EXPECTATION-MAXIMIZATION ALGORITHMS FOR FUZZY ASSIGNMENT OF GENES TO CELLULAR PATHWAYS

Liviu Popescu

Department of Computer Science, Cornell University, Ithaca, NY, 14853, US

Golan Yona *

Computer Science Department, Technion - Israel Institute of Technology
* *Email:golan@cs.technion.ac.il*

Cellular pathways are composed of multiple reactions and interactions mediated by genes. Many of these reactions are common to multiple pathways, and each reaction might be potentially mediated by multiple genes in the same genome. Existing pathway reconstruction procedures assign a gene to all pathways in which it might catalyze a reaction, leading to a *many-to-many* mapping of genes to pathways. However, it is unlikely that all genes that are capable of mediating a certain reaction are involved in all the pathways that contain it. Rather, it is more likely that each gene is optimized to function in specific pathway(s). Hence, existing procedures for pathway construction produce assignments that are ambiguous.Here we present a probabilistic algorithm for the assignment of genes to pathways that addresses this problem and reduces this ambiguity. Our algorithm uses expression data, database annotations and similarity data to infer the most likely assignments, and estimate the affinity of each gene with the known cellular pathways. We apply the algorithm to metabolic pathways in Yeast and compare the results to assignments that were experimentally verified.

1. INTRODUCTION

In the last decade an increasingly large number of genomes were sequenced and analyzed. The wealth of experimental data about genes initiated many studies in search for larger complexes, patterns and regularities. Of broad interest are studies that attempt to compile the network of cellular pathways in a given genome[1–4]. Due to the complexity of these studies, these pathways have been verified and studied extensively only in a few organisms, while in others the analysis is mostly computational. To propagate the experimental knowledge to other organisms several groups developed procedures that extrapolate pathways (and mostly metabolic pathways) based on the known association of genes with reactions in these pathways. However, many genes have unknown function and therefore the cellular processes in which they participate remain largely unknown. On the other hand, some reactions can be catalyzed by multiple genes and are associated with multiple pathways. Thus, assignments that rely just on the general functional characterization of genes are not refined enough and tend to introduce ambiguity by creating many-to-many mappings between genes and pathways. This ambiguity characterizes popular procedures for the assignment of genes to pathways[5, 6].

In a previous work[7] we presented a deterministic algorithm for pathway assignment that reduces the ambiguity by using expression data, in addition to functional characterization of genes, and selectively assigning genes to pathways such that co-expression within pathways is maximized and conflicts among pathways (due to shared assignments) are minimized. Furthermore, to complement the set of known enzymes (which is usually incomplete) our algorithm considers other genes in the subject genome that might possess catalytic capabilities based on similarity. As our tests showed, our algorithm works well on a set of test pathways, however, it assigns a single gene to each reaction. While our results generally support this assumption, it is not always the case in reality. Different genes might participate with different affinities in different pathways. Therefore, a more reasonable approach would be to assign a probabilistic measure to indicate the level of association of a given gene with a given pathway.

In this paper we present a variation over a known EM algorithm that addresses this problem, now assuming that the same gene can participate in multiple pathways, and estimates pathway assignment probabilities from expression data and sequence similarities. Our framework can be extended to include interaction data and other high-throughput data sets, each one providing information on

*Corresponding author.

different aspects of the same cellular process.

2. A PROBABILISTIC FRAMEWORK FOR ASSIGNING GENES TO PATHWAYS

In this paper we focus on metabolic pathways. In Ref. 8 a metabolic pathway is defined as "a sequence of consecutive enzymatic reactions that brings about the synthesis, breakdown, or transformation of a metabolite from a key intermediate to some terminal compound." This definition is used in many studies and by most biochemical textbooks and underlies literature curated databases such as MetaCyc[9]. We adopt this definition in our algorithm.

Our initial assumption is that the expression profiles of genes assigned to the same pathway tend to be similar which suggests that each pathway has a characteristic expression profile. Indeed, a similar assumption was employed in other studies on pathway reconstruction (see section 4). Therefore a pathway can be modeled as a probabilistic source for the expression profiles of the participating genes, having as centroid the pathway characteristic profile.

2.1. Preliminaries

In the next sections we use bold characters to represent sets or vectors and non-bolded characters to represent individual entities or measurements. The input to our algorithm is a genome $\mathbf{G}$ with N genes, enzyme families $F_1, F_2, ..F_M$ and pathways $P_1, P_2, ..., P_K$. We adhere to the set of known pathways as our algorithm is concerned with pathway assignments rather than pathway discovery (although this can be easily changed). Each pathway P contains a set enzymatic reactions. Each reaction is associated with an enzyme family F whose member genes can catalyze the reaction. We denote by $\mathbf{F}(P)$ the set of protein families that are associated with the reactions of pathway P. We use $\mathbf{G}(F_j)$ to represent the set of genes that can be assigned to enzyme family F_j based on their database records (or based on their similarity with known enzymes of family F_j, as described in section 1.2.1 of the online Supplementary Material). The set of enzymatic reactions (families) associated with gene i is denoted by $\mathbf{F}(i)$.

Our goal is to predict which genes take part in each pathway. In other words, our goal is to compute the probability $p(i|P_k)$ of gene i participating in pathway P_k as well as the posterior probability $p(P_k|i)$, which we refer to as the **affinity of gene i with pathway P_k**.

Computing the probabilities $p(i|P_k)$ and $p(P_k|i)$ is difficult since they refer to biological entities (genes) that are not observed directly but only through measurements (e.g. expression level). Therefore, we assume that each cellular process (in our case a metabolic pathway) can be modeled as a statistical source[a] generating measurable observations over genes. Each gene i is associated with a feature vector $\mathbf{x}_i$, and the conditional probability $p(\mathbf{x}_i|P_k)$ denotes the probability of the k-th source to emit $\mathbf{x}_i$. We initialize these probabilities based on prior knowledge of metabolic reactions. We then revisit these estimates and recompute these probabilities based on experimental observations until convergence to maximum likelihood solutions. However, this process is constrained so as to maintain the prior information.

The observations can be characterized in terms of different types of data (such as expression profiles, interactions, etc) that reflect different aspects of the pathway. E.g. $\mathbf{x}_i = \{\mathbf{e}_i, \mathbf{i}_i, ...\}$ where $\mathbf{e}_i$ is the expression profile of gene i, $\mathbf{i}_i$ is the interaction profile and so on. Assuming independence between these features we can decompose $p(\mathbf{x}_i|P_k) = p(\mathbf{e}_i|P_k)p(\mathbf{i}_i|P_k)\cdots$. In this work we use only expression profiles (that are generated from multiple experiments). I.e. we estimate $p(\mathbf{x}_i|P_k) \sim p(\mathbf{e}_i|P_k)$ where $p(\mathbf{e}|P_k)$ is the probability to observe expression vector $\mathbf{e}$ in pathway P_k. This approximation is based on the assumption that genes participating in the same biological process are similarly expressed. Indeed, it produces good results as we demonstrate later on. However, the algorithm can be easily generalized to include other types of data.

2.2. The EM algorithm

Our algorithm is based on the fuzzy EM clustering algorithm that assumes a mixture of Gaussian sources[10], with several modifications that are discussed in the next section. We model each pathway as a source that generates expression profiles for the pathway genes such that $p(\mathbf{e}|P_k)$ follows a Gaussian distribution $N(\mu_k, \mathbf{\Sigma}_k)$ or a mixture of Gaussian sources (assuming there are several underlying processes, intermingled together). Each path-

[a] Each pathway can also be modeled as a mixture of sources, for example, when there are multiple branches.

way has a prior $p(P_k)$.

We assume that the microarray experiments are independent of each other such that the expression vector $\mathbf{e}$ is composed of d *independent*[b] measurements $\{e_1, e_2, ..., e_d\}$. I.e. $p(\mathbf{e}|P_k) = \prod_{l=1}^{d} p(e_l|P_k)$ where each component is distributed as a one dimensional normal distribution $N(\mu_{kl}, \sigma_{kl})$. Hence the covariance matrix $\mathbf{\Sigma}_k$ is actually a diagonal matrix, whose non-zero elements are denoted by σ_k.

We seek the parameters that will maximize the likelihood of the data. We initialize the parameters of the pathway models μ_k, σ_k and $p(P_k)$ based on database annotations and similarity data, as described in section 1.2.1 in the online Supplementary Material. These parameters as well as the probabilities $p(P_k|\mathbf{e}_i)$ are modified iteratively, using an EM algorithm similar to the one described in Ref. 10, until convergence. This algorithm converges to parameters that maximize (locally) the likelihood of the data $p(\mathcal{D}/\Theta) = \prod_{i=1}^{N} p(\mathbf{e}_i|\Theta) = \prod_i \sum_k p(\mathbf{e}_i|P_k)p(P_k)$ where $\Theta = \{(\mu_1, \sigma_1), (\mu_2, \sigma_2), ..., (\mu_K, \sigma_K)\}$. For more details, see section 1 in the online Supplementary Material.

2.3. Knowledge-based clustering

Our algorithm is a variation of the EM algorithm described above, in several ways. First, our algorithm utilizes any prior information that might help to obtain more accurate assignments. Instead of random initialization of μ_k and σ_k, we use the prior information that is available from database annotations and similarity searches, to initialize the parameters. Second, we employ constrained clustering so as to minimize the number of pathways that end up with an incomplete assignment. Third, we replace the Euclidean metric with a new metric, the *mass-distance* measure, that is more effective for detecting similarity between expression profiles. Due to space limitation the details of these three elements are described in the Supplementary Material, section 1.2.

3. RESULTS

To evaluate the performance of our method we test the influence of different settings and parameters on pathway assignments and show that the algorithm produces results of biological significance. We first provide quantitative measures of performance by comparing the results we get to experimentally validated assignments. We then take a look at particular examples to illustrate the strengths of our algorithm.

Our model organism is the Yeast genome. Pathways blueprints are obtained from the MetaCyc database[9]. We used a subset of 52 metabolic pathways for which we could assign Yeast genes to all the reactions in the pathway. 23 of these were experimentally verified to exist in Yeast in SGD[11]. This set of 23 pathways serves as our test set. To assign genes to pathways we test two different expression datasets: the Cell Cycle data set of Ref. 12 and the Rosetta Inpharmatics Yeast compendium data set[13]. Genes are mapped to enzymatic reactions using Biozon[14] at `biozon.org`. Proteins that are linked with enzyme families based on their annotation are referred to as **annotated enzymes**. Proteins assigned to reactions based on similarity with known enzymes are referred to as **predicted enzymes**. For more information on the datasets used in this study see section 2 of the Supplementary Material.

To explore the influence of different options on our algorithm we ran a total of 12 experiments. We compare performance across different models (the Gaussian model vs the mass-distance model), different data sets (Cell-cycle vs. Rosetta) and different subsets of genes from the Yeast genome as outlined below. We use several performance measures as discussed in section 3 of the Supplementary Material.

Gene sets Using the prior knowledge we can restrict the set of genes we consider in our algorithm. The most constrained set of genes is **pathway genes (PG)** , i.e. the genes that can be assigned to at least one of the pathways based on database annotations and by prediction $\mathbf{PG} = \cup_k \cup_{F_j \in \mathbf{F}(P_k)} \mathbf{G}(F_j)$. The intermediate set of genes consists of **all enzymes (AE)** in the genome, annotated or predicted (including enzymes that are not associated with any of the reactions in the pathways we considered), $\mathbf{PG} = \cup_{j=1}^{M} \mathbf{G}(F_j)$. The third set of genes we consider is the entire genome or **all genes** (**AG**), $\mathbf{AG} = \mathbf{G}$.

[b]Expression profiles are typically composed of measurements taken from a set of independent experiments. For example, in time-series datasets the measurements are collected at different time-points usually spaced at relatively large time intervals during which the cell has undergone significant changes and the correlation between consecutive time points is relatively weak. Other datasets (e.g. Rosetta) are generated from experiments that are conducted practically independently under different conditions.

Table 1. Comparative results for different experimental settings. Results are reported over the test set of 23 pathways. First column lists the experimental setup. Codes used: Cell Cycle data (TS), Rosetta (ROS), pathway genes (PG), all enzymes (AE), entire genome (AG), MASS Distance (MD), Gaussian model (GM), deterministic algorithm (DET). (e.g., "ROS:AE:MD" is the experiment using Rosetta, clustering only enzymes and using the MASS distance model). In the second column we show the number of pathways with a verified (EV) assignment in the top position. The third column shows the number of pathways with violated constraints (number in parenthesis is over the entire set of 52 pathways used in clustering). The fourth and the fifth columns show the precision and recall with respect to the verified genes (where genes are assigned to a pathway if the posterior probability is greater than a threshold $\theta = 0.1$). In the last 2 columns we show the MAP with respect to the ranking of genes based on their affinity, and with respect to the ranking of all possible deterministic assignments based on their score (see section 3 of the Supplementary Material for details). The last line in the table represents the results of a model that gives a random ordering of the genes and assignments; this is equivalent to a regular pathway reconstruction algorithm.

Experiment	# of pathways with verified top assignment	# of pathways with violated constraints	precision (genes)	recall (genes)	MAP genes	MAP assignments
TS:AG:MD	**12**	12(29)	0.72	0.60	0.86	**0.7**
TS:AE:MD	10	12(28)	0.80	0.65	0.83	0.62
TS:PG:MD	8	11(24)	0.79	0.69	0.81	0.62
TS:PG:GM	9	**9(22)**	0.80	0.69	0.85	0.61
TS:AG:GM	9	12(27)	**0.84**	**0.71**	**0.87**	0.6
TS:AE:GM	6	11(27)	0.79	0.63	0.84	0.52
TS:PG:DET	10	N/A	N/A	N/A	N/A	0.68
ROS:PG:MD	**14**	10(24)	**0.85**	**0.72**	**0.94**	**0.84**
ROS:PG:GM	**14**	**11(22)**	0.83	0.71	0.94	0.82
ROS:AE:MD	13	10(24)	**0.85**	0.69	0.94	0.78
ROS:AG:MD	13	10(26)	0.84	0.69	0.93	0.77
ROS:AE:GM	12	12(25)	0.74	0.63	0.91	0.77
ROS:AG:GM	10	11(25)	0.80	0.56	0.9	0.66
ROS:PG:DET	11	N/A	N/A	N/A	N/A	0.75
random model	5.4	N/A	N/A	N/A	0.74	0.45

3.1. Summary of results

Our method is conceived as an extension of the current pathway reconstruction methods like Pathway Tools[15]. These methods do not attempt to assign genes to pathways selectively and hence cannot be compared to ours. Therefore we need some other baseline to compare our results to. We consider the random model that generates random permutations over the set of all possible assignments (this setting is similar to that of KEGG or Pathway Tools, where there is no ranking over assignments and all assignments are equally probable). For each pathway we generate 100k random permutations and compute the average MAP over the results (see section 3 in the Supplementary Material for details). We also compare the results with those of the deterministic algorithm [c] of our previous work[7].

As Table 1 shows, our algorithm is able to improve significantly over the random model under all settings, and it also improves over the deterministic algorithm. Clearly, our model exploits the information in the expression data to rank the genes effectively. When comparing the different settings our general conclusions are:

- Clearly the choice of the expression data set is important. The performance of our algorithm on the Rosetta set is significantly and consistently better than on the Cell Cycle data.
- There are no significant differences between the different model variations within each expression dataset, but there are some noticeable trends. The mass distance model has a slight advantage compared to the Gaussian model, all other being equal. The good performance even with the Euclidean metric reflects the strong correlation between pathways and expression pat-

[c] To compare the results, we ran the deterministic algorithm of Ref. 7 but skipped the last step that attempts to minimize shared assignments by looking at near-optimal assignments, since it explicitly drives the assignments towards solutions that assign a single gene to each reaction.

terns.

- Interestingly, the performance does not decline signifi cantly when we use a larger set of genes. This confi rms that pathways tend to have unique characteristic expression profi les
- Most of the pathways have zero or one violated constraints (see section 1.2.2 of the Supplementary Material) in all settings. However, there are a few pathways in which consistently most of the reactions are not satisfi ed (such as arginine biosynthesis, cysteine biosynthesis II, arginine degradation I and trehalose biosynthesis III). To some extent, this reflects the capacity (or lack thereof) of our model to cover certain pathways. However, this can also suggest that these pathways might not exist in Yeast or they might exist but in a different confi guration than the pathway blueprint. This conclusion is also reinforced by the fact that the average number of violated constraints does not seem to depend on the expression dataset.

The number of related genes (genes that have signifi cant affi nity with a pathway although they were not initially assigned to it) loosely correlates with the number of reactions in the pathway as well as with the number of genes initially assigned to the pathway (data not shown). Finally, most pathways tend to have similar performance across all experiments within a data set. This explains the small difference in performance between experiments within the datasets. The difference between the two datasets is caused by a few pathways for which the Rosetta data set is more informative.

3.2. Example - Homoserine methionine biosynthesis

In this section we present an individual case which illustrates our method. Due to the page limit we focus only on one example. Two additional examples are given in the Supplementary Material (section 4). The examples are discussed in the context of the ROS:AE:MD experiment. This setting is one of the best ones according to our performance measures and is chosen because the clustering is done with all enzymes, revealing interesting dependencies between pathways.

Methionine is bio-synthesized in this pathway from homoserine. It is part of the superpathway *Threonine and methionine biosynthesis* that consists of three pathways: *homoserine biosynthesis*, *homoserine methionine biosynthesis* and *threonine biosynthesis from homoserine* as is depicted in Figure 1a. Though related, these pathways

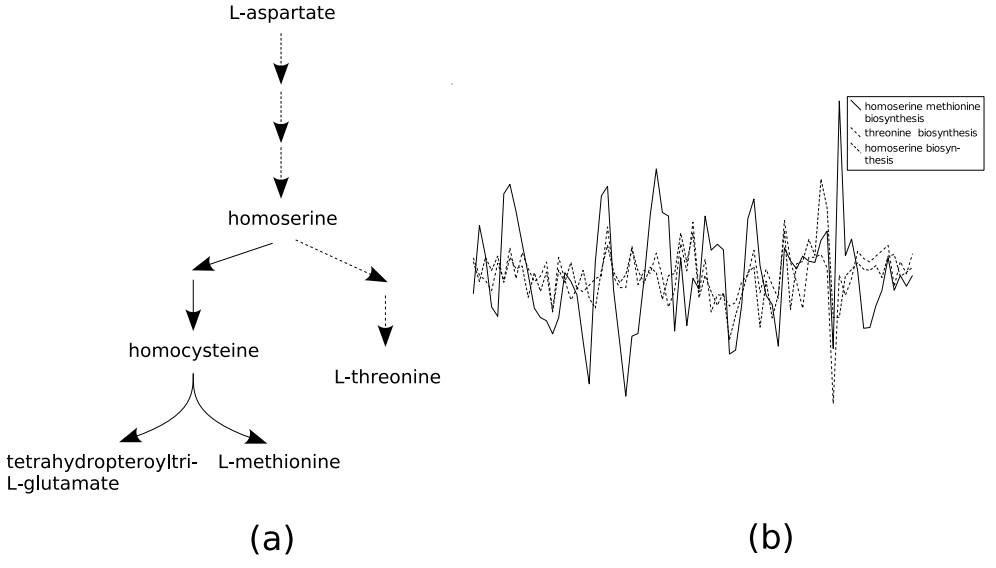

Fig. 1. The relation between *homoserine biosynthesis*, *homoserine methionine biosynthesis* and *threonine biosynthesis from homoserine*. (a). *Homoserine* is synthesized from *aspartate* in the first pathway and *methionine* and *threonine* are synthesized in the second and third pathway respectively, starting both from *homoserine*, therefore the super-pathway forks at *homoserine*. **(b)**. The characteristic profiles of the three pathways. *Homoserine biosynthesis* and *threonine biosynthesis from homoserine* are correlated while *homoserine methionine biosynthesis* is just loosely correlated.

have different characteristic expression profiles for most of the experiments. However they seem to be similar in certain experiments (Figure 1b), which suggests that they share regulation mechanisms.

This pathway has three reactions: 2.3.1.31, 4.2.99.- and 2.1.1.14. We have excluded reaction 4.2.99.- from the pathway model because it has an incomplete EC number[d]. There are seven genes that can be initially assigned to this pathway based on database annotations and function prediction (by similarity). However, only three are experimentally verified assignments: MET2, MET6 and MET17. These are also the only genes whose affinity with the pathway (posterior probability) at the end of the run is significant (1 in this case) as is shown in Table 2. Our algorithm assigns these genes such that each reaction is associated with exactly one gene. The other four unverified genes have insignificant affinity to the pathway, and no other genes are associated with the pathway. Note that all the unverified genes as well as MET17 have similar functional assignments initially (to 6 different reactions), only with different evalues. However, only MET17 makes it to final round and is assigned to the pathway by our algorithm. Furthermore, our algorithm consistently recovers the experimentally verified genes

Table 2. The probabilistic assignment of the *Homoserine methionine biosynthesis* pathway. The table lists all the genes that are potentially assigned to this pathway. The double line separates the genes that were assigned to this pathway from the ones that were rejected. In the first column we show the name of the gene or the systematic name. Second column shows the Biozon ID[14]. Third column shows the affinity to the pathway ($p(P_k|x_i)$). The forth column shows the EC family membership (number in parenthesis is the weight which reflects the confidence in this assignment). If the EC number is in bold the gene was annotated in the database as capable of catalyzing this reaction otherwise it was predicted. In the last column we list the MetaCyc ID of alternate pathways to which this gene was also assigned (number in parenthesis is the affinity to that pathway).

Gene Name	Biozon ID	Verification	Pathway affinity	EC Numbers	Alternative pathway affinity
MET2	004860000048	verified	1.00	**2.3.1.31** (1.00)	
MET6	007670000142	verified	1.00	**2.1.1.14** (1.00)	
MET17	004440000819	verified	1.00	**4.2.99.-** (1.00) **4.2.99.10** (1.00) **4.2.99.8** (0.99) **4.2.99.9** (0.45) 2.3.1.31 (0.59) 4.4.1.1 (0.37) 4.4.1.8 (0.38)	
YHR112C	003780000158	unverified	0.00	2.3.1.31 (0.09) 4.2.99.10 (0.15) 4.2.99.8 (0.09) 4.2.99.9 (0.31) 4.4.1.1 (0.24) 4.4.1.8 (0.26)	GLYOXYLATE-BYPASS (0.99)
Cys3	003940001012	unverified	0.00	**4.4.1.1** (1.00) 4.2.99.10 (0.37) 4.2.99.8 (0.24) 4.2.99.9 (0.68) 2.3.1.31 (0.26) 4.4.1.8 (0.57)	HOMOCYSDEGR-PWY (0.50) PWY-801 (0.50)
Str3	004650000171	unverified	0.00	**4.4.1.8** (1.00) 4.2.99.10 (0.22) 4.2.99.8 (0.15) 4.2.99.9 (0.45) 4.4.1.1 (0.39) 2.3.1.31 (0.15)	HOMOCYSDEGR-PWY (0.50) PWY-801 (0.50)
YFR055W	003400000153	unverified	N/A (no profile)	**4.4.1.8** (1.00) 4.2.99.10 (0.18) 4.2.99.8 (0.11) 4.2.99.9 (0.34) 4.4.1.1 (0.28)	N/A (no profile)

[d]Recently, this reaction was revised and assigned a new number 2.5.1.49.

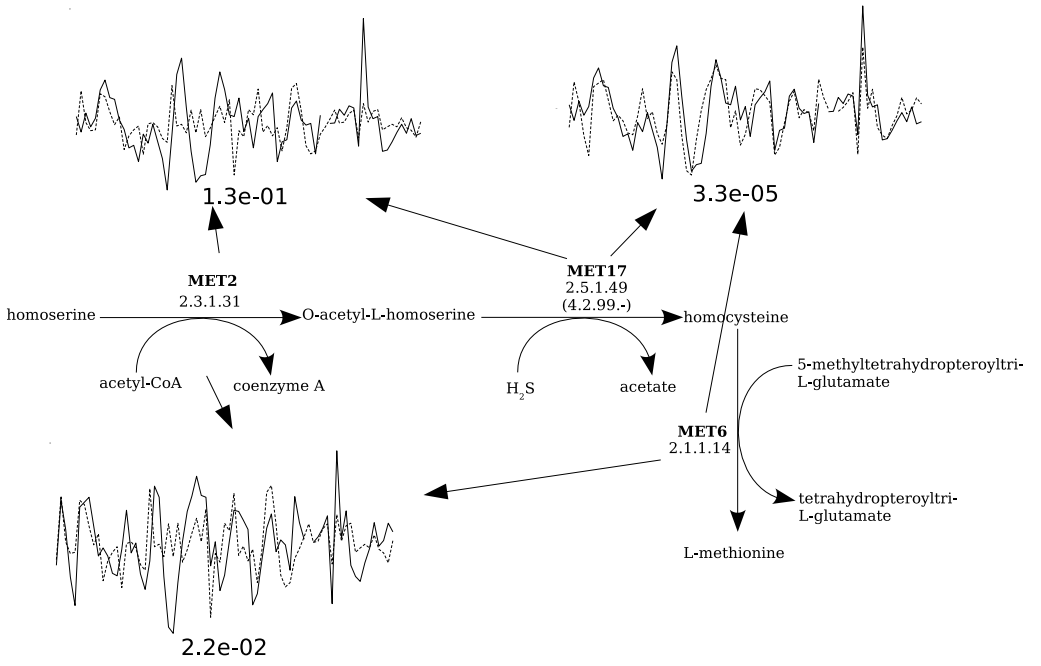

Fig. 2. ***Homoserine methionine biosynthesis*, the pathway diagram**. The diagram was obtained from the MetaCyc database[17] and augmented with genes that are experimentally verified to participate in the pathway and their expression profiles. We notice a strong correlation between the genes catalyzing last two reactions while the first gene is less correlated. (Profiles shown are for the Cell Cycle dataset).

over all experiments involving Rosetta, and most of the time-series experiments.

4. RELATED STUDIES

Metabolic pathway reconstruction has been an important direction in experimental research for many decades. This research focused on some well studied organisms like E. coli and S. cerevisiae. The knowledge thus obtained was collected in databases like EMP/MPW[18], MetaCyc[17] and KEGG[19].

Unfortunately, the experimental reconstruction of metabolic pathways is a long and costly process and the obtained information is restricted to the studied organism. The breakthroughs in DNA sequencing and thus the large number of sequenced and annotated organisms led to the development of procedures for extending metabolic knowledge from the organisms in which it was experimentally studied to newly sequenced organisms. Methods like Pathologic[5], PUMA2[20], SEED[21] and KEGG[6] use sets of blueprints of experimentally elucidated metabolic pathways, and match the reactions in these blueprints with genes in the target organism based on their functional annotations. Sometimes not all enzyme functions needed to complete the pathway can be found in the original annotation. To cope with this situation, tools for predicting the missing enzymatic activity[22–26] were added to complement the original annotation.

The analysis of the dynamic aspects of cellular processes, including metabolic pathways, was made possible by the increasing availability of high-throughput data, like expression data, interaction data and subcellular location of proteins. Clustering is one of the favorite methods for the analysis of expression data, because genes that are similarly expressed might participate in the same cellular process. Consequently a number of clustering methods were applied to expression data starting with the seminal work in Ref. 27 (see Ref. 28 and Ref. 29 for discussion of these methods).

Expression data is used in metabolic pathway analysis by a large number of studies. A large class of these studies extract active pathways by scoring them based on the expression of the assigned genes[30–34]. Clustering of expression data is also used in metabolic pathway analysis. These methods try to elucidate the function of uncharacterized genes by mapping pathways to the clusters

to which these genes belong[35, 36].

The integration of metabolic information and expression data is further used to extract active pathways and processes in several ways. Expression data and metabolic network topology are combined in Ref. 37 to define a metric that is used for extracting clusters of genes corresponding to active pathways. Similarly, active pathways and their pattern of activity are extracted[38] using a generalized form of canonical correlation analysis between kernels defined based on expression data and on the pathway graph. To predict operons this approach is extended in[39] by integrating also the position of the genes on the DNA.

A first step in the metabolic network reconstruction is the inference of the more general cellular network. Several unsupervised prediction methods used models like Bayesian networks[40] and boolean networks[41] for cellular network inference from expression data. A supervised method for cellular network inference is described in Ref. 42. The method is based on canonical correlation between a kernel function integrating expression data, interaction data, phylogenetic profiles and subcellular location and a second kernel function defined based on the experimentally validated cellular network of yeast. This work is extended in Ref. 43 by forcing chemical compatibility constraints for edges in the predicted cellular network.

Related to our study are also the the studies on *regulatory modules*. Regulatory modules[44] are sets of genes whose expression is controlled by the same group of control genes (*regulation program*). Genes in a module are assumed to have a common function. It is also commonly assumed that enzymes in the same pathway are co-regulated, thus there is an overlap between a pathway and a regulatory module. This holds for some of the known metabolic pathways, however it is not always the case and relationships between pathways and modules can be one of several types as presented in Figure 3:

(1) *one to one* – a pathway overlaps with a regulatory module, i.e. the genes participating in the pathway are co-regulated (see Figure 3a). (e.g. *Homoserine methionine biosynthesis*).
(2) *many to one (module sharing)* – a module is shared by several pathways, i.e. the genes participating in several pathways are co-regulated (see Figure 3b). (e.g. *valine biosynthesis* and *isoleucine biosynthesis*).
(3) *one to many* – a pathway overlaps several modules, i.e. not all the genes participating in a pathway are co-regulated but they can be grouped in few co-regulated groups (see Figure 3c).
(4) *mixed* – a pathway overlaps several modules and share some of them with other pathways (see Figure 3d). (e.g. *folic acid biosynthesis*).

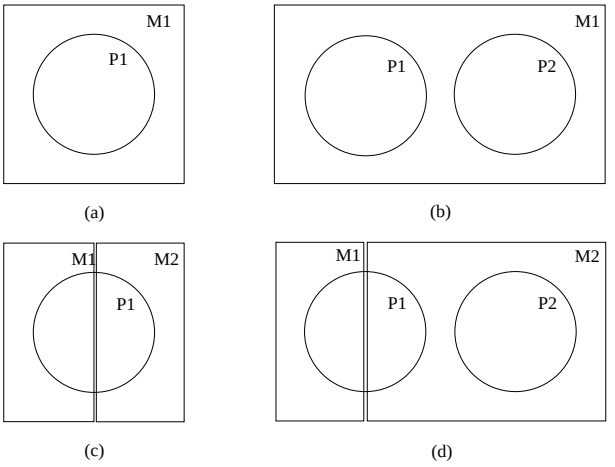

Fig. 3. Relationships between pathways and modules. a. *One to one*: the pathway P_1 overlaps module M_1. b. *Many to one (module sharing)*: The pathways P_1 and P_2 share the module M_1. c. *One to many*: The pathway P_1 overlaps both modules M_1 and M_2. d. *Mixed:* The pathways P_1 and P_2 share the module M_2 while P_1 overlaps with module M_2 as well.

Regulatory modules are explored in depth in Ref. 45, where several probabilistic models and inference algorithms are presented. For information on other related studies and extended discussion see the appendix in Ref. 7.

It is important to emphasize that these methods and others described in this section are technically different from our approach and most of them are targeted towards pathway analysis. Our current work focuses on a probabilistic framework for metabolic pathway assignment which enables us to address problems like ambiguous assignments, protein complexes and missing enzymes in the same context. In addition to expression data and metabolic knowledge, our framework can be extended to use other types of high throughput data.

5. DISCUSSION

In this paper we present an algorithm for probabilistic assignment of genes to pathways. Given a genome, our algorithm uses pathway blueprints (from MetaCyc or other sources), database annotations and similarity data (from Biozon) and genome-wide mRNA expression data to determine the characteristic expression profiles of pathways and assess the affinity of each gene with each pathway.

We test and demonstrate the power of our method on the Yeast genome. Although it is difficult to evaluate our algorithm, since the amount of experimentally validated data is limited, our results so far are significant and very encouraging and for most pathways the top assignment is also an experimentally verified one. The algorithm can also predict complexes and accommodates multifunctional enzymes.

While in this work we refer to a pathway as a well defined entity, in reality this is not the case and cellular processes are tightly related. Moreover, since cellular processes form a complex and highly connected network it is difficult to delineate the boundaries of individual pathways and the same process might be defined differently by different groups (see Ref. 7). This further motivates a probabilistic approach that assigns each gene with a certain probability to each pathway. And although we adhere to pathway diagrams that were determined in the literature, our procedures can be modified to redefine pathway boundaries so as to correlate better with regulatory modules (see section 4). Furthermore, our method can be easily applied to fill pathway holes in pathways with uncharacterized or unassigned reactions (as discussed in section 3 of the Supplementary Material).

It should be noted that while all pathways included in our analysis can be associated with Yeast genes based on annotation and functional prediction, some of the pathways might not exist in Yeast after all. An example of such a pathway is *cysteine biosynthesis II* which exists in mammals but not in Yeast. In Yeast, cysteine is obtained from homoserine and reactions making up this pathway overlap reactions in two other pathways (therefore in our analysis, all the reactions in this pathway have violated constraints). In this view, our method can also help to validate whether certain pathways exist in a given genome.

As our examples demonstrate, clustering alone cannot solve the pathway reconstruction problem and it is necessary to add additional constraints and prior knowledge to generate effective pathway models. This emphasizes the fundamental difference between our work and studies that are based on clustering of expression profiles. Furthermore, our results also indicate that even with these constraints and prior knowledge, expression data alone cannot discover all pathways and therefore additional datasets such as interaction data and subcellular location data are necessary to improve the models and we intend to integrate such datasets in future versions of our algorithm.

Besides the aforementioned extensions, there are other improvements and future directions that we would like to pursue. For example, we would like to improve function prediction. Currently, this is done based on database annotations or based on sequence similarity. However, the later is problematic and often genes are assigned based on similarity to multiple enzymatic reactions. To address this problem we intend to develop better methods to characterize enzymatic domains, using a methodology similar to the one we introduced in Ref. 46 and Ref. 24.

Finally, our algorithm can be applied to other genomes given a compatible expression dataset, and using a similar analysis to the one reported here we have started mapping pathways in the human genome.

6. SUPPLEMENTARY MATERIAL

A detailed description of our algorithm, the evaluation methodology and additional examples are available in the online supplementary material at `biozon.org/ftp/data/papers/pathway-assignment-em/`.

References

1. B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18(12):1257 – 1261, 2000.
2. Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180 – 183, January 2002.
3. N. C. Duarte, M. J. Herrgard, and B. O. Palsson.

Reconstruction and Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale Metabolic Model. *Genome Res.*, 14(7):1298–1309, 2004.

4. J. Forster, I. Famili, P. Fu, B. O. Palsson, and J. Nielsen. Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. *Genome Res*, 13(2):244–253, 2003.
5. S. M. Paley and P. D. Karp. Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. *Bioinformatics*, 18(5):715–724, 2002.
6. H. Bono, H. Ogata, S. Goto, and M. Kanehisa. Reconstruction of Amino Acid Biosynthesis Pathways from the Complete GenomeSequence. *Genome Res.*, 8(3):203–210, 1998.
7. L. Popescu and G. Yona. Automation of gene assignments to metabolic pathways using high-throughput expression data. *BMC Bioinformatics*, 6(1):217, 2005.
8. J. Stenesh. *Dictionary of Biochemistry and Molecular Biology (2nd Edition)*. John Wiley & Sons, 1989.
9. C. J. Krieger, P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee, and P. D. Karp. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, 32(Database issue):D438–442, 2004.
10. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2nd ed. edition, October 2000.
11. K. R. Christie, S. Weng, R. Balakrishnan, M. C. Costanzo, K. Dolinski, S. S. Dwight, S. R. Engel, B. Feierbach, D. G. Fisk, J. E. Hirschman, E. L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C. L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein, and J. M. Cherry. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res.*, 32 (Database issue):D311–314, 2004.
12. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Biol. Cell*, 9(12):3273–3297, 1998.
13. T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional Discovery via a Compendium of Expression Profiles. *Cell*, 102(1):109–126, July 2000.
14. A. Birkland and G. Yona. BIOZON: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics*, 7:70, 2006. URL `biozon.org`.
15. P. D. Karp, S. Paley, and P. Romero. The Pathway Tools software. *Bioinformatics*, 18(Suppl 1):S225–S232, 2002.
16. G. Yona, W. Dirks, S. Rahman, and D. Lin. Effective similarity measures for expression profiles. *Bioinformatics*, 2006. in press.
17. BioCyc. Biocyc database, 2005. URL `http://biocyc.org/`.
18. J. Selkov, E, Y. Grechkin, N. Mikhailova, and E. Selkov. MPW: the Metabolic Pathways Database. *Nucleic Acids Res.*, 26(1):43–45, 1998.
19. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, 32(Database issue):D277–280, 2004.
20. N. Maltsev, E. Glass, D. Sulakhe, A. Rodriguez, M. H. Syed, T. Bompada, Y. Zhang, and M. D'Souza. PUMA2–grid-based high-throughput analysis of genomes and metabolic pathways. *Nucl. Acids Res.*, 34(suppl 1):D369–372, 2006.
21. SEED. The seed: an annotation/analysis tool provided by fig, 2005. URL `http://theseed.uchicago.edu/FIG/index.cgi`.
22. H. Bono, S. Goto, W. Fujibuchi, H. Ogata, and M. Kanehisa. Systematic Prediction of Orthologous Units of Genes in the Complete Genomes. *Genome Inform Ser Workshop Genome Inform*, 9:32–40, 1998.
23. M. Green and P. Karp. A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5(1):76, 2004.
24. U. Syed and G. Yona. Using a mixture of probabilistic decision trees for direct prediction of protein function. In *Proceedings of the seventh annual international conference on Computational molecular biology*, pages 289–300. ACM Press, 2003.
25. I. Shah. *Predicting enzyme function from sequence*. PhD thesis, George Mason University, 1999.
26. P. Kharchenko, D. Vitkup, and G. M. Church. Filling gaps in a metabolic network using expression information. *Bioinformatics*, 20(suppl 1):i178–185, 2004.
27. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95(25):14863–14868, December 8 1998.
28. F. Valafar. Pattern Recognition Techniques in Microarray Data Analysis: A Survey. *Ann NY Acad Sci*, 980(1):41–64, 2002.
29. D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386, 2004.
30. P. Grosu, J. P. Townsend, D. L. Hartl, and D. Cavalieri. Pathway Processor: A Tool for Integrating Whole-Genome Expression Results into Metabolic Networks. *Genome Res.*, 12(7):1121–1126, 2002.
31. R. Kuffner, R. Zimmer, and T. Lengauer. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, 16(9):825–836, 2000.
32. P. Pavlidis, D. Lewis, and W. Noble. Exploring gene expression data with class scores. In *Pac Symp Biocomput*, pages 474–485, 2002.
33. S. Doniger, N. Salomonis, K. Dahlquist, K. Vranizan, S. Lawlor, and B. Conklin. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, 4(1):R7, 2003.

34. J. Rahnenfuhrer, F. S. Domingues, J. Maydt, and T. Lengauer. Calculating the Statistical Signifi cance of Changes in Pathway Activity From Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.
35. M. Nakao, H. Bono, S. Kawashima, T. Kamiya, K. Sato, S. Goto, and M. Kanehisa. Genome-scale Gene Expression Analysis and Pathway Reconstruction in KEGG. *Genome Inform Ser Workshop Genome Inform.*, 10:94–103, 1999.
36. J. van Helden, D. Gilbert, L. Wernisch, and S. Schroeder, M.and Wodak. Applications of regulatory sequence analysis and metabolic network analysis to the interpretation of gene expression data. *Lecture Notes in Computer Sciences*, 2066:155–172, 2001.
37. D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(Suppl 1):S145–154, 2002.
38. J. P. Vert and M. Kanehisa. Extracting active pathways from gene expression data. *Bioinformatics*, 19(Suppl 2): II238–II244, 2003.
39. Y. Yamanishi, J.-P. Vert, A. Nakaya, and M. Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19(Suppl):323i–330, 2003.
40. N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.
41. T. Akutsu, S. Miyano, and S. Kuhara. Algorithms for identifying boolean networks and related biological networks based on matrix multiplication and fi ngerprint function. *Journal of Computational Biology*, 7(3-4):331–343, 2000.
42. Y. Yamanishi, J.-P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20(suppl 1):i363–370, 2004.
43. Y. Yamanishi, J.-P. Vert, and M. Kanehisa. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21(suppl 1):i468–477, 2005.
44. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller1, and N. Friedman. Module networks: identifying regulatory modules and their condition-specifi c regulators from gene expression data. *Nat Genet*, 34(2):166–176, 2003.
45. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller1, and N. Friedman. Module networks: identifying regulatory modules and their condition-specifi c regulators from gene expression data. *Nat Genet*, 34(2):166–176, 2003.
46. N. Nagarajan and G. Yona. Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics*, 20: 1335-1360, 2004.
47. R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, A. E. G. Tyra G. Wolfsberg and, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2:65–73, July 1998.
48. S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, 1997.
49. E. M. Voorhees and L. P. Buckland, editors. *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, 2005.
50. J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.