Statistical Properties of Short Subsequences in Microbial Genomes and Their Link to Pathogen Identification and Evolution

Meizhuo Zhang[1], Catherine Putonti[1,2], Sergei Chumakov[3], Adhish Gupta[1], George E. Fox[2], Dan Graur[2] and Yuriy Fofanov[1,2]

[1]*Department of Computer Science, University of Houston, Houston, TX 77204-3058, USA*
[2]*Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5001, USA*
[3]*Department of Physics, University of Guadalajara, Guadalajara, Jalisco 44420, Mexico*

Abstract. Numerous sequencing projects have unveiled partial and full microbial genomes. The data produced far exceeds one person's analytical capabilities and thus requires the power of computing. A significant amount of work has focused on the diversity of statistical characteristics along microbial genomic sequences, e.g. codon bias, G+C content, the frequencies of short subsequences (*n*-mers), etc. Based upon the results of these studies, two observations were made: (1) there exists a correlation between regions of unusual statistical properties, e.g. difference in codon bias, etc., from the rest of the genomic sequence, and evolutionary significant regions, e.g. regions of horizontal gene transfer; and (2) because no two microbial genomes look statistically identical, statistical properties can be used to distinguish between genomic sequences. Recently, we conducted extensive analysis on the presence/absence of *n*-mers for many microbial genomes as well as several viral and eukaryotic genomes. This analysis revealed that the presence of *n*-mers in all genomes considered (in the range of *n*, when the condition $M<<4^n$ holds, where *M* is the genome length) can be treated as a nearly random and independent process. Thus we hypothesize that one may use relatively small sets of randomly picked *n*-mers for differentiating between different microorganisms. Recently, we analyzed the frequency of appearance of all 8- to 12-mers present in each of the 200+ publicly available microbial genomes. For nearly all of the genomes under consideration, we observed that some *n*-mers are present much more frequently than expected: from 50 to over a thousand copies. Upon closer inspection of these sequences, we found several cases in which an overrepresented *n*-mer exhibits a bias towards being located in the coding or being located in the non-coding region. Although the evolutionary reason for the conservation of such sequences remains unclear, in some cases it is plausible to believe that sequences having a clear bias for non-coding regions may be because of their role in the DNA uptake/recombination process, being parts in insertion sequences, or serving as transcription factors recognition sites. Our analysis of the frequency of appearance of 6-mers for each microbial genome revealed regions that display unusual statistical properties with respect to their own genome. After inspection of the genes contained within these regions, we believe that such regions are likely to have been acquired into the genomic sequence through horizontal gene transfer.

Keywords: pathogen identification, statistical properties, short subsequences
PACS: 87.23.–n

INTRODUCTION

Statistical analysis of the appearance of short subsequences, or *n*-mers, in different DNA sequences, from individual genes to full genomes is important for evolutionary

CP854, *Medical Physics: Ninth Mexican Symposium on Medical Physics,*
edited by H. Mercado-Uribe, G. García-Torales, J. L. Flores-Núñez, and E. Castillo-Corona

studies as well as the development of assays for organism identification and detection. Several methods in microbial identification (1-3) have been developed based upon frequency distribution of *n*-mers (2-4-mers (4-8) and 8-9-mers (2, 4)) in partial or complete genomic sequences. Furthermore, the analysis of compositional peculiarities, *e.g.* G+C content variation (9-12), extremes of codon bias (13-17), anomalies of amino acid usage (18), and complex compositional patterns (Markov models) (19) has also been employed to identify regions of horizontal gene transfer (HGT) including the identification of genomic and pathogenicity islands.

We have recently conducted a study of the ***frequency of presence*** (considering only the presence/absence of all possible *n*-mers) and the ***frequency of appearance*** (the number of times in which each *n*-mer occurs within the sequence) of *n*-mers within many microbial as well as viral and eukaryotic genomic sequences. While previous studies have focused primarily on the frequency of appearance for very short *n*-mers, $n \leq 9$, our analysis considers all *n*-mers up to 20 nucleotides in length. Based upon the results of our study, we suggest two methods for the identification and classification of organisms.

THE FREQUENCY OF THE PRESENCE OF *N*-MERS

To determine the significance of the presence/absence of *n*-mers within a genome, we calculated the frequency of the presence of all *n*-mers, $5<n<20$, for the complete genomes (both original and complementary strands) of 110 microbial, 1405 viral and 5 multicellular organisms, with sizes ranging from 0.44 kb to 2.87 Gb. The time/memory usage in a brute force algorithm increases exponentially as the size of *n*-mers increases. Thus, we have created a set of novel algorithms and data structures to efficiently perform such computations without relying on heuristic measures (20-21).

Assuming equal probabilities of appearance of every nucleotide, we estimate that the frequency of the presence of *n*-mers, f_0, in a sequence of length M to be

$$f_0 = 1 - \exp\left(-\frac{1}{x}\right), x = \frac{4^n}{M}, \qquad (1)$$

where x is the ratio of the total number of possible *n*-mers to the number of *n*-mers in the sequence, $M\text{-}n+1 \cong M$. Monte Carlo simulations using random sequences of equal distributions of A, T, C and G were found to be in full agreement with this estimation. Our analysis showed that the presence of *n*-mers in all genomes considered (for n such that $M<<4^n$ holds) can be treated as a nearly random process. The relative deviation of the number of presence of *n*-mers from the estimation suggests that larger genomes are more ordered than shorter ones, or they contain more repetitions of the genetic code.

If the presence/absence of *n*-mers is a random process, one could hypothesize that a random set of *n*-mers could be used, e.g. as microarray probes, to fingerprint and identify an organism. In order for this technique to be useful, it must distinguish between closely related organisms. Thus, we next considered how independent or correlated the appearances of *n*-mers are in different genomes. One way to approach this question is by using the well-known multiplication property for the joint probability of the intersection of events, according to which, wo events A, and B can be treated as independent if $p(A \cap B) = p(A)p(B)$. To estimate the probability of finding randomly picked *n*-mers in each genome, we have calculated the frequency of presence of *n*-mers in each genome as

well as the number of *n*-mers that appear in each pair of species genomes. Based on this we can compare the probabilities of finding randomly picked *n*-mers in each pair of genomes with probabilities calculated using the multiplication rule. The actual and calculated probabilities do not differ greatly from each other, allowing us to treat the presence/absence of randomly picked *n*-mers as independent events. We were especially interested in the range of *n* which gives rise to the frequency of presence of different *n*-mers in the genome between 5% and 50% of the total number of possible *n*-mers. This range varies with genome sizes. The value *n*=12 seems to be the most reasonable one for all microbial genomes. For viral genomes the value was found to be *n*=7. For organisms that are not close relatives of each other, the presence/absence of different 7- to 20-mers in their genomes are not correlated. For close biological relatives, some correlation of the presence *n*-mers in this range appears, but is not as strong as expected.

Because the presence of *n*-mers within known genomic sequences can be treated as a nearly random and independent process and we expect similar behavior from emerging and not yet sequenced genomes. One may use relatively small sets of randomly picked *n*-mers for differentiating between different viruses and organisms. Let us illustrate the idea through an example for two genomes. Let us randomly pick L 12-mers. Given a genome G_1 with the frequency of presence of *n*-mers p_1, we expect that $K = p_1 L$ *n*-mers present in G_1 will also appear in our random set, forming a "fingerprint" of G_1. The probability, ε, that the fingerprint of G_1 will exactly coincide with the fingerprint of some other genome G_2 (with the frequency of presence of *n*-mers p_2) is $\varepsilon = (1 - p_1 - p_2 + 2p_{12})^L$. Here p_{12} is the probability for the *n*-mer to be present in both genomes simultaneously. Given a desirable probability of error, ε^*, one can determine the appropriate size, L, of a random set of *n*-mers which can be used for reliable identification of genomes as:

$$L = \frac{\log \varepsilon^*}{\log(1 - p_1 - p_2 + 2p_{12})} \qquad (2)$$

For related organisms, let us assume that two genomes G_1 and G_2 almost coincide and differ only in m randomly located nucleotides. This situation simulates the existence of single nucleotide polymorphisms (SNPs). The value of L, necessary to distinguish the fingerprints of these two genomes with the error probability ε^*, can be estimated by:

$$L = \frac{\log \varepsilon^*}{p \log(1 - mn/N)} \leq \frac{M \,|\log \varepsilon^*|}{pmn} \qquad (3)$$

Therefore, we can use practically any sufficiently random subset of *n*-mers of an appropriate size as probe of a microarray assay to identify and distinguish between organisms. Different sizes of *n*-mers must be employed for different organisms based on their genome length. We would like to stress the logarithmic dependence of the sampling or microarray size on the error probability. The important advantage of this approach is that it can be used without *a priori* knowledge of the sequence itself.

THE FREQUENCY OF THE APPEARANCE OF *N*-MERS

Regions of Unusual Compositional Properties or RUCPS

The frequency of the appearance of *n*-mers has been applied to the identification of regions of compositional peculiarities, most commonly studies of di- and tri-nucleotides

(7-8). We have also observed correlation between regions of unusual compositional properties with the frequency of appearance of longer *n*-mers in genomic sequences. To assess the degree and pattern of similarity (or dissimilarity) between two genomic sequences of size M_1 and M_2, we divide the genomes into windows of length *w* and slide these windows along each genome with steps (the distance between the start of two neighboring windows) of size *s*. As a measure of similarity, we use the Pearson correlation coefficient between the frequencies of *n*-mers. The distribution $P(S)$ of appearances of all possible *n*-mers inside a given window is $P(S) = N_S/(w-n+1)$, where N_S and $w - n + 1 \approx w$ are, correspondingly, the number of appearances of *n*-mer *S* and the total number of *n*-mers in a window. To collect representative statistics, one has to impose the condition $w>4^n$. An application, Similarity Plot or S-plot, has been implemented to perform such comparisons between genomic sequences (20).

By first comparing a genome against itself, the degree homogeneity of a genome (a_g) can be determined as the average correlation value of all window distribution comparisons. The degree of similarity of each window with respect to its own genome (a_w) can also be calculated. Through Monte Carlo simulations, the values of a_w were found to follow a normal distribution. Thus, very few, if any, windows are expected to have an a_w value smaller than or greater than two or three standard deviations from the genomic mean, a_g. Windows that are unusually dissimilar to the rest of the genome into which they are embedded ($a_w \leq a_g - 2\sigma_g$) are of particular interest. We refer to such windows as "regions of unusual compositional properties" or RUCPs.

Identifying Horizontally Transferred Genes

A two-step procedure has been developed to identify regions that may have originated through HGT. First RUCPs are identified for a particular genome. Next, we compare the distributions of the frequency of appearance of *n*-mers in two closely related genomes. Those RUCPs that appear in both genomic sequences may or may not have been introduced through HGT. A RUCP in one sequence that does not have a corresponding RUCP in the other genome must have either been introduced through HGT into the first genome or precisely excised from the close relative after the divergence of the two species. Moreover, by comparing these suspect regions to one another, we can estimate a minimum number of sources for these putative xenologous sequences. Recently we conducted an analysis of the non-pathogenic *Escherichia coli* K12 and the closely related pathogenic *E. coli* O157:H7 (20). As a result of this study, we predict that there are at least 53 sources from which *E. coli* O157 acquired DNA through HGT.

Identifying GEIs/PAIs

Analysis of the four *Neisserial* genomes (*N. meningitidis* serogroup A, *N. meningitidis* serogroup B, *N. meningitidis* serogroup C, and *N. gonorrhoeae*) was performed looking at the distribution of the frequency of presence of all 6-mers. RUCPs were identified within each of the genomic sequences. These RUCPs contained all but two of the recently identified 24 GEIs/PAIs occurring in one, two, three or all four of the *Neisserial* species genomes. Furthermore, an additional 50 RUCPs were identified as putative GEIs/PAIs. Additional analysis suggests that these regions are members of up to 18 new

PAIs and 7 new GEIs. The presence of identical/similar RUCPs in the identical/similar chromosomal locations among the *Neisserial* species suggests that some of GEIs/PAIs may have been acquired before species divergence (Putonti et al., unpublished results).

Overrepresented Sequences

Because the appearance of *n*-mers in genomes can be treated as a nearly random process, we can estimate the number of times each *n*-mer is expected to appear for a given genome size, *M*, and *n*-mer size as $M/4^n$. To compare the expected frequency of appearance with the actual frequency of appearance, we calculated the number of times each *n*-mer sequence, $8 \leq n \leq 12$, occurs within 200+ microbial genomes. Overall, we found that the expected frequency of appearance of *n*-mers corresponds to the observed frequency of appearance. There are, however, a handful of *n*-mers which appear much less frequently than is expected ('underrepresented') and a handful of *n*-mers which appear much more frequently than is expected ('overrepresented').

For all 200+ microbial genomes, we have identified the locations of all overrepresented sequences and classified each occurrence as appearing in the coding or non-coding region. In essentially all of the 200+ genomes considered, we observed that many of the overrepresented *n*-mers showed a bias in location, preferring either the coding or non-coding region. This suggests a functional reason for the overrepresentation of particular sequences. Our hypotheses is supported by the fact that included in our sets of overrepresented sequences is functionally relevant chi sequences which have previously been identified in the literature. For example, the sequence "gctggtgg" appears over 1000 times within the four *E. coli* strains (21) and more than 90% of these occurrences are in the coding region (a typical characteristic of chi sequences). Our calculations also revealed that some sequences appear much more frequently in all genomic sequences of a taxonomical family. For example, the *Yersinia* family has 116 9-mers that appear more than 200 times in all four member genomes.

Overrepresented sequences can also be utilized for pathogen identification by employing species specific sequences as PCR primers. Pairs of overrepresented sequences known to occur in variable regions, e.g. PAIs, GEIs, drug-resistance genes, could also be used as PCR primers for the classification of strains without necessitating sequencing. The evolution of strains can also be inferred through the use of overrepresented sequences as PCR primers. An overrepresented sequence or set of overrepresented sequences which are equally distributed across the genomic sequence can be used to monitor genome rearrangement. In the event that genomic rearrangement has occurred, the amplicons produced using the primers will be of different lengths.

CONCLUSION

Based upon the results of our analyses, we present two methods for the identification of organisms which can be applied to newly emerging strains and organisms for which genomic sequence data is not yet available. Because the presence of *n*-mers in genomes can be treated as a nearly random and independent process, we believe that random sets of *n*-mers (with appropriately chosen *n*) can be effectively used to identify and distinguish between different viral, microbial and eukaryotic genomes. We also believe

that organism identification is possible by taking advantage of the fact that certain *n*-mer sequences appear to be overrepresented or underrepresented within a genomic sequence or taxonomical family of sequences. Our analyses also revealed that the frequency of the presence of *n*-mers and the frequency of the appearance of *n*-mers within genomic sequences provides insight into the evolution of strains and the identification of regions of functional and evolutionary significance.

REFERENCES

1. A. Campbell, J. Mrazek and S. Karlin, "Genome Signature Comparisons among Prokaryote, Plasmid, and Mitochondrial DNA", *Proc. Natl Acad. Sci., USA*, 1999, 96, pp. 9184–89.
2. R. Sandberg, G. Winberg, C.I. Branden, A. Kaske, I. Ernberg and J. Coster, "Capturing Whole-Genome Characteristics in Short Sequences using a Naive Bayesian Classifier", *Genome Res.*, 2001, 11, pp. 1404–09.
3. C. Putonti, C. Belapurkar, S. Chumakov, R. Mitra, G.E. Fox, R.C. Willson and Y. Fofanov, "Human-Blind Probes and Primers for Dengue Virus Identification: Exhaustive Analysis of Subsequences Present in the Human and 83 Dengue Genome Sequences", *FEBS*, 2006: pp. 398-408.
4. P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil, "Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences", *Mol Biol Evol*, 1999, 16:1391-99.
5. S. Karlin and I. Ladunga, "Comparisons of Eukaryotic Genomic Sequences", *Proc. Natl Acad. Sci., USA*, 1994, 91:12832–36.
6. S. Karlin, J. Mrazek and A.M. Campbell, "Compositional Biases of Bacterial Genomes and Evolutionary Implications", *J. Bacteriol.*, 1997, 179:3899–913.
7. H. Nakashima, K. Nishikawa and T. Ooi, "Differences in Dinucleotide Frequencies of Human, Yeast, and *Escherichia coli* Genes", *DNA Res.*, 1997, 4:185–92.
8. H. Nakashima, M. Ota, K. Nishikawa and T. Ooi, "Genes from Nine Genomes are Separated into Their Organisms in the Dinucleotide Composition Space", *DNA Res.*, 1998, 5:251–59.
9. J.G. Lawrence and H. Ochman, "Amelioration of Bacterial Genomes: Rates of Change and Exchange", *J Mol Evol*, 1997 Apr; 44(4):383-97.
10. J.G. Lawrence and J.R. Roth, in *Horizontal Transfer*, eds. M. Syvanen, & C.I. Kado, Chapman and Hall, London, 1998, pp. 208-25.
11. P. Lio and M. Vannucci, "Finding Pathogenicity Islands and Gene Transfer Events in Genome Data", *Bioinformatics*, 2000, 16:932-40.
12. N. Sueoka, "Two Aspects of DNA Base Composition: G+C Content and Translation-coupled Deviation from Intra-strand rule of A = T and G = C", *J. Mol. Evol*. 1999, 49, 49-62.
13. S. Karlin and J. Mrazek, "Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes", *J. Bacteriol*., 2000, 182, 5238-50.
14. S. Karlin, J. Mrazek and A.M. Campbell, "Codon Usages in Different Gene Classes of the *Escherichia coli* Genome", *Mol. Microbiol*., 1998, 29, 1341-55.
15. C. Medigue, T. Rouxel, P. Vigier, A. Henaut and A. Danchin, "Evidence for Horizontal Gene Transfer in *Escherichia coli* Speciation" *J. Mol. Biol*., 1991, 222, 851-56.
16. I. Moszer, E.P. Rocha and A. Danchin, "Codon Usage and Lateral Gene Transfer in *Bacillus subtilis*." *Curr. Opin. Microbiol*., 1999, 2, 524-28.
17. P.M. Sharp and W.-H. Li, "The Codon Adaptation Index--a Measure of Directional Sysonymous Codon Usage Bias,and Its Potential Applications", *Nucl. Acids Res*., 1987, 15, 1281-95.
18. M. Ventura, C. Canchaya, D. van Sinderen, G.F. Fitzgerald and R. Zink, "*Bifidobacterium lactis* DSM 10140: Identification of the atp (atpBEFHAGDC) Operon and Analysis of Its Genetic Structure, Characteristics, and Phylogeny", *Appl. Environ. Microbiol.*, 2004, 70, 3110-21.
19. W.S. Hayes and M. Borodovsky, "How to Interpret an Anonymous Bacterial Genome: Machine Learning Approach to Gene Identification", *Genome Res*, 1998, Nov; 8(11), pp. 1154-71.
20. C. Putonti, Y. Luo, C. Katili, S. Chumakov, G.E. Fox, D. Graur and Y. Fofanov, "A Computational Tool for the Genomic Identification of Regions of Unusual Compositional Properties and Its Utilization in the Detection of Horizontally Transferred Sequences". Submitted to *Molecular Biology and Evolution* (MBE). Under review.
21. R. Uno, Y. Nakayama, K. Arakawa and M. Tomita, "The Orientation of Chi Sequences is a General Tendency of G-rich Oligomers", *Gene*, 2000, 207-15