EXPLORING BIAS IN THE PROTEIN DATA BANK USING CONTRAST CLASSIFIERS

K. PENG, Z. OBRADOVIC, S. VUCETIC
Center for Information Science and Technology, Temple University, 1805 N Broad St Philadelphia, PA 19122, USA

In this study we analyzed the bias existing in the Protein Data Bank (PDB) using the novel contrast classifier approach. We trained an ensemble of neural network classifiers, called a contrast classifier, to learn the distributional differences between non-redundant sequence subsets of PDB and SWISS-PROT. Assuming that SWISS-PROT is a representative of the sequence diversity in nature while the PDB is a biased sample, output of the contrast classifier can be used to measure whether the properties of a given sequence or its region are underrepresented in PDB. We applied the contrast classifier to SWISS-PROT sequences to analyze the bias in PDB towards different functional protein properties. The results showed that transmembrane, signal, disordered, and low complexity regions are significantly underrepresented in PDB, while disulfide bonds, metal binding sites, and sites involved in enzyme activity are overrepresented. Additionally, hydroxylation and phosphorylation posttranslational modification sites were found to be underrepresented while acetylation sites were significantly overrepresented. These results suggest the potential usefulness of contrast classifiers in the selection of target proteins for structural characterization experiments.

1 Introduction

The ultimate goal of structural genomics is to determine structures for every natural protein through a large-scale structure characterization and computational analysis. However, in anticipation of the development of cost-effective techniques and protocols for large-scale experiments, current efforts in structural genomics are aimed towards determining structures of a limited portion of representative proteins to achieve a rapid coverage of the protein sequence/structure space [3]. As a common approach, the proteins are first filtered to remove those considered inappropriate for structural characterization, e.g., membrane, low complexity, and signal peptides. The remaining proteins are clustered into families based on sequence similarity. Finally, representative proteins from the families of largest biological interest are selected for structural characterization experiments. Although some progress has been made, selection of the target proteins remains an open problem in structural genomics [3].

As the main database of experimentally characterized structural information, Protein Data Bank (PDB) [1] contains more than 20,000 structures of proteins, nucleic acids and other related macromolecules characterized by methods such as X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy. However, current information in PDB is highly biased in the sense that it does not adequately cover the whole sequence/structure space. For example, membrane proteins represent a very important structural class in nature, but their structures are usually extremely difficult to determine due to the need for a lipid bilayer or substitute amphiphile [18]. In

general, PDB is positively biased towards proteins that are more amenable to expression, purification and crystallization. Another source of bias is the fact that different research groups usually have different objectives when selecting the target proteins: some aim at determining structures of proteins from a specific model organism; some may focus on proteins in a single pathway; others may be more interested in certain type of proteins, e.g., disease-related proteins. PDB is also statistically redundant due to the presence of multiple entries for highly similar or identical proteins. According to the PDB statistics available at http://www.rcsb.org/pdb/holdings.html, out of the 3,298 structures deposited in year 2001 only 204 could be considered as novel while the remaining ones were mostly minor variants of those already reported.

Understanding the bias and redundancy in PDB is crucial for selection of further structural targets as well as for various structure predictions. Several studies have been performed towards this goal. Brenner *et al.* [4] analyzed the SCOP [12] structural classification of PDB proteins and reported high skewness at all classification levels. Gerstein [6] compared several complete genomes with a non-redundant subset of PDB and concluded that the proteins encoded by the genomes were significantly different from those in the PDB with respect to sequence length, amino acid composition and predicted secondary structure composition. Liu and Rost [10] analyzed proteomes of 30 organisms and estimated that current structural information in PDB and other databases was available for only 6~38% of all proteins and found over 18,000 segment clusters suitable for structural genomics.

In this paper we provide a complementary view of the bias in PDB that explores differences in sequence properties of PDB and SWISS-PROT [2] proteins. This was accomplished by training an ensemble of neural network classifiers to distinguish between distributions of the non-redundant subsets of PDB and SWISS-PROT. Following the recently proposed contrast classifier framework [14], output of such an ensemble of classifiers measures the level to which a given sequence property is overrepresented/underrepresented in PDB as compared to SWISS-PROT. We applied the contrast classifier to analyze the bias in PDB towards numerous protein properties and to examine whether our approach can be useful in selecting the most interesting target proteins for structural characterization.

2 Methods

2.1 Datasets

Since both PDB and SWISS-PROT are statistically redundant due to the presence of large number of homologues, learning on such data could lead to biased results. Thus, non-redundant subsets were used as unbiased representatives of the two databases. The non-redundant representative of PDB used in this study was PDB-Select-25 [7]

constructed based on all-against-all Smith-Waterman alignments between PDB chains. In this subset, the maximal pairwise identity was limited to 25% since it is believed to be an appropriate compromise between reducing the sequence redundancy and preserving the sequence diversity [17]. The version used in this study was released in December 2002 and consisted of 1,949 chains. After removing chains shorter than 40 residues, the resulting set *PDB25* contained 1,824 chains with 324,783 residues.

For SWISS-PROT (October 2001, Release 40, 101,602 sequences), we applied an approach used in our previous study [20] to construct its non-redundant representative subset. Sequence similarity information from ProtoMap database [22] was used to group all SWISS-PROT proteins into 17,676 clusters using the default ProtoMap E-value cutoff of 1. A representative protein with the richest annotation in SWISS-PROT was then selected from each cluster. Similarly to *PDB25*, proteins shorter than 40 residues were removed. The resulting set *SwissRep* consisted of 16,360 proteins with 6,946,185 residues. The relatively high E-value cutoff, leading to quite aggressive redundancy reduction, was acceptable since the resulting *SwissRep* was still sufficiently large to represent the diversity of SWISS-PROT.

Table 1. *Summary of special regions in SwissRep.*

Regions	**number of regions**	**number of residues**
transmembrane	10,274	215,109 (3.1%)
low complexity	14,648	2,041,162 (29.4%)
disordered	11,332	506,229 (7.3%)

We also identified various regions of interest from *SwissRep* proteins for further analysis. Transmembrane regions were identified through the keywords (KW lines) and feature tables (FT lines) associated with each SWISS-PROT sequence. We identified transmembrane helix regions as the most distinctive among all types of membrane regions. Low complexity regions were marked by the SEG program [21] using the standard parameters K1 = 3.4 and K2 = 3.75, and a window of length 45. Disordered regions longer than 30 residues were predicted by the VL3 disorder predictor [13] with W_{in}/W_{out} = 41/1 and a threshold of 0.85. Table 1 shows the summary of these identified regions with their corresponding sizes measured as the number of regions, the number of residues and the percentage of residues at *SwissRep*.

2.2 Contrast Classifiers

Let us assume we are given dataset G obtained by unbiased sampling from a multivariate underlying distribution, and dataset H obtained by potentially biased sampling from the same distribution. This scenario could occur when objects from H are characterized by a larger set of attributes then those of G. For example, *SwissRep* is an example of unbiased dataset G that contains only protein sequence information, while *PDB25* is an example of biased dataset H that contains both protein sequence and structure information. Understanding the bias in data G is of major importance for

an appropriate analysis and inference from such data. The recently proposed contrast classifier approach [14] provides a simple but effective framework for detecting and exploring the data bias.

By $g(\mathbf{x})$ and $h(\mathbf{x})$ let us denote the probability density functions (pdf) of unbiased data G and biased data H, respectively. The contrast classifier is a classifier trained to discriminate between the distributions of datasets G and H. Using classification algorithms that are able to approximate the posterior class probability (e.g. neural networks), the output $cc(\mathbf{x})$ of a contrast classifier trained on a balanced set with the same number of examples from G (class 1) and examples from H (class 0) approximates $cc(\mathbf{x}) = g(\mathbf{x})/(g(\mathbf{x})+h(\mathbf{x}))$ [14]. With a simple transformation it follows that $h(\mathbf{x})/g(\mathbf{x}) = cc(\mathbf{x})/(1-cc(\mathbf{x}))$, and that $cc(\mathbf{x}) = 0.5$ corresponds to a data point $\mathbf{x}$ that is represented equally well in both datasets (i.e. $h(\mathbf{x})/g(\mathbf{x}) = 1$).

The contrast classifier output $cc(\mathbf{x})$ is therefore a very suitable measure for analysis of the data bias. The distribution of $cc(\mathbf{x})$ gives information about the overall level of bias in dataset H: if it is concentrated around 0.5 the bias is negligible, while if it is dispersed across the interval [0, 1] the bias is significant. Moreover, we could measure the level to which a given data point is overrepresented/underrepresented in dataset H: data points with $cc(\mathbf{x}) < 0.5$ are overrepresented, while those with $cc(\mathbf{x}) > 0.5$ are underrepresented.

2.3 Training Contrast Classifiers for Bias Detection in PDB

In this study we assume that *SwissRep* is a representative of the protein sequence space, while *PDB25* is a biased sample. Note that, while the first assumption is probably not completely correct since SWISS-PROT represents only the proteins studied with a sufficient detail, it is acceptable for the purpose of analyzing the bias in PDB. Based on the description in Section 2.2, it is evident that contrast classifiers can be used directly to explore the bias in PDB. While any classification algorithm able to approximate posterior class probability can be employed to train a contrast classifier, in this study we used feedforward neural networks with one hidden layer and sigmoidal neurons.

Since there is a large imbalance in the number of data points in *SwissRep* and *PDB25* datasets (with the proportion of approximately 21:1), learning a single neural network on balanced samples from the two datasets would not properly utilize the data diversity present in *SwissRep*. We addressed this by training an ensemble of neural networks on balanced training sets consisting of equal number of *PDB25* and *SwissRep* examples randomly sampled from the available data. Similar to bagging [5] we constructed a contrast classifier by aggregating the predictions of these neural networks through averaging. Additional benefit of using an ensemble of neural networks is that the averaging is known as a successful technique for increasing their accuracy by reducing variance while retaining low bias in prediction.

2.4 Knowledge Representation

For each sequence position, a set of relevant attributes was derived from statistics of a subsequence within a window of length W centered at the position. More specifically, given a sequence $\mathbf{s} = \{s_i, i = 1, \ldots, L\}$ of length L, for each sequence position s_i an appropriate M-dimensional attribute vector $\mathbf{x}_i = \{x_{ij}, j = 1, \ldots, M\}$ is constructed and a corresponding class label y_i is assigned. Thus, sequence $\mathbf{s}$ is represented as a set of L examples $\{(\mathbf{x}_i, y_i), i = 1, \ldots, L\}$.

Using a window of length $W = 21$, a total of 25 attributes were derived for each sequence position. These attributes were proved to be useful in various protein sequence analyses and structure prediction problems. The first 19 attributes were the amino acid frequencies within the window since it has been shown that PDB proteins exhibited unique amino acid composition patterns [6]. Only 19 of the 20 frequencies were used since the remaining one could be uniquely determined from the rest. Based on the amino acid frequencies, an attribute called *K2*-entropy [21] was calculated to measure local sequence complexity.

We also measured flexibility [19] and hydropathy [9] propensities obtained by triangular moving average window where center position had weight 1 and the most distant positions had weight 0.25. While window length was 21 for hydropathy attribute, it was only 9 for the flexibility attribute, as suggested by the previous study [19]. The final 3 attributes were outputs of the PHD secondary structure predictor [16], i.e., the prediction scores for alpha helix (H), beta strand (E) and loop (L). Finally, class labels 0 and 1 were assigned to examples from *PDB25* and *SwissRep*, respectively.

2.5 Using Contrast Classifiers to Explore Bias in PDB

Given a measure of contrast $cc(\mathbf{x})$ at each sequence position, we explored the bias in PDB towards numerous protein functional properties, as defined by SWISS-PROT keyword and feature classification. It was expected that the analysis would confirm known results (e.g. that transmembrane and low complexity regions are underrepresented in PDB) and point to some less-known sources of bias. For a set R of regions with a given functional property, mean and standard deviation of the corresponding $cc(\mathbf{x})$ were calculated to measure the direction and level of bias.

Additionally, the Kolmogorov- Smirnov goodness of fit test [11] (KS test) was used to measure the difference between the $cc(\mathbf{x})$ distributions of R and *PDB25*. The KS test measures the maximum absolute difference between the empirical cumulative distributions of the two samples and uses it to estimate the test p-value. Since $cc(\mathbf{x})$ of neighboring sequence positions are correlated due to the use of window W (=21) in attribute construction (see Section 2.4), we estimated the effective length as $L_{eff} = 1 + (L-1)/W$ for each sequence region of length L and used it in calculation of the KS test p-values.

3 Results and Discussions

3.1 Training Contrast Classifier

We built the contrast classifier as an ensemble of 50 neural networks each having 5 hidden neurons and 1 output neuron with sigmoid activation function. To reduce bias towards long sequences, a balanced training set for each neural network was selected in two steps: (a) 20 examples sampled randomly without replacement were taken from each sequence in *PDB25* and *SwissRep*, and (b) a balanced set of 8,000 examples was sampled randomly with replacement from the resulting set. Individual neural networks were trained with the backpropagation algorithm. To avoid overfitting, 80% of the balanced set was used for training and the rest was reserved to signal the training termination. If the training was not stopped after 300 epochs it was terminated automatically.

3.2 Distributions of Contrast Classifier Outputs

Comparing contrast classifier outputs at *PDB25* and *SwissRep*. A trained contrast classifier was applied to both *PDB25* and *SwissRep* sequences, and their *cc*(**x**) distributions were compared in Figure 1(a). Since *SwissRep* contained a number of *PDB25* sequences and/or their homologues, it was expected that the two distributions would overlap. However, a considerable proportion of *SwissRep* sequences had relatively large *cc*(**x**) values (e.g., larger than 0.7) while most *PDB25* sequences had smaller *cc*(**x**) values concentrated around 0.47. This result clearly illustrated the existence of bias in PDB. In the following subsections, we analyze the sources of bias in greater detail.

We also examined the distributions of another two sets in Figure 1(a): *PDB25H* of homologues of *PDB25* in *SwissRep*; and *PDB25NH* of the remaining sequences of *SwissRep*. The homologues were identified through 3 iterations of PSI-BLAST search using E-value thresholds of 0.001 for sequence inclusion in the profile and 1 for including sequences in the final selection. As expected, distribution for *PDB25H* was similar to *PDB25*, while distribution for *PDB25NH* was similar to *SwissRep*.

Distributions of 3 specific sequence regions. We examined distributions of *cc*(**x**) for transmembrane regions, low complexity regions, and predicted disordered regions from *SwissRep* (see Section 2.1). As shown in Figure 1(b), all these regions exhibited *cc*(**x**) values significantly higher than *PDB25* sequences, indicating that they were highly underrepresented in PDB. As discussed in the Introduction, transmembrane regions are typically excluded from structural characterizations. Low complexity regions have biased amino acid composition involving a few amino acid types and they often do not fold into stable 3D structure [15]. Huntley and Golding [8] performed an extensive investigation on eukaryotic proteins in PDB and reported a

large deficiency in low complexity regions. Their results indicated that even for the few low complexity regions with structural data present in PDB, tertiary structures were missing in most cases. Predicted disordered regions [13] correspond to the regions very likely to have flexible structure that could not be captured by X-ray crystallography or NMR. Since disordered proteins are hard to crystallize, it was expected that they are underrepresented in PDB.

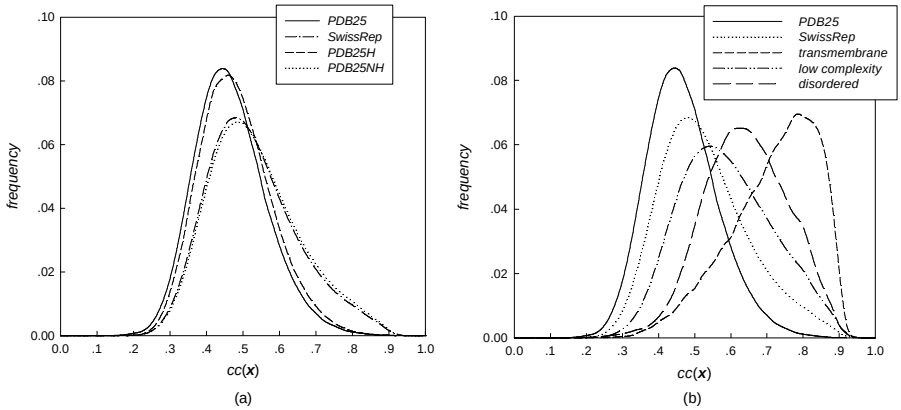

Figure 1. *Comparison of cc(x) distributions between PDB25 and other sets: (a) SwissRep, PDB25H and PDB25NH; (b) various regions of interest from SwissRep.*

Distributions of functional regions characterized by SWISS-PROT FT line. We extended the analysis to functional regions described by feature tables (FT lines in SWISS-PROT) with the FT keywords. Note that the length of functional regions could range from one (e.g. posttranslational modification sites) to a few hundred residues. In Figure 2 we plot the distributions of the 3 selected functional region types. The supplementary material with the plots of all functional regions listed in the FT lines can be accessed at *http://www.ist.temple.edu/disprot/PSB04*.

Given the explanation of the contrast classifier output discussed in Section 2.2, a positively skewed output distribution indicates that a certain type of functional site or region is underrepresented in PDB, while a negatively skewed output distribution indicates that it is overrepresented. For example, disulfide bonds (DISULFID) play important roles in stabilizing protein tertiary structure and thus should be abundant in PDB. Consistent with this fact is that their $cc(\mathbf{x})$ distribution is highly negatively skewed (see Figure 2). On the other hand, signal peptides (SIGNAL) are short segments of amino acids in a particular order that govern the transportation of newly synthesized proteins, and then cleaved from the matured proteins. Since structure characterization experiments usually target matured proteins, signal peptides are expected to be underrepresented in PDB. Accordingly, we observe a positively skewed distribution similar to that of transmembrane regions in Figure 1(b). Repeats

(REPEAT) are internal sequence repetitions and typically have low sequence complexity and thus exhibit a similar distribution to that of low complexity regions in Figure 1(b).

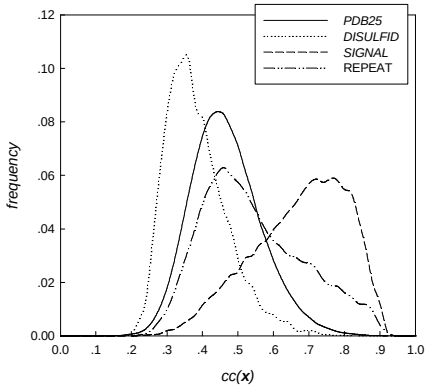

Figure 2. *Distributions of cc(**x**) of 3 selected sites or regions from SwissRep sequences.*

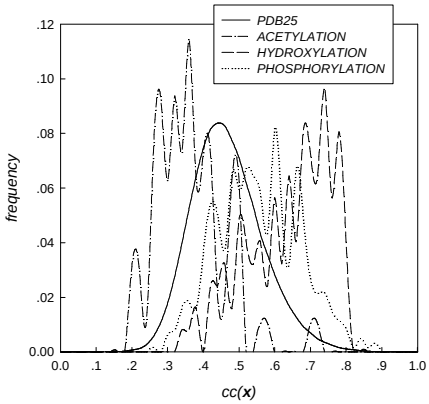

Figure 3. *Distributions of contrast classifier output cc(**x**) of 3 selected posttranslational modification sites from SwissRep sequences.*

Comparing distributions of PDB and different functional regions. The *cc*(**x**) distributions of the functional sites or regions were compared with the distributions of the *PDB25* sequences using the 2-sample Kolmogorov-Smirnov test described in Section 2.5. The FT keywords corresponding to these sites or regions were then ranked according to the resulting *p-values*, as shown in Table 2. Note that the table does not list FT keywords ACT_SITE, CA_BIND, CONFLICT, INIT_MET, LIPID, MUTAGEN, SIMILAR, SITE, THIOLEST, TRANSIT, NON_CONS, NON_TER, NOT SPECIFIED, NP_BIND, UNSURE, VARIANT, and VARSPLIC since they were either of less interest or their total effective length was less than 1000 residues. Also shown in Table 2 are means and standard deviations of *cc*(**x**) values, and the total effective length used in the KS test.

We further examined contrast classifier output *cc*(**x**) on different posttranslational modification sites identified by FT keyword MOD_RES. Results for the 5 most frequent sites are shown in Table 3. Similar to Table 2, these sites were ranked according to their Kolmogorov-Smirnov test *p-values* when compared to the distribution of *PDB25* sequences. Among the top 3 sites, phosphorylation and hydroxylation sites have positively skewed distributions, while acetylation sites have negatively skewed distribution, as shown in Figure 3. This suggests that the first 2 modification sites are underrepresented in PDB, while the acetylation sites are overrepresented.

Table 2. *Comparison of distributions of contrast classifier outputs on sites or regions of interest and PDB25 sequences. The p-values were obtained using the Kolmogorov-Smirnov 2-sample test.*

FT keyword	***p-value***	$\mu(cc)$	$\sigma(cc)$	L
DISULFID	0	0.38	0.09	11840
SIGNAL	0	0.67	0.13	4941
TRANSMEM	0	0.72	0.12	20531
REPEAT	0	0.53	0.14	13377
DOMAIN	7.93e-318	0.51	0.12	82912
CARBOHYD	1.17e-138	0.51	0.12	7015
CHAIN	5.09e-118	0.50	0.12	74737
MOD_RES	3.57e-061	0.53	0.13	1515
PEPTIDE	7.24e-061	0.53	0.13	1225
PROPEP	1.08e-060	0.51	0.12	2045
METAL	3.86e-040	0.45	0.10	1584
DNA_BIND	1.62e-021	0.50	0.09	1186
ZN_FING	5.76e-019	0.44	0.09	1465
STRAND	1.68e-018	0.45	0.10	4619
HELIX	5.42e-015	0.49	0.09	3989
BINDING	1.87e-005	0.48	0.12	4075
TURN	9.98e-005	0.47	0.09	4556
PDB25		0.47	0.10	17194

$\mu(cc)$ – mean of $cc(\mathbf{x})$, $\sigma(cc)$ – standard deviation of $cc(\mathbf{x})$, L – effective number of residues

Table 3. *Comparison of contrast classifier output distributions of different posttranslational modification sites with that of PDB25 sequences.*

modification site	***p-value***	$\mu(cc)$	$\sigma(cc)$	L
phosphorylation	3.54e-054	0.55	0.11	608
hydroxylation	1.72e-036	0.64	0.12	124
acetylation	8.84e-016	0.37	0.10	96
amidation	5.68e-015	0.55	0.14	170
methylation	8.66e-010	0.57	0.12	93
PDB25		0.47	0.10	17194

$\mu(cc)$ – mean of $cc(\mathbf{x})$, $\sigma(cc)$ – standard deviation of $cc(\mathbf{x})$, L – effective number of residues

Distributions of SCOP structural classes. According to the SCOP database [12] (release 1.61, Nov. 2002), 1,685 out of the 1,824 chains in *PDB25* can be classified into 11 structural classes, as shown in Table 4. Note that different parts of a chain might belong to different classes. We examined the $cc(\mathbf{x})$ distributions of individual structural classes and compared them with the overall distribution of *PDB25* sequences using the Kolmogorov-Smirnov test (results shown in Table 4). The most significant difference corresponded to sequences from the *small* class with a negatively skewed distribution. It is worth noting that membrane and cell surface, coiled coils, and peptide structural classes appeared to be significantly underrepresented in *PDB25*.

Table 4. *Comparison of cc*(**x**) *distributions on PDB25 proteins from different fold classes with that of all PDB25 proteins.*

Fold Class	***p-value***	***m(cc)***	***s(cc)***	***L***
small	8.74e-037	0.41	0.10	608
alpha	4.65e-018	0.49	0.10	2779
membrane and cell surface	3.86e-015	0.53	0.14	382
beta	1.82e-010	0.45	0.10	3319
coiled coils	7.71e-007	0.50	0.10	173
peptides	0.000585	0.52	0.12	85
designed	0.007401	0.55	0.10	17
alpha+beta	0.053422	0.46	0.09	3238
alpha_beta	0.272922	0.47	0.09	4210
multidomain	0.808058	0.47	0.09	469
low resolution	0.92231	0.47	0.10	47
PDB25		0.47	0.10	17194

m(cc) – mean of *cc*(**x**), ***s(cc)*** – standard deviation of *cc*(**x**), ***L*** – effective number of residues

Analysis of underrepresented proteins. Complementing the study of *cc*(x) distributions of different functional protein regions or protein types, we explored the properties of proteins that are most highly underrepresented by *PDB25*. Some of these proteins are arguably the most interesting targets for future structural determination experiments. For this study, each *SwissRep* sequence **s** was represented with a single number *cc_avg*(**s**) representing the average *cc*(**x**) over the sequence. A total of 2,814 (or 17.2% of) *SwissRep* sequences having *cc_avg*(**s**) > 0.597 were selected with the threshold chosen such that only 1% of *PDB25H* sequences satisfied the inequality. We analyzed the properties of the resulting set, called *SwissOut*, by comparing the commonness of different SWISS-PROT keywords (KW line) in *SwissOut* and *PDB25H* (see Section 3.2). By denoting $f_{SwissOut}$ and f_{PDB25H} as frequencies of proteins with a given keyword among *SwissOut* and *PDB25H*, respectively, their difference can be quantified by measuring the *Z-score* defined as $(f_{SwissOut}-f_{PDB25H})/\mathrm{sqrt}(f_{PDB25H}(1-f_{PDB25H})/N)$, where N is the number of proteins in *SwissOut*.

Table 5. *The top 6 SWISS-PROT keywords associated with underrepresented sequences.*

Keyword	$f_{SwissRep}$ **[%]**	f_{PDB25H} **[%]**	$f_{SwissOut}$ **[%]**	***Z-score***
hypothetical protein	42.38	15.64	54.34	56.52
transmembrane	17.55	14.51	42.89	42.74
complete proteome	31.64	18.24	32.55	19.66
inner membrane	1.49	1.45	3.62	9.64
chloroplast	1.36	1.22	3.13	9.23
chromosomal protein	0.66	1.29	2.74	6.82

In Table 5 we list 6 SWISS-PROT keywords with the highest Z-scores among the ones represented with more than 50 *SwissOut* proteins. By careful examination, it is evident that the obtained results are reasonable, and are another indication of a potential of the proposed contrast classifier approach. Furthermore, it is likely that *SwissOut* proteins with keyword “complete proteome” would be very interesting structural targets.

4 Concluding Remarks

We applied the contrast classifier to explore the bias existing in the Protein Data Bank towards different functional protein properties. Assuming SWISS-PROT is a representative of the protein universe while the PDB is a biased sample, we trained a contrast classifier with the non-redundant subsets of PDB and SWISS-PROT and used its output to analyze the bias in PDB. Comparing to other methods for examining bias in PDB (see the Introduction), the main strength of our approach is that it provides a quantitative measure to assess the bias in a uniform way.

Our results confirmed some well-known facts such as the lack of transmembrane, low complexity and disordered regions among PDB sequences. They have also revealed some less recognized facts such as depletion of PDB in phosphorylation and hydroxylation modification sites and overrepresentation in acetylation sites. These results are a strong indication that contrast classifiers should be considered as an attractive tool for selection of target proteins for future structural characterization experiments.

There are several immediate avenues of future research. As shown by our results, contrast classifier trained with attributes derived from simple statistics over a local window was able to successfully explore the bias in PDB. This suggests that more sophisticated choice of attributes could provide an additional insight into the sources of bias. Similarly, removing well-known underrepresented regions (e.g. transmembrane, low complexity) before the training of the contrast classifier would allow better focus on the less known sources of bias in PDB. Finally, by a slight extension of the proposed methodology contrast classifiers could be trained with sequences of known folds vs. sequences in SWISS-PROT. This could have a potential in detecting the sequences with potentially novel fold structures.

References

1. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, "The Protein Data Bank", *Nucleic Acids Res.*, **28**, 235 (2000).
2. B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003", *Nucleic Acids Res.*, **31**, 365 (2003).
3. S.E. Brenner, "Target selection for structural genomics", *Nat. Struct. Biol., Structural Genomics supplement*, **7**, 967 (2000).
4. S.E. Brenner, C. Chothia and T.J. Hubbard, "Population statistics of protein structures: lessons from structural classifications", *Curr. Opin. Struct. Biol.*, **7**, 369 (1997).
5. L. Breiman, "Bagging predictors", *Mach. Learning*, **24**, 123 (1996).

6. M. Gerstein, "How representative are the known structures of the proteins in a complete genome? A comprehensive structural census", *Fold Des.*, **3**, 497 (1998).
7. U. Hobohm and C. Sander, "Enlarged representative set of protein structures", *Protein Sci.*, **3**, 522, (1994).
8. M.A. Huntley and G.B. Golding, "Simple sequences are rare in the Protein Data Bank", *Proteins: Struc. Funct. Gen.*, **48**, 134 (2002).
9. J. Kyte and R.F. Doolittle, "A simple method for displaying the hydropathic character of a protein", *J. Mol. Biol.*, **157**, 105 (1982).
10. J. Liu and B. Rost, "Target space for structural genomics revisited", *Bioinformatics*, **18**, 922 (2002).
11. F.J. Massey Jr., "The Kolmogorov-Smirnov test of goodness of fit", *J. Amer. Statist. Assoc.*, **46**, 68 (1951).
12. A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures", *J. Mol. Biol.*, **247**, 536 (1995).
13. Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. Brown and A.K. Dunker, "Predicting intrinsic disorder from amino acid sequence", *Proteins: Struc. Funct. Gen., Special Issue on CASP5*, in press.
14. K. Peng, S. Vucetic, B. Han, H. Xie and Z. Obradovic, "Exploiting unlabeled data for improving accuracy of predictive data mining", In *Proc. Third IEEE Int'l Conf. on Data Mining*, Novemember 2003, Melbourne, FL, in press.
15. P. Romero, Z. Obradovic, X. Li, E. Garner, C.J. Brown and A.K. Dunker, "Sequence complexity and disordered protein", *Proteins: Struc. Funct. Gen.*, **42**, 38 (2001).
16. B. Rost, "PHD: predicting one-dimensional protein structure by profile-based neural networks", *Methods Enzymol.*, **266**, 525 (1996).
17. B. Rost, "Twilight zone of protein sequence alignments", *Protein Eng.*, **12(2)**, 85 (1999).
18. H. Sakai and T. Tsukihara, "Structures of membrane proteins determined at atomic resolution", *J. Biochem.* **124**, 1051 (1998).
19. M. Vihinen, E. Torkkila and P. Riikonen, "Accuracy of protein flexibility predictions", *Proteins: Struc. Funct. Gen.*, **19**, 141 (1994).
20. S. Vucetic, D. Pokrajac, H. Xie and Z. Obradovic, "Detection of underrepresented biological sequences using class-conditional distribution models", In *Proc. Third SIAM Int'l Conf. on Data Mining*, May 2003, San Francisco, CA.
21. J.C. Wootton, S. Federhen, "Analysis of compositionally biased regions in sequence databases", *Methods Enzymol.*, **266**, 554 (1996).
22. G. Yona, N. Linial and M. Linial, "ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space", *Proteins*, **37**, 360 (1999).