Comparison of Complexity Measures for DNA Sequence Analysis

Ricardo E. Monge
Escuela de Ciencias de la Computación e Informática
Universidad de Costa Rica
e-mail: ricardo.mongegapper@ucr.ac.cr

Juan L. Crespo
Escuela de Ingeniería Eléctrica
Universidad de Costa Rica
e-mail: jlcrespo@eie.ucr.ac.cr

***Abstract*—This paper looks into DNA analysis by computing and comparing complexity measures, in addition to providing a review of recent studies regarding the measurement of DNA complexity. The authors compare Shannon Entropy, Kolmogorov Complexity (approximated by Lempel-Ziv Compressibility) and statistical complexity, and observe that regions corresponding to genes have consistently different complexity measures (i.e., they are more regular) than those regions that do not have any gene associated with them. This provides insight on how to develop new tools for automated DNA analysis.**

Keywords—entropy, DNA entropy, compressibility, coding and non-coding DNA

I. Introduction

The authors' primary interest here is the analysis of complexity measures of genomic sequences as one of the preprocessing techniques that can lead to better pattern recognition and pattern inference in DNA sequences. The authors are engaged in active research regarding the development of new techniques based on bioinspired intelligence, to analyze genomic data and its different relationships with other types of biological data (as found in medical diagnosis and medical imagery).

Initial studies show that complexity metrics along a sequence show a different view of data being analyzed, and patterns may appear in the new sequence. Therefore, it is of the authors' interest to continue pursuing these types of transformation, to convey information more precisely for the computational intelligence algorithms that will be developed as part of future research.

II. Previous work in DNA complexity study

DNA complexity has been studied throughout the years, even when computational tools were not yet mainstream in genetics research. In 1974 [1], the first technique for nucleic acid complexity was developed by counting the amount of hybrid DNA-RNA sequences produced by an embryo. In 1982, Hough-Evans proposed a technique based on chromatography to reveal repetitive patterns in single-celled bacterial organisms [2]. Gusev and others evaluate genetic complexity by finding the amount of repetitive sequences (commonly interpreted as regulatory DNA) with a Lempel-Ziv measure [3]. The role of protein coding DNA and regulatory DNA has been understood only recently, and provides insight on the belief that organism complexity is related to the amount of "extra" DNA [4].

Lempel-Ziv complexity and compressibility has been used outside genomic studies. Xiao and others use Lempel-Ziv to predict a protein sequence location within a cell [5]. Ferenets and others [6] compared different types of entropy, Lempel-Ziv complexity, and fractal dimension to electroencephalograms of patients under anesthesia with the standard clinical depth of sedation score. Aboy, in 2006, used Lempel-Ziv to estimate bandwidth and harmonic variability of diverse quasi-periodic signals recorded for biomedical purposes [7].

Other studies have approached the genomic complexity problem by interpreting the nucleic acid sequence as a fractal. Berthelsen analyzed genomic sequences as a 4-dimensional random walk and computed the resulting fractal dimension, which turned out to be significantly lower than sequences of genomic data generated randomly [8]. Concerning the physical structure (and not the specific sequence), Ercolini and others computed the fractal dimension of DNA imaged under an atomic force microscope [9]. It is possible to show that complexity analysis of DNA has not been under active research during the last four years. On the other hand, research has been oriented towards the usage of next-generation sequencers, and complexity analyses have not been done yet. Complexity analysis of obtained genomic data has been done on RNA and certain types of bacterial DNA, but not on human chromosomes.

III. Comparison method

For the purposes of this study, the authors chose a short sequence of genomic data, taken from the generic Human Genome Project, of 384 kbp (kilo base pairs). Specifically, the data corresponds to Chromosome Y [10], [11] (locus NT_011896) from base number 2,781,480 to base number 3,165,480, encompassing three genes (see gene positions described in Table I). This particular section was chosen for several reasons:

978-1-4799-6174-0/14/$31.00 ©2014 IEEE

TABLE I
GENE LOCATIONS, SPECIFIC TO THE START OF THE DATASET USED

Gene Code	Location	Length
SRY	5,375	886
RNASEH2CP1	8,347	571
RPS4Y1	60,103	25,374

- Chromosome Y is a relatively small human chromosome (57 million base pairs).
- Locus NT_011896 is a block of 6.3 million base pairs without any sequencing gaps.
- The portion selected encompasses both small and large genes in a reasonable quantity (3 genes) without overlapping.
- No SNPs (single-nucleotide polymorphisms) are contemplated for the analysis (for sake of simplifying the computational process).

A moving window (local profile) analysis of 250 base pairs per computation was performed, and a graph of the variation of the specified complexity measure along the full sequence was produced. If there are noticeable differences in certain regions of the complexity measure values, those regions are matched against known information about the source. This makes it possible to determine whether complexity measures can be used to simplify the analysis, pattern matching and feature detection of genomic material. The authors also compared the resulting data with fully random data (that is, sequences of genetic symbols that do not correspond to any specific species). The analysis window was moved 50 base pairs after each computation, in an overlapping manner. The authors have analyzed the first 100,000 base pairs because that selection covers few genes (a small one and a large one), and has a reasonable amount of data to show the feasibility for the computational results (without consuming too many computing resources).

This study compares quantifiable complexity measures [12], such as a) Information Entropy, proposed by Claude Shannon; b) Kolmogorov Complexity [13], proposed by Andrey Kolmogorov and later refined by Gregory Chaitin; and c) statistical complexity [14], [15].

Information Entropy, or Shannon Entropy [16], corresponds to how much a given set of information is unpredictable. If one can predict the next set of events, given the actual information, one has a low-entropy information set. Shannon Entropy, denoted by $H(S)$, is computed by

$$H(S) = -\Sigma\, P(S_i) \log_2 (P(S_i)), \quad (1)$$

where $P(S_i)$ is the relative frequency of character i within string S.

Kolmogorov Complexity is based on the concept that if a given sequence S can be generated by an algorithm smaller than the length of the given sequence, then its complexity corresponds to the size of the algorithm. If a sequence is truly complex and random, the shortest representation is the sequence itself [17]. These models were later refined by Chaitin when *algorithmic information content* was proposed [18]. The issue is that Kolmogorov Complexity itself is not computable. In the present study, the Lempel-Ziv approximation to Kolmogorov Complexity was used. It quantifies the amount that a message can be compressed using the Lempel-Ziv-Welch algorithm [19], which was the complexity measure used by Gusev [3].

TABLE II
STATISTICAL PROPERTIES FOR GIVEN DNA SEQUENCE

Measure	μ	σ
Shannon Entropy	123.15	1.8736
Normalized Lempel-Ziv Compressibility	0.2779	0.0114
C_{LMC} statistical complexity	2637.5	506.9

TABLE III
STATISTICAL PROPERTIES FOR RANDOM DNA SEQUENCE

Measure	μ	σ
Shannon Entropy	124.87	0.1072
Normalized Lempel-Ziv Compressibility	0.2952	0.0067
C_{LMC} statistical complexity	1961.7	50.91

Regarding statistical complexity, of which several measures were studied and analyzed by Feldman and Crutchfield [14], and given the computation time constraints, the authors of this paper decided to opt for a measure that would only require statistical manipulations. The C_{LMC} measure considers both the intrinsic entropy of the data and the departure of the probability of each symbol from uniformity (referred to as disequilibrium by the authors of the measure) [15]. So, the C_{LMC} for a given string S is defined by

$$C_{LMC}(S) = H(S)\, D(S), \quad (2)$$

where H is the Shannon Entropy (defined above), and

$$D(S) = \Sigma\, (P(S_i) - n^{-1})^2 \quad (3)$$

is the disequilibrium, where n is the length of the sequence.

IV. RESULTS

Computational results have been obtained by using the given DNA sequence, for the selected complexity measures.

Figures 1, 2 and 3 contain the plots of each complexity measure for the first 100,000 base pairs. Regarding Lempel-Ziv complexity, the authors plotted normalized data (that is, dividing the numerical result by the original size of the sequence) to show the amount of reduction.

Statistical data for blocks of 20 data points (covering 1,000 base pairs) was computed, to explore gene-size regions and their statistical properties. Regarding Lempel-Ziv, the plot

corresponds to the absolute length of the compressed sequence, thus producing values between 1 and 250. For this trial, the authors chose sections of 1,000 base pairs to cover average gene-size within the selected region of 100,000 base pairs on the human chromosome, but in future work the analysis will cover the entire chromosome.

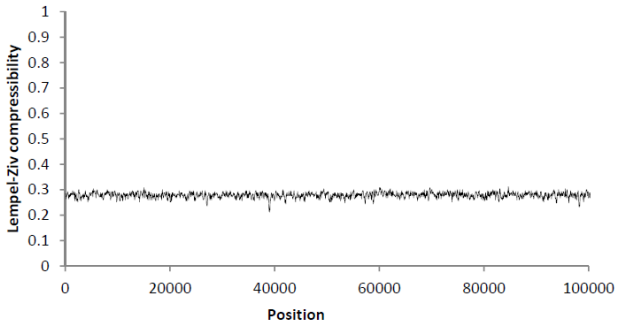

Fig. 1. Lempel-Ziv Compressibility

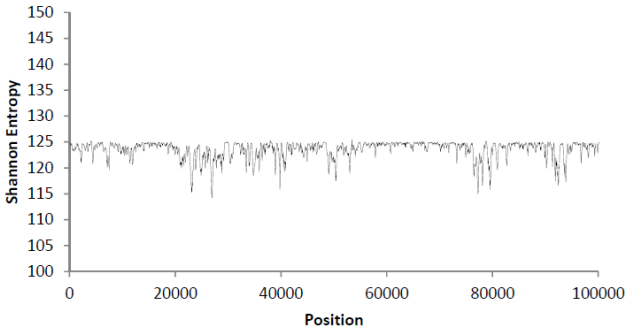

Fig. 2. Shannon Entropy

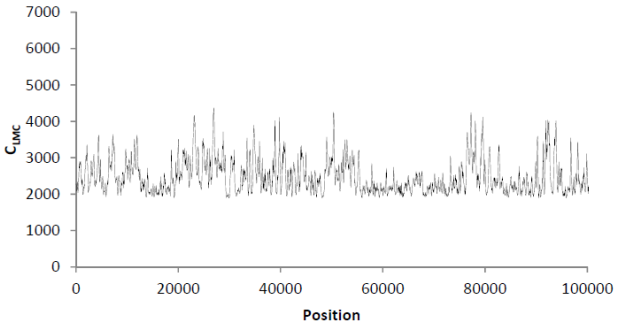

Fig. 3. Statistical complexity

V. Discussion

Genomic material contains information that is required to build life. Therefore, it was to be expected that the information content in chromosomic human DNA (the DNA tested in this study) would be slightly higher than for a generated fully random sequence of DNA nucleotides. Only one of the measures used (C_{LMC}, which measured statistical complexity) showed a significant difference between random and non-random datasets. Also, it is worth emphasizing that real DNA varies much more than the random sequence (more departures from uniform distributions), in all but one of the complexity measures used, by at least one order of magnitude. Tables II and III contain general statistical data of both the authentic genomic data and the random genomic sequence.

TABLE IV
Statistical properties for selected portions, Lempel-Ziv measure

Positions	$\|\Delta\|$	μ	σ	Gene
4000–5000	9	68.526	2.318	
5000–6000	9	72.053	2.147	SRY
6000–7000	7	70.316	2.056	
7000–8000	8	69.684	2.335	
8000–9000	10	70.789	2.800	RNASEH2CP1
9000–10000	11	69.632	3.077	
10000–11000	8	68.632	2.565	
12000–13000	8	70.474	2.294	
13000–14000	9	70.526	2.632	
14000–15000	11	70.000	2.906	
15000–16000	12	70.632	3.148	
16000–17000	8	68.684	2.110	
17000–18000	7	68.474	2.366	
18000–19000	10	69.632	2.241	
60000–61000	9	72.789	2.679	RPS4Y1
61000–62000	7	70.895	2.158	RPS4Y1
62000–63000	8	71.421	2.293	RPS4Y1
63000–64000	9	69.000	2.134	RPS4Y1
64000–65000	9	69.474	2.480	RPS4Y1
65000–66000	8	71.000	1.944	RPS4Y1
66000–67000	7	71.105	1.696	RPS4Y1
67000–68000	8	69.000	2.186	RPS4Y1
68000–69000	9	70.105	2.183	RPS4Y1
69000–70000	13	71.000	3.815	RPS4Y1
70000–71000	9	70.579	2.219	RPS4Y1
82000–83000	11	69.211	3.310	RPS4Y1
83000–84000	11	69.053	3.325	RPS4Y1
84000–85000	11	70.316	2.982	RPS4Y1
85000–86000	5	70.947	1.471	
86000–87000	8	69.158	2.410	

The Lempel-Ziv compressibility measure only shows that DNA sequences can be generated by a deterministic process, but it cannot distinguish between coding and non-coding DNA. The variation of the complexity measure resembles random noise, and as seen in Table IV, there is no significant difference between blocks corresponding to genes and those blocks that do not correspond to genes. The average compressibility measure for the data analyzed is similar to the average compressibility measure for random data, thus indicating that the Lempel-Ziv measure analyzes *algorithmic content*, not *information*.

TABLE V
STATISTICAL PROPERTIES FOR SELECTED PORTIONS, ENTROPY MEASURE

Positions	$\|\Delta\|$	μ	σ	Gene
4000–5000	4	123.587	1.295	
5000–6000	1	124.538	0.319	SRY
6000–7000	2	124.033	0.617	
7000–8000	5	122.432	1.830	
8000–9000	1	124.374	0.336	RNASEH2CP1
9000–10000	2	123.715	0.664	
10000–11000	2	123.483	0.599	
12000–13000	4	123.775	0.942	
13000–14000	1	124.502	0.305	
14000–15000	2	124.496	0.556	
15000–16000	1	124.591	0.325	
16000–17000	1	124.515	0.317	
17000–18000	1	124.574	0.300	
18000–19000	2	124.219	0.597	
60000–61000	2	124.421	0.618	RPS4Y1
61000–62000	1	124.639	0.281	RPS4Y1
62000–63000	1	124.720	0.224	RPS4Y1
63000–64000	1	124.623	0.293	RPS4Y1
64000–65000	2	124.427	0.441	RPS4Y1
65000–66000	2	124.663	0.415	RPS4Y1
66000–67000	1	124.750	0.262	RPS4Y1
67000–68000	2	124.002	0.620	RPS4Y1
68000–69000	1	124.805	0.140	RPS4Y1
69000–70000	1	124.706	0.174	RPS4Y1
70000–71000	2	124.087	0.441	RPS4Y1
82000–83000	4	123.253	1.309	RPS4Y1
83000–84000	2	124.385	0.474	RPS4Y1
84000–85000	1	124.713	0.198	RPS4Y1
85000–86000	1	124.233	0.416	
86000–87000	3	124.209	0.679	

TABLE VI
STATISTICAL PROPERTIES FOR SELECTED PORTIONS, C_{LMC} MEASURE

Positions	$\|\Delta\|$	μ	σ	Gene
4000–5000	1292	2781.628	403.539	
5000–6000	563	2162.927	156.993	SRY
6000–7000	1151	2682.088	374.100	
7000–8000	1326	2931.770	457.249	
8000–9000	544	2242.988	177.642	RNASEH2CP1
9000–10000	1071	2610.617	324.207	
10000–11000	855	2742.618	248.924	
12000–13000	1470	2537.717	368.554	
13000–14000	496	2175.645	160.778	
14000–15000	570	2124.845	175.512	
15000–16000	360	2071.461	106.019	
16000–17000	536	2133.432	149.455	
17000–18000	584	2196.890	184.021	
18000–19000	1268	2336.025	351.687	
60000–61000	774	2161.575	211.628	RPS4Y1
61000–62000	230	2130.729	67.447	RPS4Y1
62000–63000	755	2217.256	196.066	RPS4Y1
63000–64000	395	2093.803	114.886	RPS4Y1
64000–65000	502	2166.660	137.947	RPS4Y1
65000–66000	574	2169.788	150.108	RPS4Y1
66000–67000	414	2401.268	120.999	RPS4Y1
67000–68000	674	2384.215	184.940	RPS4Y1
68000–69000	298	2035.613	90.990	RPS4Y1
69000–70000	368	2155.954	112.482	RPS4Y1
70000–71000	556	2249.362	148.024	RPS4Y1
82000–83000	1179	2571.750	399.533	RPS4Y1
83000–84000	509	2208.583	154.074	RPS4Y1
84000–85000	380	2069.896	111.885	RPS4Y1
85000–86000	427	2250.024	147.913	
86000–87000	793	2225.019	220.119	

For the measure based on Shannon Entropy (see Table V and Figure 2), gene-coding regions happen to have a lower variability (identified by a low Δ and a low σ) than their non-coding regions do. For regions that correspond to genes, both statistical measures are much lower than their non-coding regions (for example, compare the Δ for the 4000-5000 block and the corresponding value for the 5000-6000 block: 4.37 > 1.103). If an experimentally determined limit of $\Delta < 1.2$ and $\sigma < 0.5$ is established, it is possible to extract all *gene coding regions* that appear in Table I. There is a block (13000-18000) that has the same characteristics, but it is not associated (yet) to any gene. Does this mean that the measure is incorrect? It might be that some gene has not yet been identified.

Statistical complexity, which is presented in Table VI and Figure 3, shows behavior similar to that of Shannon Entropy. It is partially based on Shannon Entropy, as seen in Equation 2. Base-pair blocks that do not contain gene information tend to have more variation (on average, twice the σ than for non-coding blocks). Compare, as an example, the σ for the 4000-5000 block and the corresponding value for the 5000-6000 block: 403.539 > 156.993). C_{LMC} offers a greater difference between the different types of DNA codes, but it results in more false misses, such as the case of $\Delta < 600$ and $\sigma < 200$; it will not identify the start of gene RPS4Y1 ($\sigma_{60000\text{-}61000}$ = 211.628), but it will work for the rest of the gene ($\sigma_{61000\text{-}75000} <$ 200). There is some variation at the end of gene RPS4Y1, ($\sigma_{82000\text{-}83000}$ = 399.533), as if it corresponded to a non-gene section.

Worth more research is the fact that gene-coding DNA is more regular than DNA that does not have genes or proteins linked to it. Is there a reason for that regularity? Why would non-coding regions be more *random*?

VI. CONCLUSIONS

In the first place, it is evident that complexity measures are different for gene-coding DNA than for the respective measures for regulatory DNA, even for DNA corresponding to

complex organisms (that is, not only unicellular organisms). The fact that a pattern can be determined by means of statistical operations shows that there is a research opportunity regarding pattern matching and advanced probabilistic and computational intelligence techniques to perform better global analyses of DNA.

A second interesting aspect of the study conducted is the relationship between complexity and information content. The complexity measures that were not based on information content do not provide much information for further analysis.

Finally, these results appear to indicate that future work should address analyzing larger subsets of DNA; proposing better complexity measures adapted for the short alphabet of DNA; and providing insight on pattern analysis for different types of biological and medical data, of which DNA is only part of the solution. Future work will also be directed toward studying single-nucleotide variations and assessing the effect of window size (50, at the moment) in the computational results and statistical significances.

ACKNOWLEDGMENT

The authors would like to thank the University of Costa Rica, for providing an encouraging research environment and the anonymous reviewers of this paper, for their clear and succinct suggestions regarding the study.

REFERENCES

[1] G. A. Galau, R. J. Britten, and E. H. Davidson, "A measurement of the sequence complexity of polysomal messenger RNA in sea urchin embryos," *Cell*, vol. 2, no. 1, pp. 9–21, 1974.

[2] B. R. Hough-Evans and J. Howard, "Genome size and DNA complexity of plasmodium falciparum," *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, vol. 698, no. 1, pp. 56–61, 1982.

[3] V. D. Gusev, L. A. Nemytikova, and N. A. Chuzhanova, "On the complexity measures of genetic sequences," *Bioinformatics*, vol. 15, no. 12, pp. 994–999, 1999.

[4] R. J. Taft, M. Pheasant, and J. S. Mattick, "The relationship between non-protein-coding DNA and eukaryotic complexity," *Bioessays*, vol. 29, no. 3, pp. 288–299, 2007.

[5] X. Xiao, S. Shao, Y. Ding, Z. Huang, Y. Huang, and K.-C. Chou, "Using complexity measure factor to predict protein subcellular location," *Amino Acids*, vol. 28, no. 1, pp. 57–61, 2005.

[6] R. Ferenets, T. Lipping, A. Anier, V. Jantti, S. Melto, and S. Hovilehto, "Comparison of entropy and complexity measures for the assessment of depth of sedation," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1067–1077, 2006.

[7] M. Aboy, R. Hornero, D. Abásolo, and D. Álvarez, "Interpretation of the Lempel-Ziv complexity measure in the context of biomedical signal analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 11, pp. 2282–2288, 2006.

[8] C. L. Berthelsen, J. A. Glazier, and M. H. Skolnick, "Global fractal dimension of human DNA sequences treated as pseudorandom walks," *Physical Review A*, vol. 45, no. 12, p. 8902, 1992.

[9] E. Ercolini, F. Valle, J. Adamcik, G. Witz, R. Metzler, P. De Los Rios, J. Roca, and G. Dietler, "Fractal dimension and localization of DNA knots," *Physical Review Letters*, vol. 98, no. 5, p. 058102, 2007.

[10] K. D. Pruitt, G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey *et al.*, "Refseq: an update on mammalian reference sequences," *Nucleic acids research*, vol. 42, no. D1, pp. D756–D763, 2014.

[11] I. H. G. S. Consortium *et al.*, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, 2004.

[12] M. Mitchell, *Complexity: A Guided Tour*, Oxford, 2009.

[13] M. Li and P. M. Vitányi, *An introduction to Kolmogorov complexity and its applications*, New York, 2009.

[14] D. P. Feldman and J. P. Crutchfield, "Measures of statistical complexity: Why?" *Physics Letters A*, vol. 238, no. 4, pp. 244–252, 1998.

[15] R. Lopez-Ruiz, H. L. Mancini, and X. Calbet, "A statistical measure of complexity," *Physics Letters A*, vol. 209, no. 5, pp. 321–326, 1995.

[16] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[17] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems of Information Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[18] G. J. Chaitin, "On the length of programs for computing finite binary sequences," *Journal of the ACM (JACM)*, vol. 13, no. 4, pp. 547–569, 1966.

[19] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 75–81, 1976.