World Academy of Science, Engineering and Technology
International Journal of Chemical and Molecular Engineering
Vol:5, No:6, 2011

All Proteins Have a Basic Molecular Formula

Homa Torabizadeh

***Abstract*—**This study proposes a basic molecular formula for all proteins. A total of 10,739 proteins belonging to 9 different protein groups classified on the basis of their functions were selected randomly. They included enzymes, storage proteins, hormones, signalling proteins, structural proteins, transport proteins, immunoglobulins or antibodies, motor proteins and receptor proteins. After obtaining the protein molecular formula using the ProtParam tool, the H/C, N/C, O/C, and S/C ratios were determined for each randomly selected sample. In this case, H, N, O, and S coefficients were specified per carbon atom. Surprisingly, the results demonstrated that H, N, O, and S coefficients for all 10,739 proteins are similar and highly correlated. This study demonstrates that despite differences in the structure and function, all known proteins have a similar basic molecular formula $C_nH_{1.58 \pm 0.015n}N_{0.28 \pm 0.005n}O_{0.30 \pm 0.007n}S_{0.01 \pm 0.002n}$. The total correlation between all coefficients was found to be 0.9999.

***Keywords*—**Protein molecular formula, Basic unit formula, Protparam tool.

I. Introduction

PROTEINS are macromolecules involved in many important biological structures and functions in living organisms. Examples of structural proteins are ligaments, fingernails, and hair; examples of functional proteins are digestive enzymes and muscle proteins that bring about metabolism and locomotion, respectively. Proteins are biopolymers constructed of amino acid monomers. The sequence of amino acids of a specific protein is determined by the sequence of the nucleic acid bases in the gene that encodes the protein.

The chemical properties of the constituent amino acids determine the biological activity of the protein (Fig. 1) [1]. Proteins play extremely important roles in biological systems [2]. The proteins present in single cell microorganisms, plants, animals, and humans differ greatly. The different number and sequence of amino acids result in the unique shapes, structures, and functions of proteins. Proteins are task and location-specific, and are broadly classified on the basis of structure and function. Structural classifications are based on protein folding, motifs, and protein family information. Functional classifications are based on biochemical and cellular roles, metabolic pathways, subcellular localization, and molecular interactions. Functional classifications are

H. Torabizadeh is with the Iranian Research Organization for Science and Technology, Institute of Chemical Technology, Department of Food Science and Technology, Enghelab Ave., No. 27, Forsat St., 15819 Tehran, Iran (corresponding author to provide phone: +98 88828052-7; fax:+98 2282276265; e-mail: htoraby@ alumni.ut.ac.ir).

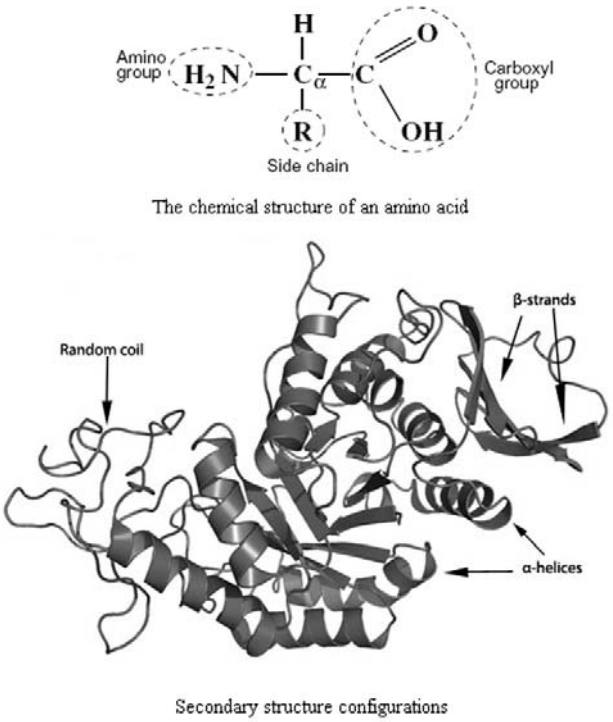

Fig. 1. Chemical structure of an amino acid and the secondary structure configurations of a protein.

further subdivided on the basis of similarities in enzyme reaction mechanisms, participation in biochemical pathways, functional roles, and cellular localization [3]. Although structural classifications are probably well defined on the basis of the criterion of molecular similarity, their overlap is surprisingly limited. On the other hand, functional classifications encompass many processes and elements, ranging from pathways to cellular compartments. These functional classifications have been shown to overlap considerably with each other [4]. Functionally, proteins are classified into the following: enzymes (proteins that catalyze chemical and biochemical reactions inside and outside living cells), storage proteins (proteins involved in storing energy that can be released during metabolic processes), hormones (proteins responsible for the regulation of many biochemical processes in organisms), signalling proteins (proteins involved in signal translation processes), structural proteins (these proteins are maintain structures of other biological components, like cells and tissues), transport proteins (proteins involved in transporting or storing chemical compounds and ions), immunoglobulins or antibodies (proteins involved in the immune response of an organism against large foreign molecules, which are introduced, for

example, by an infection), motor proteins (proteins involved in conversion of chemical energy to mechanical energy), and receptor proteins (proteins responsible for signal detection and translation into other type of signals). Each of these classified proteins acts on its own particular activities [5]. This study compares the coefficients of hydrogen, nitrogen, oxygen, and sulfur with respect to the quantity of carbon atoms in the protein. Data were obtained from bioinformatics databases using data extraction tools for more than 10,000 random protein samples. The results of this investigation indicate that a single basic molecular formula can be used to describe the relative amounts of elements present in all proteins.

II. Methods

A. Databases and Definition of the Protein Formula

In this study, the protein formula was determined for each type of protein by using bioinformatics tools and databases. It was performed by first obtaining the amino acid sequence and entering it in an available calculation tool. This allows for the formula of the individual protein to be determined. All analyses were performed using the Expasy select as a set of experimentally determined protein profiles in the UniProtKB

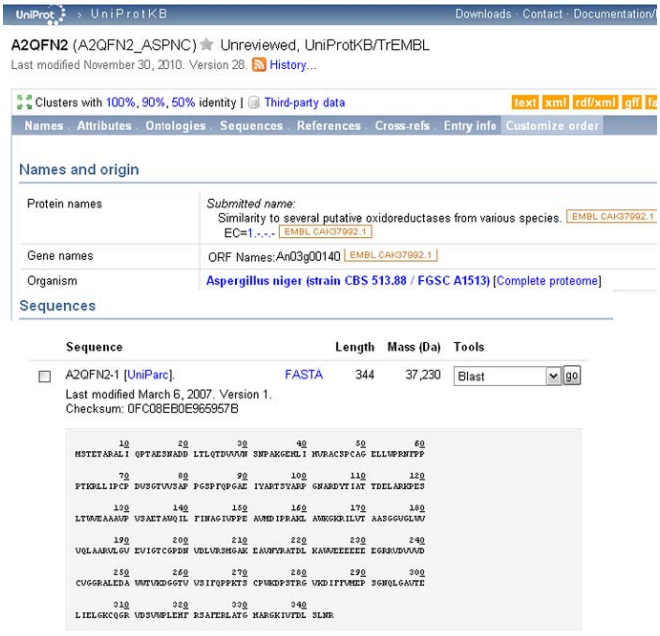

Fig. 2. Input form of the searching tool UniProtKB, for oxidoreductase from Aspergillus niger (Accession code: A2QFN2) available at http://www.uniprot.org/uniprot/A2QFN2.

(Universal Protein Resource Knowledgebase) database http://www.expasy.org/sprot/and http://www.expasy.org/tools) located in the Expasy server. For example, in the case of oxidoreductase enzymes, one of the 800 randomly selected enzymes was the oxidoreductase from *Aspergillus niger* (UniProt ID: A2QFN2) (Fig. 2). This enzyme has 344 amino acids and a molecular mass of 37,230 Dalton. At first, the amino acid sequence of the enzyme was selected and copied. The sequence was then pasted onto the calculation window of the ProtParam tool for determination of the protein formula (Fig. 3).

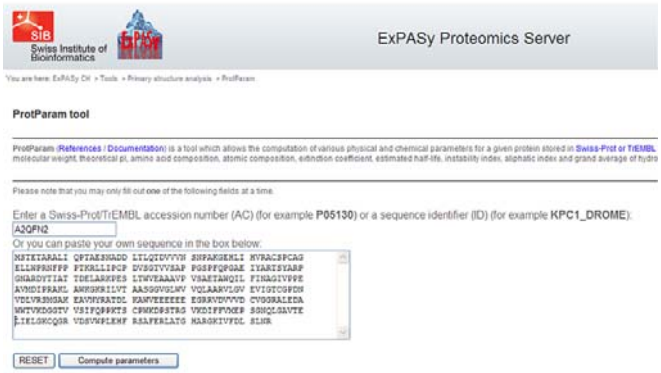

Fig. 3. Input form of the characterization tool ProtParam, available at http://www.expasy.org/tools/protparam.html.

The computed protein formula was then extracted (Fig. 4).

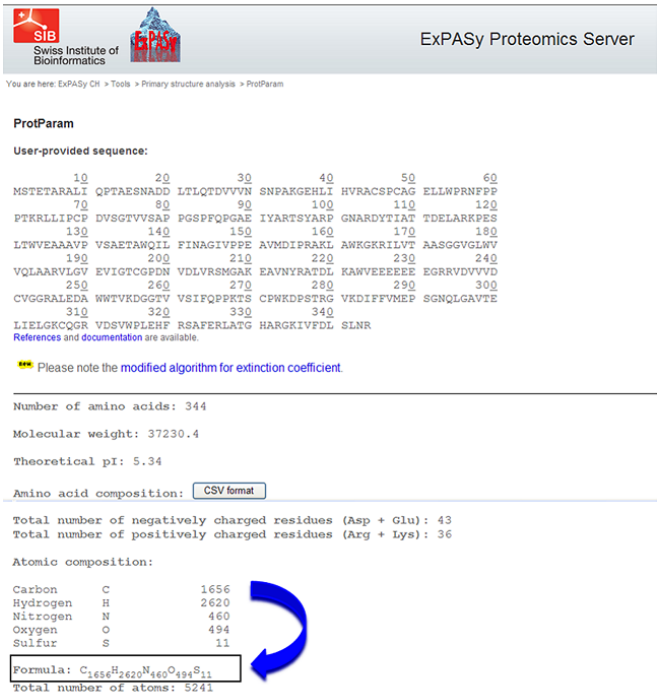

Fig. 4. Input form of the characterization tool ProtParam, available at http://www.expasy.org/cgi-bin/protparam.

In this study, 10,739 different proteins were selected randomly from among 9 groups of proteins. Starting from the N-terminus of each protein sequence, a running window of n amino acids was selected. The specified amino acid sequence of each protein was then copied and pasted onto the calculation window of the ProtParam program. Eventually, in addition to the calculated profiles of the desired proteins, a

related molecular formula was determined. The total numbers of hydrogen, nitrogen, oxygen, and sulfur atoms were then divided by the total number of carbon atoms. In this manner, the coefficients of carbon, hydrogen, nitrogen, oxygen, and sulfur were obtained. These coefficients indicate the number of hydrogen, nitrogen, oxygen, and sulfur atoms in the protein structure with respect to the number of carbon atoms. This provides the ability to assemble a basic formula for each protein sample. Table I is an example of resulted formula for oxidoreductase enzyme.

TABLE I
EXAMPLE OF RESULTED BASIC FORMULA FOR OXIDOREDUCTASE ENZYMES

Enzymes, Oxidoreductases: 800 samples		
Accession Code	Protein Formula	Basic Molecular Formula
A5D4R8	$C_{3142}H_{5015}N_{903}O_{907}S_{30}$	$C_n H_{1.60n} N_{0.29n} O_{0.29n} S_{0.010n}$
A5D560	$C_{1926}H_{3107}N_{583}O_{548}S_{17}$	$C_n H_{1.61n} N_{0.30n} O_{0.28n} S_{0.009n}$
A5D616	$C_{1712}H_{2741}N_{517}O_{481}S_{18}$	$C_n H_{1.60n} N_{0.30n} O_{0.28n} S_{0.011n}$
A5D684	$C_{1791}H_{2811}N_{483}O_{521}S_{13}$	$C_n H_{1.57n} N_{0.27n} O_{0.29n} S_{0.007n}$
A5D685	$C_{1226}H_{1925}N_{347}O_{364}S_{16}$	$C_n H_{1.57n} N_{0.28n} O_{0.30n} S_{0.013n}$
A5D686	$C_{879}H_{1426}N_{250}O_{260}S_3$	$C_n H_{1.62n} N_{0.28n} O_{0.30n} S_{0.003n}$
A5E8U7	$C_{1944}H_{3137}N_{557}O_{560}S_{11}$	$C_n H_{1.61n} N_{0.29n} O_{0.29n} S_{0.006n}$
A6SYB5	$C_{1630}H_{2616}N_{448}O_{477}S_4$	$C_n H_{1.60n} N_{0.27n} O_{0.29n} S_{0.002n}$
A6TA35	$C_{1588}H_{2514}N_{418}O_{474}S_{15}$	$C_n H_{1.58n} N_{0.26n} O_{0.30n} S_{0.009n}$
A6TAB8	$C_{1765}H_{2748}N_{518}O_{517}S_2$	$C_n H_{1.56n} N_{0.29n} O_{0.29n} S_{0.001n}$
A7K3A5	$C_{980}H_{1546}N_{258}O_{297}S_9$	$C_n H_{1.58n} N_{0.26n} O_{0.30n} S_{0.009n}$
A7K3A6	$C_{870}H_{1384}N_{240}O_{259}S_6$	$C_n H_{1.59n} N_{0.28n} O_{0.30n} S_{0.007n}$
A7K531	$C_{1741}H_{2722}N_{482}O_{521}S_{12}$	$C_n H_{1.56n} N_{0.28n} O_{0.30n} S_{0.007n}$
A7WHY8	$C_{1323}H_{2091}N_{333}O_{379}S_{10}$	$C_n H_{1.58n} N_{0.25n} O_{0.29n} S_{0.008n}$
A7ZQY9	$C_{1264}H_{2029}N_{357}O_{389}S_8$	$C_n H_{1.61n} N_{0.28n} O_{0.31n} S_{0.006n}$
A8A422	$C_{1262}H_{2027}N_{357}O_{387}S_8$	$C_n H_{1.61n} N_{0.28n} O_{0.31n} S_{0.006n}$
A9HMF9	$C_{2681}H_{4274}N_{802}O_{791}S_{18}$	$C_n H_{1.59n} N_{0.30n} O_{0.30n} S_{0.007n}$
A9HRN4	$C_{2565}H_{3995}N_{747}O_{760}S_{23}$	$C_n H_{1.56n} N_{0.29n} O_{0.30n} S_{0.009n}$
A9HX35	$C_{1804}H_{2804}N_{516}O_{532}S_{11}$	$C_n H_{1.55n} N_{0.29n} O_{0.29n} S_{0.006n}$
A9HXI4	$C_{1629}H_{2597}N_{467}O_{475}S_8$	$C_n H_{1.59n} N_{0.29n} O_{0.29n} S_{0.005n}$

III. RESULTS

The resulting coefficients allow a basic formula to be extracted for each selected protein. Finally, a basic unit formula was obtained for each classification of protein (e.g., enzymes) by determining the average of the coefficients. The results are shown in Table II.

TABLE II
FINAL AVERAGED FORMULA FOR EACH OF THE 9 PROTEIN CLASSIFICATIONS

	Proteins	Samples	Basic Molecular Formula
Enzymes	Oxidoreductase	800	$C_nH_{1.58n}N_{0.27n}O_{0.299n}S_{0.008n}$
	Transferase	1030	$C_nH_{1.58n}N_{0.28n}O_{0.29n}S_{0.008n}$
	Hydrolase	1030	$C_nH_{1.56n}N_{0.27n}O_{0.29n}S_{0.007n}$
	Lyase	1050	$C_nH_{1.59n}N_{0.28n}O_{0.30n}S_{0.008n}$
	Isomerase	280	$C_nH_{1.57n}N_{0.275n}O_{0.30n}S_{0.018n}$
	Ligase	500	$C_nH_{1.58n}N_{0.28n}O_{0.30n}S_{0.009n}$
Storage proteins		380	$C_nH_{1.57n}N_{0.27n}O_{0.30n}S_{0.009n}$
Hormones		1020	$C_nH_{1.57n}N_{0.27n}O_{0.296n}S_{0.016n}$
Signalling proteins		450	$C_nH_{1.58n}N_{0.28n}O_{0.30n}S_{0.011n}$
Structural proteins		1030	$C_nH_{1.58n}N_{0.27n}O_{0.30n}S_{0.01n}$
Transport proteins		1100	$C_nH_{1.59n}N_{0.27n}O_{0.29n}S_{0.008n}$
Immunoglobulines		612	$C_n H_{1.57n} N_{0.28n} O_{0.31n} S_{0.007n}$
Motor Proteins		693	$C_n H_{1.62n} N_{0.28n} O_{0.31n} S_{0.008n}$
Receptor Proteins		764	$C_n H_{1.56n} N_{0.27n} O_{0.29n} S_{0.01n}$
Average Formula		10739	$C_n H_{1.58n} N_{0.28n} O_{0.30n} S_{0.01n}$

The average of the coefficients that indicate the carbon, hydrogen, nitrogen, oxygen, and sulfur contents were determined, present in Table III. Then the total average of mean coefficients for each atom was calculated. As it is shown in table ш. In this table, the total correlation between the coefficients was 0.9999.

TABLE III
EXTRACTED DATASETS BASED ON THE AVERAGED COEFFICIENTS OF C, H, N, O, AND S FOR EACH OF THE 9 PROTEIN CLASSIFICATIONS

	Proteins	Coefficients C	H	N	O	S
Enzymes	Oxidoreductase	1	1.58	0.27	0.30	0.008
	Transferase	1	1.58	0.28	0.29	0.008
	Hydrolase	1	1.56	0.27	0.29	0.007
	Lyase	1	1.59	0.28	0.30	0.008
	Isomerase	1	1.57	0.28	0.30	0.018
	Ligase	1	1.58	0.28	0.30	0.009
Storage proteins		1	1.57	0.27	0.30	0.009
Hormones		1	1.57	0.27	0.30	0.016
Signalling proteins		1	1.58	0.28	0.30	0.011
Structural proteins		1	1.58	0.27	0.30	0.010
Transport proteins		1	1.59	0.27	0.29	0.008
Immunoglobulines		1	1.57	0.28	0.31	0.007
Motor Proteins		1	1.62	0.28	0.31	0.008
Receptor Proteins		1	1.56	0.27	0.29	0.010
Average		1	1.58	0.28	0.30	0.010
Total correlation		0.9999				

Now, the question may comes to mind that, is it possible that the high similarity between the coefficients of these atoms be a result of the similarity between the coefficients of the 21 known amino acids which are present as building blocks in all proteins? To finding the answer of this question and evaluate the success of the prediction, it was necessary to compare the determined coefficients of protein elements (H, N, O, and S) with the coefficients that were obtained from the formula of individual amino acids. Calculations performed on the molecular formula of individual amino acids indicate that they are collectively not described by a single basic formula. Table IV, provides the molecular and basic formula based on the calculated coefficients for each of the amino acids.

TABLE IV
CLASSIFICATION AND BASIC FORMULA EXTRACTED FROM THE MAIN AMINO ACIDS IN PROTEINS

Amino acids	Molecular formula	Basic molecular formula of amino acids
Glycine	$C_2H_5NO_2$	$C_nH_{2.5n}N_{0.5n}O_n$
Alanine	$C_3H_7NO_2$	$C_nH_{2.33n}N_{0.33n}O_{0.67n}$
Valine	$C_5H_{11}NO_2$	$C_nH_{2.2n}N_{0.2n}O_{0.4n}$
Leucine	$C_6H_{13}NO_2$	$C_nH_{2.17n}N_{0.17n}O_{0.33n}$
Isoleucine	$C_6H_{13}NO_2$	$C_nH_{2.17n}N_{0.17n}O_{0.33n}$
Serine	$C_3H_7NO_3$	$C_nH_{2.33n}N_{0.33n}O_n$
Threonine	$C_4H_9NO_3$	$C_nH_{2.25n}N_{0.25n}O_{0.75n}$
Cysteine	$C_3H_7NO_2S$	$C_nH_{2.33n}N_{0.33n}O_{0.67n}S_{0.3n}$
Cystine	$C_6H_{12}N_2O_4S_2$	$C_nH_{2.0n}N_{0.33n}O_{0.67n}S_{0.33n}$
Methionine	$C_5H_{11}NO_2S$	$C_nH_{2.2n}N_{0.2n}O_{0.4n}S_{0.2n}$
Phenylalanine	$C_9H_{11}NO_2$	$C_nH_{1.22n}N_{0.11n}O_{0.22n}$
Tyrosine	$C_9H_{11}NO_3$	$C_nH_{1.22n}N_{0.11n}O_{0.33n}$
Tryptophan	$C_{11}H_{12}N_2O_2$	$C_nH_{1.09n}N_{0.18n}O_{0.18n}$
Proline	$C_5H_9NO_2$	$C_nH_{1.80n}N_{0.2n}O_{0.4n}$
Asparagine	$C_4H_8N_2O_3$	$C_nH_{2.0n}N_{0.25n}O_{0.75n}$
Glutamine	$C_5H_{10}N_2O_3$	$C_nH_{2.0n}N_{0.2n}O_{0.6n}$
Aspartic acid	$C_4H_7NO_4$	$C_nH_{1.75n}N_{0.25n}O_n$
Glutamic acid	$C_5H_9NO_4$	$C_nH_{1.80n}N_{0.2n}O_{0.8n}$
Lysine	$C_6H_{14}N_2O_2$	$C_nH_{2.33n}N_{0.33n}O_{0.33n}$
Histidine	$C_6H_9N_3O_2$	$C_nH_{1.5n}N_{0.5n}O_{0.33n}$
Arginine	$C_6H_{14}N_4O_2$	$C_nH_{2.33n}N_{0.66n}O_{0.33n}$

The numbers presented in Table V were obtained from coefficients that were determined for each atom in the molecular formula for each amino acid. Then, the average of each column was specified and total correlation between the resulting coefficients for each of the amino acids was calculated. This correlation was found to be 0.896, (Table V).

TABLE V
EXTRACTED DATASETS BASED ON THE COEFFICIENTS OF C, H, N, O, AND S FOR EACH OF THE 21 AMINO ACIDS AND THEIR TOTAL CORRELATIONS

	Coefficients				
Amino acids	C	H	N	O	S
Glycine	1	2.50	0.50	1.00	-
Alanine	1	2.33	0.33	0.67	-
Valine	1	2.20	0.20	0.40	-
Leucine	1	2.17	0.17	0.33	-
Isoleucine	1	2.17	0.17	0.33	-
Serine	1	2.33	0.33	1.00	-

TABLE V
CONTINUE

	Coefficients				
Amino acids	C	H	N	O	S
Threonine	1	2.25	0.25	0.75	-
Cysteine	1	2.33	0.33	0.67	0.33
Cystine	1	2.00	0.33	0.67	0.33
Methionine	1	2.20	0.20	0.40	0.20
Phenylalanine	1	1.22	0.11	0.22	-
Tyrosine	1	1.22	0.11	0.33	-
Tryptophan	1	1.09	0.18	0.18	-
Proline	1	1.80	0.20	0.40	-
Asparagine	1	2.00	0.25	0.75	-
Glutamine	1	2.00	0.20	0.60	-
Aspartic acid	1	1.75	0.25	1.00	-
Glutamic acid	1	1.80	0.20	0.80	-
Lysine	1	2.33	0.33	0.33	-
Histidine	1	1.50	0.50	0.33	-
Arginine	1	2.33	0.66	0.33	-
Average	1	1.943	0.266	0.569	0.287
Total correlation	0.896				

The standard deviations of averaged coefficients of the amino acids and the averaged coefficients for all 10,739 proteins are shown in Table VI.

TABLE VI
THE STANDARD DEVIATIONS OF AVERAGED COEFFICIENTS FOR ALL PROTEIN SAMPLES AND AMINO ACIDS

	Proteins	Amino acids
H coefficients average	1.580	1.940
STDEV of H coefficients	0.015	0.410
N coefficients average	0.280	0.270
STDEV of N coefficients	0.005	0.140
O coefficients average	0.300	0.570
STDEV of O coefficients	0.007	0.260
S coefficients average	0.010	0.290
STDEV of S coefficients	0.002	0.075

Comparing the standard deviations of proteins and amino acid coefficients reveal that the data points tend to be very close to the mean as it mentions in equations 1, 2, 3, and 4. Unlike to the proteins, these values for amino acid coefficients are spread out over a larger range of values around the mean.

$$\Sigma_H/\Sigma_C = 1.58 \pm 0.015 \quad (1)$$
$$\Sigma_N/\Sigma_C = 0.28 \pm 0.005 \quad (2)$$
$$\Sigma_O/\Sigma_C = 0.30 \pm 0.007 \quad (3)$$
$$\Sigma_S/\Sigma_C = 0.01 \pm 0.002 \quad (4)$$

Thus, for 99.9% of the proteins examined, the molecular formula is:

$C_nH_{1.58 \pm 0.015n}N_{0.28 \pm 0.005n}O_{0.30 \pm 0.007n}S_{0.01 \pm 0.002n}$

For example, if the carbon coefficient be 100, the protein formula would be:

$C_{100}H_{156.50 - 159.50}N_{27.50 - 28.50}O_{29.30 - 30.70}S_{0.8 - 1.20.}$

IV. Conclusions

As a frontier research area, bioinformatics has developed substantially over the past few decades. This study takes advantage of the bioinformatics databases and tools to specify a basic molecular formula for proteins, which indicates the ratio of the constituent elements C, H, N, O, and S. This study demonstrates that despite differences in the structure and function, all known proteins are constructed on the basis of a single basic molecular formula. It is thought that, the formation of all proteins by translation from mRNA is based on a unique pattern in which the ratio of the coefficients of the contributing elements with respect to each other is kept constant. It believed that, all proteins are formed based on a similar basic molecular formula. There is significant diversity in transcription and rRNA-specific translation at the ribosome for production of proteins based on the specific genetic code because certain sequences of amino acids have completely different structural and functional properties. The survey results were quite surprising because, despite this extremely high level of diversity among the molecular structures of all proteins, for each 1.0 carbon atom there are 1.57–1.60 hydrogen atoms, 0.28–0.29 nitrogen atoms, 0.29–0.31 oxygen atoms, and 0.01 sulfur atom in all proteins. As a result, the smallest and the largest protein molecules have the same basic molecular formula. The specified formula may be applied to protein modelling prediction tasks for scientific research. Furthermore, another interesting research topic appears to be the development of a tool for designing new proteins and peptides for therapeutic, pharmaceutical, and industrial purposes. This is not a saturated research area and further research is ongoing.

Acknowledgment

Many thanks to Pegah Abdollahi and Zohre Torabian for their precious assistance in protein molecular formula determination.

References

[1] L. Jin, W. Fang, "On A comment on 'Prediction of protein structural classes by a new measure of information discrepancy," *Comput. Biol. Chem*., vol. 33, no. 6, pp. 469–470,December. 2009.
[2] L. Palopoli, S. E. Rombo, G. Terracina, G. Tradigo, P. Veltri, "Improving protein secondary structure predictions by prediction fusion," *Info. Fusion*., vol. 10, no. 3, pp. 217–232, July. 2009.
[3] C.A. Ouzounis, R.M.R. Coulson, A.J. Enright, V. Kunin, J.B. Pereira-Leal, "Classification schemes for protein structure and function," *Nat. Rev. Genet*., vol. 4, no. 7, pp. 508–519, July. 2003.
[4] S.C. Rison, T.C. Hodgman, J.M. Thornton, "Comparison of functional annotation schemes for genomes," *Funct. Integr. Genomics*., vol. 1, no. 1, pp. 56–69, May. 2000.
[5] http://proteincrystallography.org/protein/