Decoding the Design Principles of Amino Acids and the Chemical Logic of Protein Sequences

B. Jayaram

Department of Chemistry & Supercomputing Facility for Bioinformatics &

Computational Biology, Indian Institute of Technology,

Hauz Khas, New Delhi-110016, India.

Tel No.: +91-11-2659 1505; Fax: +91-11-2658 2037

Email: bjayaram@chemistry.iitd.ac.in

Website: www.scfbio-iitd.res.in

That only 20 amino acids occur naturally accounting for the structural and functional diversity of proteins remains a mystery. We show here that this is a consequence of the action of a symmetry group, identify the presence of hydrogen bond donor groups, presence of sp3 hybridized γ carbons, absence of δ carbons and linearity as properties central to side chain design and quantify the chemical logic of protein sequences.

Proteins are polymers of amino acids. An appraisal of their classification into structural proteins, protective proteins, contractile proteins, transport proteins, storage proteins, toxins, hormones, enzymes etc.[1-3] at once divulges the unlimited potential variety of structures and metabolic activities. The naturally occurring standard amino acids are 20 in number which are the building blocks or chemical templates for proteins. As to why only 20 amino acids are selected has eluded a definitive resolution thus far. From the stand point of mathematics, defining how the number 20 forms a complete set is a problem in enumeration. From a chemistry point of view, the 20 amino acid side chains constitute a near comprehensive complete chemical template (monomer) library to build polymers with the diversity of functions indicated above. From a biology perspective, the 20 amino acids contain all the information necessary for the regulation of gene expression - involving protein-nucleic acid interactions - to cite just one crucial example. The degeneracy arising from the (4^3) 64 possible codons to give 20 amino acids is a problem which is understood partly by wobble hypothesis[4] and the rule of conjugates[5]. The issue in focus here is not degeneracy in the genetic code[6] but the nature of residues viz. the underlying chemical properties and the design principles that lead to only 20 amino acids. Conventional classifications of amino acid side chains into hydrophobic, polar and charged residues have been extremely useful in understanding some aspects of the structure and function of proteins but do not help in explaining why 20 forms a complete set of chemical templates for building proteins. Also, the rationale for their design has been beyond reach.

Treating the problem as the action of a symmetry group, we found the symmetries of an equilateral triangle to fit the description quite well. For instance, if four colours (red, blue green, yellow) are used to colour the three edges (1,2,3) of a triangle, 64 coloured triangles $\{(\text{red}_1,\text{red}_2,\text{red}_3), (\text{red}_1,\text{red}_2,\text{blue}_3), (\text{red}_1,\text{blue}_2,\text{red}_3), \ldots\}$ can be made overall. By virtue of the symmetries of the triangle such as rotations and reflections (action of D_3 group)[7,8], only 20 remain as distinct and unique. Setting up the analogy between the 20 unique triangles and the 20 amino acids with the hypothesis that evolution must have an underlying logic and that the naturally occurring amino acids constitute a non-redundant complete set for all practical purposes of protein structure and function, we arrive at the following side chain design principles.

Rule 1. Amino acid side chains have evolved based on four chemical properties. A minimum of one and a maximum of three properties are used to specify each amino acid.

Rule 2. Each property occurs in exactly 10 amino acids.

Rule 3. Any two properties occur simultaneously in only four amino acids.

Rule 4. Any three properties occur simultaneously in only one amino acid.

Rule 5. Amino acids characterized by a single property occur only once.

The above rules explain not only the occurrence of exactly 20 amino acids but also the chemical logic behind their design. The challenge at this stage, to ascertain that this is not a mere coincidence, is to identify the four chemical properties and the corresponding amino acids which obey the above rules (Table 1).

One property (I) that suggests itself at once is the presence of sp^3 hybridized γ carbon atom. Only 10 amino acids {E, I, K, L, M, P, Q, R, T, V} possess this property as required by Rule 2 above. Proceeding further, another property (II) possessed again exactly by 10 amino acids {C, H, K, N, Q, R, S, T, W, Y} is the hydrogen bond donor ability. It may be noted that only four amino acids (K, Q, R, T) exhibit both properties (I & II together) as required by Rule 3. A third property (III) embedded in exactly 10 amino acids {A, C, D, G, I, M, N, S, T, V} is the absence of δ carbon. Ile is included in this set as one of the branches of its side chain is lacking in a δ carbon. Revisiting Rule 3 for consistency of this third property, I and III occur simultaneously in only four amino acids (I, M, T, V) and similarly II and III occur simultaneously in only four amino acids (C, N, S, T). Rule 4 requires that the above three properties (I, II and III) occur simultaneously in only one amino acid (T) and this conforms to the expectation. The adherence of the properties identified thus far to the rules proposed without exception is worth noting. The fourth property has been a bit elusive to identify. For a unique specification of the amino acid side chains, this property must address branching. The most likely candidate (IV) is the linearity of the side chains / non-occurrence of bidentate forks with terminal hydrogens in the side chains. This immediately pools together 10 amino acids in the set {A, D, E, F, H, K, M, P, S, Y}. Side chains with single rings are treated as without forks. The sulfhydryl group in Cys and its ability to form disulfide bridges requires it to be treated as forked. Accepting that this property (IV) satisfies Rule 2, it may be noted that Rule 3 is satisfied by I and IV (E, K, M, P); by II and IV (H, K, S, Y) and by III and IV

(A, D, M, S). Also, Rule 4 is satisfied by I, II and IV (K), by I, III and IV (M) and by II, III and IV (S). With all the four properties (I, II, III and IV) specified, it may be verified that amino acids characterized by a single property occur only once: property I (L), property II (W), property III (G) and property IV (F), consistent with Rule 5.

The attributes of amino acid side chains identified here deal with a few unique chemical properties of the side chains namely, the conformational flexibility (**g**), the hydrogen bond donor ability (**d**), the size (**s**) and the shape (**l**).

Rule 2a. Each property occurs in exactly ten amino acids: thrice in one amino acid, twice in three amino acids and once in six amino acids.

A provisional assignment consistent with the above rule is provided in the last column of Table 1 (Also please see supplementary information for details on the assignment).

The analysis presented here namely, the postulation of rules for side chain design, the identification of certain unique chemical properties consistent with the rules, an assignment of the properties to each amino acid is an attempt to crack the chemical code of amino acids. It not only answers a fundamental question concerning the occurrence of only 20 amino acids but unveils a plausible strategy adopted by nature in designing unique chemical templates for polypeptide chains for their necessary structural and functional diversity. It is conceivable that a different self-consistent set of properties satisfying the rules proposed could be discovered or the assignments further tidied. This notwithstanding, the remarkable conformity of the rules and the properties to the predictions of the symmetry group action on the set of possible chemical templates presents a compelling case for a re-examination of the 'essential' chemical properties of the amino acids and to think differently about amino acids. A case can be made that the hydrophilic and hydrophobic properties in the common classification are compound properties and not primary.

Since each amino acid could be represented as a four dimensional vector (last column of Table 1), composition rules for these vectors could be developed for mapping polypeptide chains to probe the logic of protein sequences.

A computational analysis of 159,614 protein sequences adapted from swissprot[9] was undertaken based on the chemical properties of amino acid side chains introduced above. Quite interestingly, naturally occurring protein sequences are characterized (Table 2) by a very high occurrence of amino acids with sp3 γ carbons and short side chains relative to randomly generated polypeptide sequences. Hydrogen bond donating side chains are heavily under-represented as also to a lesser extent the unbranched amino acid side chains. Pursuing these observations and utilizing the differences between protein sequences and random sequences (Table 2), an algorithm has been developed to distinguish between genes coding for proteins and non-gene regions in genomic sequences[10]. An examination of 239418 DNA sequences from 331 prokaryotic genomes annotated as genes in ncbi databank[11] was undertaken. Statistics on the DNA sequences identified as genes coding for proteins based on the chemical logic are reported in Table 3. The accuracies (prediction sensitivities >90%) attained by the chemical model are comparable to some of the popular algorithms based on sophisticated statistical and mathematical models[10,12]. While there may be some uncertainties in genome annotation itself and hence in the calculated accuracy indices of the chemical model, the overall satisfactory performance on the prokaryotic genomes is noteworthy. The gene evaluation algorithm based on the chemical logic of polypeptide sequences is web-enabled and the software made available for free access at (www.scfbio-iitd.res.in/progenie)

In a nutshell, a novel chemical analysis of amino acid side chains explains the occurrence of 20 unique chemical templates as a complete set for making functional polypeptide chains, facilitates genome analysis with very high accuracies. It is conceivable that the chemical model presenting a clue to the language of amino acids, could facilitate a better understanding of the structure and function of proteins and structure based drug design efforts.

References

1. Creighton, T. E., *Proteins: Structures and Molecular Properties*. 2nd edition, Ch-1, p-1 (W.H. Freeman and Company, New York, 1993).

2. Schulz, G. E., Schirmer, R. H., *Principles of Protein Structure*, ch-1, p-1 (Springer-Verlag, New York ,1984).

3. Branden, C., and Tooze, J., *Introduction to Protein Structure*. (Garland Publishing, Inc., New York, 1991).

4. Crick, F.H.C. *J. Mol. Biol.*, **19**, 548-555 (1966).

5. Jayaram, B., *J. Molecular Evolution*, **45**, 704-705 (1997).

6. Watson, J.D., Hopkins, N.H., Roberts, J. W., Steitz, J. A., Weiner, A. M.. *Molecular Biology of the Gene*. 4th Edition, Vol.1, ch-15, (The Benjamin Cummings, CA, 1987).

7. Fraleigh, J. B., *A First Course in Abstract Algebra*. 3rd edition, p-162 (Addison-Wesley, Reading, Massachusetts,1982).

8. Gilbert, W. J., *Modern Algebra with Applications*. p-104. (John Wiley & Sons, New York, 1976).

9. http://ca.expasy.org/sprot/

10. D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, ch-8. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001),

11. ftp://ftp.ncbi.nih.gov/genomes/Bacteria/.

12. Binnewies, T. T., Y. Motro, P. F. Hallin, O. Lund, D. La. T. Dunn, D. J. Hampson, M. Bellgard, T. M. Wassenaar, and D. W. Ussery. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics* **6**, 165-185 (2006).

Supplementary Information is linked to the online version of the paper at the nature chemical biology site.

Acknowledgements

Funding from the Department of Biotechnology and the Department of Information Technology is gratefully acknowledged. The contribution of Mr. Anuj Gupta and Mr. Debashis Sahu in web-enabling the genome analysis programs, Ms. Poonam Singhal in validating the results presented, Prof. D. L. Beveridge, Drs. N. D. Kurur, V. Haridas, N. G. Ramesh, N. Pant and Prof. I. Ghosh for a critical reading of the manuscript is gratefully acknowledged.

Table 1. The 20 amino acids and the unique chemical properties of their side chains

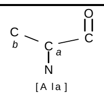
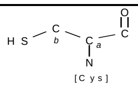
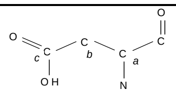
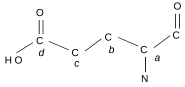
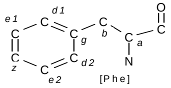
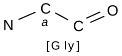
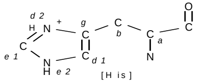
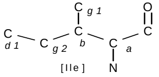
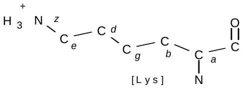
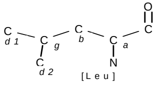
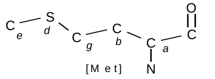
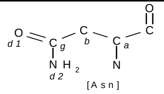
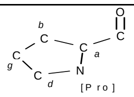
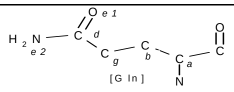
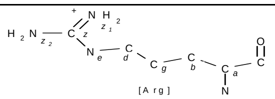
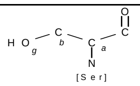
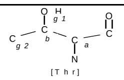
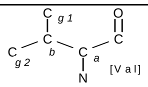
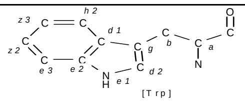
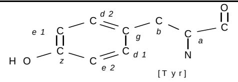
Amino acid	Molecular Structural Formula	I. sp^3 hybridized γ carbon (**g**)	II. Hydrogen bond donor group (**d**)	III. Short / Absence of δ carbon (**s**)	IV. Linearity / Absence of forks with hydrogens (**l**)	Assignment[#]
A Alanine		-	-	+	+	$\mathbf{g_0d_0}s_2l_1$
C Cysteine		-	+	+	-	$\mathbf{g_0}d_1\mathbf{s_2l_0}$
D Aspartate		-	-	+	+	$\mathbf{g_0d_0}s_1l_2$
E Glutamate		+	-	-	+	$\mathbf{g_1d_0}s_0l_2$
F Phenylalanine		-	-	-	+	$\mathbf{g_0d_0}s_0l_3$
G Glycine		-	-	+	-	$\mathbf{g_0d_0}s_3\mathbf{l_0}$
H Histidine		-	+	-	+	$\mathbf{g_0}d_2\mathbf{s_0}l_1$
I Isoleucine		+	-	+	-	$g_2\mathbf{d_0s_1l_0}$
K Lysine		+	+	-	+	$g_1d_1\mathbf{s_0}l_1$
L Leucine		+	-	-	-	$g_3\mathbf{d_0s_0l_0}$

Table 1. (continued)

M Methionine		+	-	+	+	$\mathbf{g_1d_0s_1l_1}$
N Asparagine		-	+	+	-	$\mathbf{g_0d_2s_1l_0}$
P Proline		+	-	-	+	$\mathbf{g_2d_0s_0l_1}$
Q Glutamine		+	+	-	-	$\mathbf{g_1d_2s_0l_0}$
R Arginine		+	+	-	-	$\mathbf{g_2d_1s_0l_0}$
S Serine		-	+	+	+	$\mathbf{g_0d_1s_1l_1}$
T Threonine		+	+	+	-	$\mathbf{g_1d_1s_1l_0}$
V Valine		+	-	+	-	$\mathbf{g_1d_0s_2l_0}$
W Tryptophan		-	+	-	+	$\mathbf{g_0d_3s_0l_0}$
Y Tyrosine		-	+	-	+	$\mathbf{g_0d_1s_0l_2}$

‘+’ indicates that the property is satisfied and –’ indicates that the property is not satisfied, in columns 3 to 6.

# (Column 7) Subscript refers to the number of times each property occurs in the corresponding amino acid (Please see supplementary information for details on assignment).

Table 2. A characterization of the chemical logic of 159,614 protein sequences

Property	% sequences satisfying the property	*Average value of the property	Variance (2 σ)
Excess g	93.5	6.0	1.1
Excess d	3.1	-8.6	1.9
Excess s	75.2	3.4	0.8
Excess l	35.1	-1.4	0.3
(gl-ds) >0	85.0	-	-
(gd-ls)>0	34.8	-	-
(gs-dl)>0	97.7	-	-

*Excess values of averages and standard deviation for g, d, s and l are reported relative to random sequences of length 100 amino acids for which the value for each property is 33.8. (Please see supplementary information for further details).

Table 3. An evaluation of protein (Swissprot[9]) and genic (ncbi[11]) sequences based on the chemical logic of protein sequences.

	Swissprot Sequences#	Nucleotide sequences annotated as genes*	Intergenic (Nongene) sequences*	Random polypeptide sequences
Total Number considered	157210	239418	204047	10000
Number of proteins identified	141784	227033	14699	806

#Prediction Sensitivity = **0.92;**

*Prediction Sensitivity = **0.96**; Specificity = 0.91; Correlation coefficient = 0.86

#*A definition of the statistical indices is provided in the supplementary information.