PURINE-PYRIMIDINE SYMMETRY, DETERMINATIVE DEGREE AND DNA

Diana Duplij and Steven Duplij[*†]
Kharkov National University, Svoboda Sq. 4,
Kharkov 61077, Ukraine

July 10, 2004

Abstract

Various symmetries connected with purine-pyrimidine content of DNA sequences are studied in terms of the intruduced determinative degree, a new characteristics of nucleotide which is connected with codon usage. A numerological explanation of **CG** pressure is proposed. A classification of DNA sequences is given. Calculations with real sequences show that purine-pyrimidine symmetry increases with growing of organization. A new small parameter which characterizes the purine-pyrimidine symmetry breaking is proposed for the DNA theory.

[*]E-mail: Steven.A.Duplij@univer.kharkov.ua
[†]Internet: http://gluon.physik.uni-kl.de/~duplij

Abstract investigation of the genetic code is a powerful tool in DNA models construction and understanding of genes organization and expression [1]. In this direction the study of symmetries [2, 3], application of group theory [4] and implication of supersymmetry [5] are the most promising and necessary for further elaboration. In this paper we consider symmetries connected with purine-pyrimidine content of DNA sequences in terms of the determinative degree introduced in [6].

We denote a triplet of nucleotides by xyz, where $x, y, z = \mathbf{C}, \mathbf{T}, \mathbf{A}, \mathbf{G}$. Then redundancy means that an amino acid is fully determined by first two nucleotides x and y independently of third z [1]. Sixteen possible doublets xy group in 2 octets by ability of amino acid determination [7]. Eight doublets have more "strength" in sense of the fact that they simply encode amino acid independently of third bases, other eight ("weak") doublets for which third bases determines content of codons. In general, transition from the "powerful" octet to the "weak" octet can be obtained by the exchange [7] $\mathbf{C} \overset{*}{\Longleftrightarrow} \mathbf{A}$, $\mathbf{G} \overset{*}{\Longleftrightarrow} \mathbf{T}$, which we name "star operation $(*)$" and call *purine-pyrimidine inversion*. Thus, if in addition we take into account $\mathbf{GC}$ pressure in evolution [8] and third place preferences during codon-anticodon pairing [9], then 4 nucleotides can be arranged in descending order in the following way:

Pyrimidine	Purine	Pyrimidine	Purine
$\mathbf{C}$	$\mathbf{G}$	$\mathbf{T}$	$\mathbf{A}$
very "strong"	"strong"	"weak"	very "weak"

(1)

Now we introduce a numerical characteristics of the empirical "strength" — *determinative degree* $\mathbf{d}_x$ of nucleotide x and make transition from qualitative to quantitative description of genetic code structure [6]. It is seen from (1) that the determinative degree of nucleotide can take value $\mathbf{d}_x = \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}$ in correspondence of increasing "strength". If we denote determinative degree as upper index for nucleotide, then four bases (1) can be presented as vector-row $\mathbb{V} = \begin{pmatrix} \mathbf{C}^{(4)} & \mathbf{G}^{(3)} & \mathbf{T}^{(2)} & \mathbf{A}^{(1)} \end{pmatrix}$. Then the exterior product $\mathbb{M} = \mathbb{V} \times \mathbb{V}$ represents the doublet matrix $\mathbb{M}$ and corresponding rhombic code [10], and the triple exterior product $\mathbb{K} = \mathbb{V} \times \mathbb{V} \times \mathbb{V}$ corresponds to the cubic matrix model of the genetic code which were described in terms of the determinative degree in [6]. To calculate the determinative degree of doublets xy we use the following additivity assumption

$$\mathbf{d}_{xy} = \mathbf{d}_x + \mathbf{d}_y, \tag{2}$$

which holds for triplets and for any nucleotide sequence. Then each of 64 elements (codons) of the cubic matrix $\mathbb{K}$ will have a *novel number characteristics* —determinative degree of codon $\mathbf{d}_{xyz} = \mathbf{d}_{codon} = \mathbf{d}_x + \mathbf{d}_y + \mathbf{d}_z$ which takes value in the range $\mathbf{3 \div 12}$. We can also define the determinative degree of amino acid $\mathbf{d_{AA}}$ as mean arithmetic value $\mathbf{d_{AA}} = \sum \mathbf{d}_{codon}/n_{deg}$, where n_{deg} is its degeneracy (redundancy). That can allow us to analyze new abstract amino acid properties in connection with known biological properties [6].

Let us consider a numerical description of an idealized DNA sequence as a double-helix of two codon strands connected by complementary conditions [1]. Each strand is described by four numbers $(n_{\mathbf{C}}, n_{\mathbf{G}}, n_{\mathbf{T}}, n_{\mathbf{A}})$ and $(m_{\mathbf{C}}, m_{\mathbf{G}}, m_{\mathbf{T}}, m_{\mathbf{A}})$, where n_x is a number of nucleotide x in one strand. In terms of n_x and m_x the complementary conditions are

$$n_{\mathbf{C}} = m_{\mathbf{G}},\ m_{\mathbf{C}} = n_{\mathbf{G}},\ n_{\mathbf{T}} = m_{\mathbf{A}},\ m_{\mathbf{T}} = n_{\mathbf{A}}. \tag{3}$$

The Chargaff's rules [1] for a double-helix DNA sequence sound as: 1) total quantity of purines and pyrimidines are equal $N_{\mathbf{A}} + N_{\mathbf{G}} = N_{\mathbf{C}} + N_{\mathbf{T}}$; 2) total quantity of adenine and cystosine equal to total quantity of guanine and thymine $N_{\mathbf{A}} + N_{\mathbf{C}} = N_{\mathbf{T}} + N_{\mathbf{G}}$; 3) total quantity of adenine equal to total quantity of thymine $N_{\mathbf{A}} = N_{\mathbf{T}}$ and total quantity of cystosine equal to total quantity of guanine $N_{\mathbf{C}} = N_{\mathbf{G}}$; 4) the ratio of guanine and cystosine to adenine and thymine $v = (N_{\mathbf{A}} + N_{\mathbf{T}}) / (N_{\mathbf{C}} + N_{\mathbf{G}})$ is approximately constant for each species. Usually the Chargaff's rules are defined through macroscopic molar parts which are proportional to absolute number of nucleotides in DNA [1]. If we consider a DNA double-helix sequence, then $N_x = n_x + m_x$. In terms of n_x and m_x the first three Chargaff's rules lead to the equations which are obvious identities, if complimentary (3) holds. From fourth Chargaff's rule it follows that the specificity coefficient v_{nm} for two given strands is

$$v_{nm} = \frac{n_A + m_{\mathbf{A}} + n_{\mathbf{T}} + m_{\mathbf{T}}}{n_{\mathbf{C}} + m_{\mathbf{C}} + n_{\mathbf{G}} + m_{\mathbf{G}}}. \tag{4}$$

The complementary (3) leads to the equality of coefficients v of each strand $v_{nm} = v_n = v_m \equiv v$, and v is connected with $\mathbf{GC}$ content $p_{\mathbf{CG}}$ in the double-helix DNA as $p_{\mathbf{CG}} = 1/(1+v)$.

We consider another important coefficient: the ratio of purines and pyrimidines k. For two strands from the first Chargaff's rule we obviously derive $k_{nm} = 1$. But for each strand we have

$$k_n = \frac{n_{\mathbf{G}} + n_{\mathbf{A}}}{n_{\mathbf{C}} + n_{\mathbf{T}}},\ k_m = \frac{m_{\mathbf{G}} + m_{\mathbf{A}}}{m_{\mathbf{C}} + m_{\mathbf{T}}} \tag{5}$$

which satisfy the equation $k_n k_m = 1$ following from complementary.

Let us introduce the determinative degree of *each strand* exploiting the additivity assumption (2) as

$$\mathbf{d}_n = \mathbf{4} \cdot n_{\mathbf{C}} + \mathbf{3} \cdot n_{\mathbf{G}} + \mathbf{2} \cdot n_{\mathbf{T}} + \mathbf{1} \cdot n_{\mathbf{A}}, \quad (6)$$

$$\mathbf{d}_m = \mathbf{4} \cdot m_{\mathbf{C}} + \mathbf{3} \cdot m_{\mathbf{G}} + \mathbf{2} \cdot m_{\mathbf{T}} + \mathbf{1} \cdot m_{\mathbf{A}}. \quad (7)$$

The values $\mathbf{d}_n$ and $\mathbf{d}_m$ can be viewed as characteristics of the empirical "strength" for strands, i.e. "strand generalization" of (1). Then we define summing and difference "strength" of a double-helix sequence by

$$\mathbf{d}_+ = \mathbf{d}_n + \mathbf{d}_m, \ \mathbf{d}_- = \mathbf{d}_n - \mathbf{d}_m. \quad (8)$$

The first variable $\mathbf{d}_+$ can be treated as the summing empirical "strength" of DNA (or its fragment). Taking into account the complementary conditions (3) we obtain $\mathbf{d}_+$ through one strand variables

$$\mathbf{d}_+ = \mathbf{7} \cdot (n_{\mathbf{C}} + n_{\mathbf{G}}) + \mathbf{3} \cdot (n_{\mathbf{T}} + n_{\mathbf{A}}). \quad (9)$$

We can also present $\mathbf{d}_+$ through macroscopically determined variables N_x as follows $\mathbf{d}_+ = \mathbf{7} \cdot N_{\mathbf{C}} + \mathbf{3} \cdot N_{\mathbf{A}} = \mathbf{7} \cdot N_{\mathbf{G}} + \mathbf{3} \cdot N_{\mathbf{T}}$, or through **GC** and **AT** contents as $\mathbf{d}_+ = \dfrac{\mathbf{7}}{2} \cdot N_{\mathbf{C+G}} + \dfrac{\mathbf{3}}{2} \cdot N_{\mathbf{A+T}}$.

To give sense to the difference $\mathbf{d}_-$ we derive

$$\mathbf{d}_- = n_{\mathbf{C}} + n_{\mathbf{T}} - n_{\mathbf{G}} - n_{\mathbf{A}}. \quad (10)$$

We see that the star operation obviously acts as $(\mathbf{d}_+)^* = \mathbf{d}_+$ and $(\mathbf{d}_-)^* = -\mathbf{d}_-$. From (9)-(10) it follows the main statement:

> *The biological sense of the determinative degree* $\mathbf{d}$ *is contained in the following purine-pyrimidine relations:*
>
> 1) *The sum of the determinative degrees of matrix and complementary strands in DNA (or its fragment) equals to*
>
> $$\mathbf{d}_+ = \frac{\mathbf{7}}{2} \cdot N_{\mathbf{C+G}} + \frac{\mathbf{3}}{2} \cdot N_{\mathbf{A+T}}. \quad (11)$$
>
> *2) The difference of the determinative degrees between matrix and complementary strands in DNA (or its fragment) exactly equals to the difference between pyrimidines and purines in one strand*

$$\mathbf{d}_- = n_{pyrimidines} - n_{purines}, \tag{12}$$

where $n_{pyrimidines} = n_{\mathbf{C}} + n_{\mathbf{T}}$ *and* $n_{purines} = n_{\mathbf{G}} + n_{\mathbf{A}}$*, or it is equal to difference of purines or pyrimidines between strands*

$$\mathbf{d}_- = n_{pyrimidines} - m_{pyrimidines} = m_{purines} - n_{purines}. \tag{13}$$

We can also find connection between $\mathbf{d}_+, \mathbf{d}_-$ and the coefficients k and v as follows

$$\mathbf{d}_+ = \frac{1}{2} N_{\mathbf{C+G}} (7 + 3v) = N_{\mathbf{C+G}} \left(2 + \frac{3}{2 \cdot p_{\mathbf{CG}}} \right), \tag{14}$$

$$\mathbf{d}_- = n_{pyrimidines} (1 - k_n). \tag{15}$$

If we consider one species for which $v = const$ (or $p_{\mathbf{CG}} = const$), then we observe that $\mathbf{d}_+ \sim N_{\mathbf{C+G}}$, which can allow us to connect the determinative degree with "second level" of genetic information [8]. From another side, the ratio $\frac{7}{3}$ of coefficients in (11) can play a numerological role in **CG** pressure explanations [8], and therefore $\mathbf{d}_+$ can be considered as some kind of "evolutionary strength".

Now we consider the determinative degree of double-helix sequences in various extreme cases and classify them. We call a DNA sequence *mononucleotide*, *dinucleotide*, *trinucleotide* or *full*, if one, two, three or four numbers n_x respectively distinct from zero. Properties of mononucleotide double-helix DNA sequence are in the Table 1.

Table 1. Mononucleotide DNA

n_x	$\mathbf{d}_+$	$\mathbf{d}_-$	amino acid
$n_{\mathbf{C}} \neq 0$	$7n_{\mathbf{C}}$	$n_{\mathbf{C}}$	Pro
$n_{\mathbf{G}} \neq 0$	$7n_{\mathbf{G}}$	$-n_{\mathbf{G}}$	Gly
$n_{\mathbf{T}} \neq 0$	$3n_{\mathbf{T}}$	$n_{\mathbf{T}}$	Phe
$n_{\mathbf{A}} \neq 0$	$3n_{\mathbf{A}}$	$-n_{\mathbf{A}}$	Lis

The mononucleotide sequences which encode most extended amino acids Gly and Lis have negative $\mathbf{d}_-$, and the mononucleotide sequences which encode amino acids Pro and Phe with similar chemical type of radicals have positive $\mathbf{d}_-$.

The dinucleotide double-helix DNA sequences (without mononucleotide parts) are described in the Table 2.

Table 2. Dinucleotide DNA

n_x	$\mathbf{d}_+$	$\mathbf{d}_-$	amino acid
$n_\mathbf{C} \neq 0, n_\mathbf{G} \neq 0$	$7\,(n_\mathbf{C} + n_\mathbf{G})$	$n_\mathbf{C} - n_\mathbf{G}$	Pro,Arg,Ala,Gly
$n_\mathbf{C} \neq 0, n_\mathbf{T} \neq 0$	$7n_\mathbf{C} + 3n_\mathbf{T}$	$n_\mathbf{C} + n_\mathbf{T}$	Pro,Phe,Leu,Ser
$n_\mathbf{C} \neq 0, n_\mathbf{A} \neq 0$	$7n_\mathbf{C} + 3n_\mathbf{A}$	$n_\mathbf{C} - n_\mathbf{A}$	Pro,Gly,Asn,Tur,His
$n_\mathbf{G} \neq 0, n_\mathbf{T} \neq 0$	$7n_\mathbf{G} + 3n_\mathbf{T}$	$n_\mathbf{T} - n_\mathbf{G}$	Gly,Leu,Val,Cys,Trp
$n_\mathbf{G} \neq 0, n_\mathbf{A} \neq 0$	$7n_\mathbf{G} + 3n_\mathbf{A}$	$-n_\mathbf{G} - n_\mathbf{A}$	Gly,Glu,Arg,Lys
$n_\mathbf{T} \neq 0, n_\mathbf{A} \neq 0$	$3\,(n_\mathbf{T} + n_\mathbf{A})$	$n_\mathbf{T} - n_\mathbf{A}$	Leu,Asn,Tur,TERM

The trinucleotide DNA can be listed in the similar, but more cumbersome way. The full DNA sequences consist of nucleotides of all four types and described by (9)-(10).

The introduction of the determinative degree allows us to single out a kind of double-helix DNA sequences which have an additional symmetry. We call a double-helix sequence *purine-pyrimidine symmetric*, if

$$\mathbf{d}_- = 0, \tag{16}$$

i.e. its empiric "strength" vanishes. From (10) it follows

$$n_\mathbf{C} + n_\mathbf{T} = n_\mathbf{G} + n_\mathbf{A}, \tag{17}$$

i.e. $k_n = k_m = 1$, which can be rewritten for one strand

$$n_{pyrimidines} = n_{purines} \tag{18}$$

or as equality of purines and pyrimidines in two strands

$$n_{pyrimidines} = m_{pyrimidines}, \tag{19}$$

$$n_{purines} = m_{purines}. \tag{20}$$

The purine-pyrimidine symmetry (17) has two particular cases:

$$1)\ \begin{matrix} n_\mathbf{C} = n_\mathbf{G}, \\ n_\mathbf{T} = n_\mathbf{A}, \end{matrix} \ -\text{symmetric DNA}, \tag{21}$$

$$2)\ \begin{matrix} n_\mathbf{C} = n_\mathbf{A}, \\ n_\mathbf{T} = n_\mathbf{G}, \end{matrix} \ -\text{antisymmetric DNA}. \tag{22}$$

The first case corresponds to the Chargaff's rule applied to a single strand which approximately holds for long sequences [11], and so it would be interesting to compare transcription and expression properties of symmetric and antisymmetric double-helix sequences.

We have made a preliminary analysis of real sequences of several species taken from GenBank (2000) in terms of the determinative degree. It were considered 10 complete sequences of *E.coli* (several genes and full genomic DNA 9-12 min.), 12 complete sequences of *Drosophila melanogaster* (crc genes), 10 complete sequences of *Homo sapiens* Chromosome 22 (various clones), 10 complete sequences of *Homo sapiens* Chromosome 3 (various clones). We calculated the nucleotide content $N_{\mathbf{C}}, N_{\mathbf{T}}, N_{\mathbf{G}}, N_{\mathbf{A}}$ and the determinative degree characteristics $\mathbf{d}_+, \mathbf{d}_-, q = \mathbf{d}_-/\mathbf{d}_+, k_n$ and v for every sequence. Then we averaged their values for each species. The result is presented in the Table 3.

Table 3. Mean determinative degree characteristics of real sequences

sequence	$\frac{1}{n}\sum \mathbf{d}_+$	$\frac{1}{n}\sum \mathbf{d}_-$	$\frac{1}{n}\sum q \cdot 10^3$	$\frac{1}{n}\sum k_n$	$\frac{1}{n}\sum v$
E.coli	90806	-138	-6.8	1.07	1.38
Drosophila	7325	-70	-8.9	1.09	1.31
Homo sap. Chr.22	337974	6865	1.46	0.987	1.14
Homo sap. Chr.3	806435	-1794	-2.29	1.021	1.55

First of all we observe that all real sequences have high purine-pyrimidine symmetry (smallness of parameter q). Also we see that the relation of purines and pyrimidines in one DNA strand k_n is very close to unity, therefore we have a new small parameter in the DNA theory $(k_n - 1)$ (or q), which characterizes the purine-pyrimidine symmetry breaking. This can open possibility for various approximate and perturbative methods application. Second, we notice from Table 3 that the purine-pyrimidine symmetry increases in direction from protozoa to mammalia and is maximal for human chromosome. It would be worthwhile to provide a thorough study of purine-pyrimidine symmetry and codon usage in terms of the introduced determinative degree by statistical methods, which will be done elsewhere.

Acknowledgments. Authors would like to thank G. Shepelev for providing with computer programs, S. Gatash, V. Maleev and O. Tretyakov for fruitful discussions and J. Bashford, G. Findley and P. Jarvis for useful correspondence and reprints.

References

[1] Singer M., Berg P. *Genes and genomes.* - Mill Valley: *University Science Books*, 1991. - 373 p.

[2] Findley G. L., Findley A. M., McGlynn S. P. *Symmetry characteristics of the genetic code // Proc. Natl. Acad. Sci. USA.* - 1982. - V. **79**. - 22. - P. 7061–7065.

[3] Zhang C. T. *A symmetrical theory of DNA sequences and its applications. // J. Theor. Biol.* - 1997. - V. **187**. - 3. - P. 297–306.

[4] Hornos J. E. M., Hornos Y. M. M. *Model for the evolution of the genetic code // Phys. Rev. Lett.* - 1993. - V. **71**. - P. 4401–4404.

[5] Bashford J. D., Tsohantjis I., Jarvis P. D. *A supersymmetric model for the evolution of the genetic code // Proc. Natl. Acad. Sci. USA.* - 1998. - V. **95**. - P. 987–992.

[6] Duplij D., Duplij S. *Symmetry analysis of genetic code and determinative degree // Biophysical Bull. Kharkov Univ.* - 2000. - V. **488**. - 1(6). - P. 60–70.

[7] Rumer U. B. *Sistematics of codons in the genetic cod // DAN SSSR.* - 1968. - V. **183**. - 1. - P. 225–226.

[8] Forsdyke D. R. *Different biological species "broadcast" their DNAs at different* $(\mathbf{C}+\mathbf{G})\%$ *"wavelengths" // J. Theor. Biol.* - 1996. - V. **178**. - P. 405–417.

[9] Grantham R., Perrin P., Mouchiroud D. *Patterns in codon usage of different kinds of species // Oxford Surv. Evol. Biol.* - 1986. - V. **3**. - P. 48–81.

[10] Karasev V. A. *Rhombic version of genetic vocabulary based on complementary of encoding nucleotides // Vest. Leningr. un-ta.* - 1976. - V. **1**. - 3. - P. 93–97.

[11] Forsdyke D. R. *Relative roles of primary sequence and* $(\mathbf{C}+\mathbf{G})\%$ *in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species // J. Mol. Biol.* - 1995. - V. **41**. - P. 573–581.