Application of Association Pattern Mining Techniques to Genomic Data

Vipin Kumar
University of Minnesota

kumar@cs.umn.edu
www.cs.umn.edu/~kumar

Team Members: Michael Steinbach, Rohit Gupta, Gowtham Atluri, Gang Fang, Gaurav Pandey, Tushar Garg

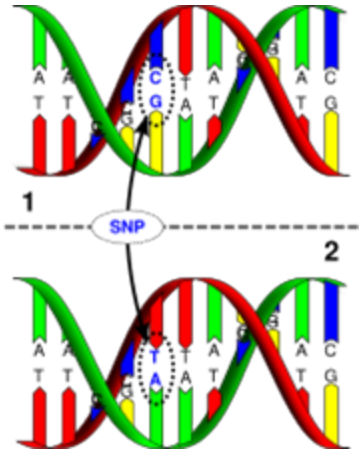
Collaborators: Brian Van Ness, Bill Oetting, Rich Mushlin (IBM)

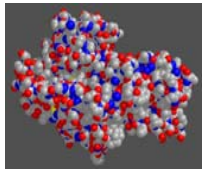
Research Supported by NSF, IBM, BICB-UMR

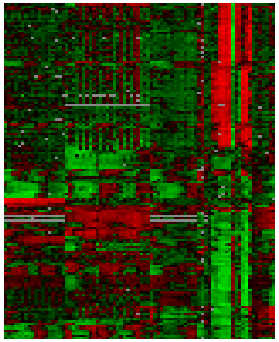
Data Mining for Biomedical Informatics

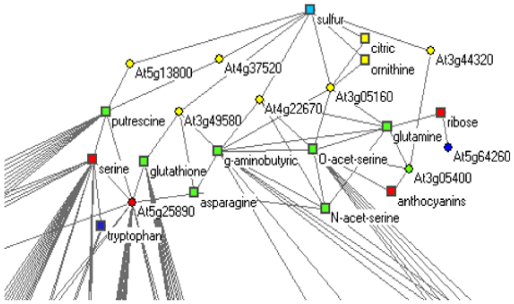
- Recent technological advances are helping to generate large amounts of genomic data
  - Gene and protein sequences
  - Gene-expression data
  - Biological networks and phylogenetic profiles
  - Single Nucleotides Polymorphisms (SNPs)

- Data mining offers potential solution for analysis of large-scale data
  - Prediction of the functions of anonymous genes
  - Identification and summarizarion of functional modules
  - Associations between genotypes and phenotype of interest









Protein Interaction Network

Association Analysis

- Association analysis: Analyzes relationships among items (attributes) in a binary transaction data
  - Example data: market basket data
  - Applications in business and industry
    - Marketing and sales promotion, inventory management
    - Telecommunication alarm diagnosis
  - Potential applications in biology
    - Identification of functional modules from protein complexes
    - Noise removal from protein interaction data
    - Genotype-phenotype associations
- Two types of patterns
  - **Itemsets:** Collection of items
    - Example: {Milk, Diaper}
  - **Association Rules:** X → Y, where X and Y are itemsets.
    - Example: Milk → Diaper

Set-Based Representation of Data

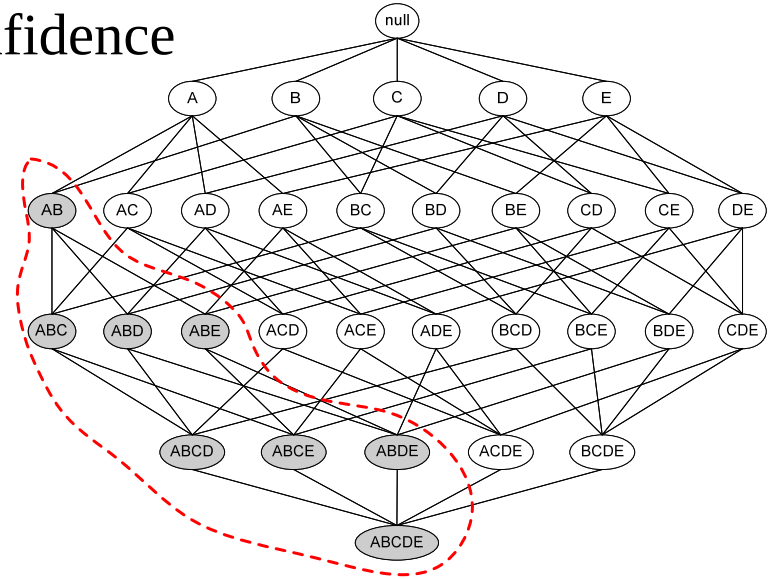
TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

$$\text{Support, } s = \frac{\#\text{ transactions that contain X and Y}}{\text{Total transactions}}$$

$$\text{Confidence, } c = \frac{\#\text{ transactions that contain X and Y}}{\#\text{ transactions that contain X}}$$

Association Analysis

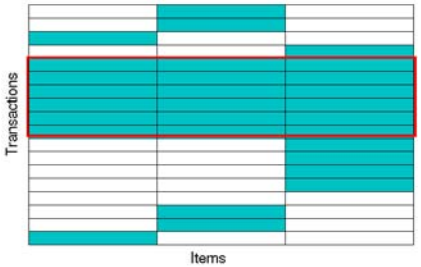
- Process of finding interesting patterns:
  - Find frequent itemsets using a support threshold
  - Find association rules for frequent itemsets
  - Sort association rules according to confidence
- Support filtering is necessary
  - To eliminate spurious patterns
  - To avoid exponential search
    - Support has anti-monotone property: $X \subseteq Y$ implies $\sigma(Y) \leq \sigma(X)$
- Confidence is used because of its interpretation as conditional probability
  - Has well-known limitations
  - Many other measures have been used as well



Given d items, there are 2^d possible candidate itemsets

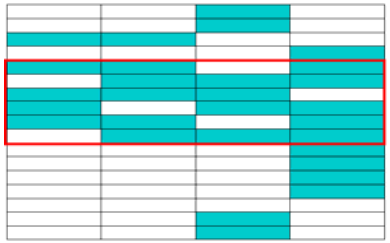
Different Association patterns

- Traditional frequent pattern

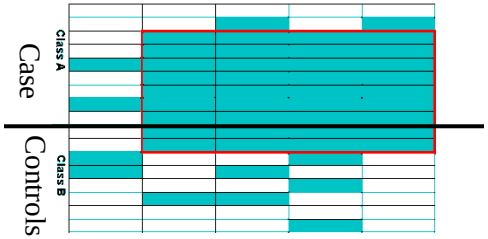


- Hyperclique pattern

- Error-tolerant frequent pattern



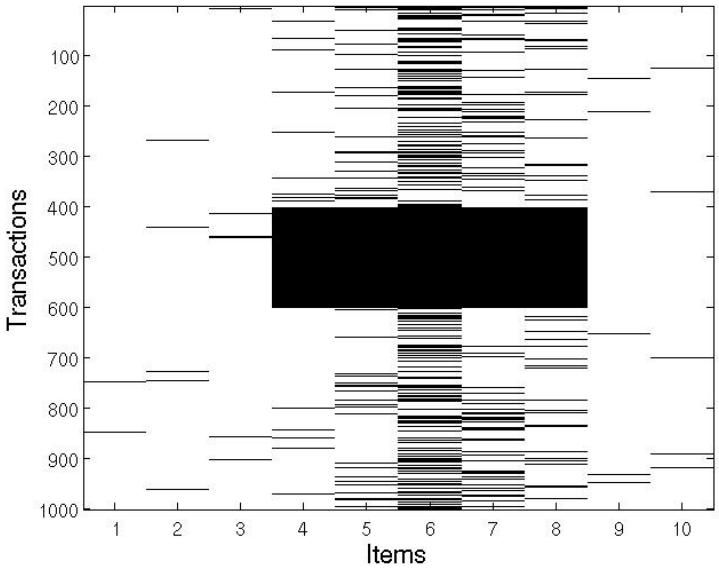
- Discriminative pattern



Applications of Hypercliques to Biological Data Sets

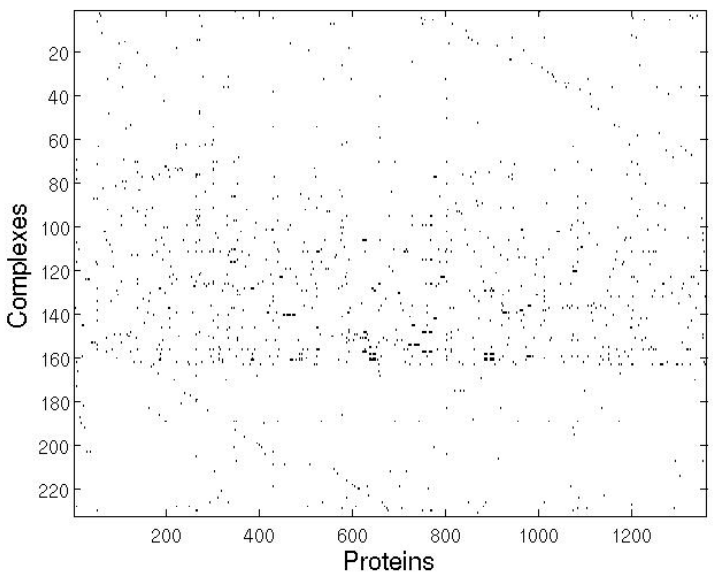
- Discovery of functional modules from protein complexes [Xiong et al, 2005, PSB]

- Noise removal [Xiong et al, 2006b, IEEE TKDE]
  - Data points not a member of any hypercliques hypothesized to be noisy
  - Improved performance of several data analysis tasks (association analysis, clustering) on several types of data sets (text, microarray data)
  - Noise-resistant transformation of protein interaction networks [Pandey et al, 2007, KDD]



Application I: Association Analysis for Identification of Protein Function Modules

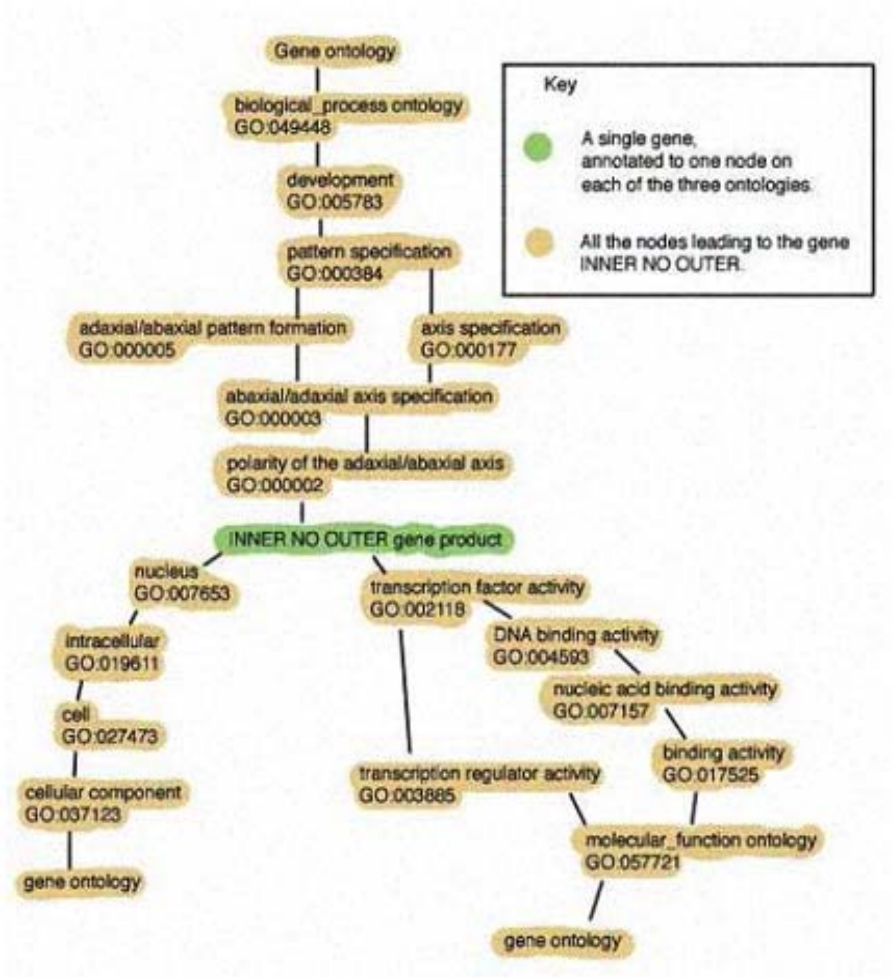
Complexes	Proteins
c1	p1, p2
c2	p1, p3, p4, p5
c3	p2, p3, p4, p6



- A **protein complex** is a group of two or more associated proteins formed by protein-protein interaction that is stable over time
- The TAP-MS dataset by Gavin et al 2002: Tandem affinity purification (TAP) – mass spectrometry (MS)
- Contains 232 multi-protein complexes formed using 1361 proteins
- Number of proteins per complex range from 2 to 83 (average 12 components)
- Hypercliques derived from this data can be used to discover frequently occurring groups of proteins in several complexes
  - Likely to constitute functional modules

Xiong et al [2005], PSB

Functional Group Verification Using Gene Ontology



- Hypothesis: Proteins within the same pattern are more likely to perform the same function and participate in the same biological process
- Gene Ontology
  - Three separate ontologies: Biological Process, Molecular Function, Cellular Component
  - Organized as a DAG describing gene products (proteins and functional RNA)
  - Collaborative effort between major genome databases

http://www.geneontology.org

Hyperclique Patterns from Protein Complex Data

- List of maximal hyperclique patterns at a support threshold 2 and an h-confidence threshold 60%. [1] Xiong et al. (Detailed results are at http://cimic.rutgers.edu/~hui/pfm/pfm.html)

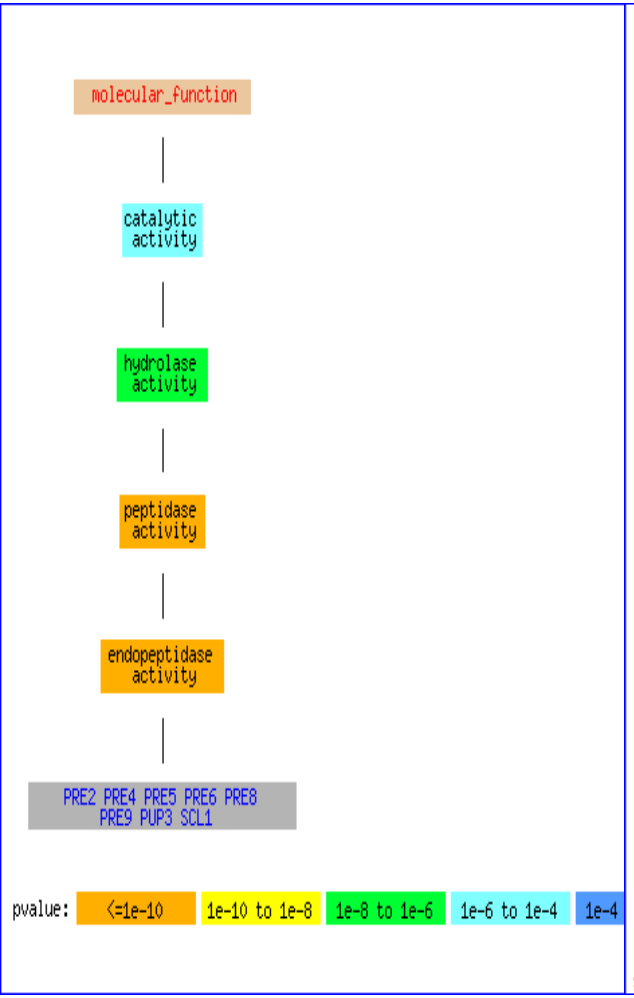
2 Tif4632 Tif4631
2 Cdc33 Snp1
2 YHR020W Mir1
2 Cka1 Ckb1
2 Ckb2 Cka2
2 Cop1 Sec27
2 Erb1 YER006W
2 Ilv1 YGL245W
2 Ilv1 Sec27
2 Ioc3 Rsc8
2 Isw2 Itc1
2 Kre33 YJL109C
2 Kre33 YPL012W
2 Mot1 Isw1
2 Npl3 Smd3
2 Npl6 Isw2
2 Npl6 Mot1
2 Rad52 Rfa1
2 Rpc40 Rsc8
2 Rrp4 Dis3
2 Rrp40 Rrp46
2 Cbf5 Kre33
3 YGL128C Clf1 YLR424W
3 Cka2 Cka1 Ckb1
3 Has1 Nop12 Sik1
3 Hrr25 Enp1 YDL060W
3 Hrr25 Swi3 Snf2
3 Kre35 Nog1 YGR103W
3 Krr1 Cbf5 Kre33
3 Nab3 Nrd1 YML117W
3 Nog1 YGR103W YER006W
3 Bms1 Sik1 Rpp2b
3 Rpn10 Rpt3 Rpt6
3 Rpn11 Rpn12 Rpn8
3 Rpn12 Rpn8 Rpn10
3 Rpn9 Rpt3 Rpt5
3 Rpn9 Rpt3 Rpt6
3 Brx1 Sik1 YOR206W
3 Sik1 Kre33 YJL109C
3 Taf145 Taf90 Taf60
4 Fyv14 Krr1 Sik1 YLR409C
4 Mrpl35 Mrpl8 YML025C Mrpl3
4 Rpn12 Rpn8 Rpt3 Rpt6
5 Rpn6 Rpt2 Rpn12 Rpn3 Rpn8
5 Ada2 Gcn5 Rpo21 Spt7 Taf60
6 YLR033W Ioc3 Npl6 Rsc2 Itc1 Rpc40
6 Dim1 Ltv1 YOR056C YOR145C Enp1 YDL060W
6 Luc7 Rse1 Smd3 Snp1 Snu71 Smd2
6 Pre3 Pre2 Pre4 Pre5 Pre8 Pup3
7 Clf1 Lea1 Rse1 YLR424W Prp46 Smd2 Snu114
7 Pre1 Pre7 Pre2 Pre4 Pre5 Pre8 Pup3
7 Blm3 Pre10 Pre2 Pre4 Pre5 Pre8 Pup3
8 Clf1 Prp4 Smb1 Snu66 YLR424W Prp46 Smd2 Snu114
8 Pre2 Pre4 Pre5 Pre8 Pup3 Pre6 Pre9 Scl1
10 Cdc33 Dib1 Lsm4 Prp31 Prp6 Clf1 Prp4 Smb1 Snu66 YLR424W
12 Dib1 Lsm4 Prp31 Prp6 Clf1 Prp4 Smb1 Snu66 YLR424W Prp46 Smd2 Snu114
12 Emg1 Imp3 Imp4 Kre31 Mpp10 Nop14 Sof1 YMR093W YPR144C Krr1 YDR449C Enp1
13 Ecm2 Hsh155 Prp19 Prp21 Snt309 YDL209C Clf1 Lea1 Rse1 YLR424W Prp46 Smd2 Snu114
13 Brr1 Mud1 Prp39 Prp40 Prp42 Smd1 Snu56 Luc7 Rse1 Smd3 Snp1 Snu71 Smd2
39 Cus1 Msl1 Prp3 Prp9 Sme1 Smx2 Smx3 Yhc1 YJR084W Brr1 Dib1 Ecm2 Hsh155 Lsm4 Mud1 Prp11 Prp19 Prp21 Prp31 Prp39 Prp40 Prp42 Prp6 Smd1 Snt309 Snu56 Srb2 YDL209C Clf1 Lea1 Luc7 Prp4 Rse1 Smb1 Smd3 Snp1 Snu66 Snu71 YLR424W

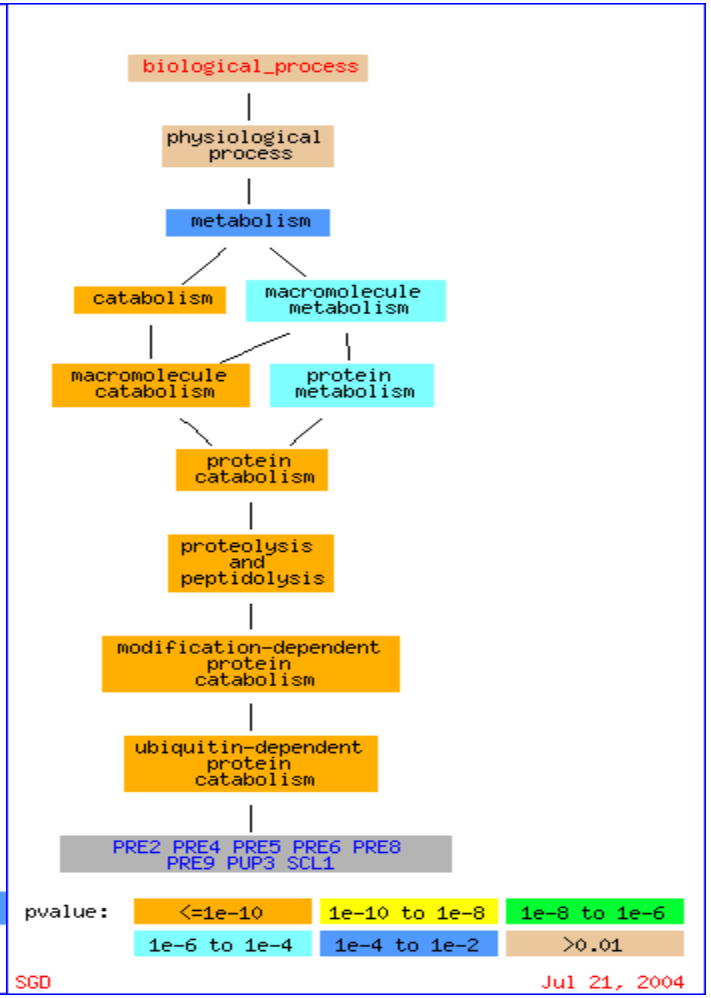
Summary

- Number of hypercliques:
  - Size-2: 22, Size-3: 18, Size-4: 3, Size-5: 2
  - Size-6: 4, Size-7: 3, Size-8: 2, Size-10: 1
  - Size-12: 2, Size-13: 2, Size-39: 1
- In most cases, proteins identified as hypercliques are found to be functionally coherent and part of same biological process when evaluated using GO hierarchies

Function Annotation for Hyperclique

{PRE2 PRE4 PRE5 PRE6 PRE8 PRE9 PUP3 SCL1}



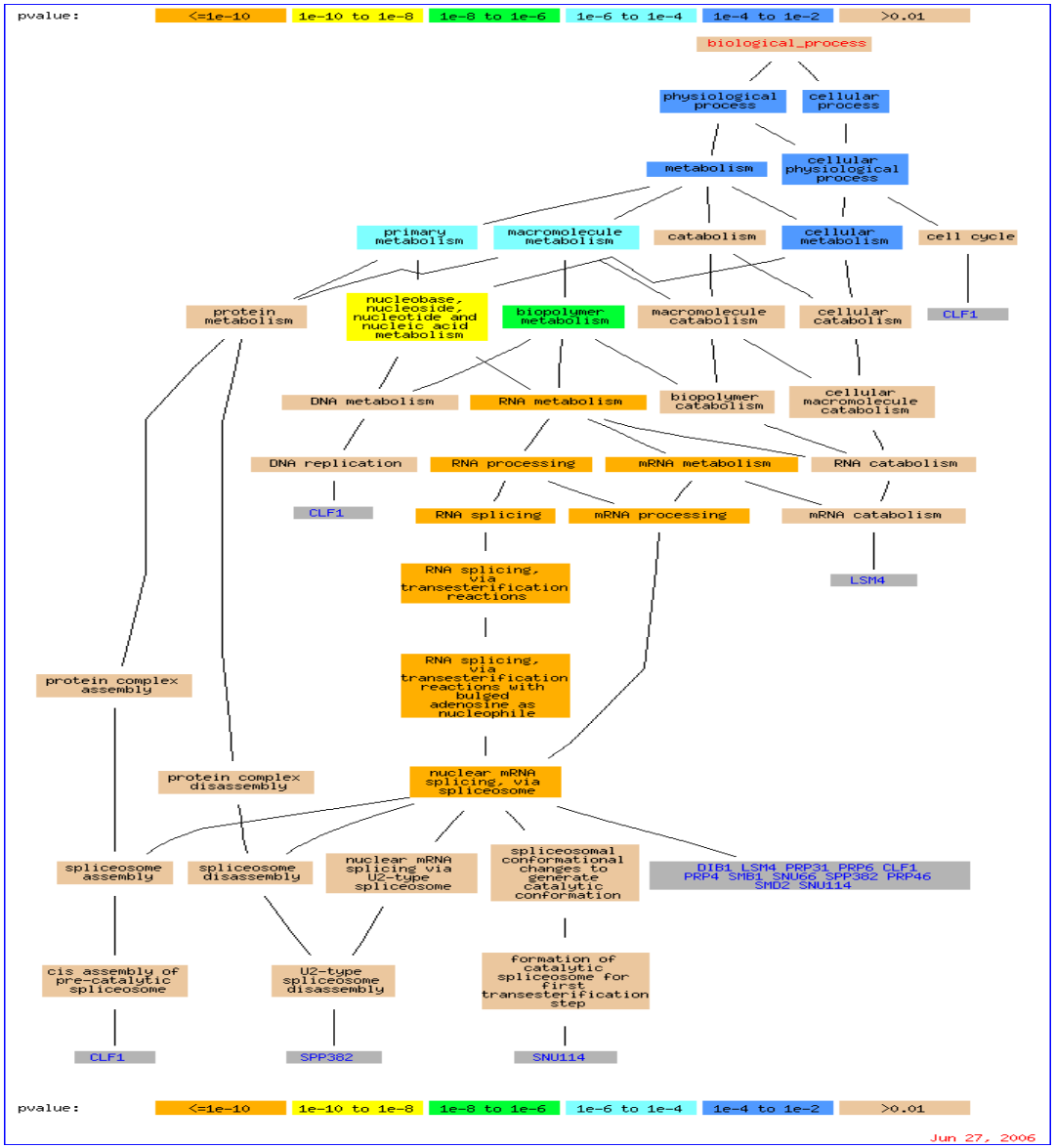


- GO hierarchy shows that the identified proteins in hyperclique perform the same function and are involved in the same biological process

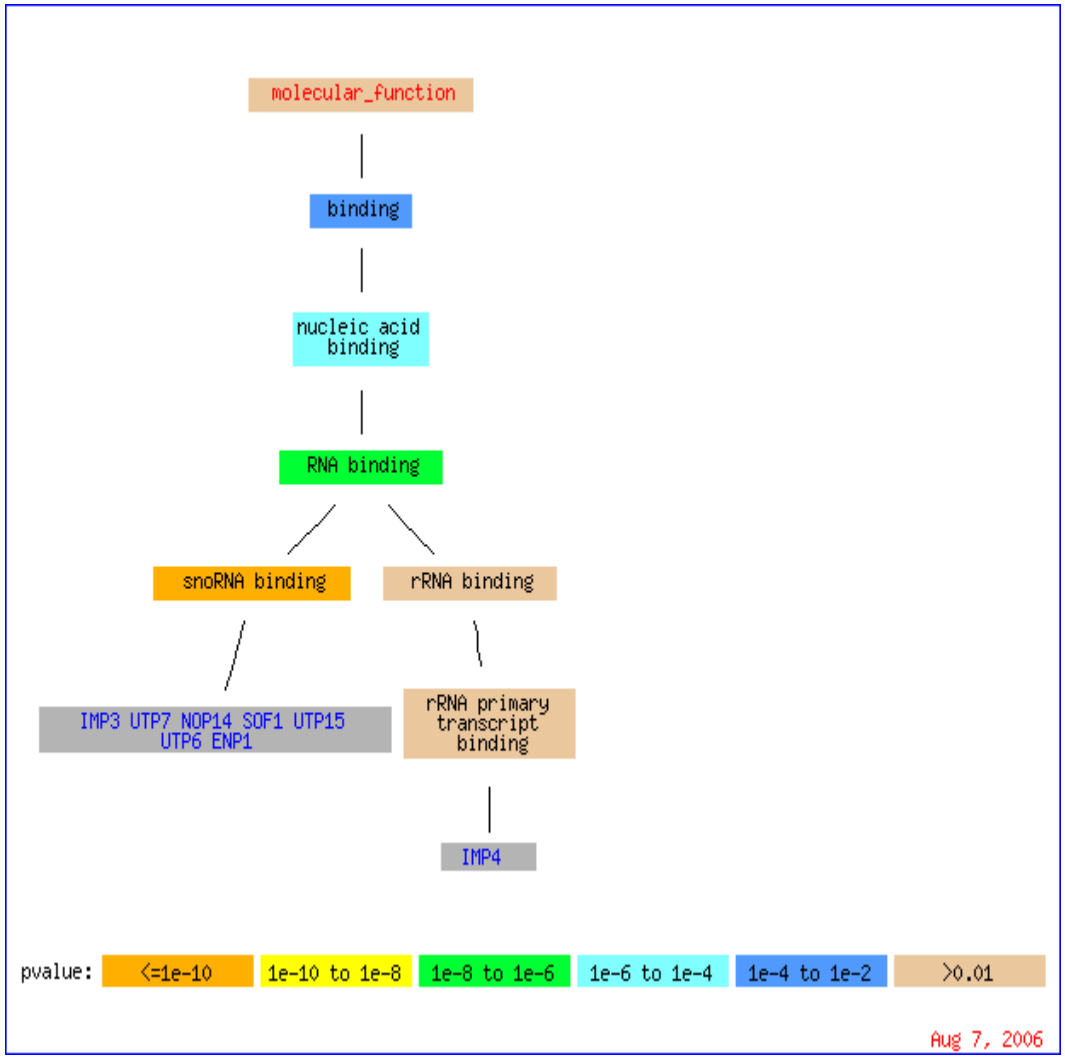
More Hyperclique Examples..

 # distinct proteins in cluster = 12

 # proteins in one group = 12



Functional Annotation of Uncharacterized Proteins

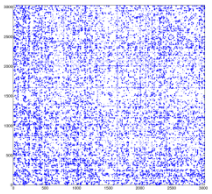


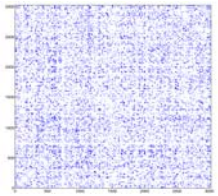
- Hyeperclique Pattern: {Emg1 Imp3 Imp4 Kre31 Mpp10 Nop14 Sof1 YMR093W YPR144C Krr1 YDR449C Enp1}
- 8 of the 12 proteins have an annotation of “RNA binding”
- Other 4 proteins have no functional annotation
- Hypothesis: Unannotated proteins have same molecular function “RNA binding”, since hypercliques tend to have proteins that are functionally coherent

Application II: Association Analysis-based Pre-processing of Protein Interaction Networks

- **Overall Objective**: Accurate inference of protein function from interaction networks
- **Challenges**:
  - Protein interaction networks are noisy and incomplete [Hart et al, 2006]
  - Adverse impact on accuracy of functional inferences [Deng et al, 2003]
- **Potential Approach**: Pre-processing of interaction networks
- Transform graph **G=(V,E,W)** into **G'=(V,E',W')**

Input PPI graph





Transformed PPI graph where P_i and P_j are connected if (P_i,P_j) is a hyperclique pattern

- Tries to meet three objectives:
  - Addition of potentially biologically valid edges
  - Removal of potentially noisy edges
  - Assignment of weights to the resultant set of edges that indicate their reliability

Validation of Final Network

- Use FunctionalFlow algorithm [Nabieva et al, 2005] on the original and transformed graph(s)
  - One of the most accurate algorithms for predicting function from interaction networks
  - Produces likelihood scores for each protein being annotated with one of 75 MIPS functional labels
- Likelihood matrix evaluated using two metrics
  - Multi-label versions of precision and recall:

$$Precision = \frac{\sum_{i=1}^{K} k_i}{\sum_{i=1}^{K} m_i} \qquad Recall = \frac{\sum_{i=1}^{K} k_i}{\sum_{i=1}^{K} n_i}$$

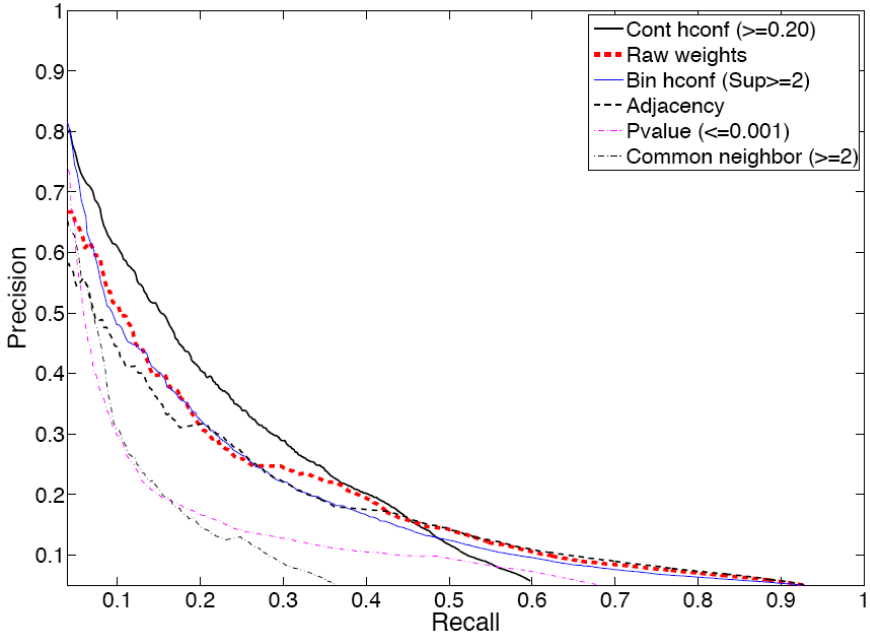
m_i = # predictions made, n_i = # known annotations, k_i = # correct predictions

  - Precision/accuracy of top-*k* predictions
    - Useful for actual biological experimental scenarios

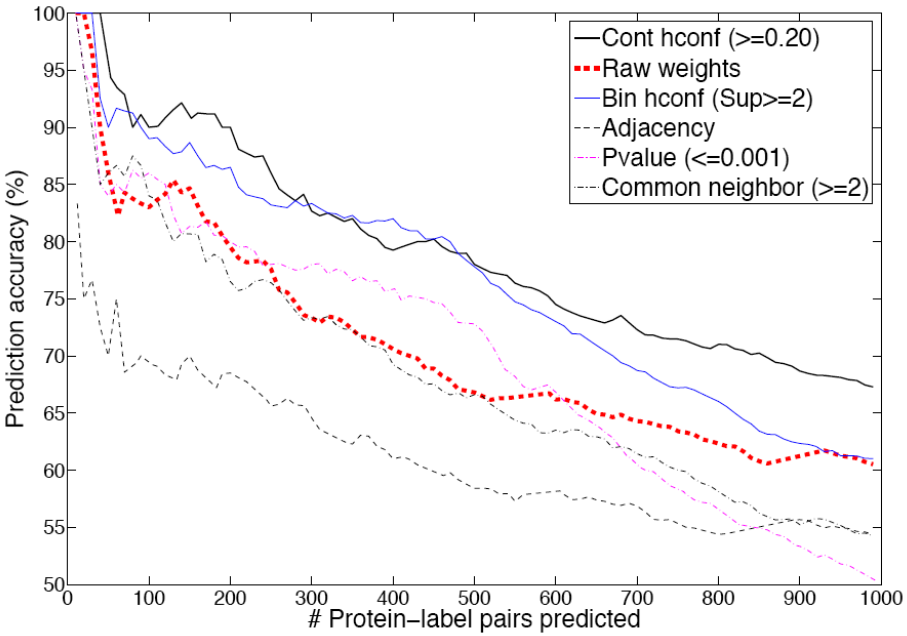
Test Protein Interaction Networks

- Three yeast interaction networks with different types of weighting schemes used for experiments
  - **Combined**
    - Composed from Ito, Uetz and Gavin (2002)'s data sets
    - Individual reliabilities obtained from EPR index tool of DIP
    - Overall reliabilities obtained using a *noisy-OR*
  - [**Krogan** et al, 2006]'s data set
    - 6180 interactions between 2291 annotated proteins
    - Edge reliabilities derived using machine learning techniques
  - **DIPCore** [Deane et al, 2002]
    - ~5K highly reliable interactions in DIP
    - No weights assigned: assumed unweighted

Results on **Combined** data set

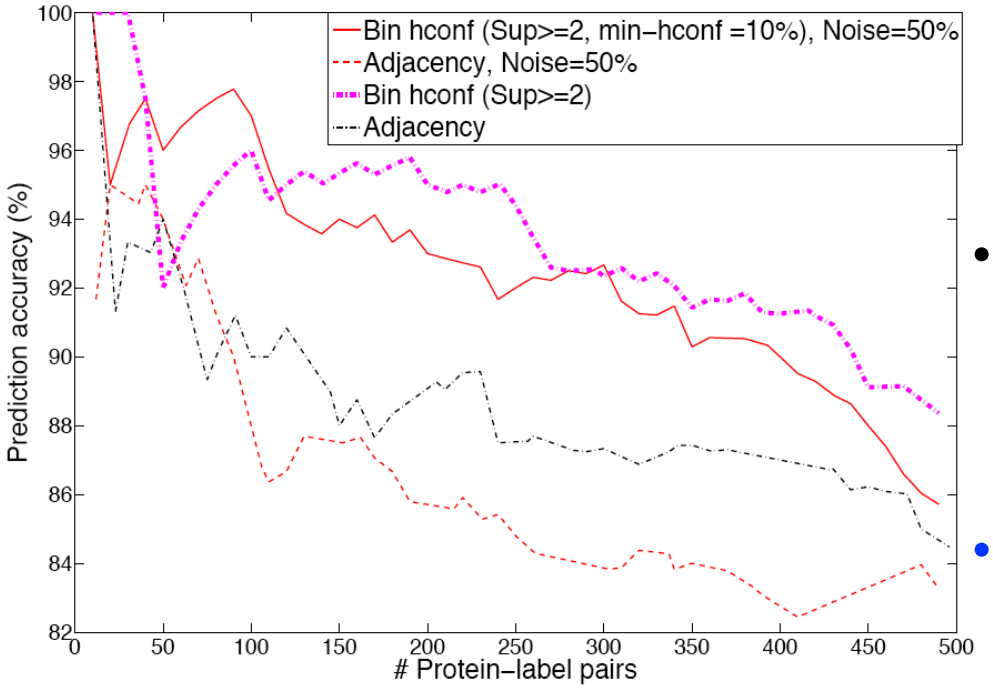


Precision-Recall



Accuracy of top-k predictions

Noise removal capabilities of H-confidence



- H-confidence and hypercliques have been shown to have noise removal capabilities [Xiong et al, 2006]
- To test its effectiveness, we added 50% random edges to DIPCore, and re-ran the transformation process
- **Fall in performance of transformed network is significantly smaller than that in the original network**

Summary of Results

- H-confidence-based transformations generally produce more accurate and more reliably weighted interaction graphs
  - Removes potentially spurious edges
  - Adds potentially biologically viable edges
  - Provides meaningful weights for the resultant edges

- Generally, the less reliable the weights assigned to the edges in the raw network, the greater improvement in performance obtained by using an h-confidence-based graph transformation

Application III: SNP Association Study

- **Given:** A data set that has genetic variations (SNPs) of a group of subjects and their associated Phenotypic (Disease).
- **Objective:** Finding a combination of genetic characteristics that best predicts the phenotype under study.

	SNP_1	SNP_2	…	SNP_M	Disease
Patient $_1$	1	1	…	1	1
Patient $_2$	0	1	…	1	1
Patient $_3$	1	0	…	0	0
…	…	…	…	…	…
Patient $_N$	1	1		1	1

Genetic Variation in Patients (SNPs) as Binary Matrix and Survival/Disease (Yes/No) as Class Label.

SNP (Single nucleotide polymorphism)

- Definition of SNP (wikipedia)
  - A **SNP** is defined as a single base change in a DNA sequence that occurs in a significant proportion (more than 1 percent) of a large population

Individual 1	A G C G T G A **T** C G A G G C T A
Individual 2	A G C G T G A **T** C G A G G C T A
Individual 3	A G C G T G A **G** C G A G G C T A
Individual 4	A G C G T G A **T** C G A G G C T A
Individual 5	A G C G T G A **T** C G A G G C T A

SNP

Each SNP has 3 values

(GG / GT / TT)

(mm / Mm/ MM)

  - How many SNPs in Human genome?
  - About 10,000,000

Why are SNPs interesting?

- In human beings, 99.9 percent bases are same.
- Remaining 0.1 percent makes a person unique.
  - Different attributes / characteristics / traits
    - how a person looks,
    - diseases a person develops.
- These variations can be:
  - Harmless (change in phenotype)
  - Harmful (diabetes, cancer, heart disease, Huntington's disease, and hemophilia)
  - Latent (variations found in coding and regulatory regions, are not harmful on their own, and the change in each gene only becomes apparent under certain conditions e.g. susceptibility to lung cancer)

Issues in SNP Association Study

- In disease association studies, number of SNPs varies from a small number (targeted study) to over a million (GWA studies)
- Number of samples is usually small
- Data sets may have noise or missing values
- Phenotype definition is not trivial (ex. definition of survival)
- Environmental exposure, food habits etc adds more variability
- Complex interaction among a number of genes may be responsible for the phenotype
- Genetic heterogeneity among individuals for the same phenotype

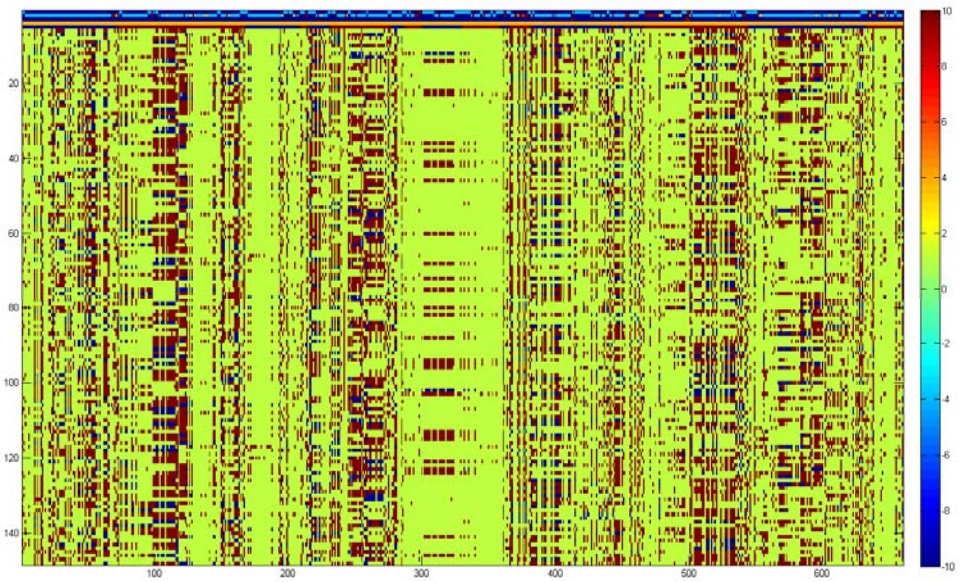
Existing Analysis Methods

- **Univariate Analysis:** single SNP tested against the phenotype for correlaton and ranked.
  - Cannot identify interacting SNPs
- **Multivariate Analysis:** groups of SNPs of size two or more are tested for possible association with the phenotype.
  - Often infeasible in practice
- Some approaches employ classification methods such as SVMs to classify cases and controls.

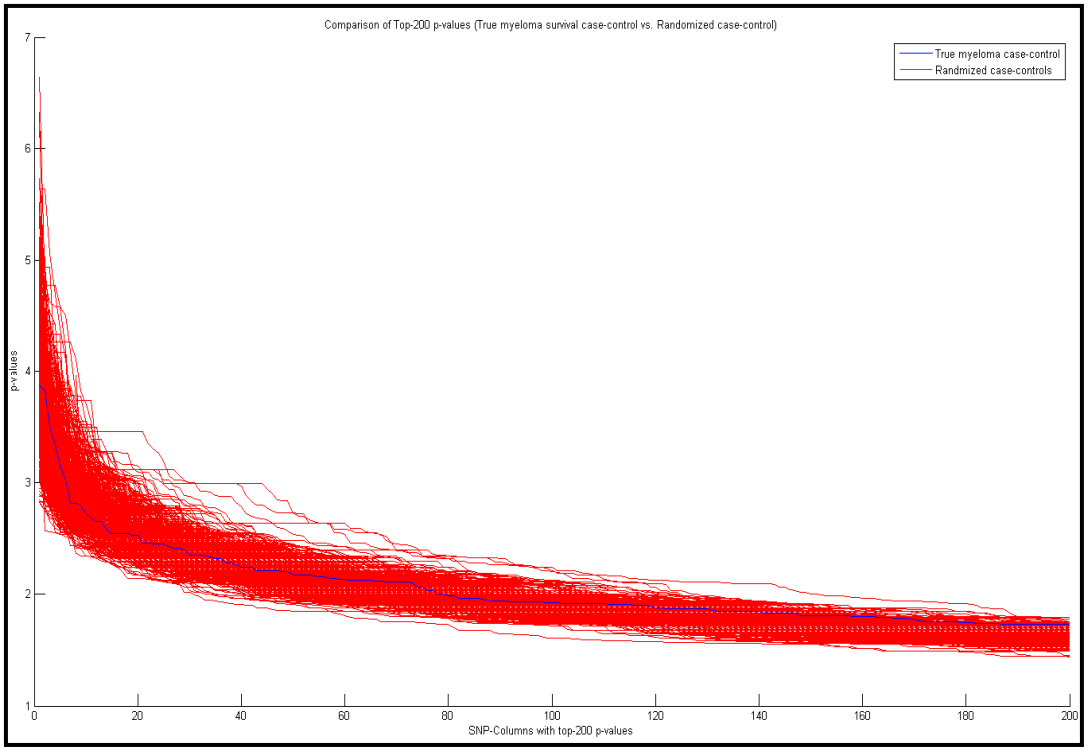
Myeloma Data

SNP data for Myeloma disease:

- 3404 SNPs (Selected according to potential relevance to Myeloma)
- 70 Cases (Patients survived shorter than 1 year)
- 73 Controls (Patients survived longer than 3 years)
- 3404 SNPs come from various regions of the chromosome, i.e., introns, synonymous , non-synonymous, 3’ UTR, 5’ UTR etc.



Univariate Analysis on Myeloma Data



- Individual SNP associations with true phenotype are not distinguishable from random permutation of phenotype.
- A combination of SNPs may be more predictive than individual SNPs.

Evaluating the Utility of Univariate Rankings for Myeloma Data

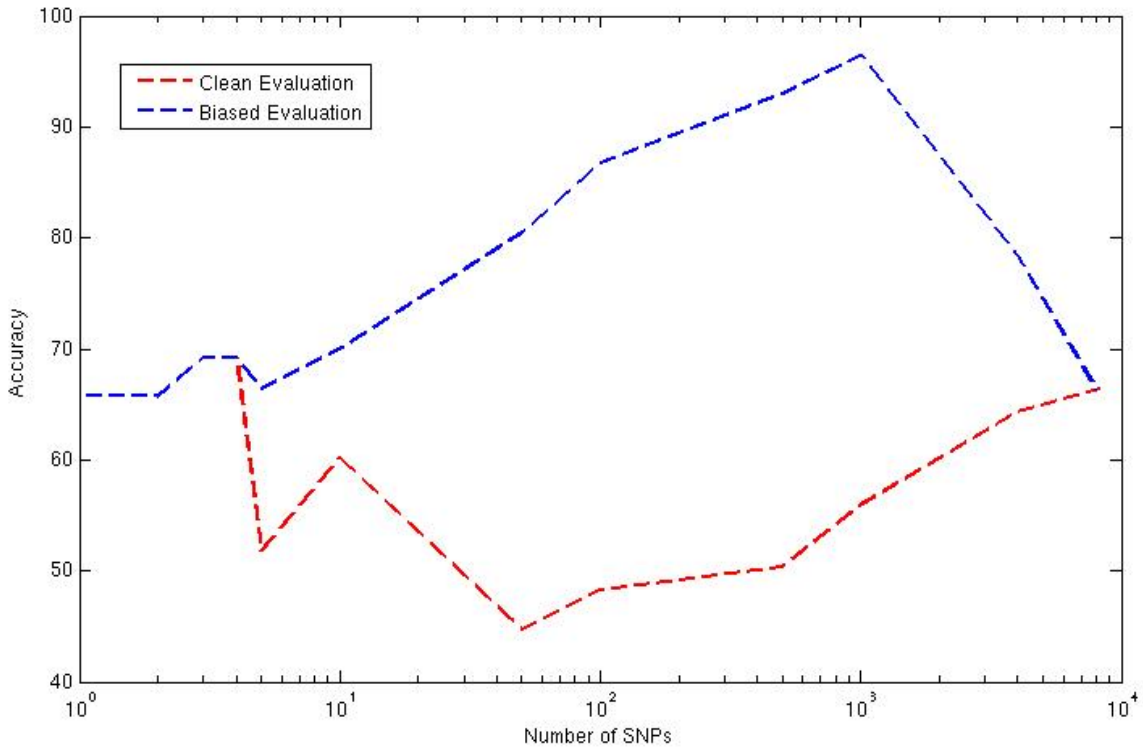
Feature Selection → Leave-one-out Cross validation With SVM

Biased Evaluation

Leave-one-out Cross validation with SVM

Feature Selection

Clean Evaluation



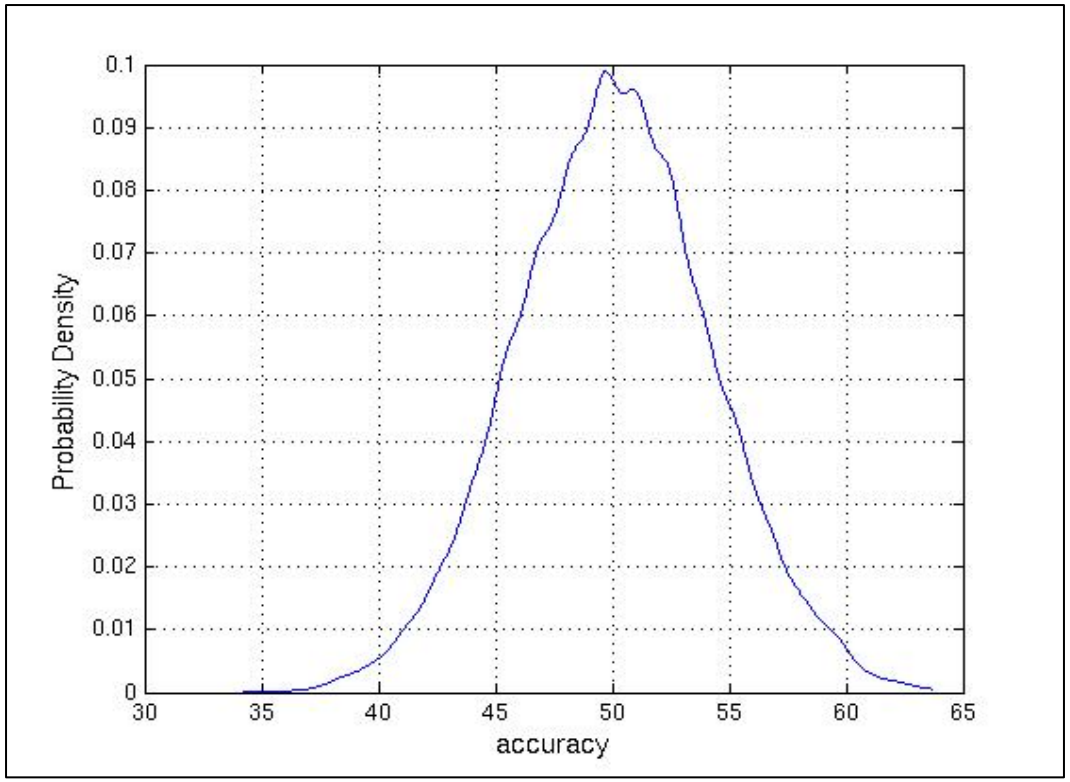
Performance of Predictive Model for Selected categories of SNPs

Non – synonymous	Introns	Synonymous	Admixture	UTR	Other Intergenic	Accuracy (%)
✓	✓	✓	✓	✓	✓	66.43
✓		✓	✓	✓	✓	58.74
		✓	✓	✓	✓	51.74
✓	✓		✓	✓		72.72
✓	✓					71.33
	✓					54.54
✓						69.99

Nonsyn + Promolign (Syn + Introns): 75.75 %

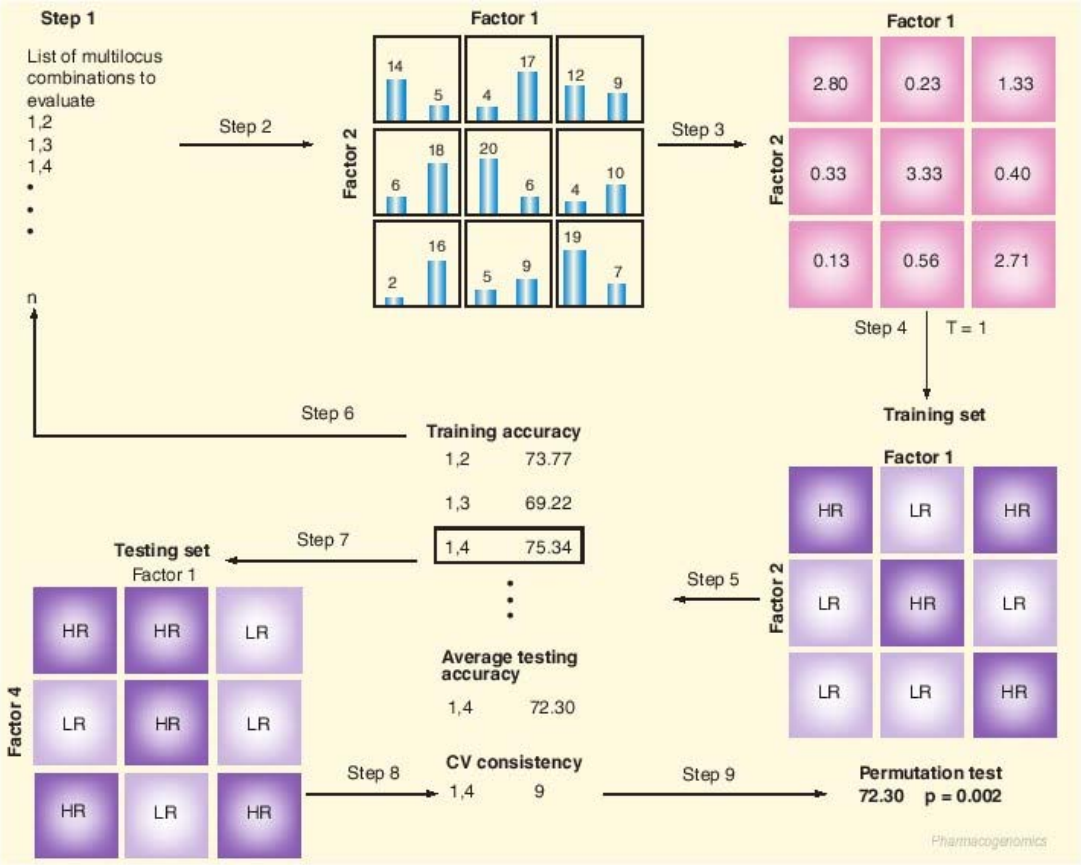
Random Permutation test

- 10,000 random permutations of real phenotype generated.
- For each one, Leave-one-out cross validation using SVM.



- Accuracies larger than 65% are highly significant. (estimated p-value is $< 10^{-4}$)

MDR



Multifactor dimensionality reduction for detecting gene–gene and gene–environm ent interactions in pharmacogenomics studies Marylyn D Ritchie et. a l. Pharmacogenomics 2005

Promise: Finds interactions between SNPs

Drawbacks: Computationally infeasible for large number of SNPs

Statistical significance tends to be low as the number of combinations are large

Error-Tolerant Discriminative Pattern

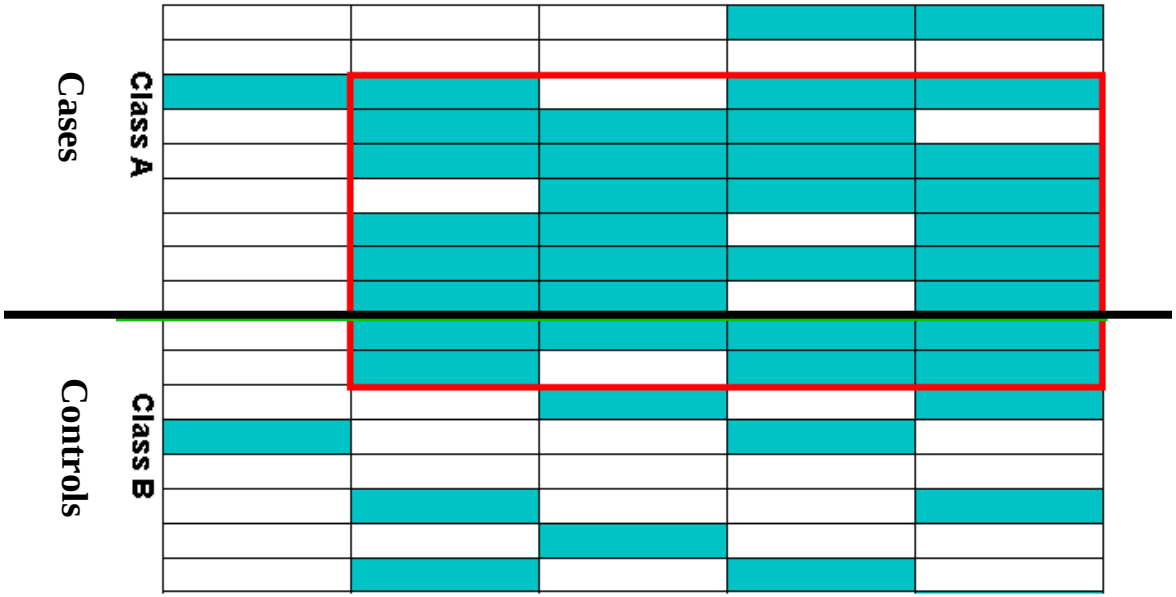
- Handles noise and makes use of the additional class information to prune high dimensional search space
- Error-tolerant support ratio of a pattern P = {i_1, i_2,…, i_m} on two classes

$$\frac{s_{\varepsilon_A}(\{i_1, i_2, \cdots, i_k\})}{s_{\varepsilon_B}(\{i_1, i_2, \cdots, i_k\})}$$

where

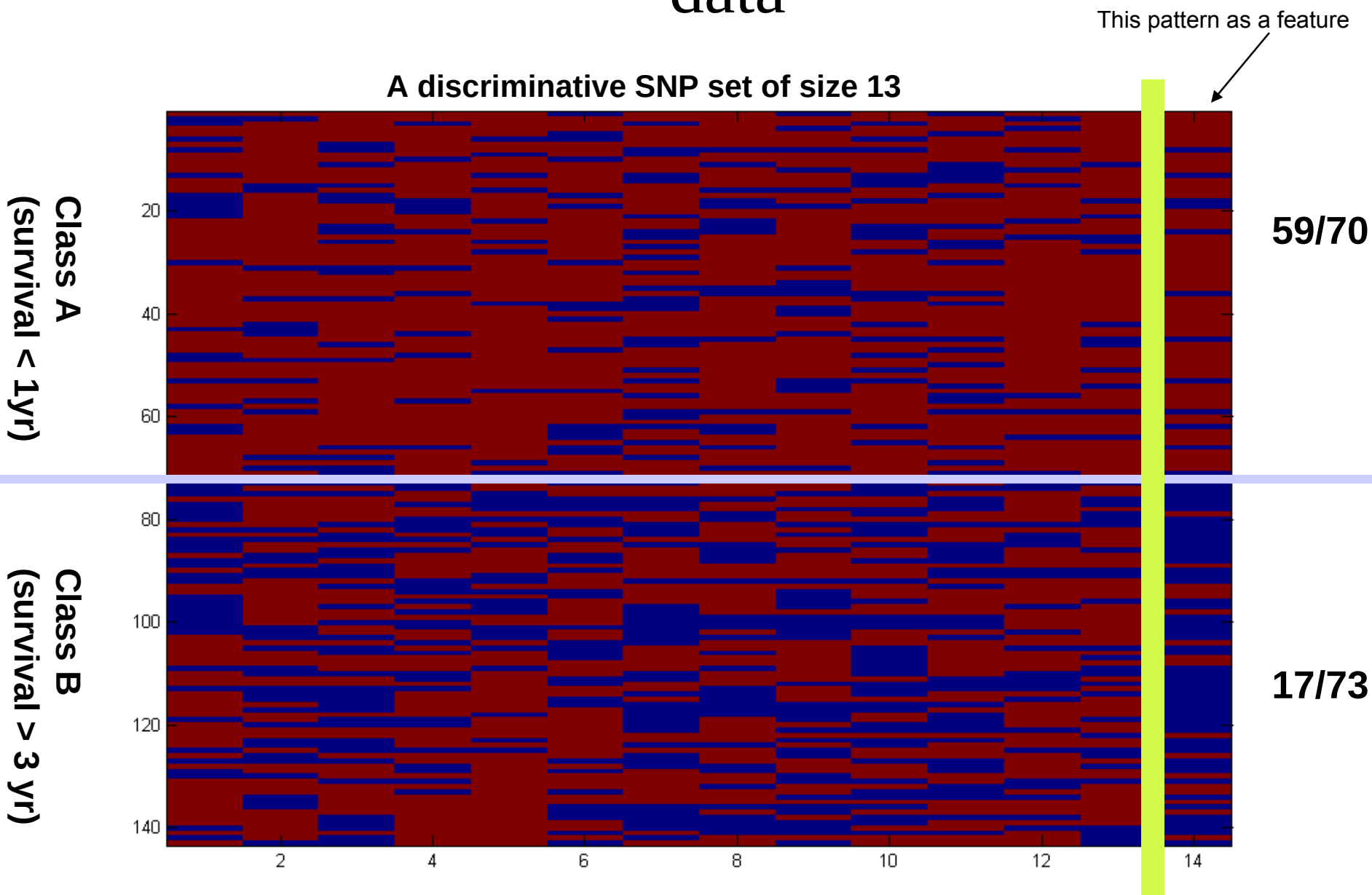
$$s_{\varepsilon}(\{i_1, i_2, \cdots, i_k\})$$

is the number of transactions having at most a fraction ε of the items missing



- A pattern P is an **error-tolerant discriminative pattern** if $s_{\varepsilon}(P) >= \delta$ (in one class) and P's support ratio >=θ, where δ is a user specified minimum support threshold and θ is a user specified minimum support ratio threshold

Error-Tolerant Discriminative Pattern in SNP data



Association Pattern Analysis Challenges

- Handling non-binary data (e.g gene expression data)
- Discovery of patterns in noisy data
  - Error-tolerant patterns
- Incorporating domain knowledge
  - Case/Control
  - Pathways
- Summarization and Evaluation of patterns

Protein Function Prediction Challenges

- Handling multiple labels and complex class structure
  - Commonly used functional classification schemes such as GO are hierarchically structured
  - A protein can have multiple functional labels
  - Traditional classification algorithms and evaluation methods tend to break down
- Integration of multiple types and sources of data
  - Integration can provide noise reduction, since the incorrectness of one data set may be reduced by combining it with other data sets
  - Data sets provide complementary information about protein function
- Effective pre-processing of biological data is needed for
  - Noise reduction
  - Handling widely differing scales for different portions of the data
  - Eliminating portions of the data irrelevant to the target class being studied