Appendix C: Encoding Schemes for DNA Sequences

Tables to represent Mapping of Nucleotide Bases to Numerical values[169] so as to represent DNA sequence as a Digital Signal.

Sr. No.	*Encoding Scheme*	*Nucleotide Base*	*Numerical Value Mapping*
1.	Single Galois Indicator[170]	A	0
		C	1
		G	3
		T	2
2.	Integer Number[171]	A	2
		C	1
		G	3
		T	0
3.	Real Number[172]	A	-1.5
		C	0.5
		G	-0.5
		T	1.5
4.	Complex Number[173]	A	1 + i
		C	-1 - i
		G	-1 + i
		T	1 - i

[169] Jennifer Kwan, Benjamin Kwan, Hon Kwan, Spectral Analysis of Numerical Exon and Interon Sequences, IEEE International Conference on BioInformatics and Biomedicine Workshop, 978-1-4244-8302-0/10/$26.00 ©2010 IEEE pp 876-877

[170] M. Akhtar, J. Epps, and E. Ambikairajah, "Signal processing in sequence analysis: Advances in eukaryotic gene prediction," IEEE Journal of Selected Topics in Signal Processing, vol. 2, pp. 310-321, June 2008.

[171] P. D. Cristea, "Genetic signal representation and analysis," in Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, vol. 4623, January 2002, pp. 77-84.

[172] N. Chakravarthy, A. Spanias, L. D. Lasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences," EURASIP Journal of Genomic Signal Processing, vol. 1, pp. 13-28, January 2004.

[173] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," J. Cell. Mol. Med., vol. 6, pp. 279-303, April-June 2002.

5.	Quaternary Code[174]	A	1
		C	- i
		G	-1
		T	I
6.	Left-rotated Quaternary Code	A	I
		C	1
		G	-i
		T	-1
7.	Electron-Ion Interaction Pseudo Potential (EIIP)[175]	A	0.1260
		C	0.1340
		G	0.0806
		T	0.1335
8.	Molecular Mass[176]	A	134
		C	110
		G	150
		T	125
9.	Atomic Number[177]	A	70
		C	58
		G	78
		T	66
10.	Paired Nucleotide Atomic Number	A	62
		C	42
		G	62
		T	42
11.	Paired Numeric or	A	1

[174] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, "New approaches to genome sequence analysis based on digital signal processing," in Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), October 2002, pp. 1-4.

175 A. S. Nair and S. P. Sreenathan, "A Coding Measure Scheme Employing Electron-Ion Interaction Pseudopo-tential (EIIP)," Bioinformation, Vol. 1, No. 6, 2006, pp. 197-202.

176 H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S, Havlin, S. M. Ossadnik, C.-K. Peng, and M. Simmons, "Statistical mechanics in biology: How ubiquitous are long-range correlations? " Physica A, vol. 205, pp. 214-253, April 1994

177 T. Holden, R. Subramaniam, R. Sullivan, E. Cheng, C. Sneider, G. Tremberger, Jr. A. Flamholz, D. H. Leiberman, and T. D. Cheung, "ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes," in Proc. of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, vol. 6694, August 2007, pp. 669417-1 to 669417-10.

	Bipolar Mapping		
		C	-1
		G	-1
		T	1
12.	Complex Paired Numeric	A	I
		C	-1
		G	-1
		T	i
13.	Dipole Moments[178] (Distribution of the polarity of a chemical bond within a molecule)	A	0.4629
		C	3.943
		G	6.488
		T	1.052
14.	UTP-IS[179] (University Technology Petronas Indicator Sequence	A	1
		C	2
		G	3
		T	4
15.	Binary Indicator Sequence[180]	A	Takes value 1, if the base exists at location n, or 0 if the base is absent at location n in the given sequence
		C	Takes value 1, if the base exists at location n, or 0 if the base is absent at location n in the given sequence
		G	Takes value 1, if

178 J. K. Meher, M. R. Panigrahi, G. N. Dash, P. K. Meher, "Wavelet Based Lossless DNA Sequence Compression for Faster Detection of Eukaryotic Protein Coding Regions ," I.J. Image, Graphics and Signal Processing 2012, 7, 47-53

179 Muneer Ahmad, Azween Abdullah, Khalid Buragga, A Novel Optimized Approach for Gene Identification in DNA Sequences

180 Dimitris Anastassiou, Genomic Signal Processing, IEEE SIGNAL PROCESSING MAGAZINE, vol 18, pp 8-20 1053-5888/01/2001 IEEE

			the base exists at location n, or 0 if the base is absent at location n in the given sequence
		T	Takes value 1, if the base exists at location n, or 0 if the base is absent at location n in the given sequence
16.	2-bit Binary Representation[181]	A	00
		C	11
		G	10
		T	01
16.	4-bit Binary Representation[182]	A	1000
		C	0010
		G	0001
		T	0100

181 R. Ranawana and V. Palade. A Neural network based multi-classifier system for gene identification in DNA sequence. Neural Computing and Applications, 2005, 14(2):122-131

182 B. Demeler, G. W. Zhou. Neural network optimization for E.coli promoter prediction. Nucleic Acids Res., 1991, 19(7):1539-1599.