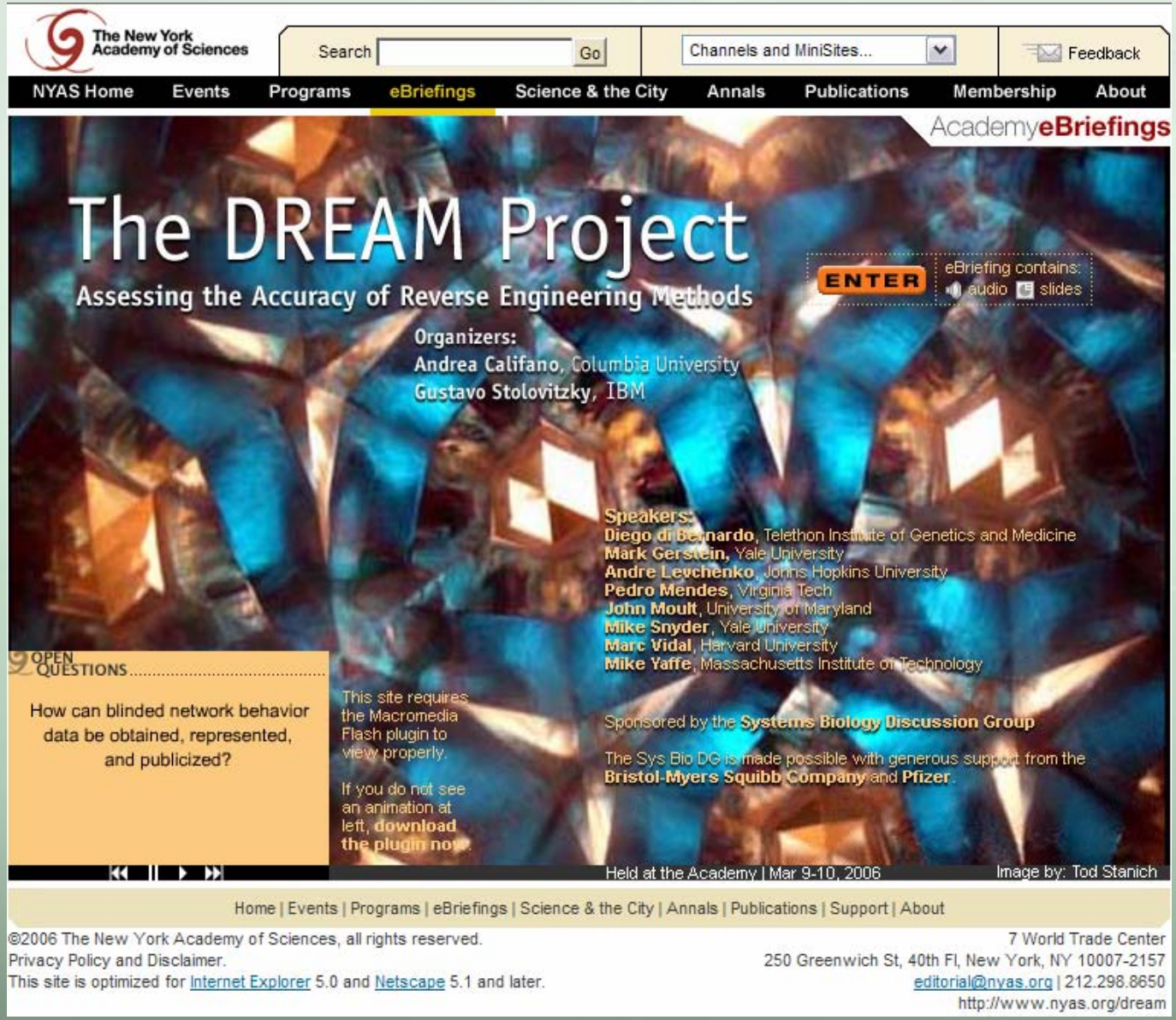
- CASP-style workshop and database to assess the quality of methods and data for the reverse engineering of cellular networks
  - Assess: against which standard?
  - Assess: what type of predictions?
    - Transcriptional
    - Signaling
    - Metabolic
    - Protein-Protein
  - Assess: which cellular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

http://www.nyas.org/ebriefreps/main.asp?intEBriefID=534



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

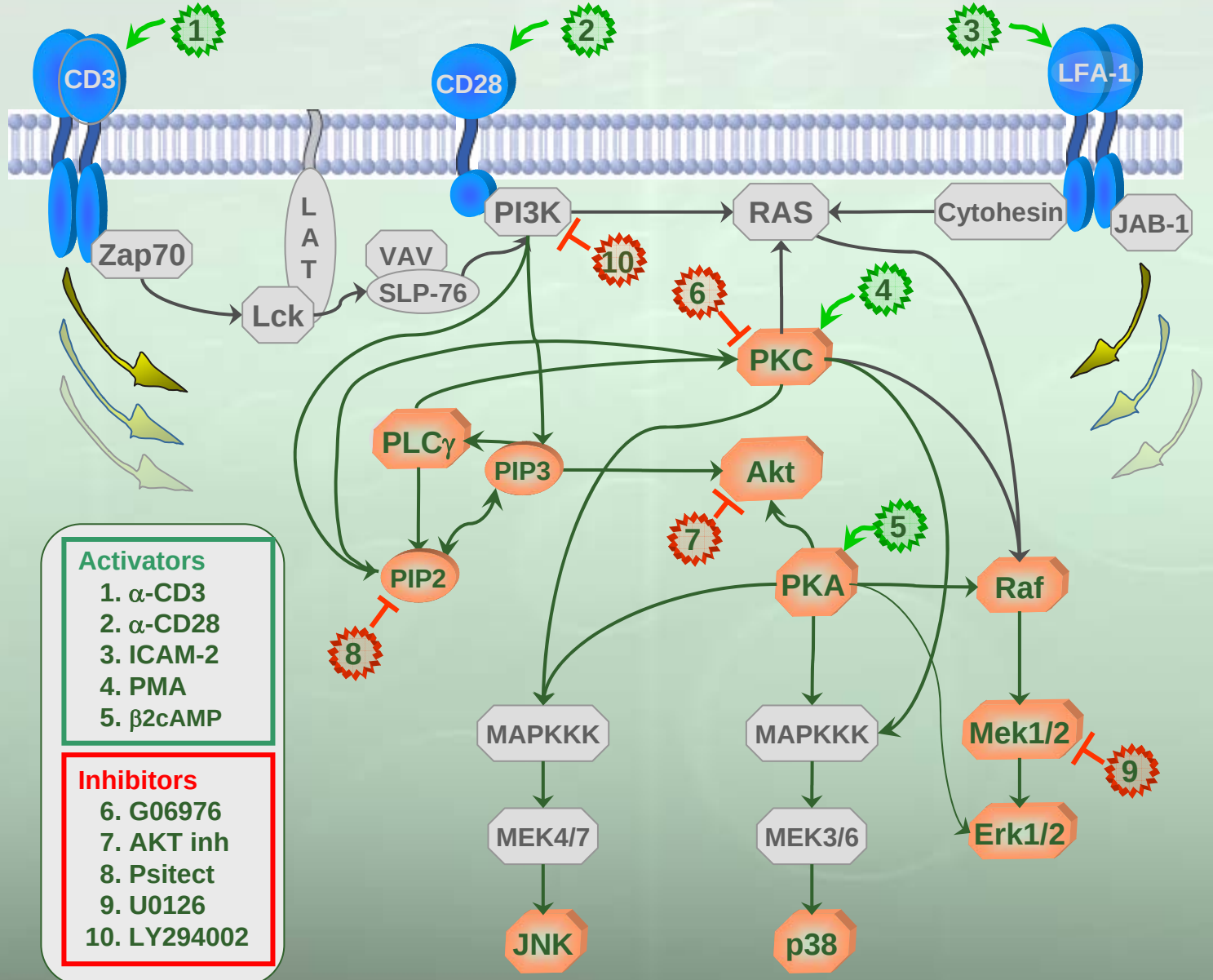
Participants	120
Invited Presentations	2+8
Submitted papers	30
Accepted papers	12
Accepted posters	13
NYAS Annals Volume	

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

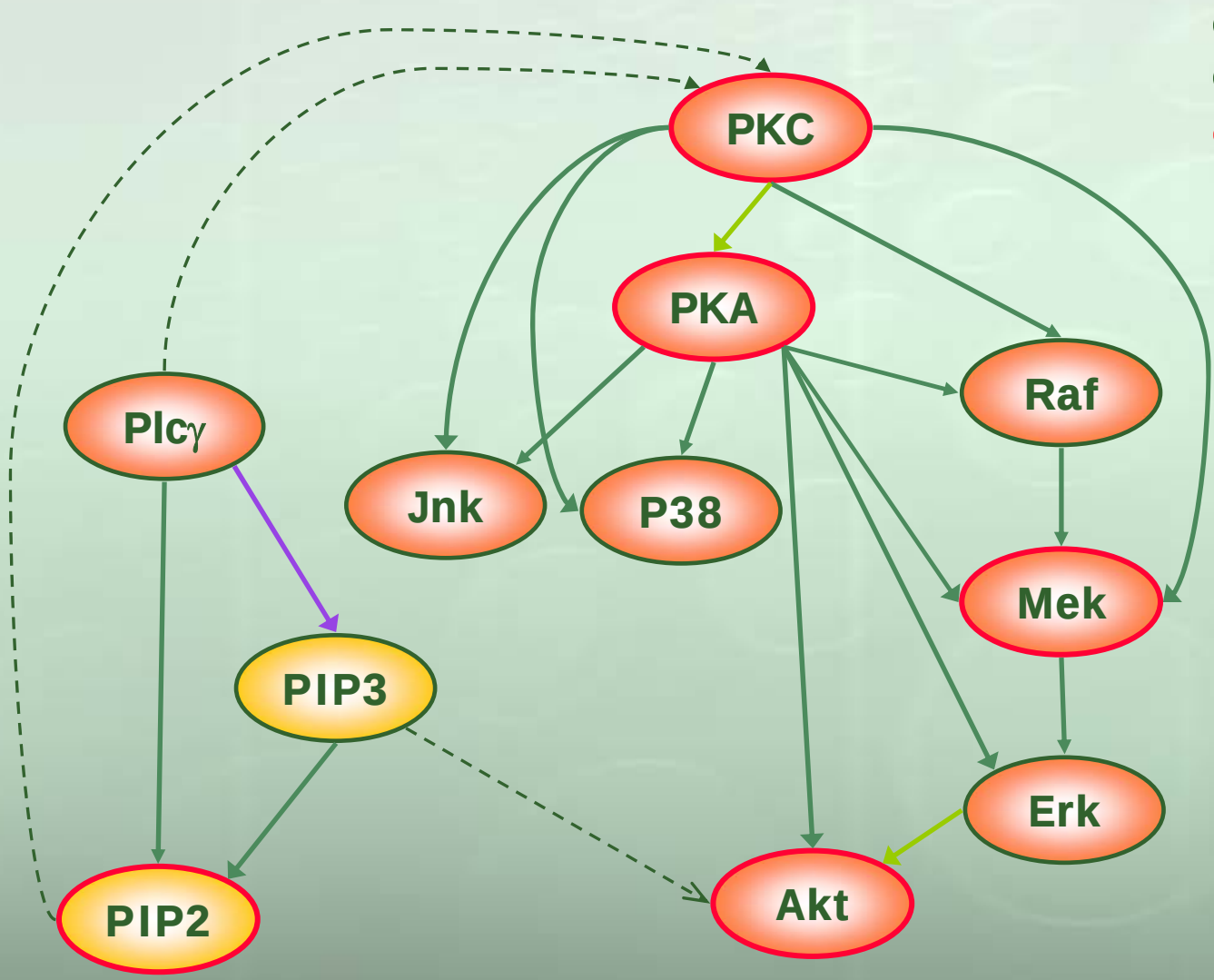
- **Submitted:**
  - I. Cantone, **D. di Bernardo**, and M.P. Cosma – Benchmarking reverse-engineering strategies via a synthetic gene network in Saccharomyces cerevisiae
  - **T. Maiwald**, C. Kreutz, S. Bohl, A.C. Pfeifer, U. Klingmüller, and J. Timmer – Dynamic pathway modeling: Feasibility analysis and optimal experimental design
  - **Theodore J. Perkins** – The gap gene system of Drosophila melanogaster: Model-fitting and validation

Interventions on a known map...



D. Pe'er

Inferred T cell signaling map

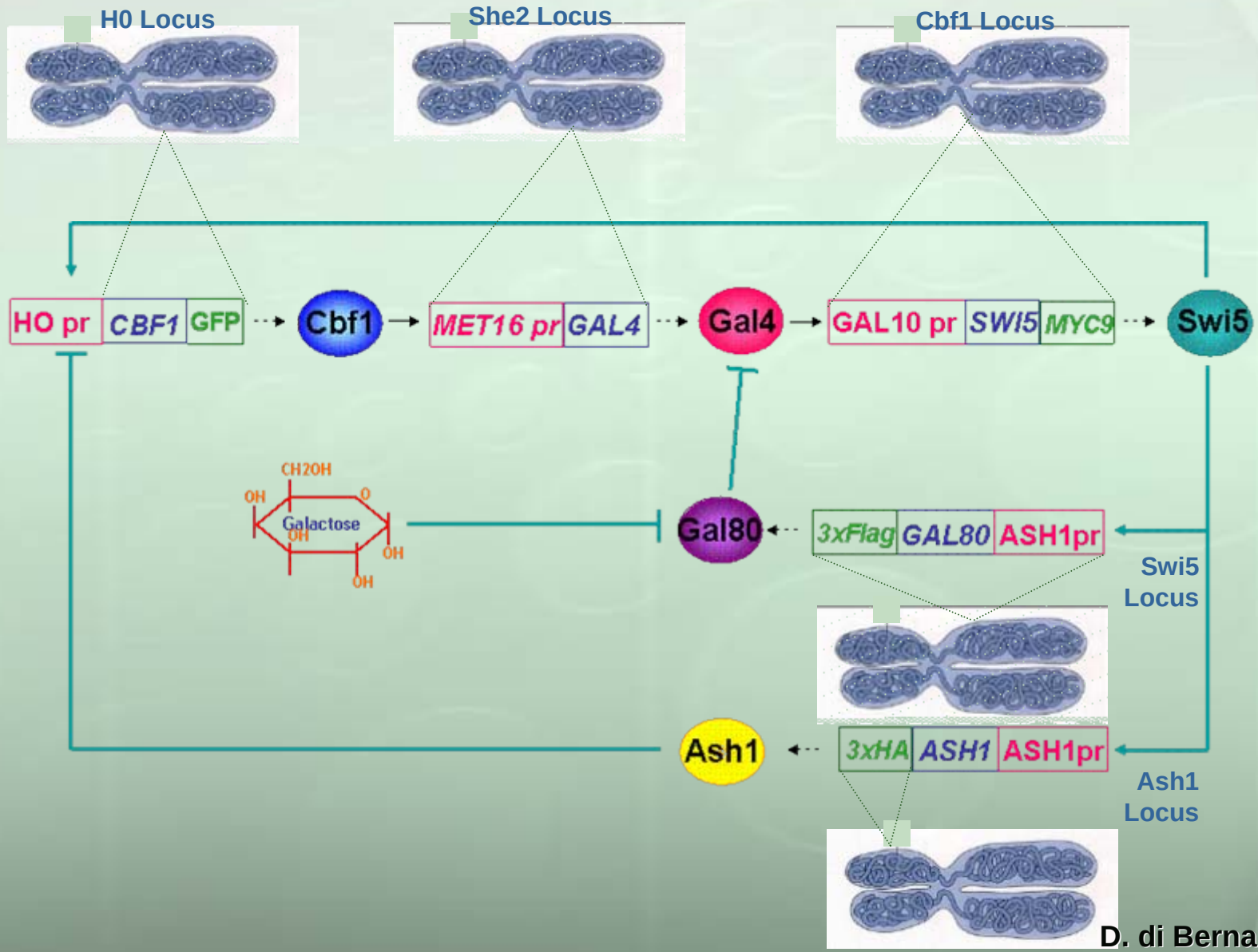


T Cells	15/17
Reported	17/17
Reversed	1
Missed	3

[Sachs *et al*, *Science* 2005]

D. Pe'er

Engineered Network topology



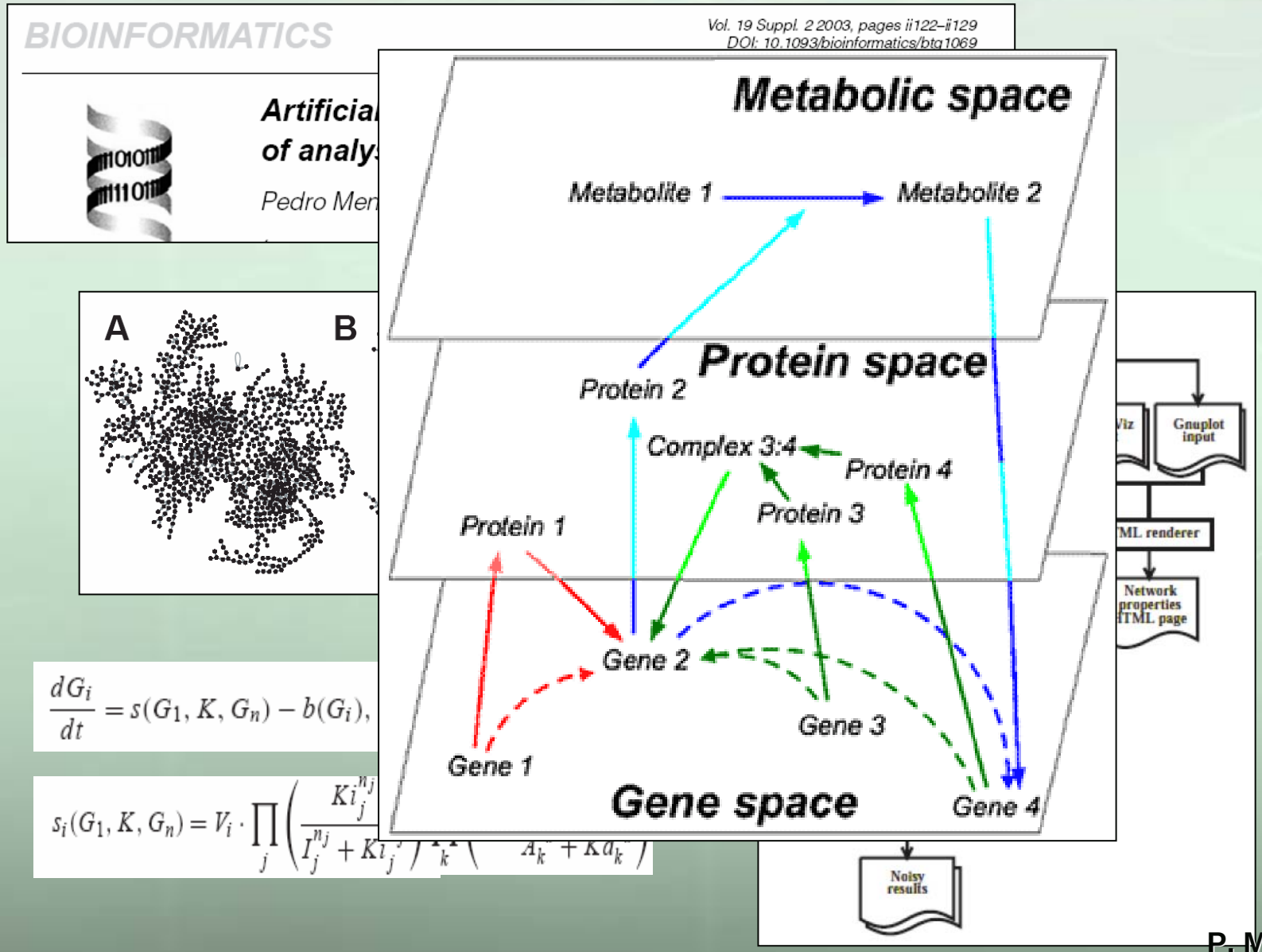
D. di Bernardo

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
  - **I. Nemenman**, M.E. Wall, G.S. Escola, W.S. Hlavacek – Reconstruction of metabolic networks from high throughput metabolic profiling data: *in silico* analysis of Red Blood Cell metabolism

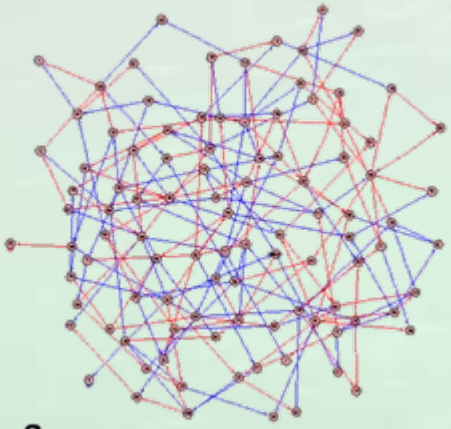
Session II: Synthetic Gold Standard



P. Mendes

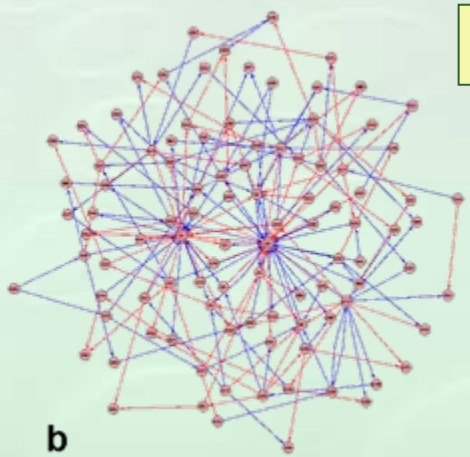
Synthetic Benchmarks:

Erdos-Renyi

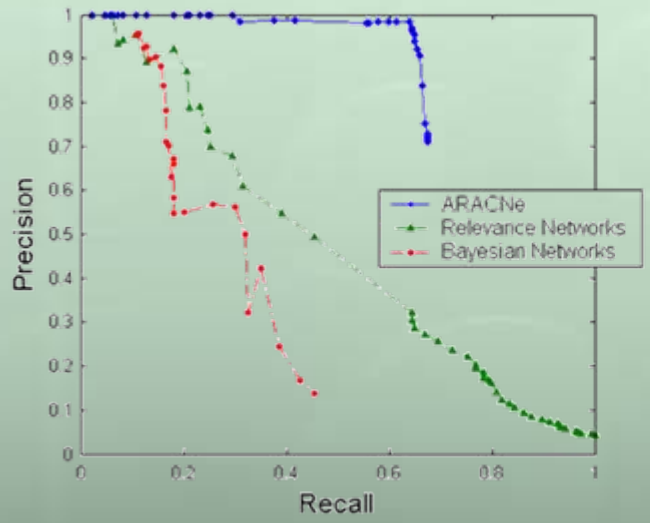


a

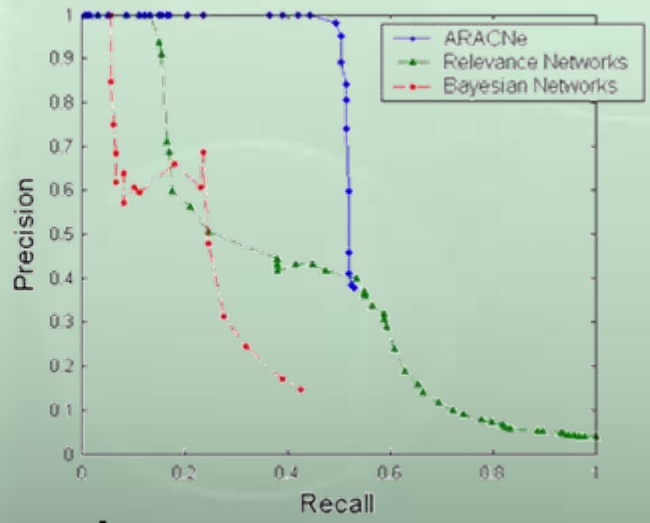
Scale-free



b



a

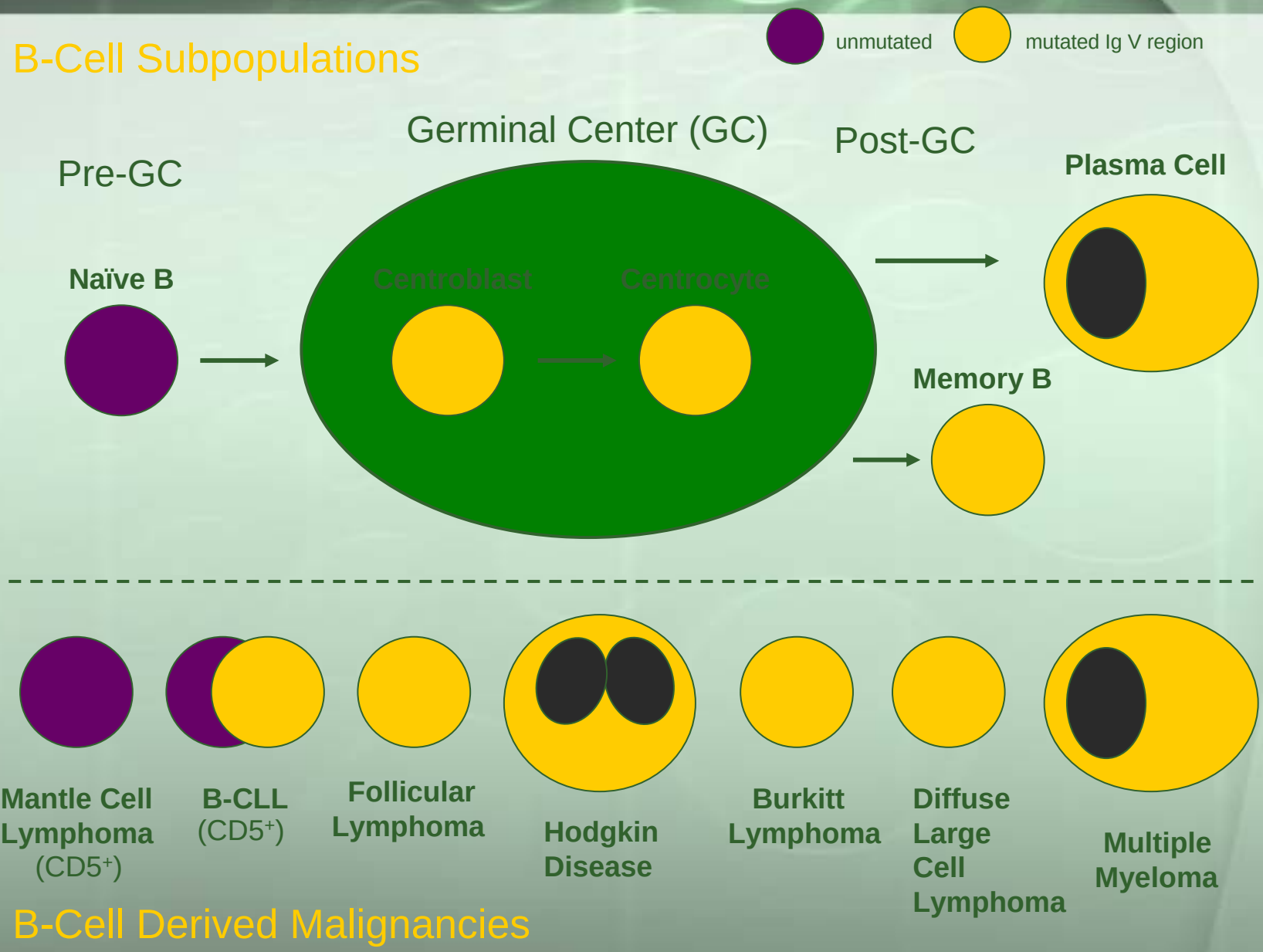


b

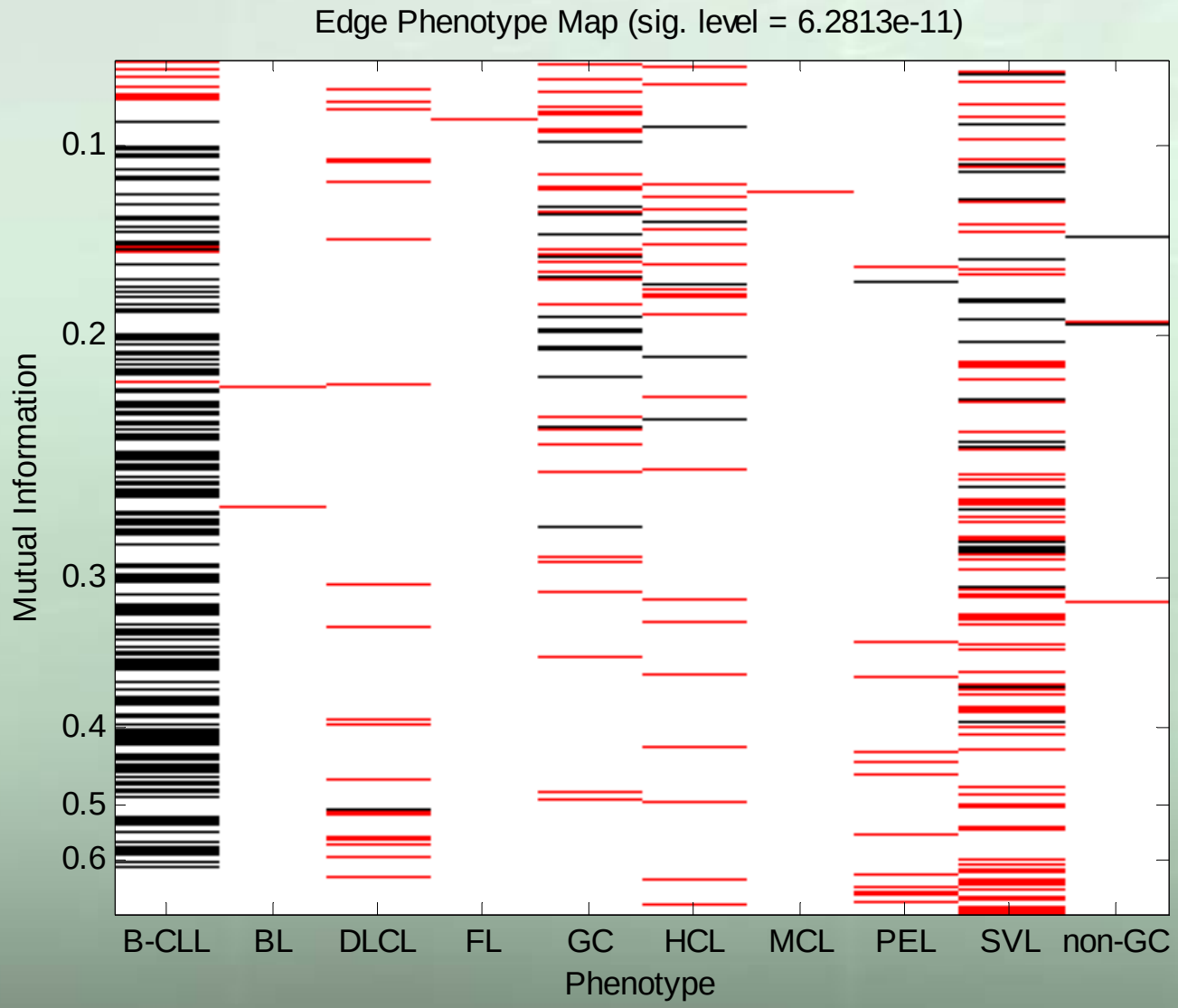
Session III: Data Generation and Validation

- **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:**
  - **B. Hayete**, J.J. Faith, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T.S. Gardner – Genome-scale mapping and global validation of the *E. coli* transcriptional network using a compendium of Affymetrix expression profiles
  - **Michael Samoilov** and Adam Arkin – Using Data Fusions and Biomolecular Modeling towards Improving the Results of Reverse Engineering in Biological Networks. The ENRICHed Approach.
  - A. Kundaje, D. Quigley, S. Lianoglou, X. Li, M. Arias, C. Wiggins, L. Zhang , and **C. Leslie** – Learning regulatory programs that accurately predict differential expression with MEDUSA

The germinal center



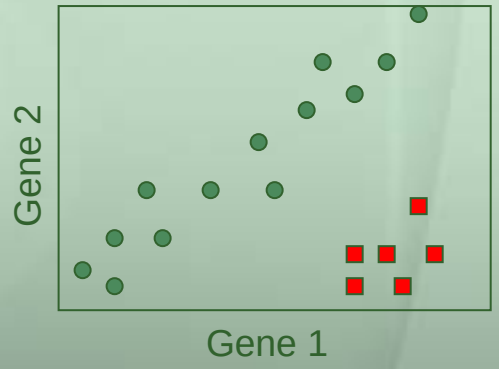
Edge Phenotype Map



- Changes occurred in 11,814 non-unique edges

MI (+) when phen. removed

MI (-) when phen. removed



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

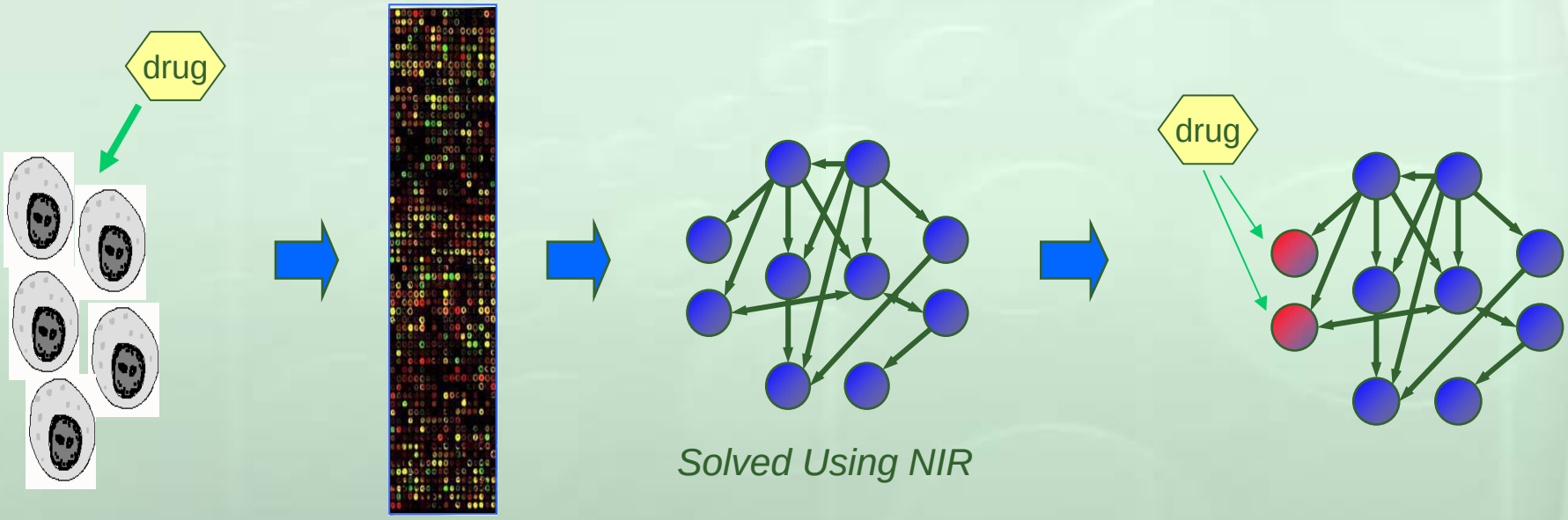
How do we use the Networks?

Treat cells with drug compound

Obtain expression profile

Filter profile using identified network

ID *direct* genetic targets of drug



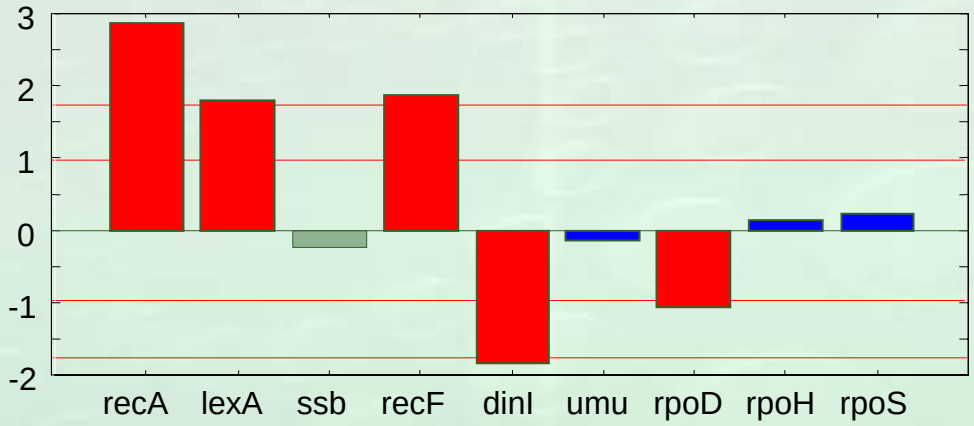
Solved Using NIR

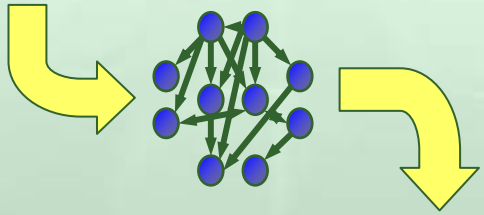
Profiling Drug Targets

J.J. Collins

Validation: recA/lexA Double Perturbation

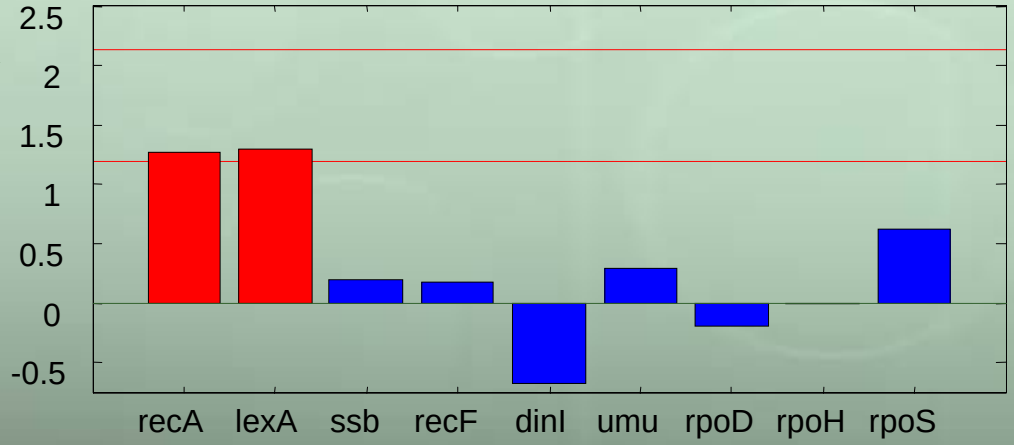
Expression changes Following recA/lexA double perturbation





Cannot distinguish affected genes using just expression data

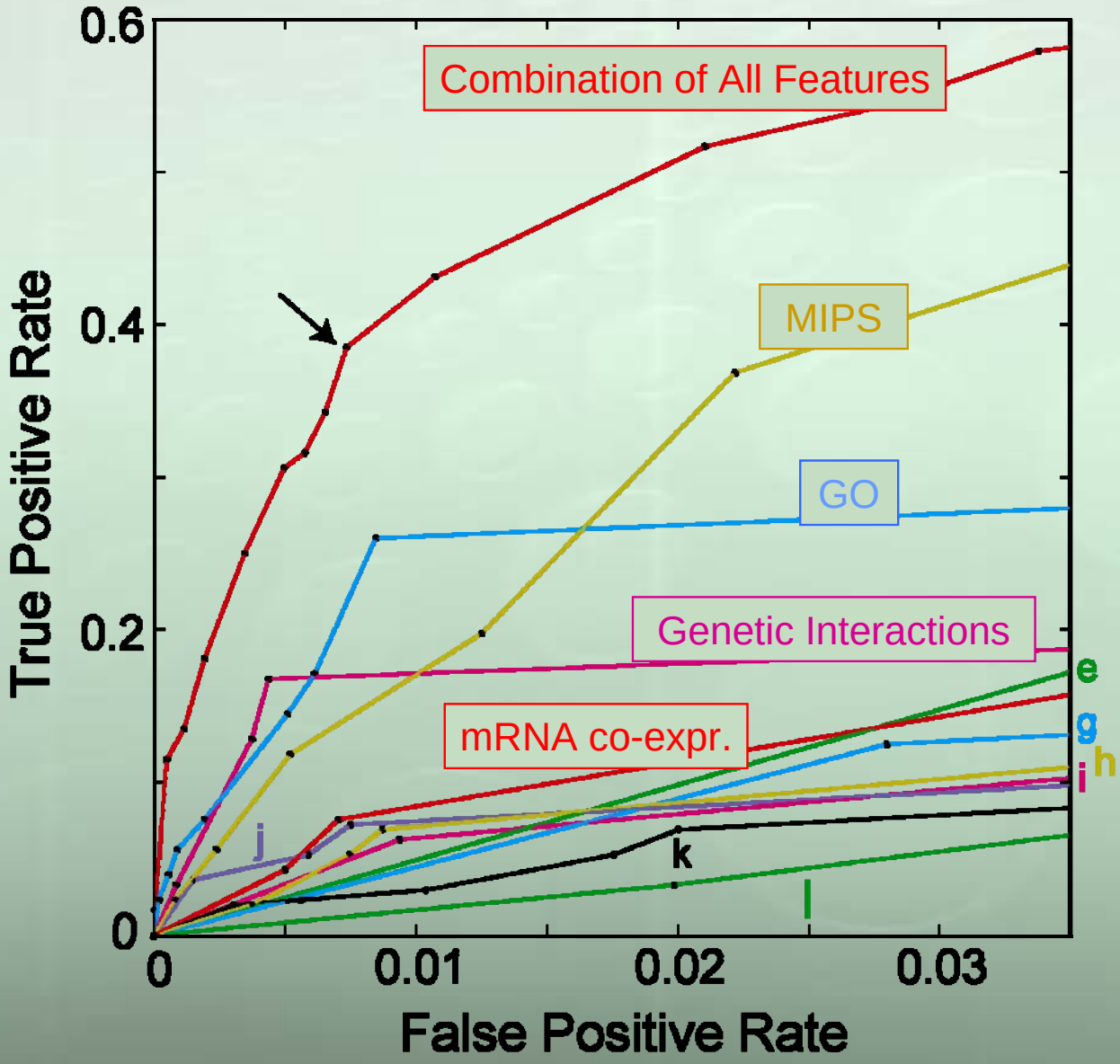
Predicted mediators: lexA and recA identified as perturbed genes by network model



Correct mediators of expression profile identified using NIR approach

J.J. Collins

Individual Features and their Integration for Yeast Membrane Protein Interaction Prediction

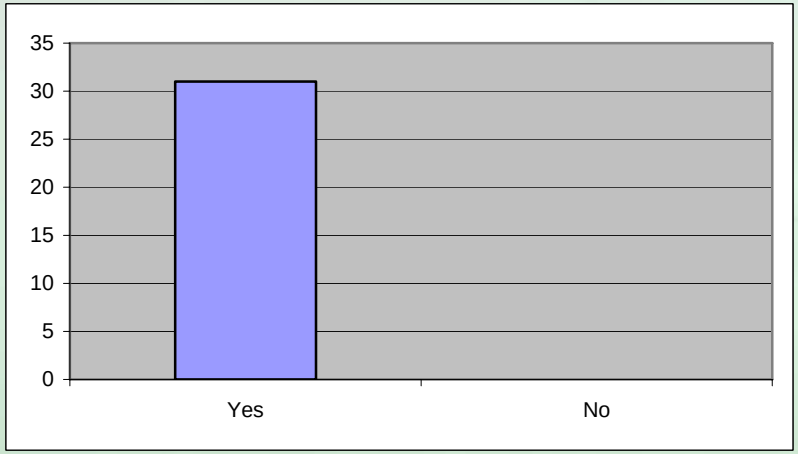


Mark Gerstein

DREAM Metrics

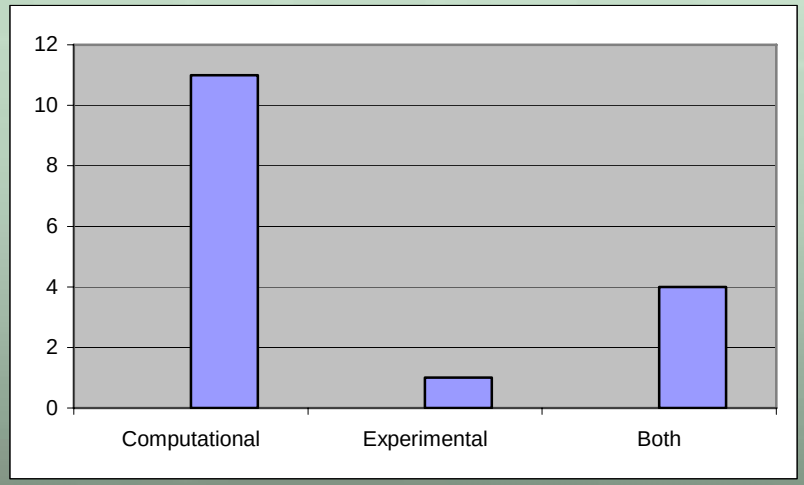
Questionnaire (30/120)

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

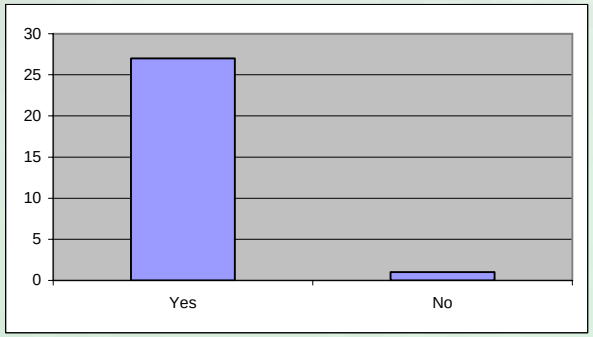


Computational	11
Experimental	1
Both	4
Total	16

Questionnaire

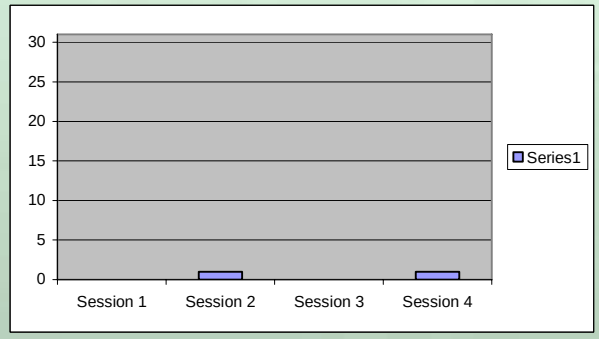
Q2 Based on the current session topics did you find that

[a] The sessions accurately represent and address the workshop goals



Yes	27
No	1
Total	28

[b] Some sessions are redundant and should be eliminated



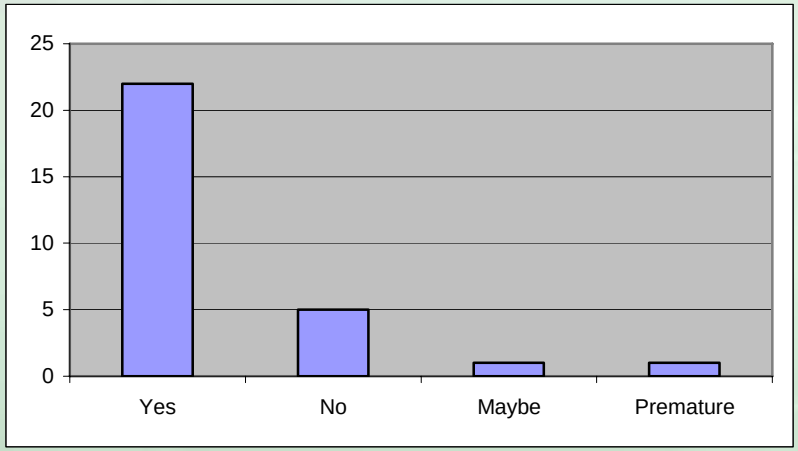
Session 1	0
Session 2	1
Session 3	0
Session 4	1
Total	2

[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
13. **More input from experimental biologists.**
14. **RE methods comparisons**
15. **Short poster presentations before poster sessions.**
16. **Competition.**

Questionnaire

Should the DREAM Workshop directly address the issue of comparing the performance of reverse engineering algorithm using blind data (similar to CASP)?

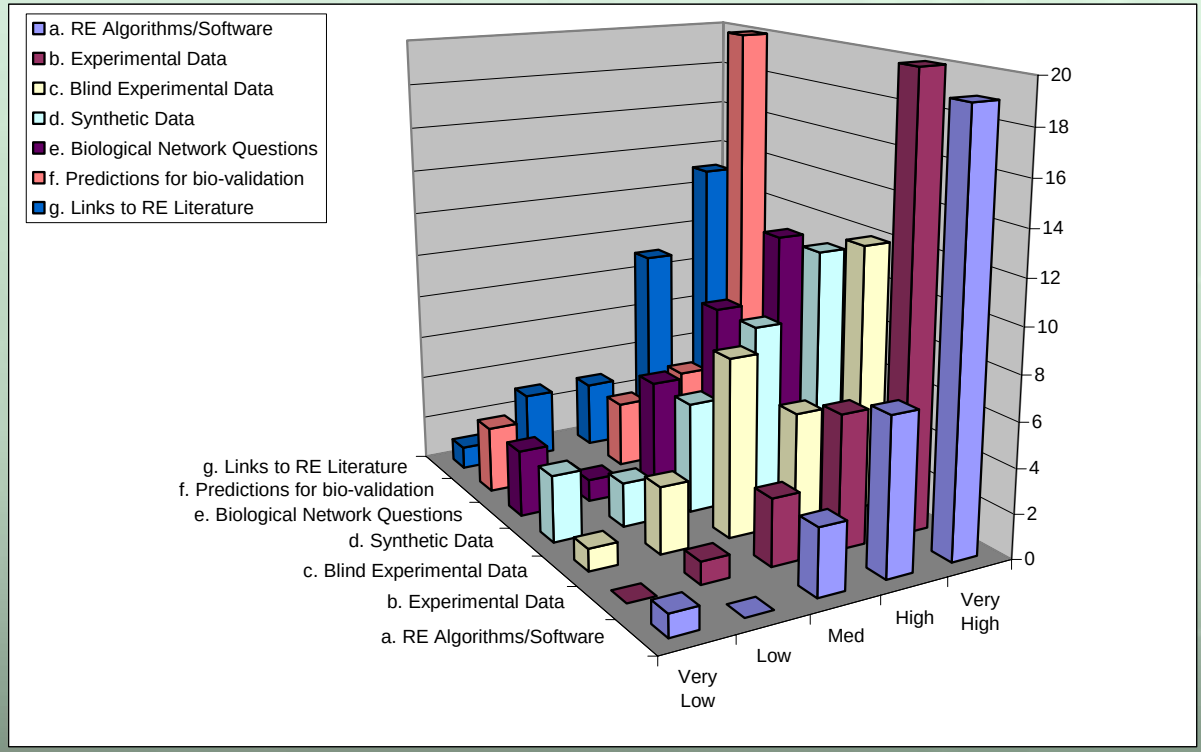


Yes	22
No	5
Maybe	1
Premature	1
Total	29

Question 4

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:

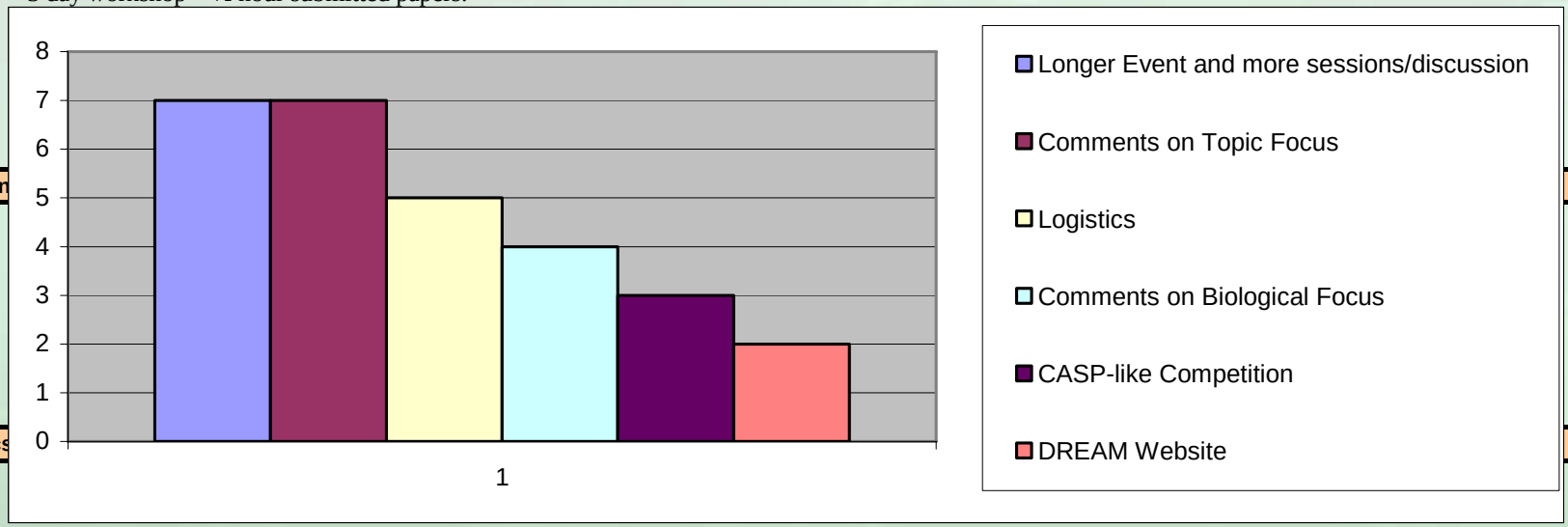
	Very Low	Low	Med	High	Very High	Total
a. RE Algorithms/Software	1	0	3	7	19	30
b. Experimental Data	0	1	3	6	20	30
c. Blind Experimental Data	1	3	8	5	12	29
d. Synthetic Data	3	2	5	8	11	29
e. Biological Network Questions	3	1	5	8	11	28
f. Predictions for bio-validation	3	0	3	4	20	30
g. Links to RE Literature	1	3	3	9	13	29



Questionnaire (Suggestions)

Longer Event and more sessions/discussion 7

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



Commen

Logistics

Common dinner.
A social mixer so that people can get to know each other better.
Integrate with GEO?

Comments on Biological Focus 4

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 3

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 assessmen

DREAM Website 2

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 an 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)