Lost Strings in Genomes: What Sense Do They Make?

Michael Sadovsky[1,3(H)], Jean-Fred Fontaine[2'4], Miguel A. Andrade-Navarro[2'4], Yury Yakubailik[3], and Natalia Rudenko[3]

[1] Institute of Computational Modelling of SB RAS, Akademgorodok, 660036 Krasnoyarsk, Russia msad@icm.krasn.ru [2] Johannes Gutenberg-Universitat Mainz, 55128 Mainz, Germany fontaine@uni-mainz.de [3] Institute of Space and Information Technologies, Siberian Federal University, Kirenskogo Str., 26, 660074 Krasnoyarsk, Russia yura yak@mail.ru, nrudnko@gmail.com [4] Institute of Molecular Biology, 55128 Mainz, Germany
http://icm.krasn.ru

Abstract. We studied the sets of avoided strings to be observed over a family of genomes. It was found that the length of the minimal avoided string rarely exceeds 9 nucleotides, with neither respect to a phylogeny of a genome under consideration. The lists of the avoided strings observed over the sets of (related) genomes have been analyzed. Very low correlation between the phylogeny, and the set of those strings has been found.

Keywords: Order · Diversity · Composition · Combinatorics · Evolution · Selection

1 Introduction

A frequency dictionary W_q of nucleotide sequences is claimed to be an entity bearing a lot of information on that latter [1-6]. A consistent and comprehensive study of frequency dictionaries answers the questions concerning the statistical and information properties of DNA sequences. Let's introduce some basic definitions. Consider a continuous symbol sequence from four-letter alphabet N = {A, C, G, T} of the length N; the length here is just the total number of symbols (nucleotides) in a sequence. The sequence is supposed to be relevant to some genetic entity (genome, chromosome, etc.). No other symbols or gaps in the sequence take place by supposition. Any coherent string $u\!>\, = v_1v_2 \ldots v_q$ of the length q makes a word. A set of all the words occurred within a sequence yields the support of that latter. Counting the numbers of copies n_u of the words, one gets a finite dictionary; changing the numbers for the frequency

$$_f = {}^{luJ} = N$$

one gets the frequency dictionary W_q of the thickness q. This is the main object of our study.

That is a common place that researchers study frequency dictionaries comprising the observed words. Here we make the hypothesis that any string $u\!>\, = v_1v_2 \ldots v_q$ of the length q to be found in a sequence may have a functional or control value. On the other hand, the total number of words of the length q (in the four-letter alphabet N) grows in capacity exponentially:

$$M(\mathrm{q}) = 4^{\mathrm{q}}\,, (1)$$

where M(q) is the number of all possible words of that length. Obviously, the value determined by (1) becomes to exceed N ,as $q > q^*$. This specific figure is determined in very simple way:

$$q^* = \max\,\{4^{\mathrm{q}} < N\}, (2)\ _q$$

and obviously is rather small for any real genetic entity.

Definition. Support Q of a frequency dictionary W_q is the set of words incorporated into the dictionary, through a search over the given text. If support Q contains all possible words of the given length q (i.e. $\|Q\| = 4^q$), we'll call it *-support. $\|A\|$ means a number of elements in A.

In general for biological sequences, a support Q is not *-support, for sufficiently long q. Suppose, then one to start to develop a series of frequency dictionaries with growing q

$$W_1, W_2, W_3, \qquad W_q. \quad (3)$$

In commonly studied biological sequences (e.g. genes or genomes), W_{-} has usually *-support. The same is true for W_2 and W_3 provided a minimal sequence length.

Let now consider some (sufficiently long) genetic sequence. That latter may be a bacterial genome, or a chromosome, if eukaryote is studied, or a genome of organella. Let now develop a series of frequency dictionaries (3) and focus on their supports, solely. Evidently, there exists the shortest length of words q so, that W_{q-1} has *-support Q(q — 1), but W_q itself has the support Q(q) that is not a *-support.

Hence, q is the minimal length of words that yields some lacunae in the support. Consider then this word (that is always easy to determine that latter). Our basic idea is that these words are not occasional, or randomly lost among the other ones; on the contrary, they are lost due to specific (anti)selection. Thus, a researcher can contribute a lot from a (detailed) study of such words.

2 Lost Strings and Evolution

We hypothesize that a set of the lost (or avoided) strings observed over a family of genetic entities is not random, but follows biologically inspired constraints; indeed, they are a matter of natural selection. This idea, in few various forms, has been formulated earlier [11,12]. Related ideas on the impact of the avoided strings on the structure (and functioning) of cancer genes is discussed in [8]; a comparative study of the avoided stings observed in assembling of a human genome is provided by [7], meanwhile, this study seems to be rather speculative. Finally, more or less theoretically charged paper [9] presents an analysis of evolution on lost strings patterns.

In our study, all genetic sequences have been downloaded from EMBL-bank; any extra symbols falling beyond the alphabet N were omitted, and the parts of a sequence split by those extra symbols were concatenated. The length of a sequence includes the eliminated symbols; an error here does not exceeds 10^{-3}.

2.1 How to Test an Interrelation Between Evolution and Lost Strings

More precisely, if the lost strings were eliminated by natural selection (not randomly) from a sequence, then one should expect that phylogenetically close species must exhibit similarity in the lost strings lists.

Reciprocally, from the analysis of mostly independent genetic entities (whatever one could understand for *mostly* and *independent*) one would be able to observe only elimination of strings.

To test this, one should develop a set of randomly selected entities, and do all the same with it. Surely, the words *randomly selected list of genetic entities* must be defined in some way precisely. For example, one might take a random sampling of the sequences from the list of mitochondrion genomes. These latter are rather short (if plant mitochondria are omitted), have identical function, and the genome consists of a single chromosome.

2.2 Random Test

Another important question is whether an observed list of the losses in the supports of various genetic entities differ from a random one. And one more question here concerns the combinatorics constraints for the "survived" words. These are two different, while strongly related questions.

What kind of a sequence model should be analyzed? Obviously, we shall not study a random non-correlated sequence; on the other hand - why not? What if even a very simple model yields a combinatorial constraints that are pretty close to those observed on some genetic entities?

If a model is not the random non-correlated sequence (Bernoulli process realization), then what type of a model is to be chosen? The very first idea is to compare to some Markov process. So, what parameters of that latter should be applied? And the most important - what is the lowest order of this Markovian process model? Some important results obtained in that direction could be found in [10,12-14].

Table 1. The figures for the length of the shortest avoided strings, for some shorter genomes. *N* is the length of a genetic entity, *L* is the least avoided string length, and *K* is the total number of the lost strings of the length L.

ID	Organism	*N*	*L*	*K*
FR775227	*Salmonella enterica subsp. mitochondrion*	17569	5	2
HQ184045	*Bos taurus isolate Mcg375 mitochondrion*	16340	5	13
JF727176	*Pan troglodytes isolate Flo mitochondrion*	16557	5	13
KC469587	*Sus scrofa domesticus breed pietrain mitochondrion*	16612	5	12
KM061558	*Canis lupus familiaris isolate Cfstp64 mitochondrion*	16730	5	13
AY217738	*Eimeria tenella*	34750	5	4
AY945289	*Fusarium oxysporum strain F11 mitochondrion*	34477	5	2
DQ508940	*Debaryomyces hansenii mitochondrion*	29462	5	2
DQ642846	*Plasmodium falciparum HB3*	29529	5	92
JQ864234	*Candida albicans strain L296 mitochondrion*	33631	5	8
EU651892	*Hemiselmis andersenii strain CCMP 644 mitochondrion*	60553	5	2
FR775213	*Salmonella enterica subsp. enterica serovar Weltevreden*	64694	6	33
FR775245	*Salmonella enterica subsp. enterica serovar Weltevreden*	63517	5	4
HG004427	*Campylobacter fetus subsp. venerealis*	61142	5	9
KF285530	*Ostreococcus tauri isolate RCC1123 chloroplast*	67681	6	13
AB042240	*Triticum aestivum chloroplast*	134545	7	399
CP00224	*Candidatus Tremblaya princeps PCIT*	138927	6	3
FR775217	*Salmonella enterica subsp. enterica serovar Weltevreden*	131230	6	19
JN861109	*Oryza sativa Indica Group cultivar Hassawi chloroplast*	134448	7	361
X86563	*Zea mays complete chloroplast genome*	140384	7	351
CP000351	*Leptospira borgpetersenii, chromosome 2*	299762	7	85
CP002163	*Candidatus Sulcia muelleri CARI,*	276511	6	99
CP007234	*Candidatus Sulcia muelleri strain TETUND*	270029	6	81
FR775191	*Salmonella enterica*	227697	6	16
FR775236	*Salmonella enterica serovar Weltevreden*	253936	6	9
AY506529	*Zea mays strain NB mitochondrion*	569630	8	1059
CP002243	*Candidatus Moranella endobia PCIT*	538294	7	6
CP003000	*Blattabacterium*	587248	6	10
CP003771	*Mycoplasma genitalium M6282*	579504	6	11
CP006771	*Mycoplasma parvum str. Indiana*	564395	6	8

Table 2. The figures for the length of the shortest avoided strings, for some longer genomic entities. Definitions are the same as in Table 1.

ID	Organism	N	L	K
AL954800	*H. sapiens,* chromosome 14	87191216	9	110
CM000265	*H. sapiens,* chromosome 14	87 316 725	9	110
CM000856	*Callithrix jacchus,* chromosome 1	210 400 635	9	3
CM000878	*Callithrix jacchus,* chromosome X	142 054 208	9	62
CM000879	*Callithrix jacchus,* chromosome Y	2 853 901	7	4
CM000001	*Canis lupus familiaris,* chromosome 1	122 678 785	9	4
CM000002	*Canis lupus familiaris,* chromosome 2	85 426 708	9	11
CM000003	*Canis lupus familiaris,* chromosome 3	91 889 043	9	41
CM000004	*Canis lupus familiaris,* chromosome 4	88 276 631	9	63

Table 3. The figures for the length of the shortest avoided strings, for some *E.coli* genomes; K is the total number of the lost shortest strings.

ID	Length	K	ID	Length	K	ID	Length	K
AE005174	5 528 445	86	AP009048	4 646 332	170	CU928163	5 202 090	118
CU928162	5 209 548	95	FM180568	4 965 553	118	U00096	4 641 652	176*
AE014075	5 231428	92	CU928160	4 700 560	199*	AM946981	4 558 947	213*
CU928161	5 032 268	101	CU928164	5132 068	134	CP001925	5 386 223	88

3 Some Preliminary Results

Here we provide some preliminary data on the behaviour of the shortest avoided strings in various genomes. Few words should be said on the structure of Table 1. It consists of six blocks, where each block contains five genomes of approximate length; that latter varies from ~ 15 000 to ^600 000 nucleotides, with approximately equal length step. The idea was to check whether the growth of L is logarithmic, or not. Of course, it might be affected by the genetic material choice; meanwhile, it brings some raw results. Indeed, Table 1 demonstrates that the growth increment is definitely less than ln2. This pattern is supported, in general, by the data shown in Table 2.

3.1 Lost Strings Sets

Both tables show significant variation in abundance of the lists of lost strings. First of all, let's focus on the dependence of L figure on the length of a sequence under consideration. The growth of L is significantly slower than a typical exponent. A 15 thousand fold growth of the length of a sequence results in a growth of L from $L = 5$ to $L = 9$. This figure seems to be universal: whatever real genetic entity is taken for analysis, one gets $L = 9$; at least for mammalian genomes (probably, some plants genomes, say larch genome, may yield $L = 10$, and one hardly could expect a greater figure).

The number of the avoided strings observed over a sequence is much more sensitive to the length of that former.

3.2 Closely Related Strains and the Avoided Strings

To investigate the impact of phylogeny on the composition of the list of the lost strings observed over a family of genomes, we comprised the lists for a set of *E.coli* genomes, of various strains (see Table 3). Table 3 comprises the genomes of bacteria *E.coli* that is well studied and widely spread object for genetic studies. Actually, EMBL-bank contains 101 genomes of various strains of *E.coli*;wehave used twelve of them, randomly chosen.

All the genomes (except three ones) exhibit $L = 8$; Table 3 shows the abundance figures of those losses. There are three genomes (these are **AM946981**, **CU928160**,and **U00096** entries) that have $L = 7$; meanwhile, it should be said, that unlike the other patterns, here a single absent string has been observed, in each genome. All these entries are marked with asterisk, at the table. This fact seems to be an exclusion itself: indeed, a list of the lost strings is generally several times longer. One might expect that the abundance of lost strings of the given length q should yield the similar figure of the part of the total number of string of the length q. A single string lost among octanucleotides is equal to 16384^{-1} that significantly differs from a typical figure observed for $q = 9$ and greater.

Anyway, a comparison of those lost 7-tipples is of a great interest. These lost strings are GTCTAGG for **AM946981,** CCTAGGA for **CU928160,**and GCCTAGG for **U00096.** First of all, GC-content for these strings is 4/7, 4/7, and 5/7, respectively. Another remarkable feature is that the strings posses a quasi-palindromic structure. All these three septanucleotides could be easily (and obviously aligned: they have the common "kern" CTAGG of the length $q = 5$.

There are rather few common lost octanucleotides among these twelve bacterial genomes. There are two strings (GAGTCTAG, GGGTCTAG) found in all twelve genomes, two strings (GGCCTAGG, GTCCTAGG) found in eleven genomes, and two strings (ACTAGTCG, ATGCCTAG) found in nine genomes. A pairwise alignment yields very good concordance, for the first and the second couples of the strings: they have common "kerns" GTCTAG and CCTAGG of the length $q = 5$, respectively. The last couple exhibits lower concordance level with only tetranu-cleotide CTAG common in them. A typical number of the octanucleotides common for several genomes varies from five to seven.

Remarkably, sequence CTAG found in all sequences discussed above is a recognition site of many bacteria or *Archaea* restriction enzymes (e.g. Bfal, CchI, Fgol or Htu [17]). Longer sequence CCTAGG is also a known recognition site of various restriction enzymes (e.g. Avrll and XmaJI).

4 Discussion

To analyze the sets of the lost strings and the properties of those sets one should start from answering the question towards the choice (or definition) of a reference sequence. There might be a number of references, and we start from that one mentioned in Sect. 2.1. Indeed, consider a sequence of the length $N$from four-letter alphabet N, with the following frequencies of symbols: /д = a, fc = в, /c = Y and /т = S,sothat a + в + γ+ S = 1. Stipulating that the symbols occur independently, and the sequence is not correlated, one has a probability (or a frequency) of a string u>of the length q to be combined from the frequencies of individual symbols through their product. Hence, the least probable string looks like a tract of the same symbol (e.g. A) of the length q with the reciprocal frequency equal to /Д.

Obviously, no one has ever observed so far such long lost tracts; this is not a trick, but another evidence of rather non-random structuredness of DNA sequences. Another approach to choose a reference sequence is to implement a Markov model surrogate sequence, with proper frequencies of к-tipples. There is no problem to develop the frequency of /-tipple from к-tipples (l > k)(see details in [1-3]):

$$7 \qquad = \underline{\prod 3 = 1\ {}^{\prime v}j\ {}^{v}j+i^{v}j+3\cdots{}^{v}j+k\text{-}i^{v}j+k}\ (.)$$

$$1\ \mathrm{l}j = 2\ {}^{/}\ {}^{v}j \qquad +\ 3\ \blacksquare\blacksquare\blacksquare{}^{v}j\ +\ k\text{-}l\ {}^{v}j\ +\ k$$

a bit more problem is to find out the string $u>$ with the minimal $/_{\text{ш}}$.

Minimal $/\mathrm{V}_l\mathrm{V}_2\mathrm{V}_3\ldots\mathrm{V}_{l\text{-}l}\mathrm{V}_l$ value from (4) yet does not provide an answer towards the question on a (minimal) lost string. An order of the reciprocal Markov chain is a point here. Still, there is no natural way to choose some specific order κ of Markov chain. Moreover, since a sequence under consideration is finite, then there always exists such order κ^*, that provides an absolutely exact matching of a simulated sequences to the given real one.

This point makes a comparative approach rather acute: one should compare the lost strings from two (or more) very closely related sequences, whose bearers are proven to be real close relatives. Such kind of study arises a question on the proximity measure of two (or more) strings. Here alignment looks rather feasible, since the strings under consideration are not longer that 10 nucleotides. Meanwhile, a diversity and specificity of various versions of alignment (both in algorithmic sense, and in software implementations) brings a problem here: probably, one has to revisit an alignment technique to find out the best one, for the short strings. Of course, such choice must be provided with a clear and concise proof of the efficiency of a method.

Another approach looking rather powerful consists in a study of so called trees of the lost strings. Suppose, we have identified the set of the shortest strings within a sequence. At the next step, we should identify the lost strings that are a symbol longer, but they do not inherit found at the first step. Consider two strings $UJ1$ and $UJ2$ so that $\backslash\ UJ_1\ | = \kappa$ and $\backslash\ UJ_2 \backslash = \kappa + 1$ (here $\backslash\ UJ_1\ |$ means the length of a string). A string $UJ2$ inherits a string w_1, if $UJ1\ C\ UJ_2$. Thus, suppose the set of the shortest lost strings A is identified; next, let us identify the set A of the lost strings that are one symbol longer, and no one s $G\ A$ inherits any $s\ G\ A$. More detail discussion of all these patterns requires further studies and falls beyond the scope of this paper.

Avoidance in bacterial genomes of palindromes related to restriction enzymes has been mentioned before [15], especially in bacteria using type II restriction-modification systems as defensive mechanism against inappropriate invasion of foreign DNA [16]. Our results show that the shortest avoided strings in 12 E. Coli genomes correspond to such palindromes. With the increasing availability of genome sequences, further work could focus on the identification of additional avoided functional sites in bacterial or animal genomes.

Let's have a more detail look at the formula (4); in particular, the version of the extension of /-tipple into $l+1$-tipple. Here the formula (4) changes for

$$7 \qquad \_ \qquad \underline{{}^{f}V\backslash V2\ Vi\cdots Vl\text{-}\backslash Vl \quad \mathrm{X} \quad f_{v2ViV\&\cdots Vl\text{-}\backslash VlVl\ +\ \backslash}}\ (5)$$

Formula (5) looks like a Markov process expression, while it is not: it is derived with no hypothesis towards the Markov property of an origin sequence (see [1-6]for details). Thus, another idea to figure out a lost string is to distinguish "inevitably lost" strings from "unexpectedly lost" ones.

Indeed, the formula (5) allows to estimate the expected frequency of (/-tipple from the real frequencies of shorter strings, in particular, from the frequency dictionary W—I. Here two options could be found:

(I) for a given lost string $u>$ of the length q, the expected frequency is

(II) for a given lost string $u>$ of the length q, the expected frequency

Obviously, the option # I holds true always, wherever one seeks for the shortest lost string. Apparently, by definition of the shortest lost string, if you get the shortest lost string at the length q, then the support of the words of the length $q - 1$ is *-support. On the contrary, the option # II may not be met for the *-support bearing the words of the length $q - 1$. Thus, one should distinguish the longer lost strings (say, a symbol or two longer than the shortest one): the biological sense of a string exhibiting the option # II may differ from that one, for a string with non-aero expected frequency.

5 Conclusion

The strings of the minimal length that are not present in a genetic entity may be as informative, as those found in that latter. Preliminary results of the study of such strings show that the sets of the shortest lost strings may not be simulated with any probabilistic, or combinatorial model of a real DNA sequence. The length of the shortest lost strings grows very slowly, as the length of a genetic sequence grows up. Next, phylogenetically close sequences yield rather proximal sets of strings. Finally, the strings comprising a set of the shortest lost strings for a given DNA sequence seem to be rather close each other, in terms of Hamming metrics, or in terms of alignment. Further studies are necessary to figure out the biological charge of such lost strings.

Acknowledgement. This study was supported by a research grant # 14.Y26.31.0004 from the Government of the Russian Federation.

References

1. Bugaenko, N.N., Gorban, A.N., Sadovsky, M.G.: Maximum entropy method in analysis of genetic text and measurement of its information content. Open Syst. Inf. Dyn. **5,** 265-278 (1998)
2. Gorban, A.N., Popova, T.G., Sadovsky, M.G., Wunsch, D.C.: Information content of the frequency dictionaries, reconstruction, transformation, classification of dictionaries, genetic texts. In: Intelligent Engineering Systems through Artificial Neural Networks. Smart Engineering System Design, vol. 11, pp. 657-663. ASME Press, New York (2001)
3. Gorban, A.N., Popova, T.G., Sadovsky, M.G.: Classification of symbol sequences over thier frequency dictionaries: towards the connection between structure and natural taxonomy. Open Syst. Inf. Dyn. **7,** 1-17 (2000)
4. Sadovsky, M.G., Shchepanovsky, A.S., Putintzeva, Y.A.: Genes, information and sense: complexity and knowledge retrieval. Theory Biosci. **127,** 69-78 (2008)
5. Sadovsky, M.G.: Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae. J. Biol. Phys. **29,** 23-38 (2003)
6. Sadovsky, M.G.: Information capacity of nucleotide sequences and its applications. Bull. Math. Biol. **68,** 156-178 (2006)
7. Garcia S.P., Pinho A.J.: Minimal absent words in four human genome assemblies. PLoS One **6**(12), e29344 (2011)
8. Alileche, A., Goswami, J., Bourland, W., Davis, M., Hampikian, G.: Nullomer derived anticancer peptides (NulloPs): differential lethal effects on normal and cancer cells in vitro. Peptides **38,** 302-311 (2012)
9. Acquisti, C., Poste, G., Curtiss, D., Kumar, S.: Nullomers: really a matter of natural selection? PLoS One **10,** e1022 (2007)
10. Aurell, E., Innocenti, N., Zhou, H.-J.: The Bulk and The Tail of Minimal Absent Words in Genome Sequences (2015). arXiv:1509.05188v1

11. Rahman, M.S., Alatabbi, A., Athar, T., Crochemore, M., Rahman, M.S.: Absent words and the (dis)similarity analysis of DNA sequences: an experimental study. BMC Res. Notes **9,** 186 (2016)
12. Garcia, S.P., Pinho, A.P., Rodrigues, J., Bastos, C.A.C., Ferreira, P.: Minimal absent words in prokaryotic, eukaryotic genomes. PLoS One **6**(1), e16065 (2011)
13. Hao, B., Xie, H., Zuguo, Y., Chen, G.: Avoided strings in bacterial complete genomes and a related combinatorial problem. Ann. Comb. **4,** 247-255 (2000)
14. Chairungsee, S., Crochemore, M.: Using minimal absent words to build phylogeny. Theoret. Comput. Sci. **450,** 109-116 (2012)
15. Gelfand, M.S., Koonin, E.V.: Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. Nucleic Acids Res. **25,** 2430-2439 (1997)
16. Fuglsang, A.: Distribution of potential type II restriction sites (palindromes) in prokaryotes. Biochem. Biophys. Res. Commun. **310**(2), 280-285 (2003)
17. Roberts, R.J., Vincze, T., Posfai, J., Macelis, D.: REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. **43,** D298-D299 (2015)