

An EM algorithm for unambiguous assignment of genes to biochemical pathways

Liviu Popescu¹ and Golan Yona¹

Keywords: cellular pathways, expression data, expectation-maximization

Introduction

Accurate assignment of genes to pathways is essential to understand the functional role of genes and to map the existing pathways in a given genome. Existing algorithms predict pathways by extrapolating experimental data from one organism to other organisms for which this data is not available. Current systems[1, 2] classify all genes that belong to a specific enzyme family to all the pathways that contain the corresponding enzymatic reaction, and thus introduce ambiguity. Here we describe an EM algorithm which solves this ambiguity by computing assignment probabilities of genes to pathways. Our algorithm uses the set of pathways from MetaCyc [3], together with statistical models of enzyme families and expression data [4] to estimate assignment probabilities which optimize the correlated co-expression. Our algorithm also identifies alternative (“backup”) genes and addresses the multi-domain nature of proteins. We apply our model to assign genes to pathways in the Yeast genome and compare the results for genes that were assigned experimentally.

Methodology

Given a genome G with N genes, enzyme families $F_1, F_2, ..F_M$ and pathways $P_1, P_2, ..., P_K$ we define the set of hidden variables X_{ijk} , $1 \leq i \leq N$, $1 \leq j \leq M$, $1 \leq k \leq K$. Each X_{ijk} denotes the probability of gene i to fill in for enzyme family F_j in pathway P_k . Our only normalization requirement is that $\sum_i X_{ijk} = 1$ i.e. multiple proteins can fill in for an enzyme family in the same pathway, however the sum of their contributions should be one (the reaction must be catalyzed).

Given the set of probabilities, one can compute the score of any global assignment \mathbf{A} by summing the scores of individual pathways:

$$score(\mathbf{A}(P_k)) = \sum_{j_1, j_2 \in \mathbf{F}(P_k)} \sum_{i_1, i_2} X_{i_1 j_1 k} X_{i_2 j_2 k} \cdot sim(\mathbf{E}_{i_1}, \mathbf{E}_{i_2})$$

where $\mathbf{F}(P_k)$ is the set of pathway families associated with P_k and $sim(\mathbf{E}_{i_1}, \mathbf{E}_{i_2})$ is a similarity measure between the expression profiles of genes i_1 and i_2 . We use an EM algorithm to iteratively compute the assignment probabilities X_{ijk} .

Initialize: We initialize the X_{ijk} by assigning non-zero probability to the gene i that based on database annotations and sequence similarity data belongs to the enzyme family j . The initial probabilities are proportional to the significance of the match between gene i and the statistical model of family j , $value(i, j)$. For database annotations we set the $value(i, j)$ to the maximum observed value and the probabilities are normalized.

Expectation step: We compute a score for each assignment \mathbf{A}_{ijk} , of gene i to enzyme family j of pathway k , taking the expectation value over all other possible assignments of genes to all other pathway families $j' \neq j$:

$$score(\mathbf{A}_{ijk}(t)) = \sum_{j' \in \mathbf{F}(P_k), j' \neq j} \sum_{i'} X_{ijk}(t) X_{i' j' k}(t) sim(\mathbf{E}_i, \mathbf{E}_{i'})$$

¹Department of Computer Science, Cornell University E-mail: liviup, golan@cs.cornell.edu

All self similarities are set to 1, to encourage recruits of multi-domain proteins while avoiding strong bias that might cause a single gene to overtake all reactions within a pathway.

Maximization step: For every pathway P_k , for every $F_j \in \mathbf{F}(P_k)$ and for every $i \in G$: if $score(\mathbf{A}_{ijk}(t)) > 0$ then compute new probability

$$X_{ijk}(t+1) = \eta * \frac{score(\mathbf{A}_{ijk}(t))}{\sum_{i'} score(\mathbf{A}_{i'jk}(t))} + (1 - \eta) * X_{ijk}(t)$$

else

$$X_{ijk}(t+1) = (1 - \eta) * X_{ijk}(t)$$

We consider only elements with $score(\mathbf{A}_{ijk}(t)) > 0$ in the sum $\sum_{i'} score(\mathbf{A}_{i'jk}(t))$. Finally, we normalize such that $\sum_i X_{ijk}(t+1) = 1$

Termination: Exit if $|X_{ijk}(t+1) - X_{ijk}(t)| < \epsilon$ for every i, j, k or after running for a maximal number of iterations *Iter*.

Results

We ran the algorithm on the Yeast genome with a set of 63 pathways, results are shown in Figure 1.

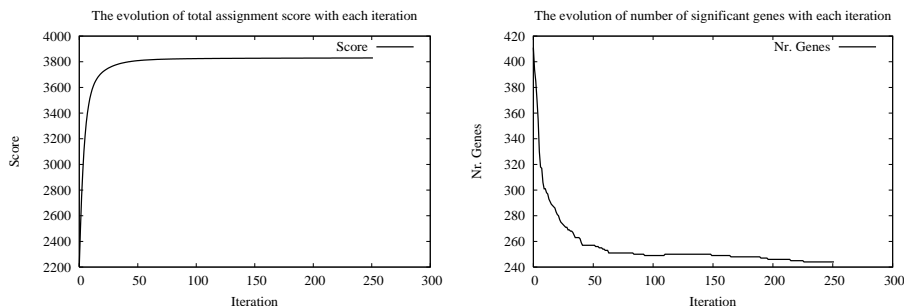


Figure 1: **EM algorithm results.** Left: The global assignment score. The graph indicates clear improvement with each iteration. Right: the number of X_{ijk} variables with values > 0.1 (i.e. genes that are instrumental in pathways). This number drops rapidly supporting our hypothesis that the many-to-many mapping is unlikely to exist in vivo.

References

- [1] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. 2004. The KEGG resource for deciphering the genome. *Nucl. Acids. Res.*, 32(90001):D277–280.
- [2] Karp, P. D., Paley, S., and Romero, P. 2002. The Pathway Tools software. *Bioinformatics*, 18(90001):225S–232.
- [3] Krieger, C. J., Zhang, P., Mueller, L. A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S. Y., and Karp, P. D. 2004. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucl. Acids. Res.*, 32(90001):D438–442.
- [4] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. 1998. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol. Biol. Cell*, 9(12):3273–3297.