

Decomposition of DNA Sequence Complexity

Pedro Bernaola-Galván,^{1,*} José L. Oliver,² and Ramón Román-Roldán³

¹*Departamento de Física Aplicada II, ETSI Industriales, Universidad de Málaga, Málaga, Spain*

²*Departamento de Genética e Instituto de Biotecnología, Universidad de Granada, Granada, Spain*

³*Departamento de Física Aplicada, Universidad de Granada, Granada, Spain*

(Received 30 November 1998)

Profiles of sequence compositional complexity provide a view of the spatial heterogeneity of symbolic sequences at different levels of detail. Sequence compositional complexity profiles are here decomposed into partial profiles using the branching property of the Shannon entropy. This decomposition shows the complexity contributed by each individual symbol or group of symbols. In particular, we apply this method to the mapping rules (symbol groupings) commonly used in DNA sequence analysis. We find that strong-weak bindings are remarkable homogeneously distributed as compared to purine pyrimidine, and that *A* and *T* are the most heterogeneous distributed bases.

PACS numbers: 87.14.Gg, 87.10.+e

Sequence compositional complexity (SCC) is a measure of the spatial heterogeneity in symbolic sequences [1]. The most outstanding feature of SCC is that it increases with the level of detail in the description [2], thus fulfilling one of the key requirements for complexity measures [3]. Such a key feature is revealed by drawing the complexity profile, a representation of SCC as a function of the statistical level of confidence (corresponding to the level of detail at which the sequence is observed) [1]. Nevertheless, the measure of complexity critically depends on the alphabet used to describe the sequence, and thus, for a given sequence, a series of measures can be obtained depending on symbol grouping (mapping rule). This problem is especially acute for DNA, where a wide variety of mapping rules are usually employed [4]. We show here, however, that such measures are related by simple relationships, thus allowing a decomposition of the overall SCC into partial complexities contributed by the different symbols in the sequence.

Symbolic sequences are currently analyzed in a binary fashion, because most of the methods work with numerical symbols, and the only way to avoid conversion artifacts is to group the original alphabet in a new binary one [4]. That is why the previous attempts to decompose DNA sequence heterogeneity were ground on base-base autocorrelation functions [5,6]. Our approach relies on the *branching* or *grouping property* of the Shannon's entropy [7], which is shared by the Jensen-Shannon divergence (JS), the entropic measure used to segment a sequence into subsequences of homogeneous base composition (domains) [8], which is a prerequisite to compute SCC [1]. Once a sequence is segmented into compositional domains according to the method described in [8], the SCC is defined as follows [1]:

$$JS = H(S) - \sum_{i=1}^n \frac{L_i}{L} H(S_i) = \sum_{i=1}^n \frac{L_i}{L} [H(S) - H(S_i)], \quad (1)$$

where S denotes the whole sequence and L its length, S_i the i th domain and L_i its length, and $H(\cdot) = -\sum f \log_2 f$

is the Shannon entropy of the relative frequencies of symbol occurrences $\{f\}$ in the corresponding (sub)sequence. In the procedure of segmentation, the key parameter is the significance level (s). In brief, s is the probability that the difference between each pair of adjacent domains is not due to statistical fluctuations (see [8] for details). So, different values of s can be seen as different levels of detail in the description of the sequence [2]. The SCC as a function of s is what we call the *complexity profile* of the sequence [1].

The complexity decomposition method.—The *branching* property can be used to group symbols, thereby forming derived sequences with reduced alphabets, in turn allowing for the decomposition of SCC into partial divergences [9]. Let us consider a partition of the symbol set (alphabet) $\mathcal{A} = \{a_1, \dots, a_k\}$ in m disjoint subsets $\mathcal{G}_1, \dots, \mathcal{G}_m$ (partial alphabets). Let us associate a new symbol g_i to each partial alphabet \mathcal{G}_i . Considering the grouped alphabet $\mathcal{G} = \{g_1, \dots, g_m\}$, we can derive a new sequence from the original one by substituting each symbol a_j by g_i , such that $a_j \in \mathcal{G}_i$. For example, in a DNA sequence we have $\mathcal{A} = \{A, T, C, G\}$ (the four nucleotides) and the disjoint subsets may be $\mathcal{G}_1 = \{A, G\}$ (purines) and $\mathcal{G}_2 = \{C, T\}$ (pyrimidines) which are usually denoted by R and Y , respectively. So, in this example $g_1 = R$ and $g_2 = Y$, the original sequence being converted into a binary sequence of purines and pyrimidines. The seven decompositions into two partial alphabets that can be done with DNA sequences are shown in Table I.

The branching property for the Shannon entropy is a well-known topic in current information theory textbooks [7]. This property states the following relation for the entropies of the grouped alphabets defined above:

$$H(S, \mathcal{A}) = H(S, \mathcal{G}) + \sum_{j=1}^m \frac{N_j}{L} H(S, \mathcal{G}_j), \quad (2)$$

where the involved alphabets are indicated explicitly. We denote by N_j the number of symbols of S belonging to \mathcal{G}_j . Expression (2) means that the entropy of the original

TABLE I. Decompositions used in DNA sequence analysis.

Type	\mathcal{G}	\mathcal{G}_1	\mathcal{G}_2
2-2 groupings:			
Purine-pyrimidine	$\begin{cases} R = A + G, \\ Y = T + C \end{cases}$	$\{A, G\}$	$\{T, C\}$
Strong-weak	$\begin{cases} S = A + T, \\ W = C + G \end{cases}$	$\{A, T\}$	$\{C, G\}$
Keto-amino	$\begin{cases} K = A + C, \\ M = T + G \end{cases}$	$\{A, C\}$	$\{T, G\}$
1-3 groupings:			
A vs TCG	$\{A, T + C + G\}$	$\{A\}$	$\{T, C, G\}$
T vs ACG	$\{T, A + C + G\}$	$\{T\}$	$\{A, C, G\}$
C vs ATG	$\{C, A + T + G\}$	$\{C\}$	$\{A, T, G\}$
G vs ATC	$\{G, A + T + C\}$	$\{G\}$	$\{A, T, C\}$

sequence can be written as the sum of the entropy of the new sequence (with the reduced alphabet) plus the weighted sum of the entropies of the sequences obtained by considering only the symbols of each partial alphabet. By substituting expression (2) in (1),

$$\begin{aligned} \text{JS}(\mathcal{A}) &= H(S, \mathcal{A}) - \sum_{i=1}^n \frac{L_i}{L} H(S_i, \mathcal{A}) \\ &= \left[H(S, \mathcal{G}) + \sum_{j=1}^m \frac{N_j}{L} H(S, \mathcal{G}_j) \right] \\ &\quad - \sum_{i=1}^n \frac{L_i}{L} \left[H(S_i, \mathcal{G}) + \sum_{j=1}^m \frac{N_{ij}}{L_i} H(S_i, \mathcal{G}_j) \right], \end{aligned} \quad (3)$$

where we denote by N_{ij} the number of symbols belonging to \mathcal{G}_j in the domain S_i . Finally, putting together the terms in \mathcal{G} and the terms in \mathcal{G}_j , we obtain that JS verifies also the branching property:

$$\begin{aligned} \text{JS}(\mathcal{A}) &= \left[H(S, \mathcal{G}) - \sum_{i=1}^n \frac{L_i}{L} H(S_i, \mathcal{G}) \right] \\ &\quad + \sum_{j=1}^m \frac{N_j}{L} \left[H(S, \mathcal{G}_j) - \sum_{i=1}^n \frac{N_{ij}}{N_j} H(S_i, \mathcal{G}_j) \right] \\ &= \text{JS}(\mathcal{G}) + \sum_{j=1}^m \frac{N_j}{L} \text{JS}(\mathcal{G}_j), \end{aligned} \quad (4)$$

where, again, the involved alphabets are indicated explicitly.

This expression shows that the divergence between the domains is decomposed into partial divergences (between the same domains) associated with the grouped alphabet \mathcal{G} and the partial ones \mathcal{G}_j .

In particular, for a DNA sequence, we have $\mathcal{A} = \{A, T, C, G\}$. Consequently, the overall SCC can also be decomposed into partial complexities, thus allowing us

to discern the respective contributions of the symbols (or symbol pairs) in a DNA sequence. We report here such a decomposition, using it to explore several of the intricacies of DNA sequence organization.

For example, Eq. (4) when applied to any one (say, $\{R, Y\}$) of the 2-2 groupings (Table I) gives

$$\begin{aligned} \text{JS}(\mathcal{A}) &= \text{JS}(\{R, Y\}) + \frac{N_A + N_G}{L} \text{JS}(\{A, G\}) \\ &\quad + \frac{N_T + N_C}{L} \text{JS}(\{T, C\}), \end{aligned} \quad (5)$$

and for any one of the 1-3 groupings (say, A vs TCG)

$$\begin{aligned} \text{JS}(\mathcal{A}) &= \text{JS}(\{A, T + C + G\}) \\ &\quad + \frac{N_T + N_C + N_G}{L} \text{JS}(\{T, C, G\}), \end{aligned} \quad (6)$$

where, obviously, $\text{JS}(\{A\}) = 0$. What is actually of interest is comparing the divergences $\text{JS}(\mathcal{G})$ to each other within the same type. Different values of them are meaningful; for example, $\text{JS}(\{A, T + C + G\}) \neq \text{JS}(\{C, A + T + G\})$ means that symbol A distributes along the domains differently than symbol C.

To compare the results corresponding to different symbol grouping, we need some criterion or reference. First, the direct comparison between divergences of the same type of grouping indicates something about the different complexities contributed by the single bases or by the pair of bases. Second, we may compare any of these with the proper mean divergence. For the 1-3 groupings we define

$$\langle \text{JS}_{1,3} \rangle \equiv \frac{1}{4} \sum_{i \neq j \neq k \neq l \in \mathcal{A}} \text{JS}(i, j + k + l), \quad (7)$$

and for the 2-2 groupings

$$\langle \text{JS}_{2,2} \rangle \equiv \frac{1}{3} \sum_{i \neq j \neq k \neq l \in \mathcal{A}} \text{JS}(i + j, k + l), \quad (8)$$

where the sums extend over all 1-3 groupings and 2-2 groupings, respectively.

Finally, it can be shown that the above mean values are approximately the third part of the whole divergence $\text{JS}(\mathcal{A})$ [10], i.e.,

$$\text{JS}(\mathcal{A}) \simeq 3\langle \text{JS}_{1,3} \rangle \quad \text{or} \quad \text{JS}(\mathcal{A}) \simeq 3\langle \text{JS}_{2,2} \rangle. \quad (9)$$

The last relationship suggests $\frac{1}{3}\text{JS}(\mathcal{A})$ as an objective, approximate reference for the partial complexities. In other words, if all alphabets have the same complexity all the profiles will be equal to each other and equal to the reference. We can now properly speak of the complexity decomposition of $\text{JS}(\mathcal{A})$:

$$\begin{aligned} \text{JS}(\mathcal{A}) &\simeq 3\langle \text{JS}_{1,3} \rangle \\ &= \frac{3}{4} \sum_{i \neq j \neq k \neq l \in \mathcal{A}} \text{JS}(i, j + k + l) \\ \text{or } \text{JS}(\mathcal{A}) &\simeq 3\langle \text{JS}_{2,2} \rangle \\ &= \sum_{i \neq j \neq k \neq l \in \mathcal{A}} \text{JS}(i + j, k + l). \end{aligned} \quad (10)$$

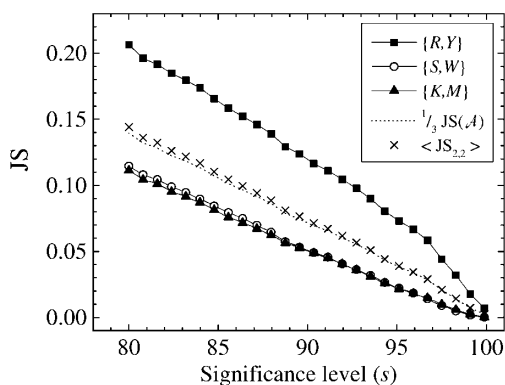


FIG. 1. SCC decomposition for the $\{R, Y\}$, $\{S, W\}$, and $\{K, M\}$ alphabets in an artificial sequence 100 000 base pairs in length obtained using the insertion-deletion model. $\frac{1}{3}JS(\mathcal{A})$ and $\langle JS_{2,2} \rangle$ are also included in order to show the validity of Eq. (9). To generate the sequence we use the same parameter values as suggested by the authors of the model [11], and we slightly modify the model to obtain a four-symbol sequence instead of a binary one (see text).

To attain the SCC decomposed profiles, the following calculations are made: (1) For each significance level s , the sequence is segmented in domains. (2) For each s and domain, the relative frequencies of the four bases are computed. From these, the relative frequencies of the grouped symbols are determined by adding together the appropriate base frequencies. (3) From those relative frequencies, the entropies and the divergences are computed for each s .

As a control experiment, we have generated an artificial sequence using the insertion-deletion model [11], whose SCC values are very close to the ones obtained in real human DNA sequences [12]. This model generates binary sequences of purines pyrimidines which can be converted into four-symbol sequences by randomly substituting each purine by an A or a G and each pyrimidine by a T or C . Obviously, in this sequence, SCC is mainly due to the alternation of purines pyrimidines. This can be observed in Fig. 1 where SCC for the grouping $\{R, Y\}$ is clearly bigger than the ones corresponding to $\{S, W\}$ or $\{K, M\}$ groupings. Note also that $\frac{1}{3}JS(\mathcal{A})$ and $\langle JS_{2,2} \rangle$ values are very close, as we stated in Eq. (9).

Application examples.—SCC decomposition was applied to reveal (1) the relative complexities of $\{R, Y\}$ versus $\{S, W\}$ derived sequences, and (2) the differential contributions of the four bases to total DNA sequence complexity.

(1) The decomposition in binary alphabets was applied to two paradigmatic examples of sequences with (HUMTCRADCV) and without (ECO110K) long-range correlations. Since the discovery of long-range correlations in DNA sequences [13], a recurrent observation has been that the correlations are stronger when the $\{R, Y\}$ mapping rule, instead of $\{S, W\}$, is used. This observation can now be readily explained through SCC decomposition. Sequences with long-range correlations show $\{R, Y\}$

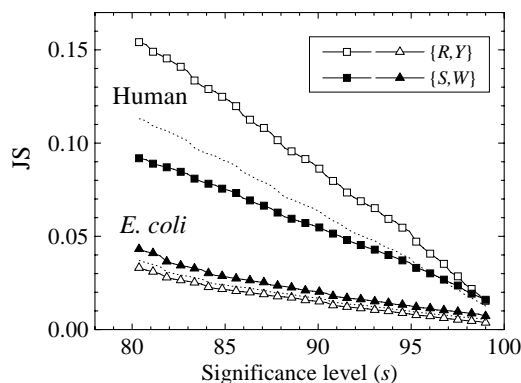


FIG. 2. SCC decomposition for the grouped $\{R, Y\}$ and $\{S, W\}$ alphabets in human (HUMTCRADCV) and *E. coli* (ECO110K) sequences. The dotted lines in both cases correspond to the reference $\frac{1}{3}JS(\mathcal{A})$.

profiles higher than the $\{S, W\}$ ones, whereas the opposite pattern was true for sequences without long-range correlations (Fig. 2). As a result, the complexity differences between both sequences are expected to be maximized when the $\{R, Y\}$ mapping rule is used, which is just what other authors have observed [13].

(2) Four-symbol decomposition, allowing one to compare single-base contributions, shows that A and T profiles were generally above the C and G ones (Fig. 3). Note that the overall complexity of the entire genome sequence of *E. coli* is very similar to that obtained with only a partial sequence (ECO110K, see Fig. 2). Greater complexities for A/T were observed not only in the *E. coli* genome, in which the GC content was around 50%, but also in genomes of more biased base composition as *Mycoplasma genitalium* (32% GC) or *Mycobacterium tuberculosis* (65% GC). The same pattern was found in eukaryotes, from yeast to man (not shown).

Another striking feature shown in Fig. 3 is the parallel or paired behavior of A and T profiles (profile pairing). Such pairing was also shown by C and G profiles, and it was detected even in those few sequences where C/G

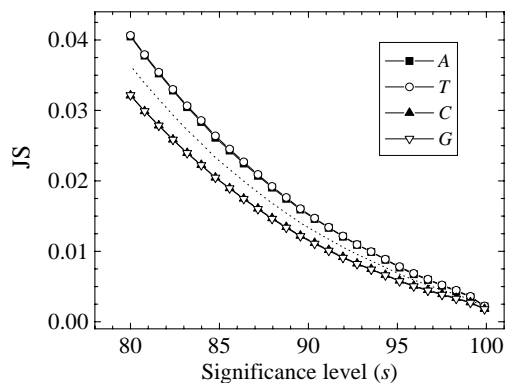


FIG. 3. Four-symbol decomposition in the complete genome of *E. coli*. The dotted line is $\frac{1}{3}JS(\mathcal{A})$.

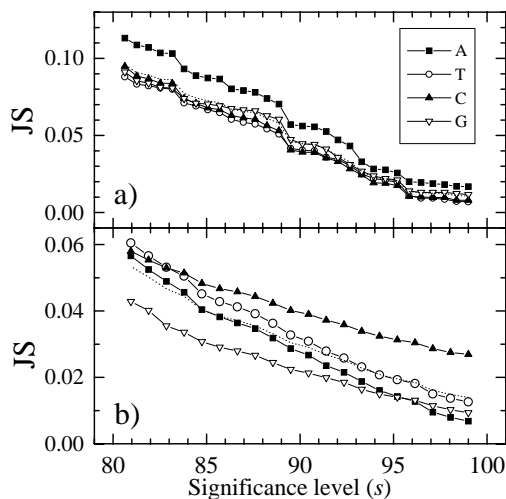


FIG. 4. Four-symbol decomposition in the complete genome of the simian retrovirus SRV-1 (a), and the human herpesvirus 7 HHV7 (b). The dotted line is $\frac{1}{3}JS(A)$.

complexities are above the A/T ones (not shown). Profile pairing may be a manifestation of the strand symmetry in the double helix structure of DNA [14], which in turn may be due to an equilibrium state between symmetric point mutations on both DNA strands [15–17]. This interpretation is supported by the observation that single-stranded retroviral RNA genomes, where strand symmetry is obviously not expected to occur, lack profile pairing (Fig. 4a). We have also observed this lack of profile pairing in double-stranded, but highly asymmetrical [18], viral DNA genomes, as the human herpesvirus 7 (Fig. 4b). Profile pairing may be related to the paired behavior observed for the whole set of base-base autocorrelation functions [5,6].

We acknowledge the critical reading of Wentian Li. This work was supported by Grants No. PB96-1414-

CO2-01 and No. MAR97-0464-CO4-02 from the Spanish Government.

*Corresponding author at Dpto. de Física Aplicada II, E.U.P., Universidad de Málaga, Plza. de El Ejido s/n, E-29071 Málaga, Spain.

Email address: rick@ctima.uma.es

- [1] R. Román-Roldán, P. Bernaola-Galván, and J. L. Oliver, *Phys. Rev. Lett.* **80**, 1344 (1998).
- [2] W. Li, *Complexity* **3**, 33 (1997).
- [3] M. Gell-Mann and S. Lloyd, *Complexity* **2**, 44 (1996).
- [4] H.E. Stanley *et al.*, *Physica (Amsterdam)* **205A**, 214 (1994).
- [5] M. Teitelman and F.H. Eeckman, *J. Comput. Biol.* **3**, 573 (1996).
- [6] W. Li, *Comput. Chem.* **21**, 257 (1997).
- [7] T. Cover and J. Thomas, *Elements of Information Theory* (John Wiley & Sons Inc., New York, 1991).
- [8] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, *Phys. Rev. E* **53**, 5181 (1996).
- [9] P. Bernaola-Galván, Ph.D. dissertation, Universidad de Granada, Spain, 1997 (in Spanish).
- [10] Numerical simulation shows that the deviation with respect to the theoretical value is under 2% for sequences with total entropy differing less than 1% from the upper bound, which holds true for typical DNA sequences.
- [11] S. Buldyrev *et al.*, *Biophys. J.* **65**, 2673 (1993).
- [12] P. Bernaola-Galván, R. Román-Roldán, J. L. Oliver, and P. Carpena, *Comput. Phys. Commun.* (to be published).
- [13] W. Li and K. Kaneko, *Europhys. Lett.* **17**, 655 (1992); C-K. Peng *et al.*, *Nature (London)* **356**, 168 (1992); R. Voss, *Phys. Rev. Lett.* **68**, 3805 (1992).
- [14] J. W. Fickett, D. C. Torney, and D. R. Wolf, *Genomics* **13**, 1056 (1992).
- [15] N. Sueoka, *J. Mol. Biol.* **40**, 318 (1995).
- [16] J. R. Lobry, *J. Mol. Evol.* **40**, 326 (1995).
- [17] W. Li, *Comput. Chem.* (to be published).
- [18] J. Mrázek and S. Karlin, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3720 (1998).