On the visualization of the DNA sequence and its nucleotide content

Aleksandr Mylläri, Tapio Salakoski, Alexey Pasechnik

Department of Information Technology, University of Turku
Lemminkäisenkatu 14a, FIN-20520, Turku, Finland
Piter Publishing House, B. Sampsonjevskij pr., 29a,
Sankt-Peterburg, Russia,

firstname.lastname@it.utu.fi,Alexey.Pasechnik@piter.com

Abstract

Visual inspection can help reveal patterns that would be computationally rather difficult to reveal. We consider three different algorithms for visualizations of a DNA sequence and its nucleotide content: random walk, fractal and visualization based on the entropy-like parameters calculated using a sliding window. We present a program that uses these three methods and visualizes either the whole of a given sequence, or specified fragments. It also provides facilities to compare visualizations obtained for different sequences/fragments. Random walk visualization considers the sequence symbol-by-symbol; the other two methods also take into account how well nucleotides are "mixed" in the sequence. It allows an easy visualization of repeated patterns, segments with a high/low content of some nucleotides, such as CG-islands, etc. The program also helps to identify regions of interest for further study.

1 INTRODUCTION

Usually the sequence of nucleotides in DNA is represented as a text based on a four-letter alphabet: {A, C, G, T}, where each symbol codes the corresponding nucleotide. Since the length of the DNA chain varies from 10^3 (viruses) to 10^9 (higher organisms) it is a complicated cognitive problem for human perception to establish similarity/dissimilarity between two DNA sequences. Most of the sequences are too long to be wholly displayed on the computer monitor, and it is very difficult to reveal some peculiarities or characteristics directly from the text string that represents DNA. A good way to visualize a DNA sequence would help in estimating nucleotide content, and in revealing repeated or missing subsequences, conserved regions, etc.

Visual inspection can help reveal patterns that would be computationally rather difficult to reveal. For example, in cluster analysis of 2D data (i.e. on the plane) it is usually much easier and faster to do clustering by visual inspection than by using some elaborate algorithm. Our experiments with visualizations using Mathematica and Maple have led us to develop a software program that visualizes a DNA sequence. The program visualizes either the whole of a given sequence, or specified fragments. It provides the option to save a selected fragment of the DNA sequence for further study. It also gives an opportunity to compare visualizations obtained for different sequences (or several different fragments of the same sequence). The sequence is assumed to be in FASTA format. The program is a tool for assisting researchers; it also inspires the formulation of new questions.

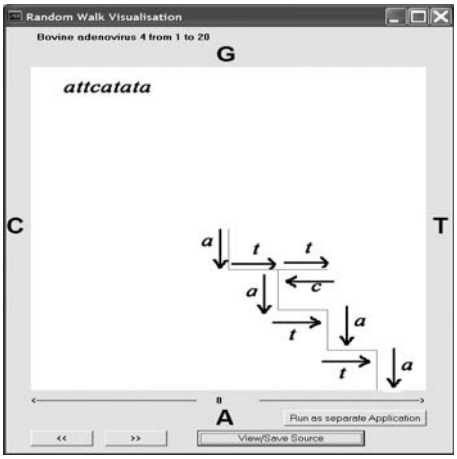

Figure 1: Random walk visualization.

2 VISUALIZATION ALGORITHMS

The program uses three different algorithms for visualizations: random walk, fractal-based, and visualization based on the entropy-like parameters calculated using a sliding window. Random walk visualization considers the sequence symbol-by-symbol, while the other two methods take into account how well nucleotides are "mixed" in the sequence and how sequences of nucleotides of different length are represented. It allows one to easily detect repeated patterns, segments with a low content of some nucleotides (or dinucleotides, triplets, etc.) Fractal-based visualization is static – it gives an idea of the distribution of words of different length (the present version considers words with length 1 - 8) for the whole considered sequence. The other two visualizations are dynamic (they give an idea of what happens along the sequence).

Different DNA sequences have different base content. One simple way to visualize this difference is as follows. Let us plot the points according to the following rule: the starting point is at the center of the chart. We read a base in the DNA sequence, and if it is 'A' we move the pen one step up, if it is 'C' we move the pen one step down, if it is 'G' we move the pen one step to the left, if it is 'T' we move the pen one step to the right. Then we read next symbol in the sequence, and so on (see Figure 1). If all bases are represented in the same proportion (approximately 25% each), the point will stay near the center; if some bases are encountered more often, the point will move from the center. This way it is easy to see which nucleotides are in abundance, which are infrequent. The program also allows one to visualize several sequences (or parts of the same sequence) simultaneously using different colors for different sequences. One such example is given in Figure 2. One can see that different organisms (here we used several DNA sequences for viruses and bacteria) produce different tracks, while homologous organisms produce similar tracks: in Figure 2, viruses produced the tracks from the center to the top-right corner, while bacteria produce the tracks from the center to the bottom-left corner, thus reflecting different content of A,T and C,G nucleotides in the sequences considered here. This visualization could be considered as a 2D random walk visualization (but different from the one used, e.g in [1].)

Another visualization algorithm is fractal-based (see, e.g. [2], [3], [4]). It allows visualization of missing (sub-)sequences in DNA. This algorithm is based on fractal addresses and is related to the game of chaos (see, e.g. [5], [6]). The work of the algorithm is illustrated in Figure 3. We split the square into four square parts corresponding to four bases. Then we split these four squares into four smaller squares, and so on. We read 8 symbols in the considered sequence and color the corresponding small square. Then we move one step forward, read next nucleotide in the sequence, color a square, and so on. In Figure 3 the square corresponding to the sequence 'GTT' is shown. If there are no triplets 'GTT' in the sequence, this

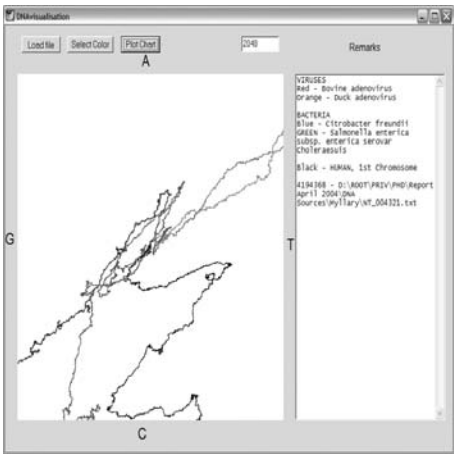

Figure 2: Random walk visualization of several sequences.

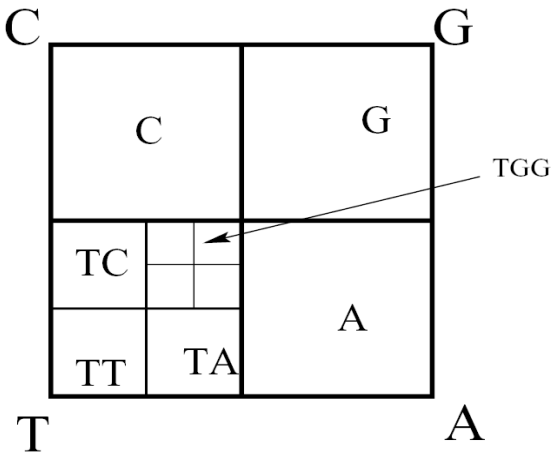

Figure 3: Fractal visualization

square will have no colored points, while if there are too many, it will be darker then others. The same is true for the squares of the "next generations" - all smaller squares - e.g. the small squares corresponding to '$GTTA$', '$GTTAC$', etc. will be also empty or have more points inside correspondingly. This way we can see how (non-)uniform is the distribution of all possible combinations of bases with length 1-8. It could be used, e.g. to reveal the presence of CG-islands that have an important biological meaning ([7], see also [8], [9].) An example of this visualization is given in Figure 4. One quick look (and a little analysis of the positions of lighter and darker squares of different sizes) reveals the different content of strings 'CG', 'CGG', etc. for two considered sequences (bacteria and part of the human chromosome). This way we have a simultaneous multilevel view on the distribution of words with length 1-8 in the DNA sequence.

The concepts of entropy and information are widely used in DNA studies (see, e.g. [10], [11], [12], [13]). For the visualization we use Shannon entropy, Markov entropy and some more entropy-like characteristics. These parameters tell us how well mixed are the bases and their chains of length 2 - 4. A uniform (well-mixed) distribution corresponds to the maxima of these characteristics, whereas minimums tell us that the distribution is far from uniform. These parameters are visualized using a sliding window: for the visualization we estimate entropies not for the whole sequence but for a fragment of a length specified by the user of the program (as a window size). These calculations are repeated with some step (one more

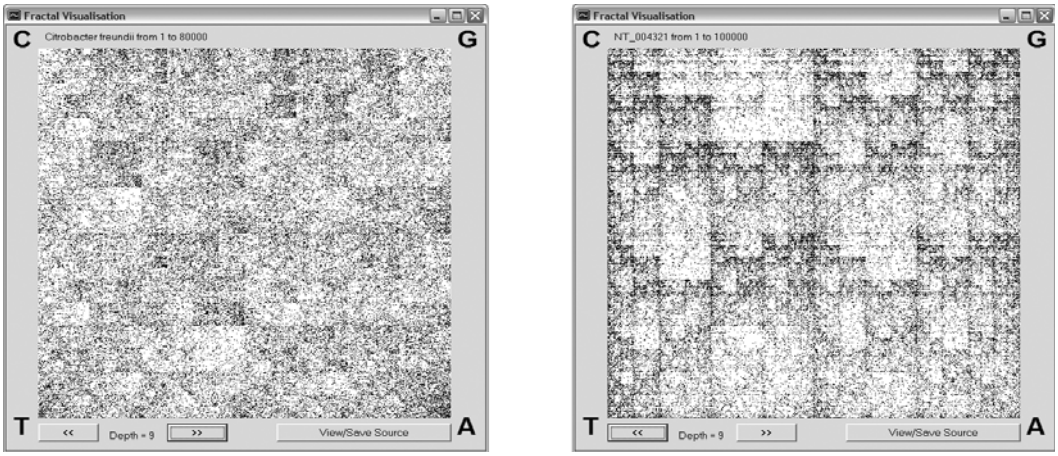

Figure 4: Example of fractal visualization: bacteria Citrobacter freundii (left) and part of the human chromosome NT 004321 (right).

parameter that can be changed), and results are plotted as four curves. It should be noticed that the window size should not be too small in order to get reliable estimates. The user can select an interesting part of the plot and re-calculate the plots for the selected fragment (and apply the other two visualizations as well) or save this fragment to a separate file for later study.

3 Conclusions

We have presented a program to visualize a DNA sequence or selected fragments of it.
More examples of the visualizations obtained with the program can be found on the poster image at `http://www.cargo.wlu.ca/casc2005/`. Many phenomena observed in these visualizations can easily be related to some biological phenomena, such as repeats or conserved regions. However, there are still several visual observations that call for biological interpretation. For example, regions where different entropies behave differently lend themselves to further study. We plan to add some more visualizations to the present set, e.g. three-dimensional trajectories ([14]) (a visualization similar to our random walk, but more informative, though at the price of the necessity of real 3D images). We also plan to add the facility to give the user the opportunity to graphically compare fractal visualizations for different sequences (different fragments of the same sequence).

4 Acknowledgements

A.M. appreciates financial support from Magnus Ehrnrooth Foundation and Turun Yliopistosäätiö.

References

[1] Salvatore Paxia, Archisman Rudra, Yi Zhou, and Bud Mishra, "A random walk down the genomes: Dna evolution in valis," *Computer*, vol. 35 (7), pp. 73–79, 2002.

[2] HJ Jeffrey, "Chaos game representation of gene structure," *Nucleic Acids Research*, vol. 18 (8), pp. 2163–2170, 1990.

[3] Bai lin Hao, H. C. Lee, and Shu yu Zhang, "Fractals related to long dna sequences and complete genomes," *Chaos, Solitons and Fractals*, vol. 11, pp. 825–836, 2000.

[4] Dan Ashlock and Jim Golden, "Evolutionary computation and fractal visualization of sequence data," in *Evolutionary Computation in Bioinformatics*, Gary B. Fogel and David W. Corne, Eds. 2003, Morgan Kaufmann Publishers.

[5] Heinz-Otto Peitgen, Hartmut Jurgens, and Dietmar Saupe, *Chaos and Fractals.New Frontiers of Science*, Springer-Verlag, New York, 1992.

[6] Michael Barnsley, *Fractals Everywhere*, Morgan Kaufmann, 2000.

[7] A. Bird, "Cpg islands as gene markers in the vertebrate nucleus," *Trends in Genetics*, vol. 3, pp. 342–347, 1987.

[8] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis. Probabilistic models of proteins and nucleic acids*, Cambridge University press, Cambridge, 2000.

[9] Pavel A. Pevzner, *Computational molecular byology*, The MIT Press, Cambridge, Massachusets, London, England, 2000.

[10] L. Gatlin, "The information content of dna," *J. Theor. Biol.*, vol. 10, pp. 281, 1966.

[11] G. W. Rowe, "On the informational content of viral dna," *J. Theor. Biol.*, vol. 101, no. 4, pp. 151, 1983.

[12] Lipman D. J. and Maizel J., "Comparative analysis of nucleic acid sequences by their general constraints," *Nucl. Acids Res.*, vol. 10, pp. 2723, 1982.

[13] Olga V. Kirillova, "Entropy concepts and dna investigations," *PLA*, vol. 273.

[14] Hsuan T. Chang, Neng-Wen Lo, Wei C. Lu, and Chung J. Kuo, "Visualization and comparison of dna sequences by use of three-dimensional trajectories," in *CRPITS '03: Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003*, Darlinghurst, Australia, Australia, 2003, pp. 81–85, Australian Computer Society, Inc.