Finding missing genes in biological pathways

CS Workshop

Ron Shamir, Didi Amar
Fall 2011
Outline

0. Workshop goals and schedule
1. Basic biology, biotechnology
2. DNA chips & microarrays
3. Gene networks
0. Workshop goals and schedule
Motivation

• Biological processes in the cell work in interconnected networks/pathways
• Understanding the pathways is critical to understanding disease processes, modifying/improving pathways
• Knowledge on pathways is incomplete
• We want to find the missing pieces in a pathway
Motivation (2)

• Modify an organism to produce a specific compound.

• Scientists in UCLA engineered cyanobacteria to convert greenhouse gas into liquid fuel.

• One must understand very well the relevant pathway that will be engineered, and exactly which genes take part in the pathway.
Motivation (3)

- Carotenoids are compounds used by plants for production of vitamin A.
- Plants are the only source for vitamin A.
- 250-500,000 malnourished children in the developing world go blind each year from a deficiency of vitamin A.
- ~half die within a year.
- Solution: enhance the level of Carotenoids in 3rd world staple crops.
- Need comprehensive understanding of the Carotenoid biosynthesis pathway.
Given a partially known target pathway, rank all candidate genes by how likely they are to belong to the pathway.

How? integrate data from several bio info sources
Input/Output

**Input:**
1. **The pathway:** A set $P$ of genes that are known to take part in a specific process or pathway
2. **Supporting data:** Obligatory and optional data sets that contain descriptors on genes and their relations.
3. **The candidates:** A set $S$ of genes (disjoint from $P$) to be ranked.

**Output:** Ranking of the candidates. A ranked list of the genes in $S$, with a score for each gene.
Training and Testing Data

In each stage a **training set** of pathways will be given for developing the methods.

After delivery of the results and the software at the end of the phase, the software will be run on an additional **test set** of new pathways that were not used in the training.
Accuracy - self rank test
Administrata

• Project should be written in Java
• Project can be done in pairs
• Pairs/singletons – please inform Didi by next Tuesday if you wish to take the workshop.
• Meetings – introductory meetings this week, next, then individually with groups (on the workshop time slot)
• Questions? Contact Didi or me
• Website:
  www.cs.tau.ac.il/~rshamir/workshop/11
Schedule

• **Stage 1:** Prediction using only one type of data set. due 6/12/11.

• **Stage 2:** Prediction using two types of data. due 10/1/12.

• **Stage 3:** A final prediction system using at least three different types of data + a final report. due 15/3/12.

• **Finale:** presentations by all groups, ~20/3/12
Grading

• 15% for stage 1 (5%/10% for accuracy on the training/test pathways)
• 25% for stage 2 (10%/15%)
• 30% for stage 3 (15%/15%)
• 20% for the implementation (modularity, clarity, documentation, efficiency)
• 10% for the final report and presentation
• Bonus/penalty
  ➢ +5% to the most accurate group at each stage
  ➢ +5% bonus for use of optional data types (provided or suggested by the group)
  ➢ -5% penalty for not meeting any stage deadline
1. Basic Biology & Biotechnology
DNA → RNA → protein

The hard disk
One program
Its output

transcription
translation
From Gene to Protein

The Human Genome: numbers

- 23 pairs of chromosomes
- ~3,000,000,000 bases
- ~20,000 protein-coding genes
- *Gene length:* 1000-3000 bases, spanning 30-40,000 bases
- up to 1,000,000 protein variants
2. DNA Chips & Microarrays
DNA chips / Microarrays

• Simultaneous measurement of expression levels of all genes.
• Global view of cellular processes.
• > 600,000 profiles available in GEO
Oligonucleotide Arrays

GeneChip expression analysis probe array

Each probe cell contains millions of copies of a specific oligonucleotide probe

Biotinylated RNA target from experimental sample

Streptavidin-phycoerythrin conjugate

Image of hybridized probe array
The Raw Data

Entries of the Raw Data matrix:

• Ratio values
• Absolute values
• Distributions...

• Row = gene’s expression pattern / fingerprint vector

• Column = experiment/condition’s profile

Expression levels,
“Raw Data”
Computational Challenges

- **Normalization**: How does one best normalize thousands of signals from same/different conditions/experiments?
- **Identify differentially expressed genes** between experiments
- **Clustering**: Partition genes into subsets that manifest similar exp. pattern
- **Biclustering**: Find subsets of genes and conditions that manifest a common exp. sub-pattern
Expression profiles via next-gen sequencing

Roche 454

Illumina Genome Analyzer II
4. Gene Networks
Biological Networks

www.biology.iupui.edu/research/bard/
http://endo.edoc.com
www.genome.ad.jp/kegg/kegg2.html
http://www.grt.kyushu-u.ac.jp/spad/pathway/tnf.html
Protein-protein interactions (PPIs)

- Low throughput measurements: accurate, scarce
- High throughput: more abundant, noisy
- Large, readily available resource
In fact, many resources...
The hairball syndrome
Gene expression

proB \rightarrow expression \rightarrow \text{gamma-glutamyl kinase} \rightarrow catalyzes \rightarrow 2.7.2.11

\text{Substrate} \rightarrow \text{glutamate} \rightarrow \text{ATP} \rightarrow \text{ADP} \rightarrow \text{gamma-glutamyl phosphate}

\text{gene}

Positive interaction

Protein

EC (reaction) number

1.5.1.2

compound

http://www.ebi.ac.uk/research/pfmp
Metabolic Pathway: Proline Biosynthesis

- **proB**
  - Expression
  - Catalyses
  - Inhibits

- **proA**
  - Expression
  - Catalyses

- **proC**
  - Expression
  - Catalyses

**Key Enzymes and Reactions**

1. **gamma-glutamyl kinase**
   - Inputs: Glutamate, ATP, ADP
   - Output: gamma-glutamyl phosphate

2. **gamma-glutamylphosphate reductase**
   - Inputs: gamma-glutamyl phosphate, NADPH, H+
   - Outputs: Glutamate gamma-semialdehyde, NADP, Pi

3. **Spontaneous**
   - Inputs: Glutamate gamma-semialdehyde, H2O
   - Outputs: 1-pyrroline-5-carboxylate

4. **1-pyrroline-5-carboxylate reductase**
   - Inputs: 1-pyrroline-carboxylate, NADPH, NADP
   - Outputs: Proline
Methionine Biosynthesis in E.coli
Signal transduction

Extracellular space:
- TNFalpha activates

Cell membrane:
- TNFR1 binds
- TRADD binds
- TRAF2 binds
- Activates

Cytoplasm:
- IKK complex activates
- IKKalpha activates
- gamma subunit
- IKKbeta
- IKBAP
- Activates

Nuclear membrane:
- NFkB/IkB complex
- NFkB/P-IkB complex
- Phosphorylation
- Dissociates
- Ubiquitination
- Degradation
- Translocates
- Regulates transcription
- NFkB

Nucleus:
- Activates

http://www.ebi.ac.uk/research/pfmp
Genetic network controlling early development of sea urchin endomesoderm.

[Image of a genetic network diagram with various genes and proteins connected by arrows and regulatory interactions.]

www.its.caltech.edu/~mirsky/endomeso.htm
Goal: “reverse engineering” of a network

Modify a known network to cure disease / improve its function / save the world

But first - find all the pieces of the network of interest!
FIN