Probabilistic frameworks for the functional interpretation of genes

Lorenz Wernisch

MRC-Biostatistics Unit
Cambridge, UK

8 Feb 2012

Functional interpretation of genes?

Biological/experimental:

- Physical-biochemical characterisation
- Biochemical function (kinase, transcription factor)
- Regulatory connections
- Pathway (signalling, metabolic)

Computational:

- GO category of gene or gene set
- Pathway membership
- Connection to annotated and characterised genes

First step: enrichment analysis

- Given collection of pathways (GO, KEGG, Biocyc)
- Priority gene list from experiments (microarrays, GWAS SNP)
- Which pathways have more than random overlap with gene list?
- Hypergeometric test or GSEA (gene set enrichment analysis, weighted Kolmogorov-Smirnov test)

Next step: more details

Want to know which are specific genes interacting with genes from list

- Direct protein-protein interaction
- In a common protein complex
- Jointly lethal (know both out cell not viable any more)
- Direct transcription regulation (binding upstream and modulating gene expression)
- Co-transcribed, co-expressed

Strategies

- Predict individual interaction partners for gene: **classifier** that specific pair interaction exists or not
- List of genes which are related to seed genes sorted by priority: **network analysis**
- Interaction partners and related genes can throw light on function of seed genes, propagate annotation

Integration of several predictors

- Do two proteins interact or not
- Several sources of evidence: experimental (Y2H), computational (literature, GO categories)
- Different quality of predictors

- Assess quality of predictors
- How to combine predictors taking quality into account?

Quality of predictors

Predictor M

Interaction $I = 1$: probability M says 'yes', true positive rate

$$P(M = 1 \mid I = 1)$$

No interaction $I = 0$: probability M says 'no', true negative rate

$$P(M = 0 \mid I = 0)$$

Both should be high for a good predictor (away from 0.5 or prior P(M))

Obtain quality from gold standard

For a gold-standard set of known interactions estimate

$$P(M = 1 \mid I = 1) = \frac{\text{number } \{I = 1 \text{ and } M = 1\}}{\text{number } \{I = 1\}}$$

Run predictor on gold standard set, count successes and failures

Possible: run without gold standard set, I hidden variable, train with Expectation-Maximisation (EM) algorithm if several predictors are available

Converges on (hidden) consensus solution among predictors

Posterior odds

If predictor says yes, how much more likely is interaction than noninteraction?

$$\frac{P(I = 1 \mid M = 1)}{P(I = 0 \mid M = 1)}$$

but we only have $P(M = 1 \mid I = 1)$ (the wrong way round)

$$P(I = 1 \mid M = 1) = \frac{P(M = 1 \mid I = 1)P(I = 1)}{P(M = 1)}$$

Bayes: posterior = likelihood * prior / normalisation to 1

Bayes factor

If predictor says interaction $M = 1$, how much more likely is interaction than noninteraction?

$$\frac{P(I = 1 \mid M = 1)}{P(I = 0 \mid M = 1)} = \frac{P(M = 1 \mid I = 1)P(I = 1)}{P(M = 1 \mid I = 0)P(I = 0)}$$

posterior odds = bayes factor * prior odds

Convert odds to probabilities

$$o = \frac{P(I = 1 \mid M = 1)}{P(I = 0 \mid M = 1)}$$

$$P(I = 1 \mid M = 1) = \frac{o}{1 + o}$$

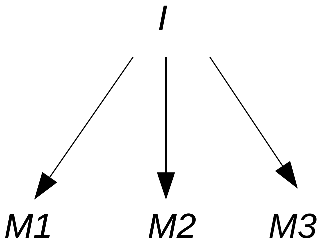
Cutoff for posterior odds

Depends on costs of false discovery C_{FD} vs false nondiscovery C_{FND}

Costs minimised for discovery whenever

$$\text{posterior odds} > \frac{C_{\mathrm{FD}}}{C_{\mathrm{FND}}}$$

Naive Bayes



Three (lousy) classifiers with

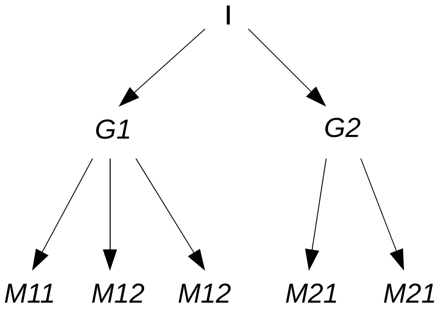
$$P(M_i = 1 \mid I = 1) = 0.7$$
$$P(M_i = 0 \mid I = 0) = 0.6$$

Assume $M_1 = 1,\ M_2 = 1,\ M_3 = 0$ and $P(I = 1) = 0.5$

$$\frac{P(I = 1 \mid M)}{P(I = 0 \mid M)} = \frac{\prod P(M_i \mid I = 1)}{\prod P(M_i \mid I = 0)} = \frac{0.7 * 0.7 * 0.3}{0.4 * 0.4 * 0.6}$$

$$P(I = 1 \mid M_1, M_2, M_3) = 1.53/(1 + 1.53) = 0.6$$

Hierarchical Naive Bayes

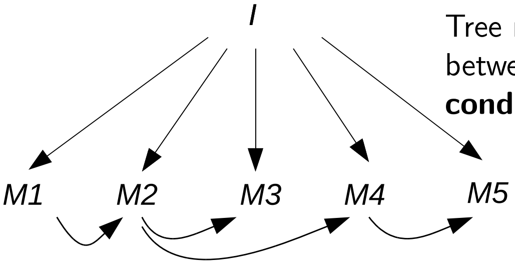


Additional dependencies since methods in *G1*, *G2* are related

Dependencies among methods (over and above interaction) distort Bayes factors

True G_1, G_2 unknown (no gold standard)
more complex algorithm (EM, variational) for estimation

Naive Bayes with tree dependence

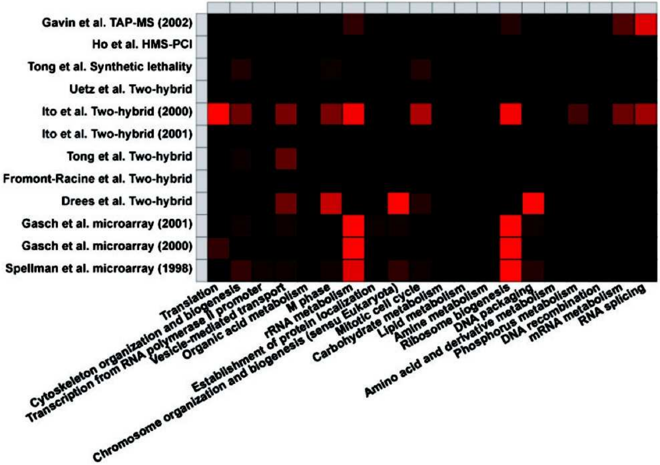


Tree models dependence between predictors **conditional on *I***

Good tree can be easily estimated by *Chow-Liu* procedure:

Complete graph with edges weighted by conditional (on *I*) mutual information
Find maximum weight spanning tree (add heaviest edge to growing forest)
This is a maximum likelihood tree!

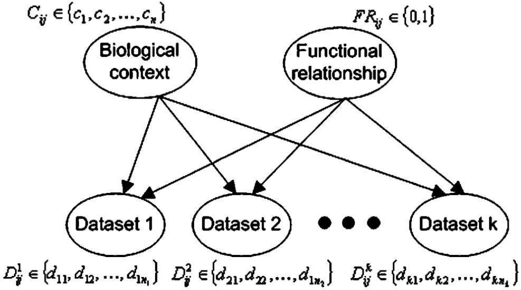
ContextPixie



(from Myers, Troyanskaya, 2007)

Intensity of red: area under ROC of classifying genes from GO sets correctly in leave-one-out

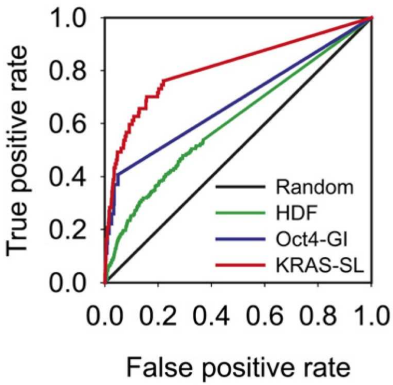
ContextPixie


$C_{ij} \in \{c_1, c_2, \ldots, c_n\}$
Biological context
$FR_{ij} \in \{0,1\}$
Functional relationship
Dataset 1
Dataset 2
● ● ●
Dataset k
$D_{ij}^1 \in \{d_{11}, d_{12}, \ldots, d_{1n_1}\}$
$D_{ij}^2 \in \{d_{21}, d_{22}, \ldots, d_{1n_2}\}$
$D_{ij}^k \in \{d_{k1}, d_{k2}, \ldots, d_{kn_k}\}$

(from Myers, Troyanskaya, 2007)

The biological context helps to pick suitable true/false positive rates $P(D_i = d_i \mid F = i, C = c)$ in naive Bayesian classifier

What can we hope for, Lee et al., 2011



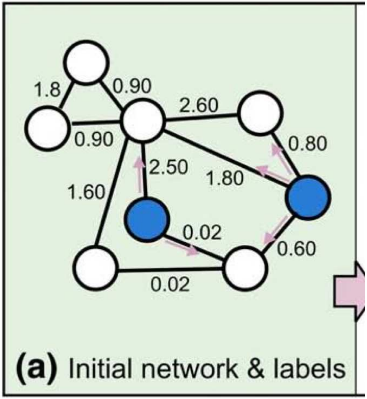
HDF: Host factors for HIV

Genes modulating Oct4 (stemness regulator)

KRAS interaction partners with lethal knockdowns in colorectal cancer cell line

HumanNet covers about 500,000 links between 87% of human proteins

Label propagation in network



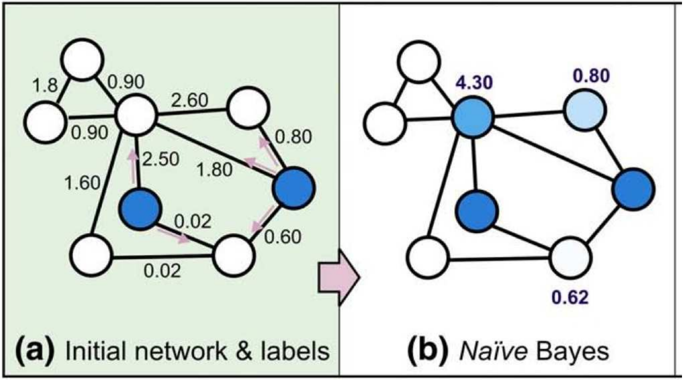
(a) Initial network & labels

(from Wang, Marcotte, 2010)

- Weighted edges between nodes
- Values on nodes (eg +1 for GO class genes, -1 for all others)
- Some values known (blue)
- Propagation to other nodes

Types of methods:
direct neighbour propagation, *indirect* neighbour propagation

Direct neighbour propagation



Direct neighbours of seed nodes get weighted average values

Indirect neighbours: iterative ranking

Basis of Google ranking (PageRank)

Graph of nodes with edge weights W_{ij}

Given background values $f_{i,0}$ for each node

Each node i gets new value f_i composed of:

- a proportion of background $\alpha f_{i,0}$
- weighted sum of neighbor values $(1-\alpha)\sum_j W_{ij} f_j$

$$f(t+1) = \alpha f(0) + (1-\alpha) W f(t)$$

Iterative ranking solution

Iterative ranking converges towards stationary solution

$$f - (1-\alpha)Wf = \alpha f(0)$$

or

$$f = \alpha(I - (1-\alpha)W)^{-1} f(0)$$

if largest absolute eigenvalue of $(1-\alpha)W$ is less than 1

(Iterative method might be more efficient than inversion of big, although sparse, matrix)

Indirect neighbours: Gaussian smoothing

Find f that minimizes

$$Q(f) = \frac{1-\alpha}{2} \sum_{i,j} W_{ij}(f_i - f_j)^2 + \alpha \sum_i (f_i - f_i(0))^2$$

$W_{ij}(f_i - f_j)^2$ encourages similar f's for neighbours with strong connection W_{ij}

$(f_i - f_i(0))^2$ ties values f to known initial values $f(0)$ (eg +1,-1 for known labels, and 0 else)

(Note: factor of $1/2$ missing in almost all papers on this topic, don't rely on equations in papers)

Gaussian smoothing solution

With $D = \mathrm{diag}(d_i) = \mathrm{diag}(\sum_j W_{ij})$

$$Q(f) = (1-\alpha)f'(D-W)f + \alpha(f - f(0))'(f - f(0))$$

after (vector) differentiation by df and setting to 0

$$(1-\alpha)(D-W)f + \alpha f - \alpha f(0) = 0$$

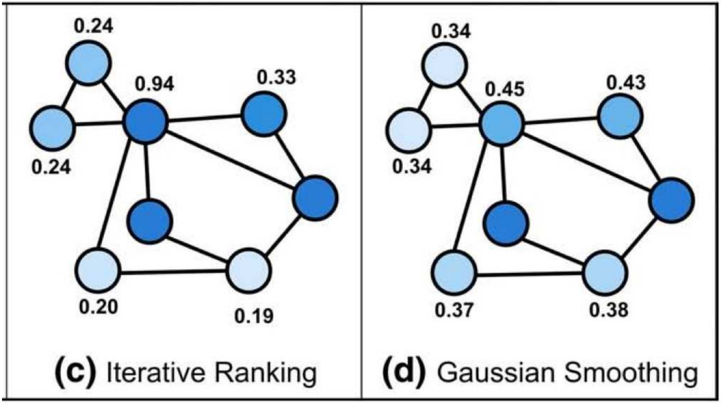
or

$$f = \alpha(S - (1-\alpha)W)^{-1})f(0)$$

with $S = \alpha I + (1-\alpha)D$

Is *iterative ranking* when $D = I$, ie $\sum_j W_{ij} = 1$

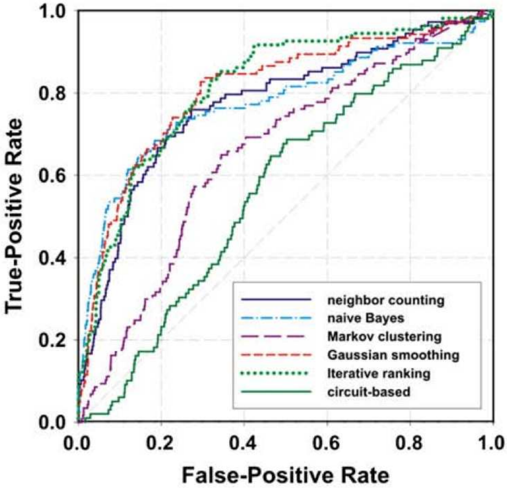
Indirect propagation results



(from Wang, Marcotte, 2010)

Gaussian smoothing basis of GeneMANIA (Mostafavi et al., 2008)

Comparison of methods, Wang, Marcotte, 2010



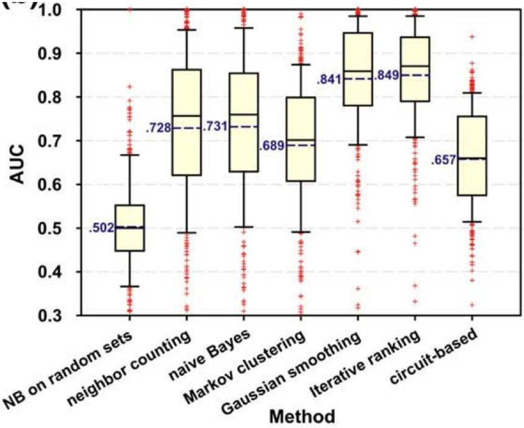
C. elegans abnormal locomotion genes

10-fold cross validation ROC curves, ie, 10 times: 90% as seed genes, 10% query genes

Indirect methods overall better, better for FP rate

Direct methods not too bad, particularly for low FP rates

Comparison of methods, Wang, Marcotte, 2010



C. elegans causal genes in 318 RNAi phenotypes

10-fold cross validation ROC curves, ie, 10 times: 90% as seed genes, 10% query genes

Similar results for yeast network

Circuit-based: voltage in circuit with weights as 1/resistance, MCL clustering based on graph flow

Combining matrices in GeneMania

Best μ_i for $K = \sum \mu_i K_i$?

- From seed genes derive matrix T with $T_i j$ positive if i, j in seed set, negative if one is in the other out, NA else
- Vectorize T into t dropping NAs by columns
- Vectorize each K_i and collect vectors as columns in matrix Ω
- Solve regularized regression by minimizing

$$(\Omega\mu - t)'(\Omega\mu - t) + (\mu - \mu_0)'S(\mu - \mu_0)$$

with some regularisation parameters μ, S since Ω is sparse

Thoughts

- Rapidly growing databases (HumanNet)
- Situation promising for key organisms (yeast, C. elegans, mouse, human)
- Usable precision recall achievable with combination of sources and networks

Still missing

- Link with more mechanistic aspects (regulation, signalling)
- Neglected organisms (most!)
- Quality control in danger of circularity (ubiquitous GO, KEGG) and networks

References

Insuk Lee, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research* 21:1109-1121, 2011

Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q., GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function, *Genome Biol.* 9 Suppl 1:S4. Epub 2008

Myers CL, Troyanskaya OG., Context-sensitive data integration and prediction of biological networks. *Bioinformatics* 23(17):2322-30, 2007

References

B Schölkopf, K Tsuda, JP Vert, *Kernel Methods in Computational Biology*, MIT Press, 2004

Wang PI, Marcotte EM, It's the machine that matters: predicting gene function and phenotype from protein networks, *Journal of Proteomics*, 73(11):2277-89, 2010