Discriminatory Power of Stoichiometry-Driven Protein Folding?

Mihaly Mezei

Department of Structural and Chemical Biology,

Mount Sinai School of Medicine,

New York, New York 10029

E-mail: Mihaly.Mezei@mssm.edu

Phone: (212) 659 5475

There is a saying that when life hands you a lemon, you make lemonade out of it. I found that this has a parallel in molecular modeling: when your test of a putative diagnostic property fails to show discriminatory power, make an invariant out of it. The paper by Mittal, Jayaram, Shenoy and Bawa [1] is a brilliant example of this. Based on three independent observations, comparison of C_α distance distributions, small standard deviation (STD) of the amino-acid propensities and the comparison of amino-acid propensities in structured and unstructured proteins, the paper concluded that the relative frequencies each amino acid occurs in natural proteins is an important contributor to protein stability. This may not necessarily be surprising, but certainly it has not been thought to be the case. Given that these relative frequencies are found to be (more or less) constant, the authors call this set of relative frequencies stoichiometry. I will comment on these observations in the order of strength (as I see it).

In my opinion, the comparison of folded and unfolded proteins in [1] is direct evidence of the importance of the close adherence to the right stoichiometry (as defined by the experimentally observed relative frequencies of amino acids). However, this comparison also points to the fact that the right stoichiometry is only a necessary condition for protein stability but probably not a sufficient one, since it is known that the probability that a random sequence will fold is vanishingly small and restricting random sequences to the right stoichiometry may still leave many that would not fold.

As for the observed STDs of propensities, I was first mildly skeptical that they indicate the importance of this particular stoichiometry, since the values were not that small. This led me to the thought experiment: what values of STDs would one expect if the residues were chosen randomly, with the only restriction that the probability of selecting a give residue is its observed propensity? Selection of residue r_i with probability p_i is described with the binomial distribution

whose STD is given as $np(1-p)$ where n is the sample size. It turns out that the observed STDs are consistently smaller than the STD from the binomial distribution. Table I shows the data from Table I of [1] extended with the STD calculated from the binomial distribution. On the average, the STD's from random sequences are 2.1 times larger than the observed values; the ratios range from 1.2 to 2.8.

Table I
The average percentage occurrence of each amino acid, their STD as observed and as calculated from the binomial distribution

	P(%)	STD (observed)	STD (random)
A	7.8	3.4	7.2
V	7.1	2.4	6.6
I	5.8	2.4	5.5
L	9.0	2.9	8.2
Y	3.4	1.7	3.3
F	3.9	1.8	3.7
W	1.3	1.0	1.3
P	4.4	2.0	4.2
M	2.2	1.3	2.2
C	1.8	1.5	1.8
T	5.5	2.4	5.2
S	6.0	2.5	5.6
Q	3.8	2.0	3.7
N	4.3	2.2	4.1
D	5.8	2.0	5.5
E	7.0	2.7	6.5
H	2.3	1.4	2.2
R	5.0	2.3	4.8
K	6.3	2.8	5.9
G	7.2	2.8	6.7

The distribution of the C_α distances has been examined in great detail and it was found that they follow a sigmoidal distribution that can be described by three parameters only. While different residues suggest higher likelihood of interactions with specific partners, no such distinctions were found in the distributions. This was also true when the short range part of the distributions

were compared – an important point since the long-range part is not expected to show significant residue-dependence. The sigmoidal shape in itself is not surprising – it is a consequence of the finite size of proteins.

Given that the importance of the right stoichiometry for protein stability has been amply demonstrated by the two arguments quoted above, I think that the residue neighborhood distributions are worth revisiting to see if there is a way to tease out data related to the contributions of specific interactions to protein stability. It is important to note what is at stake here: the development of knowledge-based potentials is based on preferential interaction between amino acids, and the determination of protein structure by modeling and simulation, based on preferential contacts and interactions, has become very common (2, 3 and references therein). There are two reasons why the C_α distance distribution is not the best measure of interaction specificity: (a) the cumulative distributions are dominated by the effect of the cubic dependence of volume with distance and (b) different side chains are of different size and thus the contact distances are different for different residue pairs. This suggests two additinal ways of analyzing the residue-residue distances: (a) instead of the cumulative distributions, calculate radial distributions, i.e., normalize by the volume available - this would generate distributions with peaks and troughs whose height and depth would be rather sensitive to small changes in propensities; and (b) calculate the number of residues of different kind that are in contact (i.e., have at least one pair of heavy atoms within VdW distance) with the test residue.

The authors draw a parallel of their observation to the seminal observation of Chargaff related to the fixed ratio of nucleotides in DNA. That ratio soon found its explanation in the double helix structure. Completing the parallel between the Chargaff rules would require the detection of the structural or mechanistic origin of the importance of the right stoichiometry. The alternative

analysis of residue neighborhoods suggested above would be one option. It would be also informative to compare the STD of the ratio of different type of residues with the STD expected from the individual residue propensity STDs.

Reference:

1. A. Mittal, B. Jayaram, S. Shenoy and T. S. Bawa, *J Biomol Struct Dyn 28,* 133-142 (2010).

2. P. Sklenovský, M. Otyepka, *J Biomol Struct Dyn 27*, 521-540 (2010).

3. M. J. Aman, H. Karauzum, M. G. Bowden, T. L. Nguyen, *J Biomol Struct Dyn 28*, 1-12 (2010).