



Computational Approaches to Functional Genomics

Olga Troyanskaya
Assistant Professor
Lewis-Sigler Institute for Integrative Genomics &
Department of Computer Science
Princeton University




Laboratory of Bioinformatics & Functional Genomics

2




A primer: Molecular biology 101

3


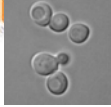


Cells are fundamental working units of all organisms




copyright Russell Knightley Media, www.rkm.com.au

4

Yeast are unicellular organisms


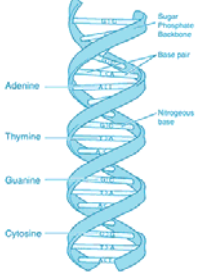


Humans are multi-cellular organisms

Understanding **how a cell works** is critical to understanding how the organism functions

5


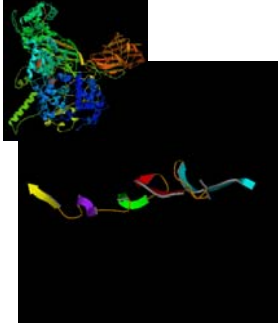
DNA

- Uses alphabet of 4 letters {ATCG}, called bases
- Encodes genetic information in triplet code
- Structure: a double helix

6

Proteins

- A sequence of amino acids (alphabet of 20)
- Each amino acid encoded by 3 DNA bases
- Perform most of the actual work in the cell
- Fold into complex 3D structure

Courtesy of the Zhou Laboratory, The State University of New York at Buffalo

7 **How does a cell function?**

GENES
Carries recipes for proteins

DNA
Information in DNA

PROTEIN MACHINE
Proteins act alone or in complexes to perform all cellular functions

PROTEINS

DNA is a sequence of bases (A, T, C, G)
TAT-CGT-AGT

Each 3 bases of DNA encode 1 amino acid

Proteins consist of amino acids, whose sequence is encoded in DNA
Tyr-Arg-Ser

Courtesy U.S. Department of Energy Genomes to Life program

8 **DNA-RNA-protein**

replication

DNA $\xrightarrow{\text{transcription}}$ mRNA $\xrightarrow{\text{translation}}$ protein

Polypeptide: val his leu thr

mRNA: G U G C A U C U G A C U

DNA: C A T G C A T A G C T A G C T A G C T

anti-sense strand (template) sense strand

Code for leucine Code for threonine

9 **Genes vs. proteins**

- Genes are units of inheritance
- They are static blueprints
- It's proteins (dynamic) that do most of the work
- The process of making mRNA, and then protein from a gene (or genes) is called GENE EXPRESSION
- It's the control of gene expression that causes most phenotypic differences in organisms

10 **Gene Regulatory Circuit**

If C then D

If B then NOT D

If A and B then D

If D then B

- Genes =? wires
- Motifs =? gates

11 **The "greatness" of genomics...**

- Biological systems are complex
- Many biological processes & diseases result from complex changes on molecular level
- Need to observe & model cellular processes on a systems level

High-throughput technologies have lead to an explosion of data in biology in hopes of understanding biological systems

12 **... And its "downfall"**

Explosion of functional genomic DATA

KNOWLEDGE of components and inter-relationships that lead to function

13 **Why have genomic data not been utilized fully?**

Challenges:

- Genomic data are noisy
- Genomic data are heterogeneous
- Coverage/accuracy varies by biological process

14 **Computation is a tool for functional genomics**

Computational methods (and targeted experiments) can greatly aid in extracting knowledge from biological data, but several challenges must be addressed:

Our approach:

- (1) Integrated analysis of diverse data
- (2) Probabilistic methods to battle noise in data
- (3) Integrating computation and experiments
- (4) Accessibility and usefulness to community (bringing experts into the analysis loop and feedback to experimental biology)

15 **Story #1: predicting function of unknown proteins**

16 **Predicting gene function using the Gene Ontology hierarchy**

- Could improve accuracy by enforcing Hierarchical consistency

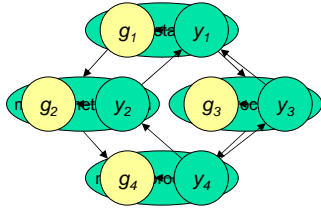
17 **Hierarchical Consistency**

<u>TRAINING</u>	<u>EVALUATION</u>
All genes	cell proliferation YES
All genes	cytokinesis NO
All genes	bud site selection YES

18 **Our Method**

- Individual classifiers for each class
 - Inconsistent predictions allowed
 - Any classification algorithm can be used
 - Parallel evaluation
- Bayesian combination of predictions
 - Inconsistencies resolved globally
 - Any inference algorithm can be used

A Bayesian Framework



Given predictions $g_1 \dots g_N \in \mathcal{Y}$, find true labels $y_1 \dots y_N \in \{0,1\}$ that maximize

$$P(y_1 \dots y_N | g_1 \dots g_N) = \alpha P(g_1 \dots g_N | y_1 \dots y_N) P(y_1 \dots y_N)$$

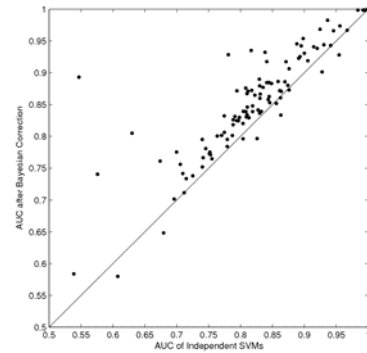
Data Types *(for Saccharomyces cerevisiae)*

- The Gene Ontology
 - 105 "meaningful" nodes selected
- Pairwise Interaction (*GRID*)
 - Affinity Precipitation
 - Affinity Chromatography
 - Two-Hybrid
 - Purified Complex
 - Biochemical Assay
 - Synthetic Lethality
 - Synthetic Rescue
 - Dosage Lethality
- Colocalization
 - *O'Shea*
 - *Curated Complexes* (152 features)
- Transcription Factor Binding Sites
 - *PROSPECT* (39 features)
- Microarrays (*SMD*)
 - Spellman et al., 1998
 - Gasch et al., 2000, 2001
 - Sudarsanam et al., 2000
 - Yoshimoto et al., 2002
 - Chu et al., 1998
 - Shakoury-Elizeh et al., 2003
 - Ogawa et al., 2000 (342 features)

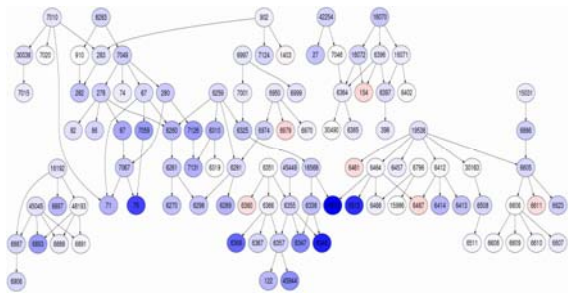
Does hierarchical consistency help?

- For each class, 10 linear SVMs trained by bootstrapping
- Median of unthresholded outputs used (bagging)
- Area under the ROC curve (AUC) for evaluation
- 93 of 105 nodes (86%) are improved by Bayesian correction.
 - Best $\Delta AUC = +0.346$ (+63% of old AUC)
 - Worst $\Delta AUC = -0.031$ (-3% of old AUC)
 - Average $\Delta AUC = +0.033$ (+4% of old AUC)

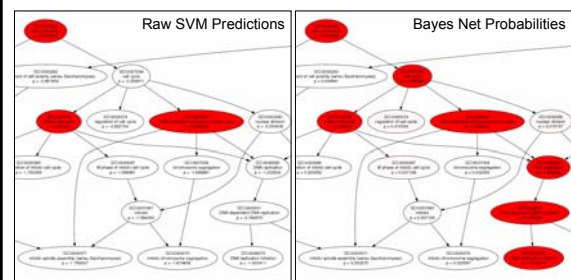
Most processes improve in accuracy (AUC Scatter Plot)



AUC Changes



Held-out Example: YNL261W



25

Verification: New Data

- GO since our April 2004 snapshot
 - 105 new annotations for 88 genes
- Predictions over the 88 genes on our data
 - Independent SVMs
 - 32% precision, 7% recall
 - Bayesian correction
 - 32% precision, 20% recall
 - 51% precision, 7 % recall

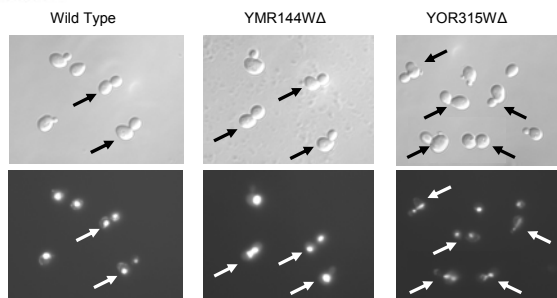
26

Predictions of novel proteins involved in mitosis

- Lab testing of some predictions for mitosis
 - YMR144W - "mitotic chromosome segregation"
 - Large-budded YMR144WΔ cells -> frequent nuclear defects
 - YOR315W - "mitotic spindle assembly"
 - Cells were fixed and
 - Large-budded YOR315WΔ cells -> frequent misaligned spindles (anti- α -tubulin antibody) and nuclear defects.
 - YMR299C - "mitotic cell cycle"
 - Lee et al. (2005) showed YMR299C protein that is part of a dynein pathway
- Independent SVMs miss these.

27

Experimental validation



28

Summary

- Using multiple information sources helps prediction accuracy
 - Multiple diverse data sources
 - Using gene ontology hierarchy
- Probabilistic and machine learning approaches can generate experimentally testable predictions
- Our hierarchical consistency approach increases accuracy and generates novel predictions

29

Story #2: predicting biological networks

Specific goal: building biological networks from experimental data

- Gene expression
- Physical protein-protein interactions
- Genetic interactions
- Cellular localization
- Sequence
- ...

Functional genomic DATA

Key ideas:

- Integration: combine information from all available sources in a robust way
- Understand/use information on biological context
- Building a practical system that directly involves biologists in the prediction process and can direct further experiments

bioPIXIE – a system for discovery & analysis of biological networks (in specific biological context)

*For *S. cerevisiae*: integrates data from ~6500 publications
 *Other organisms coming

31

System overview

bioPIXIE: Pathway Inference from eXperimental Interaction Evidence

32

Bayesian context-specific integration

$$P(FR_{i,j} | D_{1,i}, D_{2,i}, \dots, D_{n,i}, C_{i,j}) = \alpha P(D_{1,i} | FR_{i,j}, C_{i,j}) P(D_{2,i} | FR_{i,j}, C_{i,j}) \dots P(D_{n,i} | FR_{i,j}, C_{i,j})$$

where α is a normalization constant.

- We
- 174 observable nodes (datasets grouped by publication and by assay)
- Naive bayes
- (compares favorably against more sophisticated alternatives, e.g. TAN)
- Training set: GO biological process co-annotated proteins

33

System overview

bioPIXIE: Pathway Inference from eXperimental Interaction Evidence

Muys et al. Discovery of biological networks from diverse functional genomic data. *Genome Biology* (2005) 34

From integrated pairwise data to process-specific networks

YLL170W	YKHL66C	0.718544
YLR270W	YCR001C	0.654169
YEL009C	YHL011C	0.658789
YHR209W	YLR270W	0.658789
YHR209W	YAL096C	0.658789
YHR209W	YCR833C	0.658789
YHR209W	YLR346W	0.658789
YHR209W	YLR299W	0.658789
YHR209W	YKLO64W	0.658789
YHR209W	YCR269W	0.658789
YHR209W	YK9078W	0.658789
YHR209W	YHR104W	0.658789
YHR209W	YHR84C	0.658789
YHR209W	YCR436W	0.658789

→ use existing knowledge: Expert-driven discovery

35


Experts can drive the search process

- Rad23 entered with Rad4, Rad3, and Rad24
- The resulting network is enriched (22 of 44) for DNA repair proteins (GO:0006281)

36

Network recovery algorithm

- Query: Rad23 with proteasome components Pup1, Pre6, Rpn12
- Recovered network is enriched (36 of 44) for ubiquitin-dependent catabolism proteins and only contains 2 DNA repair proteins (Rad6 and Rad23).


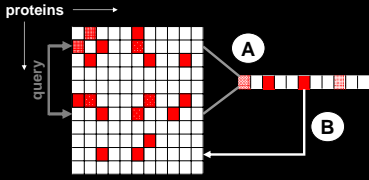


37

Network recovery algorithm

Basic idea: local search in the PPI network centered at the query

Which proteins should we extract as a single, functionally coherent group?

A: determine a "characteristic" interaction profile for the query set
 B: search the remaining set of proteins for the closest matches to the characteristic profile

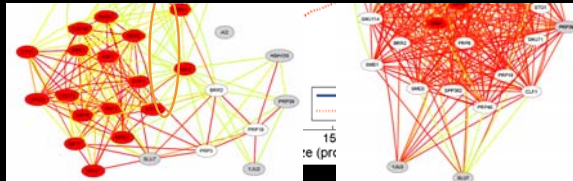
38

Evaluation: the importance of biological context

RNA splicing: same as protein-protein interactions

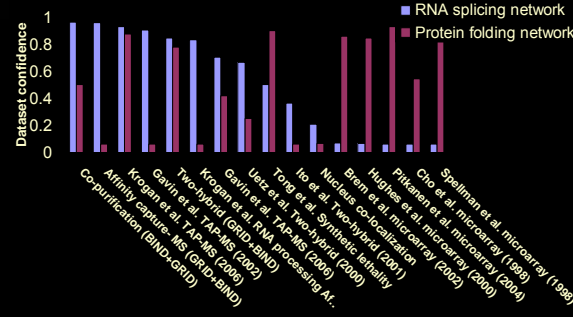
Global network: 22 FP (False Positives)

Context-specific integration improves 44/53 evaluated bio. process GO terms an average of 25%



15
e (pr)

RNA splicing dataset relevance



Dataset confidence

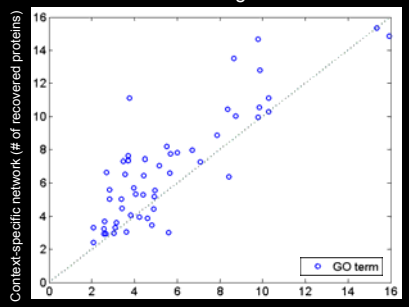
RNA splicing network
Protein folding network

(16 of 174 input datasets)

40

A consistent improvement

- Context-specific integration improves 44/53 evaluated bio. process GO terms an average of 25%



Context-specific network (# of recovered proteins)

Global network (# of recovered proteins)

GO term

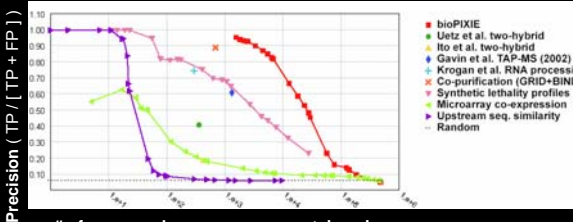
10-protein query; each point: average of 50 trials

41

General network recovery evaluation

- How accurately can we recover known network components?
- How much does integration of diverse data help?

Evaluation: measure how often observed data connects functionally related proteins (e.g. shared GO annotations)

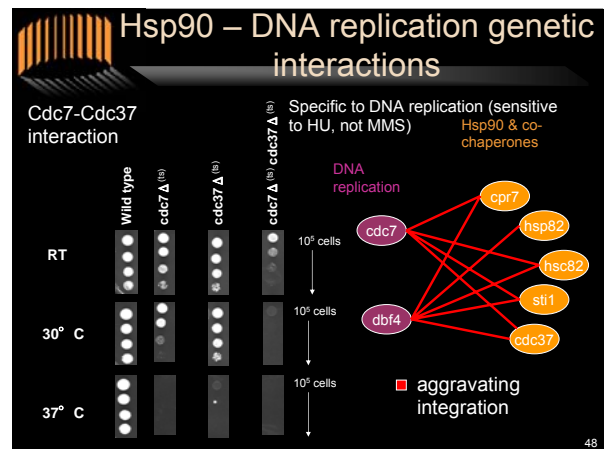
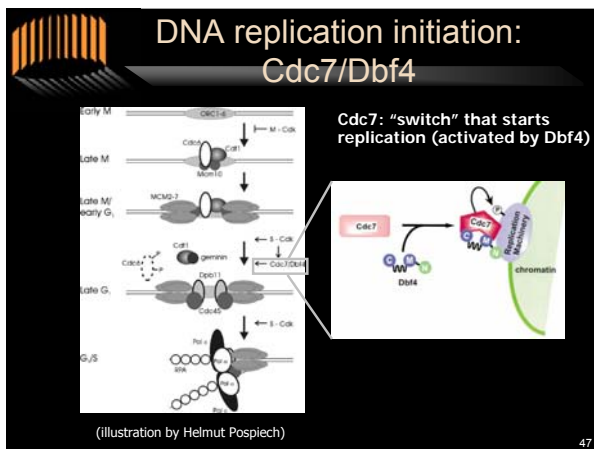
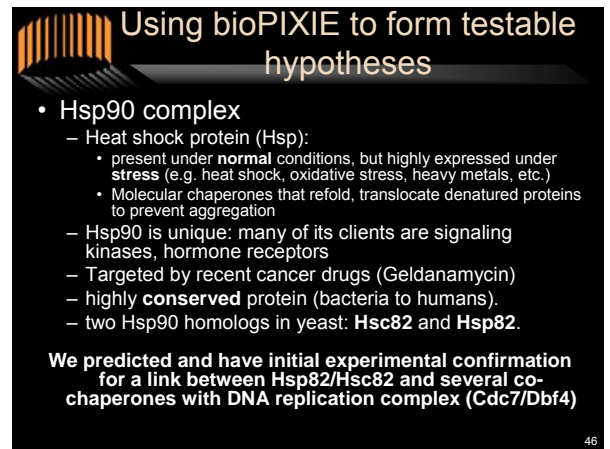
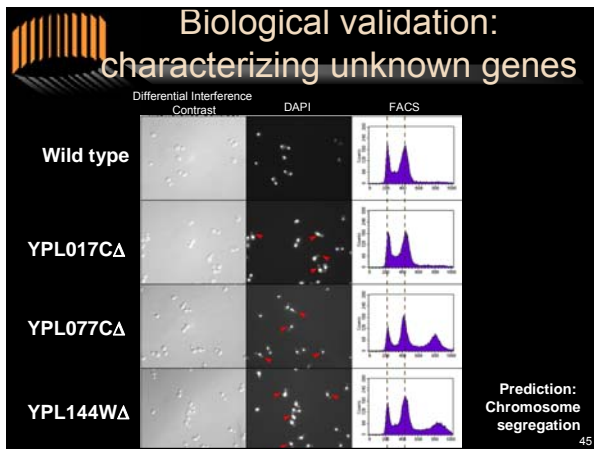
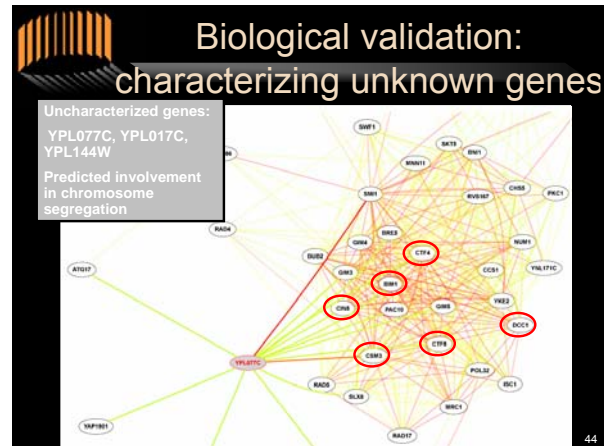
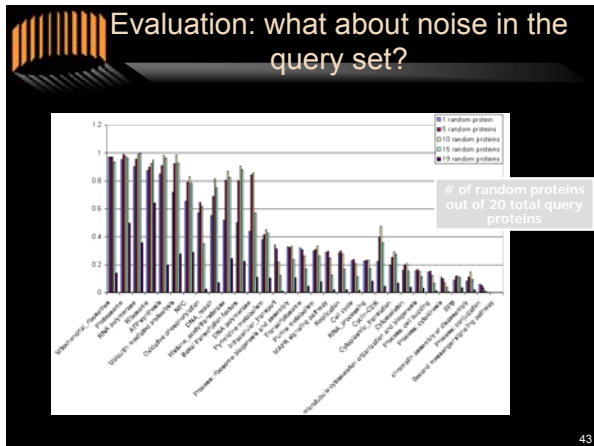


Precision (TP / [TP + FP])

of recovered same-process protein pairs

(8 of 174 input datasets)

42





A (possible) bigger picture



So what?

- Analysis of integrated genomic data can direct generation of testable, non-trivial hypotheses
- Important to integrate data and to take into account biological process