Predicting Gene Function Using SVMs and Bayesian Networks

Laura Miron, Benoit Pit--Claudel
{lmiron, bpitcla}@stanford.edu

Introduction / Related Work

- Determining the function of genes experimentally is often costly in time and money.
- Machine learning has been used to predict gene function, using features such as sequence, pairwise interaction, histone markers, and more.
- Most previous work handles protein functions independently, ignoring the structure between functions.
- Our work trains svm classifiers on individual GO nodes, then feeds the output into a Bayesian network representing the relationship between nodes.
- Our work aims at reproducing and improving on a method developed by Barutcuoglu & al [1].
  → Focus on Saccharomyces cerevisiae

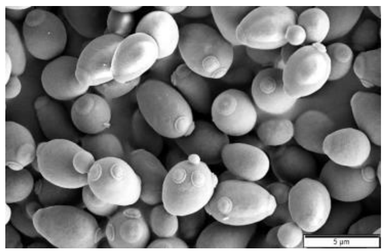

Figure 1. Saccharoyces cerevisiae (yeast) [2]

SVM Results

- Before running the SVM on all go nodes, we perform a parameter search on the most represented GO node (GO:0045944, positive regulation of transcription by RNA polymerase II)
- Below, graphs comparing kernels (linear and rbf with different γ values) and penalty parameters C of the error term.

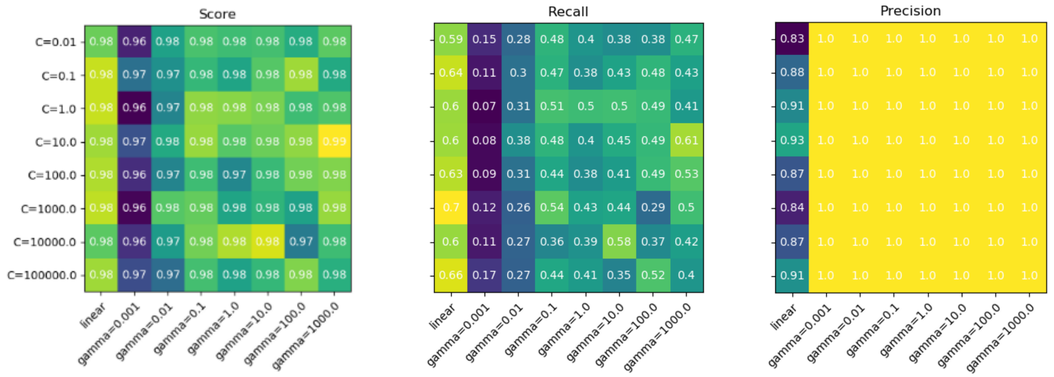

Figure 2. Results of parameter search on linear kernel and GO:0045944

- Very good accuracy to be expected since the proportion of positive examples for each classifier is very low
- Most important metric should be recall: a false positive should be compensated by the Bayesian network, whereas the false negative could have more impact
- Average accuracy for all nodes over 97.7%

Methods

- In our final classifier, we train one 10-Ensemble SVM per gene ontology node, using a linear kernel and $C = 1.0$
- One challenge in gene prediction is the small number of examples overall, and in particular the small number of positive examples for each GO node; we therefore use bootstrapped samples with replacement to train the classifiers

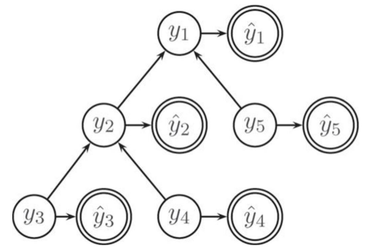

Figure 5. Bayes net structure [Barutcuoglu]

- We create a Bayes net where $\hat{y}_i$ represents the svm prediction for label *i*, and y_i represented the true value for membership in *i*
- $P(\hat{y}_i \mid y_i)$ is calculated during svm training using *maximum likelihood estimation*
- $P(y_i O \mid ch(y_i))$ is inferred by counting from the training labels
- Finally, for a given training example/ assignment to all $\hat{y}_i$, we use Bayesian exact inference to find the most likely assignment to all y_i

Data and Features

Labels

- Boolean membership in each of **95** selected Gene Ontology (GO) classes [3].

Pairwise Interactions

- BioGRID protein interaction data for **5395** proteins [4],
- **5394** Boolean features.

Microarray Expression Levels

- Microarray gene expression levels for same proteins obtained through [5],
- Microarray data has one or more missing columns for each example, which we complete using KNN,
- **161** float features

--> 5555 features total, mix of floats and Booleans.

Bayesian Results

- Due to bugs in the library **pgmpy,** we are currently unable to make inferences on the full net of 95 nodes
- We predict on the nodes shown in fig.3, and, as shown in fig. 4, obtain higher accuracy than the svm alone in all cases

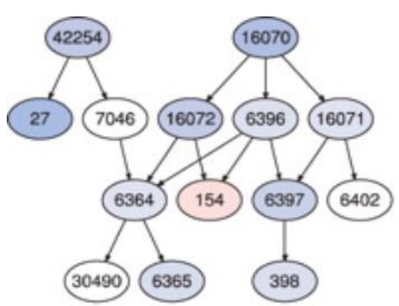

Figure 3. Hierarchical relationship between GO nodes, *Barutcuoglu, et al.*

Node	Svm Accuracy	Bayes Accuracy
GO:0042254	0.984	1.0
GO:0016070	0.953	1.0
GO:0016072	0.984	.998
GO:0000154	0.956	1.0
GO:0030490	0.998	1.0
GO:0006402	0.995	1.0
GO:0016071	0.987	1.0
GO:0000027	0.994	1.0
GO:0000398	0.995	0.998
GO:0006364	0.988	0.998

Figure 4. Individual svm accuracy vs. Bayes net accuracy for selected GO nodes

Conclusions

- Improved accuracy on invidual svm classifiers compared to Barutcuoglu & al [1]
- Where Barutcuoglu & al had better results with C->+∞ and using an rbf kernel, we obtained better results with a linear kernel.
- Refinements in GO classification between 2006 and today might explain better results.
- Their bootstrapping procedure is effective for dealing with very few positive examples and possibly uncertain examples
- Useful for newly discovered species where little data is available

References

[1] Zafer Barutcuoglu & al., Hierarchical multi-label prediction of gene function, Bioinformatics, 22(7):830–836, 2006.
[2] By Mogana Das Murtey and Patchamuthu Ramasamy - [1], CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=52254246
[3] Ashburner M & al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.
[4] Stark C & al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. 2005;34(Database issue):D535-9.
[5] NBCI Gene Expression Omnibus, Tanya Barrett & al. Ncbi geo: archive for functional genomics data sets–update. Nucleic Acids Research, 41(D1):D991–D995, 2013.