

The significance of nucleotides within DNA codons: a quantitative approach.

Alejandro Guerra-Hernández, Miguel Angel Jiménez-Montaño, Carlos Rubén de la Mora-Basáñez
Universidad Veracruzana
Facultad de Física e Inteligencia Artificial
Departamento de Inteligencia Artificial
Sebastián Camacho No. 5, Xalapa, Ver.,
México 91000

E-mail: {aguerra, ajimenez, cdelamora}@uv.mx

Abstract

The identity of the expressed amino acids coded by triplets of nucleotides (codons) in the Genetic Code, appears to depend on the nucleotide position within a codon, as well as its physico-chemical features. Although different orders of significance for the nucleotide position and its physicochemical properties has been reported in literature [8, 3], they are established using largely qualitative criteria. Following the work of D.A. Mac Dónaill [4], we propose a quantitative approach to test the importance of patterns relating nucleotide features to amino acids. Our implementation of this approach in Lisp produced the same results obtained by Mac Dónaill using the PAM250 [1] similarity matrix. We extend these results considering other similarity matrices measuring not only sequence, but structural similarity.

La identidad de los aminoácidos expresados por los codones (tripletes de nucleótidos) en el Código Genético, parece depender de la posición de los nucleótidos en el codón, así como de sus propiedades físico-químicas. Aunque en la literatura se proponen diferentes ordenes de relevancia para las posiciones de los nucleótidos en el codón y sus propiedades [8, 3], éstos se basan ampliamente en criterios cualitativos. Siguiendo la estrategia de D.A. Mac Dónaill [4], proponemos un enfoque cuantitativo a la verificación de la importancia de patrones en los codones, que relacionan propiedades de los nucleótidos con los amino ácidos. Nuestra implementación en Lisp produce el mismo resultado obtenido por D.A. Mac Dónaill, usando la matriz de similitud PAM250 [1]. Estos resultados son confirmados usando otra matriz basada en similitud entre secuencias; y extendidos, al usar matrices basadas en similitud estructural.

1. Introduction

The relationship between nucleotide sequences and protein composition is captured in the Genetic Code. A codon is defined as a triplet of nucleotides $\langle N_1, N_2, N_3 \rangle$ with $N_i \in \{A, C, G, T, \}$. The significance of a nucleotide N_i , is strongly dependent of its position within the codon, generally following the order $N_2 > N_1 > N_3$. For instance, in many cases, changing the third nucleotide makes no difference to the expressed amino acid, and when it does, the resulting amino acid often has similar psycho-chemical properties.

Alternatively it is possible to represent the Genetic Code as a 6-tuple of the form $\langle C_1, H_1, C_2, H_2, C_3, H_3 \rangle$, where C_i represents the chemical nature of nucleotide N_i , i.e., pyrimidine (Y) or a purine (R); and H_i represents the hydrogen bonding strength between the nucleotide N_i and its complement, i.e., strong (S) or weak (W).

M.A. Jiménez-Montaño et al. [3] advocated the following order of significance: $C_2 > H_2 > C_1 > H_1 > C_3 > H_3$. However this suggested hierarchical order, as well as the order proposed by R. Swanson [8]: $C_2 > C_1 > H_2 > H_1 > C_3 > H_3$, are essentially qualitative in nature, making difficult the comparison between competing models. We propose, following D.A. Mac Dónaill [4], a quantitative approach to establish such hierarchical orders of significance.

2. State of Art

The approach proposed by D.A. Mac Dónaill [4] is quite simple, it involves changing a selected nucleotide feature

and observing and quantifying the effect of the change. This introduces the concept of amino acid similarity, where a particular reference is the PAM250 [1] matrix, a measure widely employed in sequence similarity analysis. An element s_{ij} of this matrix is a score of the substitution probability of amino acid i by j over an evolution period, and ultimately reflects the similarity between the amino acids i and j , where less positive and negative scores reflect dissimilarity or difference.

In the BLOSUM62 [2] matrix, every possible identity and substitution is assigned a score based on the observed frequencies of such occurrences in alignments of related proteins. Identities are assigned the most positive scores. Frequently observed substitutions also receive positive scores and seldom observed substitutions are given negative scores. Both, PAM250 and BLOSUM62 are based on sequence similarity.

A. Prlić et al. [6], proposed a matrix based on structure alignments, instead of sequence alignments as it is the case for the PAM and BLOSUM matrices. Structural equivalence is defined by close distances corresponding to residues that occupy similar positions in the structures and resemble each other in their side-chain orientation. Pairs of amino acids were chosen to have high structural but low sequence similarity, and they were used to generate a substitution matrix.

Miyazawa and R.L. Jernigan [5] proposed a matrix which scores the contact energy and the number of contacts for each type of amino acid pair, supplemented with base mutation rates and the effects of the genetic code.

R. Sánchez et al. [7], proposed a Hamming distance between codons to build a Hasse diagram which reflect the codon order, i.e., given two amino acids, it computes the mean distances between their respective codons. Amino acids with large differences have high Hamming distance values.

3. Methodology

Two experiments are possible. First, the significance of the position of the nucleotide N_i within the codon is explored:

1. A nucleotide position $i \in \{1, 2, 3\}$ is selected.
2. Twelve possible nucleotide mappings are possible: $T \rightarrow \{C, A, G\}$; $C \rightarrow \{A, G, T\}$; $A \rightarrow \{G, T, C\}$; $G \rightarrow \{T, C, A\}$.
3. The mappings are applied to all 64 codons in the Genetic Code table $GCode$, at the nucleotide position i . In many instances, the mapping changes the effect of the codons, resulting in a modified genetic code table $GCode'$.

4. The elements of a similarity matrix, e.g. PAM250, BLOSUM62, Prlić, etc., s_{ij} give a numerical measure of the similarity of amino acids i and j .
5. The overall significance S of a mapping may be estimated from the values s_{ij} , summing across the 64 codons:

$$S = \sum_{i,j=1\dots 64} s_{GCode_i GCode'_j} \quad (1)$$

6. An average significance $\sum_{i=1\dots 3} S_i/12$ is obtained considering the twelve possible mappings for each of the three nucleotide positions.

Then, a second experiment considering the physico-chemical features of the codons is performed:

1. A nucleotide position $i \in \{1, 2, 3\}$ and a nucleotide feature C_i or H_i are selected.
2. Two mappings are possible, depending on the feature selected: $f_C : \{R \rightarrow Y, Y \rightarrow R\}$ or $f_H : \{W \rightarrow S, S \rightarrow W\}$. Basically, they correspond to flipping the bit of the selected nucleotide feature.
3. The corresponding mapping is applied to all 64 codons in the Genetic Code table $GCode$, resulting in a modified Genetic Code table $GCode'$. We proceed as we did in the previous experiment (steps 4 and 5), obtaining the overall similarity (eq. 1) for each nucleotide feature.

4. Results

As expected, results for the nucleotide position in the first experiment (table 1) validate the order of significance $N_2 > N_1 > N_3$ for all the matrices considered in our experiments. For Sánchez [7], higher values reflect higher significance (longer distances in the Hasse diagram). For the rest of the matrices lower values reflect higher significance (lower substitution probability).

Matrix	N_1	N_2	N_3
PAM250 [1]	235.91	220.08	281.08
BLOSUM62 [2]	242.25	220.83	292.17
Prlić [6]	157.48	142.38	185.40
Miyazawa [5]	27.26	21.91	31.96
Sánchez [7]	87.69	90.88	73.14

Table 1. Average significance for nucleotide position

Matrix	C ₁	H ₁	C ₂	H ₂	C ₃	H ₃
PAM250 [1]	-16	16	-60	-32	122	276
BLOSUM62 [2]	-2	1	-76	-66	162	302
Prlić [6]	18.14	29.34	-26.93	-30.83	111.54	189.84
Miyazawa [5]	7.06	7.02	-18.71	-13.70	21.06	30.70
Sánchez [7]	173.79	135.56	196.88	140.2	100.42	70.88

Table 2. Overall significance for physicochemical features

The overall significance for the physicochemical features is shown in table 2. The orders of significance induced for these values are shown in table 3.

Matrix	Order
PAM250 [1]	$C_2 > H_2 > C_1 > H_1 > C_3 > H_3$
BLOSUM62 [2]	$C_2 > H_2 > C_1 > H_1 > C_3 > H_3$
Prlić [6]	$C_2 > H_2 > C_1 > H_1 > C_3 > H_3$
Miyazawa [5]	$C_2 > H_2 > H_1 > C_1 > C_3 > H_3$
Sánchez [7]	$C_2 > C_1 > H_2 > H_1 > C_3 > H_3$

Table 3. Orders of significance induced from physicochemical features

5. Conclusions and perspectives

In this work, we take into account, besides the usual PAM250 [1] considered by D.A. Mac Dónaill [4], the BLOSUM62 [2] also based in sequence similarity, validating the same order or relevance.

Also, a matrix obtained from comparison of three dimensional protein structures [6], validates the same orders. A matrix that scores the contact energy and the number of contacts for each type of amino acid pair, supplemented with base mutation rates and the effects of the genetic code [5], gives almost the same order ($H_1 > C_1$, but they obtain almost the same value). These two matrices differ radically in their construction from PAM250 and BLOSUM62 matrices. Therefore the validation of the results obtained by Mac Dónaill [4] is far from trivial.

The matrix proposed by R. Sánchez [7] was expected to induce a different order, since they obtained an alternative hyper-cube from the one proposed by M.A. Jiménez-Montaño [3]. There are only two classes of hyper-cubes for the Genetic Code, so that the order induced by the matrix proposed by R. Sánchez, corresponds to the order of significance proposed by Swanson [8].

The methodology proposed in this work does not establish which order of significance is better, but given a sim-

ilarity matrix, it offers a quantitative approach to establish the associated order of significance.

Acknowledgments

The second author is supported by Fondo Sectorial de Investigación para la Educación SEP-CONACYT, project SEP-2003-CO2.44625.

References

- [1] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. In M.O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, pages 345–352. National Biochemical Research Foundation, Washington, D.C., 1978.
- [2] S. Henikoff and Henikoff J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.*, 89:10915–10919, 1992.
- [3] M.A. Jiménez-Montaño et. al. The hypercube structure of the genetic code explains conservative and non conservative amino acid substitutions *in vivo* and *in vitro*. *Biosystems*, 39(117), 1996.
- [4] D.A. Mac Dónaill and M. Manktelow. Molecular informatics: Quantifying information patterns in the genetic code. *Molecular Simulation*, 30(5):267–272, 30 April 2004.
- [5] S. Miyazawa and R.L. Jernigan. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Prot. Eng.*, 6:267–278, 1993.
- [6] A. Prlić, S.F. Domingues, and M.J. Sippl. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Engineering*, 13(8):545–550, 2000.
- [7] R. Sánchez, E. Morgado, and R. Grau. The genetic code boolean lattice. *MATCH Commun. Math. Comput. Chem*, 53:29–46, 2004.
- [8] R. Swanson. A unifying concept for the amino acid code. *Bull. Math. Biol.*, 46(187), 1984.