Journal of Macromolecular Science, Part A
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597274

The Use of Observed Amino Acid Composition in the Proteins to the Analysis of the Sense and Antisense Strands Of DNA

J. Seetharaman[a]; R. Srinivasan[b]
[a] Dept. of Biochemistry, McMaster University, Hamilton, ONT., Canada [b] Department of Crystallography and Biophysics, University of Madras, Madras, India

To cite this Article Seetharaman, J. and Srinivasan, R.(1995) 'The Use of Observed Amino Acid Composition in the Proteins to the Analysis of the Sense and Antisense Strands Of DNA', Journal of Macromolecular Science, Part A, 32: 1, 1237 — 1243
To link to this Article: DOI: 10.1080/10601329508020345
URL: http://dx.doi.org/10.1080/10601329508020345

MACROMOLECULAR REPORTS, A32(SUPPL. 8), 1237-1243 (1995)

THE USE OF OBSERVED AMINO ACID COMPOSITION IN THE PROTEINS TO THE ANALYSIS OF THE SENSE AND ANTISENSE STRANDS OF DNA

J. Seetharaman* and R. Srinivasan
Department of Crystallography and Biophysics
University of Madras, Madras-25, India

ABSTRACT

An analysis of the amino acid composition (via codons) vs. molecular weight in the sense and antisense strands of DNA, in both normal and reverse directions (5' to 3' and 3' to 5' respectively) has been done using the Nucleotide Sequence Data Bank. The amino acid composition in the sense strand in the normal direction (5' to 3', called as SSP1) showed that most of the amino acids follow the inverse correlation between molecular weight and frequency of occurrence. A similar feature is observed in the case of the antisense strand in the normal direction (5' to 3' of antisense strand, called as ASP2) also. There are more similarities in amino acid composition between SSP1 and ASP2. The other two amino acid composition obtained by 3' to 5' reading of sense and antisense strands (called as SSP1' and ASP2' respectively) do not show any correlation between them.

INTRODUCTION

Several studies have been carried out by different workers to understand the tertiary, secondary and primary structure of proteins [1-8]. The randomness or otherwise the frequency of the occurrence of the amino acids in proteins have been treated statistically [8-12]. Most of these studies have been done with respect to specific organisms or group of organisms. So far there is no consolidated study of the frequency of occurrence of codons (amino acids) over the two strands (sense and antisense) of DNA and also over a wide range of organisms which is now possible due to the availability of large data of protein and nucleic acid sequences. It has been shown earlier that there exists a linear inverse correlation of amino acid

* Present address: Dept. of Biochemistry, McMaster University, Hamilton, ONT. L8N 3Z5 Canada.

composition with molecular weight [2] using protein sequence data for 32 proteins. This linear inverse correlation has been examined using the available nucleic acids sequence data [13]. The amino acid composition for this investigation has been obtained from the codons of the sense strand in the normal direction (5' to 3').

It is interesting to note that peptides corresponding to the usually ignored anti-sense DNA code have been reported to have properties of potentially biological and biotechnological importance [14-21]. It has been pointed out that there is coding capacity in both DNA strands [20, 22] and furthermore that both strands can be transcribed simultaneously in the same cell [23]. Because of the increasing importance of the antisense strand, the amino acid composition of the antisense strand has been examined for the statistical inverse correlation. The amino acid composition for this analysis is obtained from the codons of the antisense strand in the normal direction (5' to 3').

Furthermore, as there are reports of 3' to 5' reading of the nucleic acids sequences [16] the above statistical study has been extended to the sense and antisense strands in the reverse direction also (3' to 5').

MATERIALS AND METHOD

The nucleic acid sequence data provided by Wada et al., [13] gives the codon usage for various individual organisms of all the 64 codons of the sense strand of DNA in the normal direction (5' to 3'). A global data (GLO) is obtained from all these data. To generate the codons of the sense strand in the reverse direction and of antisense strand in both normal and reverse directions a transformation matrix table (Table.1) has been constructed from the codons of sense strand (5' to 3') as follows.

Consider the first amino acid Arg. Its first codon is CGA. i) When this codon, CGA is read in the Sense Strand in the normal direction , 5' to 3' (SSP1) it codes for the first codon of Arg (R1) itself. ii) When the same Arg codon CGA is read in the Sense Strand in the reverse direction, 3' to 5' (SSP1') the code is read as AGC and this codes for fifth codon of Ser (S5). iii) When the same codon CGA is read in the Antisense Strand in the normal direction, 5' to 3' (ASP2) it becomes GCU and this codes for fourth codon of Ala (A4). iv) When the same Arg codon CGA is read in the Antisense Strand in the reverse direction, 3' to 5' (ASP2') it becomes UCG and codes for the third codon of Ser (S3).

Similarly all the elements of the transformation matrix are obtained from the sense codons of SSP1. The amino acid composition of one particular amino acid, say Arg is obtained by summing all the six codons of Arg (i.e., R1 + R2 + R3 + R4 + R5 + R6). This is done for each amino acid and for each case SSP1, SSP1', ASP2 and ASP2'.

RESULTS AND DISCUSSION

The frequency of occurrence of each amino acid obtained from the sense strand in the normal direction , 5' to 3' (SSP1) is plotted against its molecular weight and a least squares line is obtained between them (ref. Fig. 1) Figure. 1 corresponds to the global data (GLO) and it

TABLE-1

Transformation matrix table of 1)SSP1-->ASP2, 2)SSP1-->ASP2' and 3)SSP1-->SSP1'

	Codon	SSP1	ASP2	ASP2'	SSP1'		Codon	SSP1	ASP2	ASP2'	SSP1'
ARG	CGA	R1	A4	S3	S5		GGG	G3	P2	P2	G3
	CGC	R2	A3	A3	R2		GGU	G4	P1	T2	W
	CGG	R3	A2	P3	G2	VAL	GUA	V1	H2	Y1	M
	CGU	R4	A1	T3	C1		GUC	V2	Q2	D1	L3
	AGA	R5	S4	S4	R5		GUG	V3	H1	H1	V3
	AGG	R6	S2	P4	G1		GUU	V4	Q1	N1	L6
LEU	CUA	L1	D2	Z2	I2	LYS	AAA	K1	F2	F2	K1
	CUC	L2	E2	E2	L2		AAG	K2	F1	L4	E1
	CUG	L3	D1	Q2	V2	ASN	AAC	N1	L6	V4	Q1
	CUU	L4	E1	K2	F1		AAU	N2	L5	I3	Z1
	UUA	L5	N2	Z1	I3	GLN	CAA	Q1	V4	L6	N1
	UUG	L6	N1	Q1	V4		CAG	Q2	V2	L3	D1
SER	UCA	S1	S6	Z3	T4	HIS	CAC	H1	V3	V3	H1
	UCC	S2	R6	G1	P4		CAU	H2	V1	M	Y1
	UCG	S3	S5	R1	A4	GLU	GAA	E1	L4	F1	K2
	UCU	S4	R5	R5	S4		GAG	E2	L2	L2	E2
	AGC	S5	S3	A4	R1	ASP	GAC	D1	L3	V2	Q2
	AGU	S6	S1	T4	Z3		GAU	D2	L1	I2	Z2
THR	ACA	T1	C2	C2	T1	TYR	UAC	Y1	M	V1	H2
	ACC	T2	W	G4	P1		UAU	Y2	I1	I1	Y2
	ACG	T3	C1	R4	A1	CYS	UGC	C1	T3	A1	R4
	ACU	T4	Z3	S6	S1		UGU	C2	T1	T1	C2
PRO	CCA	P1	G4	W	T2	PHE	UUC	F1	K2	E1	L4
	CCC	P2	G3	G3	P2		UUU	F2	K1	K1	F2
	CCG	P3	G2	R3	A2	ILE	AUA	I1	Y2	Y2	I1
	CCU	P4	G1	R6	S2		AUC	I2	Z2	D2	L1
ALA	GCA	A1	R4	C1	T3		AUU	I3	Z1	N2	L5
	GCC	A2	R3	G2	P3	MET	AUG	M	Y1	H2	V1
	GCG	A3	R2	R2	A3	TRP	UGG	W	T2	P1	G4
	GCU	A4	R1	S5	S3	TER	UAA*	Z1	I3	L5	N2
GLY	GGA	G1	P4	S2	R6		UGA*	Z2	I2	L1	D2
	GGC	G2	P3	A2	R3		UGA*	Z3	T4	S1	S1

* The terminating codons UAA, UAG and UGA are named Z1, Z2 and Z3.

shows that the amino acids gly, ala, ser, val, thr, pro, ile, asp, asn, gln, lys, phe, tyr and trp have their distribution close to the least squares line. However, some amino acids like leu, glu, arg, cys, met and his show considerable deviation from the least squares line. An analysis of the above results show that most of the amino acids that occur with high frequency have low molecular weight and in turn the amino acids that occur with low frequency have high molecular weight. Therefore it is clear that there is a linear inverse correlation between the frequency of occurrence and molecular weight substantiating the earlier results [2].

Figure.2 shows a similar plot for ASP2 (Antisense Strand 5' to 3') and it is inferred that the amono acids gly, ala, ser, ile, asp, gln, his, glu and trp spread along the least squares line with minimal deviation and amino acids leu, arg, val, phe, met lys and cys show considerable deviation from the least squares line. An analysis of the above results show that most of the amino acids that occur with high frequency have low molecular weight and in turn the amino

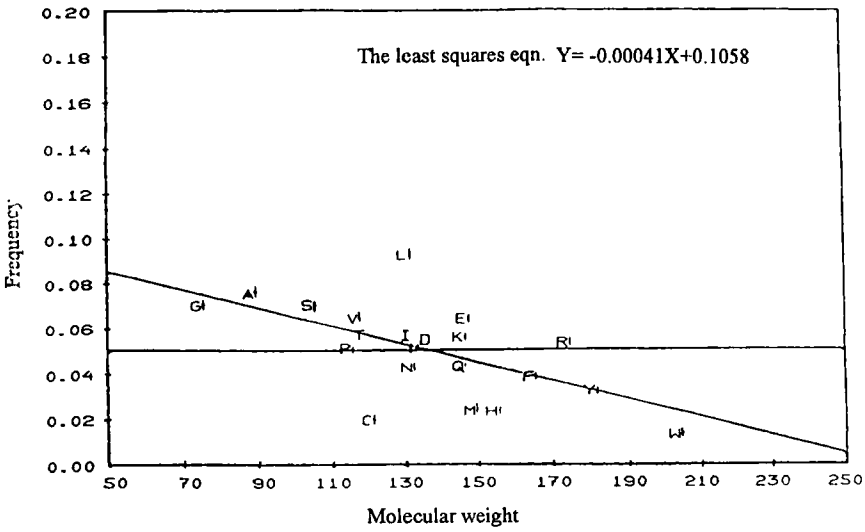

Fig.1. Frequency of occurrence of amino acids plotted against molecular weight for GLO in SSP1 (Sense Strand in 5' to 3' direction). The vertical bars correspond to the probable errors. The error bars are marked on the points to show the statistical spread of the observed data. The length of the bars in each side represent the standard error which is $\varepsilon = \sqrt{2}/\pi\sigma(x)$ where $\sigma(x)$ is the standard deviation. The horizontal line represents the theoretical expected frequency of random occurrence (i.e., 0.05%) and oblique line represents the least squares fit. Single letter amino acid code is used to represent the amino acids.

acids that occur with low frequency have high molecular weight. Therefore it is clear that there is a linear inverse correlation between the frequency of occurrence and molecular weight as in SSP1.

It is observed from Fig. 3 and 4 for SSP1' and ASP2' respectively that most of the amino acids show wide spread of frequency of occurrence and they do not show any correlation between molecular weight and frequency.

A compelling reason for an amino acid to occur with high or low frequency in a protein would be its suitability or unsuitability for protein structure and function. An example would be its ability to form ordered secondary and tertiary structures. It is possible that some physical and chemical properties such as solubility, thermal stability, photochemical stability, selective absorption on clays or selective binding to particular nucleotide sequence of polynucleotides can play a significant role in determining its higher or lower frequency of occurrence. Although these factors may have played a role in the selection process of amino acids it is unlikely that a single reason could account for most or least abundance of the amino acids.

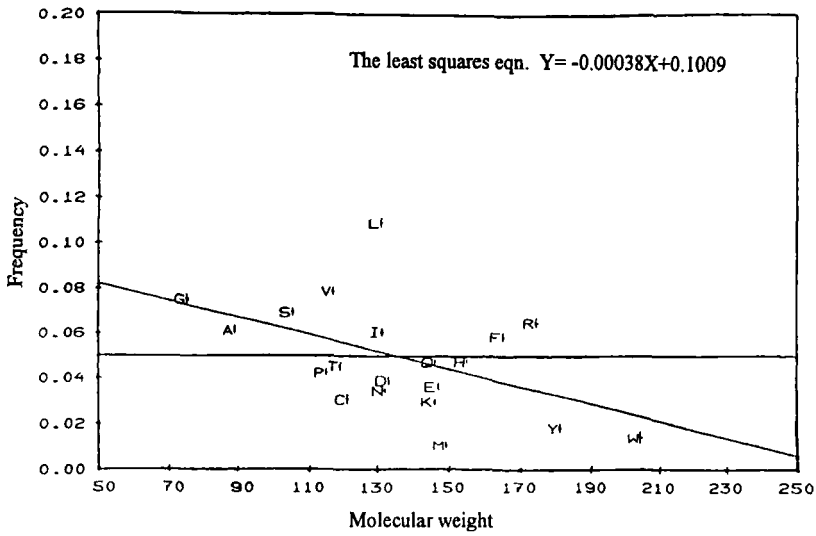

Fig.2. Frequency of occurrence of amino acids plotted against molecular weight for GLO in ASP2(Antisense Strand in 5' to 3' direction)

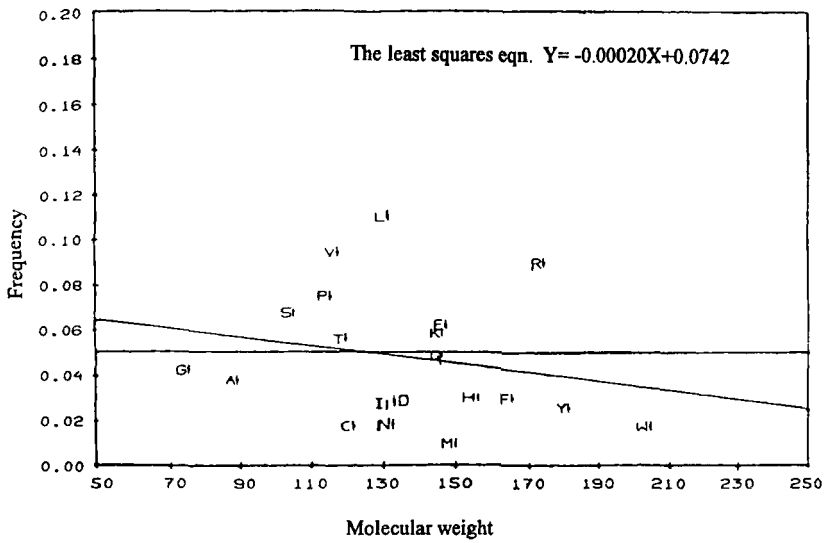

Fig.3. Frequency of occurrence of amino acids plotted against molecular weight for GLO in SSP1' (Sense Strand in 3' to 5' direction)

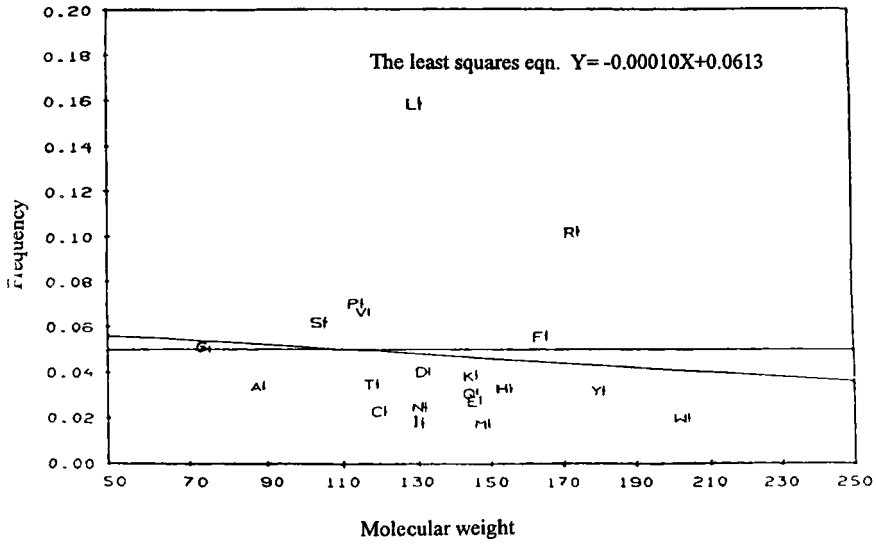

Fig.4. Frequency of occurrence of amino acids plotted against molecular weight for GLO in ASP2'(Antisense Strand in 3' to 5' direction)

The reason for high frequency of occurrence of the amino acids with low molecular weight and vice versa is recognised when the importance of the folding process of protein is considered. It is clear that during folding process of protein, their accessible surface area decreases [24]. It has been reported that the accessible surface area of the folded protein chain is a function of molecular weight of the polypeptide [25, 26]. Therefore, the molecular weight of each amino acid is a variable that natural selection may take into account. The amino acids in proteins would appear to indicate that the natural selection tend to reduce the accessible surface area. Hence the amino acid that may favor the folding of proteins (i.e., less molecular weight amino acids) occur with high frequency and accordingly are assigned the number of codons (i.e, high molecular weight amino acids are assigned less number of codons and low molecular weight amino acids are assigned more codons) in the genetic code.

The probable reasons for the high or low frequency of occurrence of the amino acids in ASP2 depend on the amino acids in 5' to 3' direction of the sense strand which get transformed to yield the amino acids. For instance, the high frequency of leu is due to the transformation of A4, A3, A2, A1, S4 and S2 (6 codons of leu in ASP2) (ref. Table. 1). The codons transformed from SSP1 to yield leu in ASP2 are more used in the sense strand and hence correspondingly high frequency of occurrence is observed (i.e., L1, L2, L3, L4, L5 and L6 in ASP2' is equal to A4, A3, A2, A1, S4 and S2 in SSP1). Similar arguments can be extended for the high or low frequency of occurrence of the amino acids in ASP1' and ASP2' also.

CONCLUSIONS

It is observed that there are more similarities in amino acid composition between the sense strand and antisense strand when they are read in the normal direction (5' to 3'). It reveals the potential importance of the usually ignored antisense strand. Both of them show an inverse correlation between molecular weight and frequency of occurrence. The amino acid composition obtained from the sense and antisense strands by the reading the sequences in the reverse direction (3' to 5') do not show any correlation between them.

REFERENCES

[1] Tristram. G. R. and Smith. R. H (1963) Adv. in Protein Chem. **18**, 227.
[2] Rajan. S. S. and Srinivasan. R (1976) Curr. Sci. **45**, 859.
[3] Chou. P.Y. and Fasman. G.D (1977) J. Mol. Biol. **115**,135.
[4] Srinivasan. R (1978) Ind. J. Biochem & Biophy. **15**, 75.
[5] Mesrob. B. K. Stoeva. St. P (1983) Int. J. Pept. & Protein Res. **21**, 369.
[6] Bernardi. G (1985) J. Mol.Evol. **22**, 363.
[7] Graur. D (1985) J. Mol. Evol. **22**, 53.
[8] Shpaer. E. G (1989) Protein Seq. Data Anal. **2**, 107.
[9] Holmquist. R (1975) J. Mol. Evol. **4**,277.
[10] Holmquist. R (1978) J. Mol. Evol. **11**, 349.
[11] Laird. M. and Holmquist. R (1975) J. Mol. Evol. **4**, 261.
[12] Hanai. R. and Wada. A (1988) J. Mol. Evol. **27**, 321.
[13] Wada. K, Wada. Y. Doi Hh. Ishibashi. F. Gojobori .T and Ikemura. T (1991) Nucl. Acids. Res. **19**, 1981.
[14] Blalock. J. E. and Smith. E. M (1984) Biochem. Biophys. Res. Commun. **121**, 203.
[15] Bost. K. L. Smith, E. M and Blalock. J. E (1985) Proc. Natl. Acad. Sci. USA **82**, 1372.
[16] Blalock. J.E. and Bost. K.L (1986) Biochem. J **234**, 679.
[17] Shai. Y. Flashner. M. and Chaiken. I. M (1987) Biochemistry. **26**, 669.
[18] Blalock J.E (1990) Tibtech. **8**, 140.
[19] Slootstra. J (1990) Tibtech. **8**, 279.
[20] Brentani. R (1990) TIBS **15**, 463.
[21] Brentani. R (1990) J. Mol. Evol. **31**, 239.
[22] Bren tani. R. R (1988) J. Theor. Biol. **135**, 495.
[23] Miyajima, N., Horiuchi, R., Shibuya, Y., Fukushige, S., Matsubara, K., Toyoshima, K and Yamamoto, T (1989) Cell. **57**, 31.
[24] Lee. B and Richards. F.M (1971) J. Mol. Biol. **55**,379.
[25] Chothia. C (1975) Nature. **254**, 304.
[26] Teller. D.C (1976) Nature. **260**, 729.