Databases and ontologies

In silico gene function prediction using ontology-based pattern identification

Yingyao Zhou1,*, Jason A. Young2, Andrey Santrosyan1, Kaisheng Chen1, S. Frank Yan1 and Elizabeth A. Winzeler1,2

1Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA and
2Department of Cell Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

Received on July 9, 2004; revised on October 25, 2004; accepted on October 29, 2004
Advance Access publication November 5, 2004

ABSTRACT

Motivation: With the emergence of genome-wide expression profiling data sets, the guilt by association (GBA) principle has been a cornerstone for deriving gene functional interpretations in silico. Given the limited success of traditional methods for producing clusters of genes with great amounts of functional similarity, new data-mining algorithms are required to fully exploit the potential of high-throughput genomic approaches.

Results: Ontology-based pattern identification (OPI) is a novel data-mining algorithm that systematically identifies expression patterns that best represent existing knowledge of gene function. Instead of relying on a universal threshold of expression similarity to define functionally related groups of genes, OPI finds the optimal analysis settings that yield gene expression patterns and gene lists that best predict gene function using the principle of GBA. We applied OPI to a publicly available gene expression data set on the life cycle of the malarial parasite Plasmodium falciparum and systematically annotated genes for 320 functional categories based on current Gene Ontology annotations. An ontology-based hierarchical tree of the 320 categories provided a systems-wide biological view of this important malarial parasite.

Availability: A web accessible P. falciparum e-annotation database containing the results of this study can be accessed online at http://carrier.gnf.org/publications/OPI

Contact: zhou@gnf.org

INTRODUCTION

The observation that functionally related genes tend to share similar mRNA expression profiles, i.e., guilt by association (GBA), is one of the cornerstones for functional genomic research based on microarray technology (Walker et al., 1999; Quackenbush, 2003). The typical GBA data analysis procedure clusters genes based on the similarities among their expression profiles. Using these clusters, it is then hypothesized that uncharacterized genes may potentially share functional roles with annotated genes in the same cluster. The principle of GBA has been widely applied and has yielded interesting biological discoveries (Le Roch et al., 2003; Wu et al., 2002; Brown et al., 2000; Eisen et al., 1998). However, its success has also been greatly limited due to a series of open questions.

In the initial data preprocessing stages alone, several issues directly affect the results of GBA analyses. Is it beneficial to perform a data transformation such as the logarithmic transformation (Geller et al., 2003; Pan et al., 2002)? How does one filter out genes with trivial expression patterns in order to boost the signal-to-noise ratio (Herrero et al., 2003; DChip manual, Filter Genes, http://www.biostat.harvard.edu/complab/dchip/)? An over-stringent filtering process excludes valuable biological information, while on the other hand an over-relaxed approach retains noisy data and degrades the data quality. Unsupervised clustering analyses are controversial as well. What is the best similarity metrics? Should it be Pearson correlation coefficient-based or Euclidean distance-based (DChip manual)? Is a partitioning-based clustering method such as k-means better off and what is the right k (Yeung et al., 2001)? Or does a hierarchical clustering method provide more flexibility, and what is the similarity threshold to use in order to identify the right sub-trees for functional analysis (Allocco et al., 2004)?

Although significant progress has been made by computational biologists to address these issues, universal answers to the above questions are not warranted. In practice, researchers may favor one method over another based on a trial-and-error approach with their particular data sets. One cannot be sure which gene lists generated using various analyses on a given data set will provide the optimal insight into the underlying biological processes, because of the inability to quantitatively study the impact of various analysis methods and related method parameters chosen during a GBA study.

The key contribution of the Ontology-based pattern identification (OPI) algorithm described in this study is to provide a tool for computationally identifying the optimal analysis pipeline and corresponding parameters given existing pieces of related biological knowledge. The optimal analysis pipeline is defined as the one that finds the best association between patterns in gene expression data and the known functional classifications for the genes. With the above questions in mind, we applied OPI to a published life cycle expression microarray data set of the malaria parasite Plasmodium falciparum. This resulted in the construction of a web-accessible public database (http://carrier.gnf.org/publications/OPI) that contains 320 statistically significant gene expression patterns for various gene ontology categories. These gene ontology categories encompassed 167 biological processes, 84 molecular functions and 69 cellular components. We present here the details of the OPI algorithm and address some of the open questions discussed above. Compared to our previously published k-means results, OPI gives richer and more precise biological interpretations to the same underlying data set (Le Roch, 2003).

*To whom correspondence should be addressed.
et al., 2003). Both the OPI and our database should contribute to advancing understanding of the biology of this important disease organism.

METHODS
Mathematical framework and implementation

Hereafter, a method $M$ refers to the combination of a data analysis pipeline and its parameter set. Application of an $M$ to a data set $X$, with hints from some existing knowledge $G$, yields a discovery $D$. $P_D$ is a scoring function that quantifies the interestingness of discovery $D$ in the context of $G$. For the sake of convenience, we define the sign of $P_D$, so that lower scores correspond to more interesting discoveries. The goal of a computational study is to minimize:

$$P_D(D(M, X, G), G) \quad \forall M \in \mathcal{M},$$

where $\mathcal{M}$ is the ensemble of all possible valid analysis methods applicable to the knowledge discovery process.

A possible implementation of the above framework in the task of gene expression-based functional annotation can be illustrated as the following. $X$ is an $n \times m$ matrix of gene expression profiles, where $n$ is the total number of genes in the system of study and $m$ is the total number of experiments. An example of an $M$ could be the whole process of applying the log-transformation to $X$ first, followed by using an ANOVA $p$-value cut-off of 0.05 in the data-filtering process, and then find all the genes whose patterns show a minimal similarity of 0.05 in the data-filtering process, and then find all the genes whose pattern-mining program recursively traversed every vertex in the GO acyclic graph (DAG), with each vertex denoting a functional group. The current gene annotations for $P. falciparum$ are mostly maintained by the Sanger Institute (UK). We downloaded the March 2004 version of the GO database and excluded GO evidences that had not been manually approved by a curator. The malaria microarray consists of probes for genes possessing annotations for 696 biological processes, 716 molecular functions and 225 cellular components. Out of 5159 genes in our data set, 2096 have certain annotations (including trivial categories such as ‘biological process unknown’). For a particular function group of interest, $G$ comes in the form of a list of genes. For example, PFE1150w, MAL3P1.7 and PF14_0455 are the three genes currently possessing the annotation ‘Response to Drug’ (GO:0042493).

The hyper-dimensional space of analysis methods

Our pattern-mining program recursively traversed every vertex in the GO graph from the three root categories and applied OPI algorithm for each functional group independently. It is crucial for OPI to consider a diversified method space, since it is not known in advance which method will be the best of choice for a particular group (more details in Discussion). Therefore our method ensemble, $\mathcal{M}$, used in this study was chosen to be a five-dimensional parameter space, which spanned three discrete binary dimensions and two continuous real dimensions:

$$\mathcal{M} : P_{\text{ANOVA}} \times \left(\text{Linear Log} \times \begin{cases}
\text{Unweighted} \\
\text{Weighted}
\end{cases}\right) \times \begin{cases}
Q_{\text{Single}} \\
Q_{\text{Average}}
\end{cases} \times S_0,$$

$P_{\text{ANOVA}}$, is the cut-off $p$-value threshold used to filter out any non-differentially expressed genes in our data set based on a probe-level one-way ANOVA test (Le Roch et al., 2002). In our study, we limited our $P_{\text{ANOVA}}$ search within the range of 0.05 to 0.2 purely for the sake of keeping computational processing time to a reasonable length. Pearson correlation was applied to calculate the similarity between two expression profiles with four variances depending on: (1) whether expression data were log-transformed or not; (2) experiments were weighted or not. If the weighted option was applied, the 14 experiments in the erythrocyte cycle were given a weight of 1/14 each to sum up to one, the gametocyte and the sporozoite experiment was each assigned a weight of 0.5 (a weight of 1.0 seemed inferior compared to 0.5 based on initial trials). The weighted Pearson correlation coefficient between two arbitrary real vectors $x$ and $y$, with a weighting vector $w$, was defined as:

$$r_w(x,y) = \frac{\sum w_i x_i y_i - (1/\sum w_i) \sum x_i \sum y_i}{\left(\sum w_i x_i^2 - (1/\sum w_i) (\sum w_i x_i)^2\right)^{0.5} \left(\sum w_i y_i^2 - (1/\sum w_i) (\sum w_i y_i)^2\right)^{0.5}}$$

The weights aimed at evening the contributions of gene expression measurements for each crucial stage in the parasite life cycle and it was later proven to be a helpful technique. Two methods were implemented in constructing the query vector $Q$, which acted as a bait to fish out both known and novel genes for a particular functional group. In an analysis, the $N_G$ known genes surviving the filtering process controlled by $P_{\text{ANOVA}}$ were ranked based on their average Pearson correlation coefficient to the rest $N_D - 1$ genes in the descending order. If the single top known gene was chosen as the bait, the
Genes were rediscovered by OPI, log \(G\) of a gene was encountered (denoted as filled boxes in Fig. 1). As more annotated genes were averaged instead, the query vector was denoted as \(Q_{\text{Average}}\). \(S_0\) was the minimal similarity score required for a gene to be included into the final discovery list. \(S_0\) ranged between \(-1\) and 1.

**Global optimization in the method space**

Due to the fact that knowledge \(G\) comes in a discrete form in our study, only a finite number of discrete values in both \(P_{\text{ANOVA}}\) and \(S_0\) dimensions correspond to local minima of \(P_D\). Using this observation, instead of working on continuous real dimensions, our optimization program exhaustively searched all possible local maxima in the parameter space, with the only exception that a subset of evenly spaced \(P_{\text{ANOVA}}\) values were sampled due to the limitation of computational resources. To illustrate this process, Figure 1 depicts the analysis of the GO category ‘Cell–Cell Adhesion’ (GO:0016337) group as it is searched along the \(S_0\) method dimension with the other four dimensions fixed (\(P_{\text{ANOVA}} = 0.17, \text{Linear-transform, Unweighted, } Q_{\text{Average}}\)). There were 10 annotated genes under GO:0016337: PF07_0051, PFD0635c, PFD0630c, PFD0625c, PF08_0103, PFB1055c, PF07_0049, PFB0935w, PFC0120w and PFC0110w. All the genes were ranked based on their correlation coefficients to the reference expression pattern in descending order. The dark curve in Figure 1 shows the knowledge score \(P_D\) (represented in the log_{10} scale) as a function of the number of top genes selected. The similarity threshold \(S_0\) was lowered, more genes down the list were included. \(P_D\) values decreased and formed local minima whenever an annotated gene was encountered (denoted as filled boxes in Fig. 1). As more annotated genes were rediscovered by OPI, \(P_D\) values decreased until PF07_0049, erythrocyte membrane protein 1, PfEMP1, the 23rd gene in the sorted list, was included. This global minimum, \(P_D\) of \(-13.66\), corresponded to an \(S_0\) of 0.55. The remaining three known cell–cell adhesion genes occurred after position 132, with correlation coefficients less than 0.32. Hence, their expression profiles were found to be too different from the previous seven genes to be included in the final list. Therefore, OPI identified a cluster of 23 genes most likely to be involved in the process of cell–cell adhesion with a TPR of 70% \((N_C/N_G = 7/10)\). Since only \(N_P = 10\) out of the list of 23 identified genes have GO annotations, a moderate FDR of 30% is assigned as the lower bound based on the formula mentioned previously: \((N_D - N_C)/N_D = (10 - 7)/10)\).

**False positive rate controlled by permutation studies**

If a biologist studies the data set with one particular functional group in mind, the \(P_D\) score reflects the type I family error in the statistical significance of the resultant gene expression pattern. However, the OPI method is most powerful when it is applied to systematically discover interesting patterns for all function groups in a data-mining approach; therefore the \(P_D\) value has to be corrected for the ‘multiple testing problem’ (Benjamini and Hochberg, 1995; Storey, 2002). In correcting \(P_D\), we are not aware of any readily available mathematical model that can accurately capture both the complexity of a method \(M\) and the DAG structure of the GO family. Although Bonferroni and other types of corrections have been proposed for GO terms (Readme: MGI GO Term Finder. http://www.spatial.maine.edu/~mdolan/Readme_for_GO_Term_Finder.html), they are too conservative for the problem under consideration. Therefore, we resorted to a computationally intensive permutation-based method to estimate the true statistical significance. Despite the disadvantage of having a long running time, this approach had the advantage of taking into account all factors involved \((M, X \text{ and } G)\) and thereby giving the true \(p\)-value of a gene list occurring by chance. In a particular permutation simulation, gene expression profiles were randomly reassigned to other genes; the GBA relationship was broken and became a true null hypothesis. The whole OPI process was repeated for \(R_N\) such permutation runs. If in a total number of \(R_N\) simulations, OPI discovered gene expression patterns with a better score than the original pattern under examination, the original pattern was assigned a \(p\)-value of \(R_0/R_N\). Remember that \(R_N\) simulations only yield a resolution of 1/R_N, therefore \(p < 1/R_N\) if a better pattern was not identified in any of the permutation tests. Better statistics can be easily derived since the results follow a simple binomial distribution. In the example of the cell–cell adhesion group, the gray dots in Figure 1 show the results from five independent permutation runs. It is clear that the \(P_D\) of \(-13.66\) was fairly unlikely to occur by chance.
RESULTS

GNF-TSRI Malaria e-Annotation Database

After applying a total of 100 permutation runs, 320 GO categories were identified to have a permutation-based P-value less than or equal to 0.05, which meant that no more than 5% of the random runs lead to a better P_D score. These results included 167 biological processes, 84 molecular functions and 69 cellular components. Among them, 104 categories had an FDR below 50%. Here, we constructed a public GNF-TSRI Malaria e-Annotation Database with a web interface to enable the research community to visualize and analyze both our analysis results and the underlying data sets (http://carrier.gnf.org/publications/OPI/). The database contains all the 320 categories for the sake of completeness; users may use the sorting interface to apply any threshold appropriate for their purpose of study. The key features of the web application include: (1) searching by a GO/gene identifier or by description keywords; (2) visualizing both heat maps and scatter plots of each significant GO category; (3) exporting capabilities for all the results for further study.

The OPI analysis is a dynamic process, so that as the GO database annotation quality improves and more experimental data becomes available over time, the database will also evolve and become a valuable complement to PlasmoDB (http://www.plasmodb.org).

Systematic view of coordination among functional categories

Biological process, molecular function and cellular component are three independent and complementary organizing principles used by the GO database. Studying how these categories network together provides a better understanding of complex biological events at the systems level. Since our OPI analysis identified a representative gene expression pattern for each category, we clustered the 320 GO categories hierarchically based on the similarities among their representative query expression patterns (Fig. 2). Unlike previous gene expression analyses, to our knowledge this is the first time gene expression data were clustered at the functional-group level instead of individual genes or samples. For example, the segment C in Figure 2 contains the following biologically similar functional groups: ‘Antigenic Variation’ (GO:0020033), ‘Passive Immune Evasion’ (GO:0042782), ‘Defense Response’ (GO:0006952), ‘Host Cell Plasma Membrane’ (GO:0020002), ‘Cell Surface Antigen Activity, Host-Interacting’ (GO:0042280), ‘Cell–Cell Adhesion’ (GO:0016337), and ‘Rosetting’ (GO:0020013).

Validation of functional annotation for antigenic variation group

In addition to the excellent agreement with previous observations as discussed below (Le Roch et al., 2003), our in silico gene function predictions are supported by other studies as well as by sequence homology evidence (Young et al., to be published). However, due to the lack of knowledge of the total number of true positives and true negatives in each GO category, along with the inability of assessing whether a new predicted functional annotation is a real biological discovery or a mistake, systematic computational validation is not feasible (Troyanskaya et al., 2003). Here we use ‘Antigenic Variation’ (GO:0020033) as an example to illustrate the quality of our electronic annotation. When OPI was first developed, the October 2003 version of GO database was used and 67 genes were predicted to associate with antigenic variation. This group had a TPR of 74%, an FDR of 58%, and a log P_D of −21.75. There were 50 genes in the list that did not contain any direct evidence based on the GO database at the time. In the latest study, the March 2004 GO database was used and we found 246 genes with a TPR of 81%, FDR of 58%, and a much better log P_D of −39.03. Therefore, the later GO version enabled us to discover more antigenic genes without observing an increase in the FDR. Retrospective analysis showed that 12 out of the 50 genes originally predicted in silico by the October 2003 study are now confirmed. This also shows that the FDR formula based on annotated genes gives a conservative upper bound estimation; the quality of OPI results are actually often times much better. OPI is a functional annotation algorithm that will likely only improve as knowledge of P. falciparum biology becomes more comprehensive.

Pattern similarity and functional granularity

No matter what analysis method is chosen, the similarity threshold is probably one of the most important parameters directly controlling the quality of the resultant gene list. FDRs were plotted against correlation coefficient thresholds S_0 for all 320 functional groups (Fig. 3). All three GO root categories (BP, MF and CC), represented by different shapes, showed similar trends, i.e., as the expression profiles of the genes became more similar, the quality of functional annotation got better. However, it was also observed that the GBA relationship was merely a general trend; for any single FDR value, there was a significant spread in S_0. This means that there is no universal quantitative formula relating FDR to expression similarity; the threshold S_0 must be set specifically for each individual GO category. Although a recent study concluded that a minimal similarity of 0.85 is required for genes to have more than 50% probability of sharing co-regulated binding sites (Allocco et al., 2004), this study demonstrates that such a ‘one-size-fits-all’ approach is likely misleading. For example, a much lower similarity threshold is required in order to assign a gene to a generic broad functional category such as ‘Protein Metabolism’ (GO:0019538, S_0 = 0.38), while a more stringent criterion should be imposed in order for a gene to be assigned to a narrower group such as ‘Cation Transport’ (GO:0006812, S_0 = 0.81). Therefore, the similarity threshold for GBA principle cannot be generalized to one single value for all functional categories.

To further understand the relationship between minimum pattern similarity S_0 and underlying functional granularity, the S_0 for all ancestor-descendant functional-category pairs among the 320 significant clusters were plotted descendant S_0 (y-axis) versus ancestor S_0 (x-axis) (Fig. 4). The fact that the data points overwhelmingly lie below the diagonal coincides with our expectation that OPI is able to adjust S_0 automatically based on the intrinsic functional granularity of the underlying category. The S_0 value, therefore, naturally serves as a continuous measure for GO term levels. Simple integer GO term levels have once been employed in a study as a quantitative measurement of biological granularity (Allocco et al., 2004). We believe that the S_0 value will serve a better metric in such studies. S_0 may also be used to prioritize targets as genes belonging to those pathways with highly conserved expression patterns will arguably serve as better drug targets or as more reliable biomarkers. We anticipate these results will find their uses in the near future.

Comparison of analysis methods

Figure 5 depicts the number of times each of the eight methods yielded a group with an FDR less than 50%, resulting in a total of 104
Ontology-based pattern identification

Fig. 2. A hierarchical tree of 320 gene ontology categories based on the similarities among their representative query expression patterns. The tree offers a functional view of the *P. falciparum* life cycle at the systems level.
Fig. 3. The functional annotation quality (FDR) plotted against the optimal correlation coefficient threshold $S_0$. As a general trend in agreement with GBA, FDR decreases as $S_0$ increases. The size of the marker represents the size of a resultant gene list, which also anti-correlates well with $S_0$. The exact relationship between FDR and $S_0$ is highly category-specific.

Fig. 4. Comparison of the optimal correlation coefficient threshold $S_0$ among descendant GO categories and ancestor GO categories. It shows that $S_0$ identified by the OPI algorithm agrees well with the underlying biological granularity. As a GO category becomes more specific in characterizing gene functions, GBA imposes a more stringent requirement for the similarities among gene expression patterns. The size of the marker represents the distance between paired categories simply measured by subtracting their integer depth of GO levels. The fact that most data points falling above the diagonal are close ancestor-descendant pairs indicates these outliers are likely explained by statistical noise.

groups. Query method patterns by $Q_{\text{Average}}$, usually outperformed those using $Q_{\text{Single}}$ with 80 average-winning categories versus 24 single-winning categories. The statistical significance of the preference for $Q_{\text{Average}}$ is supported by a $p$-value of $10^{-38}$ estimated based on a binomial distribution. One must be cautious not to overemphasize the advantage of $Q_{\text{Average}}$ though, since $Q_{\text{Single}}$ was actually preferred in 23% of the cases, which is still a fairly significant portion.

Despite the common practice of applying log-transformation during data preprocessing, we did not observe log-transformation showing any statistically significant advantage ($p$-value = 0.12) over linear-transformation in our analysis (Geller et al., 2003; Pan et al., 2002). It was also shown that applying weighting factors to the data was undesirable for the majority of categories ($p$-value = $10^{-105}$). This indicates that a great deal of information likely arises from the
Intraerythrocytic cycle data. These observations support our previous k-means study, where linear-transformation and uniform weighting were applied.

**DISCUSSION**

In applying GBA for expression-based functional genomic studies, unsupervised learning methods such as hierarchical clustering (Eisen et al., 1998), k-means clustering (Le Roch et al., 2003; Tavazoie et al., 1999), self-organization map (Tamayo et al., 1999) are commonly used. Such methods provide efficient hypothesis-generation tools whose results are less prone to bias. However, both the quantity of biologically significant expression patterns and the qualities of these clusters and functional annotations are difficult to control in unsupervised learning, due to the many subjective parameters involved. Recently, Mootha et al. designed an algorithm called gene set enrichment analysis (GSEA), which relies on using only annotated genes in the GO database to enrich weak differentially expressed signals (Mootha et al., 2003). Using this approach, they were able to successfully identify the PGC-1α-responsive pathway to be involved in type 2 diabetes mellitus. The GSEA algorithm, however, takes a supervised approach and totally relies on their biological annotations for generating clusters. OPI overcomes the limitations of these approaches by taking a promiscuous path. While genes are still grouped based on their expression similarities as in the case in unsupervised learning methods, the degree of similarity required, namely $S_0$, is controlled based on the quality of annotated genes according to existing biological knowledge, which is similar to supervised learning methods. The novelty of the OPI method enables it to bridge the two opposite camps of learning approaches.

In comparison to the robust k-means method we previously employed (Le Roch et al., 2003), OPI identified functional gene clusters with higher statistical confidence when applied to the same expression data set. OPI identified gene clusters for nearly all of the 17 GO functional categories described by the 15 k-means clusters with the only exceptions being ‘Lipid Metabolism’ (GO:0006629) and ‘Fatty Acid Metabolism’ (GO:0006631). However, genes with annotation for lipid metabolism and fatty acid metabolism were only slightly enriched in cluster four of the k-means study, with marginal $p$-values of 0.029 and 0.048, respectively. OPI did however identify statistically significant clusters for the related GO categories ‘Membrane Lipid Metabolism’ (GO:0006643) and ‘Membrane Lipid Biosynthesis’ (GO:0046467) with $p$-values of $6 \times 10^{-6}$ and $2 \times 10^{-5}$, respectively. For all of the other GO functional categories described by the k-means clusters, however, OPI generated a corresponding cluster with comparable or greater statistical significance. The largest improvement in $p$-values for a GO functional category from a k-means cluster to an OPI cluster was seen for ‘Antigenic Variation’ (GO:0020033), which went from a $p$-value of $4 \times 10^{-9}$ to a $p$-value of $9 \times 10^{-10}$. The only instance where the statistical significance of an OPI generated cluster was lower than for the k-means cluster was for the GO functional category ‘26S Proteasome’ (GO:0005837), whose k-means cluster $p$-value was $2 \times 10^{-9}$ compared to a similar $p$-value of $1 \times 10^{-8}$ for the OPI cluster.

In addition to identifying functional gene clusters with greater statistical confidence than was possible with a robust k-means analysis, OPI also allowed the generation of more function-specific clusters. For example, in the k-means study, cluster 15 comprised of 130 genes whose expression patterns described cell invasion, the apical complex, and the rhoptry organelles. However, due to the nature of the k-means clustering method, it was impossible to identify sub-patterns...
within this cluster that might better describe each of these functional categories individually. OPI analysis allowed the resolution of these function-specific sub-patterns with the identification of a cell invasion cluster of 53 genes, an apical complex cluster of 82 genes and a rhpopy cluster of 6 genes. Furthermore, due to the recursive nature of OPI, genes were allowed to fall into more than one of these clusters if their expression pattern warranted so, thus suggesting possible multi-functional roles. For example, the hypothetical gene PFE1410c was included in the cell invasion and apical complex OPI clusters, but not in the rhpopy OPI cluster. Furthermore, this gene was also included in the phosphoprotein phosphatase activity OPI cluster. Analysis of the predicted protein sequence of this hypothetical gene indicated that it contains a possible phosphoenolpyruvate-dependent sugar phosphotransferase system EIIA 2 domain. In bacteria, this system is involved in the transfer of phosphate groups to incoming sugars as they are translocated across the cell membrane, a process that conceivably could be occurring at the apical complex of the parasite during the cell invasion processes. This example demonstrates precisely how specific multi-functional gene roles can be predicted using OPI in a manner not possible using other standard clustering algorithms such as a robust k-means approach.

One potential criticism of OPI is that it identifies only one representative expression pattern for each GO category, which certainly is an approximation especially for those categories involving several different regulation mechanisms. However, this is not a limitation if one considers the recursive nature of the gene ontology. For example, regulation of ‘DNA Transposition’ (GO:0000337) consists of two child categories: ‘Positive Regulation of DNA Transposition’ (GO:0000336) and ‘Negative Regulation of DNA Transposition’ (GO:0000335). If applied recursively, OPI can identify two distinct gene expression patterns for GO:0000335 and GO:0000336, respectively. A representative expression pattern for GO:0000337 will come from whichever mechanism dominates the parent category in the given species and the data set under study. The OPI algorithm can be applied in parallel to each individual GO category as well as each independent permutation simulation; therefore it is straightforward to take advantage of the distributed computing environment such as a Linux cluster. At GNF, OPI has been successfully utilized to study microarray data sets of more complicated organisms such as humans and mice. It is still true, however, that certain genes will not co-cluster because they are involved in regulatory processes that may be constitutively expressed or may be orthologously expressed relative to the process they regulate. Therefore, in some cases, genes pertinent to some biological processes are likely to be overlooked using only OPI analysis. Despite these limitations, however, OPI provides starting points for further investigation of many genes involved in biological functions of which little is known, which is especially useful in a largely uncharacterized organism such as *P. falciparum*.

Another frequently encountered criticism of OPI is the incompleteness of the GO database itself. It is understood that gene annotation is an ongoing effort of the biological community and we expect OPI results to improve as gene annotations improve over time. It should also be pointed out that OPI can easily be applied using other knowledge databases such as the protein family database or even customized categories derived from the user’s own biological expertise. Additionally, the application of OPI is not limited solely to gene expression data, as it can also be directly applied to many other types of genome-wide biological data sets such as those arising from proteomic studies.

We have summarized several open informatics challenges particularly related to GBA-based functional analyses. The answers to these pressing issues are generally not available because of the lack of both perfect data sets with known biological truths and benchmarking standards. We do not believe a universal answer exists to each of the above questions; the answers will be different depending on the data sets and biological questions being studied. Therefore, instead of solely judging methods based on, for instance, their statistical rigor or process, OPI benchmarks analysis methods and parameters based on their capability of producing computational discoveries that best conform to existing biological knowledge. By using a pair-wise similarity measurement, we avoided the ambiguity of whether to use the k-means clustering or the hierarchical clustering method. By considering a large five-dimensional method space, we took into account the effect of data filtering (represented by the PANOVA dimension) and data transformation (either linear- or log-transformation). The effect of the single-linkage or average-linkage in a hierarchical clustering method was mimicked by choosing between Q_Single and Q_Average, and sample dependency was considered by choosing between weighted or unweighted similarity measurement schemes. Our observation was that each method held an advantage in certain categories, but never in all categories. Therefore, the answer to the open questions is that one should not commit to any single analysis method, but rather evaluate all methods simultaneously and identify the most successful analysis approach in a category-specific manner. OPI represents a new and unique tool aptly suited for this task.

During the preparation of our manuscript, two related studies have been published. Breitling et al. (2004) introduced an iterative group analysis algorithm and demonstrated a different application of the similar ontology-based analysis approach in identification of differentially expressed gene classes. Toronen (2004) applied another ontology-based analysis method to identify best scoring clusters (sub-trees) on top of an expression-based hierarchical gene tree, where the results independently confirmed the necessity of using different scoring thresholds (analogous to S_0) for different gene classes. Compared to these studies, OPI not only identifies significant gene classes and representative gene expression patterns, but also offers a unique advantage of studying biological implications of various expression analysis algorithms as discussed above.

To summarize, OPI is an efficient pattern-mining algorithm that enables genome-wide in silico function annotation of uncharacterized genes based on existing ontology knowledge. We have successfully applied OPI to a *P. falciparum* life cycle gene expression data set and constructed a public database that will be a valuable tool for the malaria research community. OPI provides a novel mathematical framework for the benchmarking and understanding of the performance of many gene expression analysis methods, and for identifying the best method in a functional category-specific manner.

ACKNOWLEDGEMENTS

We thank Ogi Boras for setting up the Linux cluster that enabled this study, as well as Guangzhou Zou, Karine Le Roch, Quinton Fifelman, Peter Blair, David Baker, Daniel Carucci and the anonymous referees for helpful comments.

REFERENCES

Ontology-based pattern identification


