Filling missing components in yeast metabolic pathways using heterogeneous data

Gal Chechik[1], **Aviv Regev**[2] **and Daphne Koller**[1]
[1] Computer science department, Stanford, [2] Center for Genome Research, Harvard

Abstract

The set of cellular metabolic reactions forms a complex network of interactions, but even in well studied organisms the resulting pathways contain many unidentified enzymes. We study how 'network relations' among genes in the yeast metabolic pathway are manifested in functional properties of genes and their products, including mRNA expression, protein domain content and cellular localizations. We develop compact and interpretable probabilistic models for representing protein-domain co-occurrences and gene expression time courses. The former can provide predictions relating domains and gene functions. The latter reveals relations between the activation of genes and the usage of their protein products in the pathways. These models are then combined and used for completing unidentified enzymes in the pathways, achieving accuracy that is superior to existing state-of-the-art approaches.

1 Introduction

Metabolic pathways present a unique opportunity to use machine learning techniques for pathways analysis and reconstruction. Large parts of the pathways are already well known, making it possible to use supervised learning for learning properties of the known sub-pathways and then use them to reconstruct unidentified ones [7, 8]. Metabolic processes are of the most fundamental processes in the cell, and are involved in virtually all other processes. As such, their reconstruction is crucial for understanding various metabolic diseases such as diabetes and obesity. However, even in intensively studied organisms like yeast, 20% of the reactions still have unidentified enzymes (30-80 % in other organisms). The most common approach for reconstructing unknown sub-pathways is by sequence homology with other organisms, and this approach is successfully applied in many cases. The current paper takes an orthogonal approach: we learn how network relations between enzymes are reflected in various of their properties, and then use these to find the best candidates that would catalyze a reaction.

Integration of multiple data sources is found to be particularly important in the task described here. However, we find that even data integration is not sufficient unless each specific data type is carefully modeled for best extraction of the relevant signals. This paper therefore follows the following format: we start by describing our general approach, and then Sections 3 and 4 describe efficient representations of two types of data: gene expression time courses and protein domains. We condlude by describing results with the combined data types and discuss that combinations.

2 Functional data and network motifs

Enzymes located within a small neighborhood in the pathway, typically participate in the same process. As a result, various properties of neighboring enzyme pairs are dependent. This is true for example for their expression profiles, cellular localization and protein domains found in neighboring enzymes. We therefore aim to learn how properties of neighboring enzymes are co-related, and specifically, to learn a weighted measure of how each gene is related to its neighbors.

Importantly, different types of neighborhood relations between enzyme pairs lead to different relations of their properties. For example enzymes in a linear chain, depend on the preceding enzyme product as their substrate. Hence it is expected that the corresponding genes are co-expressed, and their RNA levels co-vary across conditions [6]. On the other hand, enzymes in forking motifs (same substrate, different products) were found to have anti-correlated expression profiles [5].

To preserve the distinction between different neighbor relations, we defined a set of network motifs, including *forks* (same input different outputs), *funnel* (same output different inputs), *linear chains* (enzymes at various distances on a linear path). Enzyme relatedness is measured separately for each type of neighbor relation. The specific measures of relatedness depends on the type of data (expression, localization), and is discussed in details below. For a given reaction, we calculate a set of relatedness features for any enzyme which is a candidate to catalyze the reaction. This set of features therefore varies both by type of neighbor and type of data.

Unfortunately, not all enzymes share the same types of neighbors. For example, enzymes in a linear chain have *upstream* neighbors but may not have any *forking* neighbors. As a result, some of the features of each sample are structurally missing. This is an inherently different case from features that are missing due to noise, since each sample lies in a subspace of the full features space. The solution to this problem is different from setting the values of the missing x entries to zero, because the norm of w must be correctly normalized in each projected subspace. We show however that the classical SVM formulation can be extended to handle structural zeros, by correct normalization of each dimension of w, which can be viewed as a modified kernel.

With this representation of enzymes' relatedness to their neighborhood, we train a classifier using a set of positive samples created from known enzymes with their known neighbors, and a set of negative samples, created from known neighborhoods, but with an impostor enzyme plugged in, instead of the correct enzyme that catalyses the reaction. We trained an SVM, using standard 20-80 cross validation procedures for optimizing over standard kernels (polynomial, RBF), and parameters. All results reported below are averages over five validation sets.

Unfortunately, using naive similarity measures between enzymes and their corresponding genes resulted in relatively low accuracy of 60%. We therefore proceed to discuss more efficient representation of two of the data types that we have used.

3 Modeling gene expression time courses

Gene expression experiments are often conducted by imposing cells to environmental perturbations, and then measuring RNA levels at several time points. This type of RNA expression data poses a difficult modeling challenge, since one aims to learn about the complex underlying biophysical dynamics from a few measurements taken at irregular intervals. Previous approaches (e.g. [1]) used splines to encode continuous gene expression profiles, but since splines are a fully general function representation, they are more prone to overfit-

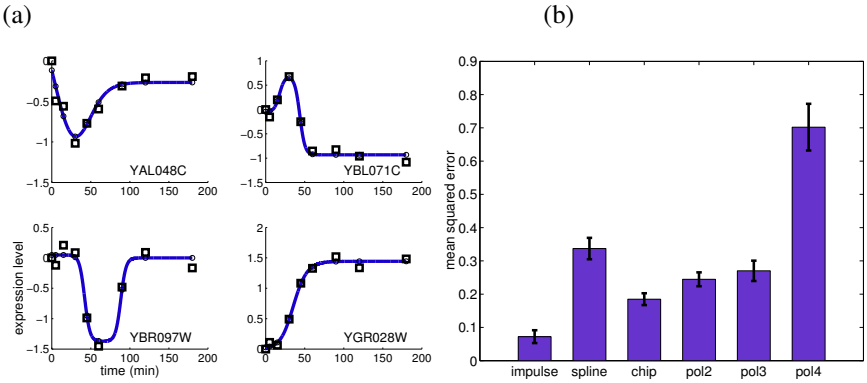

Figure 1: (a) Examples of impulse model fit (solid line) to expression of genes (squares) responses to $1M$ sorbitol stress, as described in [3]. (b) Quantitative evaluation, using the following procedure: The model is fit separately for each time course (gene) after hiding one of the data points, and then used to predict the missing value. This is repeated across all measurements and genes. The impulse model is found to out-perform other models, including polynomials, splines and Hermit interpolation. Other evaluations show that impulse representation is largely effective in measuring similarity between gene profiles (data not shown due to lack of space).

ting. As a consequence, model parameters are estimated using clusters of genes, making it inapplicable for the current task.

Here we make a deliberate decision to use a restricted model capturing a limited set of biologically motivated temporal profiles, which follow an impulse-like time course. The models are rich enough to capture a wide variety of expression behaviors but simple enough to be learned from sparse data, and fit a temporal pattern even to the expression data of a single gene

When a cell's environment is perturbed, a clear pattern is often observed among the responding genes; A gene's transcript level may rise, peak at a certain point, and then level off at a new steady state. We use an *impulse model* designed to encode precisely this type of behavior. The impulse model encodes this behavior as a product of two sigmoid functions, one that captures the onset response, and another that models the offset one. The model function has six free parameters, $\theta = [h_0, h_1, h_2, t_1, t_2, \beta]$, encoding the heights, the transition times and the slope, and is formally written as $f_\theta(x) = \frac{1}{h_1} \cdot s_1(x) \cdot s_2(x)$, where $s_1(x) = h_0 + (h_1 - h_0)S(+\beta, t_1)$ is the first (rescaled) sigmoid, and $s_2(x) = h_2 + (h_1 - h_2)S(-\beta, t_2)$ is the second. Fig. 1a plots several examples of the model's fit to real data. and Fig. 1b presents a quantitative evaluation of its quality.

Once the impulse model is fit, each set of measurement in time is transformed to parameters' space. We then used Mahalanobis distances in the impulse space to quantify the measure of similarity and relatedness between the temporal profiles of two enzymes in the pathway.

4 proteins and domains co-occurrences

Protein domains are preserved sequences corresponding to protein substructures that often correspond to specific enzymatic functions. The distribution of domains across enzymes pairs could therefore carry a strong predictive signal. In practice, protein similarity measures based on their raw domains are unreliable, because the large number of domains leading to very sparse protein-domains distribution. As a result, naive similarity measures yield poor predictions.

We address this problem by mapping proteins and domains to a common low dimensional

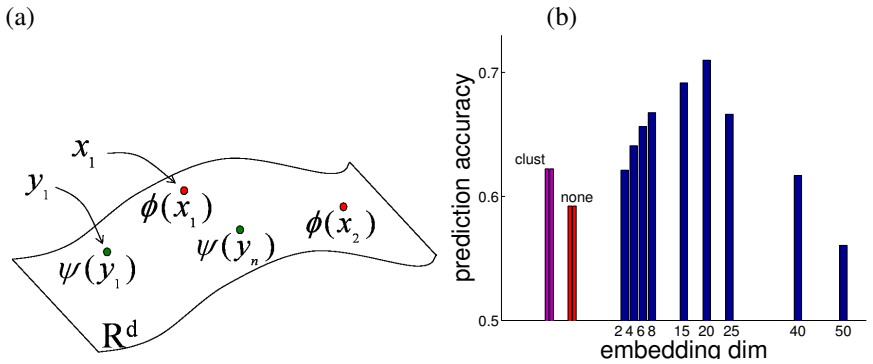

Figure 2: (a) Illustration of embedding of categorical data in a lower dimensional manifold. (b) Classification accuracy as a function of embedding dimensionality using CODE.

manifold, based on their observed co-occurrence. Using *co-occurrence data embedding* (CODE) [4]. we model the empirically observed co-occurrences as depending on the distance between embedded objects. According to the model, the conditional probability of observing a domain y in a protein x decays with the distance between their embeddings

$$p(y|x) = \frac{\bar{p}(y)}{Z(x)} e^{-d^2_{x,y}} \quad \forall x \in X, \forall y \in Y \tag{1}$$

where X is the set of proteins, Y is the set of domains, and $d^2_{x,y}$ is the distance between the embedding locations $\phi(x)$ and $\psi(y)$. The parameters of the models are the embedding locations $\vec{\phi}$ and $\vec{\psi}$, which are optimized to maximize the log likelihood of the observed data $\max_{\vec{\phi},\vec{\psi}} l(\vec{\phi},\vec{\psi}) = -\sum_{x,y} \bar{p}(x,y) d^2_{x,y} - \sum_x \bar{p}(x) \log Z(x)$.

We used CODE to co-embed proteins and domains from the Prosite domains database `http://us.expasy.org/prosite/`. The Euclidean distance between proteins in the embedded space was be as a measure of their relatedness. Figure Fig. 2b plots the classification accuracy obtained when using this similarity measure, as a function of the embedding dimensionality. Using CODE improves accuracy to 72 %, using domains data only.

5 Results

We first tested the classification accuracy in a binary classification task: a balanced set of positive and negative samples was created as explained above. For the negative set, genes that are known to code for metabolic enzymes were used, but plugged into the wrong pathway (network neighborhood). Expression data consisted of 642 experimental conditions, including 17 time courses from [3] and [2]. Cellular localization data was taken from [].

Fig. 3a compares the classification accuracy obtained using different methods. While using naive measures of similarity only achieves chance level prediction accuracy, combining the more detailed models of gene expression time courses and protein similarity by domains content achieves 88 % accuracy.

Fig. 3b shows the performance of our method in a ranking task. For each reaction, a set of candidate genes are ranked according to their relatedness to the neighboring enzymes. When candidates were chosen from the set of known metabolic enzymes, the correct enzyme was ranked first or second in almost 20% of the cases. When candidates are chosen from the set of non metabolic genes (an easier task), 85 % of the genes were ranked in the top 1% of the list (first 50). This is considerably better than previous results in the literature, where only 25 % of the genes where ranked in the top 1%. [7]. (We are aware however of better yet unpublished results achieving 65 % for the same task).

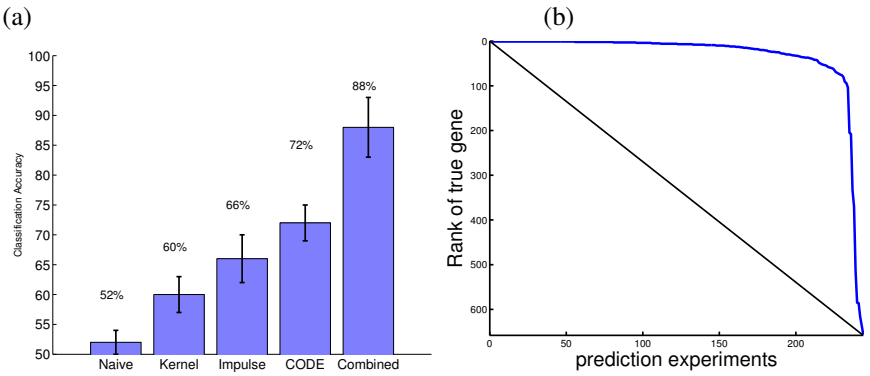

Figure 3: (a) classification accuracy of several models. **Naive** using corelation for expression similarity, D_{KL} for domains similarity and hamming distance for localization similarity. **Kernel** Same as naive but with normalization kernel. **Impulse** Using gene expression only, similarity measured as Mahalanobis distance in impulse space **Code** using domains data only. Similarity measured as Euclidean distance in embedded space. **Combined** using expression, time-courses, domains and localization (b) ranking results. The rank of the true gene in 250 ranking experiments

6 Discussion

Heterogeneous data typically improves learning because the integration of multiple noisy data sources successfully averages out the noise. Interestingly in the current problem the noise is unevenly distributed across the features space, since different data types have have a variable reliability for different network motifs. Specifically, we analyzed the signal's strength obtained from temporal patterns of gene expression, and find that it largely contributes to prediction of linear chain relations but less so to parallel and forking relations. On the other hand Protein domains are often found to repeat in enzymes with a forking relation. This supports the idea that measuring relatedness separately for different network motifs is highly important in the current task.

References

[1] Z. Bar-Joseph, G. Gerber, D. Gifford, T. Jaakkola, and I. Simon. Continuous representations of time series gene expression data. *Journal of Computational Biology*, 10(3-4):241–256, 2003.

[2] A.P. Gasch, M. Huang, S. Metzner, S.J. Elledge, D. Botstein, and P.O Brown. Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p. *Mol. Biol. Cell*, 12(10):2987–3003, 2001.

[3] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11:4241–4257, 2000.

[4] A. Globerson, G Chechik, F Pereira, and N. Tishby. Eucledan combedding of co-occurrence data. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *NIPS 17*, Vancouver, Canada, 2004.

[5] J. Ihmels, R. Levy, and N. Barkai. Principles of transcriptional control in the metabolic network of saccharomyces cerevisiae. *Nature Biotechnology*, 22:86–92, 2003.

[6] P. Kharchenko, GM Church, and D. Vitkup. Expression dynamics of a cellular metabolic network. *Molecular Systems Biology*, 2005.

[7] P. Kharchenko, D. Vitkup, and GM Church. Filling gaps in a metabolic network using expression information. *Bioinformatics*, 2003.

[8] JP. vert and Y. Yamanishi. Supervised graph inference. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *NIPS 17*, Vancouver, Canada, 2004.