Image and Fractal Information Processing for Large-Scale Chemoinformatics, Genomics Analyses and Pattern Discovery

Ilkka Havukkala, Lubica Benuskova, Shaoning Pang, Vishal Jain, Rene Kroon, and Nikola Kasabov

Knowledge Engineering and Discovery Research Institute,
Auckland University of Technology Auckland,
New Zealand
ilkka.havukkala@aut.ac.nz
www.kedri.info

Abstract. Two promising approaches for handling large-scale biodata are presented and illustrated in several new contexts: molecular structure bitmap image processing for chemoinformatics, and fractal visualization methods for genome analyses. It is suggested that two-dimensional structure databases of bioactive molecules (*e.g.* proteins, drugs, folded RNAs), transformed to bitmap image databases, can be analysed by a variety of image processing methods, with an example of human microRNA folded 2D structures processed by Gabor filter. Another compact and efficient visualization method is comparison of huge amounts of genomic and proteomic data through fractal representation, with an example of analyzing oligomer frequencies in a bacterial phytoplasma genome. Bitmap visualization of bioinformatics data seems promising for complex parallel pattern discovery and large-scale genome comparisons, as powerful modern image processing methods can be applied to the 2D images.

1 Introduction

Massive amounts of information keep accumulating into many complex chemical structure databases, including protein and RNA structures, drug molecules, drug-ligand databases, and so on. Surprisingly, there is no commonly accepted standard for recording and managing chemical structure data, *e.g.* drug molecules, suitable for automated data mining [1]. Also genomic data are accumulating at increasing speed, with almost 2,000 microbial and eukaryotic genomes listed in the Genomes OnLine Database (GOLD), either completed or being sequenced [2]. Increasing interest is now being focused on characterizing various genomes, especially for their repetitive DNA and repeated DNA motifs, especially in the non-coding regions, important for chromatin condensation and gene regulation [3].

We present and illustrate two promising approaches to handle large-scale chemoinformatics and genomics data, based on visualization as bitmaps and applicable to standardized pattern analysis and knowledge discovery.

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 163–173, 2006.

2 Protein, RNA and Other Chemoinformatics Databases

2.1 Current Analysis Methods

There are currently some 35,000 databased protein structures (X-ray and NMR) in the Protein Data Bank PDB [4], and many more structures have been estimated by computational comparison of amino acid sequences to secondary and tertiary structures, either by *ab initio* folding programs or supervised methods involving sequence threading to a known protein structure. A large number of web servers are available on the internet to compare protein structures with each other, see *e.g.* [5]. The underlying structural alignment algorithms are crucial for drug design, *e.g.* ligand to protein binding simulation. However, these algorithms currently cannot handle simultaneous comparison and classification of large numbers of structures, except by brute force, using very large distributed computing infrastructures, like FightAIDS@Home on the World Community Grid, which performs AutoDock analysis of drug and HIV virus target matching on thousands of PCs around the world [6]. However, currently there is no efficient solution for matching, clustering and classifying large numbers of molecular structures efficiently.

Amino acid sequence similarity has been used as a proxy to compare similar protein structures, but a minimum of 30% sequence identity and a known structure is needed for modelling protein structures. For accurate drug design, up to 60% sequence identity is needed to ensure proper ligand binding models. Also, in this respect the current set of protein structures do not yet cover sufficiently the natural protein structure space [7]. In addition, protein structure is known to be clearly more conserved than sequence similarity.

Similarly to proteins, the folding of the RNA molecules is also known to be often more conserved than their sequence, and most recent estimates suggest that the number of non-coding genes with stable 2D RNA structures of transcripts is in the thousands [8], and may match the total number of protein coding genes in eukaryotic genomes.

There is thus a need for new efficient methods for comparing and clustering of large numbers of macromolecule structures that could avoid the use of complicated and detailed data structures pertaining to the 3D atomic coordinates of proteins, RNAs, and organic molecules. Such an alternative approach advocated in this paper is to generate 2D projections of molecular structures, transform the data into bitmap images and then analyze the bitmap images using a variety of advanced methods developed in the artificial intelligence community for face recognition, fingerprint classification and so on. An example of using this approach for RNA structures is described below.

2.2 Bitmap Image Processing Approach to Clustering and Classification of Folded RNAs

RNA molecules commonly self-assemble, resulting in more or less stable specific conformations in which nucleotide pairs A–U and C–G are formed for a reduced

free energy level. The conformations are characteristic of the different RNAs, *e.g.* eukaryotic ribosomal RNAs, microbial riboswitches, human microRNAs and so on. With the latest algorithms, secondary 2D structures can be computed quite fast and reliably from RNA sequence [9] Normally only the most stable structure with the lowest thermodynamic energy ($\Delta\mathbf{G}$) is considered, but there can also be several other more or less likely conformations, collectively known as the Boltzmann ensemble, which can nowadays also be computed with reasonable accuracy [10]. Ideally, these alternative conformations should be taken into account in comparative analysis of different RNAs.

Consensus structure comparisons for a set of RNA sequences have been previously made in three basic ways: 1) multiple alignment of sequences, followed by structure folding of the consensus, 2) Sankoff method of simultaneously aligning sequences and folds and 3), folding sequences to structures, followed by structural alignment, as reviewed in [11].

The first method may not cluster together all related sequences, as RNA structure is more conserved than its sequence. With the Sankoff method it is not easy to cluster large numbers of sequences/structures and the method is also computationally very demanding for large-scale use. The third method is a novel field, and demands a very good method to align structures to start with. Several approaches have been introduced, including RNA as topological graphs or trees. Representative algorithms in this field are RNAFORESTER and MARNA, reviewed in [9] and TREEMINER [12]. Their performance in analysing and clustering very large RNA sets is not yet known.

A new generic approach proposed by us [13] for large-scale analysis of RNA structures consists of first computing the 2D structures for the set of RNA sequences, followed by transformation of the structures into bitmap images and analysis of the image set with a suitable image processing algorithm (Fig. 1).

2.3 Example of Human MicroRNAs Analysed by Gabor Filter Method

MicroRNAs are short, 80-150 basepairs long RNAs that do not code for protein, but fold into hairpin structures and exert their effect on gene regulation by binding to matching sequences of messenger RNAs of protein-coding genes, reviewed in [14]. They are now known from plants, mammals and many lower eukaryotes as well. In a first case study of the general bitmap image analysis approach [13], the set of 222 known human microRNAs was folded by RNAFold algorithm of the Vienna package [15] and transformed into bitmap images, which were then used to extract classificatory information using Gabor filter method. Gabor filter produces rotation-invariant features, which are used to calculate measures of similarity to compare images. Greyscale bitmaps of 512x512 pixels were used, with low-resolution spatial frequencies and four angular directions.

Fig. 2 (top middle and right) shows two examples of Gabor filter transformed bitmap images of folded RNA (top left) at low angular resolution. From the transformed images, feature vectors were obtained, and Manhattan distances between vectors of all pairs of microRNAs calculated. The heat map of all ver-

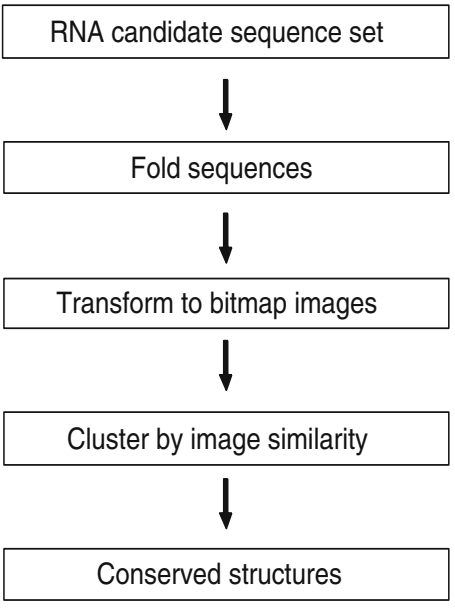

Fig. 1. Bitmap image analysis approach to RNA structure classification

sus all comparisons of the 222 microRNAs (Fig. 2, bottom) shows clearly the diagonal of similar items (the microRNAs were ordered by known microRNA families) or structural motifs together.

In the heatmap colour scaling blue pixels show the most similar microRNA pairs, and red pixels the least similar ones. In addition, many other putative similarities between microRNAs that do not share sequence similarity are also indicated for a large number of other microRNA pairs. For more details, see [13]. These additional similarities are worth exploring further, because they may correlate with specific structures in the folded RNAs. Thus the bitmap image similarity could help in sequence pattern discovery by providing additional information for clustering RNAs with weakly similar sequences.

2.4 Further Improvement of the Approach

For improving the bitmap utilization method, other ways of visualizing the 2D structure could be used, *e.g.* by using different colours or shapes for different bases or basepairs. Subsequently, various other image feature extraction methods could be used to derive informative colour/shape/contour/curvature data for clustering and classification of the microRNA structure images. The approach is a general one, applicable to all kinds of macromolecules for which an informative 2D structure representation is easily computed. This method could reveal relevant features not previously considered by chemists or biologists, or it could be used as a prefiltering step in very large databases of molecular structures. Then the challenge is to develop the image clustering methods to handle large num-

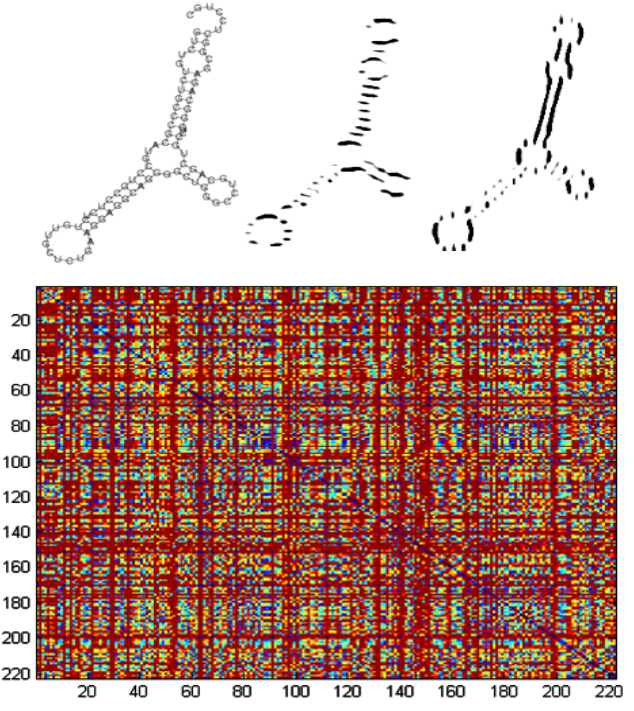

Fig. 2. Gabor filter analysis of microRNA structures. Top: left, a sample folded RNA structure; middle, Gabor filtered image at $\theta = 0$ rotation angle; right, image at $\theta = \pi/2$. Bottom: Heatmap matrix of Gabor filter feature vector Manhattan distance similarities of 222 human microRNAs. x- and y-axes: microRNA identification number, heatmap colourscale: Blue (dark): most similar, Red (light): least similar.

bers of bitmap images efficiently. Automation of the procedure involves suitable cutoffs for similarity measures for desired statistically significant clustering of the similar structures.

3 Genomics Databases

3.1 Current Analysis Methods

Similarly to the expansion of chemoinformatics related databases, genomic and proteomic data is stretching bioinformaticians to develop efficient large-scale methods for pattern identification, knowledge discovery and easily accessible and queriable databasing. Multiple alignments of many genomes (utilizing BLAST or other fast string comparisons) are already used for interspecies comparisons [16],[17], but more compact data summarization methods are needed. Analyzing whole genomes to quickly reveal their salient features and to extract new knowledge is an essential goal for biological sciences. We advocate the solution of compressing information about oligomer frequencies in long sequences into

small, coloured fractal representations in 2D or 3D space. This can achieve compression of genome data by a million times or more.

3.2 Fractal Representation Approach for DNA Sequences

Fractals in the form of iterated function system (IFS) and Chaos Game Representation have been used to visualize short DNA [18] or protein [19] sequences of genes, even complete genomes [20],[21], and in principle any symbolic sequences [22]. The iterated function system transforms DNA sequences to unique points in 2-dimensional space. The principle here is to map all oligomers of fixed size of N bases contained in the genome to a 2D space with 2Nx2N elements.

An important characteristic of the representation space is that there are so-called attractor points in the space, *e.g.* in the corners, representing subsequences AAAA, CCCC, and so on. Similar oligomers are situated spatially close to each other in the representation space. (Fig. 3 illustrates the IFS principle. Equation (1) shows the four transformations in the rectangular coordinate space in successive basepairs of the DNA, with x and y axes ranging from 0 to 1.

$$\begin{aligned} \omega_T(x,y) &= (0.5x + 0.5, 0.5y) \\ \omega_A(x,y) &= (0.5x, 0.5y + 0.5) \\ \omega_G(x,y) &= (0.5x, 0.5y) \\ \omega_C(x,y) &= (0.5x + 0.5, 0.5y + 0.5) \end{aligned} \tag{1}$$

Every transformation contracts coordinates to its quarter of a unit square. A limit set of points emerging from an infinite application of the IFS is called the IFS attractor. End positions of all the oligomers are marked on the grid, and their frequency in each cell counted, and the frequencies displayed by greyscale or colour scale. We show an example with a microbial phytoplasma genome.

3.3 Example of Phytoplasma Genome Octamers Visualized in Fractal Space

Phytoplasmas are wall-less prokaryotic microbes and obligate parasites of plants, with genome sizes below one million basepairs. They belong to Mollicutes, known to have AT-rich genomes. The Aster Yellow Witches' Broom genome has recently been sequenced, and is used here as an example of a new unexplored genome [23]. All octamer oligonucleotides of the whole genome (*ca.* 700 kilobases) were plotted in fractal space of 256x256 = 65,536 pixels, and their frequencies are shown as a colour heatmap (Fig. 4). As expected, the AT-diagonals have high frequencies of octamers, and the abundance of A-rich and T-rich sequences at opposite corners is immediately evident. This was verified by using RepeatScout algorithm [24] to calculate the most abundant non-overlapping octamers (including tandem ones) in the genome. Fig. 5 illustrates that the most abundant octamers indeed are AT-rich. What is not easy to find out from these octamer frequency listing is that there are two approximately equally abundant types of these oligomers, A-rich and T-rich, as shown by the red-orange clusters in the corners A and T, respectively.

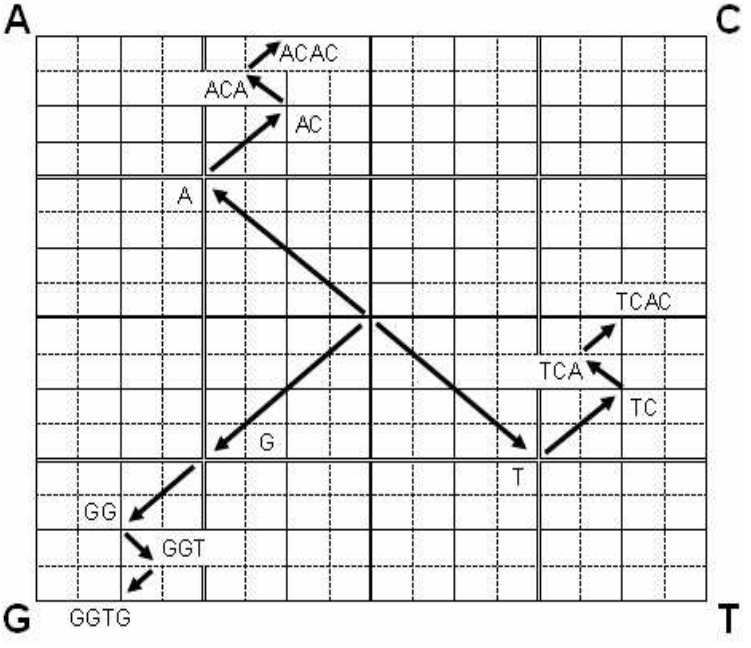

Fig. 3. Principle of mapping N-mers for a fractal space. Here all tetramer polynucleotides are mapped to unique positions in a 16 x 16 coordinate grid. Three end positions for three tetramers are shown.

The basic difference of the fractal method to counting and comparison of frequencies of tandem and interspersed repeats is also that overlapping oligomers are enumerated exhaustively. This is important in terms of RNAi and transcription factor regulating mechanisms of gene expression and chromatin remodelling., which rely on the presence of suitable binding site oligomers in any relevant genome location.

Another finding easily seen in the fractal representation is the cluster in the middle of bottom border between G and T corners, which suggests an abundance of GT-rich octamers. Such repetitive motifs might have a special function in the phytoplasma for its host relationship. Indeed, it has been suggested that repetitive DNA is important in prokaryotes for genome plasticity, especially in host-parasite interactions [25]. For example, in *Neisseria* bacterium octamer repeats are specifically enriched, suggesting a special mechanism for their generation and instability [26].

Short direct tandem repeats (microsatellites) seem to be rare in the closely related onion yellows phytoplasma genomes, based on searching the Microorganisms Tandem Repeats Database [27],[28]. Thus the common GT-rich octamers mentioned above are most likely interspersed multicopy sequences of unknown function.

In summary, the fractal histogram plot seems very useful to show simultaneously over-represented and under-represented oligomers that may be under

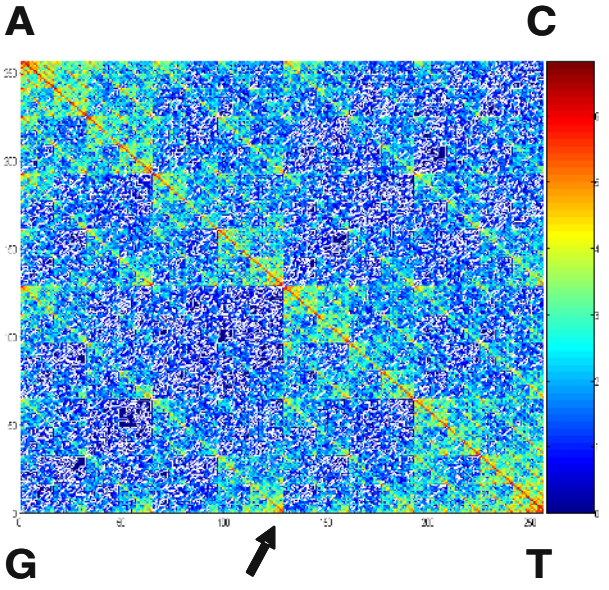

Fig. 4. Aster Yellows Witches' Broom phytoplasma genome octamers visualized in a 256 x 256 (28 x 28) grid as a heatmap, red colour means higher frequency. The abundance of A and T rich octamers is obvious on the red diagonal and in the top left and bottom right corners. An arrow points to a cluster of GT-repeats in the middle of bottom border between G and T corners.

special evolutionary selection pressures. A specific feature of the fractal representation is that the oligomers cluster based on their similarity starting from the beginning of the sequence, so that sequences with the same beginning but different suffixes are near each other.

3.4 Further Improvement of the Approach

For a more detailed analysis of any genome, one would draw fractal histograms with different oligomer lengths to identify specific repeated interspersed motifs in the genome. Overlaying/substracting from a plot of similar length random sequence with same ratios of A/T/C/G could show statistically significant differences according to a specific cutoff.

Successive sections of the genome could be analyzed separately, so that one could find out repeat-rich regions, coding and non-coding regions and so on in the genome. Keeping track of the oligomer coordinates as well would enable one to map specific oligomer groups to specific locations in the genome. Such a tool could thus be a very versatile method of visual exploration and comparison of genomes. Similarly, comparing two or more genomes by overlaying could be easily accomplished, to pinpoint the relevant changes in abundant or under-represented oligomers in the genome. This would be effective for immediate and informative genome scale visual comparisons.

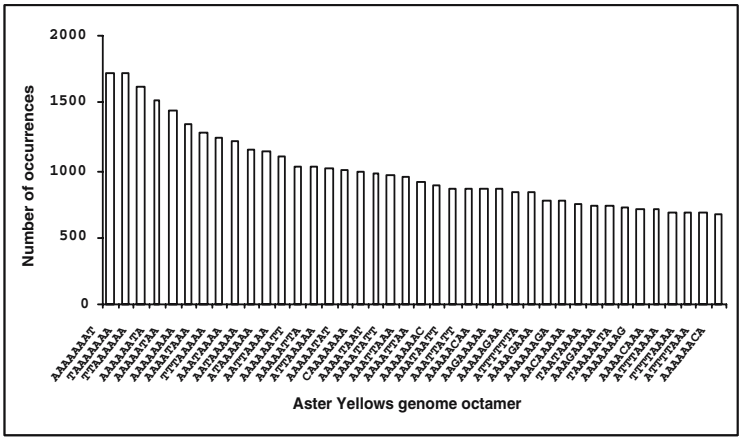

Fig. 5. The most abundant Aster Yellows Witches' Broom phytoplasma genome octamers obtained by RepeatScout algorithm. The abundance of A and T rich oligomers is clear.

The oligomer lengths could be variable, depending on the scale of interest, up to oligomer size 20 or so, which would map all unique single-copy sequences in a separate grid cell. The fractal spaces of the different length oligomers could be viewed successively as a moving colour video track for quick visualization of the relevant features, with several genomes shown side by side in synchrony.

When analysing longer sequences where a small trivial difference may appear in the beginning of the string, leading to a quite different location in the fractal space. This could be mitigated by mapping strings in the reverse direction also. To achieve a sequence similarity based clustering like in BLAST, one would need a different ordering of the similar strings in the fractal space. A specific application for short oligomer based microarray technology is visualization of the set of oligos (the "oligome") on an array, and comparison between arrays and the target transcriptomes/genomes for completeness of coverage of possible hybridization sites. Further extension to larger alphabets to encompass also complete proteomes, rather than short single protein sequences is also an interesting possibility. Finally, automation of the method could be accomplished by image processing of the overlayed/subtracted images to highlight/extract the oligomer clusters of interest in the fractal space, down to the specific most common oligomers differing in frequency between the genomes.

4 Discussion

We have presented two promising visualization and classification methods, both based on transforming a bioinformatics problem to the image analysis domain, to deal with large sets of molecular structures and oligomer motifs in large genomes and proteomes. An example on transforming folded RNA molecules to 2D struc-

ture bitmaps was given, but the approach applies to several domains, including complex organic molecule databases and even protein secondary structure diagrams. For fractal coding of genome oligomer distribution, an example of phytoplasma genome showed that specific types of repeats can be visualized effectively. Various extensions of the fractal method seem worth pursuing for novel types of DNA sequence pattern clustering and classification. Finally, moving the bioinformatics domain symbolic data into bitmap representation domain makes it possible to use the wide variety of bitmap image analysis methods developed in other fields outside biology. This interdisciplinary approach should be both interesting and fruitful for informative visualization, data mining and knowledge discovery in bioinformatics and chemoinformatics datasets.

Acknowledgments. Supported by the Knowledge Engineering and Discovery Research Institute, Auckland University of Technology and the FRST NERF Fund (AUTX0201), New Zealand.

References

1. Banville, D.L.: Mining the chemical structural information from the drug literature. Drug Discovery Today **11(1/2)** (2006) 35–42
2. http://www.genomesonline.org/
3. Vinogradov, A.E.: Noncoding DNA, isochores and gene expression: nucleosome formation potential. Nucleic Acids Res. **33(2)** (2005) 559–63
4. http://www.rcsb.org/pdb/holdings.do
5. Vlahovicek, K. *et al.*: CX, DPX and PRIDE: WWW servers for the analysis and comparison of protein 3D structures. Nucleic Acids Res. **1(33)** (2005) W252–254
6. http://fightaidsathome.scripps.edu/index.html
7. Vitkup, D. *et al.*: Completeness in structural genomics. Nature Struct. Biol. **8** (2001) 559–566
8. Washietl, S. *et al.*: Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nature Biotechnol. **23(11)** (2005) 1383–1390
9. Washietl, S., Hofacker, I.L., Stadler, P.F.P., Tino, P.: Fast and reliable prediction of noncoding RNAs. PNAS USA **102(7)** (2005) 2454–2459
10. Ding, Y., Lawrence, C.E.: A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31** (2003) 7280–7301
11. Gardner, P.P., Giegerich, R.: A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics **5(140)** (2004) 1–18.
12. Zaki, M.J.: Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications. IEEE Trans. Knowl. Data Eng. **17(8)** (2005) 1021–1035
13. Havukkala, I., Pang, S.N., Jain,V., Kasabov, N.: Classifying microRNAs by Gabor filter features from 2D structure bitmap images on a case study of human microRNAs. J. Comput. Theor. Nanosci. **2(4)** (2005) 506–513
14. Mattick, J.S., Makunin,I.V.: Small regulatory RNAs in mammals. Hum. Mol. Genet. **14(1)** (2005) R121–132
15. Hofacker, I.: Vienna RNA secondary structure server. Nucleic Acids Res. **31** (2003) 3429–3431

16. Frazer,K.A. *et al.*: VISTA: computational tools for comparative genomics. Nucleic Acids Res. **32**(2004) W273-279
17. Brudno, M. *et al.*: Automated whole-genome multiple alignment of rat, mouse, and human. Genome Res. **14(4)** (2004) 685–692
18. Jeffrey, H.J.: Chaos game representation of gene structure. Nucleic Acids Res. **18(8)** (1990) 2163–2170
19. Fiser, A., Tusn??ady, G.E., Simon, I.: Chaos game representation of protein structures. J. Mol. Graphics **12** (1994) 302–304
20. Hao, B., Lee, H., Zhang, S.: Fractals related to long DNA sequences and complete genomes Chaos, Solitons and Fractals **11(6)** (2000) 825–836
21. Almeida, J.S. *et al.*: Analysis of genomic sequences by Chaos Game Representation Bioinformatics **17(5)** (2001) 429–37
22. Tino, P.: Spatial representation of symbolic sequences through iterative function systems. IEEE Trans. Syst. Man Cybernet. **29** (1999) 386–393
23. Bai, X. *et al.*: Living with genome instability: the adaptation of phytoplasmas to diverse environments of their insect and plant hosts. J. Bacteriol. **188** (1999) 3682–3696
24. Price, A.L., Jones, N.C., Pevzner, P.A.: De novo identification of repeat families in large genomes. Bioinformatics. , **21(Suppl. 1)** (2005) i351–i358
25. Aras, R.A. *et al.*: Extensive repetitive DNA facilitates prokaryotic genome plasticity. Proc. Natl. Acad. Sci. USA **100(23)** (1999) 13579–135784
26. Saunders, N.J. *et al.*: Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58. Molecular Microbiology, **37(1)** (2000) 207–215
27. Denoeud, F., Vergnaud, G.: Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource. BMC Bioinformatics. **5** (2004) 4
28. http://minisatellites.u-psud.fr/