Baseline Statistics

Jack Ellenberger, Sean Griesemer, Vishal Prasad
(Dated: June 5, 2015)

I. Introduction

Hydrophobic groups aid in the stabilization of hydrogen bonds along the main chains of protein structures by insulating the bonds from water molecules. This "wrapping" plays an important role in the interactions between protein molecules. Under-wrapped hydrogen bonds (UWHB's, also known as *dehydrons*), are thermodynamically favorable sites for protein-protein and protein-ligand associations. The study of how dehydrons influence these interactions has numerous potential applications in the design of pharmaceuticals.

It is thus pertinent to develop a statistical overview of dehydron placement within a typical protein. If dehydrons are indeed promoters of sub-molecular interactions, then at the very least one might expect dehydron placement to follow some non-uniform distribution. However, it is not currently well known how dehydrons are actually distributed. In order for future investigations to identify the role dehydrons play in protein interactions, they will need a summary of the baseline statistics on dehydron locations to be used as a comparison.

In this report, we present our computational approach for identifying dehydrons within proteins, giving a brief outline of the pipeline of our analysis. We then give a complete comparison of the "empirical" distribution of dehydron placement with respect to other atoms in the protein with a distribution of "uniformly" sampled dehydrons, in order to identify how the empirical distribution differs. We do so by locating the hydrogen bonds and dehydrons within each protein in a subset of the NRPDB set of PDB files, and then comparing the probabilities of finding actual dehydrons near arbitrary atoms in the protein with the corresponding probabilities of finding randomly sampled hydrogen bonds. We also provide comparisons of the distribution of dehydron centers between the uniform and empirical model by measuring the dispersion.

II. Overview of the Dehydron

A dehydron is a hydrogen bond containing a less-than-normal amount of wrappers, or nonpolar carbonaceous groups within its desolvation domain. A schematic of a hydrogen bond is shown in Figure 1. The geometric criteria for two mainchain residues to contain a hydrogen bond are listed below:

1. $\overline{HA} \leq 2.5\mathring{A}$
2. $\overline{DA} \leq 3.5\mathring{A}$

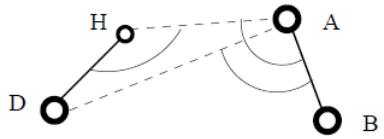

FIG. 1. Geometric model for hydrogen bonds. D is the donor, H is the hydrogen atom, A is the acceptor, and B is adjacent to the acceptor.

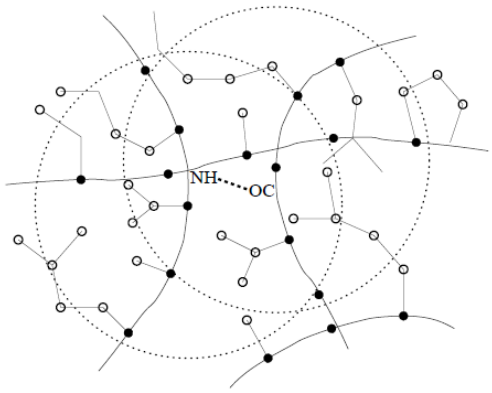

FIG. 2. Illustration of the desolvation domain, with the hydrogen bond depicted at the center.

3. $\angle DHA \geq 90°$
4. $\angle HAB \geq 90°$
5. $\angle DAB \geq 90°$

The atoms in question are $[D, H, A, B] = [N, H, O, C]$.

For any mainchain hydrogen bond, one can construct a "desolvation domain" by drawing spheres of radius R around the C_α carbons; the union of the spheres is the desolvation domain (see Figure 2). The radius $R = 6.5\mathring{A}$ is chosen because it gives the volume at which maximum local residue packing is obtained; however, other radii specifying different sizes of interaction domain can be experimented with.

Within the desolvation domain, the wrappers are the hydrophobic nonpolar carbonaceous groups which "wrap" the hydrogen bond from contact with water molecules. The nonpolar carbonaceous groups for each residue are listed in Table 8.2 of the class textbook. The dehydron is under-wrapped, meaning it has relatively few wrappers. The number of wrappers in a standard mainchain hydrogen bond is normally distributed, with a mean of roughly 25 and a standard deviation of 6. Typically, the dehydron is defined somewhat arbitrarily

to have fewer than 19 wrappers, or one standard deviation below the mean. Other numbers for wrapper cutoffs could be possible, if a more physically justifiable wrapper cutoff is discovered.

III. Dehydron Finder

In order to locate dehydrons within our numerous protein database files, we employed the ProDy python library developed by the University of Pittsburgh (`http://prody.csb.pitt.edu`). We opted against using the available WRAPPA program in order to extract interesting data features that were not immediately available through WRAPPA's processing. In addition, we chose to structure outputs in the form of .csv's which, while less visually appealing that WRAPPA's outputs, were much easier to parse and probe for data trends.

Our program operated on the input of the 3,842 .pdb files in the NRPDB data set, though not all proteins were used. Due to computational complexity limits, we ignored any files that had an atom count greater than 10,000, as well as any protein in which the number of C_α "neighbors" was greater than 4,000. We defined C_α neighbors as pairs of C_α carbons that were within one desolvation domain radius of each other. This would both weed out the .pdb files that were too large to calculate in a responsible amount of time, as well as ridding our data set of malformed .pdb files. When we parsed each .pdb we only mined data from the first provided model, however there was evidence that without our two limitations, some multi-model proteins were sneaking in. With the restrictions in place, there was no chance of several similar protein chains layering on top of one another, skewing the data.

From the main set of 3,842 proteins, we ultimately found 1,874 proteins that fit our limitations with the desolvation domain radius set at $6.5\mathring{A}$. For each one of these proteins, after being run through the Reduce program to infer hydrogen locations, the number of legitimate C_α neighbors are found. For each pairing of C_α carbons within one desolvation domain radius of each other, the mainchain atoms $[N, H, O, C]$ are identified, and the residues are checked for a hydrogen bond using the five criteria. Then, if a residue pair contains a hydrogen bond, the nonpolar carbons are counted within the desolvation domain, and if this number is less than 19, then the hydrogen bond is defined to be a dehydron.

ProDy's part in this process was first to parse the protein into an AtomGroup. From each AtomGroup, C_α carbons could be selected with a Selection.calpha call, and the C_α neighbors could be identified using AtomGroup.findNeighbors. Once a pair of C_α carbons had been identified as a hydrogen bond, nonpolar carbons could be found using a dictionary of residue names and their constituent nonpolar atoms. At this point we also found the number of atoms within each desolvation domain, and added that number to a running total for dehydrons or non-dehydron hydrogen bonds where applicable.

Once all dehydrons had been identified, we created a random model of dehydron distribution for each protein (to be described further in Section IV) and wrote our findings to five output files for later analysis in either R or Python.

Sample output of our program for the PDB file 4XED.pdb is shown in Figure 3. The column headings are explained below:

fname: PDB file name

ca1RNum, ca2RNum: Residue number of the 1st/2nd C_α center of a given desolvation sphere

ca1RName, ca2RName: Residue name of the 1st/2nd C_α center of a given desolvation sphere

ca1X/ca1Y/ca1Z, ca2X/ca2Y,ca2Z: Location of the 1st/2nd C_α center in Cartesian (XYZ) coordinates, in $\mathring{A}$

cX/cY/cZ: Location of the desolvation domain center in Cartesian (XYZ) coordinates, in $\mathring{A}$. The desolvation domain center is defined by the midpoint between the two C_α locations.

NPC#: Number of nonpolar carbons found in the dehydron or H-bond desolvation domain

AtomsInHBDD: Number of atoms found in the H-bond desolvation domain

AtomsInDDD: Number of atoms found in the dehydron desolvation domain

isDehydron: True if H-bond is a dehydron, False otherwise

HB#: Number of H-bonds found in file

HB#less1SD: Number of H-bonds which have nonpolar carbon counts under 1.0 standard deviations less than the mean of 25

HB#less1.1SD: Number of H-bonds which have nonpolar carbon counts under 1.1 standard deviations less than the mean of 25

HB#less1.5SD: Number of H-bonds which have nonpolar carbon counts under 1.5 standard deviations less than the mean of 25

dehydron#: Number of dehydrons found in file

totalAtomsInPDB: Total number of atoms found in PDB file

totalAtomsInDDD: Total number of atoms found in the desolvation domains of located dehydrons

totalAtomsInHBDD: Total number of atoms found in the desolvation domains of H-bonds

There are four tables outputted by the program: (a) a per h-bond output, (b) a per dehydron output, (c) a per random dehydron output, and (d) a per file output. Each of these tables contains all information that might

a) Per H-bond Output

fname	ca1RNum	ca1X	ca1Y	ca1Z	ca2RNum	ca2X	ca2Y	ca2Z	cX	cY	cZ	ca1RName	ca2RName	NPC#	AtomsInHBDD	isDehydron
e4XED.pdb	10	3.17	16.36	1.03	24	0.42	18.58	4.88	1.79	17.47	2.95	THR	THR	25	148	FALSE
e4XED.pdb	10	3.17	16.36	1.03	24	0.42	18.58	4.88	1.79	17.47	2.95	THR	THR	25	148	FALSE
e4XED.pdb	69	20.25	12.15	6.79	94	18.78	16.73	4.69	19.51	14.44	5.74	GLY	MSE	19	149	FALSE
e4XED.pdb	69	20.25	12.15	6.79	94	18.78	16.73	4.69	19.51	14.44	5.74	GLY	MSE	19	149	FALSE
e4XED.pdb	93	16.62	14.42	2.55	13	11.66	16.46	0.18	14.14	15.44	1.36	THR	THR	20	159	FALSE
e4XED.pdb	17	19.52	20.97	8.26	96	23.77	19.73	4.79	21.65	20.35	6.53	ILE	LYS	14	133	TRUE
e4XED.pdb	19	15.43	24.85	8.87	16	18.74	22.41	4.89	17.08	23.63	6.88	THR	ASN	15	130	TRUE

b) Per Dehydron Output

fname	ca1RNum	ca1X	ca1Y	ca1Z	ca2RNum	ca2X	ca2Y	ca2Z	cX	cY	cZ	ca1RName	ca2RName	NPC#	AtomsInDDD
e4XED.pdb	17	19.52	20.97	8.26	96	23.77	19.73	4.79	21.65	20.35	6.53	ILE	LYS	14	133
e4XED.pdb	19	15.43	24.85	8.87	16	18.74	22.41	4.89	17.08	23.63	6.88	THR	ASN	15	130
e4XED.pdb	5	-7.77	5.27	3.96	28	-10.64	10.22	4.70	-9.20	7.74	4.33	ASN	THR	15	158
e4XED.pdb	29	-12.57	7.54	6.52	32	-17.06	9.66	7.52	-14.82	8.60	7.02	ASP	GLY	14	137
e4XED.pdb	54	-12.74	15.56	10.68	57	-9.53	17.76	6.87	-11.14	16.66	8.78	ALA	GLU	15	132

c) Per Random Dehydron Output

fname	ca1RNum	ca1X	ca1Y	ca1Z	ca2RNum	ca2X	ca2Y	ca2Z	cX	cY	cZ	ca1RName	ca2RName	NPC#	AtomsInDDD
e4XED.pdb	90	7.44	10.66	2.17	73	8.18	10.55	7.31	7.81	10.60	4.74	LEU	VAL	30	168
e4XED.pdb	75	1.58	8.73	8.78	88	1.57	6.86	3.86	1.58	7.79	6.32	ALA	GLY	34	179
e4XED.pdb	37	-3.07	10.52	13.12	49	-2.59	9.82	17.75	-2.83	10.17	15.44	TYR	TRP	35	175
e4XED.pdb	63	9.28	19.79	11.74	21	9.49	22.82	6.99	9.38	21.31	9.36	HIS	TYR	29	155
e4XED.pdb	50	-3.12	13.56	17.45	37	-3.07	10.52	13.12	-3.09	12.04	15.29	ILE	TYR	30	165

d) Per File Output

fname	HB#	HB#less1SD	HB#less1.1SD	HB#less1.5SD	dehydron#	totalAtomsInPDB	totalAtomsInDDD	totalAtomsInHBDD
e4XED.pdb	39	6.19	5.30	2.61	5	1602	690	1430

FIG. 3. Our program's output for the file 4XED.pdb. These are tables of (a) information for each located hydrogen bond, (b) information for each located dehydron, (c) information for each randomly-sampled dehydron, (d) total hydrogen bond, dehydron and atom counts per file. Each row is a distinct dehydron or hydrogen bond in tables (a),(b) and (c), and each row is a distinct file in table (d).

```
WRAPPA OUTPUT: BONDS
PDB NAME:      e4XED
DATE:          2015/05/17              NOTE: All distances are given in Angstroms. All angles are given in degrees.

--BND-- ---------------DONOR---------------- --------------ACCEPTOR-------------- ---DST--- -------ANGLES------- --WRP--
     ID Res C  Seq I  DSC *  D   *  H   * SS Res C  Seq I  DSC *  A   *  B   * SS   DA    HA     DHA     DAB     HAB   W  **
-------------------------------------------------------------------------------------------------------------------------
HB_0001 THR A    19    CA     N      H        ASN A   16    CA  A  O   A  C   A    3.09 2.12 163.83 141.19 140.90  16   2
HB_0009 ASN A     5    CA     N      H        THR A   28    CA     O      C        2.92 1.98 157.14 169.82 166.83  18   6
HB_0010 GLY A    32    CA     N      H        ASP A   29    CA     O      C        2.89 1.92 163.97 120.96 116.77  17   6
HB_0011 GLU A    57    CA     N      H        ALA A   54    CA     O      C        3.14 2.16 168.10 132.46 132.51  17   4
HB_0012 ILE A    17    CA     N      H   A    LYS A   96    CA     O      C        2.71 1.71 172.64 135.75 133.26  16   4
```

FIG. 4. WRAPPA Bonds file output for 4XED.pdb. There were actually 12 dehydrons identified, but 7 of them were removed from this list because they were corresponding duplicates from various models. Only 5 distinct dehydrons were found.

be relevant for comparing the distributions of dehydrons between the uniform and empirical models.

To check whether our program was finding dehydrons correctly, we compared our output with that of the WRAPPA program for the file 4XED.pdb. The Bonds file output is displayed in Figure 4. Our program identified the same five dehydrons as the WRAPPA program (compare Figure 4 with Figure 3b), giving us confidence that our dehydron finder functions properly.

IV. Results and Discussion

Please see the following link for our complete results: `https://github.com/jackellenberger/DigitalBiology2015BaselineStatistics/`

The purpose of this project is to compare dehydrons in real proteins against dehydrons which are "uniformly distributed" in a model protein. Our method of collecting information about real dehydrons was discussed above in Section III, but how do we reason about a "uniform distribution" in a fictional protein? Our first thoughts on this matter were to

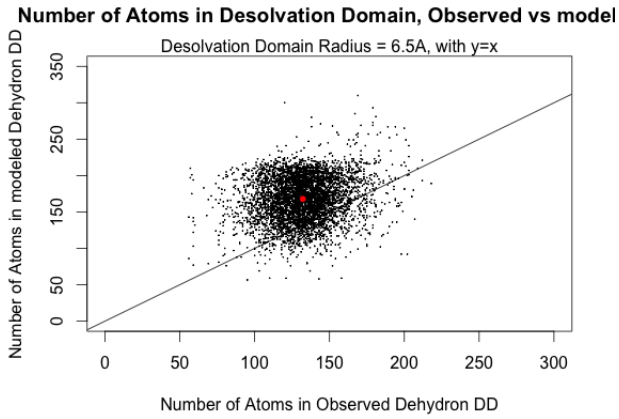

FIG. 5. Scatter plot of the number of atoms found within desolvation domains of located dehydrons (x-axis) and randomly sampled dehydrons (y-axis). The red dot, roughly (140,160), is the centroid of the data. The line is $y = x$.

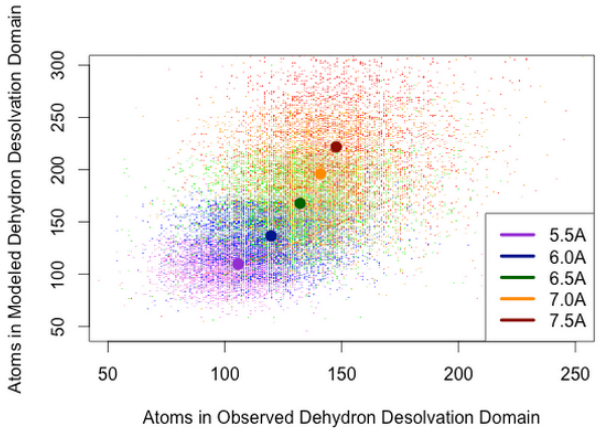

FIG. 6. Scatter plots of the number of atoms found within desolvation domains of located dehydrons (x-axis) and randomly sampled dehydrons (y-axis) for five domain radii. The large dots are the centroids of the data.

generate a random protein, and then calculate the number and distribution of dehydrons found in said protein. The problem with this approach is that the concept of a "random protein" is poorly defined, and the methods for constructing such an object are unclear. The approach we chose instead, and which is employed throughout this paper, is for a given protein **Pr**:

1. Identify the set of all hydrogen bonds H and the set of all dehydrons D of **Pr**
2. Call D the "observed distribution" of dehydrons in **Pr**
3. Randomly select $|D|$ hydrogen bonds, and call them the "uniformly modeled distribution" of dehydrons in **Pr**
4. Compare the observed and modeled distribution of dehydrons

The number of the dehydrons in the observed and the modeled distribution is the same. All that changes is the distribution of dehydron positions.

In Figure 5, we compare distributions of the number of atoms found within desolvation domains of observed dehydrons vs. modeled dehydrons. This joint distribution appears to take the form of a Gaussian peak. The centroid tells us how the mean number of atoms between the two distributions differs. The centroid is located just left of the $y = x$ diagonal, indicating that dehydron domains, on average, contain fewer atoms than typical hydrogen bond domains. In Figure 6 we plot the domain radii dependence of these distributions. As the domain radius decreases, the centroid appears to approach the $y = x$ diagonal, and the variance becomes narrower.

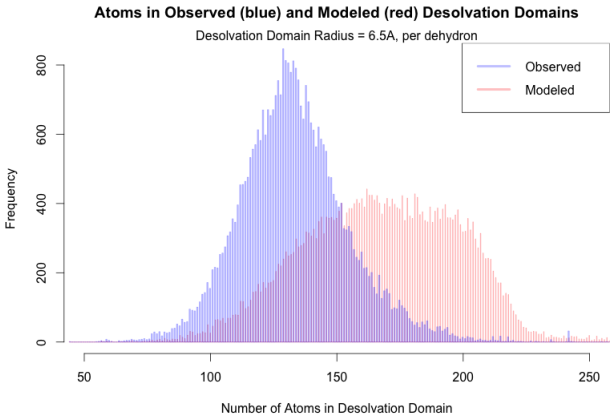

FIG. 7. Histogram of the number of atoms found in the desolvation domains of dehydrons, taken for both the observed and modeled distributions.

Figure 7 similarly reveals the difference in the numbers of atoms within empirical (blue) vs. modeled (red) dehydron domains. In addition, we find that the modeled dehydron distribution is strikingly flatter and wider than the empirical distribution. The means are 133.66 atoms per observed dehydron and 167.77 atoms per modeled dehydron.

A simple interpretation of these results is that dehydrons have fewer wrappers in their domains than typical hydrogen bonds, and therefore have fewer atoms overall. Since a dehydron be definition has fewer than 19 wrappers in its desolvation domain, whereas the typical hydrogen bond has 25, then there should be a difference of no less than 6 atoms. As there are many other atoms occupying the domains besides wrappers, our measured difference of 20 atoms is reasonable.

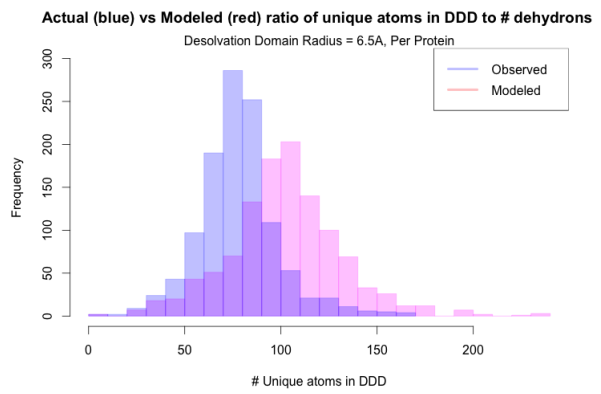

FIG. 8. Histogram of the number of atoms found in the desolvation domain of any dehydron in a given protein. Unlike in Figure 7, the data here was taken per protein rather than per dehydron. We can think of this data as being the number of atoms found within the union of the dehydron desolvation domains of all dehydrons in a given protein.

However, Figures 5-7 only present the numbers of atoms within each dehydron domain, and fail to consider whether these atoms are shared among domains. In Figure 8, we restrict per-file desolvation domain atom numbers to account only for unique atoms, and disregard shared atoms. In this plot, we still see a difference in the distributions, but the difference is less dramatic. The modeled distribution has a much more Gaussian appearance, a correction from Figure 7. The means are closer together than they were in Figure 7: 88.89 atoms per observed dehydron, and 102.36 atoms per modeled dehydron. Thus the desolvation domains of the modeled proteins still tend to contain more atoms. Indeed, the probability that a given atom is found within the desolvation domain of some dehydron is 0.4627 for observed data, and 0.5863 for the modeled data. This shows that the number of atoms in observed dehydrons is less than the number of atoms in modeled dehydrons. However, the distributions are centered much closer together. Comparing the distributions with those of Figure 7, we conclude that shared atoms are much more common in modeled dehydrons than in observed dehydrons.

With this in mind, we are led to believe that modeled dehydron domains tend to intersect more, causing them to share more atoms. To further justify this idea, we computed the number of shared atoms in observed vs. modeled dehydron domains, and displayed the results in Table I. We find this number is higher in the modeled data than in the observed data.

The fact that the difference between the two distributions in Figure 7 is less dramatic in Figure 8 is accounted for by our finding in Table I that the desolvation domains in the uniform model tend to intersect

	Observed	Model
Mean	1286.948	1778.207
Median	917	1129

TABLE I. The number of atoms found in the intersection of two or more desolvation domains. Both the mean and median number of shared atoms is lower in the observed dehydrons than modeled dehydrons, suggesting that modeled dehydron domains have a greater degree of intersection.

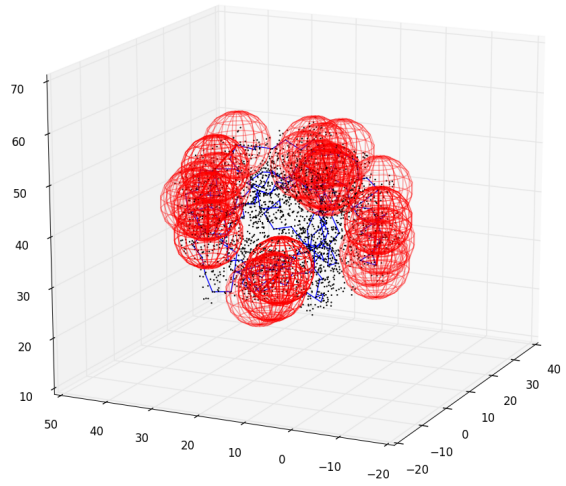

FIG. 9. The observed desolvation domains of protein 3FLL. In blue is the mainchain; in red, the desolvation domains of dehydrons; the dots are atoms. Note that i) the desolvation domains tend to be on the surface, and ii) because the dehydrons are located on the surface, the density of atoms in the desolvation domains is low, and hence the number of atoms found in some desolvation domain is lower than it would be if the dehydrons were uniformly distributed as in the model protein.

more. Because of the high number of intersections in the uniform model, in Figure 7 many atoms are double counted. Redundant counts of atoms disappears in Figure 8. Nonetheless, both figures suggest that dehydrons, when compared with hydrogen bonds, tend to be located in less populated parts of the protein or areas with relatively low numbers of atoms. These areas would serve as hotspots for protein-protein interactions.

In fact, a 3D visual of the protein 3FLL shown in Figure 9 gives a clear indication that observed dehydrons are most often located at the outer edges of the protein. These outskirts are much less populated with atoms than the bulk of the protein. In Figure 10, we compare 3D visuals of desolvation domain locations between observed dehydrons (left) and modeled dehydrons (right) for the protein 2Q2A. The observed dehydrons are almost all located at the outer edge, while the modeled dehydrons are predominantly clustered within the bulk.

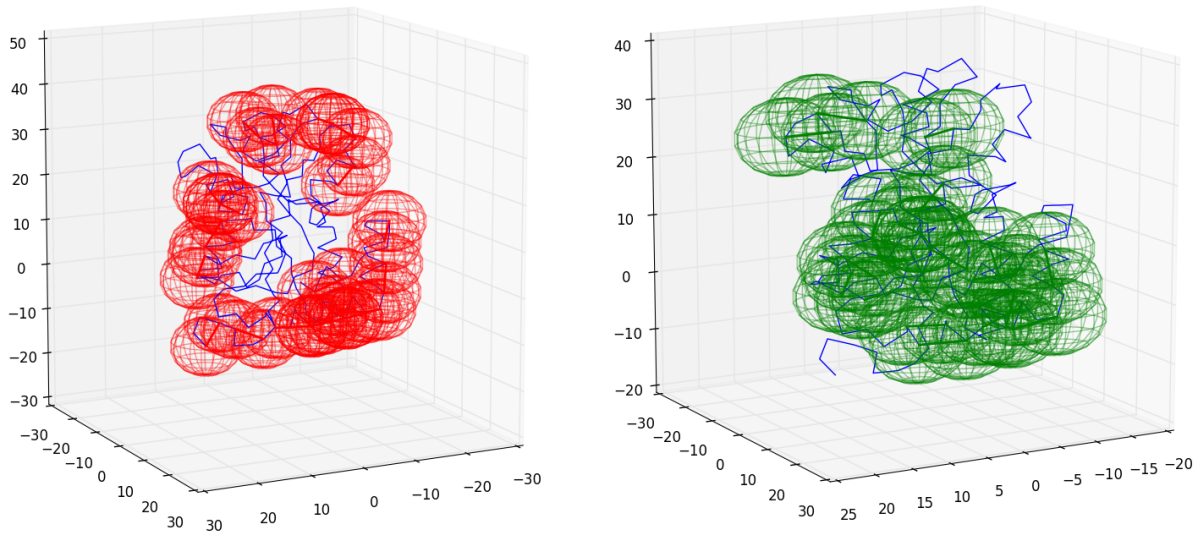

FIG. 10. LEFT: Desolvation domains in red of the observed dehydron distribution of protein 2Q2A. Notice that the desolvation domains are arranged around the surface. In blue is the mainchain. RIGHT: desolvation domains in green of the modeled dehydron distribution of protein 2Q2A. The dehydrons are distributed in the interior of the protein as well as the surface.

Now, we would like to be able to quantitatively determine the distribution of dehydron locations within proteins. This problem is difficult because every protein is different, and our measure must be able to adequately draw normalized comparisons of the location distributions between observed and modeled dehydrons. Our method of choice was to compare the index of dispersion I_D of dehydron locations within proteins between the empirical and uniform model. The index of dispersion is a normalized measure of the dispersion of data points, and is given by:

$$\text{1D: } I_D = \frac{\sigma^2}{\mu}, \text{ 3D: } I_D = \frac{|\mathbf{\Sigma}|}{\mu}$$

where $|\mathbf{\Sigma}|$ is the determinant of the 3D covariance matrix of dehydron locations, and μ is the mean of the locations. The physical intuition here is that $|\mathbf{\Sigma}|$ gives the volume spanned by the data points. If we compare I_D between observed vs. modeled dehydrons, we should find that the dispersion is higher for observed dehydrons if they are more spread apart.

In Figure 11, we plot distributions of the ratios of I_D of modeled over observed dehydrons per protein. For domain radius 6.5$\mathring{A}$, the median ratio is 0.354, and the vast majority of ratios are much less than 1. Observed dehydrons therefore have a larger index of dispersion on average, and so they are more spread out. There is also a large dependence of the ratios on the domain radii, yet for all radii studied we found that the median ratio was less than 1.

It is of further interest to look at the distribution of dehydron locations linearly along the protein's mainchain. To do so, we used the residue numbers of dehydron locations as an indicator of their linear position along the mainchain. As dehydron centers are defined by the locations of their two C_α's, we took the average of the C_α residue numbers to represent the dehydron location. Some C_α pairs were quite far apart, so we chose to filter any dehydrons with C_α's more than 20 residues apart out of this analysis.

In Figure 12, 1D plots of the distribution of residue numbers of observed vs. modeled dehydrons are shown for the protein 3FLL. In this case, the observed dehydrons are positioned close to the tails of the mainchain, while the modeled dehydrons are more evenly distributed. However, after taking many random samples of dehydrons, we found that their average I_D was only slightly lower than that of the observed dehydrons. On the 1D plots, modeled dehydrons were frequently spread out like the observed dehydrons.

Quantitative comparisons of the indices of dispersion of observed vs. modeled dehydron residue numbers are given in Figure 13. For radius 6.5$\mathring{A}$, the median dispersion ratio was 0.938, meaning that the observed dehydrons were slightly more spread out along the mainchain. Between radii 5.5-7.5$\mathring{A}$, the median ratio decreased monotonically from 0.975 to 0.827. We thus conclude that, while dehydrons are predominantly located at the surfaces of proteins in 3D space, there is only a modest preference for dehydrons to be spread

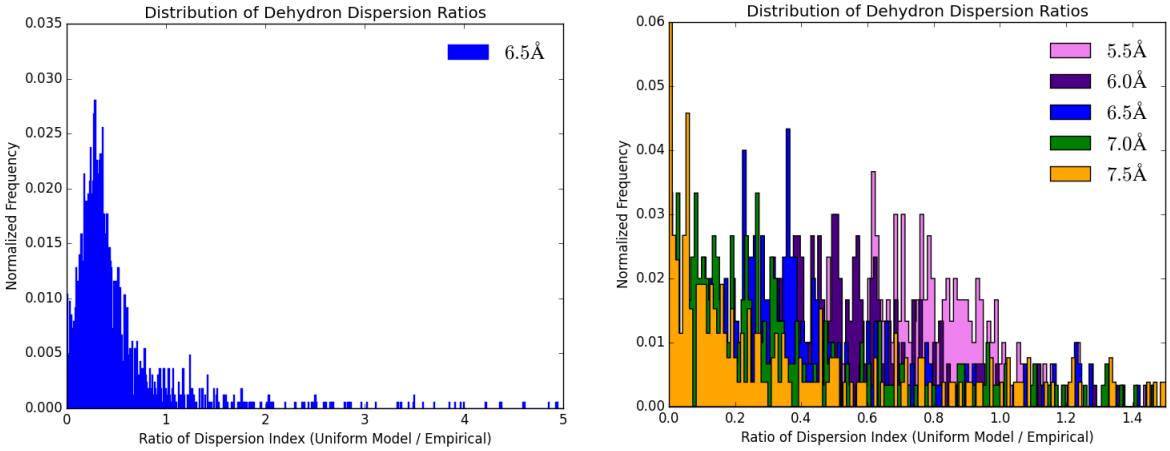

FIG. 11. LEFT: Histogram of the ratios of index of dispersion of observed dehydrons over modeled dehydrons per protein, for domain radius 6.5$\mathring{A}$. RIGHT: Same histograms, but for several domain radii. The dehydron locations were represented by the xyz coordinates of their centers, and the 3D index of dispersion formula was used.

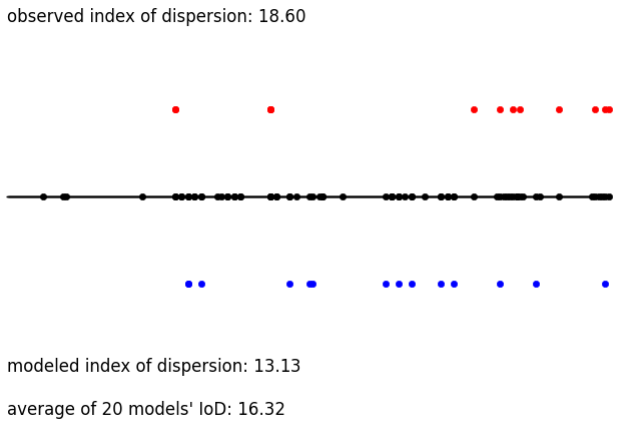

FIG. 12. 1D plots of residue numbers for (top red) observed dehydrons, (middle black) hydrogen bonds, and (bottom blue) modeled dehydrons, for the protein 3FLL. The indices of dispersion I_D for the observed vs. modeled dehydrons are displayed in text. At the very bottom, the average I_D of 20 sets of randomly-sampled dehydrons from this protein is also given.

apart along the mainchain.

V. Conclusions

In this report, we identified dehydrons in 1,874 proteins from the NRPDB set and performed statistical analyses of their relative locations and proximity to other atoms. As a baseline, we constructed a "uniform" model of dehydrons by randomly sampling hydrogen bonds to act as dummy dehydrons, and compared the empirical dehydron distributions with the uniformly sampled dehydron distributions. We found that observed dehydrons are frequently located at the outer surfaces of proteins and are thereby less likely to intersect and less likely to be near many atoms, whereas modeled dehydrons are more clumped within the bulk of the proteins and are located at areas of high atom density. Furthermore, we demonstrated that observed dehydrons were only slightly more disperse than modeled dehydrons along the protein's mainchain. These findings may offer insight into future investigations on the probing of protein interaction sites for use in bioengineering applications.

VI. Appendix

In Figure 14, we compare the number of located dehydrons to the number we'd expect if 15.9% of the hydrogen bonds were dehydrons (following the normal distribution of nonpolar carbon counts, with 19 NPC's defining the dehydron cutoff). The purpose of this plot is to check that most if not all points lie along the $y = x$ diagonal line, indicating that the number of located dehydrons roughly equals the anticipated number for each file. It is interesting in and of itself to see whether certain PDB files have small or large numbers of dehydrons. Certain proteins may have structures which favor the formation of an unusually large number of dehydrons. We find from Figure 14 that most points follow the $y = x$ diagonal. The linear fit has slope very nearly equal to one and y-intercept near zero. There are, however, a large number of points far outside of the diagonal, on both sides. It is probable that these proteins have a low or high preference for dehydrons due to their structure.

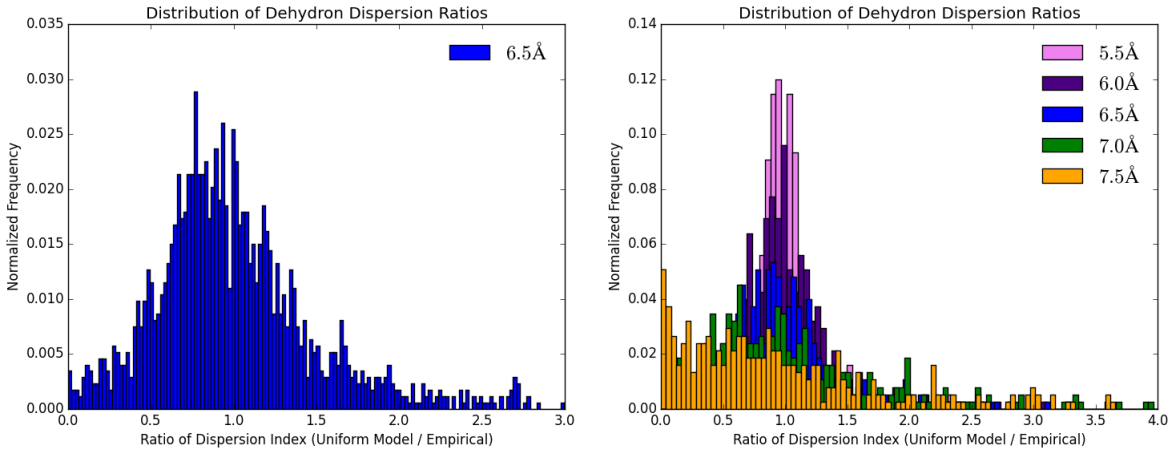

FIG. 13. LEFT: Histogram of the ratios of index of dispersion of observed dehydrons over modeled dehydrons per protein, for domain radius 6.5$\mathring{A}$. RIGHT: Same histograms, but for several domain radii. Here, the dehydron locations were represented by their residue numbers, and the 1D index of dispersion formula was used.

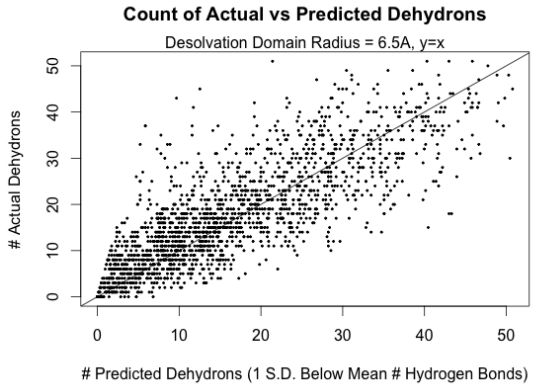

FIG. 14. Scatter plot of the number of located dehydrons per PDB file versus the number of anticipated dehydrons, assuming that the anticipated number is given by 15.9% of the number of located hydrogen bonds (from the 68-95-99.7 rule). The line is a fitted linear regression.

[1] A. Fernandez and L. R. Scott. Dehydron: A structurally encoded signal for protein interaction. *Biophysical Journal*, 85:1914–1928, 2003.

[2] A. Fernandez. Epistructural tension promotes protein associations. *Physical Review Letters*, 108:188102, 2012.

[3] C. M. Fraser, A. Fernandez, and L. R. Scott. Wrappa: A screening tool for candidate dehydron identification. 2014.

[4] L. R. Scott and A. Fernandez. Digital biology. 2014.

[5] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology*, 285:1735–1747, 1999.