Bioinformatics Advance Access published April 25, 2014

Redundancy-weighting for better inference of protein structural features

Chen Yanover [1,*], Natalia Vanetik [2], Michael Levitt [3], Rachel Kolodny [4], and Chen Keasar [5,*]

[1]Machine Learning for Healthcare and Life-Sciences, IBM Research Laboratory, Haifa, Israel; [2]Department of Software Engineering, Shamoon College of Engineering, Israel; [3]Stanford University School of Medicine, Stanford, CA; [4]Department of Computer Science, University of Haifa, Mount Carmel, Haifa, Israel; [5]Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Associate Editor: Prof. Anna Tramontano

ABSTRACT

Motivation: Structural knowledge, extracted from the Protein Data Bank (PDB), underlies numerous potential functions and prediction methods. The PDB, however, is highly biased: many proteins have more than one entry, while entire protein families are represented by a single structure, or even not at all. The standard solution to this problem is to limit the studies to non-redundant subsets of the PDB. While alleviating biases, this solution hides the many-to-many relations between sequences and structures. That is, non-redundant data-sets conceal the diversity of sequences that share the same fold and the existence of multiple conformations for the same protein. A particularly disturbing aspect of non-redundant subsets is that they hardly benefit from the rapid pace of protein structure determination, as most newly solved structures fall within existing families.

Results: In this study we explore the concept of redundancy-weighted data-sets, originally suggested by Miyazawa and Jernigan. Redundancy-weighted data-sets include all available structures and associate them (or features thereof) with weights that are inversely proportional to the number of their homologs. Here, we provide the first systematic comparison of redundancy-weighted data-sets with non-redundant ones. We test three weighting schemes and show that the distributions of structural features that they produce are smoother (having higher entropy) compared with the distributions inferred from non-redundant data-sets. We further show that these smoothed distributions are both more robust and more correct than their non-redundant counterparts.

We suggest that the better distributions, inferred using redundancy-weighting, may improve the accuracy of knowledge-based potentials, and increase the power of protein structure prediction methods. Consequently, they may enhance model-driven molecular biology.

Contact: cheny@il.ibm.com, chen.keasar@gmail.com

1 INTRODUCTION

Mining the riches of experimentally-determined data in the Protein Data Bank (PDB) Berman *et al.* (2013a,b); Bernstein *et al.* (1977) has been a major source of structural knowledge over that last four decades. In a reverse engineering-like fashion it allows the derivation of rules and methods for the prediction of secondary structure (Chou and Fasman, 1974; Garnier *et al.*, 1978; Rost, 1996; McGuffin *et al.*, 2000), solvent accessibility (Rost, 1996; Karplus, 2009), and trans-membrane regions (Rost, 1996; McGuffin *et al.*, 2000); as well as knowledge-based potentials (Tanaka and Scheraga, 1976; Sippl, 1993; Lüthy *et al.*, 1992; Eisenberg *et al.*, 1997; Samudrala and Moult, 1998; Summa and Levitt, 2007). These methods and potentials have had a considerable impact on our understanding of the protein universe, and accelerated progress in biology, chemistry and medicine. This study aims to enhance these data mining efforts by attacking their major obstacle, data redundancy.

The underlying premise behind data mining of protein structures is that recurring patterns may result from physical aspects of protein folding and stability (e.g., the hydrophobic effect) (Miyazawa and Jernigan, 1985). However, the data, which is available for mining, is not a uniform sample of sequence and structure spaces. Certain folds (e.g., TIM barrels and G-protein-coupled receptors) are far more abundant than others in genomes (see Goldstein (2008) for some suggestions of why it is so). The PDB content is further skewed by research interests of the contributing experimentalists and by methodological constraints. This bias, often referred to as the "PDB redundancy", may amplify or diminish the signal of recurring patterns. For example, the stabilizing effect of the beta-alpha-beta supersecondary structure may be overestimated due to the abundance of folds that include it (e.g., TIM barrels).

The common solution to data redundancy is to use a non-redundant subset of the PDB, composed of family representatives. Kabsch and Sander (1983) pioneered this solution using a 62-membered subset of the 75 high-quality entries of the PDB, leaving out homologs with sequence identity of 50% or higher (a rather promiscuous threshold by current standards). This approach has been adopted by numerous studies, and became the field's norm, with publicly-available and standardized tools for data culling (Hobohm and Sander, 1994; Wang and Dunbrack, 2003, 2005). Yet, notwithstanding the evident utility of non-redundant PDB subsets, they have inherent limitations in scalability and descriptive power. First and foremost, they do not benefit from

*to whom correspondence should be addressed

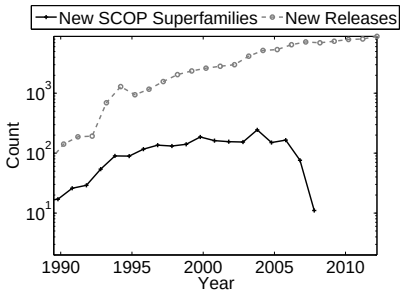

Fig. 1. The rapid pace of protein structure determination hardly affects the size of non-redundant databases. The PDB growth rate (gray) has accelerated in recent years but the rate of new SCOP superfamily reports (black) dwindled in the last three releases of the database (versions $1.71 - 1.75$).

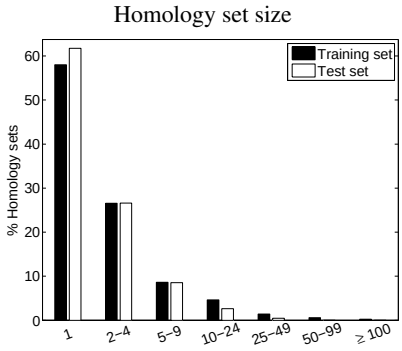

Fig. 2. Homology subset sizes for the polypeptide structures in the training (black) and test (white) sets.

the rapid growth of the PDB since almost all new entries are mapped to already known folds (Figure 1; and see also Levitt (2009)). More importantly, non-redundant data-sets conceal much of the complexity of protein universe. Specifically, they hide the compatibility of diverse sequences with the same fold and the flexibility of protein structures (Kosloff and Kolodny, 2008). Our working hypothesis is that this oversimplified perspective manifests itself in artificially "bumpy" distributions of the measurable features.

An important alternative was presented more than a decade ago by Miyazawa and Jernigan (1996, 1999), who weighted structures in their knowledge-based potentials. In those studies they considered all the PDB structures (of sufficient quality and length), yet assigned non-uniform weights to protein chains, so that the weights of chains with many homologs are scaled down. Thus, all the data is considered, yet biases are alleviated. Somewhat surprisingly, we are unaware of any recent studies that use this approach. Further, to the best of our knowledge, no study has yet compared its performance with the standard, representative subset, approach.

Interestingly, the redundancy, which is removed from structural data-sets, is valuable when investigating families of homologous sequences. There, evolutionary conserved patterns characterize a family and deviations from these patterns shed light on the uniqueness of specific members. The distribution of sequence similarities across a family is typically uneven; that is, there may be subsets of sequences which are more closely related. This might bias the analysis towards patterns that appear in such highly similar subsets. In this context, however, a non-redundant subset (i.e., one without recognizable similarities) would consist of a single sequence, and be devoid of any information. Instead, scholars have successfully used various weighting schemes that down-scale contributions from large subsets of sequences (see Altschul *et al.* (1997), and references therein), most notably for multiple sequence alignment (Thompson *et al.*, 1994) and sequence search (Altschul *et al.*, 1997). Structural data mining is a very different computational task: most importantly, it does not focus on a single protein family but rather must deal with multiple families and account for their varying sizes. Nonetheless, the success of including more, and even all, available proteins in the context of search and alignment tasks, suggests that similar approaches may be valuable in structural data mining as well.

Our study revisits the redundancy-weighting approach and provides the first systematic assessment of its correctness and robustness. To this end, we compare inter-atomic distance distributions, a central component in knowledge-based potentials, sampled from either non-redundant or redundancy-weighted data-sets; for the sake of completeness, we also consider distributions that were sampled from an unweighted, redundant data-set. We estimate the complexity of these distributions by their entropy and observe that the redundancy-weighted data-sets have a higher entropy than their non-redundant counterparts. We further demonstrate that the higher entropy distributions are more correct and robust. Our observations suggest that structure prediction methods could benefit considerably from training on redundancy-weighted data-sets rather than on non-redundant ones. This in turn, can improve our understanding of the forces that shape the protein structure universe.

2 METHODS

2.1 Structure data-set

Our data-set includes 7,307 structures of polypeptide chains (in 6,956 PDB files), at least 40 residues long, solved at 1.5Å resolution or better. To allow a straightforward definition of local structural features, we broke each chain into (non-overlapping) segments of peptide-bonded amino acids, where two consecutive residues are considered "bonded" if the distance between their Cα atoms is within 10% of the canonical peptide bond distance (2.95 or 3.8Å for cis and trans peptide bond, respectively) and none is listed as "missing" (in PDB Remark 465, missing residues, or Remark 470, missing atoms). The resulting set consists of 8,876 segments longer than 20 residues.

This data is split into training and test subsets. The training subset consists of 6,299 structures (7,635 segments) released by the end of 2011, whereas the remaining 1,008 chain structures (1,241 segments), released from January 2012 to April 2013, constitute the test set. We compare the feature distributions obtained from the larger training set to those inferred using the smaller test set.

This study focuses on Cα distance distributions of the 441 possible type pairs (all combinations over 20 amino acids + the wild-card amino acid X) of residues separated by predefined distances along the sequence (e.g., 20 residues apart). We compare distributions that were inferred using various weighting schemes, and over different sets of protein structures. In general, the reliability of distributions improves as the number of instances from which they were derived, grows. Figure 4D depicts the number of residue pairs (20 residues apart) in the training set. Note the two orders of magnitude difference between the rarest pair (two Tryptophan residues, 304 instances) and the most common pair (two Leucine residues, 10,163).

2.2 Weighting schemes

For each chain we maintain a list of homologs and their aligned regions, as found by the FASTA (Pearson and Lipman, 1988) sequence alignment

tool (with e-value $\leq 10^{-4}$). The sets of chains and homology assignments constitute the vertices and edges, respectively, of a neighbors graph (Figure 3). We consider connected components of this graph as homology subsets; the size distributions of these subsets in the training and test sets are shown in Figure 2. We follow a common practice (see, for example, Li and Godzik (2006); Bull *et al.* (2013)) and generate a non-redundant data-set by picking a random representative from each homology subset. In contrast, redundant and redundancy-weighting data-sets use all available structures.

In this study we consider five schemes (Figure 3): non-redundant, redundant, and three ways for redundancy-weighting. The standard, non-redundant (*NR*) scheme effectively assigns a weight of 1 to each feature instance (here, a distance between two Cα atoms separated by a certain number of residues) from a representative subset of structures. The redundant (*R*) scheme, too, effectively associates a weight of 1 to each feature, but includes all structures in the set.

The redundancy-weighting schemes use all structures available in the data-set, while balancing the contributions of protein families with many PDB entries and ones that have fewer. To this end, these schemes define the weight of a feature instance to be inversely proportional to the number of its homologs.

We implement three redundancy-weighting schemes. The first, per-sequence weighting scheme (*RWs*) associates each structure s with a set of close neighbors, ones with sequence similarity $\geq$ 40% and coverage $\geq$ 50%. The structure's weight, which is assigned to all its features, is $1/\left(N(s)+1\right)$, where $N(s)$ is the number of these neighbors. The second weighting scheme, *RWf*, assigns the weights individually to each feature instance. The weight of a structural feature, in this scheme, is inversely proportional to the number of homologous chains that align with all of its positions, without gaps. The last scheme, *MJ*, follows Miyazawa and Jernigan (1996, 1999) sample weighting. Briefly, sequence identity between each pair of structures is computed as the fraction of identical residues in their alignment (Needleman and Wunsch, 1970); eigenvalue decomposition of the sequence identity matrix defines per-sequence weights which are approximately equal to the inverse of the number of similar sequences. This weight is assigned to all the features of the structure.

2.3 Performance assessment

We quantify the similarity between two distance distributions, R and Qm by their Jensen-Shannon divergence (JSD) defined as:

$$JSD(R;Q) = \frac{1}{2}\left(\text{dKL}(R,M) + \text{dKL}(Q,M)\right) \quad (1)$$

where $M = \frac{1}{2}(R+Q)$ and dKL is the Kullback-Leibler divergence (Lin, 1991, eq. 4.1 and setting R and Q's weights to 0.5). For each feature distribution (e.g., Euclidean distances between Tryptophan Cα at position i and Histidine Cα at position $i+20$; see Figure 4A-C), we define the *robustness* of a weighting scheme, $\mathcal{W}$, as the complement of the Jensen-Shannon divergence between $\mathcal{W}$-induced feature distributions in the training ($P_{\mathcal{W}(\text{Train})}$) and test ($P_{\mathcal{W}(\text{Test})}$) sets:

$$\text{Robustness}(\mathcal{W}) = 1 - JSD\left(P_{\mathcal{W}(\text{Train})}; P_{\mathcal{W}(\text{Test})}\right) \quad (2)$$

Similarly, the *correctness* of a weighting scheme, for a given feature distribution, is the complement of the divergence between the scheme's test distribution and the standard, non-redundant (*NR*) feature distribution, inferred over the training set:

$$\text{Correctness}(\mathcal{W}) = 1 - JSD\left(P_{NR(\text{Train})}; P_{\mathcal{W}(\text{Test})}\right) \quad (3)$$

We also compute the Shannon entropy of $\mathcal{W}$-induced feature distribution, $P_{\mathcal{W}}$:

$$\text{Entropy}(P_{\mathcal{W}}) = -\sum_j p_j \cdot log_2(p_j), \quad (4)$$

where j ranges over all distribution bins.

Finally, we define the gain in either correctness, robustness, or entropy, as the increase in that measure obtained using weighting scheme $\mathcal{W}_1$ compared to $\mathcal{W}_2$; for example:

$$\text{Correctness gain}(\mathcal{W}_1,\mathcal{W}_2) = \text{Correctness}(\mathcal{W}_1) - \text{Correctness}(\mathcal{W}_2). \quad (5)$$

3 RESULTS

Distributions of inter-atomic distances lie at the heart of knowledge-based potentials and as such are among the most studied features in PDB data mining. Here, we compare data-mining of such distances from three types of data-sets: redundant, non-redundant, and redundancy-weighted. The redundant training and test data-sets include all proteins that were solved before and after the end of 2011, respectively, at 1.5Å resolution or better. The training and test non-redundant data-sets are subsets of the redundant ones, containing a single representative from each set of homologous structures. Finally, the three pairs (training and test) of redundancy-weighted data-sets include the same structures as their redundant counterparts but the contributions of these structures, or features thereof, are weighted. The three redundancy-weighting schemes (see Methods and Figure 3) are: (1) a scheme that assigns a single weight to each polypeptide chain – and all the feature instances computed from its structure – based on the number of its homologs (*RWs*); (2) a scheme that computes a per-feature weight based only on homologs that cover all the positions in a feature (*RWf*); and (3) the algebraic sample weighting scheme of Miyazawa and Jernigan (1996, 1999) (*MJ*).

We focus on distributions of Cα-Cα distances, as a proof of concept. Figure 4, upper panel, shows training and test distributions of Cα distances between three types of amino acid pairs located 20 positions apart along the sequence: Tryptophan (W) and Histidine (H, left), Tryptophan and Tyrosine (Y, center), and Tryptophan and Leucine (L, right). We compute the distributions using the standard, non-redundant scheme (*NR*) or a redundancy-weighting scheme, *RWf*. The insets in Figure 4A-C depict the similarity of these distributions. We assess the performance of a weighting scheme $\mathcal{W}$ by two quantities: *correctness*, which measures how similar is $\mathcal{W}$'s test distribution to the non-redundant distribution over the training set (inset's top row); and *robustness*, which measures is how similar are $\mathcal{W}$'s training and test distributions (inset's diagonal; see Methods for formal definitions).

The lower panel of Figure 4 shows the correctness of Cα-Cα distance distributions over all pairs of amino acids when using a non-redundant data-set and a redundancy-weighted one. Predictably, correctness for both schemes is lower for amino acid pairs with a relatively small number (i.e., hundreds) of samples (top left corner of both heat-maps) compared to pairs with many more (i.e., tens and hundreds of thousands) samples (see Figure 4D for sample counts of each amino acid pair); the Spearman's rank correlation between the number of samples and the correctness of *NR* and *RWf* is 0.96, P-value $< 10^{-100}$. Evidently, the distance distributions derived using the redundancy-weighted scheme are, for the most part, more correct than those inferred using the non-redundant scheme (Figure 4G); out of 441 distributions, one for each amino acid pair, 422 (95.7%) redundancy-weighted distributions are more correct. Moreover, the improvement in correctness, or the *correctness gain*, is greater for amino acid pairs with lower number of samples (Spearman's rank correlation $= -0.7$, P-value $< 10^{-65}$) for which, as we indicated above, the correctness is poorer.

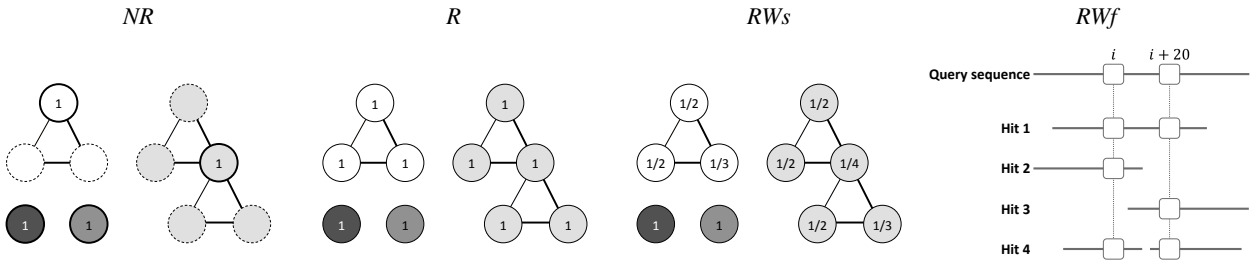

Fig. 3. Weighting schemes. A homology graph represents the similarity relations (edges) between chain sequences (nodes); each homology set, i.e., a connected component in this graph, is depicted in a different shade. The non-redundant (*NR*) and redundant (*R*) schemes use uniform weights (shown inside the circles), assigned to a set of representatives and to all available structures, respectively. The two redundancy-weighting schemes, *RWs* and *RWf*, associate each feature (e.g., a specific pair of Cα atoms, 20 residues apart), with a weight, which is inversely proportional to its redundancy. *RWs* considers redundancy at the chain level, that is all the features of a given chain have the same weight, which is inversely proportional to the number of chains with which it has considerable sequence similarity and coverage (schematically illustrated as thicker lines). In contrast, *RWf* considers redundancy at the feature level. Two features are considered redundant if they are part of an alignment of high identity level and are both complete and continuous. For example, the query feature instance shown in the illustration is given a weight of 0.5 as its only counterpart is Hit 1.

A more complete picture is presented in Figure 5, which compares the standard non-redundant scheme with the four nonstandard ones, over three performance criteria: correctness, entropy, and robustness. As expected, the non-redundant Cα_i – Cα_{i+20} distributions are significantly more correct (one-tailed, paired-sample Wilcoxon signed rank test P-value $< 10^{-6}$) and associated with greater entropy (P-value $< 10^{-16}$) than the corresponding redundant distributions. This agrees with the common practice of preferring non-redundant data-sets to redundant ones. Yet, importantly, all three redundancy-weighting schemes perform significantly better than the non-redundant scheme in terms of correctness and robustness, and obtain higher entropy scores; for example: the median improvement in correctness using *RWs* compared to *NR* is 0.89%, P-value $< 10^{-68}$.

Finally, Figure 6 compares the performance of all weighting schemes over Cα distance distributions of residues 5, 10, 20, 50, and 100 amino acids apart. In all these cases, and consistent with the results for the Cα_i – Cα_{i+20} distributions (Figure 5), all three redundancy-weighting schemes perform better in terms of their correctness, entropy, and robustness than the standard non-redundant and redundant alternatives. Among the redundancy-weighting schemes, *RWf* is the most correct in 4 out of 5 amino acid distances and *MJ* significantly more correct than *RWs* for Cα atoms 50 and 100 amino acids apart; *RWf*'s entropy values are the greatest for the longer amino acid ranges and *MJ*'s for the shorter ones; and, consistently, *MJ* is the most robust, followed by *RWf*. Comparing the non-redundant and redundant distributions reveals that the latter is always more robust, while the former is more correct in all but the longest range (where *R* is more correct, P-value = 0.0002); and obtains greater entropy values in all but the shortest range. Indeed, these results show that the redundancy-weighting performs better than the widely-used non-redundant data-sets, in all these measures.

4 DISCUSSION

The dominant approach for data mining of the PDB is to extract the knowledge from a relatively small subset of non-homologous structures and consider their homologs (all other structures) as redundant and, thus, non-informative. Although this approach circumvents much of the inherent biases in the PDB, it also artificially reduces the variability of the structural landscape. An alternative approach, which we term here *redundancy-weighting*, considers all available structures but assigns them (or features thereof) lower weights proportionally to the number of homologs they have in the data-set.

Redundancy-weighting, originally proposed by Miyazawa and Jernigan almost two decades ago (Miyazawa and Jernigan, 1996), has been rarely used since and, to the best of our knowledge, never been systematically benchmarked against the standard approach. Here, we compare the correctness, complexity, and robustness of feature distributions, which were inferred using non-redundant and redundant data-sets, and three redundancy-weighting schemes: the algebraic sample weighting of Miyazawa and Jernigan (*MJ*) (Miyazawa and Jernigan, 1996), as well as our novel per-sequence (*RWs*) and per-feature (*RWf*) weightings. To this end, we quantify and compare the correctness, complexity, and robustness of distance distributions, which were derived from training and test data-sets according to these schemes.

Our most significant contribution is demonstrating that the three redundancy-weighting schemes outperform the "standard" non-redundant approach in all tested metrics. The *MJ* scheme, which is the most robust, is probably not scalable enough for practical use. This scheme requires an eigenvalue decomposition of an all-against-all similarity matrix. As the PDB is quickly approaching the 100,000 structures milestone, such decomposition becomes challenging both in terms of the computational requirements (space and time), and the numerical stability (Heath, 2002). *RWf*, which is the most correct scheme, has a milder limitation: it requires re-computing of weights for each structural feature. The *RWs* scheme is less accurate and robust than the other two, but does not suffer from these limitations. Thus, it may serve as a simple, off-the-shelf, general weighting scheme, which performs better than non-redundant data-sets.

Here, we focused on a relatively small set (only a few thousands) of the PDB's most accurately solved structures (with resolution better than 1.5Å). Focusing on this set had several advantages: Most importantly, the computational cost of exploring alternative weighting schemes is tractable, and in particular we could easily calculate the eigenvalue decomposition needed for the MJ weighting scheme. Notice that homology is not too common within this set, which includes more than 50% singletons (see Figure 2). Thus, one could expect that refining the weighting scheme will not have any impact on the accuracy and robustness when inferring structural

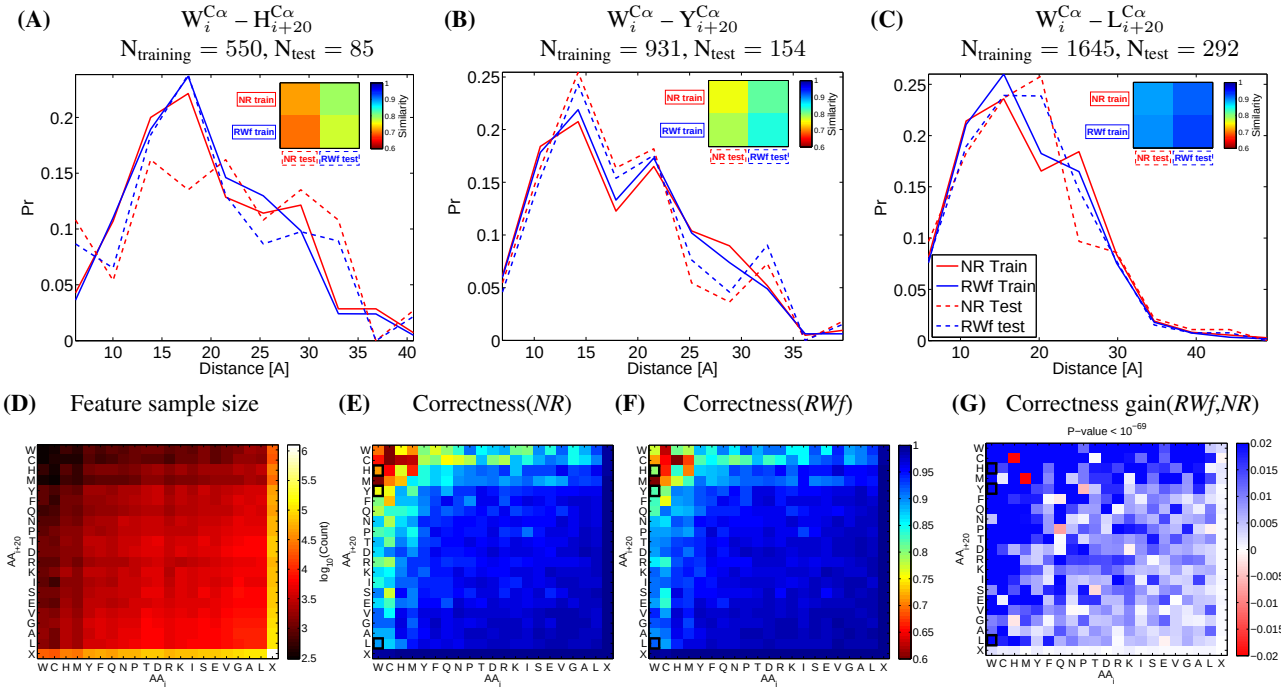

Fig. 4. Upper panel: Comparing feature distributions. Cα-Cα distance distributions between Tryptophan ($W_i^{C\alpha}$) and Histidine ($H_{i+20}^{C\alpha}$, A), Tryptophan and Tyrosine ($Y_{i+20}^{C\alpha}$, B) or Tryptophan and Leucine ($L_{i+20}^{C\alpha}$, C) 20 positions apart calculated using non-redundant (*NR*, red) and redundancy-weighted (*RWf*, blue) schemes over the training (solid) and test (dashed) sets; amino acid pairs are ordered, from left-to-right, by the number of their instances in these sets (as indicated above each plot). The insets show the similarity of the specified distributions, where hotter colors indicate greater divergences. Weighting scheme performance is described in terms of correctness and robustness. Correctness here is the similarity of a weighting scheme-induced test distributions (dashed) to the one inferred using the non-redundant training data-set (solid, red); these values are shown in the top row of each inset. Robustness indicates the similarity of a weighting scheme-induced distributions over the training and test sets (plots with the same color); these values are shown along the inset diagonals. Note the improvement (higher similarity) in both measures as the number of instances increases. Lower panel: **D**. The number of Cα pairs with 20 residues separation in the training set (note that the color scheme is exponential). The amino-acid types are ordered by prevalence in the Swiss-Prot data bank (from left to right and top to bottom); X denotes all residue types. **E-G**. Correctness of *NR* (E) and *RWf* (F) weighting schemes for all pairs of amino acids is shown as a heat-map; entries corresponding to distributions shown in the upper panel are highlighted (black boxes). Correctness gain, the difference between the correctness of *RWf* and *NR*, is positive (G, blue shades) for the the vast majority of amino acid pairs, indicating an improvement obtained using the former scheme; the one-tailed, paired-sample Wilcoxon signed rank test P-value is shown on the top.

features from the set. However, it does. Thus, we believe that such weighting schemes can similarly benefit others, even more so when considering larger subsets of the PDB, culled using more lax experimental quality thresholds.

While most residue-residue contact potentials (e.g., Miyazawa and Jernigan (1996)) consider "un-directional" residue pairs (thus combining pairs of two same residues with different orders to a single unique pair), we chose to focus on directional pairs in order to increase the dynamic range of the explored features (Figure 4D). Notably, the trends shown in Figure 6 are preserved when un-directional pairs are considered (data not shown).

Importantly, our study shows that the gain in accuracy and robustness is proportional to the rarity of the studied feature (Figs. 4 and 5). Current applications of PDB data mining (e.g., derivation of pairwise potentials and secondary structure prediction) are dominated by highly prevalent features (e.g., contacts between specific residues and the three major secondary structure elements). However, in more subtle applications (e.g., multi-body potentials (Gniewek *et al.*, 2011) and fragment prediction (Gront *et al.*, 2011; Kalev and Habeck, 2011; Shen *et al.*, 2013)) the number of relevant features grows while their prevalence in the non-redundant data-set drops. We speculate that shifting to the redundancy-weighting paradigm may be essential for the advancement of computational structural biology beyond pairwise potentials and prediction of simple features.

Although two of the redundancy-weighting schemes presented here are already useful, this study is mostly a proof-of-concept. A promising route for improvement is to take advantage of rapid structural search methods (Budowski-Tal *et al.*, 2010) and replace the currently used sequence alignments by more reliable and sensitive structural alignments. Another possible direction is to focus on domains rather than peptide chains, which often harbor several, repeating domains (e.g., Zn fingers) and are thus redundant by nature. Finally, the weighting scheme may apply some evolutionary model to gain better estimate of the inter-dependencies of homologous proteins (Thompson *et al.*, 1994).

In a wider context, redundant data-sets are abundant in other domains of knowledge, e.g., genomics, natural language processing and computer vision. While the current work focuses on protein structure data-sets, we hope insights obtained in this domain can have impact on data mining in other, unrelated ones.

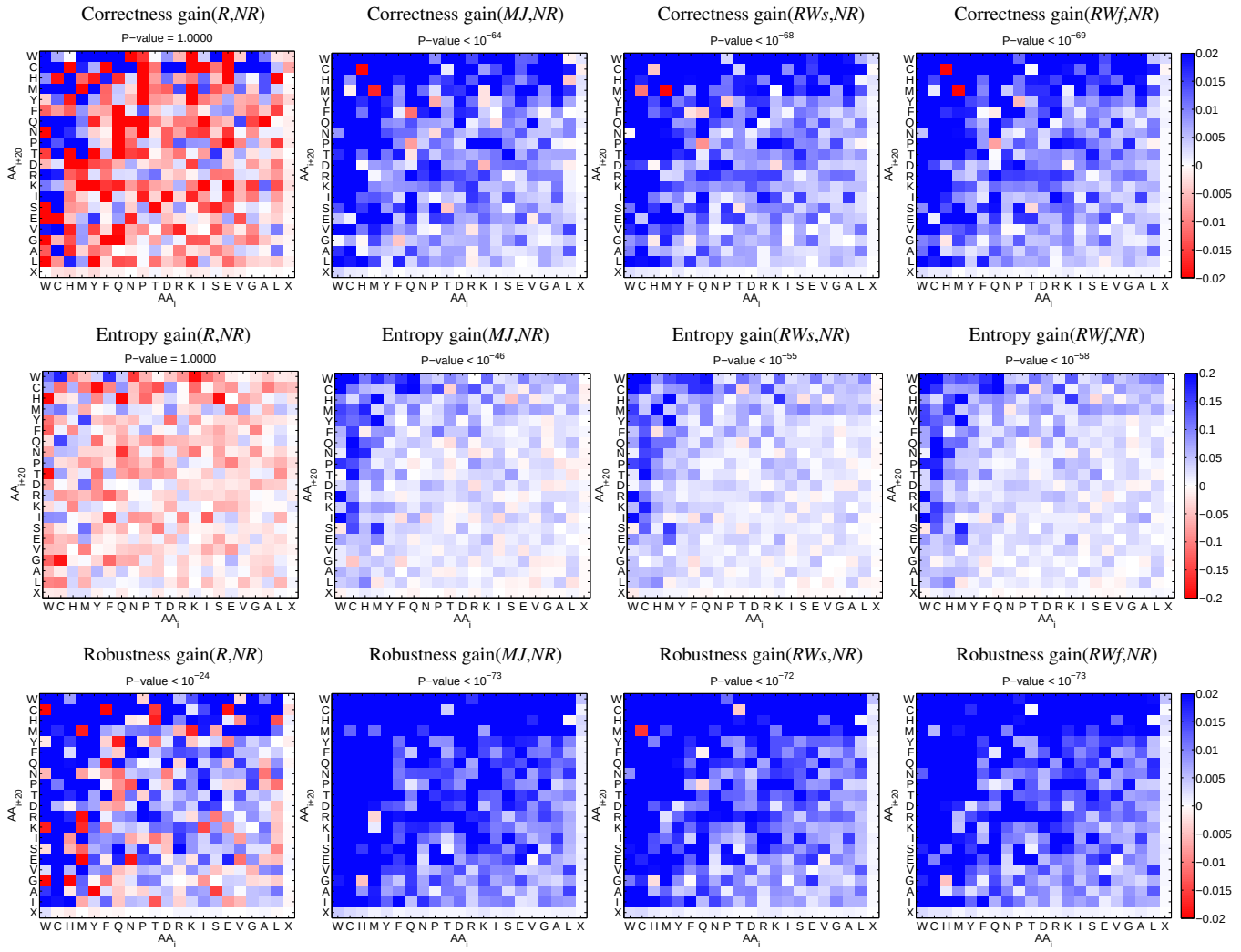

Fig. 5. Weighting scheme performance. Correctness (top row), entropy (middle), and robustness (bottom) gain of various weighting schemes compared to the standard, non-redundant scheme. One-tailed, paired-sample Wilcoxon signed rank test P-values are shown above each heat-map; see Figure 4 for more details.

ACKNOWLEDGEMENT

CK and ML are supported by BSF gramt No. 2009432. CY and CK are grateful to Eitan Bachmat for generous support. RK and CY would like to thank Yaron Rothman and Udi Feldman for introducing them to one another.

REFERENCES

Altschul, S. F. *et al*. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Berman, H. M. *et al*. (2013a). The future of the protein data bank. *Biopolymers*, **99**(3), 218–222.

Berman, H. M. *et al*. (2013b). Trendspotting in the protein data bank. *FEBS Letters*, **587**(8), 1036 – 1045.

Bernstein, F. C. *et al*. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, **112**(3), 535 – 542.

Brenner, S. E. *et al*. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences*, **95**(11), 6073–6078.

Budowski-Tal, I. *et al*. (2010). FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proceedings of the National Academy of Sciences*, **107**(8), 3481–3486.

Bull, S. C. *et al*. (2013). Maximising the size of non-redundant protein datasets using graph theory. *PloS one*, **8**(2), e55484.

Chou, P. Y. and Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, **13**(2), 222–245.

Cook, D. (1967). The relation between amino acid sequence and protein conformation. *Journal of Molecular Biology*, **29**(1), 167 – 171.

Eisenberg, D. *et al*. (1997). VERIFY3D: Assessment of protein models with three-dimensional profiles. In R. M. S. Charles W. Carter Jr., editor, *Macromolecular Crystallography Part B*, volume 277 of *Methods in Enzymology*, pages 396 – 404. Academic Press.

Garnier, J. *et al*. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, **120**(1), 97 – 120.

Gniewek, P. *et al*. (2011). Multibody coarse-grained potentials for native structure recognition and quality assessment of protein models. *Proteins: Structure, Function, and Bioinformatics*, **79**(6), 1923–1929.

Goldstein, R. A. (2008). The structure of protein evolution and the evolution of protein structure. *Current Opinion in Structural Biology*, **18**(2), 170 – 177. ¡ce:title¿Theory and simulation / Macromolecular assemblages¡/ce:title¿.

Gront, D. *et al*. (2011). Generalized fragment picking in Rosetta: design, protocols and applications. *PLoS One*, **6**(8), e23294.

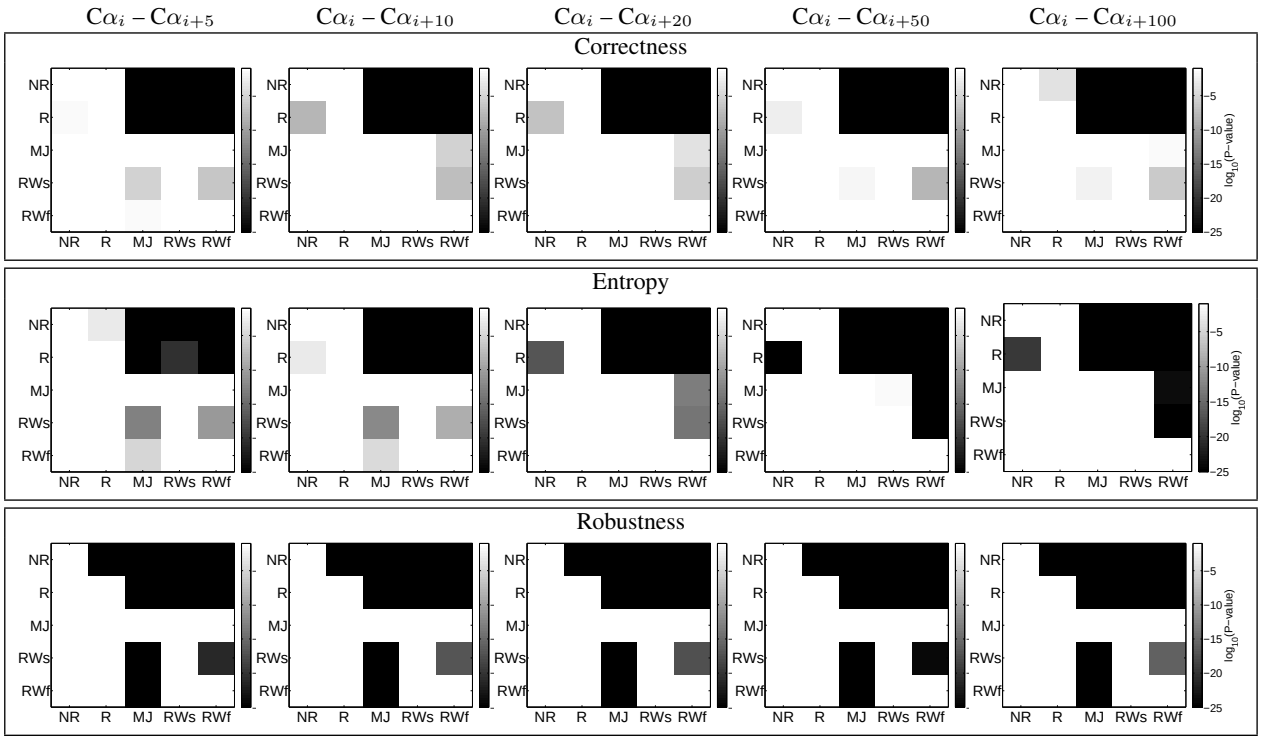

Fig. 6. Performance summary. Significance of correctness (top), entropy (middle), and robustness (bottom) gain of various weighting schemes, all-against-all (x-axis scheme over y-axis one), for distance distributions of Cα atoms 5 (left column) to 100 (right column) amino acids apart. The base-10 logarithm of one-tailed, paired-sample Wilcoxon signed rank test P-value is shown, with stronger colors indicating more significance gains. *NR*: non-redundant; *R*: redundant; *MJ*: Miyazawa and Jernigan (1996, 1999) sample weighting; *RWs*: per-sequence redundancy-weighting; *RWf*: per-feature redundancy-weighting.

Heath, M. T. (2002). *Scientific computing: an introductory survey*. The McGraw-Hill Companies Inc, New York, second edition.

Hobohm, U. and Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science*, **3**(3), 522–524.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**(12), 2577–2637.

Kalev, I. and Habeck, M. (2011). HHfrag: HMM-based fragment detection using HHpred. *Bioinformatics*, **27**(22), 3110–3116.

Karplus, K. (2009). SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Research*, **37**(suppl 2), W492–W497.

Kolodny, R. *et al*. (2013). On the universe of protein folds. *Annual Review of Biophysics*, **42**(1), 559–582. PMID: 23527781.

Kosloff, M. and Kolodny, R. (2008). Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins: Structure, Function and Bioinformatics*, **71**(2), 891–902.

Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences*, **106**(27), 11079–11084.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–1659.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, **37**(1), 145–151.

Lüthy, R. *et al*. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**(6364), 83–85.

McGuffin, L. J. *et al*. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, **16**(4), 404–405.

Miyazawa, S. and Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**(3), 534–552.

Miyazawa, S. and Jernigan, R. L. (1996). Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, **256**(3), 623 – 644.

Miyazawa, S. and Jernigan, R. L. (1999). Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins: Structure, Function, and Bioinformatics*, **34**(1), 49–68.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443 – 453.

Osadchy, M. and Kolodny, R. (2011). Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proceedings of the National Academy of Sciences*, **108**(30), 12301–12306.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, **85**(8), 2444–2448.

Prothero, J. (1966). Correlation between the distribution of amino acids and alpha helices. *Biophysical Journal*, **6**(3), 367.

Rost, B. (1996). PHD: Predicting one-dimensional protein structure by profile-based neural networks. In R. F. Doolittle, editor, *Computer Methods for Macromolecular Sequence Analysis*, volume 266 of *Methods in Enzymology*, pages 525 – 539. Academic Press.

Samudrala, R. and Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, **275**(5), 895 – 916.

Shen, Y. *et al*. (2013). Detecting protein candidate fragments using a structural alphabet profile comparison approach. *PloS one*, **8**(11), e80493.

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Bioinformatics*, **17**(4), 355–362.

Summa, C. M. and Levitt, M. (2007). Near-native structure refinement using in vacuo energy minimization. *Proceedings of the National Academy of Sciences*, **104**(9), 3177–3182.

Tanaka, S. and Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**(6), 945–950.

Thompson, J. D. *et al*. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**(22), 4673–4680.

Wang, G. and Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**(12), 1589–1591.

Wang, G. and Dunbrack, R. L. (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, **33**, W94–8.