W686–W691 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl040

GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences

Feng Gao and Chun-Ting Zhang*

Department of Physics, Tianjin University, Tianjin 300072, China

Received December 19, 2005; Revised and Accepted January 18, 2006

ABSTRACT

In order to understand the evolution, structure and function of genomes, it is important to know the general compositional features of DNA sequences. Based on the quadratic divergence, a new segmentation algorithm to partition a given genome or DNA sequence into compositionally distinct domains has been put forward. With the aid of the technique of cumulative GC profile, the distribution of segmentation points can be displayed intuitively. We have therefore developed them into GC-Profile, an interactive web-based software system, which can be used to segment prokaryotic and eukaryotic genomes. GC-Profile provides a quantitative and qualitative view of genome organization. Based on the obtained results, the relationships between the G+C content and other genomic features, such as distributions of genes and CpG islands, can be analyzed in a perceivable manner. It shows that GC-Profile would be an appropriate starting point for analyzing the isochore structure of higher eukaryotic genomes, and an intuitive tool for identifying genomic islands in prokaryotic genomes. GC-Profile is freely available at the website http://tubic.tju.edu.cn/GC-Profile/. In addition, precompiled binaries, together with examples and documentation, can also be freely downloaded for a local execution.

INTRODUCTION

With the advent of high-throughput DNA sequencing, genomic sequences of numerous prokaryotic and eukaryotic organisms have become publicly available. In order to understand the evolution, structure and function of genomes, it is important to know the general compositional features of DNA sequences. Delineating compositionally homogeneous G + C domains in DNA sequences can provide much insight into the understanding of the organization and biological functions of genomes. Furthermore, quantitative analysis of compositional heterogeneity of genome sequences reveals important statistical properties that are useful to locate the origin and terminus of replication in bacterial (1) and archaeal (2) genomes, and to detect horizontally transferred genes and genomic islands (3).

Historically, many windowless methods have been developed to calculate the G + C content, which are usually given the name of 'segmentation of DNA sequences'. Among them, the methods of entropic segmentation (4,5), hidden Markov model (HMM) (6,7) and wavelet shrinkage technique (8) should be mentioned. Recently, a computer program (IsoFinder), based on a modified version of the entropic compositional segmentation algorithm, has been available online and can be used to identify isochores (9).

Our group has developed a suite of segmentation programs. The first program is the cumulative GC profile (10), which has been applied successfully to prokaryotes (3) and eukaryotes (11). Recently, we also developed a new segmentation algorithm for DNA sequences, which is based on the quadratic divergence (12). We have since developed them into GC-Profile, an interactive web-based software system, available as a public resource at http://tubic.tju.edu.cn/GC-Profile/.

METHODS

A new segmentation algorithm of DNA sequences

The genome order index S is defined by (13)

$$S \equiv S(P) = a^2 + c^2 + g^2 + t^2, \quad \mathbf{1}$$

where a, c, g and t denote the occurrence frequencies of A, C, G and T, respectively, in a genome or a DNA sequence, and S can serve as an appropriate divergence measure to quantify the compositional difference between two DNA sequences (12). Consider a genome with N bases. Let n be an integer,

*To whom correspondence should be addressed. Tel: +86 22 2740 2987; Fax: +86 22 2740 2697; Email: ctzhang@tju.edu.cn

$2 \leqslant n \leqslant N-1$. For a given n, the genome sequence is partitioned into two sub-sequences, one left and the other right. The compositional difference between the right and left sub-sequences can be quantified by the quadratic divergence, as described in the following. Let $w_1 = n/N$ and $w_2 = (N-n)/N$ be two weight coefficients. Let $P_l = (a_l,c_l,g_l,t_l)$ and $P_r = (a_r,c_r,g_r,t_r)$ be two vectors, where a_l, c_l, g_l, t_l and a_r, c_r, g_r, t_r are the occurrence frequencies of bases A, C, G and T in the left and right sub-sequences, respectively. Define the quadratic divergence

$$\Delta S(P_l, P_r) = w_1 S(P_l) + w_2 S(P_r) - S(w_1 P_l + w_2 P_r), \quad \mathbf{2}$$

where $S(P)$ is defined by Equation 1. The segmentation algorithm proposed here is based on the quadratic divergence. Suppose that n^* is a point, at which $\Delta S(P_l, P_r)$ reaches maximum, then n^* is a compositional segmentation point of the genome found first. The new algorithm is also recursive as in (4,5), i.e. after n^* is determined, the same procedure is applied to both the resulting left and right sub-sequences, respectively. Recursively apply the procedure until the halting parameter is less than a given threshold t_0, or the resulting sub-sequence is shorter than a given minimum length (12).

Cumulative GC profile

We define

$$z_n = (A_n + T_n) - (C_n + G_n), \quad n = 0, 1, 2, \ldots, N,\ z_n \in [-N, N], \quad \mathbf{3}$$

where A_n, C_n, G_n and T_n are the cumulative numbers of the bases A, C, G and T, respectively, occurring in the sub-sequence from the first base to the nth base in the DNA sequence inspected. Here z_n is the z-component of the Z-curve, which is a three-dimensional curve that uniquely represents a DNA sequence (14,15). Usually, for an AT-rich (GC-rich) genome, z_n is approximately a monotonously increasing (decreasing) linear function of n. To amplify the deviations of z_n, the curve of $z_n \sim n$ is fitted by a straight line using the least square technique,

$$z = kn, \quad \mathbf{4}$$

where (z, n) is the coordinate of a point on the straight line fitted and k is its slope. Instead of using the curve of $z_n \sim n$, we will use the z' curve, or cumulative GC profile, hereafter, where

$$z'_n = z_n - kn. \quad \mathbf{5}$$

Let $\overline{\mathrm{G+C}}$ denote the average G + C content within a region Δn in a sequence, we find from Equations 3–5

$$\overline{\mathrm{G+C}} = \frac{1}{2}\left(1 - k - \frac{\Delta z'_n}{\Delta n}\right) \equiv \frac{1}{2}(1 - k - k'), \quad \mathbf{6}$$

where $k' = \Delta z'_n/\Delta n$ is the average slope of the z' curve within the region Δn. The above method to calculate the G + C content is called a windowless technique (10).

SERVER IMPLEMENTATION

The web server GC-Profile is implemented on Apache server and the web interface is designed using Common Gateway Interface (CGI) Perl scripts. The segmentation algorithms, which are based on the quadratic divergence and cumulative GC profile, are written in the language of C++. The output graphs are generated by gnuplot graphic routine (http://www.gnuplot.info/).

INPUTS/OUTPUTS OF THE WEB SERVER

Input options

GC-Profile has a user-friendly and intuitive input interface. Users can choose to paste the sequence in the box or upload the sequence (FASTA format) in a file.

The following inputs are required for the web server GC-Profile.

(i) Halting parameter t_0 for segmentation. The default value is 1000, but this can be changed according to the requirements of users. Note that $t_0 \geqslant 0$ (12).
(ii) Minimum length. Generally, the minimum length is set to be 1000 bp for prokaryotic genomes and 3000 bp for eukaryotic genomes (12).
(iii) Gap size to be filtered. The default value is 1% of the input sequence, i.e. gaps more than 1% of the input sequence are retained, otherwise they are simply deleted. Other values are also provided to satisfy user's need.
(iv) The graph size to output. It defaults to medium (800×600 pixels). User can change the size from small (640×480 pixels) to giant (2400×1800 pixels).
(v) Whether to label the coordinates of segmentation points to the cumulative GC profile.
(vi) Whether to plot z' curve instead of $-z'$ curve. By default $-z'$ curve is plotted.
(vii) Whether to set as multiplot mode, in which plots are placed on the same page.
(viii) Whether to upload a data file containing density distribution of genes (CpG islands; and other genomic elements). With this option the corresponding distribution will be plotted against the G + C content.
(ix) Whether to upload a data file containing absolute coordinates in the input sequence. This option allows users to label the positions of some interesting genes, e.g. horizontally transferred genes, to the cumulative GC profile. It is very useful to reveal the genomic context of these genes.

Outputs

By default GC-Profile generates four files for each job: two tables and two figures. The output web page shows the process of GC-Profile, and provides links to the results of sequence segmentation: (i) coordinates, sizes and G + C contents of the segmented domains as an HTML table (Figure 1A); (ii) number, coordinates, segmentation strength, segmentation times and segmented contig of the segmentation points as an HTML table (Figure 1B); (iii) the cumulative GC profile and (iv) the GC content of the input sequence in PNG format (Figure 1C and Figure 2). If upload options are chosen, the density distribution or the coordinates points labeled to the cumulative GC profile can also be obtained.

APPLICATIONS OF GC-PROFILE TO THE ANALYSIS OF DNA SEQUENCES

The potential applications of GC-Profile are presented here and will be utilized to demonstrate how GC-Profile may be used and what kind of information GC-Profile can provide. Each application is demonstrated by a concrete example. Additional examples are accessible from the website http://tubic.tju.edu.cn/GC-Profile/.

Visualization of the isochore organization of eukaryotic genomes

The nuclear genomes of vertebrates are mosaics of isochores, very long stretches (>300 kb) of DNA that are fairly homogeneous in base composition [for reviews, see (16,17)]. The large-scale variation in base composition affects both coding and non-coding sequences and seems to reflect a fundamental level of genome organization (18). This isochore organization shows marked variation in a number of important biological properties, including gene density, chromosome bands, patterns of codon usage, gene length, replication timing, recombination rate and the distribution of transposable elements etc. For more details, see (16,17).

As an example, the isochore map of chicken chromosome 28 is shown (Figure 1). The draft chicken genome sequence, release galGal2, and the associated CpG island data were downloaded from http://genome.ucsc.edu/. To display the global G + C content distribution along the chromosome, gap size to be filtered was set to be 1% of the chromosome size. Applying the segmentation algorithm to the resulting

Halting parameter = 300.00 Filtered gap size = 47314 bp Minimum length = 3000 bp

A

Start (bp)	Stop (bp)	Length (bp)	GC content (%)
1	52132	52132	52.51
52133	722658	670526	45.79
722659	1595651	872993	53.76
1595652	1698610	102959	48.08
1698611	1757620	59010	55.10
1757621	1879393	121773	45.48
1879394	2021042	141649	55.63
2021043	2644230	623188	37.08
2644231	2687837	43607	60.13
2687838	2733050	45213	44.46
2733051	3028793	295743	52.87
3028794	3228500	199707	46.47
3228501	3363805	135305	58.30
3363806	3389525	25720	40.89
3389526	3447846	58321	52.99
3447847	3720812	272966	44.12
3720813	4025924	305112	50.23
4025925	4120934	95010	39.79
4120935	4231479	110545	46.30
4231480	4731479	500000	-

B

No.	Segmentation points	Segmentation strength	Segmentation times	Segmented contig
1	52132	419.03	3	1-4231479
2	722658	3766.84	2	1-4231479
3	1595651	502.76	3	1-4231479
4	1698610	358.13	6	1-4231479
5	1757620	354.60	5	1-4231479
6	1879393	954.21	4	1-4231479
7	2021042	4503.77	1	1-4231479
8	2644230	11612.92	2	1-4231479
9	2687837	396.45	6	1-4231479
10	2733050	527.29	7	1-4231479
11	3028793	910.32	5	1-4231479
12	3228500	599.10	4	1-4231479
13	3363805	824.29	5	1-4231479
14	3389525	519.44	6	1-4231479
15	3447846	2100.04	3	1-4231479
16	3720812	1060.18	5	1-4231479
17	4025924	493.64	4	1-4231479
18	4120934	427.89	5	1-4231479

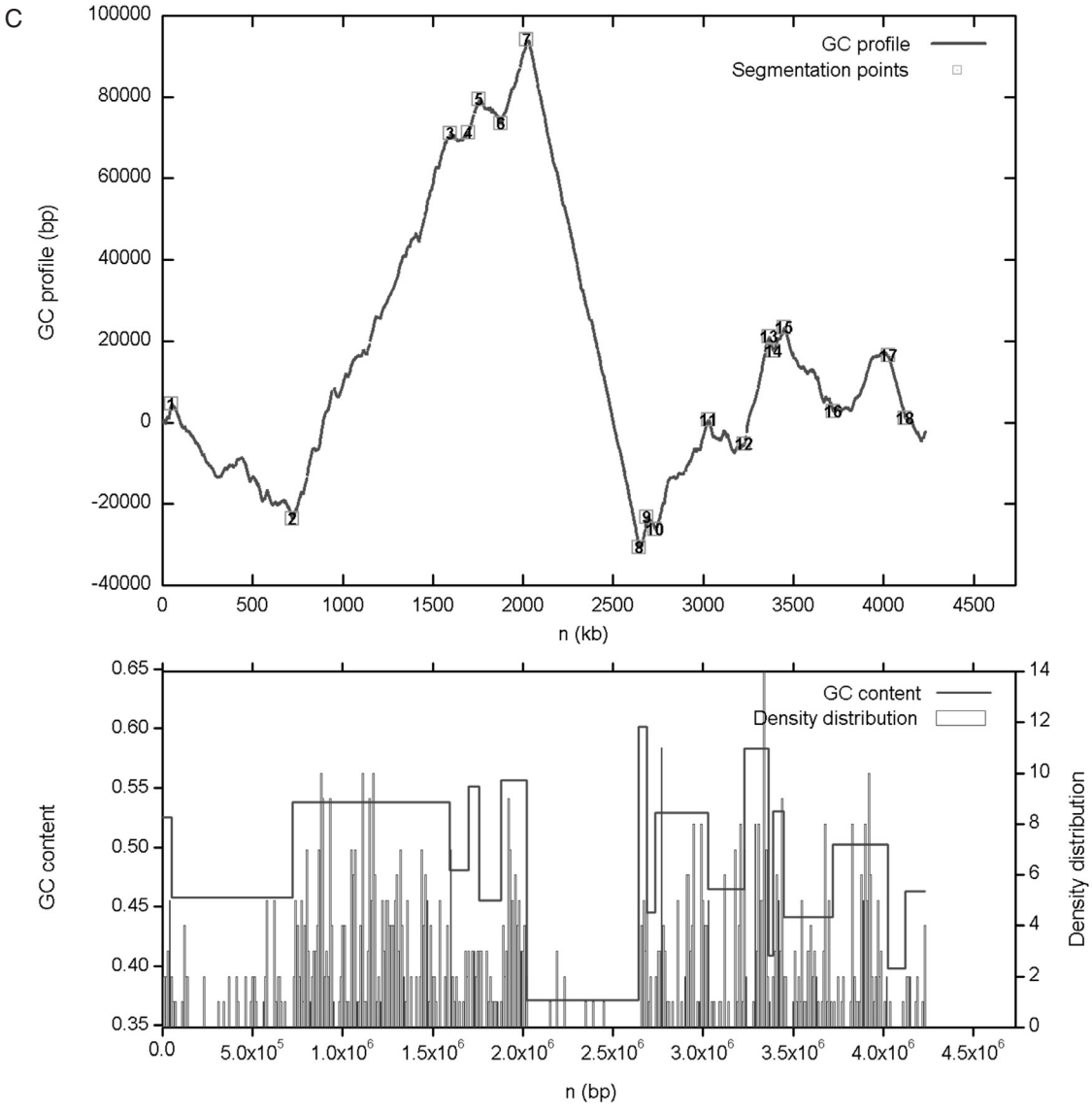

Figure 1. An example of output pages of GC-Profile when the input is the sequence of chicken chromosome 28. (**A**) Coordinates, sizes and G + C contents of the segmented domains as an HTML table. (**B**) Number, coordinates, segmentation strength, segmentation times and segmented contig of the segmentation points as an HTML table. (**C**) The negative cumulative GC profile for chicken chromosome 28 marked with the segmentation points obtained. The lower plot shows the distributions of the G + C content and CpG islands along chicken chromosome 28. The G + C content is calculated for the domains segmented at $t_0 = 300$. Here, the halting parameter t calculated for each segmentation point is also referred to as the segmentation strength, which is defined based on the quadratic divergence instead of the Jensen–Shannon divergence.

contig, eighteen segmentation points were obtained at $t_0 = 300$ (Figure 1). The region from 2 021 042 (point 7) to 2 644 230 (point 8) bp was deemed as an isochore. The G + C content of this isochore is 37.08%, the lowest G + C content among the resulting regions. As shown in Figure 1C, this region is a desert region of CpG island distribution, which was calculated in 10 kb long, non-overlapping windows. It is also shown that the obtained segmentation points have clear biological implications. Note that the distribution of CpG islands is closely correlated to the segmented regions with distinct G + C content. It is worthwhile to point out that the segmentation points obtained here are exactly the boundaries of the related regions. For example, there is an abrupt decrease (increase) of the density of CpG islands at the first (second) boundary of the G + C-poorest region between 2 021 042 (point 7) and 2 644 230 (point 8) bp on chicken chromosome 28 (Figure 1C). Similar phenomena are observed in other G + C distinct regions. The cumulative GC profiles and the corresponding isochore coordinates for the latest release of human, mouse, rat and chicken genomes (hg17, mm6, rn3 and galGal2, respectively) at UCSC are also accessible from the website http://tubic.tju.edu.cn/GC-Profile/.

Identification of genomic islands in prokaryotic genomes

Horizontal gene transfer is recognized as a major force for microbial evolution, as it leads to 'evolution in quantum leaps' (19,20). Genomic islands are formerly

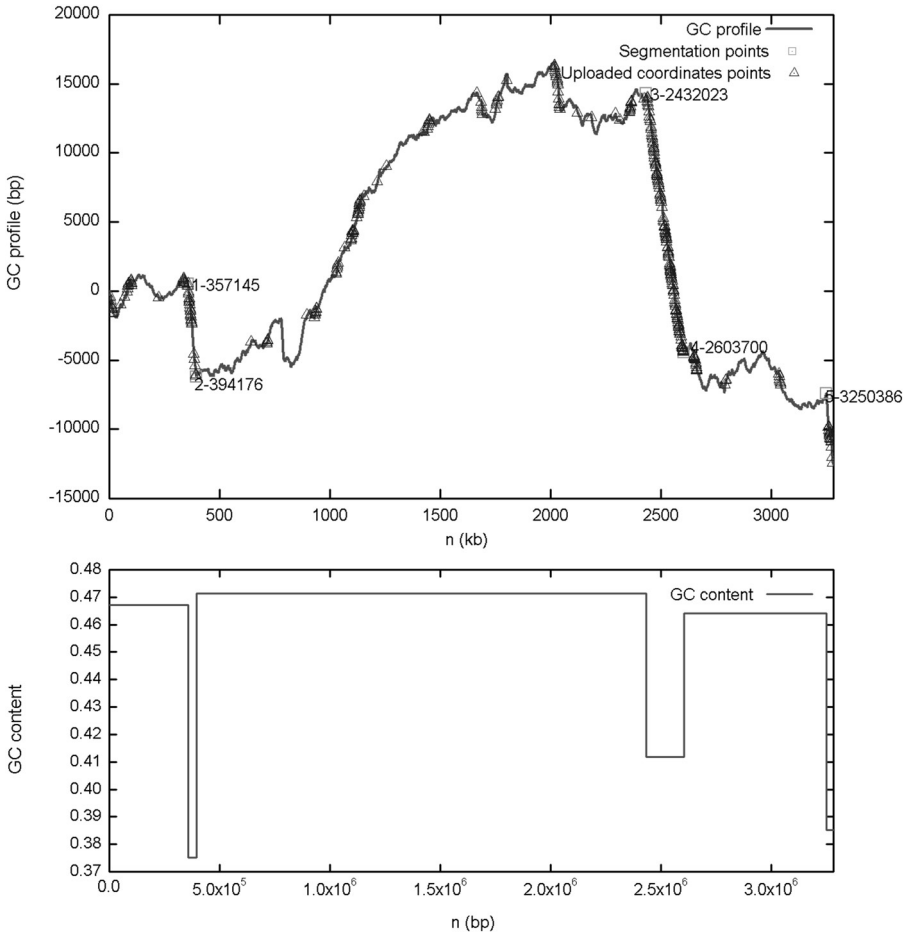

Figure 2. The negative cumulative GC profile for the genome of *V.vulnificus* CMCP6 chromosome I marked with the segmentation points obtained. It shows that from 357 145 to 394 176 bp, 2 432 023 to 2 603 700 bp and 3 250 386 to 3 281 945 bp, there are three regions of low GC content, which are recognized as genomic islands. The segmentation points are obtained at $t_0 = 100$. Here, we also mapped the horizontally transferred genes from HGT-DB to the negative cumulative GC profile. It can be seen that the three regions contain clusters of horizontally transferred genes, which strongly suggests that these regions are horizontally transferred genomic islands.

mobile genetic elements that have been acquired by the core genomes via horizontal gene transfer (21,22). They often consist of DNA regions that differ from the core genome in their G + C content and codon usage (22). Depending on the functions they encode, genomic islands can be classified further as pathogenicity islands, metabolic islands, secretion islands, resistance islands and symbiosis islands (21–23).

Below we show the negative cumulative GC profile for the genome of *Vibrio vulnificus* CMCP6 chromosome I marked with the obtained segmentation points (Figure 2). The segmentation results show that from 357 145 to 394 176 bp, 2 432 023 to 2 603 700 bp and 3 250 386 to 3 281 945 bp, there are three regions of low GC content, which are recognized as genomic islands. These regions have been designed as VVGI-1, VVGI-2 and VVGI-3, respectively in (3). In Figure 2, the negative cumulative GC profile for the genomic islands is distinct from that of the rest of the genome, in that the genomic islands have relatively low GC content, as reflected by abrupt drops in the negative cumulative GC profile at the regions of the genomic islands identified. The abrupt drop in the negative cumulative GC profile indicates that there are clear boundaries between the genomic islands and the surrounding regions. In addition, these three regions have many conserved features of genomic islands. For example, VVGI-1 and VVGI-2 have integrase genes at the 5′ end. VVGI-3 has unusual GC content, codon usage and amino usage, and eight transposase genes. For more details, please refer to (3). Here, we also mapped the genes in horizontal gene transfer database (HGT-DB) (24) to the negative cumulative GC profile. It can be seen that the three regions contain clusters of horizontally transferred genes, which strongly suggests that these regions are horizontally transferred genomic islands.

CONCLUSION

In this article, we present a publicly available, interactive web-based platform, GC-Profile, which is dedicated to analyzing the compositional heterogeneity of DNA sequences. GC-Profile implements a new segmentation algorithm based on the quadratic divergence, and integrates a windowless method for the G + C content computation, known as the cumulative GC profile. The integration of cumulative GC profile with the coordinates of segmentation points leads to a clear graphical representation of the G + C content variation along a genome or chromosome and enables us to establish the

relationships between the G + C content and other genomic features, such as distributions of genes and CpG islands. It shows that GC-Profile would be an appropriate starting point for analyzing the isochore structures of higher eukaryotic genomes, and an intuitive tool for identifying genomic islands in prokaryotic genomes. The advantage of the technique is that an investigator is able to study the variation of GC content in a perceivable and precise manner. The precise boundary coordinates obtained by the segmentation algorithm and the associated cumulative GC profile provides a useful platform to analyze a genome or chromosome.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Ren Zhang, Jian-Hui Zhang and Yan Lin for invaluable assistances. The present work was supported in part by NNSF of China Grant No. 90408028. Funding to pay the Open Access publication charges for this article was provided by the National Natural Science Foundation of China (Grant No. 90408028).

Conflict of interest statement. None declared.

REFERENCES

1. Lobry,J.R. (1996) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, **78**, 323–326.
2. Zhang,R. and Zhang,C.T. (2004) Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea*, **1**, 335–346.
3. Zhang,R. and Zhang,C.T. (2004) A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics*, **20**, 612–622.
4. Oliver,J.L., Bernaola-Galvan,P., Carpena,P. and Roman-Roldan,R. (2001) Isochore chromosome maps of eukaryotic genomes. *Gene*, **276**, 47–56.
5. Li,W., Bernaola-Galvan,P., Haghighi,F. and Grosse,I. (2002) Applications of recursive segmentation to the analysis of DNA sequences. *Comput. Chem.*, **26**, 491–510.
6. Churchill,G.A. (1992) Hidden Markov chains and the analysis of genome structure. *Comput. Chem.*, **16**, 107–115.
7. Peshkin,L. and Gelfand,M.S. (1999) Segmentation of yeast DNA using hidden Markov models. *Bioinformatics*, **15**, 980–986.
8. Lio,P. and Vannucci,M. (2000) Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics*, **16**, 932–940.
9. Oliver,J.L., Carpena,P., Hackenberg,M. and Bernaola-Galvan,P. (2004) IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res.*, **32**, W287–W292.
10. Zhang,C.T., Wang,J. and Zhang,R. (2001) A novel method to calculate the G + C content of genomic DNA sequences. *J. Biomol. Struct. Dyn.*, **19**, 333–341.
11. Zhang,C.T. and Zhang,R. (2004) Isochore structures in the mouse genome. *Genomics*, **83**, 384–394.
12. Zhang,C.T., Gao,F. and Zhang,R. (2005) Segmentation algorithm for DNA sequences. *Phys. Rev. E*, **72**, 041917.
13. Zhang,C.T. and Zhang,R. (2004) A nucleotide composition constraint of genome sequences. *Comput. Biol. Chem.*, **28**, 149–153.
14. Zhang,C.T. and Zhang,R. (1991) Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.*, **19**, 6313–6317.
15. Zhang,R. and Zhang,C.T. (1994) Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn.*, **11**, 767–782.
16. Bernardi,G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**, 3–17.
17. Bernardi,G. (1995) The human genome: organization and evolutionary history. *Annu. Rev. Genet.*, **29**, 445–476.
18. Eyre-Walker,A. and Hurst,L.D. (2001) The evolution of isochores. *Nature Rev. Genet.*, **2**, 549–555.
19. Koonin,E.V., Makarova,K.S. and Aravind,L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.*, **55**, 709–742.
20. Groisman,E.A. and Ochman,H. (1996) Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, **87**, 791–794.
21. Hentschel,U. and Hacker,J. (2001) Pathogenicity islands: the tip of the iceberg. *Microbes Infect.*, **3**, 545–548.
22. Hacker,J. and Carniel,E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity: a Darwinian view of the evolution of microbes. *EMBO Rep.*, **2**, 376–381.
23. Hacker,J. and Kaper,J.B. (2000) Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.*, **54**, 641–679.
24. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, **31**, 187–189.