Biological Sciences, Applied Biological Sciences, Genetics

Similarity searches in genome-wide numerical data sets

Galina Glazko, Michael Coleman, and Arcady Mushegian*

Stowers Institute for Medical Research, 1000 E 50th St., Kansas City MO 64110, and *Department of Microbiology, Molecular Genetics, and Immunology, University of Kansas Medical Center, Kansas City, KS 66160, USA

Correspondence: Galina V. Glazko. E-mail: gvg@stowers-institute.org

The number of text pages: 22

The number of figures: 7

The number of tables: 6

The number of words in the abstract: 175

The total number of characters in the paper: 38,247

Clustering approaches are commonly used to navigate and interrogate gene expression profiles, protein-protein interaction information, and other large genomic datasets. However, many biological questions that are investigated by analysis of the genome-wide measurements are not global-clustering problems at all. Rather, a frequent problem is to find the neighbors of a query, which is a vector in the multidimensional measurement space, and to rank these neighbors by similarity to the query. To address this local-clustering problem, we developed an iterative pattern-matching program called psi-square. The program searches the space of genome-wide vectors, finds a group of highly similar vectors, derives a probabilistic model of that group, and repeats database search using this model as a query. We applied the method to several pathway-discovery problems, which use three types of genome-wide datasets, namely gene content in microbes, gene expression in the blood stage of malaria parasite, and protein-protein interactions in yeast. The unified method of analysis is generally more sensitive and in many cases also more specific than each of the specialized methods applied to these data before.

Genome era produces large multidimensional datasets, which need to be analyzed in robust, quantitative ways. The first-aid response to the advent of gene expression data and other genome-scale measurements was cluster analysis. The techniques of global partitioning of the data, such as K-means, partitioning around medoids, various flavors of hierarchical clustering, and self-organized maps (1-4) , have provided the initial picture of similarity in the gene expression profiles, and helped to infer functional links between genes. However, cluster analysis has its drawbacks. Typically, once gene is assigned to a cluster, it remains in that cluster, even though many genes participate in more than one pathway. Furthermore, the degree of intra-cluster similarity between expression profiles may not be the same for every set of functionally linked genes, which puts limitations on the use of cutoffs and on the number of clusters that can be predicted with confidence. Several approaches have been suggested to overcome these problems, for example, iterative clustering and iterative maximization of the partition quality (5).

Another approach to finding functionally relevant groups of genes is network derivation, which has been popular in the analysis of gene-gene and protein-protein interactions (6-10), and is also applicable to gene expression analysis (11, 12). This class of methods overcomes the inflexibility of hierarchical clustering/partitioning approaches. However, network definition is also confronted with the issue of estimating statistical significance, and, as with partitioning approaches, the significance threshold can be different in different parts of the same network (13). In addition, visualization and navigation of links in the highly connected network poses its own set of computational challenges.

Although the general picture of dependencies between genes and their products can be obtained by these methods, in fact many biological questions asked of the genome-wide

measurements have little to do with global clustering or with delineation of the whole network. Rather, a commonly encountered task is to discover the neighbors of a point, which represents a set of measurements associated with a gene or a protein. Finding such groups does not require the knowledge of the complete set of genome-wide correlations – the fundamental task here is to discover and rank similarities that are local with regards to that network. Pathway reconstruction and discovery of functional links belong to this class of tasks: we are given one or a few members of a pathway, and would like to infer the other, functionally linked components of the same pathway. Functional links may be discovered, for example, by similarity of expression profiles (2, 14, 15), or by similarity between the set of protein-protein interaction partners (16, 17), or by co-inheritance of groups of genes across different genomes (18). If a query belongs to a functionally and evolutionarily defined module, we want to find as many members of this module as possible. At the same time, many - perhaps most - entities in the measurement space are not involved in the module of our interest and, with correctly chosen statistics, should display only the random-level similarity to the query.

This logic has been exploited for decades, and with considerable success, in another area of computational biology, i.e., in sequence similarity-based prediction of biopolymer structure, function, and evolutionary origin. Nowadays, it is standard to begin studying new sequence with a database search, performed by a program like BLAST (19) or PSI-BLAST (20). If there is a similarity between an uncharacterized query sequence and a better-studied sequence in the database, this information can be used for structural, functional, and evolutionary inferences. At the same time, the similarities between sequences that are unrelated to the query are not of interest, and there is often no need to examine them at all.

In this work, we apply the same logic to searching in the multidimensional space of genome-wide numeric datasets. The search is performed by an iterative pattern-matching program that was inspired by PSI-BLAST, and is called psi-square (as in "pseudo-PSI"). The idea of the algorithm is to start with a numerical pattern of interest (gene expression profile, gene occurrence pattern, protein interaction list, or any other), to find group of highly similar patterns, to derive a probabilistic model of that group, and to repeat database search using this model as a query. In the rest of this paper, we describe the psi-square algorithm and software, and apply it to several pathway-discovery problems, which make use of three very different types of genome-wide datasets.

Materials and Methods

Algorithm. The summaries of genome-wide measurements associated with a given gene have been called "profiles" and "patterns" (e.g., "phyletic patterns" (21, 22) or "expression profiles" (15)). For the sake of generality, we will call a set of numbers (measurements) associated with the i^{th} gene "a gene vector". In different experiments, the same gene can be associated with a phyletic gene vector, an expression gene vector, a protein-protein interaction gene vector, etc. Different measurements for the same gene can, in principle, be combined. In this study, however, we are concerned with the examples for which each coordinate of each vector represents one and the same type of measurement.

A gene vector space, or vector database, is a set of vectors $\mathbf{X_i}=(x_{i1},.x_{i2},\ldots,x_{iN})$, where $i=1,..,M$ and $j=1,\ldots,N$, and M, N indicate, respectively, the number of genes and the number of data points/experimental conditions associated with each gene. We assume that a vector of interest, called "query", is known (either produced by actual measurements, or made up), and

we want to find similar vectors in the database. The query may represent a set of relative or absolute measurements, as with gene expression data; or it may consist of numerically encoded discrete states, such as gene presence-absence, gene expression or lack thereof; or it can be a probabilistic model derived from a series of related vectors. We will use "profile" to refer to the set of all probabilities associated with every coordinate in a vector (23) and will use "condition-specific scoring matrix" (CSSM) as a synonym for profile.

The psi-square algorithm searches the database for vectors that are similar to the query vector (or query profile/CSSM, i.e., a probabilistic model of several related vectors). The program takes query vector X_i as its input and produces a set of similar gene vectors (a subset of the vector database) as its output. The logic of the algorithm is reminiscent of iterative sequence similarity search and has two iterative steps: (1) compare the profile formed from the query vector and, perhaps, other closely related vectors to the entire vector database; (2) update the profile based on the high-scoring matches, producing the CSSM of scores s_{kj}, where $k=1,..,K$ and $j=1,..N$, and K is an additional parameter that corresponds to the number of discrete categories (see below).

Conditions may represent different treatments, different time points in gene expression experiments, different genomes in the phyletic pattern space, etc. The number of conditions (N) is the number of vector coordinates. A vector or a vector set of interest are called target vector set, T, and the complete vector database is called background vector set B. Every element s_{kj} of the matrix is the log-odds ratio $s_{kj} = \log\{\Pr(a_k,c_j|T)/\Pr(a_k,c_j|B)\}$, where $\Pr(a_k,c_j|T)$ is the probability of observing the value a_k under condition c_j in T, and $\Pr(a_i,c_j|B)$ is the probability of observing a_k under the same condition in B. The probability is estimated as the frequency (f^T_{kj} or f^B_{kj}) of the given observation under the specific condition in the target

or background vector sets, respectively. This scoring scheme is intuitive and familiar from the theory of sequence comparison. In the context of sequence similarity searches, the log-odds scores derived from the target dataset are known to be optimal for signal recovery (24). High sensitivity of this scoring scheme in our hands (see below) suggests that log-odds scores may be likewise close to optimal when applied to different types of gene vectors, though this proposition remains to be formally proven.

A high-scoring pair (HSP) is a pair of vectors such that their similarity exceeds a certain threshold. The similarity measure and significance threshold may be derived from empirical observations, or from a process model of some kind. In this article, we focus on a similarity/distance measure derived from the correlation coefficient, while psi-square software allows one to choose from several distance measures.

The algorithm proceeds through the following steps:

1. Initialize the program with a vector or a group of related vectors (initial value of target vector set).

2. Construct the scoring matrix (CSSM) of the form $s_{kj} = \log(f^{T}_{kj}/f^{B}_{kj})$, where k varies over the number of possible values of vectors (or transformed vectors, see below) and j varies over the set of conditions.

3. Use the CSSM as a query at the next iteration of the search. Score similarity between CSSM and each database vector as follows: $S(\text{vector})=\sum s_{kj}$, where s_{kj} is the score of value k under condition j in a given vector. Vectors with higher similarities to CSSM get higher scores. Construct the empirical distribution of these scores. Record vectors with scores from a given percentile (e.g., 95%) of the total score distribution as new high-scoring matches.

4. Add these vectors to the target vector set; update the CSSM.

5. Repeat step 3. The process terminates when we cannot find new matches at step 3.

The vectors' coordinates can be either discrete or continuous. Discrete coordinates often have only two states, e.g., "turned on-turned off" or "present-absent", but they may be multistate. The present algorithm assumes a finite number of states, which can be achieved by discretizing continuous variables. Discretization simplifies the data representation, and some machine-learning algorithms have been shown to perform better with discrete-valued attributes, even though they can also handle continuous attributes (25, 26). The number of states that each coordinate can take after discretization is designated by *K*, a parameter that can be either dictated by an *ad hoc* scientific hypothesis, or computed on the fly. We use equal-width interval binning and set *K* value globally, assuming that the coordinates of all vectors of the same type come from the same probability space. For every vector in the database, the range of its values, E_{max} and E_{min}, observed over all conditions, is calculated with step $\delta = E_{max}-E_{min}/K$. Each vector is transformed to receive a set of discretized coordinates, where its i^{th} value is replaced by the attribute $(tr^{k}{}_{x_i})$ .The number of intervals depends on the data set. For example, in sequence similarity analysis, the number of initial states for nucleic acids may be naturally set for five – four nucleotides and the gap. For coded binary character states, such as presence/absence, *K* is 2. For other types of data, the value of *K* is estimated from the initial target vector set (see Supporting Text for details).

There are two more parameters that have to be specified, the similarity threshold for the inclusion in the target vector set, *r*, in step 1, and the percentile of the score distribution that is used as the inclusion cutoff *s*, in step 3. The optimal values of *r* and *s* depend on the

sample size and signal-to-noise ratio in the data, and their selection is similar to the decisions commonly made in sequence database searches.

Phyletic vectors. Gene presences and absences are summarized in the COG database (http://www.ncbi.nlm.nih.gov/COG/new). There were 4873 COGs from 66 complete genomes of unicellular organisms in the COG database, as of September 21, 2004 (22). 284 fungi-specific COGs were not considered in this study. Each *i*th COG ($i = 1,\ldots,4589$) is a phyletic vector, where the *j*th coordinate ($j = 1,\ldots,63$) is set at 1 if it is represented in the *j*th genome and 0 if it is not (we ignore some details, such as the presence of in-paralogs in some COGs – see (22) for discussion). In this case K is set at 2, corresponding to two possible values of binary coordinates, 0 and 1; r and s parameters were adjusted interactively.

Protein-protein interaction vectors. We used the tandem-affinity purification (TAP) data set from Gavin et al. (27) and removed purifications that only retrieved the bait itself. This retains 455 purifications, containing 1361 proteins. The K parameter is naturally set at 2; r and s parameters were adjusted interactively.

Gene expression vectors. Gene expression data for the asexual intraerythrocytic developmental cycle (IDC) of the malaria parasite *P. falciparum* are from Bozdech et al. (28) (Quality Control data set, 5081 vectors with 46 coordinates). Missing data and outliers (coordinates deviating more than 3 s.d. from the mean value for a given vector) were replaced by the mean; this is called "the IDC set" in the sequel. The parameters for this data

were chosen iteratively, in order to maximize the number of new matches and minimize the average number of matches (see details below).

Specificity and sensitivity estimates. When the training sample (list of proteins with desired properties) is available, we compare the sensitivity and specificity of psi-square with the performance of the approaches used in the literature for each analysis. Specificity is computed as TP/(TP+FP) and sensitivity as TP/(TP+FN), where TP denotes true positives (genes/proteins included in the training sample); FP denotes false positives (genes/proteins not included in the training sample), and FN denotes false negatives (genes/proteins included in the training sample but not found by the approach). For simplicity, we treated all genes found by only one approach as false positives (overestimating FPs, because this does not account for the novel predictions that may be ultimately proven correct).

Results

Phyletic vectors. Information about phyletic distribution of orthologous genes, i.e., presence and absence of orthologs in completely sequenced genomes, is of interest because functionally linked proteins tend to be co-inherited in the same subsets of genomes (29). Informally, co-inheritance has been approximated by low Hamming distance (e.g, three bits or less) between phyletic vectors (18), but a more systematic analysis indicated that other distance measures, in particular those based on correlation, can greatly improve the sensitivity of functional inference from co-inheritance (30). One case study in this work is the search for new components of flagellae in bacteria based on their co-inheritance with the known flagellar components.

Flagellae, the sensory and locomotive organs, are found in 23 bacteria out of 50 in the COG database. Two parasitic bacteria, (*Chlamidia trachomatis* and *Chlamidophila pneumoniae*), and two symbionts (*Buchnera sp. APS* and *Yersinia pestis*) do not have flagellae, but contain several genes orthologous to flagella assembly factors in other species, presumably because these genes have additional functions, such as assembly of other extracellular protein complexes (31, 32). The genomic signature of the flagellar biosynthetic and structural genes is represented by a vector with 27 coordinates set to one and 23 coordinates set to zero (Fig. 1, supporting information). There are only 6 COGs characterized by such a phyletic vector, yet at least 37 bacterial COGs are known or inferred to be directly involved in bacterial flagella biogenesis and function (Fig. 1). Thus, phyletic vectors of at least 31 flagella-related genes mismatch the query constructed on the basis of flagella genomic signature – in the extreme case, by 22 conditions (Fig. 1). The likely reasons for these mismatches include differential gene losses and functional takeovers by unrelated genes, and, probably, existence of several modules within the flagella apparatus, some of which are capable of functioning independently, as in the aforementioned aflagellate bacteria (31, 32). Regardless of the reason for patchy distribution of flagellar components, many of them can not be sensitively and specifically discovered by exact matching to a made-up genomic signature, nor with naïve methods of Hamming distance-based matching.

Levesque et al. (33) have suggested a series of algorithms that make functional predictions on the basis of phyletic vectors and set theory. The threshold of similarity between subsets is an adjustable parameter. This “Trait to Gene” software (TTG in the sequel) identifies 33 COGs as associated with flagella phenotype at the most sensitive similarity threshold 0.65 (Fig. 1 and see ref. (33) for details). Among those, 27 COGs have

annotations indicating their involvement in flagella. Thus, the approach results in at least 82% true positives and recovered 73% of the 37 known flagellar COGs (Table 1).

Another approach for functional prediction from phenotype has been suggested by Jim et al. (34). Their method computes the phenotype propensity (PP), i.e., the ratio of two frequencies, that of the genomes that have both phenotype and protein of interest, and of all genomes which have the same protein. Proteins that appear only in genomes with given phenotype have the highest propensities. The PP approach identifies 46 COGs, corresponding to 60 *E.coli* proteins, with the highest propensities to flagella phenotype. Twenty-two of them (59%) overlap with 37 known flagellar COGs (Table 2, supporting information).

We applied the psi-square algorithm to find vectors most similar to the flagella genomic signature vector (Fig. 1). Using COG1298, one of the six COGs perfectly matching the flagellate phenotype, as a query with r set at 0.6, we recovered 45 COGs. Twenty-nine of these COGs are involved in flagella assembly or function (this corresponds to 78% of all known flagellar proteins). Thus, the naïve psi-square approach had higher sensitivity, but lower selectivity, than TTG (Fig. 2, Table 1), and exceeded the PP approach in both specificity and sensitivity (Fig. 3, supporting information; Table 2).

To supplement the naïve psi-square search, we collected 29 flagella-associated COGs found at the first step of the analysis and used them as queries in further rounds of psi-square searches, with the r parameter set more conservatively at 0.7. The union of all newly found matches gives 73 vectors, with 34 true positives, i.e. 92% of the known flagellar COGs (Fig. 4, supporting information). Seven flagellar components were predicted by psi-square at this

step, but were missed by TTG (Fig. 4, supporting information), indicating higher sensitivity of psi-square towards these outlying vectors (Fig 1).

Thirty-four COGs were predicted by psi-square only (Fig. 4c). Phyletic patterns of these COGs were much "patchier" than the flagella genomic signature (Fig. 2e, 4c) . Five of the proteins found only by psi-square, COG2160, COG3154,COG2356, COG0854, COG3154, appear to be unrelated to flagella function and biogenesis (Fig. 4c). On the other hand, among the 34 genes uniquely identified by psi-square, nine are involved in cell division, shape determination, and chemotaxis; these are most likely not spurious matches, as the recent literature suggests several linkages between these processes and flagellar function (35, 36). We expect that several of the remaining COGs, for example some of the transcriptional regulators (COG1221, COG3829, COG3835) and signal transduction proteins (COG3852, COG3605) are also involved in the regulation of flagellar biogenesis. Moreover, 3 proteins found by psi-square (COG1699, COG2257, COG3034, Fig. 4c) may have previously unreported connections to flagellar phenotype, based on contextual information from STRING database (Table 3, supporting information).

Gene expression vectors. The lifecycle of the malaria parasite includes three stages: the mosquito, liver and blood stages. The blood stage is responsible for all of the malaria symptoms and mortality in humans and is therefore an important target for vaccine development (37). Despite much effort, an effective malaria vaccine is still unavailable (38). Recently, the transcriptional program of the asexual intraerythrocytic development cycle (IDC) of *P.falciparum* has been characterized (28). The parasite-specific genes, especially

those related to the initiation of the IDC (merozoite invasion), may be good candidates for vaccine development.

Several candidate antigens have been identified in *P.falciparum*. Most of them are expressed on the parasite cell surface, in particular within apical organelles involved in merozoite invasion (37). Among the best-studied invasion proteins are seven malaria vaccine candidates, AMA1, MSP1, MSP3, MSP5, EBA175, RAP1 and RESA1. Their expression profiles undergo sharp induction during the mid- to late schizont stage. In order to find additional vaccine candidates, Bozdech et al. (28) compared the Euclidean distances between expression profiles of seven antigens and the rest of plasmodium transcriptome, and the 5% of this distribution with the lowest distance (5%ED) was proposed as a plausible set of vaccine candidates. The 5%ED set of 262 ORFs included virtually all known merozoite-associated genes.

We used the psi-square approach to find proteins involved in merozoite invasion in the IDC set. Seven independent searches were initiated with seven antigens as queries. (When one ORF was represented by multiple probes on the chip, we chose the vector with the highest average correlation to the other vectors). We tried several thresholds for correlation and several values of the *K* parameter, with 24 parameter settings altogether (Table 4, supporting information). The correlation threshold 0.9 and K=15 maximized the number of iterations and new matches. Fig. 5 presents matches found during several iterations of psi-square for queryPFA0110w (ring-infected erythrocyte surface antigen precursor). In sum, psi-square and 5%ED identified, respectively, 596 and 419 probes; there were 409 probes found by both approaches, 187 found only by psi-square, and 10 probes found only by 5%ED.

The average maximum time of expression for 187 unique probes, corresponding to 151 unique ORFs found by psi-square, matched 30 hours, i.e., the beginning of the schizont stage. Among them were several already known *P.falciparum* antigenes, such as RESA-H3 (PFB0915w), MSP8 (PFE0120c), octapeptide-repeat (ORA) (PFL0035c), PF70 (PF10_0025), membrane protein ag-1 (PFD0255w), RESA-2 (PF11_0512), tryptophan/threonine-rich antigen (PF08_0003), and transmission-blocking target antigen (PF13_0247). None of these proteins have been identified by the 5%ED method.

For proteins involved in merozoite invasion, we cannot estimate specificity and sensitivity within the previously established framework: not only is the list of true positives unknown, but we intentionally tuned parameters of psi-square so as to find more candidates. Therefore, to compare biological relevance of two approaches, we examined the sequence properties of the two sets of hypothetical proteins (HP), found either by the psi-square approach only (HP_{s1}, 108 proteins), or by both the psi-square and 5%ED methods (HP_{s12}, 154 proteins). The structural properties of the proteins in these two non-overlapping sets are nearly the same, and at the level of predicted molecular function, the two groups of proteins exhibited many common features as well (Tables 5 and 6, supporting information). Both sets were depleted of the housekeeping genes involved in genome expression, in intermediate metabolism, and in signal transduction from cytoplasm to the nucleus. Among the proteins with predicted enzymatic activity, there is a clear prevalence of domains involved in lipid biosynthesis and membrane remodeling. Also seen in both sets are proteins with chaperone activity, components of cytoskeleton and of secretory vesicles, and multiple protein kinases and phosphatases (Table 6). These observations are compatible with the idea of regulated changes in the cell surface and cell shape upon transitioning to the merozoite phase.

Interestingly, HP_{s1} and HP_{s12} recover different bona fide antigen-related proteins (RESA in the case of HP_{s1} and AMA-1 and MSP7 in the case of HP_{s12}).

These results indicate that psi-square is quite specific towards the putative proteins involved in merozoite invasion. At the same time, psi-square is more sensitive than 5%ED method: psi-square has recovered many ETRAMPs, expressed mostly at early ring stage and located at the parasite-host cell interface, as well proteins identified by MudPIT as parasite proteins on the surface of the infected erythrocyte (PIESPs, Florens et al. (39)), none of which was detected by 5%ED. Psi-square also identified PFE0340c, an ortholog of the rhomboid protease involved in adhesin cleavage during invasion of another apicomplexan parasite, *Toxoplasma gondii* (40). Identification of this and other enzymes involved in membrane remodeling and signal transduction suggest an additional strategy of anti-malaria drug development, namely to screen for small-molecule inhibitors of these merozoite invasion-related enzymes.

Protein-protein interaction data. The majority of cellular processes are carried out by multiprotein complexes (41), and analysis of their composition is of great interest. Screening of protein-protein interaction (PPI) at a large scale can be done with yeast two-hybrid technology (42), which registers only pairwise PPI, and with various affinity purification schemes (27), which record the protein content of a complex but not individual interacting pairs. High-throughput screens are noisy because of non-specific binding, fragmentation of the whole complex into subcomplexes (17), low reproducibility (27) and other factors; true protein complexes must be discerned by a combination of analytical biochemistry and computational techniques (27).

We used the psi-square strategy to identify protein complexes in yeast affinity purification data from Gavin et al. (27) . The PPI vector space can be set up in several ways. For example, purification vectors can be compared in the space of protein coordinates, or else protein vectors can be compared in the purifications' space. In the former case, the search result would be the set of purifications similar to the purification of interest; in the latter case, the result is the set of proteins co-purifying with the query protein.

We applied our approach to recover the contents of the protein complex responsible for post-transcriptional maturation of the 3'-end of eukaryotic pre-mRNA. This reaction occurs in several steps, including site-specific cleavage, polymerization of the poly(A) tail, and trimming of adenylate residues to mature length (43). In yeast, the major components of these processes are poly(A)-binding protein (Pab1p), poly(A) nuclease (PAN), and three multidomain complexes, CFIA, CFIB, and CPF (44). Using Pta1 as the first bait, Gavin et al. (27) experimentally identified 12 of the 13 known components of the polyadenylation complex and 7 new putative components.

The psi-square search of interaction vectors initiated with Pta1 converged in one iteration (r=0.6), detecting 10 known components of the polyadenylation machinery (Cft1p, Cft2p, Glc7p, Pap1p, Pfs2p, Pta1p, Ysh1p, Fip1p, Yth1p, Rna14p) and two putative components, Ref2p and YKL059C, which have been also identified by Gavin and co-workers. We then applied the same strategy as with flagella proteins, running 13 psi-square searches, one for each already found component, and taking the union of all newly found vectors. This strategy led to identification of 5 additional components (Ssu72p, YOR179C, Clp1p, Pcf11p, Rna15p) which were also found in TAP-purification analysis (27). In sum,

our analysis identified all components found by TAP, except for two, Pab1p and YKL018W (Fig. 6a, supporting information).

The orthogonal search, initiated with a purification vector of all proteins retrieved when Pta1 was used as a bait, converged at one iteration (r=0.6), resulting in 11 similar purifications. These purifications included 33 proteins (Fig. 6b, Fig. 7: supporting information). Thus, the protein-based query retrieves a set of proteins virtually identical to the original complex found by Gavin et al. (27), whereas the purification-based query discovers many additional proteins. Sequence analysis indicates that among these new findings there are two RNA helicases Has1p and Dbp4p, putative RNA modification enzymes Cbf5p (pseudouridylate synthase-like) and Nop1p (methyltransferase-like), as well as nucleolar proteins Nop56p, Nop58p, and Rsa3p. Many of these proteins are more familiar as components of processosome, the complex that is responsible for maturation of ribosomal RNAs. Recent evidence, however, suggests the existence of extensive cross-talk between processing of rRNA and mRNA (45), and our results point in the same direction.

Discussion

Many global-clustering approaches tend to underestimate functional relationships among gene vectors (46-49). To address the limitations of global clustering, we propose the similarity search program, psi-square, which is applicable to recognition of any kind for patterns represented in vector form. The use of profiles is familiar to molecular biologists from such tasks as prediction of gene structure by homology, delineation of protein families, and fold recognition. The same intuition applies to iterative search of vector spaces for similarities between gene vectors. Because query vectors are converted into probabilistic

models that can be iteratively updated, the resulting sensitivity of the method is higher than in more naïve similarity searches.

The performance of psi-square depends on the choice of distance measure and several search parameters. The optimal choice of distance measure in genome-wide datasets is examined elsewhere (50).

Some of the ideas that were used in psi-square algorithm have been discussed before. Most notably, Zhou and co-authors (48) have introduced the shortest path concept, which seeks to find series of closest neighbors in genome-wide data in an iterative fashion. In contrast to their approach, psi-square does not rely on pre-computed network, but uses a query to interrogate unordered vector space and to produce a probabilistic model of the query. Our approach also estimates the significance of observed similarities from the background data, similar to how it is done in sequence database searches.

Software availability: psi-square code and formatting utilities are at http://research.stowers-institute.org/bioinfo

Acknowledgment. We are grateful to Dennis Shasha for the modified version of the TTG program.

References

1. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat Genet* **22,** 281-5.
2. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc Natl Acad Sci U S A* **95,** 14863-8.
3. Ben-Dor, A., Shamir, R. & Yakhini, Z. (1999) *J Comput Biol* **6,** 281-97.
4. Toronen, P., Kolehmainen, M., Wong, G. & Castren, E. (1999) *FEBS Lett* **451,** 142-6.
5. Varma, S. & Simon, R. (2004) *BMC Bioinformatics* **5,** 126.
6. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002) *Science* **298,** 824-7.
7. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. (2004) *Science* **303,** 1538-42.
8. Bader, G. D. & Hogue, C. W. (2003) *BMC Bioinformatics* **4,** 2.
9. King, A. D., Przulj, N. & Jurisica, I. (2004) *Bioinformatics*.
10. Ramani, A. K., Bunescu, R. C., Mooney, R. J. & Marcotte, E. M. (2005) *Genome Biol* **6,** R40.
11. Bergmann, S., Ihmels, J. & Barkai, N. (2004) *PLoS Biol* **2,** E9.
12. Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. & Pavlidis, P. (2004) *Genome Res* **14,** 1085-94.
13. Brun, C., Herrmann, C. & Guenoche, A. (2004) *BMC Bioinformatics* **5,** 95.
14. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. (2003) *Science* **302,** 249-55.
15. DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997) *Science* **278,** 680-6.
16. Bader, G. D. & Hogue, C. W. (2002) *Nat Biotechnol* **20,** 991-7.
17. Krause, R., von Mering, C. & Bork, P. (2003) *Bioinformatics* **19,** 1901-8.
18. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999) *Proc. Natl. Acad. Sci USA* **96,** 4285-4288.
19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J Mol Biol* **215,** 403-10.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res* **25,** 3389-402.
21. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997) *Science* **278,** 631-637.
22. Tatusov, R. L., Fedorova, N. D., Jackson, J. J., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., *et al.* (2003) *BMC Bioinformatics* **4,** 41.
23. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc Natl Acad Sci U S A* **84,** 4355-8.
24. Altschul, S. F. (1991) *J Mol Biol* **219,** 555-65.
25. Pfahringer, B. (1995) in *Proceedings of the 12th International Conference on Machine Learning.*, pp. 456-463.
26. Catlett, J. (1991) in *Proceedings of the European working session on learning on Machine learning*, pp. 164-178.

27. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., *et al.* (2002) *Nature* **415,** 141-147.
28. Bozdech, Z., Llinas, M., Pulliam, B. L., Wong, E. D., Zhu, J. & DeRisi, J. L. (2003) *PLoS Biol* **1,** E5.
29. Bowers, P. M., Cokus, S. J., Eisenberg, D. & Yeates, T. O. (2004) *Science* **306,** 2246-9.
30. Glazko, G. V. & Mushegian, A. R. (2004) *Genome Biol* **5,** R32.
31. Young, G. M., Schmiel, D. H. & Miller, V. L. (1999) *Proc Natl Acad Sci U S A* **96,** 6456-61.
32. Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B., Sebaihia, M., James, K. D., Churcher, C., Mungall, K. L., *et al.* (2001) *Nature* **413,** 523-7.
33. Levesque, M., Shasha, D., Kim, W., Surette, M. G. & Benfey, P. N. (2003) *Curr Biol* **13,** 129-33.
34. Jim, K., Parmar, K., Singh, M. & Tavazoie, S. (2004) *Genome Res* **14,** 109-15.
35. Motaleb, M. A., Corum, L., Bono, J. L., Elias, A. F., Rosa, P., Samuels, D. S. & Charon, N. W. (2000) *Proc Natl Acad Sci U S A* **97,** 10899-904.
36. Macnab, R. M. (2004) *Biochim Biophys Acta* **1694,** 207-17.
37. Good, M. F., Kaslow, D. C. & Miller, L. H. (1998) *Annu Rev Immunol* **16,** 57-87.
38. Good, M. F. (2001) *Nat Rev Immunol* **1,** 117-25.
39. Florens, L., Liu, X., Wang, Y., Yang, S., Schwartz, O., Peglar, M., Carucci, D. J., Yates, J. R., 3rd & Wub, Y. (2004) *Mol Biochem Parasitol* **135,** 1-11.
40. Brossier, F., Jewett, T. J., Sibley, L. D. & Urban, S. (2005) *Proc Natl Acad Sci U S A* **102,** 4146-51.
41. Alberts, B. (1998) *Cell* **92,** 291-4.
42. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000) *Nature* **403,** 623-627.
43. Proudfoot, N. & O'Sullivan, J. (2002) *Curr Biol* **12,** R855-7.
44. Mangus, D. A., Smith, M. M., McSweeney, J. M. & Jacobson, A. (2004) *Mol Cell Biol* **24,** 4196-206.
45. Beggs, J. D. & Tollervey, D. (2005) *Nat Rev Mol Cell Biol* **6,** 423-9.
46. Hunter, L., Taylor, R. C., Leach, S. M. & Simon, R. (2001) *Bioinformatics* **17 Suppl 1,** S115-22.
47. Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. & Gerstein, M. (2001) *J Mol Biol* **314,** 1053-66.
48. Zhou, X., Kao, M. C. & Wong, W. H. (2002) *Proc Natl Acad Sci U S A* **99,** 12783-8.
49. Balasubramaniyan, R., Hullermeier, E., Weskamp, N. & Kamper, J. (2005) *Bioinformatics* **21,** 1069-77.
50. Glazko, G., Gordon, A. & Mushegian, A. (2005) *Bioniformatics* (accepted).

FIGURE LEGENDS

Figure 2. Phyletic vectors and COGs associated with flagella phenotype, identified by psi-square and TTG algorithms (45 COGs and 49 COGs, respectively), with COG1298 used as a query. a) 27 COGs in benchmark (see text), also found by psi-square and TTG; b) 5 COGs found by psi-square and TTG; c) 2 COGs found by psi-square and in benchmark and one COG found by TTG only; d) 8 COGs found in benchmark only; e) 11 COGs found by psi-square only. COG numbers and functional annotations are shown in the right-hand column.

Figure 5. Expression vectors for the closest matches retrieved by psi-square with query PFA0110w in Plasmodium IDC dataset. Two best matches per iteration (nine iterations before convergence) are shown.

Table 1. Sensitivity and specificity of psi-square and TTG algorithms in prediction of flagellae components.

	Psi-square: single query	Psi-square: combined query	TTG
False Positives (FP)	11	34	1
True Positives (TP)	29	34	27
False Negatives (FN)	8	3	10
SPECIFICITY[a]	**0.725**	**0.500**	**0.964**
SENSITIVITY	**0.784**	**0.919**	**0.730**
Number of predicted proteins:	45	73	33

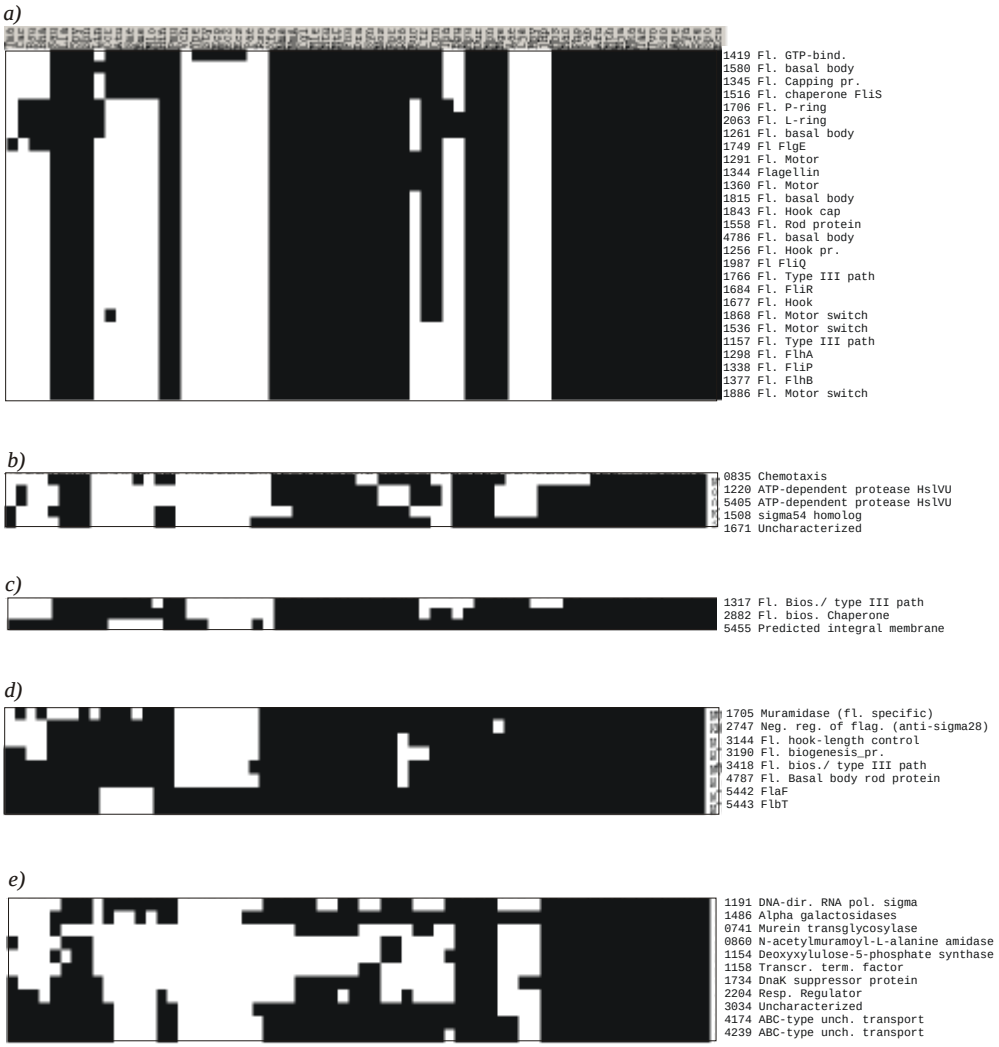
a)
1419 Fl. GTP-bind.
1580 Fl. basal body
1345 Fl. Capping pr.
1516 Fl. chaperone FliS
1706 Fl. P-ring
2063 Fl. L-ring
1261 Fl. basal body
1749 Fl FlgE
1291 Fl. Motor
1344 Flagellin
1360 Fl. Motor
1815 Fl. basal body
1843 Fl. Hook cap
1558 Fl. Rod protein
4786 Fl. basal body
1256 Fl. Hook pr.
1987 Fl FliQ
1766 Fl. Type III path
1684 Fl. FliR
1677 Fl. Hook
1868 Fl. Motor switch
1536 Fl. Motor switch
1157 Fl. Type III path
1298 Fl. FlhA
1338 Fl. FliP
1377 Fl. FlhB
1886 Fl. Motor switch
b)
0835 Chemotaxis
1220 ATP-dependent protease HslVU
5405 ATP-dependent protease HslVU
1508 sigma54 homolog
1671 Uncharacterized
c)
1317 Fl. Bios./ type III path
2882 Fl. bios. Chaperone
5455 Predicted integral membrane
d)
1705 Muramidase (fl. specific)
2747 Neg. reg. of flag. (anti-sigma28)
3144 Fl. hook-length control
3190 Fl. biogenesis_pr.
3418 Fl. bios./ type III path
4787 Fl. Basal body rod protein
5442 FlaF
5443 FlbT
e)
1191 DNA-dir. RNA pol. sigma
1486 Alpha galactosidases
0741 Murein transglycosylase
0860 N-acetylmuramoyl-L-alanine amidase
1154 Deoxyxylulose-5-phosphate synthase
1158 Transcr. term. factor
1734 DnaK suppressor protein
2204 Resp. Regulator
3034 Uncharacterized
4174 ABC-type unch. transport
4239 ABC-type unch. transport

Figure 2

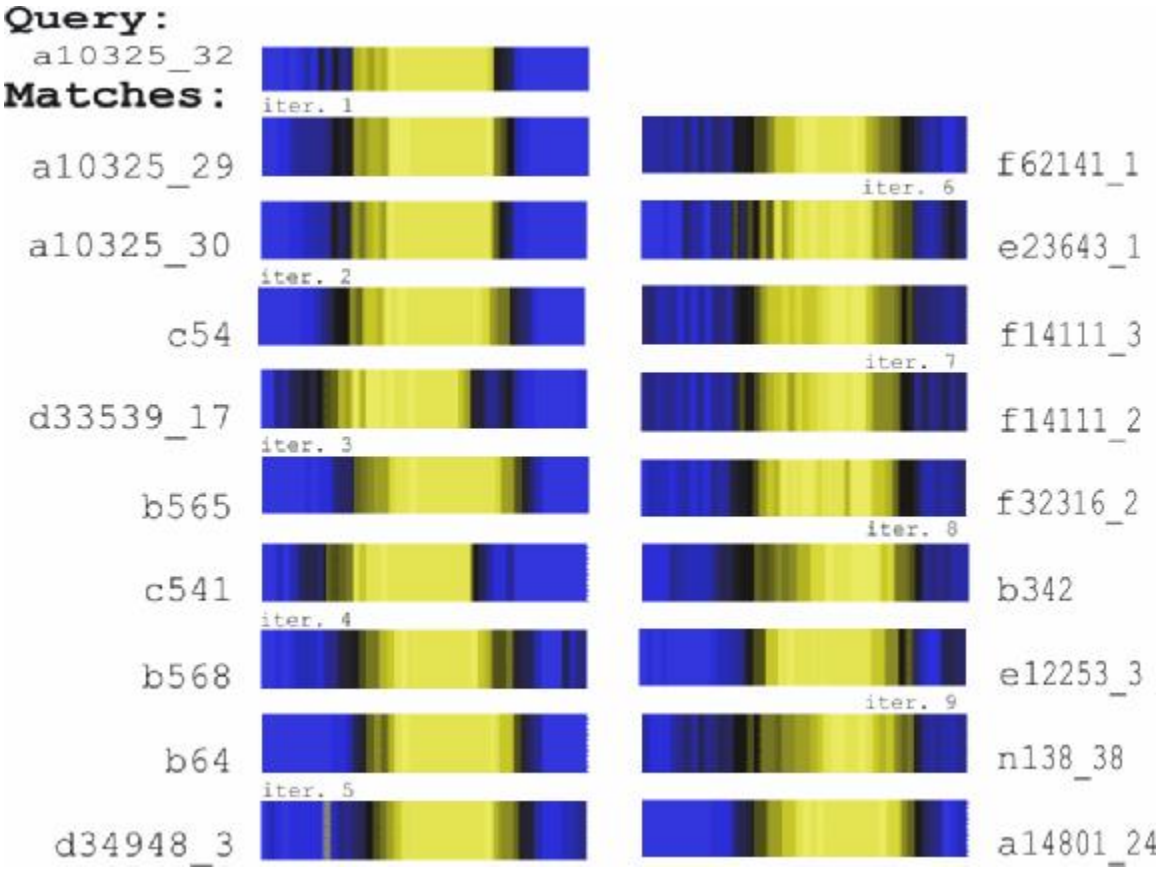
Query:
a10325_32
Matches:
iter. 1
a10325_29
a10325_30
iter. 2
c54
d33539_17
iter. 3
b565
c541
iter. 4
b568
b64
iter. 5
d34948_3
iter. 6
f62141_1
e23643_1
f14111_3
iter. 7
f14111_2
f32316_2
iter. 8
b342
e12253_3
iter. 9
n138_38
a14801_24

Figure 5.