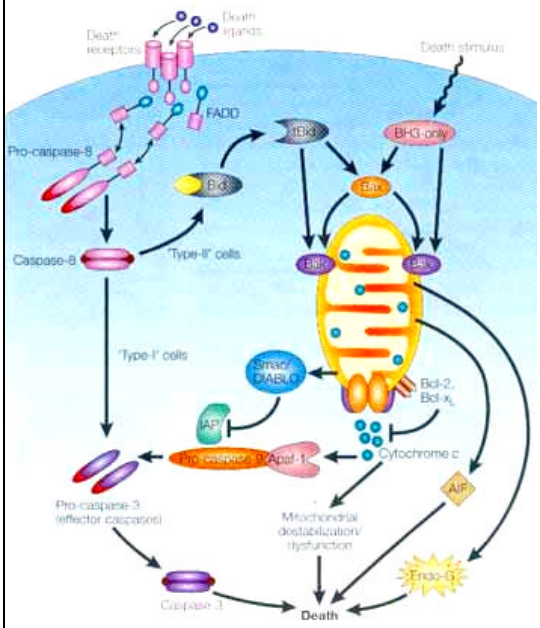
Children's Hospital Informatics Program | Harvard Medical School

Adding the Arrows: Beyond Pair-wise Comparisons

Isaac S. Kohane

Children's Hospital Informatics Program | Harvard Medical School



What we are hoping to re-create

Translocation

Induction

Inhibition

…

An engineer, a physicist, a mathematician, a computer scientist, and a statistician are on a train heading north, and had just crossed the border into Scotland. They look out the window and see a black sheep for the first time.

The engineer exclaims, “Look! Scottish sheep are black!”

The physicist yells, “No, no. Some Scottish sheep are black.”

The mathematician looks irritated and says, “There is at least one field, containing at least one sheep, of which at least one side is black.”

The computer scientist says, “Oh, no, a special case!”

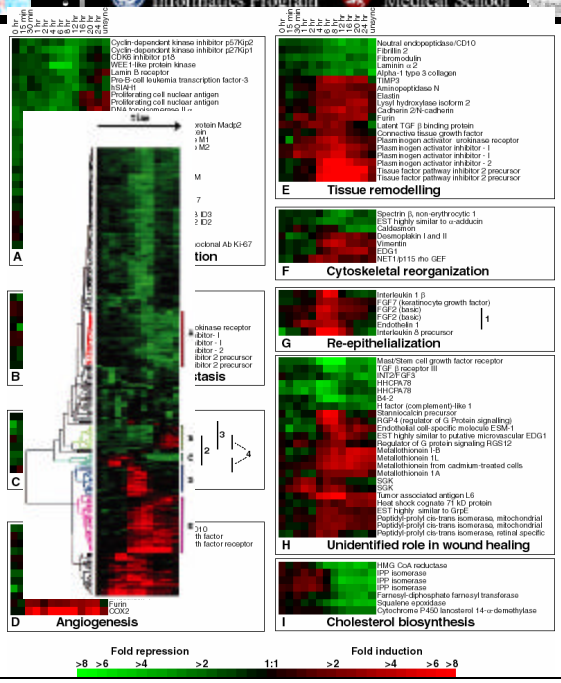
Finally, the statistician says, “It is not statistically significant!”

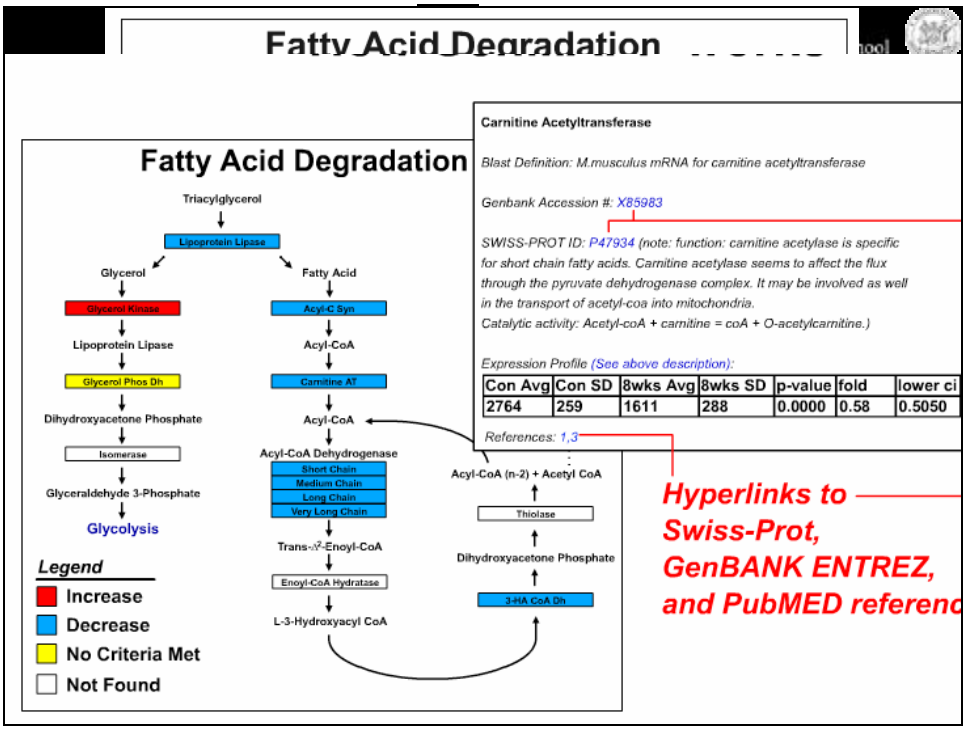
Children's Hospital Informatics Program | Harvard Medical School

Pathways in Dendrograms?

Dendrograms

Comparison on a cDNA analysis of fibroblast response to serum (Iyer et al., 1999)





Causality *ab initio*

- If dendrograms do not give arrows and…
- If hand-coded pathway maps limit what is illuminated

- **Can we get closer to causal pathways/maps from data alone?**

Time and causality

- **Three Conditions for Inferring Cause:**
  - ✓ **Covariation** the cause and effect have to be related.
    - ☞"Guilt by association"
  - ✓ **Time precedence of the cause** , the cause had to precede the effect in time
  - ✓ **No plausible alternative explanation of the effect**

John Stuart Mills (System of Logic, 1843).

Children's Hospital Informatics Program | Harvard Medical School

Time and Functional Genomics

- Slopes vs. Statics
- Signal Processing Approach
- Making order and time matter in clustering

Time and causality

- **Three Conditions for Inferring Cause:**
  - ✓ **Covariation** the cause and effect have to be related.
  - ✓ **Time precedence of the cause** , the cause had to precede the effect in time
  - ✓ **No plausible alternative explanation of the effect**

John Stuart Mills (System of Logic, 1843).

Children's Hospital Informatics Program Harvard Medical School

Interim Summary

- Traditional biological knowledge sources (a.k.a. pathways and literature) are useful transducers of noisy data.
- Time series experiments will bring us closer to mechanistic underpinnings.
- Systematically explore alternate models explaining data
  - ✓ Probabilistic framework is useful to this end.

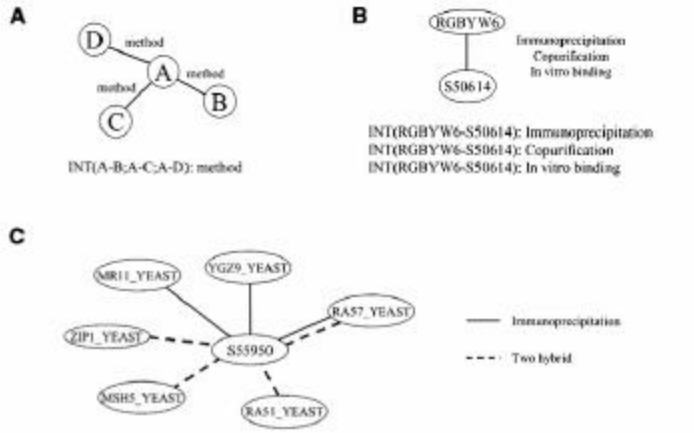


Figure 1. Three illustrations of protein interactions and their descriptions by machine readable text of the sort that would automate database entry. (**A**) A set of four proteins (A, B, C and D) have been studied and shown to interact by a given method. The condensed text is shown below. (**B**) Example of the SWI6–SWI4 interactions observed by Siegmund and Nasmyth (15) using immunoprecipitation, copurification and *in vitro* binding. In this case the protein PIR codes are used RGBYW6 for SWI6p and S50614 for SWI4p. (**C**) Example of the Zip3 interactions demonstrated by two-hybrid screen and immunoprecipitation by Agarwal and Roeder (16). In this case S55950 (Zip3p) can be immunoprecipitated with RA57p (RA57_YEAST), MR11p (MR11_YEAST) and Zip2p (YGZ9_YEAST). Two-hybrid screen was performed and showed interactions of Zip3p with ZIP1p, MSH5p, RAD57p, RAD51p.

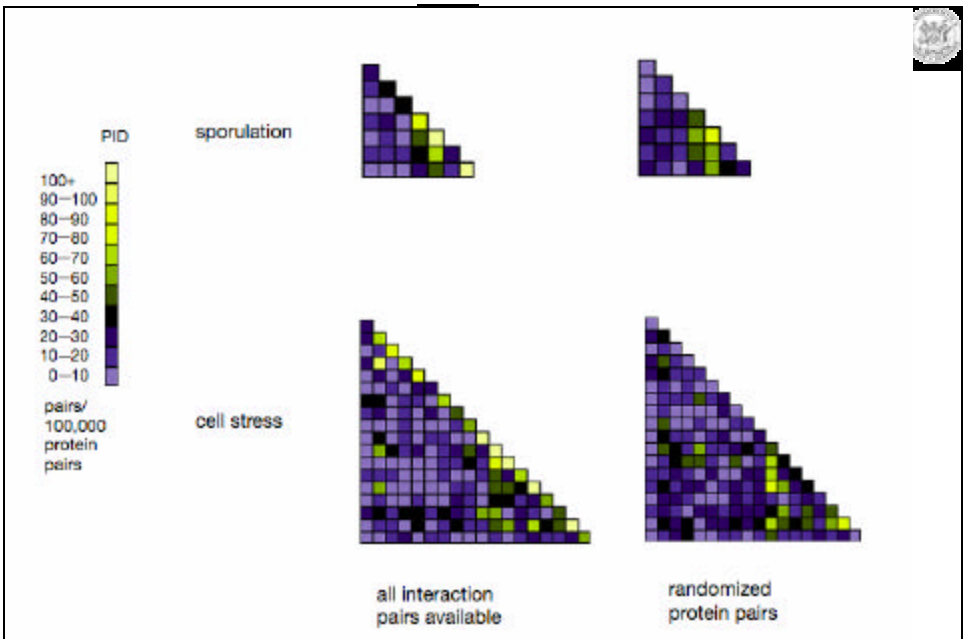
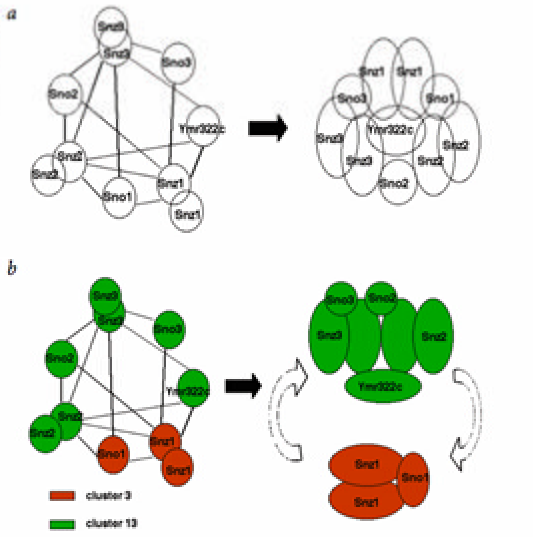


Fig. 3 Transcriptome–interactome correlation maps for sporulation and cell-stress data sets (see legend of Fig. 2). We constructed transcriptome–interactome correlation maps using sporulation (top) and cell-stress (bottom) profiling clusters and protein interaction data currently available.

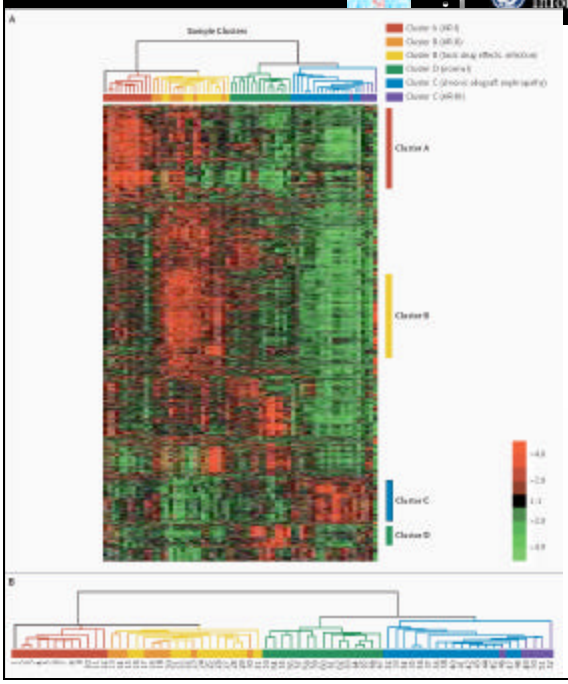
The leverage of prior knowledge

Fig. 4 Improved model from the integration of transcriptome and interactome data. **a**, Interaction network obtained from the combined data set of protein–protein interactions. Circles represent proteins and lines represent two-hybrid interactions. **b**, Improved model obtained by taking into account expression profiles.



Children's Hospital Informatics Program | Harvard Medical School

Let's Revisit Causality at Genomic Level



The New Histopathology

- New Diagnosis?
- B-cell signature (CD20+) in Cluster A (and others)

ORIGINAL ARTICLE

Molecular Heterogeneity in Acute Renal Allograft Rejection Identified by DNA Microarray Profiling

Minnie Sarwal, M.D., Ph.D., Mei-Sze Chua, Ph.D., Neeraja Kambham, M.D., Sze-Chuan Hsieh, M.S., Thomas Satterwhite, B.S., Marilyn Masek, M.S., and Oscar Salvatierra, Jr., M.D.

Children's Hospital Informatics Program | Harvard Medical School

Table 3. Clinical Correlates of CD20 Status in Renal-Biopsy Samples from Patients with Acute Rejection.

Variable	Retrospective Series of Biopsy Samples (N=31)	Biopsy Samples Included in the Microarray Analysis (N=20)
CD20+ on staining — no./total no. (%)	9/31 (29)	9/20 (45)
Graft loss		
In patients with CD20+ sample — no./total no. (%)	7/9 (78)	8/9 (89)
In patients with CD20– sample — no./total no. (%)	9/22 (41)	1/11 (9)
P value	0.11	<0.001
Glucocorticoid resistance		
In patients with CD20+ sample — no./total no. (%)	8/9 (89)	4/9 (44)*
In patients with CD20– sample — no./total no. (%)	1/22 (5)	1/11 (9)
P value	<0.001	0.01

Finding

- New Prognosis
- New Therapeutic Opportunity
- New Research Opportunity

Where to go?

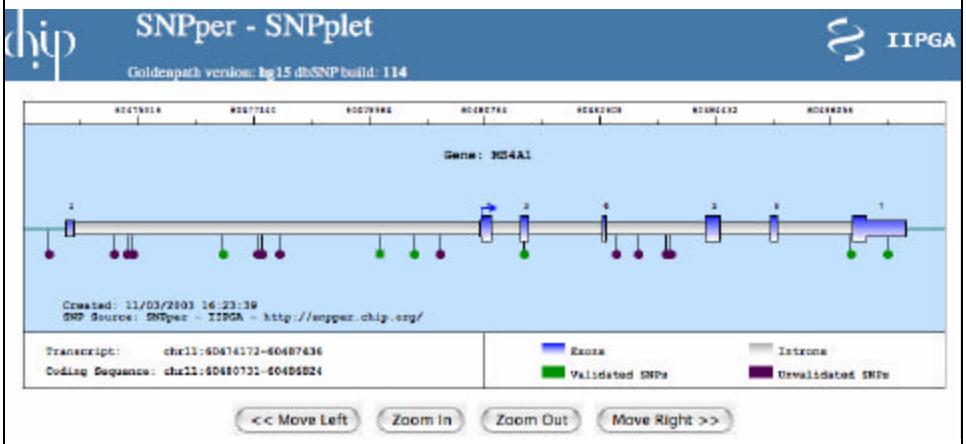
ACT ⟶ ATT

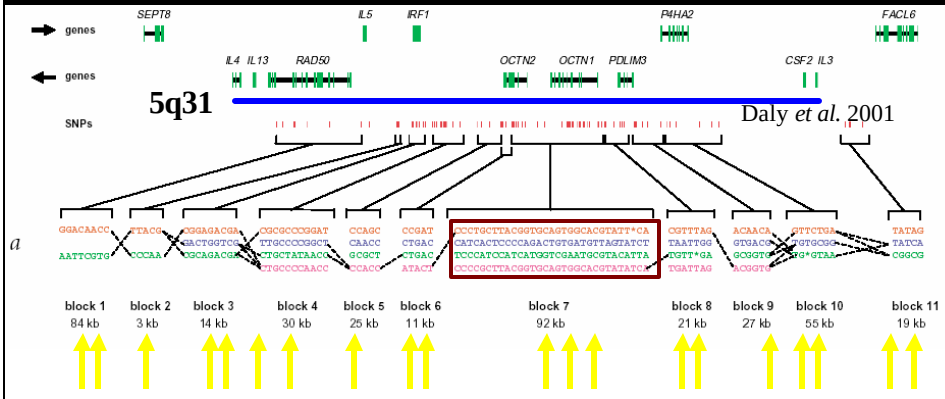
- How can we study how variants in genes (e.g. CD20) influence disease?
  - ✓What the the variants?
- How can we perform association studies more cost-effectively?

Children's Hospital Informatics Program
Harvard Medical School

CD20

- Which variant to study?





Haplotype Tagging SNP's (htSNP's)

chip Children's Hospital Informatics Program Harvard Medical School

Typical Approach

- Probabilistic framework
  - ✓ Probabilistic dependencies
  - ✓ "80% of variation" associated with the following htSNP's
- OR
  - ✓ Enumerate all possible combinations of SNP's that can account for the haplotypes and assess coverage.
- OR
  - ✓ A new deterministic algorithm to determine the provably optimal minima set of htSNP's
  - ✓ Best Enumeration of SNP Tags– BEST
    - ☞ http://www.chip.org/home/resources.cgi

BEST(Best Enumeration of SNP Tags)

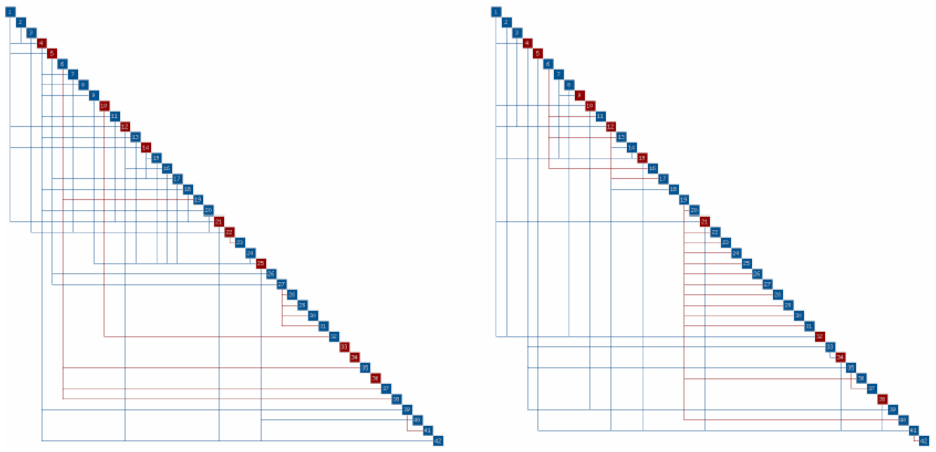
- Deterministic
- Fast (takes < 1 second for any of the available data sets)
- Provably optimal

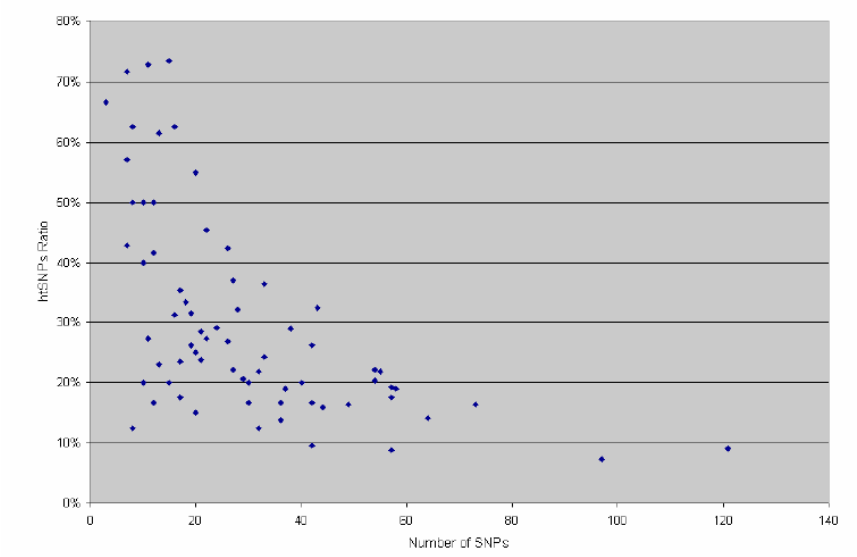
```
procedure BEST(S)
  H←∅;
  for i ≤ |S| do
    if not DERIVABLE(S_i,(S − S_i)) then H←H ∪ S_i;
  end
  R←∅;
  for i ≤ |S| do
    if S_i ∉ H and not DERIVABLE(S_i,H) then R←R ∪ S_i;
  end
  NB←∅; B←{H};      ;;; Note that B is a set of sets now initialized as the single set H.
  while ∃B_i ∈ B||DERIVED(B_i)| ≤ |C| do
    for i ≤ |C| do
      for j ≤ |B| do
        if R_i ∉ B_j then NB←(C_i ∪ B_j) ∪ NB;
      end
    end
    B←NB_i|max DERIVED(NB_i); NB←∅;
  end
  if ∃b_i|∃B_i ∈ B| DERIVABLE(b_i,(B_i \ b_i)) then
    BEST(B_i) ∀B_i ∈ B
  else return B_i ∈ B|min|B_i|
end
```

Ramoni, PNAS, 03

Children's Hospital Informatics Program | Harvard Medical School

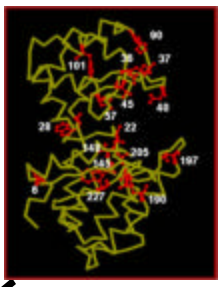
The Solution



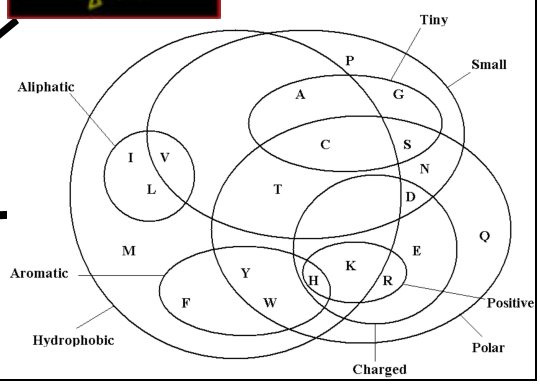


European Americans

Children's Hospital Informatics Program
Harvard Medical School



Direct Digital Assessment of Variants



Exploit Phylogenetic information

An example of PFAM alignments:

```
143Z_HUMAN   KNELVQKAKLAEQAERYDDMAACMKSVTEQGA
143Z_MOUSE   KNELVQKAKLAEQAERYDDMAACMKSVTEQGA
143Z_SHEEP   KNELVQKAKLAEQAERYDDMAACMKSVTEQGA
143B_MOUSE   KSELVQKAKLAEQAERYDDMAAAMKAVTEQGH
143B_RAT     KSELVQKAKLAEQAERYDDMAAAMKAVTEQGH
143B_HUMAN   KSELVQKAKLAEQAERYDDMAAAMKAVTEQGH
143B_BOVIN   KSELVQKAKLAEQAERYDDMAAAMKAVTEQGH
1433_XENLA   -------AKLSEQAERYDDMAASMKAVTELGA
O57469       KNELVQKAKLAEQAERYDDMAACMKRVTEEGG
143Z_XENLA   KNELVQKAKLAEQAERYDDMAACMKRVTEEGG
1434_CAEEL   KEELVNRAKLAEQAERYDDMAASMKKVTELGA
```

- How conserved is the position where the SNP/mutation occurs in the protein domain: **positional entropy**
  - How often a specific allele appears in the cross-species alignment **f_A and f_B, (f_A - f_B), or f_B/f_A**

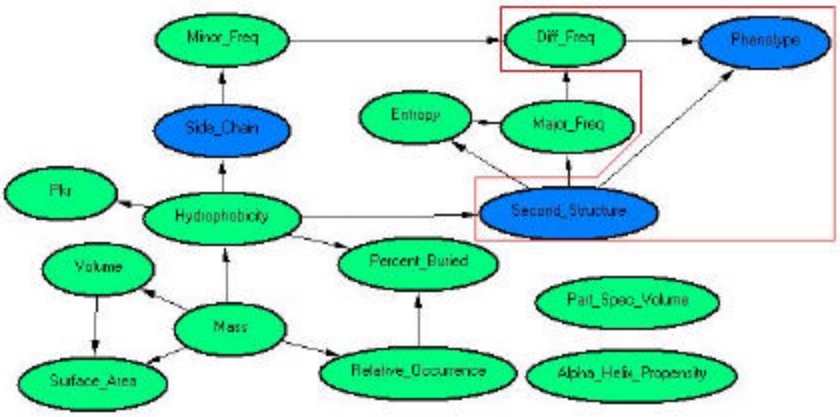
Children's Hospital Informatics Program | Harvard Medical School

Microbial Mutation Datasets

- **Large missense mutation studies:** up to 13 different mutations at each position, 6059 in total.
- **Phenotypic screen:** biochemical measurements, categorized relative to WT function.

Study	Species	Target	Method	total # mutations	# mutatoins /position	Phenotypic Screen
lacI (Suckow et al, 1996)	**E. coli**	entire protein	Nonsense suppression	3744	12-13	beta-galactosidase assay (fold of inhibition)
T4 lysozyme (Rennell et al, 1991)	**Bacteriophage**	entire protein	Nonsense suppression	2015	12-13	plaque-forming ability
HIV protease (Loeb et al., 1989)	**HIV-1**	entire protein	saturation mutagenesis	336	1-10	western blot assay
f1 gene V (Zabin et al, 1991)	**Bacteriophage**	entire protein	saturation mutagenesis	313	1-11	cell growth inhibition

Automatic (Bayesian) Discovery of Predictor of Phenotype Changes



Children's Hospital Informatics Program | Harvard Medical School

Can we make the leap from microbe to human?

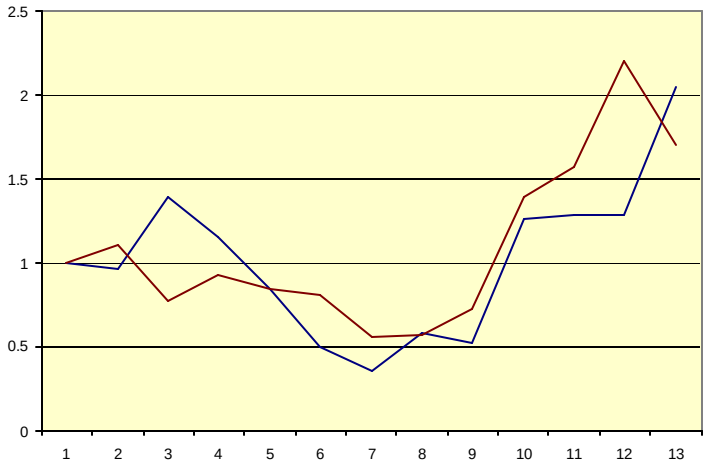
- 200 OMIM genes: variants annotated to cause disease
- Same genes but no disease associated with variants (yet)
- Specificity: 70-80%, Sensitivity 50-60%
  - ✓ That is: if the variant is identified as "pathogenic", 4 out of 5 times it is disease-causing
- So, why don't we just apply the algorithm to CD20?

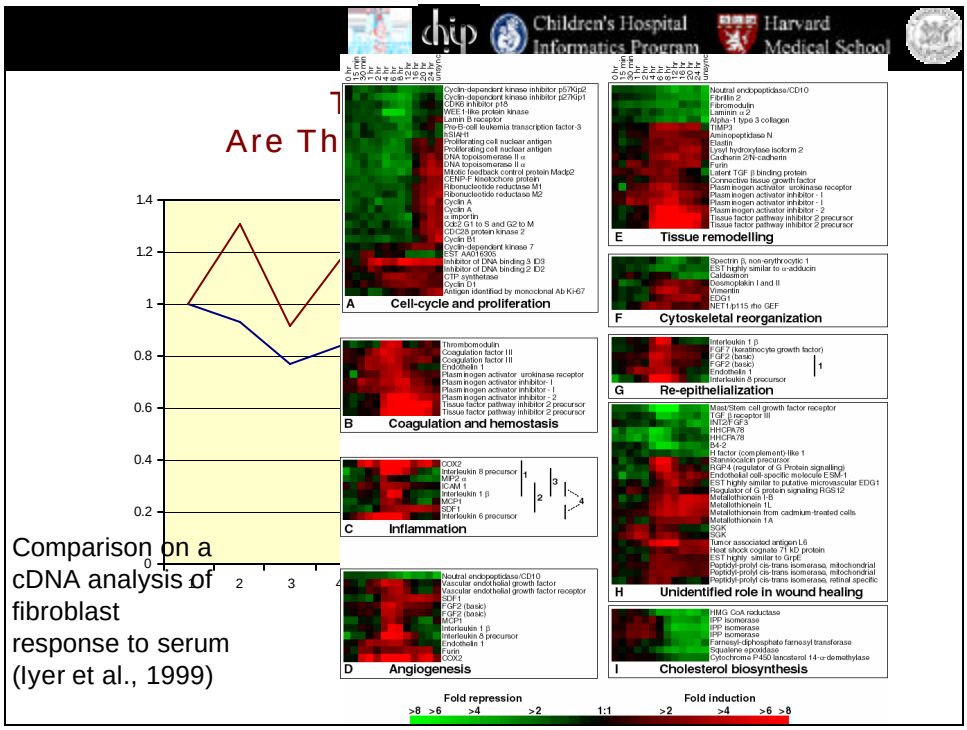
Dynamics

- Neglected until recently in functional genomics?
- Why?
- What do we lose by ignoring dynamics analyses in our current data sets?
- Work of Kohane, Shahar, Musen, and other in clinical informatics suggests that:
  - ✓ We lose discrimination
  - ✓ We lose insights into processes

Children's Hospital Informatics Program | Harvard Medical School

Are They Similar?





Comparison on a cDNA analysis of fibroblast response to serum (Iyer et al., 1999)

Autoregressive Models

- Take a time series, of dependent observations:

$$x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \ldots$$

- Approximate with an autoregressive model:

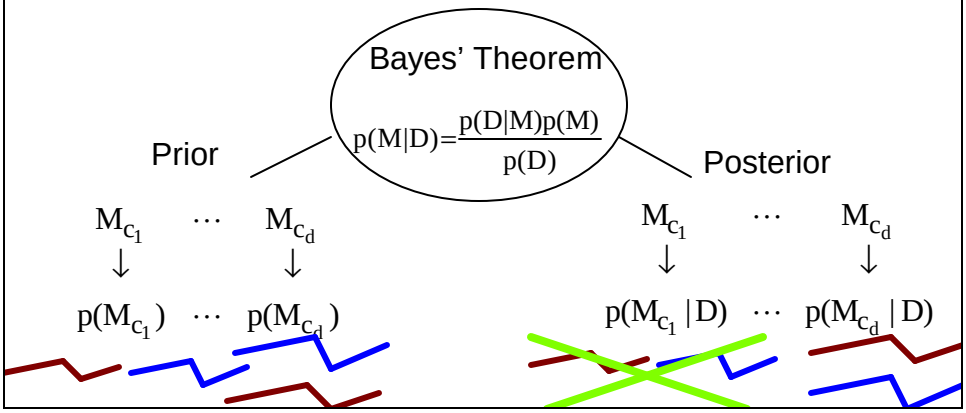
$$P(x_t \mid x_0, x_1, \ldots, x_{t-1}) \quad \text{——} \quad P(x_t \mid x_{t-p}, \ldots, x_{t-1})$$

the basic assumption is that t_0 is independent of the remote past given the recent past.

- The length of the recent past is the Markov Order *p*.

Bayesian Model Selection

- Clustering is a statistical model selection problem.
- We are looking for the model with maximum posterior probability given the data.



Children's Hospital Informatics Program | Harvard Medical School

Marginal Likelihood

- We want the most probable model given the data:

$$p(M_i \mid \Delta) = \frac{p(M_i, \Delta)}{p(\Delta)} = \frac{p(\Delta \mid M_i)\, p(M_i)}{p(\Delta)}$$

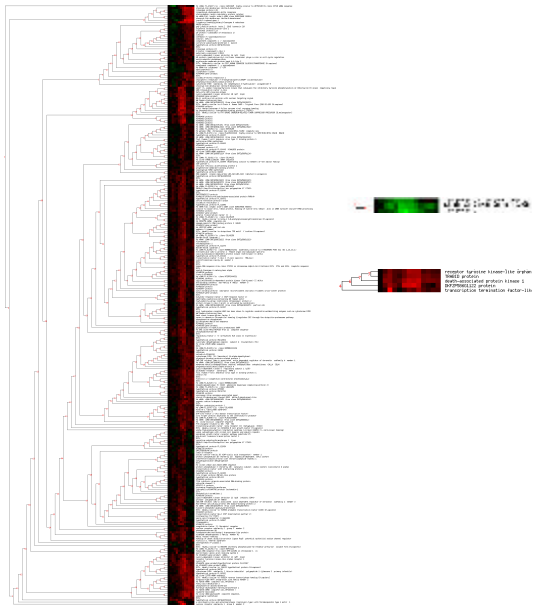
- But we use the same data for all models:

$$p(M_i|D) \propto p(D \mid M_i) p(M_i).$$

- We assume all models are a priori equally likely:

$$p(M_i|D) \propto p(D \mid M_i).$$

- This is the marginal likelihood, which gives the most probable model generating Δ.



Four Clusters

Children's Hospital Informatics Program
Harvard Medical School

Cluster Members

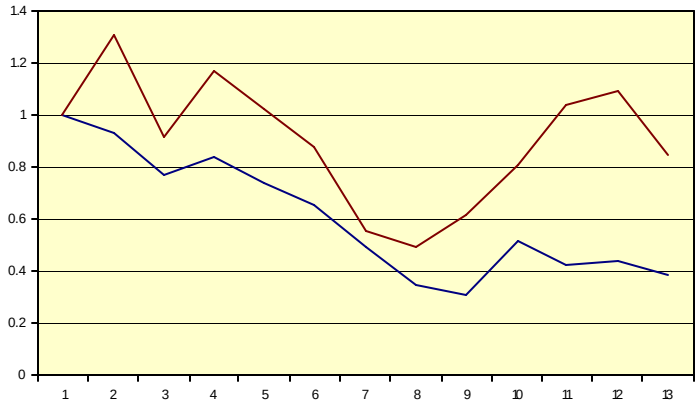
Cytokine Cluster	*Cluster 2*	*Cluster 3*	*Apoptosis Cluster*
Interleukin 8	D	A	Tyrosine kinase-like orphan receptor 2
Interleukin 6 (interferon beta 2)	E	B	TRAF-binding protein domain
Prostaglandin-endoperoxide synth 2	F	C	Death-associated protein kinase
	G		Transcription termination factor-like protein
	H		DKFZP586G1122 protein
	I		
	J		

Validation

- Back in 1999, 238 out of 517 genes were unknown.
- We relabeled the genes according to the current state of the art and less than 20 are left unknown.
- There are 19 *repeated genes* in the dataset:
  - ✓ Original clustering puts 4 of these in different clusters;
  - ✓ We put 1 of these in two different clusters.
- Interestingly enough, if we run the clustering with Markov order 0 (assuming uncorrelated iid data), we get 4 "misplacements" as well, albeit of other genes.
- Conclusion:
  - ✓ Temporal order provides more accurate insights into concerted behavior of gene expression.
  - ✓ Sound statistical basis for models rather than the "looks right" test
  - ✓ Now being applied to several time-series in heart disease model, brain development, transplant rejection, insulin effect

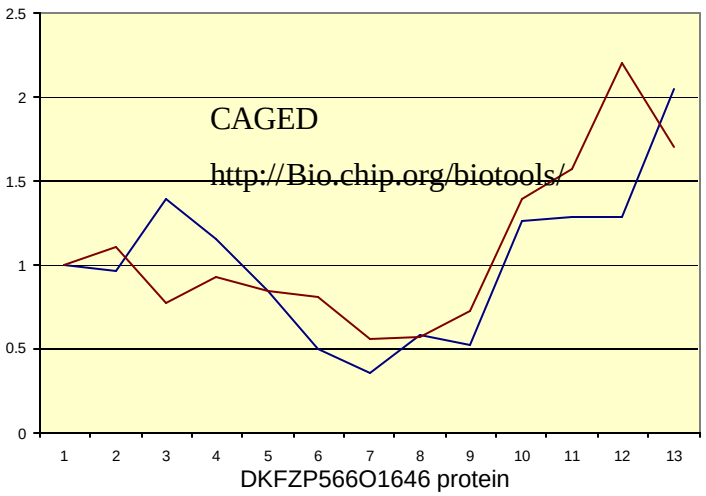
Children's Hospital Informatics Program
Harvard Medical School

They are the same



Tax1 (human T-cell leukemia virus type I) binding protein 1

Are they really the same?

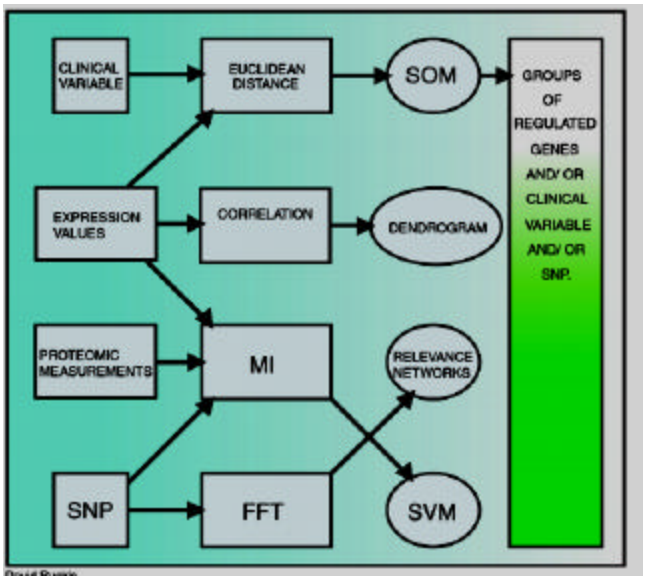


Children's Hospital Informatics Program
Harvard Medical School

Dissimilarity Metrics

Finding the **right representation** for genomic data

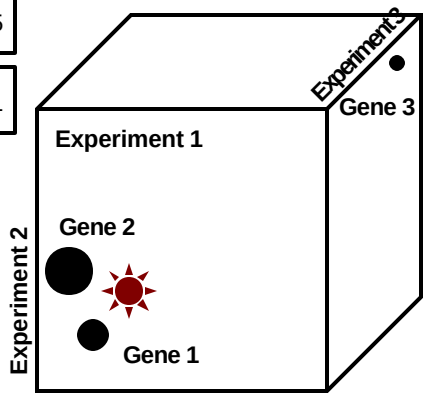
At least as important as employing an optimal analytic algorithm



Framework

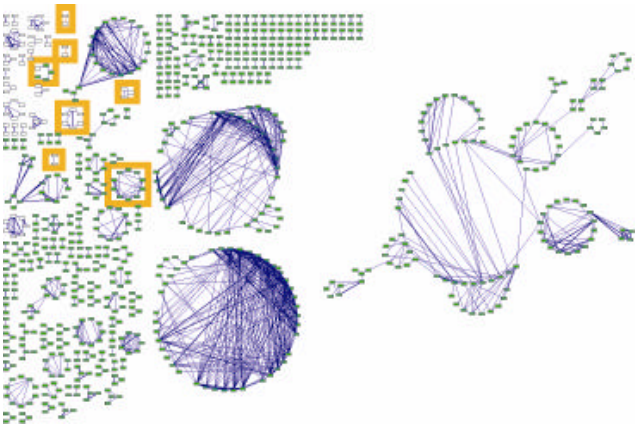
Already addressed: Euclidean Distance

	RNA Expr Gene 1	RNA Expr Gene 2	RNA Expr Gene 3
Experiment 1	0.7	0.3	7.3
Experiment 2	1.2	1.9	6.5
Experiment 3	1.1	0.9	8.1



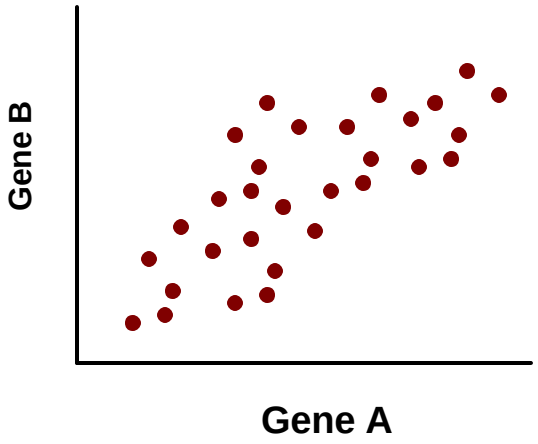
Already addressed:correlation coefficient

- Useful when
  - ✓ Negative correlations are of interest
  - ✓ When Euclidean distance is inadequate

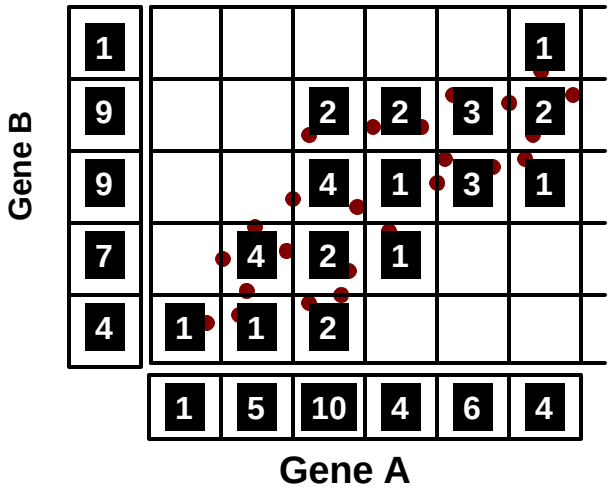


Children's Hospital Informatics Program
Harvard Medical School

Mutual Information



Mutual Information

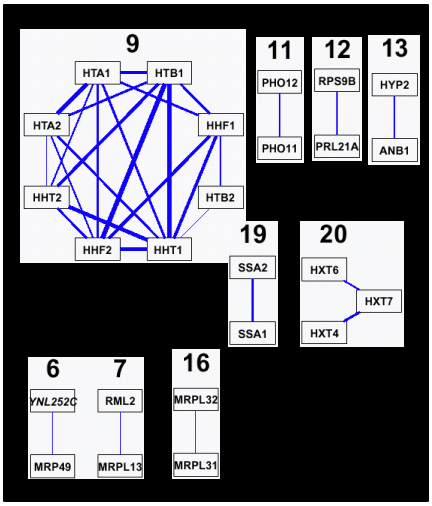


Mutual Information (A,B) =
Entropy (A) + Entropy (B) – Entropy (A and B)

Children's Hospital Informatics Program | Harvard Medical School

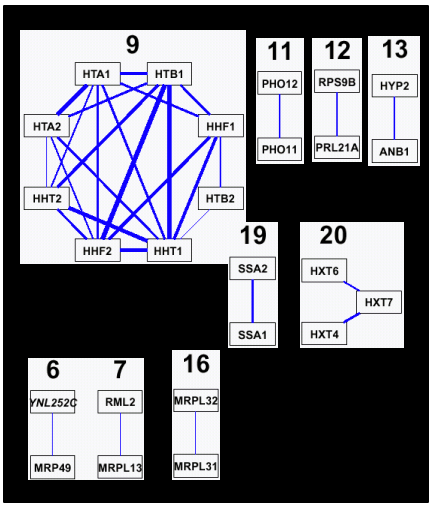
Similar Function Networks

- Nine networks with associations linking genes with similar functions
  - ✓ Histones
  - ✓ Acid phosphatases
  - ✓ Ribosomal proteins
  - ✓ Translation initiation
  - ✓ 70 kDa heat shock proteins
  - ✓ Hexose transporters
  - ✓ Mitochondrial ribosomal proteins

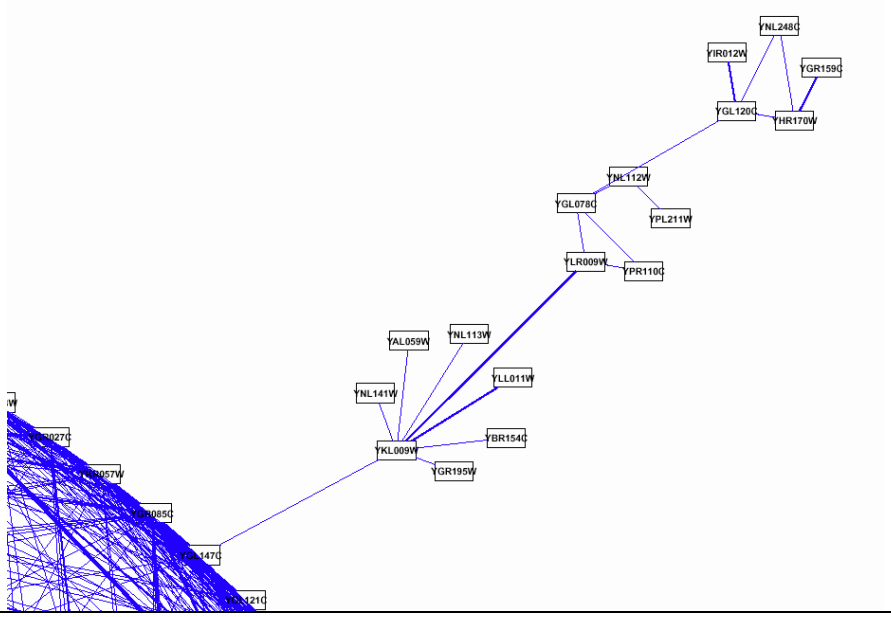


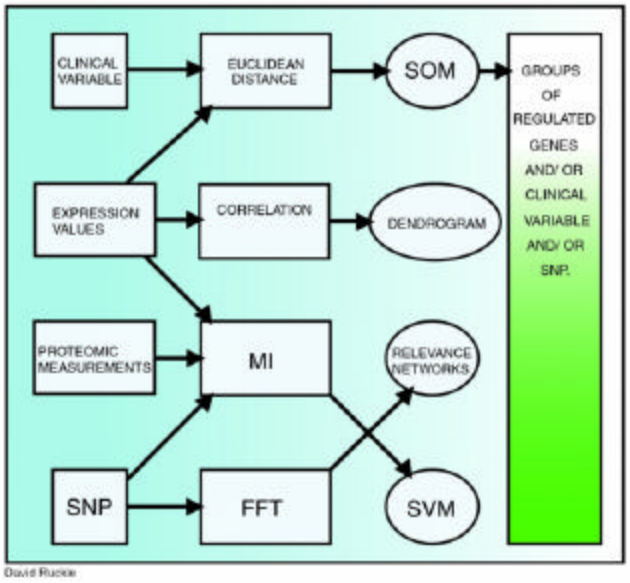
Similar Function Networks

- Nine networks with associations linking genes with similar functions
  - ✓ Histones
  - ✓ Acid phosphatases
  - ✓ Ribosomal proteins
  - ✓ Translation initiation
  - ✓ 70 kDa heat shock proteins
  - ✓ Hexose transporters
  - ✓ Mitochondrial ribosomal proteins

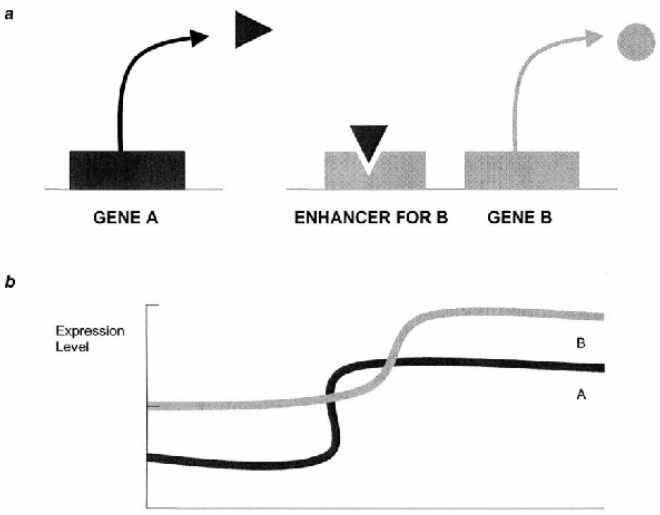


Other Networks





Framework



Metrics that Capture Dynamics

Point by point slopes

- Different networks identified than in static networks (why?)
- Artifactual high correlations near rest

Slope(n,n+1) =

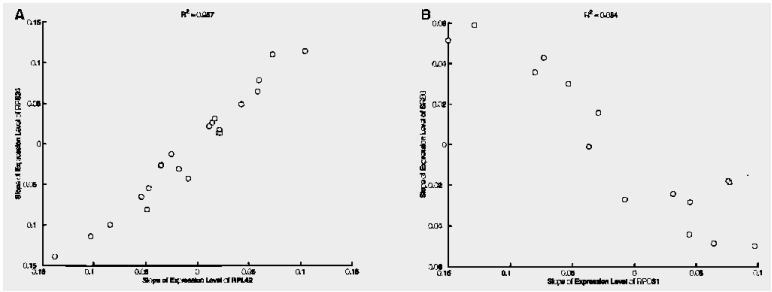
$$\frac{(\text{expression\_level}_{n+1} - \text{expression\_level}_{n})}{(\text{time}_{n+1} - \text{time}_{n})}$$

Children's Hospital Informatics Program | Harvard Medical School

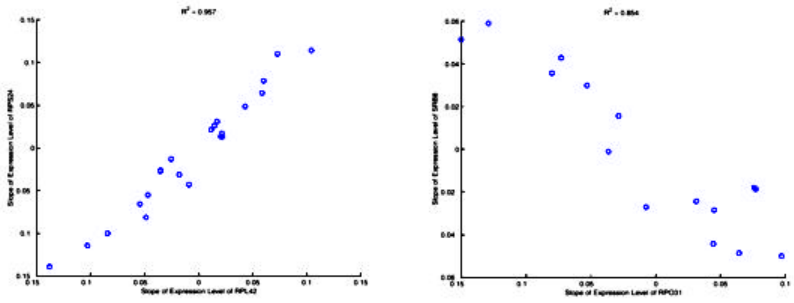
Slope Results

r^2=0.057

r^2=0.054



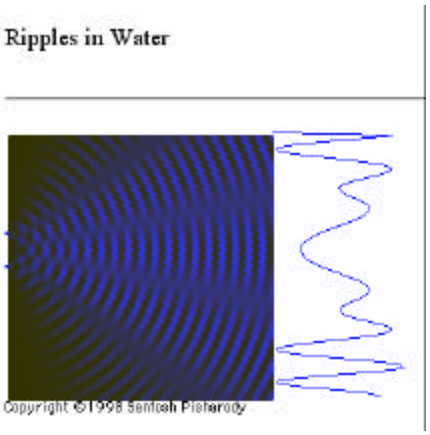
Slope dynamics



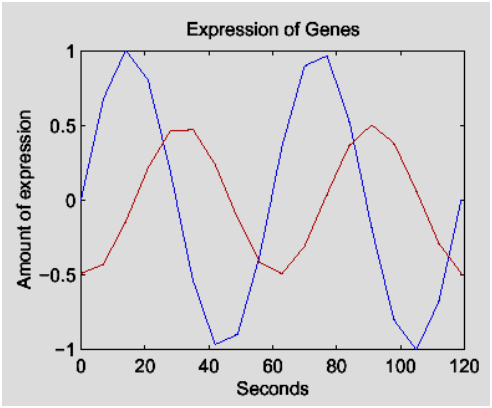
Strong relationships not identified by static dissimilarity measures

Children's Hospital Informatics Program | Harvard Medical School

Reconstructing the source waveform



- So far, our potential genetic networks are undirected
- Temporality is first step to determining causality
- Multi-frequency analysis, assuming nothing about genomic frequency

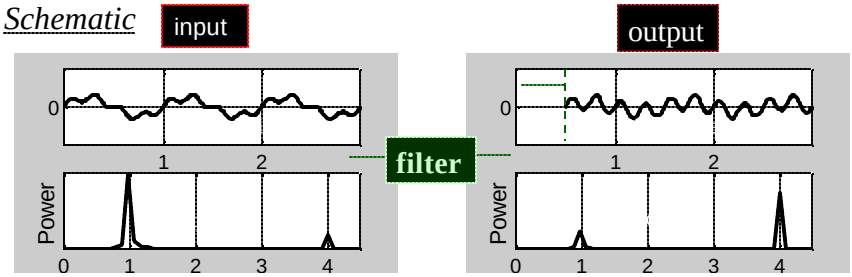


Gene A
Transmitter

Biological System

Gene B
Receiver

Schematic

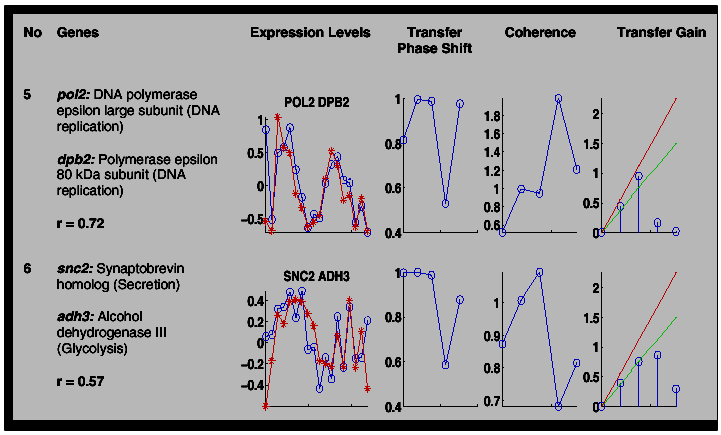


At each frequency, the relationship between oscillations of in input and output are quantified by:

- ✓ **Transfer gain**: amplitude modulation from input to output.
- ✓ **Transfer phase**: time lead or time lag from input to output.
- ✓ **Coherence**: linearity of relationship between signals.

Signal Processing Finds Gene Regulation

- Given enough compute-power, digital signal processing can be successfully applied to all possible pairs of genes



Butte AJ, et al. Comparing the Similarity of Time-Series Gene Expression Using Signal Processing Metrics, Submitted.

Children's Hospital Informatics Program | Harvard Medical School

How do you determine what is the right dissimilarity measure

What is the figure of merit?

- For classification?
- For clustering?

Table 1. Comparison of error rates for various classification methods

Class	Method	FP	FN	TP	TN	S(M)
TCA	D-p 1 SVM	18	5	12	2,432	6
	D-p 2 SVM	7	9	8	2,443	9
	D-p 3 SVM	4	9	8	2,446	12
	Radial SVM	5	9	8	2,445	11
	Parzen	4	12	5	2,446	6
	FLD	9	10	7	2,441	5
	C4.5	7	17	0	2,443	−7
	MOC1	3	16	1	2,446	−1
Resp	D-p 1 SVM	15	7	23	2,422	31
	D-p 2 SVM	7	7	23	2,430	39
	D-p 3 SVM	6	8	22	2,431	38
	Radial SVM	5	11	19	2,432	33
	Parzen	22	10	20	2,415	18
	FLD	10	10	20	2,427	30
	C4.5	18	17	13	2,419	8
	MOC1	12	26	4	2,425	−4
Ribo	D-p 1 SVM	14	2	119	2,332	224
	D-p 2 SVM	9	2	119	2,337	229
	D-p 3 SVM	7	3	118	2,339	229
	Radial SVM	6	5	116	2,340	226
	Parzen	6	8	113	2,340	220
	FLD	15	5	116	2,331	217
	C4.5	31	21	100	2,315	169
	MOC1	26	26	95	2,320	164
Prot	D-p 1 SVM	21	7	28	2,411	35
	D-p 2 SVM	6	8	27	2,426	48

Bake-off

- What does this mean?

Brown, PNAS '00

Concordance

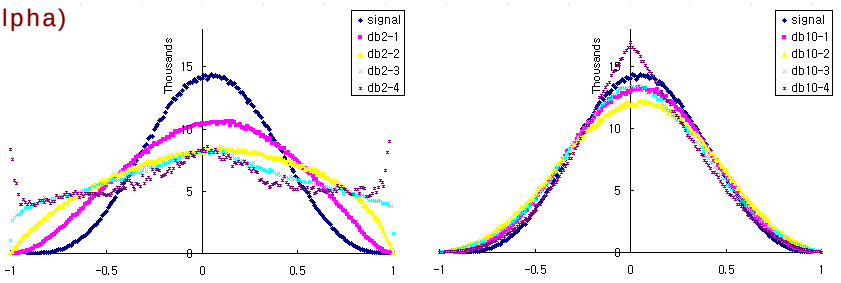
	Rater 1 - Yes	Rater 1 - No
Rater 2 - Yes	A	B
Rater 2 - No	C	D

The Jaccard, J = A/(A+B+C)

Children's Hospital Informatics Program
Harvard Medical School

Permutation Test

Distribution of R (alpha)



Multiple Evaluation/Scoring Techniques

- Robustness under noise
- Matching known associations/relationships
  - ✓ Literature
  - ✓ Gene Ontology
- Comparisons to alternative "wet" measurement results.

Children's Hospital Informatics Program | Harvard Medical School

Visual Inspection

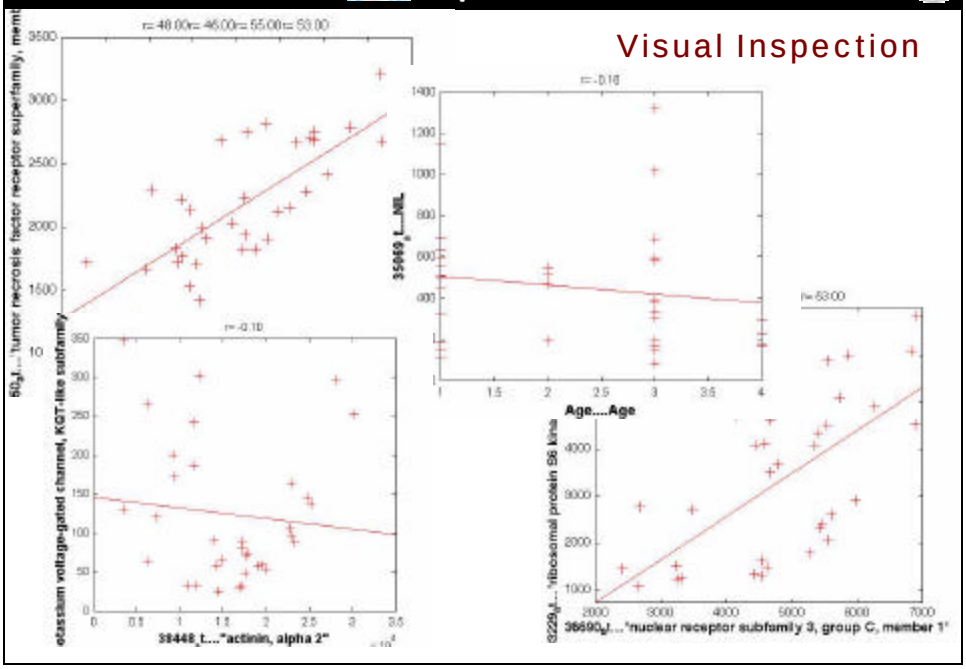
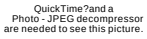


Figure of Merit has to be found in the Scientific Method

- Example Biological Validation

Children's Hospital Informatics Program Harvard Medical School

A short introduction to functional genomic causality with Bayesian Networks



Comprehensive Bioinformatics Approach: All Data Are Grist

Children's Hospital Informatics Program
Harvard Medical School

How do we combine all these data types

- If functional genomics is going to be meaningful we will have to include:
  - ✓ Phenotype data
  - ✓ Environmental data
  - ✓ Genomic data (RNA, SNP, Proteomic etc)
- Each data type has its own characteristics.
- How do we add evidence from different sources in a principled way that is statistically sound?

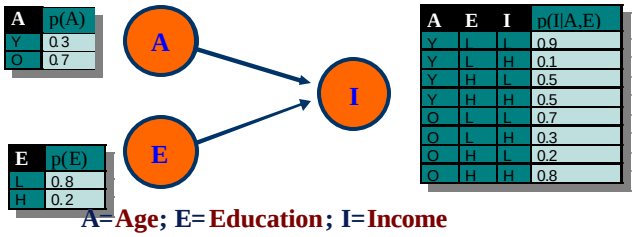
Bayesian Network: A Framework for Probabilistic Relationships

Qualitative: A dependency graph made by:

Node: a variables X, with a set of states $\{x_1,\ldots,x_n\}$.

Arc: a dependency of a variable X on its parents Π.

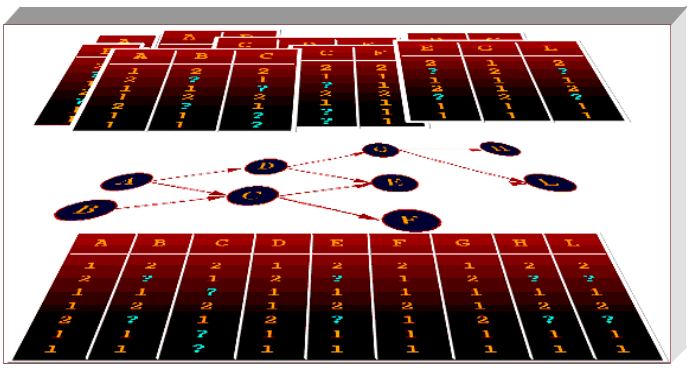
Quantitative: The distributions of a variable X given each combination of states π_i of its parents Π.



A=Age; E=Education; I=Income

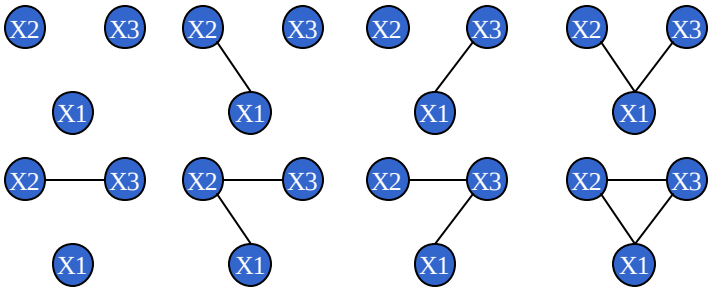
Finding the Most-likely Dependency Graph

- Goal: find the global model that best reflects the data.



Search

Problem: Searching for all the possible models is intractable.

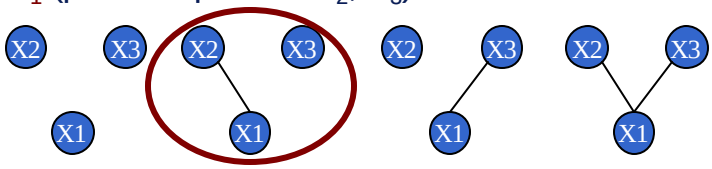


Children's Hospital Informatics Program | Harvard Medical School

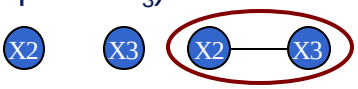
Growth in Number of Possible Models with Variables

Local Model Selection

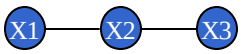
X_1 (possible parents X_2; X_3):



X_2 (possible parent X_3).



The model:



Add "parents" only if it increases the marginal likelihood of the model

i.e. **plausible alternative explanation of the effect**

Children's Hospital Informatics Program | Harvard Medical School

Application

Cases: 41 patients affect by leukemia.

- Genomic: expression measures on 72 genes;
- Clinical: 38 clinical phenotypes (3 used).

Representational Risks:

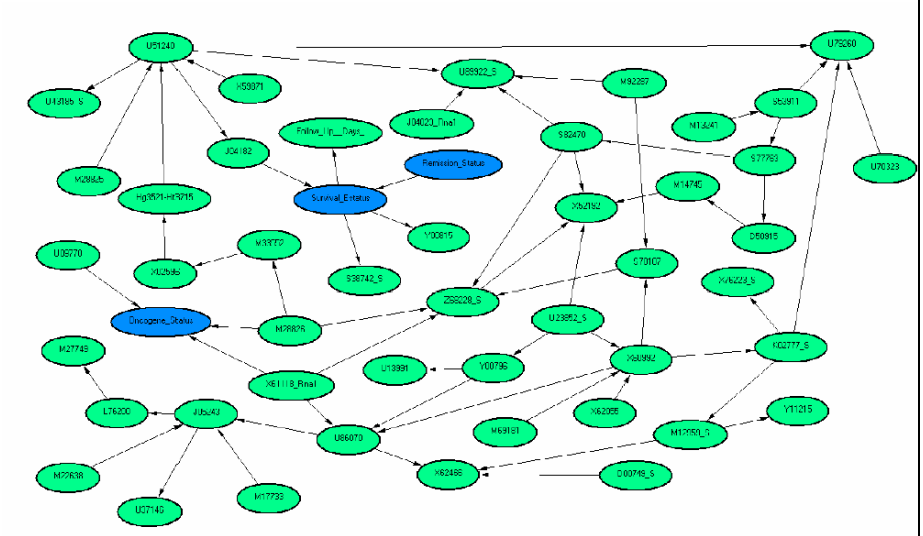
- Deterministic links: hide other links more interesting.
- Overfitting: Too many states for the available data.

Transformations:

- Definitional dependencies: if suspected, removed.
- Sparse phenotypes: consolidated (oncogene status).

With Adolfo Fernando Thomas Look, DFCI, Harvard Medical School.

Relationship between gene and phenotype



Children's Hospital Informatics Program
Harvard Medical School

Dependency Strength

Bayes factor: ratio between the probability of 2 models.

Threshold: To add a link, we need to gain at least 3 BF.

Oncogene_Status

X61118_Rna1	M28826	U09770		1
U09770	X61118_Rna1	M28826	S53911	7
U09770	X61118_Rna1	M28826	M69181	56
X61118_Rna1	M28826	M17733		63
X61118_Rna1	M28826	S77763		315
U09770	X61118_Rna1	M28826	U43185_S	447
U09770	X61118_Rna1	M28826	J05243	447
X61118_Rna1	M28826			973
X61118_Rna1	M28826	S38742_S		1016
X61118_Rna1	M28826	J05243		1534
U09770	X61118_Rna1	M28826	M17733	1804
U09770	X61118_Rna1	M28826	Z68228_S	1807
U09770	X61118_Rna1	M28826	J04823_Rna1	3558
U09770	X61118_Rna1	M28826	U37146	3564
U09770	X61118_Rna1	M28826	X62055	3564
U09770	X61118_Rna1	M28826		3570
X61118_Rna1	M28826	Y11215		3933
X61118_Rna1	M28826			4254
U09770	X61118_Rna1	M28826	Survival_Estatus	7369
X61118_Rna1	M28826	Survival_Estatus		11093

Validation

Cross-validation: A form of predictive validation.

1. For each case, remove it from the database;
2. Use these data to learn the probability distributions of the network;
3. Use the quantified network to predict value on a variable of the removed case.

Validation parameters:

Correctness: Number of cases correctly predicted;

Coverage: Number of cases actually predicted;

Average Distance: How uncertain is a prediction.

Results

25 variables were isolated.

Cross validation conducted on two variables.

Oncogene Status:

Coverage: 100% (41).

Accuracy: 97.56% (40).

Average Distance: 0.03339.

Survival Status:

Coverage: 97.56% (40).

Accuracy: 100% (40).

Average Distance: 0.004146.