DYNAMICAL VISUALIZATION OF THE DNA SEQUENCE AND ITS NUCLEOTIDE CONTENT

Alexey Pasechnik[1], *Aleksandr Mylläri*[2] *and Tapio Salakoski*[2]

[1] Piter Publishing House, B. Sampsonjevskij pr., 29a,
Sankt-Peterburg, Russia, Alexey.Pasechnik@piter.com
[2] Department of Information Technology, University of Turku and
Turku Centre for Computer Science TUCS,
Lemminkäisenkatu 14 A, FIN-20520 TURKU, Finland, firstname.lastname@it.utu.fi

ABSTRACT

Visual inspection can help reveal patterns that would be computationally rather difficult to reveal. We present a program to visualize a DNA sequence and its nucleotide content. The program visualizes either the whole of a given sequence, or specified fragments. It also provides facilities to compare visualizations obtained for different sequences. The program uses three different algorithms for visualizations: track-based, fractal-based and visualization based on the entropy-like parameters calculated using a sliding window. Track-based visualization considers the sequence symbol-by-symbol; the other two methods also take into account also how well nucleotides are "mixed" in the sequence. It allows an easy revealing of the repeated patterns, segments with a low content of some nucleotides, etc. Current version of the program works under Microsoft Windows.

1. INTRODUCTION

"A picture is worth a thousand words." – The same or similar idiom exists in many languages. You can describe something by drawing just one picture as well as you can by writing or saying a lot of words. Moreover, 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. One should also mention that even the most complicated pattern-matching algorithms need to know what kind of patterns to look for. So the search for patterns is inherently limited by the expressive power of the pattern representation language used, whereas visualization can help to find patterns of new kind. Finally, visualization exploits the opportunity of human perception to work on several abstraction levels simultaneously. We present a program to visualize a DNA sequence and its nucleotide content. 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 can also use a description file using ncbi human genome annotation. We would like to stress that our program is a tool for assisting the researcher, not the ultimate answer to all questions in DNA studies (number '42' doesn't appear in the code naturally [1]), moreover, it inspires formulation of new questions.

2. VISUALIZATION ALGORITHMS

The program uses three different algorithms for visualizations: track-based, fractal-based, and visualization based on the entropy-like parameters calculated using sliding window. Track-based 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 reveal 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 word's length 1 - 8) for the whole considered sequence. Other two visualizations are dynamic – they give an idea of what happens along the sequence.

2.1. Track visualization

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: starting point is in 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. 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. If in one fragment of the sequence there is a deficit of some nucleotide(s), and later

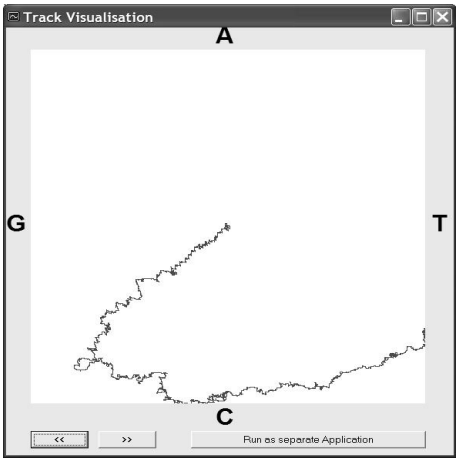

Figure 1. Example of track visualization.

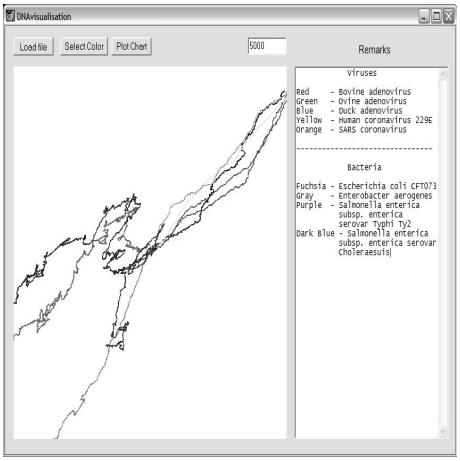

Figure 2. Visualization of several sequences.

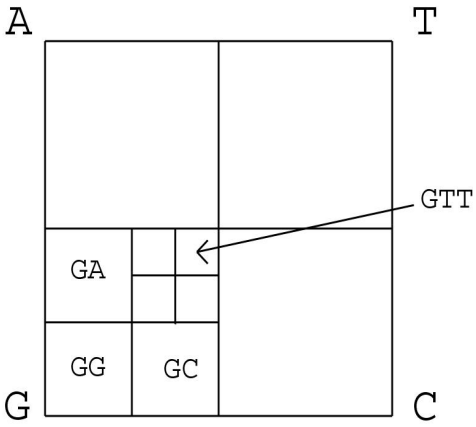

Figure 3. Fractal visualization

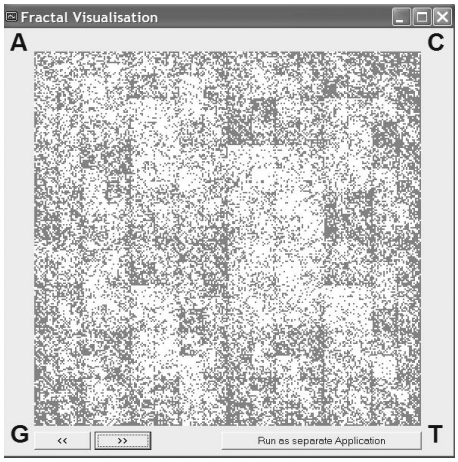

Figure 4. Example of fractal visualization.

an abundance, it will be easily seen on the image. One example of this visualization is presented on Figure 1. It can be seen that in the beginning there is a lack of *T* and *A* nucleotides, later *T* and *A* start to prevail, and there is much less of *G*. The program also allows to use this visualization module as an independent application, in this case it is possible to visualize several sequences (or parts of the same sequence) simultaneously using different colors for different sequences. One such example is given on 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: on Figure 2 viruses produce the tracks from the center to the top-right corner, the 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. Track visualization could be considered as a 2D random walk visualization (but different from the one used, e.g in [2].)

2.2. Fractal-based visualization

Another visualization algorithm is fractal-based (see, e.g. [3], [4], [5]). It allows to visualize missing sequences in DNA. This algorithm is based on fractal addresses and is related to the game of chaos (see, e.g. [6], [7]). 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 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 square will have no colored points, if there are too many - it will be darker then others. Same is true for the squares of the "next generations" - all smaller squares - e.g. 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 ([8], see also [9], [10].) 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 that there is a lack of

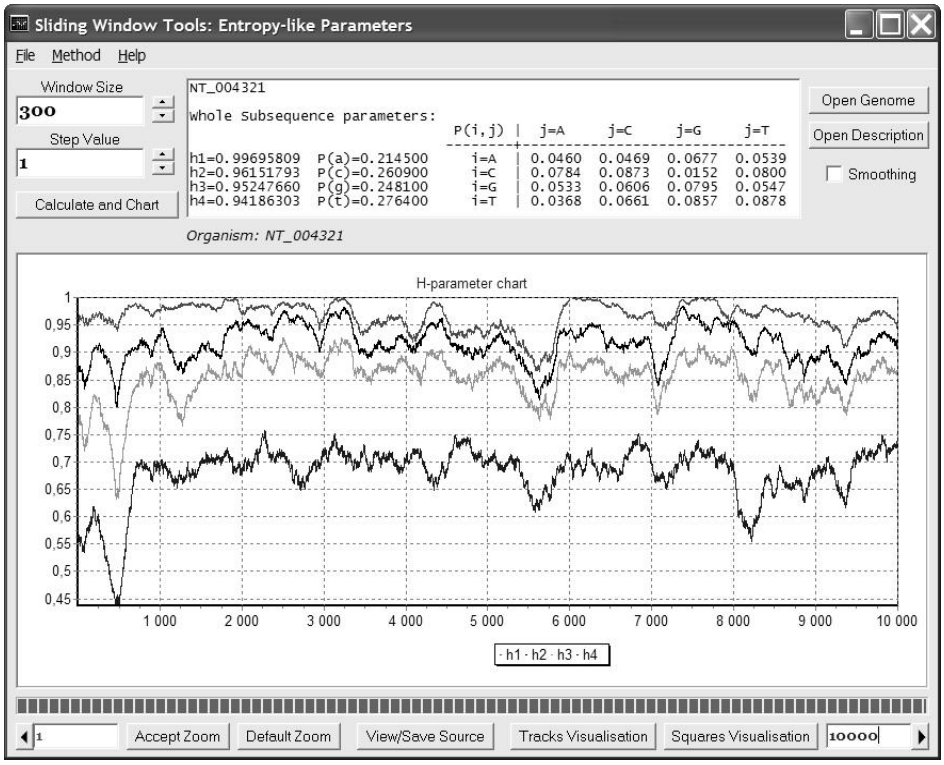

Figure 5. Sliding window entropy visualization.

strings '*CG*', '*CGT*', '*CGTA*', '*CGAT*', etc. This way we have simultaneous multilevel view on the distribution of words of length 1-8 in the DNA sequence.

2.3. Sliding window entropy visualization

Concepts of entropy and information are widely used in DNA studies (see, e.g. [11], [12], [13], [14]). First concept of entropy was proposed by C. Shannon in application to the theory of information transmission [15]. Entropy in that context implies the measure of heterogeneity of a set of symbols. It is defined as:

$$H = -\sum_i p_i \log p_i, \tag{1}$$

where p_i is the probability of appearance of the i-th symbol. This is the classic Shannon entropy. DNA is represented as a long sequence of symbols of the alphabet consisting of four letters, but the analysis of letter groups - so called words - seems also interesting. Shannon's formula becomes

$$H_n = -\sum_{i=1}^{a^n} p(C_i) \log p(C_i), \tag{2}$$

where C_i is a block of symbols of a length n; a is the number of letters in the alphabet (in the case of DNA $a = 4$), a^n is the number of all possible combinations of length n. For the visualization we use Shannon entropy

$$h_1 = -\sum_i p_i \log p_i, \tag{3}$$

and Markov entropy

$$h_2 = -\sum_i \sum_j p_i q_{ij} \log q_{ij}, \tag{4}$$

where

$$p_i = \frac{N_i}{N}, \quad q_{ij} = \frac{N_{ij}}{N_i}.$$

(3) and (4) are sometimes called Markov entropy of the zero and first order respectively. We also consider two more characteristics:

$$h_3 = -\sum_i \sum_j \sum_k p_i q_{ij} r_{ijk} \log r_{ijk} \tag{5}$$

and

$$h_4 = -\sum_i \sum_j \sum_k \sum_l p_i q_{ij} r_{ijk} s_{ijkl} \log s_{ijkl} \tag{6}$$

where

$$r_{ijk} = \frac{N_{ijk}}{N_{ij}}, \quad s_{ijkl} = \frac{N_{ijkl}}{N_{ijk}}.$$

Parameters h_1 - h_4 tell us how well mixed are the bases and their chains of length 2 - 4. Uniform (well-mixed) distribution corresponds to the maxima of these characteristics, minimums tell us that the distribution is far from uniform. These four parameters are visualized using sliding window: for the visualization we estimate h_1 - h_4 not for the whole sequence (values for the whole

considered fragment/sequence are given in the information window in the center, see Figure 5), but for a fragment of a length specified by user of the program (as a window size). These calculations are repeated with some step (one more 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 to get reliable estimates for h_2 - h_4. 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. As can be seen in Figure 5, the behavior of the four curves is usually correlated. According to our studies, the clear minima typically correspond to biological repeat regions while the more subtle changes are more difficult to interpret and might bear some relationship to e.g. regulatory regions or other areas where certain patterns occur frequently. Moreover, there are places with different behavior of the four parameters (see, e.g. the region around 4000 - 4700 in Figure 5), which might be interesting to study further.

3. USING THE PROGRAM

The main (starting) module for the program is called DNA-Tools.exe. When it is started, a window similar to the one on Figure 5 appears. The user can open genome (upper button on the right) and start visualizations using default values for the parameters, or choose own values for the parameters (start and end points of the considered fragment; window size and step for the sliding window visualization). There is also an opportunity to smoothen the curves for parameters h_1 - h_4 by checking the corresponding box, but it slows down the calculations. The user can select an interesting fragment on the h_1 - h_4 plots using the mouse click-and-drag, and the selected fragment will be zoomed. Double click returns back to the old scale, or the user can accept zoom by clicking on the corresponding button. In this case, the coordinates for the fragment under study will be changed. It is possible to save selected fragment of the sequence to a separate file for later studies.

4. CONCLUSION

We have described the program to visualize DNA sequence or selected fragments of it. Many phenomena observed on 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 parameters h_1 - h_2 and h_3 - h_4 behave differently lend themselves to further study.

The present version of the program works under Microsoft Windows, but in nearby future we plan to have a LINUX version. The program was made using Borland Delphi 6 Personal Edition, several procedures were done using the free Borland C++ Compiler. Executable files can be obtained by e-mail request to authors. Authors will be thankful for comments and suggestions.

5. REFERENCES

[1] Douglas Adams, *The Hitch Hiker's Guide to the Galaxy*, Heinemann, London, 1996.

[2] 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.

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

[4] 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.

[5] 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.

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

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

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

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

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

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

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

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

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

[15] Shannon C. E., "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379, 1948.