Running Title: Mining the Structural Genomics Pipeline

Mining the Structural Genomics Pipeline: Identification of Protein Properties that Affect High-Throughput Experimental Analysis

Chern-Sing Goh[1,9], Ning Lan[1,9], Shawn Douglas[1,3,9], Baolin Wu[4], Nathaniel Echols[1,9], Andrew Smith[1,3,9], Duncan Milburn[1,9], Gaetano T. Montelione[6,7,8,9], Hongyu Zhao[4,5], and Mark Gerstein[1,3,9]*

[1] Molecular Biophysics and Biochemistry
[2] Molecular, Cellular, and Developmental Biology
[3] Computer Science
[4] Department of Epidemiology and Public Health
[5]Department of Genetics
Yale University, 266 Whitney Ave
New Haven, CT 06520

[6]Center for Advanced Biotechnology and Medicine,
[7]Dept. of Molecular Biology and Biochemistry, Rutgers University, and [8]Dept. of Biochemistry, Robert Wood Johnson Medical School, UMDNJ
Piscataway, NJ 08854

and

[9]Northeast Structural Genomics Consortium

corresponding author: mark.gerstein@yale.edu

Abstract

Structural genomics projects represent major undertakings that will change our understanding of proteins. They generate unique datasets that, for the first time, present a standardized view of proteins in terms of their physical and chemical properties. By analyzing these datasets here, we are able to discover correlations between a protein's characteristics and its progress through each stage of the structural genomics pipeline -- from cloning, expression, purification, and ultimately to structural determination. First, we use tree-based analyses (decision trees and random forest algorithms) to discover the most significant protein features that influence a protein's amenability to high-throughput experimentation. Based on this, we identify potential bottlenecks in various stages of the structural genomics process through specialized "pipeline schematics". We find that the properties of a protein that are most significant are: (i) whether it is conserved across many organisms, (ii) the percent composition of charged residues, (iii) the occurrence of hydrophobic patches, (iv) the number of binding partners it has, and (v) its length. Conversely, a number of other properties that one may have thought important such as nuclear localization signals did not turn out to be significant. Thus using our tree-based analyses, we are able to identify combinations of features that best differentiate the small group of proteins for which a structure has been determined from all the currently selected targets. This information may prove useful in optimizing high-throughput experimentation. Further information is available from http://mining.nesg.org/.

Keywords: **structural genomics, COGs, charged residues, hydrophobicity, decision trees**

Introduction

With the advent of the post-genomic era, the next challenge is to determine the structure of encoded proteins[1] which can lead to functional annotation of previously uncharacterized gene products [2-5]. The structural genomics effort has demonstrated the possibility of rapid structure determination on a genome-wide scale and is expected to generate a considerable amount of data. However, there are several challenges that can deter the process of proteins through the structural genomics pipeline[6-9] - from target cloning, expression, purification, to structural determination.

In addition to a growing collection of crystal and NMR structures, structural genomics is generating new and novel datasets where proteins are subject to uniform conditions for expression. Never before has it been possible to gain access to such a large amount of standardized experimental protein data, generated for thousands of targets from many organisms, at multiple sites over various structural genomics consortia. These data sets can be mined to look for correlations between a protein's properties and its performance in these standardized experiments. For instance, we might imagine that proteins that have more hydrophobic sequences might be harder to express or that proteins that interact with partner proteins might be less able to crystallize or fold correctly. These questions can be answered now through these new structural genomics datasets.

The SPINE database was created not only as an information repository for the Northeast Structural Genomics (NESG) Consortium, but also as a vehicle to integrate and manage data in a standardized fashion that makes it accessible to systematic data analysis[10,11]. Bertone *et al*[10] demonstrated the potential data mining capabilities of the SPINE database by developing a decision tree algorithm that was used to infer whether a protein was soluble from a dataset of 562

M. thermoautotrophicum protein expression constructs. Here we used information from all the targets from TargetDB (http://targetdb.pdb.org/), amounting to over 27,000 selected targets from over 120 organisms, to systematically correlate biophysical properties of proteins to their sequence features in order to determine their amenability to high-throughput experimentation. This work has three values. First of all, it utilizes a unique dataset generated under relatively uniform conditions. Second, it can tell us more about the properties of proteins in a systematic fashion and, thirdly, it can generate information needed to optimize protocols and conditions for effective high-throughput structural genomics.

Results and Discussion

Our overall approach to the data mining analysis is twofold. First, we employ two types of tree-based algorithms, random forest and decision tree analysis, to identify features most influential in determining whether a protein is amenable to high-throughput experimental analysis. Random forest analysis[12,13] is a robust algorithm particularly useful for calculating the importance of features by measuring the effect of permutations of each feature on prediction accuracy. It uses two techniques: bagging (bootstrap aggregating) and random feature selection. In combination, these methods have been shown to improve the stability and accuracy of prediction over a single tree model. While the random forest method is a robust technique for ranking features in terms of their importance, it is more difficult to interpret. In order to measure the frequencies of proteins containing certain features and understand how combinations of these protein properties can affect their amenability for experimental analysis, we use decision trees[14,15], a commonly used machine learning method. In general, we partition the initial sample, consisting of positives and negatives, into different subsets depending on a particular feature, such as amino acid

composition or protein binding partners. If the feature preferentially separates the positives and negatives, this is readily apparent and the most selective rules appear at the top of the decision trees[16]. We use these decision trees to identify and view features and combinations of features that are particularly selective. In a second type of analysis, which we call pipeline analysis, we diagram the way particular features change over the structural genomics pipeline and identify bottlenecks or stages in the pipeline where these features show the largest change.

Tree-Based Analysis

As of February 2003, sequence and experimental progress information for 27,267 protein targets were collected from the TargetDB and used in the tree analyses. We performed the tree analyses on all the targets found in the TargetDB in order to discover protein features that are the strongest predictors for whether a protein can be structurally determined. The protein properties used in the analysis are listed in Table 1. These properties comprise of general sequence composition, and other protein characteristics such as COG assignment, length of hydrophobic stretches, number of low complexity regions, and number of interaction partners.

Based on the current data in TargetDB, 1.3% (370/27711) of all targets are structurally determined. Results from the random forest (Table 2) and decision tree (Figure 1a) analyses suggest that protein properties such as COG[17] (clusters of orthologous groups) assignment; percent composition of charged, polar, and nonpolar residues; and length of the protein correlate with a tendency to be structurally determined.

Within the high-throughput structural genomics pipeline, there are many stages in the process that contribute to the attrition of proteins. Each step has its own selective conditions that affect whether a protein target advances to the next step. In order to identify protein

characteristics most influential in achieving the next level in the high-throughput experimental determination, random forest and decision tree analyses were run on proteins that were successfully cloned, expressed, or purified (Figures 1 and 2). The tree nodes in figures 1 and 2 represent the probability that proteins that satisfy the rule will be successful. The numbers to the right and left of the node are the numbers of proteins that are successful and those that aren't. The sum of the two numbers at each node is the total number of proteins that satisfy the rule or set of rules. To further aid experimental design, an additional decision tree was created (figure 2) using the same datasets as in figure 1 but without "meta-descriptors" such as COG analysis or binding partner information. The evolution of these features is traced through the pipeline figures (figure 3). Some protein features such as COG assignment and protein sequence length are found in both the pipeline analyses and in the overall structure decision tree. It is noted that at each stage of the protein determination pipeline, certain features appear to be more influential than others.

Tree-based analysis on targets that have been expressed suggests COG assignment, sequence length, and pI values are important features that affect the outcome. These results are reflected in the pipeline figure analysis (figure 3) where there are significant differences in these features between cloned but not expressed proteins and expressed proteins. Out of the 14,385 cloned protein targets in this analysis, 3764 have a COG assignment and a pI value below 5.9 (Figure 1c). The decision tree analysis suggests that cloned proteins meeting these criteria have a better chance (73%) of being expressed compared to all cloned targets (58%) that are expressed.

In contrast to these findings found for expressed proteins, purified proteins (Figure 1d) have different determining characteristics that include the percent composition of charged

residues such as aspartic acid and glutamic acid, and the percent sequence composition of asparagine and glutamine amino acids. This suggests that an optimal combination of aspartic acid, glutamic acid, asparagine, glutamine, and lysine sequence composition can increase the chance of expressed proteins to become purified from 46% to 77% (p-value = 4.1×10^{-50}).

Similarly, the tree analyses identify protein features such as methionine and alanine percent sequence composition that can affect the outcome of whether a purified protein becomes structurally determined (Figure 1e). The decision tree analysis also shows that proteins with very low methionine composition (less than 0.3%) and alanine percent composition less than 8.5% have a 67% chance of being crystallized (p-value = 3.4×10^{-8}).

Solubility

Since solubility is an important determinant for whether a protein is amenable to structural determination, an analysis was performed to find protein characteristics that influence the outcome of a protein's solubility. Serine percent composition is shown to be the major determinant in determining solubility. Other predictors of solubility such as conservation across organisms (COGs) and charged residue composition are similar to the other tree analyses performed on the various stages of the structure determination pipeline, confirming the significance of a protein's solubility in its amenability to high-throughput experimentation.

Analysis of Specific Structural Genomics Centers

Decision tree analysis was performed on six separate structural genomics centers: the Northeast Structural Genomics Consortium (NESG), the Joint Center for Structural Genomics (JCSG), the Mycobacterium tuberculosis Structural Genomics Consortium (TB), the Midwest

Center for Structural Genomics (MCSG), the Montreal-Kingston Bacterial Structural Genomics Initiative (BSGI), and the Berkeley Structural Genomics Center (BSGC). These groups each have their own separate initiatives with differing methods of target selection, cloning, and purification[8]. The decision trees in Figure 4 were performed to identify important protein properties that would influence a target's amenability to be structurally determined within each consortium. The resulting diverse trees illustrate the unique approach that each of the consortia has taken.

It is notable that more than half of the targets that are not structurally determined can be selected by the top three rules in each of the consortia decision trees. The results suggest that each consortium has its own methods of target selection, cloning, and protein production. The decision tree analysis is able to highlight patterns of successes for these consortia. For example, the NESG (Figure 4a) seems to be more successful with proteins that have more than 12% aspartic and glutamic acid composition and protein lengths of less than 112 amino acids. The NESG targets are comprised mostly of small (<340 aa) prokaryotic and eukaryotic proteins, and, generally, smaller proteins tend to have a higher concentration of charged residues than larger ones. Similarly, the BSGC (Figure 4f) seems to achieve better results for protein targets with protein lengths less than 346 amino acids and proteins that are conserved across organisms. Since the data collected from TargetDB for each consortium does not distinguish between targets that have not yet been structurally determined and those that cannot be structurally determined, these results may have alternately served to highlight certain targets within each consortium that have progressed further through the pipeline.

Main Protein Features

Conservation Across Organisms

Protein features commonly found in the decision trees are indicated in Table 1. Table 1 also illustrates the differences between the subset containing all the targets and the subset consisting of targets that have been structurally determined. These results highlight which features differ the most between the two subsets. Most of the protein characteristics commonly found in the decision trees also contrast markedly between the two subsets. The percentage of proteins that have COGs rises from 59% in all the targets to 85% in targets that have been structurally determined (p-value = 1.4×10^{-24}).

Pipeline figures (Figure 3) can show where common bottlenecks occur in each step of protein structure determination process. At each stage of the pipeline, the number of total proteins is represented in parenthesis, and the values of the characteristics of interest are shown. Of the 27,711 protein targets, so far 14,767 have been cloned, 8587 expressed, 4115 purified and 370 structurally determined. The results from the tree-based analyses corroborates with the data shown in the pipeline figures. For example, the tree analyses indicate that COG assignment is a major determinant for whether a protein will be expressed. For the COG assignment protein characteristic (Figures 3a), we can see in the pipeline that a major bottleneck occurs at this stage. Of proteins that are cloned but not expressed, 53% have COGs compared to 70% in expressed proteins (p-value = 2.2×10^{-92}). Most proteins are expressed in bacteria so targets with bacterial counterparts have a better chance of being expressed than those that don't. Eukaryotic proteins expressed in bacterial vectors typically have a lower success rate than bacterial proteins.

COGs are families of proteins found in many organisms. Generally, more studies have been performed on these proteins due to their presence in various organisms, which can increase

the probability that these proteins will be structurally determined. Studying proteins that belong to COGs is one of the most successful methods of tackling crystal structure determination[18] since these proteins can be cloned and purified from many organisms. This study highlights the utility of multiplex gene expression system analysis.

Hydrophobicity

In Table 1, the percentage of small hydrophobic protein residues (GAVLI) increase from 34.6% in the subset containing all the target proteins to 38.4% in the subset consisting of proteins that are structurally determined (p-value=9.6×10^{-19}). However, the average hydrophobicity (hphobe) and the average number of hydrophobic residues within hydrophobic stretches (hp_aa) are shown to decrease between the subsets of all target proteins and structurally characterized proteins. Small hydrophobic residue composition (GAVLI) was found to be one of the determinants for whether proteins could be structurally determined. Structurally determined proteins had an average of 38.4% GAVLI composition as opposed to purified but not structurally determined proteins that had an average of 36% GAVLI composition (p-value = 3.6×10^{-7}).

Through target selection, most proteins with predicted transmembrane regions have been removed. However, target proteins that are more hydrophobic are usually less amenable to high-throughput experimentation probably due to solubility issues. Overall, 30% of all the targets have hydrophobic stretches with minimum hydrophobicity scores below –1 (based on the GES scale[19], see methods) compared to the 21% that are structurally determined (p-value = 1.4×10^{-4}). Figure 3(c) highlights two bottlenecks in the pipeline that could occur based on the protein's hydrophobicity features. 36% of proteins that are cloned but not expressed have hydrophobicity scores below –1 compared to 25% of expressed proteins (p-value = 7.2×10^{-36}). Similarly, while

29% of proteins that are expressed but not purified have low hydrophobicity scores, only 20% of purified proteins contain highly hydrophobic patches (p-value = 3.1×10^{-22}). This seems to confirm the idea that hydrophobic proteins are less likely to be both expressed and purified due to their decreased solubility.

In general, all the protein targets had an average of 15 hydrophobic residues within hydrophobic stretches compared to structurally determined proteins that had an average of 7 hydrophobic residues (p-value = 3.8×10^{-5}). For proteins with a high number of hydrophobic residues within hydrophobic stretches, the pipeline analysis suggests that two bottlenecks can occur at the expression stage and the purification stage. The decision tree for the purification stage also indicates that this feature is a strong determinant for whether a protein can be purified. The expressed proteins that do not become purified have a high number of 16.3 hydrophobic residues compared to the purified proteins that have only 6.5 hydrophobic residues within hydrophobic stretches (p-value = 7.7×10^{-30}). This suggests that proteins with large or many hydrophobic stretches may not fold properly in the experimental conditions being used. The cloned proteins have an average of 16.3 hydrophobic residues within a hydrophobic stretch (p-value=9.3×10^{-8}). Cloned proteins that were not expressed had an average of 22.7 hydrophobic residues compared to expressed proteins that had an average of 11.6 hydrophobic residues (p-value = 1.0×10^{-63}), indicating that proteins with many hydrophobic stretches may have more difficulty to be expressed due to their decreased solubility.

Protein Length

Protein length was another feature that decreased from 291 in the data set containing all the target proteins to 243 in the data set consisting of proteins that were structurally determined

(p-value = 3.2×10^{-4}). The tree analyses suggested that the outcome of whether a protein was expressed could be correlated to its sequence length. The pipeline analysis results corroborated this finding. While cloned proteins had an average length of 276 residues (p-value = 4.8×10^{-25}), the average protein length of cloned but not expressed proteins was considerably higher at 305 residues compared to expressed protein that had an average protein length of 254 (p-value = 2.6×10^{-34}). One explanation is that large proteins could contain multiple domains, which increases the difficulty of experimental analysis.

The histogram of the protein lengths (Figure 5) demonstrates that the distribution of protein lengths is different between proteins that are expressed to those that are cloned but not expressed. There are more expressed than non-expressed proteins that have a length of less than 400 amino acids. However, there is a larger number of cloned but not expressed proteins for proteins with lengths of 400 to 2000 amino acids. The analysis suggests a possible correlation between a protein's length and its amenability to high-throughput structural determination.

Composition of Specific Amino Acids

The percent composition of small negatively charged residues (DE) increased from 12.7% in all the protein targets to 14.2% in structurally determined targets. The bottleneck for this feature, indicated in figure 2(f), was found in the purification stage where purified proteins had 14.2% DE (Asp/Glu) composition as opposed to 12.6% expressed proteins that were not purified (p-value = 3.3×10^{-101}). The percent composition of DE was highlighted in several tree analyses including structure determination, purification, and solubility. Highly charged amino acids interact favorably with solvent molecules so proteins with a higher percent of acidic amino acids have a higher probability of being soluble. This is corroborated in the results of the

decision tree analyses where this feature was highlighted in the purification and solubility determination trees.

The percent composition of serine on average decreased from 6.7% in all the protein targets to 5.2% in structurally determined targets (p-value=3.0×10^{-22}). Serine was highlighted in the solubility tree analyses suggesting its influence on a protein's solubility. However, it is not clearly understood how the decrease in serine composition affects a protein's amenability to high-throughput experimental analysis.

Number of Binding Partners

Based on the interaction information found in the MIPS complex catalog (complex_partners), the average number of binding partners for proteins that are structurally determined is less than all the protein targets, suggesting that some proteins may require the presence of their binding partners in order to fold properly and in turn be structurally determined.[20-23] The pipeline figure analysis indicates that a bottleneck for this feature occurs at the purification stage where proteins that are expressed but not purified have an average of 0.96 binding partners compared to those that become purified which have an average of 0.7 binding partners (p-value = 0.06). The average number of binding partners decreases even further from the stage when a target is purified to the stage that it becomes structurally determined, suggesting that it is easier to purify and crystallize proteins that have few binding partners. The reported average number of binding partners for all of these proteins is less than one due to the fact that less than 4% of these proteins are functionally shown to have binding partners. As more functional information becomes available, the correlation between the number of binding

partners and a protein's amenability to high-throughput structural determination will most likely become more pronounced.

Occurrence of Signal Peptides

Signal peptides control the entry of proteins into secretory pathways[24-26]. As the protein is translocated through the membrane, the signal sequence is cleaved off releasing the mature protein. The presence of a signal sequence was not highlighted as a determining factor in any of the decision tree analyses. However, there was a general decrease from 14.5% in all the targets to 8.2% in the structurally determined proteins (p-value = 5.3×10^{-4}). The pipeline analysis suggests that expressed proteins have a lower percent (13%) of signal sequences than proteins that are not expressed (16.5%) (p-value = 1.2×10^{-7}). Similarly, 10.6% of purified proteins have signal sequences compared to 15.7% of expressed but not purified proteins (p-value = 9.3×10^{-12}). Since secreted proteins are exported, their native environment is most likely extracellular. However, the current expectation profile for proteins that are expected to be successful are proteins that are both cytoplasmic and monomeric.

Features not Found to be Predictors

All the target proteins have an averaged normalized low complexity value of 14.4 compared to structurally determined proteins that have a value of 13.6 (p-value = 0.07). While there is a small decrease, this difference is probably not significant. However, cloned proteins have an average low complexity value of 14.6 compared to expressed proteins that have a value of 12.8 (p-value = 2.1×10^{-54}). Through target selection, most proteins with long low complexity

regions have been removed[4,27], which may indicate why there is little change in this feature between all the targets and those that become structurally determined.

There is little difference between the percentage of proteins containing an NLS (nuclear localization signal) motif (PS00015) in the structurally determined subset and all the targets. There is a small increase from 4.3% of all the targets to 4.9% of the structurally determined targets that contain the NLS motif. However, this result is not found to be significant (p-value=0.52).

Other features such as cysteine composition and KR (Lys/Arg) composition did not show a distinct change between the structurally determined subset and the total targets subset. It was thought that proteins containing disulfide bridges might be more difficult to crystallize. Similar to proteins with signal sequences, proteins with disulfide bridges are usually extracellular. However, the amount of cysteine in proteins that were crystallized compared to those that were not did decrease, but not substantially. Additionally, the percent composition of the charged residues, DE (Asp/Glu), increased in proteins that were structurally determined compared to all the protein targets. We had thought to observe a change in the percent composition of other charged residues such as KR (Lys/Arg). However, there was very little change in this feature between the structurally determined subset and all the targets.

Statistical Issues using Structural Genomics Datasets

Since the data collected from the TargetDB is basically a “snapshot” of the structural genomics progress, it is not possible to distinguish between targets that have failed at a certain stage from those targets that are yet to be studied. The subset of successful proteins can be partitioned into a "white" (or successful) subset. However, the remainder of the proteins is

partitioned into a "gray" subset, rather than a completely unsuccessful ("black") subset, since some of these proteins could potentially be successful if attempted. This can create more difficulty in making accurate conclusions when trying to determine which features are important to the amenability of a protein to high-throughput experimentation. While the presence of potential false negatives can decrease the strength of prediction, our analysis shows a statistically significant correlation between certain protein features and their amenability to being successfully determined. The issue of having false negatives would manifest itself more greatly if the successful subset of data was just as or almost the same as the negative subset of data with respect to the features being studied, thereby creating non-statistically significant differences. However since the protein features in this subset of proteins are statistically distinct from the protein features found in the rest of the population, this indicates that there is enough information to distinguish the two (unbalanced) sets even from this smaller successful subset. With the accumulation of more “successful” data, more information will be learned to increase the ability for predictions.

Using these machine-learning techniques, we are able to identify common trends or correlations between specific protein properties and the success of an outcome at each stage of the structural genomics pipeline. This analysis uses standard tools that have been widely employed in such fields as econometrics and epidemiology and produces straightforward, robust statistical conclusions. However, these techniques cannot be directly used to make causal relationships. Incorporating general biological information, we are able to utilize statistical information to make inferences about the relationships between factors and outcomes and determine whether these relationships are plausible.

More specifically, the correlations found in these high-throughput experiments can be defined as either causal or non-causal. The causal correlations are based on protein properties and features of a particular experiment. For instance, the results show that less hydrophobic proteins have a better chance to be purified. For the purification stage, the rate of success is highly dependent on a protein's fundamental properties. Alternatively, non-causal correlations are affected by other factors such as the rate that scientists can start analyzing proteins at each stage of the process or biases in the way targets have been selected or prioritized for cloning, expression, crystallization, etc. These non-causal relationships have less effect in later stages of the structural genomics pipeline where there are fewer proteins. However, they seem to be more predominant in earlier stages of the pipeline such as in the cloning step where the number of proteins selected outweighs the amount of resources available.

Interpretation of Pipeline Figures

Because of the distinction between causal and non-causal correlations that we elaborated on above, we have to be particularly careful in interpreting the pipeline figures with respect to cloning. It is known that cloning of proteins is usually believed to be successful. This issue has ramifications for interpreting the pipeline figures. A small analysis we have done within the NESG consortium has shown that most proteins can be cloned with approximately 95% success (i.e. somewhat higher cloning success rates for prokaryotic proteins and somewhat lower success rates for eukaryotic proteins). Therefore, most of the statistical differences that we observed between the selected and cloned parts of the pipeline do not reflect intrinsic properties of proteins. Rather they reflect subtle sociological biases in the way the various structural genomic centers have gone about picking their targets for cloning. Thus, in interpreting the pipeline

figures, if we want to understand the appropriate statistics of each stage we can start looking at the pipeline figures from the cloning stage on up to expressed, purified, and so forth. Each of these steps represents a real statistical difference attributed to the properties of proteins.

However, there is another way that we can look at the pipeline semantics. In a very global manner, we can compare the properties of proteins that have structures to the entire universe that is tackled by structural genomics. The latter is essentially the proteins of TargetDB. In this way, we can compare the very top of the pipeline schematic all the way to the bottom, taking into account that almost all of the proteins in the selected pool could be cloned if attempted. This gives us an idea of the overall statistical differences between proteins that are broadly targeted by the centers versus those that have been successfully solved.

Important Discoveries for Future Data Collection in TargetDB

From this comprehensive analysis we can glean a number of points that will help us to better gather data for TargetDB in the future. First, it is important to adequately gather negative as well as positive information. Second, it is critical to distinguish between proteins with negative information in the pipeline and proteins that are waiting in the pipeline. In particular, it is critical to track (i) what is attempted, (ii) what is successful, and (iii) what fails, at each step of the structure production pipeline. Therefore, it might be useful to add a number of categories to TargetDB including more detail about the status of each protein.

As the recognition in the importance of characterizing structural genomics information increases, more detailed mechanisms for capturing this data will greatly improve and enable further data mining efforts. Here we have shown the general utility of performing this type of analysis on information gathered from the structural genomics efforts. This paper identifies

protein features that correlate with successful outcomes at each stage of the pipeline. We demonstrate plausible consistencies between these identified protein properties and the effect that they may have in determining the outcome of protein's progress through the structural genomics pipeline. The results of this analysis can aid researchers to choose better target proteins and, therein, can increase the efficiency of high-throughput experimentation.

Conclusions

The structural genomics initiative will produce a vast amount of experimental information that can provide insights into protein structure and function. As the numbers of solved structures are gradually increasing, data collected from these efforts can aid in optimizing and accelerating the structure determination process. This study suggests that several key protein characteristics including protein length, composition of negatively charged and polar residues, hydrophobicity, presence of a signal sequence, and COG assignment can determine whether a protein will progress through the stages of the structure determination pipeline. Proteins with an optimal combination of these features can be rapidly selected and moved through the pipeline more efficiently. Additionally, using these parameters, it is possible that less ideal proteins can be re-engineered to increase their chance for being structurally determined.

Methods

Targets

The data set of protein targets was collected on February 9, 2003 from TargetDB (http://targetdb.pdb.org/), a target registration database that includes target data from worldwide

structural genomics and proteomics projects. This subset consisted of 27,711 proteins and was inserted into the SPINE database for further analysis.

Random Forest Analysis

The random forest analysis combines two powerful ideas in machine learning techniques: bagging and random feature selection. Bagging (bootstrap aggregating) uses the final vote of bootstrap replicates to create a classifier. The random forest analysis grows a random tree for each bootstrap sample by choosing the best split at each node from a small number of randomly selected predictors.

For each data set, 5000 bootstrap samples were used and seven features (the square root of the total number of features) were randomly selected at each node split. One third of the original training set was omitted from each bootstrap sample. These out-of-bag (oob) samples were placed back into the trees and a final prediction was based on the votes of these tree predictions. The error of each data set was calculated based on the percent of incorrect predictions made out of the total number of predictions. The error rates for these data sets range on average between 15-30%.

The importance for each variable was measured by permuting all the values for the variable in the oob samples. When these values were placed back in the tree, a new test set error was computed. The amount that the test error differs from the original test error was defined as the importance of the variable.

Decision Tree Analysis

Decision trees were constructed using the R tree model software[28,29] with default parameters of minimum node deviance of 0.1, minimum node size of 10, and cost complexity pruning of 5. Targets with missing values were omitted resulting in 27,267 total protein targets analyzed. The overall structure determination tree used a subset of proteins that were determined in comparison to all the rest of the proteins that were not. Correspondingly, the cloning determination tree used a subset of proteins of all cloned target proteins compared to all target proteins that were not cloned. Decision trees were also constructed for the expressed versus cloned but not-expressed, purified versus expressed but not-purified, and structure versus purified but not-structurally determined subsets. Under no associations, the distribution of the positive outcome out of the total number of samples (ie 14385 cloned out of the total 27267 sample) was randomly assigned to each of the terminal nodes. The mean and variance of the terminal nodes were calculated and the approximate p-value for each terminal node was derived from its Z-score.

Cross-validation is a commonly used technique to estimate the error rate of future predictions. The ipred package[30] in R was used to perform a 10-fold cross-validation on each of the decision trees, where each successive application of the learning procedure used a different 90% of the data set for the training and the remaining 10% for testing. The estimated error was calculated by taking the sum of the number of incorrect classifications obtained from each one of the ten test subsets and divided that sum by the total number of instances that had been used for testing. The average prediction success over all the decision trees in figure 1 was 76%. For the

decision trees in figure 4, the cross-validation approach resulted in an overall prediction success of 96%.

Data Analysis

Amino Acid Composition

Amino acid compositions were calculated by taking the fraction of the total number of the amino acid residue by the total number of amino acids in the whole sequence.

Hydrophobicity

Hydrophobicity scores were measured using the GES scale[19] where the lower the score, the more hydrophobic the amino acid is. Hydrophobic residues within hydrophobic stretches were calculated by counting all the residues within a sliding twenty amino acid window with a hydrophobicity score below –1.0 kcal/mol on the GES hydrophobicity scale. Minimum hydrophobicity scores were calculated using the minimum hydrophobicity score of all the sliding twenty amino acid windows for each protein.

Signal Sequences

Signal sequences were measured by implementing a pattern match for sequences containing a charged residue within the first seven amino acids followed by a stretch of 14 hydrophobic residues.

Low Complexity Scores

Entropic low complexity scores were calculated using the SEG[31] program. Long low complexity regions were identified with SEG using standard parameters with a trigger complexity K(1) of 3.4, an extension complexity K(2) of 3.75, and a sequence window length of 45. Short low complexity regions were identified using a trigger complexity K(1) of 3.0, an extension complexity K(2) of 3.3, and a sequence window length of 25.

Motifs

Prosite motifs were identified using the ps_scan[32] program to scan the Prosite[33] database for known motifs.

Binding Partners

A commonly used method for identifying protein-protein interactions is to utilize known binding information in one species to predict interactions of proteins in another species[34-36]. Interologs are pairs of potential orthologs of known interacting partners. We employed interolog information to identify target binding partners by mapping the known binding partners of a target's yeast interolog to itself. The complex_partners values were measured by calculating the average number of known binding partners for the yeast interolog found in the MIPS[37-41] complex catalog and mapping it to the protein target. Comparatively, the any_partners values identified the average number of known binding partners for a specific target's yeast interolog using various sources[37-51] of experimentally determined information including the MIPS database.

Statistical Significance for Pipeline Figures

To test whether the features illustrated in Figure 3 show an increasing or decreasing trend from the total target population of 27711 proteins to those 370 proteins with identified structures, we conducted trend tests in the following form: $T = \sum_{i=1}^{C} w_i y_i$, where C is the number of classes above the baseline total population, w_i is the weight for the ith class, and y_i is the observed mean or ratio related to the feature of interest in the ith class. In our case, $C = 4$, which corresponds to the cloned proteins, expressed proteins, purified proteins, and proteins with identified structures. The value of y_i is the proportion of the proteins with a given feature in the ith class for 3(a), 3(c), and 3(h), whereas the value of y_i is the average feature value in the ith class for the other features in Figure 3. The weights are 1, 2, 3, 4 for an increasing trend for features in 3(a), 3(b), 3(f), and 3(i). The weights are 4, 3, 2, and 1 for a decreasing trend for the other features.

To assess the statistical evidence of a trend in the data based on T, we calculated the mean and variance of T under the null hypothesis of no trend conditional on the feature distribution in the baseline total population consisting of 27711 proteins. It can be shown that when the feature of interest is binary, i.e. a given protein either has or does not have this feature,

$$E(T) = \sum_{i=1}^{C} w_i y_0 \text{ and } Var(T) = \left(\sum_{i=1}^{C} \beta_i \right) y_0(1-y_0) - \sum_{i=1}^{C} \beta_i v_i,$$

where

$$\beta_i = \left(\sum_{j=i}^{C} w_j \right)^2 \frac{N_{i-1} - N_i}{N_{i-1} - 1} \frac{1}{N_i}, v_1 = 0, v_{i+1} = (1-\alpha_i) v_i + \alpha_i y_0(1-y_0), \text{ for } i \geq 1, \alpha_i = \frac{N_{i-1} - N_i}{N_{i-1} - 1} \frac{1}{N_i},$$

N_i is the number of proteins in the *i*th class, N_0 is the number of proteins in the total population, and y_0 is the proportion of proteins having a given feature in the total population. The statistical significance of the observed increasing trend is

$$1-\Phi(\frac{T-E(T)}{\sqrt{Var(T)}}),$$

and the statistical significance of the observed decreasing trend is

$$\Phi(\frac{T-E(T)}{\sqrt{Var(T)}}),$$

where Φ is the cumulative function of the standard normal distribution.

When the feature of interest is a continuous one, it can be shown that, under the null hypothesis of no trend,

$$E(T)=\sum_{i=1}^{C} w_i y_0 \text{ and } Var(T)=\left(\sum_{i=1}^{C}\gamma_i\right)\sigma_0^2,$$

where $\gamma_i=\left(\sum_{j=i}^{C} w_j\right)^2 \frac{N_{i-1}-N_i}{N_{i-1}}\frac{1}{N_i}$, y_0 and σ_0^2 are the mean and variance of the given feature in the total population, respectively. For the observed trend test statistic, we can use the above mean and variance under the null hypothesis to assess statistical significance level for an increasing or decreasing trend as above for the binary case.

Acknowledgements

This work was supported in part by grant 5P50GM062413-03 from the Protein Structure Initiative of the Institute of General Medical Sciences, National Institutes of Health and grant DMS-0241160 (to HYZ) from the NSF. We thank Tom Acton for helpful discussions.

References

1. Service, R.F. Structural genomics offers high-speed look at proteins. 2000 *Science*, **287,** 1954-1956.
2. Brenner, S.E. and Levitt, M. Expectations from structural genomics. 2000 *Protein Sci*, **9,** 197-200.
3. Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M.A., Madhusudhan, M.S., Mirkovic, N. and Sali, A. Protein structure modeling for structural genomics. 2000 *Nat Struct Biol*, **7 Suppl,** 986-990.
4. Brenner, S.E. Target selection for structural genomics. 2000 *Nat Struct Biol*, **7 Suppl,** 967-969.
5. Brenner, S.E. A tour of structural genomics. 2001 *Nat Rev Genet*, **2,** 801-809.
6. Service, R.F. Structural genomics. Tapping DNA for structures produces a trickle. 2002 *Science*, **298,** 948-950.
7. Pedelacq, J.D., Piltch, E., Liong, E.C., Berendzen, J., Kim, C.Y., Rho, B.S., Park, M.S., Terwilliger, T.C. and Waldo, G.S. Engineering soluble proteins for structural genomics. 2002 *Nat Biotechnol*, **20,** 927-932.
8. Terwilliger, T.C. Structural genomics in North America. 2000 *Nat Struct Biol*, **7 Suppl,** 935-939.
9. Chance, M.R., Bresnick, A.R., Burley, S.K., Jiang, J.S., Lima, C.D., Sali, A., Almo, S.C., Bonanno, J.B., Buglino, J.A., Boulton, S., Chen, H., Eswar, N., He, G., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M. and Wang, L.K. Structural genomics: a pipeline for providing structures for the biologist. 2002 *Protein Sci*, **11,** 723-738.

10. Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T. and Gerstein, M. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. 2001 *Nucleic Acids Res*, **29,** 2884-2898.

11. Goh, C.S., Lan, N., Echols, N., Douglas, S.M., Milburn, D., Bertone, P., Xiao, R., Ma, L.C., Zheng, D., Wunderlich, Z., Acton, T., Montelione, G.T. and Gerstein, M. SPINE 2: a system for collaborative structural proteomics within a federated database framework. 2003 *Nucleic Acids Res*, **31,** 2833-2838.

12. Breiman, L. Random Forests. 2001 *Machine Learning*, **45,** 5-32.

13. Breiman, L. (2002), *IMS Wald Lecture 2*.

14. Quinlan, J.R. Simplifying decision trees. 1987 *Int. J. Man-Machine Stud.*, **27,** 221-234.

15. Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

16. Dash, M.a.L., H. Feature Selection for Classification. 1997 *Intelligent Data Anal.*, **1,** 131-156.

17. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. A genomic perspective on protein families. 1997 *Science*, **278,** 631-637.

18. Savchenko, A., Yee, A., Khachatryan, A., Skarina, T., Evdokimova, E., Pavlova, M., Semesi, A., Northey, J., Beasley, S., Lan, N., Das, R., Gerstein, M., Arrowmith, C.H. and Edwards, A.M. Strategies for structural proteomics of prokaryotes: Quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. 2003 *Proteins*, **50,** 392-399.

19. Engelman, D.M., Steitz, T.A. and Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. 1986 *Annu Rev Biophys Biophys Chem*, **15,** 321-353.

20. Wright, P.E. and Dyson, H.J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. 1999 *J Mol Biol*, **293,** 321-331.

21. Dyson, H.J. and Wright, P.E. Coupling of folding and binding for unstructured proteins. 2002 *Curr Opin Struct Biol*, **12,** 54-60.

22. Yokoyama, S. Protein expression systems for structural genomics and proteomics. 2003 *Curr Opin Chem Biol*, **7,** 39-43.

23. Dunker, A.K. and Obradovic, Z. The protein trinity--linking function and disorder. 2001 *Nat Biotechnol*, **19,** 805-806.

24. Gierasch, L.M. Signal sequences. 1989 *Biochemistry*, **28,** 923-930.

25. von Heijne, G. Protein targeting signals. 1990 *Curr Opin Cell Biol*, **2,** 604-608.

26. Rapoport, T.A. Transport of proteins across the endoplasmic reticulum membrane. 1992 *Science*, **258,** 931-936.

27. Sali, A. Target practice. 2001 *Nat Struct Biol*, **8,** 482-484.

28. Ihaka, R. and Gentleman, R. R: A language for data analysis and graphics. 1996 *Journal of Computational and Graphical Statistics*, **5,** 299-314.

29. Team, R.D.C. R: A language and environment for statistical computing. 2003 *http://www.R-project.org*.

30. Peters, A. and Hothorn, T. Improved Predictors. 2003 *http://cran.r-project.org/src/contrib/PACKAGES.html#ipred*.

31. Wootton, J.C. and Federhen, S. Analysis of compositionally biased regions in sequence databases. 1996 *Methods Enzymol*, **266,** 554-571.

32. Gattiker, A., Bienvenut, W.V., Bairoch, A. and Gasteiger, E. FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. 2002 *Proteomics*, **2,** 1435-1444.

33. Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. and Bucher, P. PROSITE: a documented database using patterns and profiles as motif descriptors. 2002 *Brief Bioinform*, **3,** 265-274.

34. Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N. and Vidal, M. Protein interaction mapping in C. elegans using proteins involved in vulval development. 2000 *Science*, **287,** 116-122.

35. Walhout, A.J. and Vidal, M. Protein interaction maps for model organisms. 2001 *Nat Rev Mol Cell Biol*, **2,** 55-62.

36. Yu, H., Luscombe, N.M., Zhu, X., Chung, S., Goh, C.-S. and Gerstein, M. Annotation transfer for genomics: assessing the transferability of protein-protein and protein-DNA interactions between organisms. 2004 *Genome Research*.

37. Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. MIPS: a database for genomes and protein sequences. 2002 *Nucleic Acids Res*, **30,** 31-34.

38. Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S. and Weil, B. MIPS: a database for genomes and protein sequences. 2000 *Nucleic Acids Res*, **28,** 37-40.

39. Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. and Frishman, D. MIPS: a database for genomes and protein sequences. 1999 *Nucleic Acids Res*, **27,** 44-48.

40. Mewes, H.W., Hani, J., Pfeiffer, F. and Frishman, D. MIPS: a database for protein sequences and complete genomes. 1998 *Nucleic Acids Res*, **26,** 33-37.

41. Mewes, H.W., Albermann, K., Heumann, K., Liebl, S. and Pfeiffer, F. MIPS: a database for protein sequences, homology data and yeast genome information. 1997 *Nucleic Acids Res*, **25,** 28-30.

42. Tong, A.H., Drees, B., Nardelli, G., Bader, G.D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C.W., Fields, S., Boone, C. and Cesareni, G. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. 2002 *Science*, **295,** 321-324.

43. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. 2000 *Nature*, **403,** 623-627.

44. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. 2001 *Proc Natl Acad Sci U S A*, **98,** 4569-4574.

45. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M. and Eisenberg, D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. 2002 *Nucleic Acids Res*, **30,** 303-305.

46. Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M. and Eisenberg, D. DIP: the database of interacting proteins. 2000 *Nucleic Acids Res*, **28,** 289-291.

47. Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., Marcotte, E.M. and Eisenberg, D. DIP: The Database of Interacting Proteins: 2001 update. 2001 *Nucleic Acids Res*, **29,** 239-241.

48. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. 2002 *Nature*, **415,** 141-147.

49. Bader, G.D., Betel, D. and Hogue, C.W. BIND: the Biomolecular Interaction Network Database. 2003 *Nucleic Acids Res*, **31,** 248-250.

50. Bader, G.D. and Hogue, C.W. BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. 2000 *Bioinformatics*, **16,** 465-477.

51. Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T. and Hogue, C.W. BIND--The Biomolecular Interaction Network Database. 2001 *Nucleic Acids Res*, **29,** 242-245.

Figure 1. Decision tree analysis of two subsets of protein targets - (a) structurally determined versus not structurally determined, (b) cloned versus not cloned, (c) expressed versus cloned but not expressed, (d) purified versus expressed but not purified, (e) structurally determined versus purified but not structurally determined, and (f) soluble versus not soluble. The boxed values in the terminal nodes show the probability of a successful outcome at each node. To the right of the node, the value represents the number of successful proteins and the value to the left of the node denotes the number of proteins that were unsuccessful. The bracketed number at the root of the tree is the number of total proteins analyzed.

Figure 2. Decision tree analysis of two subsets of protein targets with only fundamental descriptors.

Figure 3. Pipeline figure analysis of (a) percent of proteins that belong to COGs, (b) average percent of GAVLI percent composition, (c) percent of proteins that have minimum hydrophobicity scores below –1 on the GES scale, (d) number of hydrophobic residues within hydrophobic stretches below –1 on the GES scale, (e) average protein length, (f) average percent of DE percent composition, (g) average number of binding partners based on the MIPS complex catalog, (h) percent of proteins that contain signal sequences, and (i) average normalized low complexity values. For proteins containing these features, stages where possible "bottlenecks" can occur are presented in the pipeline figures. The pipeline figures show the mean values and their standard errors. The numbers in parentheses are the actual number of proteins at each stage in the structural genomics pipeline.

Figure 4. Decision tree analysis of two subsets of data from targets that are not structurally determined and those that are. This analysis is performed on data from (a) the Northeast Structural Genomics Consortium (NESG), (b) the Joint Center for Structural Genomics (JCSG), (c) the Mycobacterium tuberculosis Structural Genomics Consortium (TB), (d) the Midwest Center for Structural Genomics (MCSG), (e) the Montreal-Kingston Bacterial Structural Genomics Initiative (BSGI), and (f) the Berkeley Structural Genomics Center (BSGC). The boxed values in the terminal nodes show the probability of a successful outcome at each node. To the right of the node, the value represents the number of structurally determined proteins and the value to the left of the node denotes the number of proteins that were unsuccessful. The bracketed number at the root of the tree is the number of total proteins analyzed.

Figure 5. Histogram of expressed compared to non-expressed protein lengths.

Table 1. Protein Features Analyzed

Protein Feature	Protein Description	Number of Trees Feature is Found	All Targets	Structurally Determined Targets	Meta Descriptor
COG	Percent that have COGs	>1	59%	85%	Yes
GAVLI	Average GAVLI Percent Composition	>1	34.6%	38.4%	No
DE	Average DE Percent Composition	>1	12.7%	14.2%	No
SCTM	Average SCTM Percent Composition	>1	15.5%	13.6%	No
any_partners	Ave Number of Known Binding Partners of the Yeast Homolog from many sources*	>1	3.8	4.1	Yes
pI	Average pI Value	>1	7.1	6.3	No
length	Average Length	>1	291	243	No
Q	Average Glutamine Percent Composition	>1	3.6%	3%	No
W	Average Tryptophan Percent Composition	>1	1.1%	1%	No
K	Average Lysine Percent Composition	>1	6.7%	6.7%	No
hp_aa	Average Number of Hydrophobic Residues within a Hydrophobic Stretch below a Threshold of –1.0 kcal/mol	1	15	7	No
sheet	Average Beta-Strand Percent Composition	1	15.8%	19.8%	No
cplx_s	Average Normalized Low Complexity Value – Short	1	14.6%	13.3%	No
S	Average Serine Percent Composition	1	6.7%	5.2%	No
E	Average Glutamic Acid Percent Composition	1	7.4%	8.6%	No
NQ	Average NQ Percent Composition	1	7.6%	6.5%	No

DENQ	Average DENQ Percent Composition	1	19.7%	20.5%	No
I	Average Isoleucine Percent Composition	1	5.9%	6.2%	No
AILV	Average AILV Percent Composition	1	29.4%	31.6%	No
ST	Average ST Percent Composition	1	11.5%	10%	No
A	Average Alanine Percent Composition	1	7.2%	8.1%	No
C	Average Cysteine Percent Composition	1	1.5%	1%	No
KR	Average KR Percent Composition	1	12.5%	12.6%	No
M	Average Methionine Percent Composition	1	2.5%	2.4%	No
P	Average Proline Percent Composition	0	4.7%	4.4%	No
V	Average Valine Percent Composition	0	6.9%	7.8%	No
N	Average Asparagine Percent Composition	0	4%	3.5%	No
hphobe	Average Minimum Hydrophobicity Score on the GES Scale	0	2.9	-0.6	No
complex_partners	Ave Number of Known Binding Partners of the Yeast Homolog based on the MIPS complex catalog [38-41]	0	0.74	0.47	Yes
cplx_l	Average Normalized Low Complexity Value - Long	0	20.5	21.4	No
helix	Average Helix Percent Composition	0	40.1%	39%	No
coil	Average Coil Percent Composition	0	44%	41.1%	No
LM	Average LM Percent Composition	0	11.6%	11.8%	No
R	Average Arginine Percent Composition	0	5.9%	5.9%	No
DEKR	Average DEKR Percent Composition	0	26.8%	25.2%	No
HKR	Average HKR Percent Composition	0	14.3%	14.5%	No

D	Average Aspartic Acid Percent Composition	0	5.3%	5.6%	No
F	Average Phenalalanine Percent Composition	0	4.1%	3.7%	No
Y	Average Tyrosine Percent Composition	0	3.1%	2.9%	No
T	Average Threonine Percent Composition	0	5.1%	4.9%	No
H	Average Histidine Percent Composition	0	2.3%	2%	No
G	Average Glycine Percent Composition	0	6.5%	7.3%	No
FWY	Average FWY Percent Composition	0	8.4%	7.6%	No
signal	Percent that have Signal Sequences	0	15%	8%	No
PS00015	Percent that Contain PS00015 (Nuclear Localization Signal Peptide) Prosite Motif	0	6.2%	5.4%	Yes
PS00013	Percent that Contain PS00013 (Membrane Lipoprotein Peptide) Prosite Motif	0	0.5%	0.4%	Yes
PS01129	Percent that Contain PS01129 (enzyme involved in RNA metabolism) Prosite Motif	0	0%	0%	Yes
PS00018	Percent that Contain PS00018 (EF-hand calcium-binding domain) Prosite Motif	0	0.4%	1.4%	Yes
PS00030	Percent that Contain PS00030 (RNA recognition) Prosite Motif	0	0.1%	0.8%	Yes

Table 1 reports the number of times that a feature is found in the decision tree figures 1 and 2. Some of the features that appear in more than one tree may still exhibit no distinct difference between all the targets and those that have been structurally determined (columns 4 and 5). This occurs because certain features have more effect in different stages of the structural genomics

pipeline, such as expression and purification, but not necessarily as great an influence on whether a protein can become structurally characterized.

*Sources include BIND[49-51], DIP[45-47], MIPS[37-41], Cellzome (http://www.celzome.com) databases and datasets from various yeast two-hybrid experiments[42-44].

Table 2. Random Forest Analysis

		Structure vs No Structure	Cloned vs UnCloned	Expressed vs Cloned	Purified vs Expressed	Structure vs Purified	Soluble vs Insoluble
Importance Ranking	1	GAVLI	I	COG	DE	GAVLI	S
	2	DE	Q	length	NQ	A	DE
	3	SCTM	AVILM	Hphobe	pI	C	COG
	4	S	sheet	DE	COG	M	SCTM
	5	DENQ	length	pI	GAVLI	pI	length

Figure 1

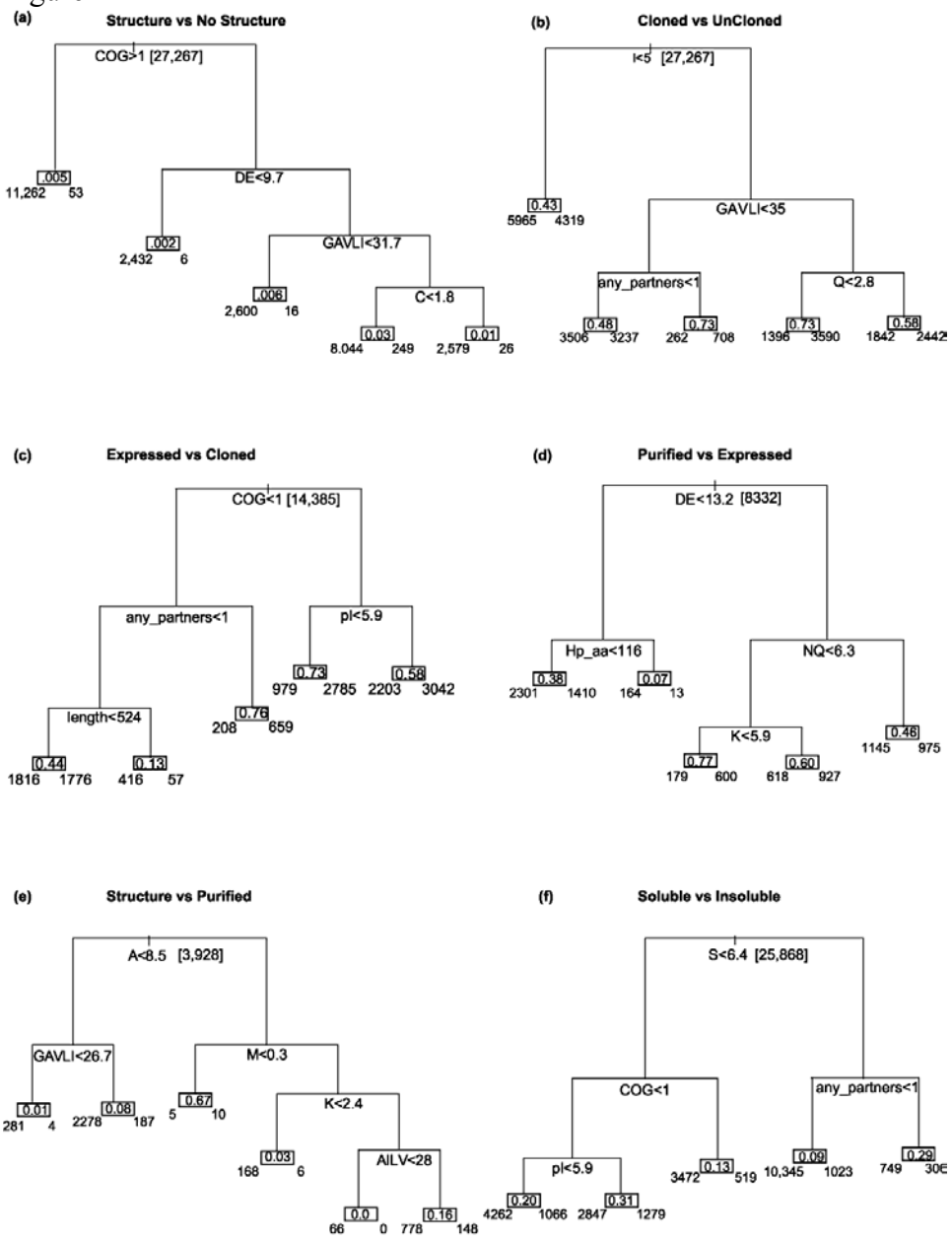

(a)
Structure vs No Structure
COG>1 [27,267]
.005
11,262 53
DE<9.7
.002
2,432 6
GAVLI<31.7
.008
2,600 16
C<1.8
0.03
8.044 249
0.01
2,579 26
(b)
Cloned vs UnCloned
I<5 [27,267]
0.43
5965 4319
GAVLI<35
any_partners<1
0.48
3506 3237
0.73
262 708
Q<2.8
0.73
1396 3590
0.58
1842 2442
(c)
Expressed vs Cloned
COG<1 [14,385]
any_partners<1
pI<5.9
0.73
979 2785
0.58
2203 3042
length<524
0.76
208 659
0.44
1816 1776
0.13
416 57
(d)
Purified vs Expressed
DE<13.2 [8332]
Hp_aa<116
0.38
2301 1410
0.07
164 13
NQ<6.3
K<5.9
0.77
179 600
0.60
618 927
0.46
1145 975
(e)
Structure vs Purified
A<8.5 [3,928]
GAVLI<26.7
0.01
281 4
0.08
2278 187
M<0.3
0.67
5 10
K<2.4
0.03
168 6
AILV<28
0.0
66 0
0.16
778 148
(f)
Soluble vs Insoluble
S<6.4 [25,868]
COG<1
any_partners<1
pI<5.9
0.20
4262 1066
0.31
2847 1279
0.13
3472 519
0.09
10,345 1023
0.29
749 306

Figure 2

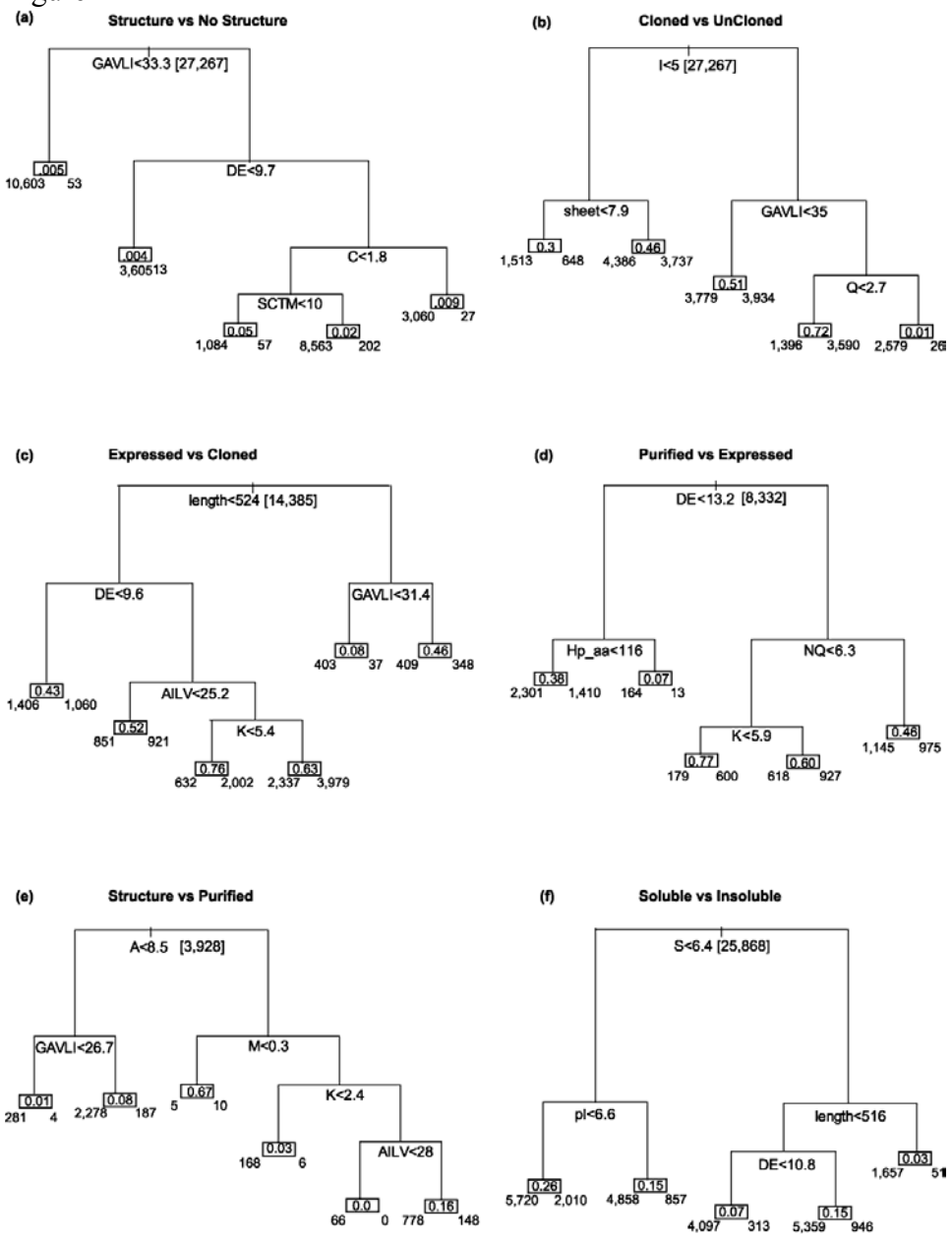

(a)
Structure vs No Structure
GAVLI<33.3 [27,267]
.005
10,603 53
DE<9.7
.004
3,60513
C<1.8
SCTM<10
.009
3,060 27
0.05
1,084 57
0.02
8,563 202
(b)
Cloned vs UnCloned
I<5 [27,267]
sheet<7.9
GAVLI<35
0.3
1,513 648
0.46
4,386 3,737
0.51
3,779 3,934
Q<2.7
0.72
1,396 3,590
0.01
2,579
(c)
Expressed vs Cloned
length<524 [14,385]
DE<9.6
GAVLI<31.4
0.43
1,406 1,060
AILV<25.2
0.08
403 37
0.46
409 348
0.52
851 921
K<5.4
0.76
632 2,002
0.63
2,337 3,979
(d)
Purified vs Expressed
DE<13.2 [8,332]
Hp_aa<116
NQ<6.3
0.38
2,301 1,410
0.07
164 13
K<5.9
0.48
1,145 975
0.77
179 600
0.60
618 927
(e)
Structure vs Purified
A<8.5 [3,928]
GAVLI<26.7
M<0.3
0.01
281 4
0.08
2,278 187
0.67
5 10
K<2.4
0.03
168 6
AILV<28
0.0
66 0
0.16
778 148
(f)
Soluble vs Insoluble
S<6.4 [25,868]
pI<6.6
length<516
0.26
5,720 2,010
0.15
4,858 857
DE<10.8
0.03
1,657
0.07
4,097 313
0.15
5,359 946

Figure 3


(a) Percent that Belong to COGS
(370)
85% STRUCTURES
75% PURIFIED (4115)
70% EXPRESSED (8587)
63% CLONED (14767)
59% TOTAL (27711)
75% NO STRUCTURES (3745)
65% NOT PURIFIED (4472)
53% NOT EXPRESSED (6180)
54% NOT CLONED (12844)
(b) Percent of GAVLI Composition
(370)
38.4%±0.34% STRUCTURES
36.3%±0.14% PURIFIED (4115)
35.8%±0.1% EXPRESSED (8587)
35.7%±0.07% CLONED (14767)
34.6%±0.05% TOTAL (27711)
36%±0.15% NO STRUCTURES (3745)
35.4%±0.13% NOT PURIFIED (4472)
35.5%±0.1% NOT EXPRESSED (6180)
33.3%±0.06% NOT CLONED (12844)
(c) Percent with Hydrophobicity Scores < -1.0 kcal/mol
(370)
21% STRUCTURES
20% PURIFIED (4115)
25% EXPRESSED (8587)
29% CLONED (14767)
30% TOTAL (27711)
20% NO STRUCTURES (3745)
29% NOT PURIFIED (4472)
36% NOT EXPRESSED (6180)
31% NOT CLONED (12844)
(d) Number of Hydrophobic Residues
(370)
6.6±0.74 STRUCTURES
6.5±0.28 PURIFIED (4115)
11.6±0.38 EXPRESSED (8587)
16.3±0.36 CLONED (14767)
15.1±0.24 TOTAL (27711)
6.5±0.3 NO STRUCTURES (3745)
15.3±0.67 NOT PURIFIED (4472)
22.7±0.65 NOT EXPRESSED (6180)
13.8±0.33 NOT CLONED (12844)
(e) Protein Sequence Length
(370)
243±8.9 STRUCTURES
241±2.2 PURIFIED (4115)
254±1.7 EXPRESSED (8587)
276±1.8 CLONED (14767)
291±1.55 TOTAL (27711)
241±2.2 NO STRUCTURES (3745)
266±2..5 NOT PURIFIED (4472)
305±1.6 NOT EXPRESSED (6180)
308±2.6 NOT CLONED (12844)
(f) Percent of DE Composition
(370)
14.2%±0.2% STRUCTURES
14.2%±0.06% PURIFIED (4115)
13.4%±0.04% EXPRESSED (8587)
12.9%±0.03% CLONED (14767)
12.7%±0.02% TOTAL (27711)
14.2%±0.06% NO STRUCTURES (3745)
12.6%±0.06% NOT PURIFIED (4472)
12.3%±0.06% NOT EXPRESSED (6180)
12.4%±0.04% NOT CLONED (12844)
(g) Number of Binding Partners
(370)
0.47±0.2 STRUCTURES
0.7±0.09 PURIFIED (4115)
0.84±0.07 EXPRESSED (8587)
0.84±0.05 CLONED (14767)
0.74±0.04 TOTAL (27711)
0.72±0.09 NO STRUCTURES (3745)
0.96±0.1 NOT PURIFIED (4472)
0.86±0.09 NOT EXPRESSED (6180)
0.62±0.06 NOT CLONED (12844)
(h) Percent that Contain Signal Sequences
(370)
8.2% STRUCTURES
10.6% PURIFIED (4115)
13.3% EXPRESSED (8587)
14.6% CLONED (14767)
14.5% TOTAL (27711)
10.9% NO STRUCTURES (3745)
15.7% NOT PURIFIED (4472)
16.5% NOT EXPRESSED (6180)
14.3% NOT CLONED (12844)
(i) Normalized Low Complexity Value
(370)
13.3±0.9 STRUCTURES
12.4±0.3 PURIFIED (4115)
12.8±0.2 EXPRESSED (8587)
14.6±0.1 CLONED (14767)
14.6±0.1 TOTAL (27711)
12.3±0.3 NO STRUCTURES (3745)
13.2±0.3 NOT PURIFIED (4472)
16.9±0.3 NOT EXPRESSED (6180)
14.7±0.1 NOT CLONED (12844)

Figure 4

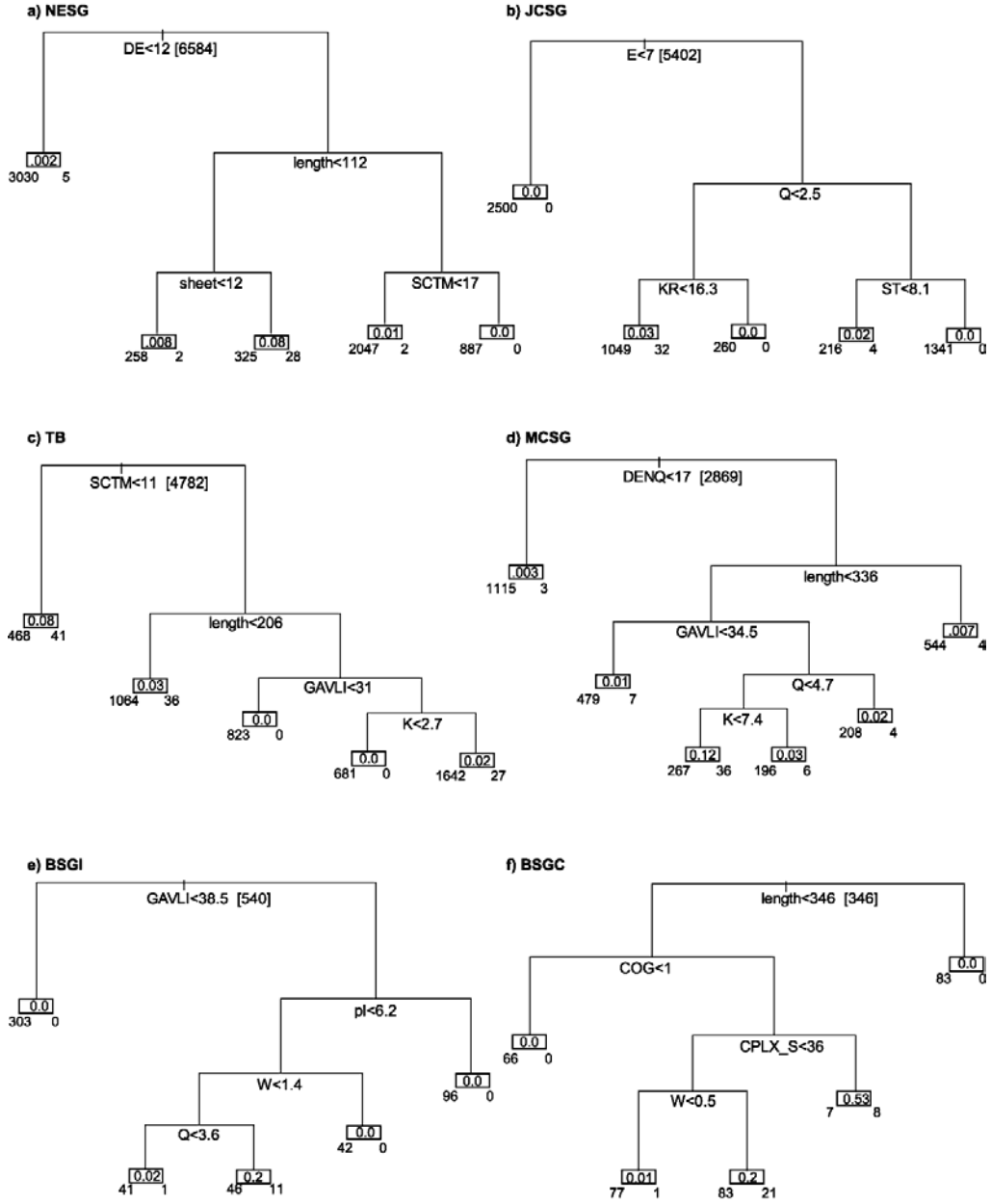
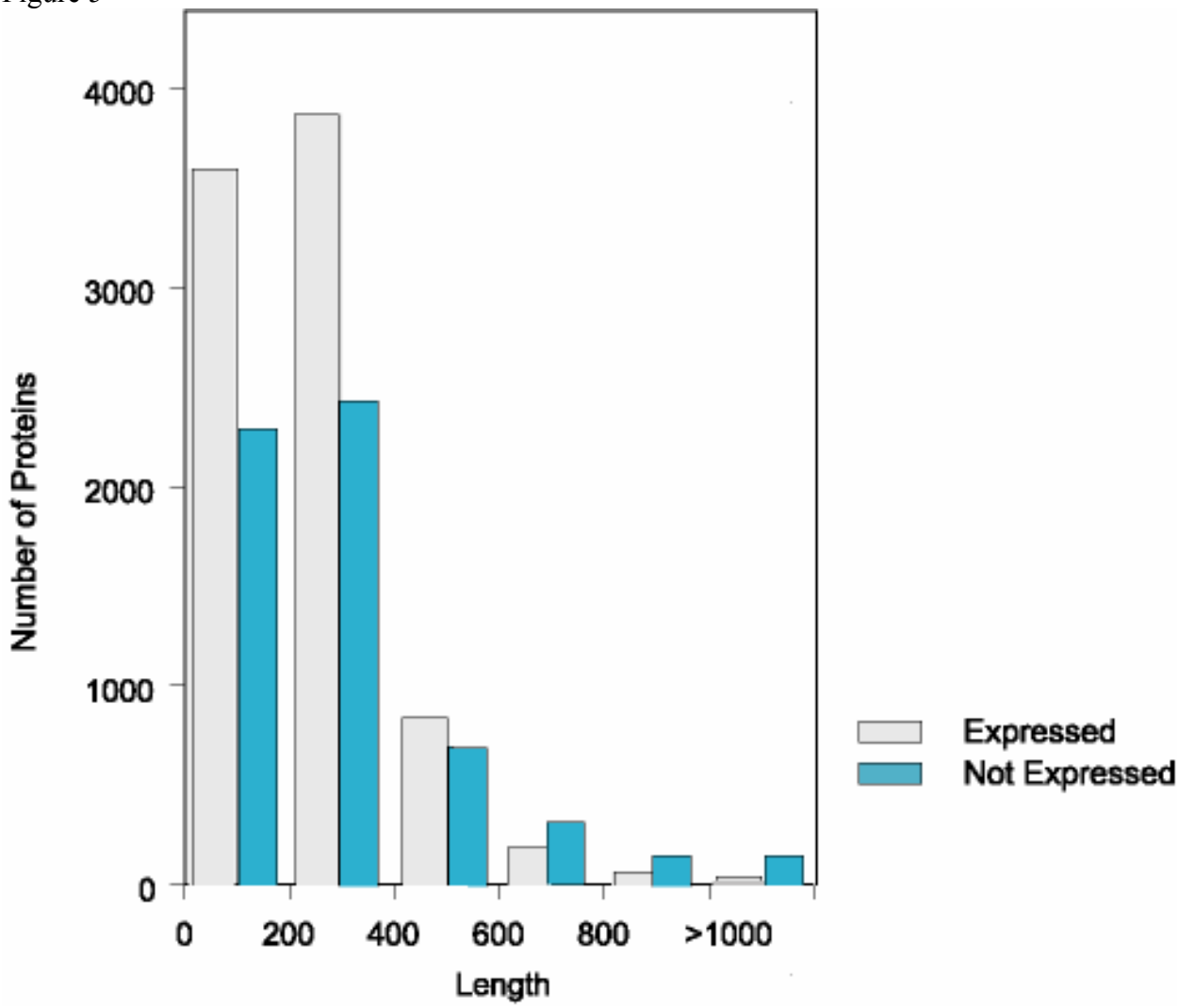

a) NESG
DE<12 [6584]
.002
3030 5
length<112
sheet<12
.008
258 2
0.08
325 28
SCTM<17
0.01
2047 2
0.0
887 0
b) JCSG
E<7 [5402]
0.0
2500 0
Q<2.5
KR<16.3
0.03
1049 32
0.0
260 0
ST<8.1
0.02
216 4
0.0
1341 0
c) TB
SCTM<11 [4782]
0.08
468 41
length<206
0.03
1064 36
GAVLI<31
0.0
823 0
K<2.7
0.0
681 0
0.02
1642 27
d) MCSG
DENQ<17 [2869]
.003
1115 3
length<336
GAVLI<34.5
0.01
479 7
Q<4.7
K<7.4
0.12
267 36
0.03
196 6
0.02
208 4
.007
544 4
e) BSGI
GAVLI<38.5 [540]
0.0
303 0
pI<6.2
W<1.4
Q<3.6
0.02
41 1
0.2
48 11
0.0
42 0
0.0
96 0
f) BSGC
length<346 [346]
COG<1
0.0
66 0
CPLX_S<36
W<0.5
0.01
77 1
0.2
83 21
0.53
7 8
0.0
83 0

Figure 5


4000
3000
2000
1000
0
Number of Proteins
0
200
400
600
800
>1000
Length
Expressed
Not Expressed