Statistically Rigorous Automated Protein Annotation

Werner G. Krebs and Philip E. Bourne[1,]*

San Diego Supercomputer Center
and
[1]Department of Pharmacology

University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0505, USA

*To whom correspondence should be addressed.
Email: bourne@sdsc.edu Fax: +1 858-354-8301 Voice: +1 858-534-8301

Revision Friday, October 31, 2003

Abstract

Motivation: Assignment of putative protein functional annotation by comparative analysis using pre-defined experimental annotations is performed routinely by molecular biologists. The number and statistical significance of these assignments remains a challenge in this era of high-throughput proteomics. A combined statistical method that enables robust, automated protein annotation by reliably expanding existing annotation sets is described. An existing clustering scheme, based on relevant experimental information (e.g., sequence identity, keywords, or gene expression data) is required. The method assigns new proteins to these clusters with a measure of reliability. It can also provide human reviewers with a reliability score for both new and previously classified proteins.

Results: A dataset of 27,000 annotated PDB polypeptide chains (of 36,000 chains currently in the PDB) was generated from 23,000 chains classified *a priori*.

Availability: PDB annotations and sample software implementation are freely accessible on the Web at http://pmr.sdsc.edu/go.

Contact: bourne@sdsc.edu

Introduction

Putative annotation of protein function that is both broad and accurate is a major problem to which bioinformatics can contribute (Guigo, 1997; Rust *et al.*, 2002). The genome projects and subsequent gene finding techniques provide us with a wealth of functionally uncharacterized proteins. Many methods for automatic functional assignment exist (Bazzan *et al.*, 2002; Biswas *et al.*, 2002; Blaschke and Valencia, 2002; Cohen *et al.*, 2000; Eisenhaber and Bork, 1999; Kaplan *et al.*, 2003; Kasukawa *et al.*, 2003; Kretschmann *et al.*, 2001; Sasson *et al.*, 2003; Sonnhammer *et al.*, 1997; Tavazoie *et al.*, 1999). Functions should be stated in terms of a consistent nomenclature as, say, offered by the Gene Ontology (GO; Ashburner *et al*., 2000), the MIPS ontology (Mewes *et al.*, 1999), or the Enzyme Commission numbers (IUBMB, 1999; Webb and IUBMB, 1992) and be based on a good statistical foundation. In turn, clusters of proteins not statistically related to terms from an existing nomenclature should suggest that new nomenclature needs to be defined.

Although a large number of sequences have been already annotated within frameworks such as GO, many important sequences remain unannotated. Our goal here was to use the available annotations and propagate them to even more sequences using statistical techniques and consistency checks. We describe a statistical method that reliably assigns annotations to clusters of proteins by applying rigorous statistical tests with guaranteed reliability to discover interrelationships between the clusters and an existing set of annotations. The method can be used to validate and significantly enhance annotations derived from other automatic methods, such as those based on text mining (Bazzan *et al.*, 2002; Biswas *et al.*, 2002; Blaschke and Valencia, 2002; Kaplan *et al.*, 2003; Rust *et al.*, 2002). Typically, these methods use some form of data-mining rule (e.g., linguistic analysis or correlation based on sequence motifs) to associate sequences with GO terms. The method is described here is different because it takes, as input, sequences previously annotated with GO terms, and propagates these to additional sequences by examining the statistical interrelationships between the prior annotations and clusters generated from relevant external data, such as sequence identity. Our method is therefore complimentary

to previously published studies employing other methods of automatic annotation. The initial application of the method, described here, was to reliably assign Gene Ontology (GO) terms to PDB data (Berman *et al.*, 2000) using sequence-based clusters to a depth and reliability not previously achieved.

System and Methods

Statistical Methods

Statistical Selection Criteria

Our method proposes and then examines putative new GO annotations by applying rigorous statistical criteria on a term-by-term basis. There is no need statistically to distinguish between biological processes, biological functions, and biological location as assigned by GO. It is true that sequence will be most closely related to biological function, however, the method compensates for any reduction in statistical correlation between sequence and the other two ontology types. In most implementations of the GO database itself, the three different ontology types are distinguished by an internal flag and otherwise treated in an identical fashion. Our term-by-term methodology likewise handles all three ontological types in a single pass and does not by itself distinguish between them.

GO is organized as a Directed Acyclic Graph (DAG) or network-like structure; this could conceivable cause problems for automatic annotation schemes. A benefit to the term-by-term processing approach is that it is not impacted by the DAG structure of GO. If there is sufficient statistical evidence for the method to infer only a leaf annotation it is left to the end-user to infer any annotations implied by the DAG. The method generates annotation at all depths in the GO tree and is untroubled by the fact that most proteins have annotations at multiple depths in the GO tree. Indeed, the method excels in this type of environment. For example, it will sometimes correctly conclude, based on the statistical evidence from its sequence neighbors, that a particular protein is a “membrane protein”, but find that there is not evidence to discern its exact function.

Hypergeometric p-value criterion

In statistics, the p-value is the probability that an associated null hypothesis is true given a particular set of observations. Typically, this is the probability that a particular set of observations can be explained entirely by chance. A cut-off is normally set below which the p-value indicates that the null-hypothesis is false and it is accepted that the observations cannot be explained by chance alone.

Assume a particular proteome has g annotated proteins of which f have the specific classification of interest (e.g., belong to a specific GO term). Then the probability that a randomly selected protein has that classification is $p = f/g$.

If a particular cluster has n classified proteins, of which k have the classification of interest, we wish to determine the probability of observing k or more random events of probability p from a set of n. Thus, intuitively, the p-value may be computed from the cumulative binominal distribution:

$$P(n \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

However, the binominal distribution computes the probability of i draws from a bin of size N allowing for repeat draws. A protein in our clustering experiment may only be assigned a particular classification once, i.e., repeat draws of the same classification are not allowed. Consequently, we use a close relative of the binominal distribution, namely, the cumulative hypergeometric distribution. We have developed an efficient Perl language implementation that uses a number of optimizations to rapidly compute the cumulative hypergeometric distribution even for relatively large values of the hypergeometric distribution:

$$P(n \geq k) = \sum_{i=k}^{n} \binom{f}{i} \binom{g-f}{n-i} \Big/ \binom{g}{n}$$

If a particular cluster and classification combination pass the p-value criterion, this indicates we accept statistically that the number of observed occurrences of the

classification in the cluster cannot be explained by chance, i.e., the cluster is statistically biased towards the classification.

For sequence-based clustering we required the p-value to be less than 0.005 before we accepted that a cluster was selectively biased towards a particular classification. A significantly smaller p-value criterion would be required for clusters based on less reliable data, such as keyword or gene expression data. Because the relationship between sequence and function is so well established empirically, the null hypothesis is implicitly false, and less statistical evidence is required to establish selective bias than with other clustered data. The value of 0.005 was found by manual examination of borderline classification outcomes on a large test dataset; values for other types of clustered data can be determined by similar mechanisms. It is suggested that different p-value criteria be used depending on the reliability of the clustered data. This can be done within the same run as a result of using hierarchical clusters.

Bayesian criterion

If the p-value criterion is passed, we apply a second statistical test. We compute the conditional Bayesian probability that a particular polypeptide chain has the specific classification given that it is a member of the cluster in question as simply k/n. We require this to be larger than some suitable cut-off condition. In our study we required the Bayesian probability to be 95% before we assigned the classification to polypeptide chains in the cluster. The Bayesian probability is the "best-guess" or expected value of the classical probability based on current information and can therefore be computed and used even on very small sample sizes. Our Bayesian criterion effectively requires the "average" or expected error on new annotations to be five percent or less.

Confidence criterion

We reinterpreted the Bayesian conditional probability previously described as a classical sampling proportion. In this treatment, the n previously classified genes in the cluster are treated as a random sample of n genes in a larger population of N genes (where N is the total size of the cluster, including both classified and unclassified genes).

We can extrapolate the results of the random sampling to the entire population by expecting that:

$$\frac{k}{n} \pm 1.96 \sqrt{\frac{\left(n/N\right)\left(1 - n/N\right)}{n}}$$

is the fraction of N that can be expected with 95% confidence or above to have the classification. We can interpret this interval as a 95% confidence interval containing the "true" or classical probability that the putative classification in this cluster is indeed true. We can combine this with our previous criteria by requiring the lower bound on this confidence interval to be above some reasonable value; we use 80%. (An optional implementation is to compute the maximum of either the above estimated lower bound, or k/N, which is the theoretical lower bound on the classical probability.)

By requiring this lower bound to be 80% or greater, we effectively guarantee that the maximum error rate on any new annotations can be no more than 20%. The application of the Bayesian criterion from the previous section further ensures that the expected, or average, error rate in new annotations will be below 5% assuming the seed dataset is 100% accurate. The true error rate is difficult to estimate but will be considerably below either of these values (likely close to the error rate in the seed dataset), because we have previously required that the clustering method selectively biases the cluster towards the particular classification to very high statistical confidence.

By combining these three simple statistical tests we have created a conservative virtual statistical test that can be argued as appropriate for our application on theoretical grounds, yet is computational tractable and produces intuitively meaningful statistics. Manual examination of the PDB GO assignments empirically supports these statements. Refer to http://pmr.sdsc.edu/go for a demonstration interface and computed values on specific PDB identifiers.

Hierarchical clustering of genes

For the purposes of our experiment we clustered all 36,000 polypeptide chains in the PDB into a hierarchy of five sets of sequence-based clusters using cd-hit (Li *et al.*, 2001) with sequence identity thresholds set at 95%, 90%, 70%, 50% and 40%,

respectively. Protein chains with sequences of less than 40 amino acids in length were removed from the clustered sets.

The statistical approach applied here may be applied to clusters of differing levels of reliability, such as protein classifications based on keywords and expression data. Indeed, we show here, using a hierarchy of clusters of decreasing experimental reliability, that different clustering schemes with varying levels of reliability (equivalent to different experimental methods) may be robustly combined within our method with only variation in choice of threshold parameters. Principally, a more stringent p-value criterion threshold should be used when our method is applied to clustering schemes based on data derived from experimental techniques whose relation to the classifications in question is known to be less reliable.

Hierarchical processing

In order to classify an unknown polypeptide chain, each set of clusters is examined sequentially in order of reliability to see if clusters exist in which the unknown protein has classified neighbors. The polypeptide chain is rejected from the result pool if neighbors were found but the polypeptide chain, cluster, and classification did not pass the statistical criteria outlined above. For example, if the polypeptide chain had classified neighbors at the 95% sequence identity level, but these did not pass the statistical significance tests, the polypeptide chain was not considered further. An alternative approach would use clusters at different levels of sequence identity. For example, the polypeptide chain that fails the statistical test at 95% identity could be re-examined at 90% or 70% sequence identity to see if the classification that can be deduced from these clusters pass the statistical test. In our experience, if the gene had neighbors at 95% sequence identity, but the cluster did not pass statistical tests, examination of the larger 90% or 70% clusters did not improve matters appreciably, but with cluster hierarchies that use different experimental techniques this refinement of perceived cluster reliability may be desirable.

If a polypeptide chain was found to belong to multiple clusters within the same experimental set, the clusters were combined into a super-cluster with duplicates eliminated. This super-cluster was then processed as described above.

Inputs and Output

The method's input consists of available GO annotations for a set of 'seed' sequences (this section) as well as new sequences that we desire to annotate. The output of the method consists of annotations for some of these new sequences (Results section below).

The public Compugen dataset (Xie *et al.*, 2002) was used to map Gene Ontology (GO) terms (Ashburner *et al.*, 2000) to SwissProt (Bairoch and Apweiler, 2000) identifiers (Table 1). At the time of this study SwissProt identifiers mapped to PDB (Berman *et al.*, 2000) identifiers, representing complete structures, but not to individual PDB polypeptide chains. A PDB chain corresponds to a protein sequence present in a protein structure, whereas a PDB identifier maps to a solved structure that may refer to several polypeptide chains and other components. We used additional information from the PDB, namely the DBREF records mapping PDB identifiers to predominantly GenBank identifiers, which were in turn mapped by SwissProt directly. Additionally, FASTA searches were used as necessary to determine the PDB chain identifiers corresponding to each SwissProt identifier. The software that determines the mapping between SwissProt and PDB chains is available by contacting the authors.

Previously, Compugen, Inc., published a dataset (Xie *et al.*, 2002) that used text mining as well as other techniques to map SwissProt identifiers to GO accession numbers. By merging the public Compugen dataset (available from the GO website – http://www.geneontology.org/cgi-bin/GO/downloadGOGA.pl/gene_association.compugen.Swissprot) and our own SwissProt to PDB chain mappings, we were able to produce a dataset mapping approximately 23,000 PDB chains to GO terms. To verify the mappings were accurate and statistically meaningful, we manually examined a random subset of 460 chains and were unable to find any errors, suggesting (via the classical sampling formula) that less than one percent of the chains in this dataset contain errors, consistent with Xie et al.'s claims (Xie *et al.*, 2002). In addition, we also applied our statistical methods (described below) to already classified PDB chains; the resulting p-value and Bayesian probability

values served as a convenient error check for both the Compugen mapping and the GO assignments themselves.

Results

Using the hierarchical clustering and statistical significance tests described above we were able to extend the Gene Ontology (GO) classification to cover 27,226 protein sequences, out of approximately 36,000, represented by individual polypeptide chains in the PDB. This added approximately 4,000 classified protein sequences to our seed database of approximately 23,000 protein sequences obtained by combining the Compugen dataset with our SwissProt to PDB chain mapping dataset. Manual examination of this dataset, including special examination of borderline classifications with p-value and Bayesian values just inside or outside the acceptable region, revealed empirically that the resulting automatically generated dataset was of high quality.

Table 2 provides the number of unique combinations of new GO term accession numbers and PDB chains found using this method. The Bayesian statistics represent the simple *k/n* test. A feature of the combined method is that it has roughly similar attenuation statistics, as does the more traditional lone p-value test (column 2 versus columns 3 and 4 of Table 2). The difference between a p-value of 0.001 and 0.1 without the Bayesian test is approximately 3,000 GO term/PDB chain combinations, whereas keeping the p-value threshold constant but altering the Bayesian cut-off from 0% through 95% increases the attenuation by between 1,000 and 2,000 GO term/PDB chain combinations in this sample, depending on the initial p-value threshold selected for the test. Note that the test with a p-value threshold of 0.001 alone produces approximately the same number of GO term/PDB chain combinations as the more complicated combination test combining a less stringent p-value threshold of 0.005 together with a second test requiring the Bayesian probability to exceed 95%. As stated in the Materials and Methods section, on theoretical grounds the combination of the p-value test with the Bayesian criterion is a better test in the sense that the results will have fewer false positives (i.e., a more “correct” results set) than the simple p-value test alone with more stringent cut-off. Table 2 provides empirical evidence to support this conjecture. The p-value test alone at

threshold 0.001 would have accepted some additional 814 GO terms (12,678-11,834) that had 20% or more of the annotated PDB chains in the same cluster lacking that annotation in the preexisting seed dataset—effectively all false positives if one assumes our preexisting GO annotations are 100% complete and accurate. The same data are illustrated in Figure 1. The surface is actually a smooth curve but appears segmented due to the linear interpolation. Nevertheless, it can be seen that the surface is more sloped in the Y-direction at lower (less stringent) p-value thresholds than at higher p-value thresholds.

Since GO annotations are not 100% complete and accurate we developed a database and associated Web front-end to manually examine individual cases. The web tool makes available the statistical scores providing the evidence for these automatic classifications. It uses p-values to sort these classifications, a feature we find very helpful in providing insights into the underlying data. The Web site is accessible at http://pmr.sdsc.edu/go.

This method of statistical based clustering can be used to flag existing database classifications for probable errors by treating a previously classified protein as unclassified and computing the p-value and Bayesian probability. Existing classification/protein combinations with high p-values or low Bayesian probability can be flagged for further examination. We were able to use this method to find errors in existing databases.

Although we only used sequence-based clustering in the generation of the dataset in question, we also experimented with gene annotation and keyword-based clustering of the data. We found that by using clustering based on experimental techniques known *a priori* to be strongly predictive of function, larger p-values may be tolerated, because the null hypothesis is known *a priori* to be false due to non-statistical, experimental considerations and thus not as much statistical confidence is needed. The appropriate p-value cut-off to use can be estimated empirically for a particular experimental technique by examining the decisions made by the statistical algorithm—in particular borderline outcomes—on a test dataset that has already been classified using other techniques. Our test is commutative, so the Bayesian threshold (with or without the sampling correction) can be applied first for a fixed percentage threshold (95% without sampling correction or

80% with sampling correction seems to work well with GO) and p-values can then be examined for those chain and annotation combinations that pass the Bayesian test. It was found that large p-values could safely be tolerated for experimental techniques known to be highly predictive of function—95% sequence identity clustering, for example—and smaller p-value thresholds were needed for less predictive experimental techniques, such as functional clustering based on protein expression data. This is consistent with the body of scientific literature on protein expression analysis, which typically uses the hypergeometric test in conjunction with much smaller p-values (Tavazoie *et al.*, 1999). For example, we undertook simple k-means clustering to generate ten large, disjoint clusters using S-plus software (Krause and Olson, 2000) on unnormalized Affymetrix yeast data (Wodicka *et al.*, 1997). We confirmed that the much smaller p-values used for the MIPS ontology (Tavazoie *et al.*, 1999) are also appropriate for GO and our methodology applied to nosier data, such as that typically associated with unnormalized protein expression data, is used as input. Normalization and more sophisticated techniques for clustering gene expression data may reduce this need. Based on an examination of clusters derived from individual SwissProt keywords, we anticipate data quality intermediate between 40% sequence homology clusters and protein expression data. In general, higher quality experimental clusters require less stringent statistical filters, and produce better results with our methodology.

To provide further empirical evidence that the Bayesian percentage adds additional value beyond the p-value criterion, we computed the least-squares linear regression correlation between the p-value and Bayesian percentage in a variety of ways. The three criteria we have presented are all statistical significant on similar data, and therefore some correlation can be expected and is present (Table 2). However, in spite of the expected correlation, straight linear regression of either the Bayesian statistic or the lower bound on the classical percentage confidence interval statistic produces extremely poor Pearson correlation statistics with R-squared values less than 0.10. In order to improve the fit we attempted to massage the data using standard statistical techniques, including exclusion of outliers. We attempted to correlate the reciprocal of p-values greater than 10^{-100} with both the Bayesian and lower bound statistics with similar results. We had the greatest success correlating the log of the p-values with the two percentages.

However, the fit is extremely poor at the end of the ranges when the Bayesian or lower-bound statistics reach 100%. Consequently, we considered only rejected lower-bound statistics (e.g., those observations where the lower bound was less than 80%) and only Bayesian statistics where the Bayesian statistic was less than 100%, which improved the r-squared factor. However, the correlations remained poor despite massaging of the variables and elimination of outliers (Table 3). The best correlation was between the log of the Bayesian statistic (excluding values at 100%) and p-values (excluding p-values greater than 1e-100) which had an R-squared of 0.347; this is not surprising given that the statistics share the same set of variables (*n*,*k* for Bayesian and *n*,*k*,*f*,*g* for the p-value statistic). Both the p-value and Bayesian statistic provide intuitively useful information, and the low correlation demonstrates that the value of the p-value statistic provides little new information about the Bayesian statistic, hence both should be computed and used. The lower bound on the classical percentage confidence interval statistic involves an important additional variable *N* not available to the p-value statistic, and therefore it is not surprising that it is even less correlated than the Bayesian statistic. Due to the poor correlation between p-values and these error estimates, these error estimates cannot be safely inferred from the p-value alone but need to be explicitly computed.

Discussion

Recent approaches to automatic Gene Ontology (GO) annotation have sought to utilize additional sources of information such as sequence and keyword information to extend classifications in the GO and other databases in a process of putative automated protein annotation. Here we use a combined statistically sound metric to distinguish between reasonable classifications and more questionable inferences.

The combined statistical methods introduced here were originally motivated by considerations of experimental data less reliably indicative of protein function, such as gene expression data. (Tavazoie *et al.*, 1999; Cohen *et al.*, 2000) introduced a version of the technique presented here to compute the p-values of clusters of genes in relation to MIPS terms (Mewes *et al.*, 1999). They do not compute the Bayesian probabilities or the classical confidence limits proposed here.

Figure 2 illustrates what the methodology proposed here sets out to achieve using a hierarchy of clusters. A given set of annotations (blue circle) typically represents a subset of a given ontology (green dashed circle). The goal is to add data for which there is no additional reliable experimental evidence. Such data are added only if it can be shown to be statistically meaningful within various stringency criteria. This creates a hierarchy of clusters at different p-value thresholds with a different number of members. From this hierarchy of clusters the method selects only the most reliable data available to make statistical inferences. The method may examine less reliable experimental data (e.g., clusters at lower sequence identity or data from keyword or protein expression data) when the more reliable experimental data is insufficient to assign a classification in a statistically robust way. The use of a hierarchy of clusters allows the application of multiple p-value cut-off criteria. Experimental techniques for which there is *a priori* knowledge that clusters are closely related require less statistical evidence to disprove the null hypothesis and may be allowed a larger p-value cut-off criteria then less reliable experimental techniques, such as protein expression data (Figure 2).

A further improvement comes from the use of one or both of our probability criteria, which ensures that only statistically likely classifications are accepted. A first criterion requires the Bayesian estimate of the probability that a classification is correct be above a certain threshold. A second, related criterion requires that a lower bound on an estimate of the classical probability be above a certain threshold. These criteria are only weakly correlated with one another (Table 3) and we recommend both be employed.

Of the three criteria, the p-value statistics are sensitive to the largest number of variables, and are able to discern small differences in statistical reliability. Consequently, we found it useful to use p-values to sort inferred classifications in our web tool. However, we find that all three statistics are useful, and find empirically that at least two of them (p-values and Bayesian probabilities) must be used together to achieve the best accuracy.

The method proposed here is general and can be used to assign putative functional annotation to unknown protein sequences. As such, the method could accelerate the manual curation of protein sequences and be used to seek out errors in already annotated collections.

Conclusion

We have applied hypergeometric p-values and Bayesian statistics to extend a seed dataset of Protein Data Bank (PDB) protein polypeptide chains to Gene Ontology (GO) mappings in a statistically rigorous fashion, resulting in a mapping of 27,226 PDB sequences to Gene Ontology terms. This represents an improvement of 4,000 sequence assignments over the starting seed assignment. Most recently we have received a SwissProt to GO mapping as a result of the InterPro project. On an on-going basis we will apply this initial mapping to PDB polypeptide chains identified as having SwissProt counterparts to provide a comprehensive GO mapping for sequences in the PDB. This will enable browsing of the PDB by molecular and biochemical function as well as cellular location. Initial access to these assignments can be found at http://pmr.sdsc.edu/go. However good functional assignments become, the method described here provides a useful, additional quality assurance check.

Acknowledgements

The automatic Gene Ontology classifier was supported by the U.S. National Science Foundation (NSF) National Partnership for Advanced Computational Infrastructure (NPACI) and the U.S NSF Division of Biological Infrastructure grant DBI 0111710. The authors would also like to thank Ilya Shindalov, Wolfgang Bluhm, Eliot Clingman, and David Stoner for useful discussions. We would also like to thank UCSD Technology Transfer for assistance with intellectual property matters related to this work.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* **25**, 25-9.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res,* **28**, 45-8.

Bazzan,A.L., Engel,P.M., Schroeder,L.F. and Da Silva,S.C. (2002) Automated annotation of keywords for proteins related to mycoplasmataceae using machine learning techniques. *Bioinformatics,* **18 Suppl 2**, S35-43.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.,* **28**, 235-242.

Biswas,M. *et al.* (2002) Applications of InterPro in protein annotation and genome analysis. *Brief Bioinform,* **3**, 285-95.

Blaschke,C. and Valencia,A. (2002) Automatic ontology construction from the literature. *Genome Inform Ser Workshop Genome Inform,* **13**, 201-13.

Cohen,B.A., Mitra,R.D., Hughes,J.D. and Church,G.M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet,* **26**, 183-6.

Eisenhaber,F. and Bork,P. (1999) Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics,* **15**, 528-35.

Gasteiger,E., Jung,E. and Bairoch,A. (2001) SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr Issues Mol Biol,* **3**, 47-55.

Guigo,R. (1997) Computational gene identification: an open problem. *Comput Chem,* **21**, 215-22.

IUBMB. (1999) Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999). *Eur J Biochem,* **264**, 610-50.

Kaplan,N., Vaaknin,A. and Linial,M. (2003) PANDORA: keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Res,* **31**, 5617-26.

Kasukawa,T. *et al.* (2003) Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Res,* **13**, 1542-51.

Krause,A. and Olson,M. (2000) *The basics of S and S-Plus*. 2nd edit. Statistics and computing, Springer, New York.

Kretschmann,E., Fleischmann,W. and Apweiler,R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics,* **17**, 920-6.

Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics,* **17**, 282-3.

Mewes,H.W. *et al.* (1999) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.,* **27**, 44-48.

Rust,A.G., Mongin,E. and Birney,E. (2002) Genome annotation techniques: new approaches and challenges. *Drug Discov Today,* **7**, S70-6.

Sasson,O. *et al.* (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res,* **31**, 348-52.

Sonnhammer,E., Eddy,S. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins,* **28**, 405-20.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat Genet,* **22**, 281-5.

Webb,E.C. and IUBMB. (1992) *Enzyme Nomenclature*, Academic Press, San Diego.

Wodicka,L. *et al.* (1997) Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat Biotechnol,* **15**, 1359-67.

Xie,H. *et al.* (2002) Large-scale protein annotation through gene ontology. *Genome Res,* **12**, 785-94.

Tables

Database	*URL*	*Description*
Protein Data Bank (PDB) (Berman *et al.*, 2000)	http://www.rcsb.org	Definitive database of protein structures. Source of some gene product identifiers (PDB DBREF records) linking PDB identifiers with GO gene product identifiers.
Gene Ontology (GO) (Ashburner *et al.*, 2000)	http://www.godatabase.org	Cellular location, molecular and biochemical function. Directly classifies about 1,200 PDB identifiers, another 1,200 via SwissProt, and another 19,000 chains via PDB DBREF records.
SwissProt (Bairoch and Apweiler, 2000; Gasteiger *et al.*, 2001)	http://www.expasy.ch/sprot/	Curated non-redundant protein sequences providing functional description, domains structure, post-translational modifications, variants, etc.
Compugen public SwissProt to GO mapping (Xie *et al.*, 2002)	http://www.geneontology.org/cgi-bin/GO/downloadGOGA.pl/gene_association.compugen.Swissprot	Maps GO accession numbers to SwissProt IDs.

Table 1

Description of the external data sources at the time of this work.

p-value Threshold	0% Bayesian Cut-off (p-value only)	80% Bayesian Cut-off	95% Bayesian Cut-off
0.1	15,915	14,581	14,108
0.05	15,679	14,371	13,898
0.01	14,481	13,453	12,980
0.005	14,215	13,233	**12,760**
0.001	**12,678**	11,834	11,361

Table 2

Comparison of new assignments for different values of p-value thresholds and Bayesian cut-off thresholds.

Regression			R-squared	Observations (after exclusion of outliers from 20,264 observations)
Y-variable	**Truncation Filter**	**Selection Filter**		
Log of lower bound of percentage confidence interval	p-value > 1e-100	Lower bound of percentage confidence interval < 80% rejects only	0.133	7902
Log of Bayesian probability	p-value > 1e-100	Lower bound of percentage confidence interval < 80%	0.030	7902
Log of Bayesian probability	p-value > 1e-100	Bayesian probability < 100%	0.347	1615

Table 3

Regression Statistics. The first three columns specify the transformation applied to the dataset prior to least-squares linear regression. The first column specifies computation of the transformed Y variable used in each regression, and requires taking the logarithm of one of the statistics used in the criteria. The second column specifies the first of two filters used to remove outliers from the transformed dataset in order to further improve the correlation statistic. The third column specifies the second of two filters used to remove outliers from the transformed dataset. The fourth column gives the square of the traditional Pearson regression correlation statistic, commonly denoted as R-squared. The final column describes the number of observations remaining in the transformed and filtered dataset used in the regression analysis.

Figures

Figure 1

3D surface graph of the linear interpolation of data in Table 2. The Z axis (height of the surface) represents the number of new GO term/PDB chain combinations, the Y-axis represents the Bayesian cut-off, and the X-axis represents the p-value threshold used in the test.

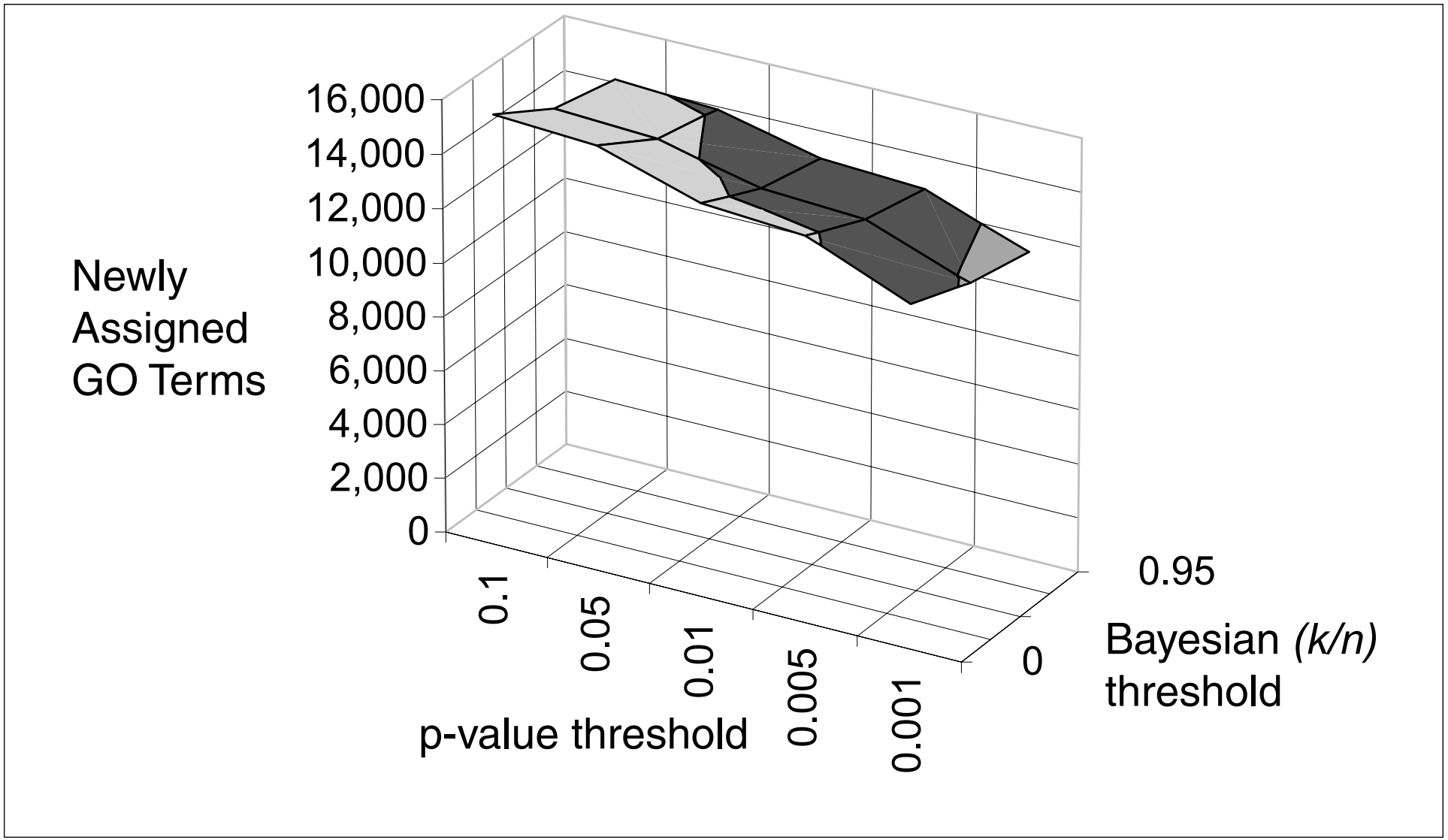
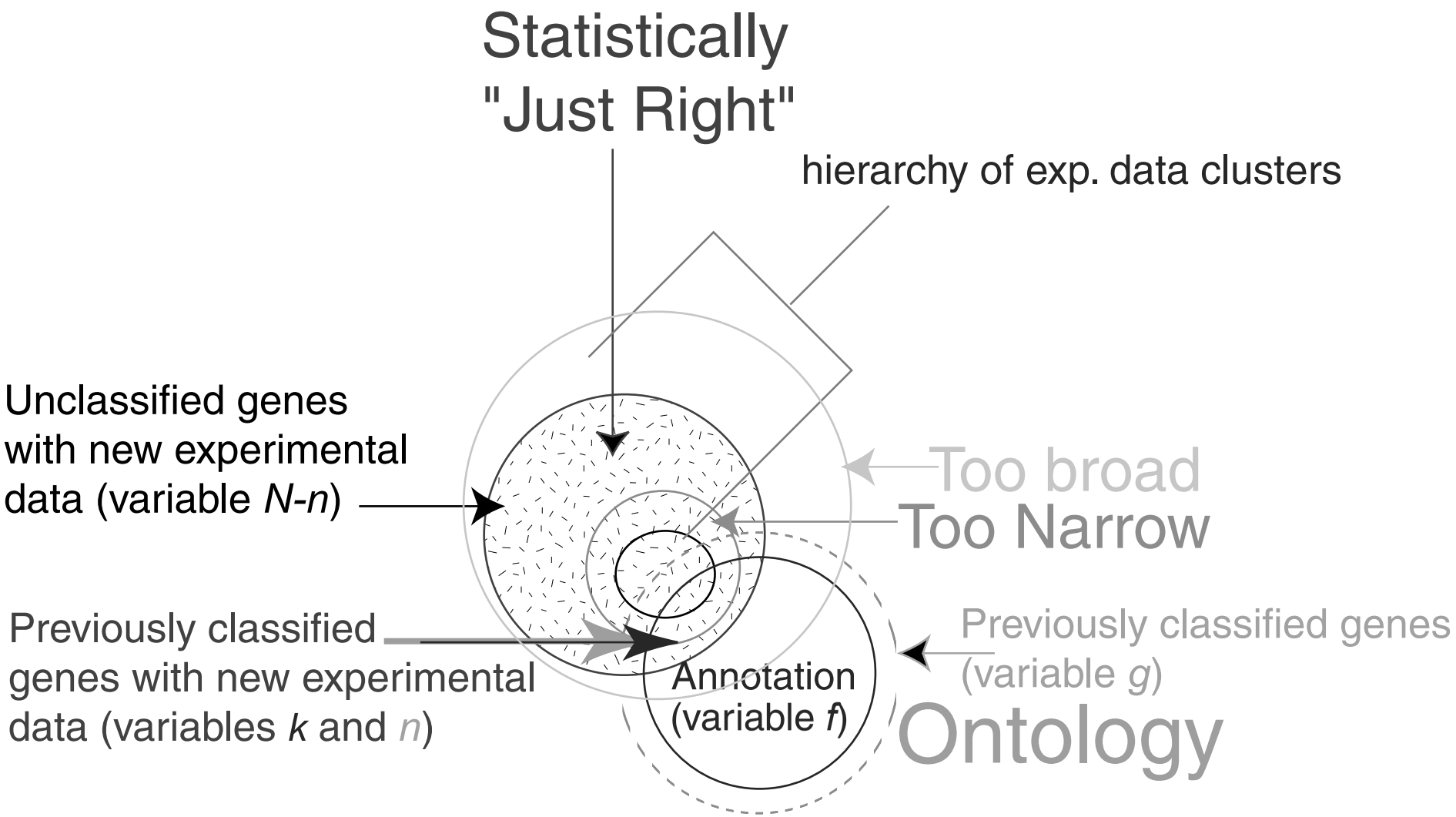

Newly
Assigned
GO Terms
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
0.1
0.05
0.01
0.005
0.001
p-value threshold
0.95
0
Bayesian (k/n)
threshold

Figure 2

Variables associated with the experimental data clusters.


Statistically
"Just Right"
hierarchy of exp. data clusters
Unclassified genes
with new experimental
data (variable N-n)
Too broad
Too Narrow
Previously classified
genes with new experimental
data (variables k and n)
Previously classified genes
(variable g)
Annotation
(variable f)
Ontology