Optimizing Protein Representations with Information Theory

Julian Mintseris[1] julianm@bu.edu
Zhiping Weng[1,2] zhiping@bu.edu

[1] Bioinformatics Program, Boston University, Boston MA 02215, USA
[2] Biomedical Engineering Department, Boston University, Boston MA 02215, USA

Abstract

The problem of describing a protein representation by breaking up the amino acids atoms into functionally similar atom groups has been addressed by many researchers in the past 25 years. They have used a variety of physical, chemical and biological criteria of varying degrees of rigor to essentially impose our understanding of protein structures onto various atom-typing schemes used in studies of protein folding, protein-protein and protein-ligand interactions, and others. Here, instead, we have chosen to rely primarily on the data and use information-theoretic techniques to dissect it. We show that we can obtain an optimized protein representation for a given alphabet size from protein monomers or protein interface datasets that are in agreement with general concepts of protein energetics. Closer inspection of the atom partitions led to interesting observations pointing to the greater importance of the hydrophobic interactions in protein monomers compared to interfaces and, conversely, greater importance of polar/charged interaction in protein interfaces. Comparing the atom partitions from the two datasets we show that the two are strikingly similar at alphabet size of five, proving that despite some differences, the general energetic concepts are very similar for folding and binding. Implications for further structural studies are discussed.

Keywords: atom types, mutual information, protein-protein complexes, amino acid properties

1 Introduction

The ever-increasing wealth of protein structural information has, in turn, spurred researchers to develop a host of new ideas and approaches that use this information to infer, predict and make new hypotheses about protein structure, function and evolution as well as relationships between proteins and other molecules in the cell. Applications have included statistical potentials to study and predict protein folding, protein docking, protein-small molecule and protein-nucleic acid interactions and 3D structural motif studies, to name a few. In most studies of protein structure, and indeed in any other modeling study, the first question that a researcher has to face is how to represent the problem. We will use statistical potentials as an example familiar to many readers.

Knowledge-based statistical potentials have been used in structural studies for almost three decades starting with their first introduction by Levitt and Warshel [6] and Tanaka and Scheraga [17] and further development by Miyazawa and Jernigan [13], Sippl [16] and many others. Discussions of practical and theoretical advantages and shortcomings of statistical potentials have been numerous, but currently they remain widely used in the areas of protein structure prediction, protein-protein interaction studies and docking, as well as protein-small molecule studies. The interpretation of statistical potentials most often relies either on statistical Bayesian approaches or on the more energetically appropriate Inverse Boltzmann Law, but regardless of the details, the general formulation is as follows:

$$E_{structure} = \sum_{i,j} Observed\ Structure_{ij} \log\left(\frac{Observed\ Database_{i,j}}{Reference_{i,j}}\right) , \quad (1)$$

where i and j are amino acid or atom types. Numerous variations on this theme exist in the literature.

Despite the extensive work on knowledge-based statistical potentials, we believe there is an aspect that so far has defied rigorous treatment - specifically the definition of i and j in Eq.(1). Essentially, this is an issue of finding an appropriate way to represent protein structures. Many studies use the 20 amino acids - the intuitive choice. Earliest efforts in this area started in 1978 - shortly after statistical potentials for protein folding were first introduced [22]. Considerable theoretical work exists on the simplest of alphabets involving just two groups - the so-called HP model [4, 18]. On the other extreme, recent explosion in the number of available nonredundant protein structures made possible studies that used the largest possible set of atoms, where every atom of every amino acid is defined separately - a total of 167 heavy atom types [15]. Protein-protein and protein-small molecule studies are not able to take advantage of the extended atom types due to insufficient data. In fact, even for protein structures, choosing a subset of the whole protein structure space, such as a fold or a large superfamily, as the Observed Database would require the use of a reduced atom type set in order to accurately estimate all the parameters. In this paper we focus on the derivation of such reduced representations and discuss their implications for our understanding of protein structure energetics. We use two distinct datasets of protein monomers and protein complexes and demonstrate how protein representations can point very specifically to the similarities and differences between the forces underlying the biophysical processes of folding and binding.

2 Information-Theoretic Approach to Optimizing Protein Atom Types

2.1 Datasets

The protein monomer dataset consists of 808 protein structures with sequence identity $\leq 20\%$ and resolution of 1.8 Å or better from the PISCES database [20]. This nonredundant set is sufficient to calculate the extended 167x167 pairwise matrix of atomic contacts within a 6 Å cutoff and subsequent calculations do not require any corrections for under-sampling. In order to get an unbiased dataset, we did not count contacts with chain neighbors, where chain neighbors for an atom of residue i were defined as all atoms belonging to residues $i-4$ to $i+4$ in the protein sequence. To ensure correct chain neighbor assignment, we carefully renumbered the residues in all structures using the S2C database [19].

The protein complex dataset includes 327 non-redundant protein complexes involving interaction between non-identical interacting partners and is an extension of a dataset derived previously [10]. Redundancy is defined on the level of SCOP [14] family, i.e. the dataset was grouped by family-family interactions and one representative complex was chosen from each group. In addition, we include 122 homodimers from a recently published dataset, non-redundant at 25% sequence identity level [2].

2.2 Information Theory

Cline *et al.* [3] used an information-theoretic approach to analyze the contributions of several traditional amino acid-based alphabets using mutual information. Note that the following definition of mutual information (MI) is similar to that of a statistical potential above:

$$MI = \sum_{i,j} P(i,\ j) \log \frac{P(i,j)}{P(i)P(j)} \tag{2}$$

where $P(i,\ j)$ is the probability that an atom of type i forms a contact with an atom of type j, and $P(i)$ and $P(j)$ are the marginal probabilities.

Interpretation of the reduced representation problem in information-theoretic terms is straightforward. Mutual information between two variables I and J (representing a grouping of the protein atom types) is a measure of how much information one variable reveals about the other [5]. If i and j are

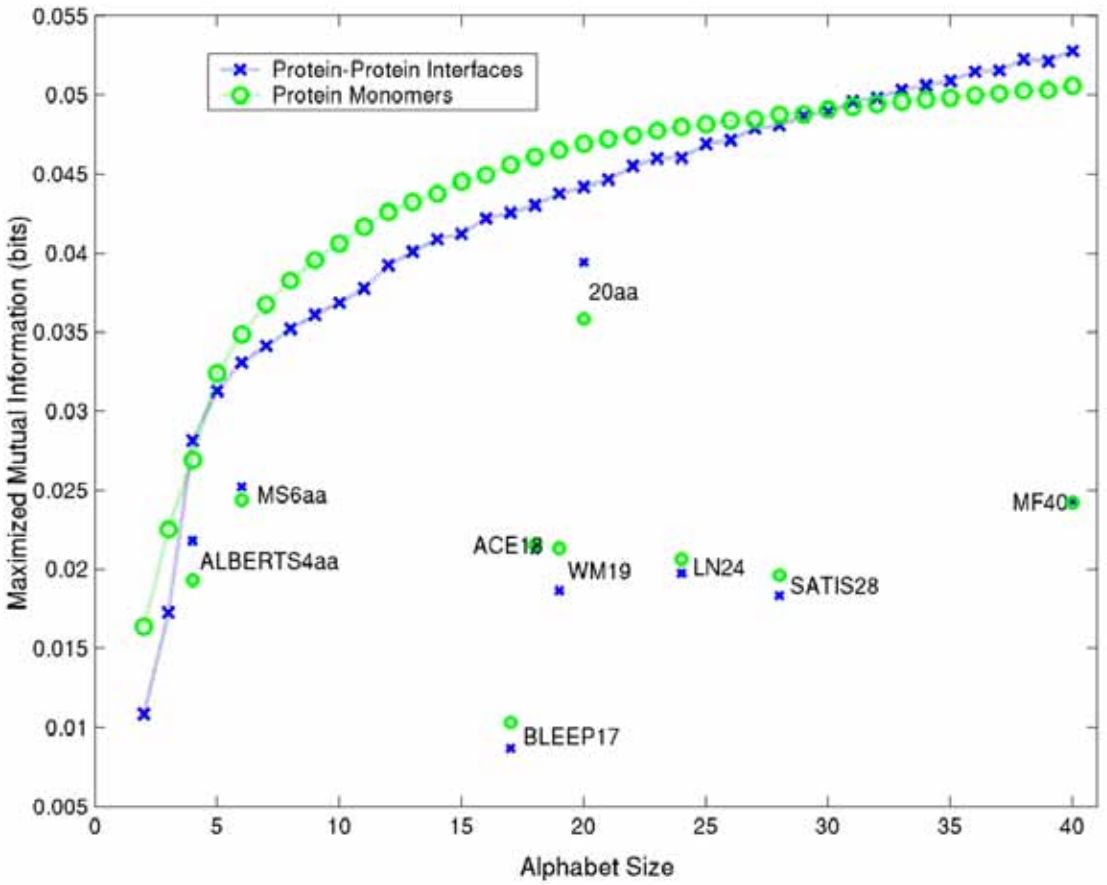

Figure 1: Maximized mutual information for all reduced representations from 2 to 40 atom types starting with 167 heavy atoms. The same optimization procedure was performed using a dataset of protein monomers and protein-protein complexes. In addition, we calculated mutual information for other published reduced alphabets: ×ALBERTS4 - 4 atom type representation based on the standard textbook amino acid grouping of Alberts *et al.* [1]; ×MS6aa - 6 atom types based on amino acid grouping by Mirny and Shakhnovich [11] ×ACE18 - 18 atom type representation from Zhang *et al.* [23]; ×WM19 - 19 atom type representation based on Warme and Morgan [22]; ×20aa - 20 atom types based on standard amino acids ×LN24 - 24 atom type representation based on Li and Nussinov [7]; ×SATIS28 - 28 atom type representation based on Snarey *et al.* [12], and ×MF40 - 40 atom type representation based on Melo and Feytmans [9].

instances of I and J, where the number of such instances is governed by the size of the atom type alphabet, we want to define i and j such that the mutual information is maximized. Each instance i or j is a grouping of protein atoms of one type. It is easy to see from Eq. (2) that if i and j are chosen randomly, the probability of the joint distribution would be equal to the product of marginal distributions resulting in zero mutual information. On the other extreme, the maximum possible mutual information for a given alphabet size can be determined if we take $P(i, j) = P(i) = P(j)$. Eq. (2) then reduces to:

$$MI_{max} = \sum_{i,j} P(i) \log \frac{P(i)}{P(i)P(j)} = \log(alphabet\ size) \tag{3}$$

Another way to think about this is to realize that grouping atoms with similar biochemical properties - atoms that are commonly found in protein structures in similar environments, tends to increase mutual information by increasing the certainty that a specific atom type will occur in a given protein environment. Thus mutual information is a rigorous and intuitive measure suitable for optimization.

Notice that mutual information is also a measure of independence. If the variables i and j are randomly distributed, they reveal no information about each other, as shown above. Assuming under a null hypothesis (H_0) that i and j are independent and an alternative hypothesis (H_1) that they are not, it can be shown that a log likelihood ratio test is exactly equivalent to the definition of mutual information [3].

In the statistical context of the test of independence, the objective of finding the representation with maximum mutual information is equivalent to maximizing the significance of the test of independence between the atom types. The problem of finding such an optimal reduced protein representation for a given target alphabet size is essentially equivalent to maximum likelihood estimation. We have a model as described above and a comprehensive nonredundant dataset of proteins and protein complexes. The distributions of heavy atoms into a given number of atom types, or the probabilities of membership of each heavy atom in an atom type group are the parameters to estimate. An exhaustive solution to this problem is impossible: the number of ways of distributing k objects into m bins grows very quickly. We use Monte Carlo methods to estimate the best reduced representations by randomly perturbing the bin memberships and accepting/rejecting based on the Metropolis criterion. This is in many ways similar to K-means clustering but here we optimize a metric global to the whole dataset as opposed to a distance from every point to some cluster center.

3 Results and Discussion

3.1 Comparison of Reduced Representations

We perform a million steps of Monte Carlo optimization five times for every alphabet size and for each dataset and keep the maximum. The results are shown in Figure 1. In addition to our results we have also calculated and included in Figure 1 mutual information for several published atom-typing schemes. Many authors have made efforts to group protein atoms based on a variety of criteria such as atoms size (radius), chemical connectivity, semi-rigorous observations of contact correlation, and general biophysical common sense. Of course, the most popular representation is the one based on the 20 standard amino acids and there has been at least one residue-level clustering study [21]. Much work has also been done with the simple HP models, grouping the standard amino acids into hydrophobic and hydrophilic ones. Common textbooks describe 4 groups - positive, negative, polar uncharged, and non-polar. Note that the atom-typing schemes referred to in Figure 1 represent only a small sample of such work. Sometimes the researchers make arbitrary choices distinguishing main-chain and side-chain atoms, and the classification of such amino acids as glycine and proline often differ substantially in different papers. In other words, there seem to be as many atom-typing schemes as there are investigators approaching the issue. Using mutual information as a measure of "quality", the published "hand-made" atom typing schemes including the traditional 20 amino acid representation

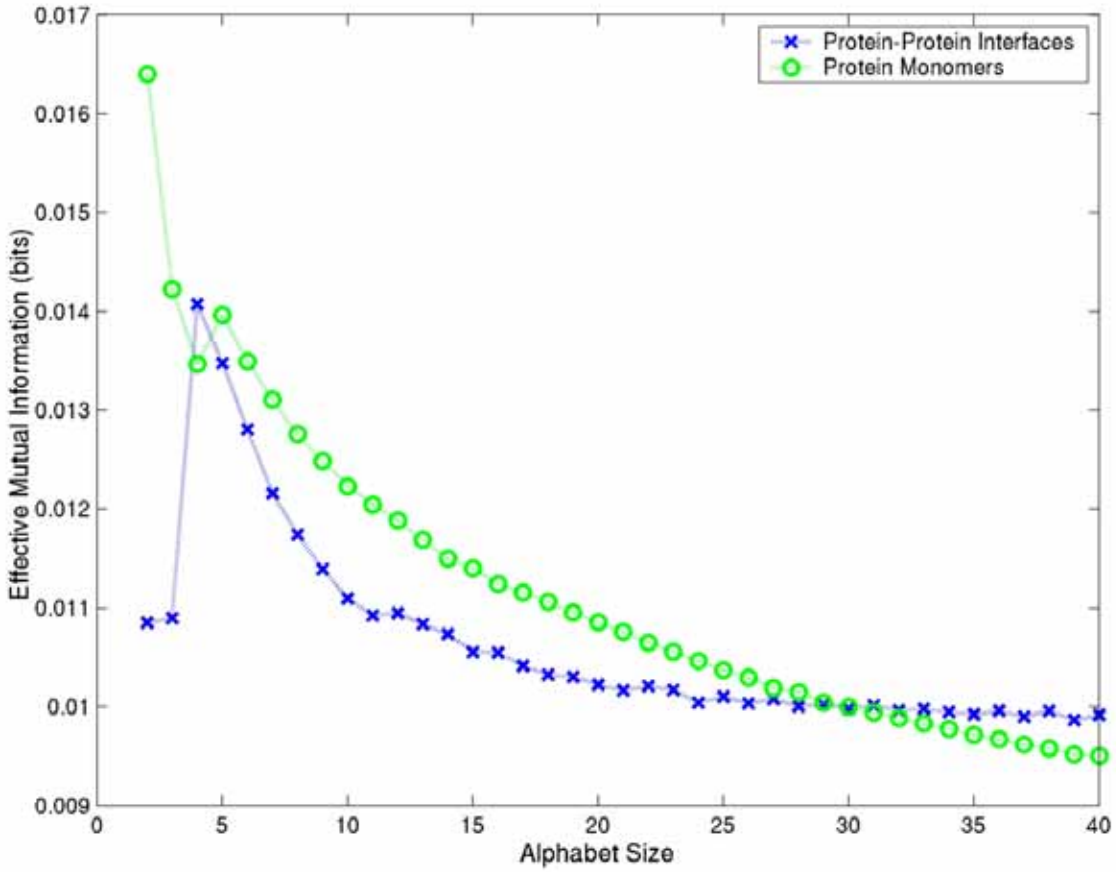

Figure 2: Mutual information of optimized protein representations from Figure 1 calculated as fraction of total mutual information possible given the alphabet size (Eq. 3).

as well as "textbook" categorizations are sub-optimal. For alphabet size 20, the mutual information obtained with the optimal reduced representation for protein complexes is 12.2% higher than the natural configuration where each of the 167 heavy atoms is assigned to its residue (here we refer to this configuration as the '20aa' scheme). The analogous figure for protein monomers is 30.6%. Note also, that this standard representation has substantially higher mutual information than all of the other published atom type schemes.

Figure 1 shows optimized mutual information monotonically increasing with increasing alphabet size. This is to be expected given the definition of mutual information in Eq. 2. Figure 2 gives better sense of how the optimized mutual information relates to maximum possible mutual information for a given alphabet size as calculated by Eq. 3. We call the ratio of optimized mutual information to this maximum, *effective mutual information.* This plot demonstrates some crucial differences as well as similarities between the two datasets and underlying structural energetics.

For protein monomers, the highest effective mutual information is attained with only two atom types - the HP model. This is in agreement with many prior observations of the importance of the hydrophobic effect in protein folding. There is another sharp peak at 5 atom types before effective mutual information steadily decreases. For protein interfaces, there is only one significant peak at 4 atom types.

Before considering some optimized protein representations in more detail, it is instructive to compare the results for protein monomers and protein interfaces. This involves a comparison between two partitions of the set of 167 heavy atoms. This problem has been approached in a variety of ways in the

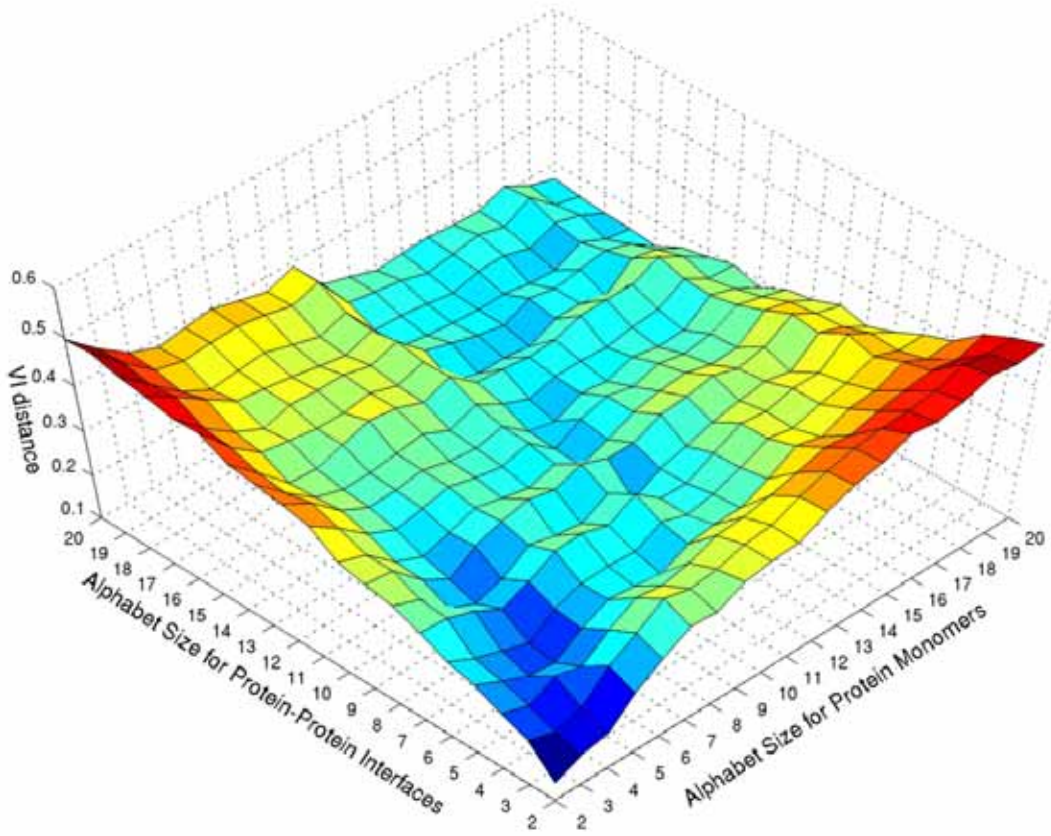

Figure 3: Comparison of protein representations optimized for protein monomers and protein-protein complexes.

past. Here, we choose a recent paper that proposes a measure of distance between two partitions of a set that, unlike most other measures, is a true mathematical metric called the Variation of Information or VI [8]. This metric also relies on concepts from information theory. A plot of all-against-all VI distance between optimized atom types from the two datasets is shown in Figure 3. It shows that the greatest similarity between the optimized representations of the two datasets occurs with both datasets at alphabet size 2. In other words, the Hydrophobic/Polar subdivision is very similar, even though we have seen in Figure 2 that the role of the hydrophobic effect is quite different in the context of protein monomers versus that of protein interfaces. Figure 4 shows a detailed breakdown of the HP model derived from the two datasets. It is evident that the atom type assignments are very similar and both generally make sense in light of the commonly accepted division of amino acids into hydrophobic/polar categories. While all atoms of most amino acids belong to the same class, a few, such as tyrosine and tryptophan are split. In both monomers and complexes, the atoms around the polar, hydrogen-bonding groups of tyrosine and tryptophan (including most main-chain atoms) are classified as hydrophilic, while the other portions are hydrophobic.

The other significant VI distance minimum in Figure 3 occurs with both datasets partitioned into alphabets of size 5. Note that in Figure 2, both protein monomers and protein interfaces exhibited high effective mutual information for this alphabet size. While the largest peak for protein interfaces occurs at alphabet size 4, alphabet of size 5 has the second highest effective mutual information. Figure 5 provides a closer look at the two optimized representations for alphabet of size 5 for each of the datasets. First, note the remarkable similarity between the atom type assignments derived

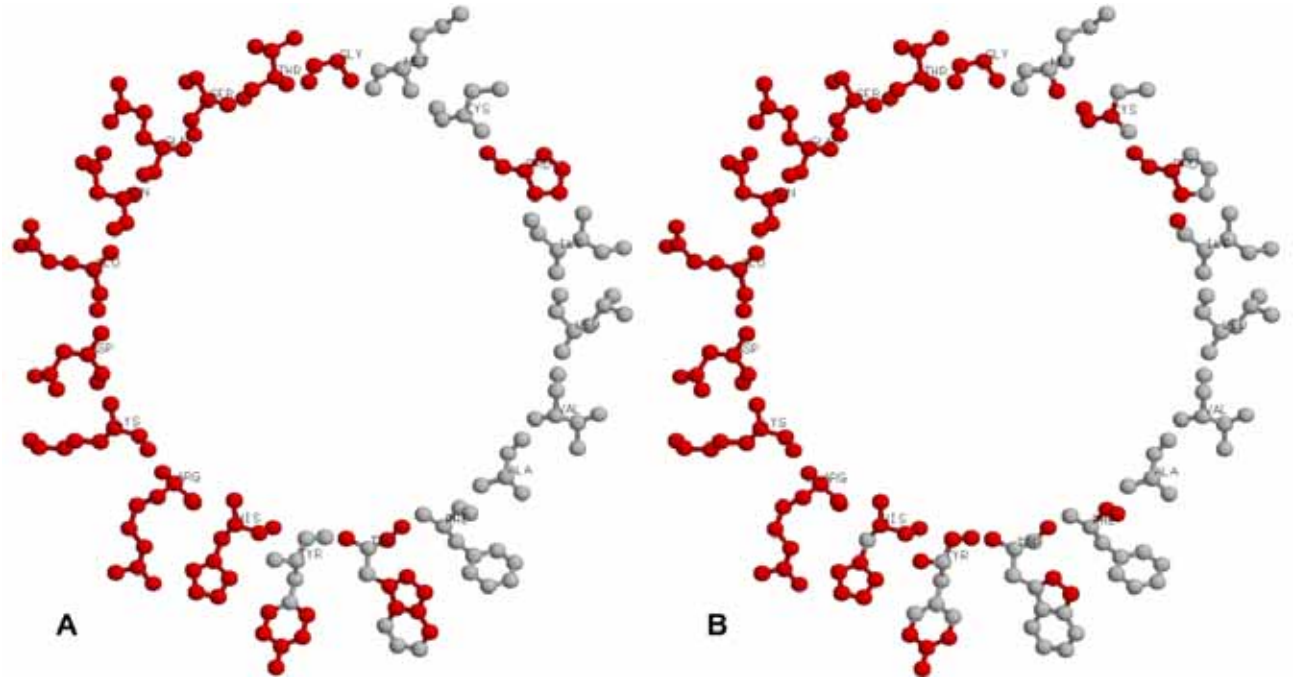

Figure 4: Protein representation for an alphabet size 2 optimized for **A**: 808 non-redundant protein monomers; **B**: 327 non-redundant protein interfaces.

from two completely different datasets. The main difference between the two occurs in the histidine side chain, the distal end of which is grouped together with the positively charged side chains in the context of protein monomers and with polar uncharged atoms in the context of protein interfaces. This duality of histidine is not at all surprising considering that this side-chain may or may not be positively charged depending on the pH of its environment. Our results here indicate that it perhaps may be more likely to carry the positive charge in protein monomers than in interfaces.

The atom type breakdown in Figure 5 is in excellent agreement with our biochemical knowledge of protein structures and confirms our understanding of the importance of hydrophobic, charge, and polar uncharged interactions to the stability of protein monomers and protein complexes. The five groups can be classified roughly as follows: (1) hydrophobic side-chains of I, L, V, M and F; (2) positively charged side-chains of K and R (and H for interfaces); (3) negatively charged side-chains of D and E (including main-chain atoms for interfaces); (4) all atoms of polar residues N, Q, S, and T ; main-chain atoms of charged amino acids; G; polar portions of side-chains of Y and W; (5) A, C; main-chain atoms of hydrophobic amino acids I, L, V, M and F; non-polar portions of Y and W.

The specificity and detail of atom partitions grow with increasing alphabet size and mutual information. Thus, for instance, in Figure 4, with alphabet of size 2, all atoms of both lysine and arginine belong to the same group suggesting that the positively charged side-chains are the dominant features of these amino acids. In Figure 5, with alphabet of size 5, both lysine and arginine atoms are divided into side-chain and main-chain. However, with alphabet size increasing to about 10 (except 6 for R in interfaces), both amino acids are split into three groups: (1) main-chain; (2) non-polar aliphatic portion of the side chain; (3) positively charged distal end of the side chain. The fact that this division appears at higher alphabet size, compared to such residues as tyrosine and tryptophan, which are divided at size 2, could mean that the latter distinction plays a more important role than the hydrophobic role of aliphatic portions of the K/R side-chains. Thus when choosing an appropriate atomic-level protein representation for a particular task, one may want to find a balance between the number of parameters and the level of detail.

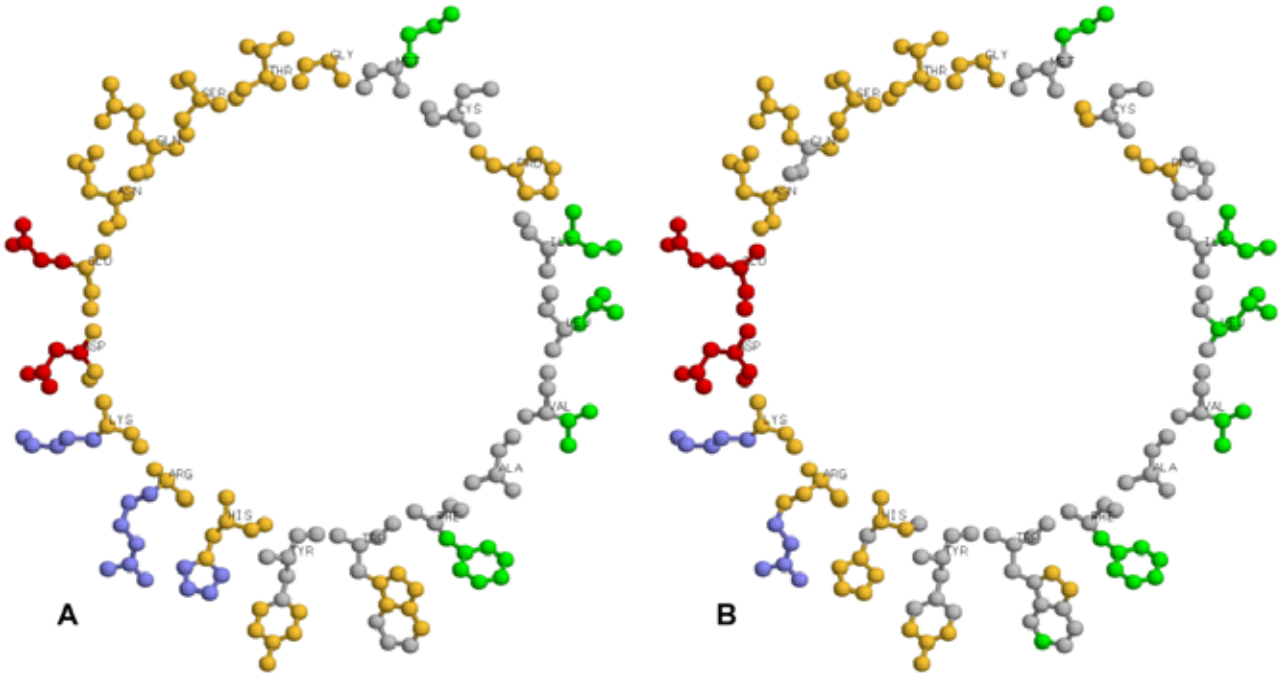

Figure 5: Protein representation for an alphabet size 5 optimized for **A**: 808 non-redundant protein monomers; **B**: 327 non-redundant protein interfaces.

4 Summary

We have developed methods for atom-typing, or finding an optimized protein representation, using information theory that are both rigorous and intuitive. The ability to define such atom groups can be useful for some aspects protein folding, such as side-chain placement, as well as protein-protein interactions studies and docking. Furthermore, we believe our methods will be useful in defining atom types for small molecule ligands, which is an important problem in protein-ligand interaction and QSAR studies. Protein structural motif searches may also benefit from our results. In the process of our analysis, the atom types have shed light on some aspects of protein folding and binding confirming the importance of hydrophobic effect and suggesting that polar/charged interactions are more important in binding. Despite some differences, it is clear from comparing atom partitions from protein monomer and interface datasets that the forces behind folding and binding are mostly very similar.

Acknowledgments

The authors are grateful to Dr. Temple F. Smith for helpful discussions. This work was partially funded by NSF grants DBI-0078194, DBI-0133834, and DBI-0116574. J.M. is supported by Krell Institute/DoE fellowship.

References

[1] Alberts, B., *Molecular Biology of the Cell*, New York: Garland Science, 2002.

[2] Bahadur, R.P., Chakrabarti, P., Rodier, F., and Janin, J., Dissecting subunit interfaces in homodimeric proteins, *Proteins*, 53(3):708–719, 2003.

[3] Cline, M.S., Karplus, K., Lathrop, R.H., Smith, T.F., Rogers, R.G., Jr., and Haussler, D., Information-theoretic dissection of pairwise contact potentials, *Proteins*, 49(1):7–14., 2002.

[4] Huang, E.S., Subbiah, S., and Levitt, M., Recognizing native folds by the arrangement of hydrophobic and polar residues, *J. Mol. Biol.*, 252(5):709–720, 1995.

[5] Kullback, S., Keegel, J.C., and Kullback, J.H., *Topics in Statistical Information Theory*, Berlin; New York: Springer-Verlag, 1987.

[6] Levitt, M. and Warshel, A., Computer simulation of protein folding, *Nature*, 253(5494):694–698, 1975.

[7] Li, A.J. and Nussinov, R., A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking, *Proteins*, 32(1):111–127., 1998.

[8] Meila, M., *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory*, Washington, DC, USA (Springer), 2777, 2003.

[9] Melo, F. and Feytmans, E., Assessing protein structures with a non-local atomic interaction energy, *J. Mol. Biol.*, 277(5):1141–1152, 1998.

[10] Mintseris, J. and Weng Z., Atomic contact vectors in protein-protein recognition, *Proteins*, 53(3):629–639, 2003.

[11] Mirny, L. and Shakhnovich, E., Evolutionary conservation of the folding nucleus, *J. Mol. Biol.*, 308(2):123–129, 2001.

[12] Mitchell, J.B.O., Alex, A., and Snarey, M., SATIS: Atom typing from chemical connectivity, *J. Chem. Inform. Comp. Sci.*, 39(4):751–757, 1999.

[13] Miyazawa, S. and Jernigan, R.L., Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading, *J. Mol. Biol.*, 256(3):623–644, 1996.

[14] Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247(4):536–540, 1995.

[15] Samudrala, R. and Moult, J., An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction, *J. Mol. Biol.*, 275(5):895–916, 1998.

[16] Sippl, M.J., Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins, *J. Mol. Biol.*, 213(4):859–883., 1990.

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

[18] Thomas, P.D. and Dill, K.A., Statistical potentials extracted from protein structures: how accurate are they?, *J. Mol. Biol.*, 257(2):457–469, 1996.

[19] Wang, G., Arthur, J.W., and Dunbrack, R.L. S2C:A database correlating sequence and atomic coordinate numbering in the Protein Data Bank, `http://www.fccc.edu/research/labs/dubrack/s2c`, 2002.

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

[21] Wang, J. and Wang, W., A computational approach to simplifying the protein folding alphabet, *Nature Structural Biology*, 6(11):1033–1038, 1999.

[22] Warme, P.K. and Morgan, R.S., A survey of atomic interactions in 21 proteins, *J. Mol. Biol.*, 118(3):273–287, 1978.

[23] Zhang, C., Vasmatzis, G., Cornette, J.L., and DeLisi, C., Determination of atomic desolvation energies from the structures of crystallized proteins, *J. Mol. Biol.*, 267(3):707–726, 1997.