ANALYSIS OF PARITY RATIO OF PROTEIN SEQUENCES: A NEW APPROACH BASED ON CHARGAFF'S RULE

K. MANIKANDAKUMAR, K. GOKUL RAJ**, R. SRIKUMAR***, S. MUTHUKUMARAN*****

*Department of Physics, Bharathidasan University College (W), Orathanadu – 614 625, Tanjavore, Tamil Nadu, India, bioinfokm@gmail.com

**Department of Computer Science, Jamal Mohamed College, Tiruchirappalli – 620 020, Tamil Nadu, India

***Department of Microbiology, Bharathidasan University College (W), Orathanadu – 614 625, Tanjavore, Tamil Nadu, India

****Department of Physics, Arignar Anna Government Arts College, Attur – 636 121, Salem District, Tamil Nadu, India

Abstract. Amino acid residues are classified in many groups such as hydrophobic, polar, charged, etc. Nevertheless, these groups are not saying any uniqueness under statistical results. Different classifications have different statistical results. Many of the groupings of amino acid residues analyses are not related to Chargaff's rule. But the nucleotide favorable amino acid grouping is unique under statistical calculation. It is also related to the Chargaff's rule of protein nucleotide content. The basis of present explanations of the manner in which the amino acids are activated before being assembled to make a protein; they are being invoked incessantly in attempts to unravel the nucleotide code which is thought to be responsible for specifying the amino acid sequence of proteins. Finally, we observed that the parity ratio of protein sequences ranges between 0.96% and 1.04% for most of the protein families. The +0.04% variation may be rectified from the alignment of the sequencing. The difference between the parity ratio of these families of protein sequences and the unity of Chargaff rule range from –0.3 to +0.3. We propose that the protein parity ratio for nucleotide favorable amino acid grouping has similar statistical calculation in all protein sequences. This statistical calculation is similar to Chargaff's rule. The approach proposed in the present paper may be further pursued to show how different classifications of amino acids can complement each other to yield the properties of extant protein sequences.

Key words: protein sequences, parity ratio, Chargaff's rule, amino acids.

INTRODUCTION

Although there are 64 possible triplet codons, there are only 20 different amino acids coded by them. Thus, most amino acids are inserted into a growing polypeptide chain in response to two or more different triplets forming the mRNA.

Received: October 2009.
in final form March 2010.

For example, proline is coded by 5'-CCN-3' and alanine by 5'-GCN-3' (where N can be any of the four bases). In other cases the pattern may be more complex, such as for isoleucine, which is coded by 5'-AT(T, C or A)-3'. Note that certain amino acids (e.g. serine) may be coded by two distinct groups of triplets, which cannot be adequately represented as 5'-(T or A) (C or G) N-3'. It is to be noted that synthetic oligonucleotide probes for detecting protein-coding sequences often involve the preparation of 'mixed probes'. Due to the redundancy of the genetic code, mixtures of two (or more) nucleotides are sometimes used for every single position in the oligonucleotide. It is anticipated that a single-letter code might be used to designate such mixtures.

The most famous of Chargaff's rules is that in DNA, the proportion of A equals that of T, and C that of G [3]. This nucleotide balance is governed by complementary base pairing rules fundamental to the structure of the double helix [12]. Astonishingly, the nucleotides preserve almost the same equality balance in the single strands of DNA [11] and this phenomenon is sometimes named Chargaff's second parity rule [1, 2, 4, 5, 8, 9, 10]. In [7] it is analyzed the matrix frequency of *Oryza sativa* (japonica cultivar group) complete genomes.

Textbooks of biochemistry state the base ratio 1:1 between the number of adenine (A) and thymine (T) molecules and between the number of guanine (G) and cytosine (C) molecules. This is referred as Chargaff's rule in honor of Erwin Chargaff, who discovered this regularity. With an approximation, Chargaff's rule (%A = %T; %G = %C) applies to single-stranded DNA too. In long nucleotide sequences, Chargaff’s values are also found with approximately equal frequencies. This applies to all species studied. However, species differ in base composition. With the goal of understanding the protein sequence groupings, we have compared the groupings of amino acid residues in long sequences and their shuffled counterparts. Among the 67 protein families studied there is a hierarchy of groupings which is influenced by both residue composition and residue order.

The sequence of amino acids is uniquely specified by the sequence of nucleotides. Similarly, it is possible to convert an amino acid sequence to a linear order of residues. But this raises problems with the codons for leucine, arginine, serine and stop codons. With leucine, for example, the coding triplets are precisely specified by CTN and TTR, but combining these gives YTN, which also includes two phenylalanine codons, TTT and TTC. Thus, information may be lost when an amino acid sequence is converted into a single sequence of base-uncertainty symbols. To avoid ambiguity, therefore, it is important to specify whenever the triplet YTN for a leucine residue, for example, occurs that it does not include TTT or TTC as possibilities, etc. The normalized frequencies for each confirmation (e.g., P_α, P_β, P_t) were calculated from the fraction of residues of each amino acid that occurred in that confirmation, divided by this fraction for all residues. Random occurrence of a particular amino acid in a conformation would give a value of unity.

Many of the groupings of amino acid residues are not applying to Chargaff's rule. We propose that nucleotide favorable amino acids method might be sufficient to explain the phenomenon. To test this, we consider that every amino acid is coded by certain nucleotide codons. In this way, the vast majority of naturally occurring DNA molecules will evolve special properties. We presume that any DNA can be made to evolve to intra-strand parity through a process of inversion, and that deviations from parity are rare in evolution. In nature, a single sequence contains all the information necessary to dictate the fold of the protein.

To analyze the amino acid residues, it is necessary to know their characteristics for grouping them into different types like hydrophobic, polar, and charged, etc. These characteristics stand for the conformational models like alpha helix, beta sheet, coiled coil, etc. of the protein secondary structure. In this study, we developed a new analysis of amino acid residues based on nucleotide sequences. The triplet nucleotide sequences are indicated by their related amino acid residues (the triplets are constituted by 1st base, 2nd base and 3rd base of nucleotide) [6]. In the analysis proposed in this paper, we ignore the first and third bases, and consider second base only. We identify the amino acid residue with the 2nd base of nucleotide and consequently we group the amino acid residues in nucleotide related categories like: adenine favorable (A_{aa}), thymine favorable (T_{aa}), guanine favorable (G_{aa}) and cytosine favorable (C_{aa}) amino acids. This method is used to uniquely identify the protein sequence patterns and characteristics.

MATERIAL AND METHODS

GROUPING OF AMINO ACID RESIDUES

We classify the amino acid residues into four groups [6], namely, adenine favorable, thymine favorable, guanine favorable, and cytosine favorable amino acids. They are denoted as A_{aa}, T_{aa}, G_{aa} and C_{aa} respectively. The list of amino acids selected in each of these four groups is provided in Table 1.

The grouping of amino acids is based on the second base of the nucleotide triplet. We ignore the 1st and 3rd bases of nucleotide triplets, because they are not related to our grouping of amino acids. We eliminate the unknown residues from the calculations.

Table 1

Grouping of amino acid residues in nucleotide favorable categories

Name of the group	Related amino acids
Adenine favorable (A_{aa})	Asp (D), Glu (E), His (H), Lys (K), Asn (N), Gln (Q), Tyr (Y)
Thymine favorable (T_{aa})	Phe (F), Ile (I), Leu (L), Met (M), Val (V)
Guanine favorable (G_{aa})	Cys (C), Gly (G), Arg (R), Trp (W)
Cytosine favorable (C_{aa})	Ala (A), Pro (P), Ser (S), Thr (T)

For example: the 2nd base of nucleotide triplet coding the serine amino acid can be G or C. We ignore the AGT and AGC codons. We take only TCT, TCC, TCA and TCG codons into consideration for the serine residue.

DETAILS OF DATA USED

There are (among others) four major classes of proteins according to the Structural Classification of Proteins (SCOP), namely, All alpha, All beta, alpha plus beta and alpha by beta. In the present study, we have chosen few representative families of proteins for each of these four classes. The selected protein sequences are downloaded from the Protein Data Bank (PDB) website using SCOP option (http://www.rcsb.org/pdb/browse/browse.do?t=11&useMenu=no).

RESULTS AND DISCUSSION

We analyzed 67 families of protein sequences from the SCOP option of the PDB database. We carefully investigated them from the point of view of their composition in nucleotide favorable, amino acids. Figure 1 shows the ratio of the adenine plus guanine favorable amino acid residues to the thymine plus cytosine favorable amino acids in 19 different families of proteins. Thus it is shown that all families of protein sequences may belong to the new grouping of amino acid residues in nucleotide favorable, i.e. this grouping of amino acid residues has similar statistical distribution in all protein sequences. It is shown that all families of proteins sequences are characterized by a ratio of nearly one. Few families of protein sequences have minor variations. But the ratio of them is above 0.95. This statistical calculation is similar to Chargaff's rule. So, we may follow this kind of groupings in further analysis. Figure 2 shows the difference between the ratio of the nucleotide favorable amino acid residues and Chargaff's ratio in different families of protein sequences. As shown, the difference of ratios has values that range from –0.3 to +0.3. This difference is very low for most of the families.

We have taken 19 families of protein sequences representative for the four structural categories of proteins. Table 2 shows the method of calculation of the $A_{aa}T_{aa}G_{aa}C_{aa}$ analysis. Table 2 shows first the percentage of $A_{aa,}$ $T_{aa,}$ G_{aa}, and C_{aa} grouped amino acids from the total number of amino acids in every protein sequence. Further, we have calculated the percentage for A_{aa}+G_{aa} and T_{aa}+C_{aa} groups of amino acid residues. Finally, we calculated the ratio of the A_{aa}+G_{aa} and T_{aa}+C_{aa} sums. For most of the protein families analyzed in this study, the ratio lies between 0.96% and 1.04%. Up to +0.04 % variation may be rectified from the alignment of the sequence. The A_{aa} group of amino acid residues has values from 29.75% to 36.94%. The T_{aa} group of amino acid residues ranges from 18.75% to 29.75%, but the majority range from 24% to 25%. The G_{aa} group of amino acid residues ranges from 13.04% to 20.25%. The C_{aa} group of amino acid residues ranges from 20.65% to 30.23%, but the majority has the values from 22.60% to

25.64%. The A_{aa}+G_{aa} group ranges from 48.88% to 51.00%. The T_{aa}+C_{aa} group ranges from 48.98% to 51.12%. The ratio of the (A_{aa}+G_{aa}) and (T_{aa}+C_{aa}) ranges from 0.96% to 1.04%. Finally, we give the difference of nucleotide favorable ratio of each protein sequence against the unity coming from Chargaff's rule (i.e., 1-protein parity ratio).

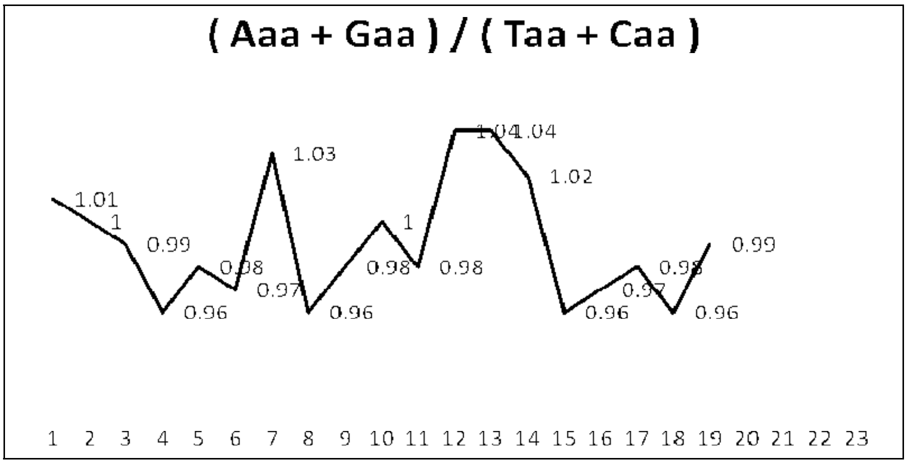

Fig. 1. Parity ratio of the nucleotide favorable amino acid residues in different families of proteins.

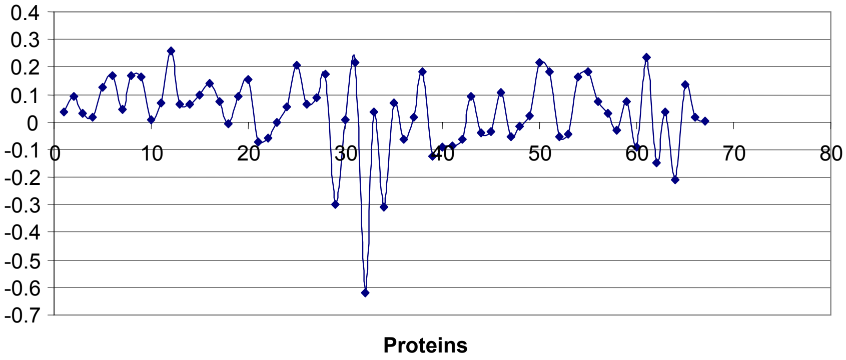

Fig. 2. Difference between the unity of Chargaff's ratio and the ratio of the nucleotide favorable amino acid residues for different families of protein sequences.

ALL ALPHA CLASS OF PROTEIN SEQUENCES

We have studied 19 families of all alpha category of proteins. The highest percentage of A_{aa} grouped amino acids was found in EF Hand-like proteins (38.5%). The lowest percentage was found in spectrin repeat-like proteins (25.03%). The highest percentage of T_{aa} grouped amino acids was found in a protein family of nuclear receptor proteins, ligand-binding domain (33.85%). The lowest percentage of T_{aa} grouped amino acids was found in the protein family of phospholipase A2 (PLA2) (16.89%). The highest percentage of G_{aa} grouped amino acids was found in spectrin repeat-like proteins (31.63%), while the lowest percentage of G_{aa} grouped amino acids was found in nuclear receptor proteins, ligand-binding domain (11.7%). The highest percentage of C_{aa} grouped amino acids is present in the protein family of lambda repressor-like proteins, DNA-binding domains (28.0%). The lowest percentage of C_{aa} grouped amino acids is in protein family of EF Hand-like (20.3%). The highest percentage of A_{aa}+G_{aa} grouped amino acids is in phospholipase A2 (61.79%), while the lowest percentage of this category is present in the nuclear receptor protein family, ligand-binding, domain (43.95%).

Table 2

Calculation of parity ratio for different protein sequences

S.N.	Name of protein	No. of protein	Total No. of A. A.	No. unknown	A_{aa}	T_{aa}	G_{aa}	C_{aa}	$A_{aa} + G_{aa}$	$T_{aa} + C_{aa}$	$(A_{aa} + G_{aa})/(T_{aa} + C_{aa})$	$1 - (A_{aa} + G_{aa})/(T_{aa} + C_{aa})$
Proteins (All alpha)												
1.	Alpha/alpha toroid	455	194686	0.00	33.80	25.27	16.39	24.56	50.19	49.83	1.01	–0.01
2.	Ferritin-like	586	140426	0.00	36.94	27.64	13.04	22.39	49.98	50.03	1.00	0.00
3.	Multiheme cytochromes	159	51228	0.01	32.28	23.51	17.47	26.76	49.75	50.27	0.99	0.01
4.	SAM domain-like	647	97125	0.01	33.13	26.90	15.95	24.04	49.08	50.94	0.96	0.04
Proteins (All beta)												
5.	6-bladed beta-propeller	244	115464	0.00	30.96	24.87	18.44	25.73	49.40	50.60	0.98	0.02
6.	Galactose-binding domain-like	298	181741	0.01	30.92	25.20	18.22	25.64	49.14	50.84	0.97	0.03
7.	Glycosyl hydrolase domain	297	161709	0.00	33.42	25.39	17.29	23.91	50.71	49.30	1.03	–0.03
8.	Prealbumin-like	554	168228	0.00	32.91	24.61	16.16	26.32	49.07	50.93	0.96	0.04
9.	Supersandwich	329	258264	0.00	32.60	25.82	16.93	24.64	49.53	50.46	0.98	0.02
10.	Trypsin-like serine proteases	1970	299404	0.10	29.63	24.77	20.25	25.34	49.88	50.11	1.00	0.00
Proteins (alpha + beta)												
11.	Ferredoxin-like	2132	529957	0.95	30.29	26.67	19.29	23.76	49.58	50.43	0.98	0.02
12.	RNase A-like	273	30762	0.06	36.07	18.75	14.93	30.23	51.00	48.98	1.04	–0.04
13.	SH2-like	314	36056	0.19	34.89	26.51	16.02	22.60	50.91	49.11	1.04	–.0.04
14.	Zincin-like	431	140560	0.10	36.34	26.02	14.08	23.57	50.42	49.59	1.02	–0.02
Proteins (alpha / beta)												
15.	Dihydrofolate reductases	160	32943	0.00	34.67	29.74	14.37	21.23	49.04	50.97	0.96	0.04
16.	Flavodoxin-like	1368	385194	1.12	29.75	26.58	19.46	24.21	49.21	50.79	0.97	0.03
17.	UDP-Glycosyltransferase/ glycogen phosphorylase	181	113105	0.00	34.86	29.75	14.74	20.65	49.60	50.40	0.98	0.02
18.	Periplasmic binding protein-like II	468	150426	0.00	34.46	25.67	14.42	25.45	48.88	51.12	0.96	0.04
19.	Ribonuclease H-like motif	448	92157	0.00	35.18	27.36	14.57	22.90	49.75	50.26	0.99	0.01

ALL BETA CLASS OF PROTEIN SEQUENCES

We have investigated a total of 19 protein sequences in All beta structural category of proteins. The highest percentage of A_{aa} grouped amino acids was found in lipocalins family (38.75%), while the lowest percentage in streptavidin-like proteins (26.59%). The highest percentage of T_{aa} grouped amino acids was found in acid proteases family (30.63%), while the lowest percentage in streptavidin-like proteins (17.45%). The highest percentage of G_{aa} grouped amino acids was found in SH3-like barrel proteins (26.86%), while the lowest percentage in lipocalins proteins (13.4%). The highest percentage of C_{aa} grouped amino acids was found in streptavidin-like proteins (36.26%), while the lowest percentage was found in lipocalins proteins (21.52%). The highest percentage of A_{aa}+G_{aa} grouped amino acids is consecutively found in protein family of SH3-like barrel proteins (54.72%), while the lowest percentage is in nucleoplasmin-like/VP proteins (43.4%). The highest percentage of T_{aa}+C_{aa} grouped amino acids was found in the protein family of nucleoplasmin-like/VP proteins (56.59%), while the lowest percentage in SH3-like barrel proteins (45.29%). The biggest parity ratio is in SH3-like barrel proteins (1.21%). The smallest ratio is in nucleoplasmin-like/VP proteins (0.77%).

ALPHA PLUS BETA CLASS OF PROTEIN SEQUENCES

We studied 14 protein families from alpha plus beta structural category of proteins. The highest percentage of A_{aa} grouped amino acids was found in thymidylate synthase/dCMP hydroxymethylase proteins (36.41%), while the lowest percentage was found in ferredoxin-like (30.29%). The highest percentage of T_{aa} grouped amino acids was found in kinase-like proteins (30.16%), while the lowest percentage in RNase A-like proteins (18.75%). The highest percentage of G_{aa} grouped amino acids was found in lysozyme-like proteins (21.2%), while the lowest percentage in kinase-like proteins (13.78%). The highest percentage of C_{aa} grouped amino acids was found in RNase A-like proteins (30.23%), while the lowest percentage in thymidylate synthase/dCMP hydroxymethylase proteins (19.5%). The highest percentage of A_{aa}+G_{aa} grouped amino acids was found in lysozyme-like proteins (52.88%) and the lowest in glyceraldehyde-3-phosphate dehydrogenase-like proteins, C-terminal domain (45.02%). The highest percentage of T_{aa}+C_{aa} grouped amino acids was found in glyceraldehyde-3-phosphate dehydrogenase-like proteins, C-terminal domain (54.99%). The lowest percentage of T_{aa}+C_{aa} grouped amino acids was found in lysozyme-like protein family (47.12%). In the latter there is the biggest parity ratio (1.12%), while the smallest was found in glyceraldehyde-3-phosphate dehydrogenase-like proteins, C-terminal domain (0.82%).

ALPHA BY BETA CLASS OF PROTEIN SEQUENCES

We studied 15 protein sequences from alpha by beta structural category of proteins. The highest percentage of A_{aa} grouped amino acids was found in ribonuclease H-like motif (35.18%) and the lowest in subtilisin-like proteins (26.37%). The highest percentage of T_{aa} grouped amino acids was found in UDP-glycosyltransferase/glycogen phosphorylase proteins (29.74%),and the lowest in subtilisin-like proteins (22.77%). The highest percentage of G_{aa} grouped amino acids was found in flavodoxin-like proteins (19.46%)and the lowest in thioredoxin fold proteins (13.94%). The highest percentage of C_{aa} grouped amino acids was found in RNase A-like proteins (30.23%) and the lowest in UDP-glycosyltransferase/glycogen phosphorylase proteins (20.65%). The highest percentage of A_{aa}+G_{aa} grouped amino acids was found in ribonuclease H-like motif (49.75%) and the lowest in subtilisin-like proteins (42.61%). The highest percentage of T_{aa}+C_{aa} grouped amino acids was found in subtilisin-like proteins (57.39%) and the lowest percentage in ribonuclease H-like motif (50.26%). The biggest parity ratio is in ribonuclease H-like motif (0.99%) and the lowest in subtilisin-like proteins (0.74%).

CONCLUSION

In our parity ratio studies on protein sequences, we have investigated up to 67 protein sequences in four structural categories. For most protein families, parity ratios are lying between 0.96% and 1.04%. The +0.04% variation may be rectified from the alignment of the sequence. The A_{aa} group of amino acid residues represents from 29.75% to 36.94% of the total of amino acid residues. The T_{aa} group of amino acid residues represents from 18.75% to 29.75% of the total of amino acid residues. But the majority of protein families has 24% to 25% T_{aa} amino acids. The G_{aa} group of amino acid residues represents from 13.04% to 20.25%, while the C_{aa} group of amino acids represents 20.65% to 30.23% of the total of amino acid residues. But the majority of protein families has 22.60% to 25.64% C_{aa} amino acids. The A_{aa}+G_{aa} group has values from 48.88% to 51.00%. The T_{aa}+C_{aa} group ranges from 48.98% to 51.12%. The ratio of the (A_{aa}+G_{aa}) and (T_{aa}+C_{aa}) groups of amino acids ranges from 0.96% to 1.04%. Finally, we have given the difference of this ratio against the unity of Chargaff's rule (i.e., 1-protein ratio). The difference ranged from –0.3 to +0.3. In summary, we showed that the parity ratio for nucleotide favorable amino acid grouping has similar statistical calculation for all protein sequences. This statistical calculation is similar to Chargaff's rule. The approach presented in the paper may be further pursued to show how different classifications of amino acids can complement each other to yield the properties of extant protein sequences.

REFERENCES

1. ALBRECHT-BUEHLER, G., Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions, *Proc. Natl. Acad. Sci. USA*, 2006, **103**, 17828–17833.
2. BAISNÉE, P.F., S. HAMPSON, P. BALDI, Why are complementary DNA strands symmetric?, *Bioinformatics*, 2002, **18**, 1021–1033.
3. CHARGAFF, E., Structure and function of nucleic acids as cell constituents, *Fed. Proc*., 1951, **10**, 654–659.
4. FICKETT, J.W., D.C. TORNEY, D.R. WOLF, Base compositional structure of genomes, *Genomics*, 1992, **13**, 1056–1064.
5. FORSDYKE, D.R., J.R. MORTIMER, Chargaff's legacy, *Gene*, 2000, **261**, 127–137.
6. HARTL, D.L., ELIZABETH W. JONES, *Genetics – Analysis of Genes and Genomes*, Sixth edition, Jones and Bartlett Publishers, Sudbury, Massachusetts, 2005.
7. MANIKANDAKUMAR, K, S. MUTHU KUMARAN, R. SRIKUMAR, Matrix frequency analysis of *Oryza sativa* (japonica cultivar group) complete genomes, *Journal of Computer Science & Systems Biology*, 2009, **2**, 159–166.
8. MITCHELL, D., R. BRIDGE, A test of Chargaff's second rule, *Biochem. Biophys. Res. Commun*., 2006, **340**, 90–94.
9. PRABHU, V.V., Symmetry observations in long nucleotide sequences, *Nucleic Acids Res*., 1993, **21**, 2797–2800.
10. QI, D., A.J. CUTICCHIA, Compositional symmetries in complete genomes, *Bioinformatics*, 2001, **17**, 557–559.
11. RUDNER, R., J.D. KARKAS, E. CHARGAFF, Separation of *B. subtilis* DNA into complementary strands, *Proc. Natl. Acad. Sci. USA*, 1968, **60**, 921–922.
12. WATSON, J.D., F.H. CRICK, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid, *Nature*, 1953, **171**, 737–738.