W692–W695 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl234

CorGen—measuring and generating long-range correlations for DNA sequence analysis

Philipp W. Messer* and Peter F. Arndt

Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received February 14, 2006; Revised March 1, 2006; Accepted March 28, 2006

ABSTRACT

CorGen is a web server that measures long-range correlations in the base composition of DNA and generates random sequences with the same correlation parameters. Long-range correlations are characterized by a power-law decay of the auto correlation function of the GC-content. The widespread presence of such correlations in eukaryotic genomes calls for their incorporation into accurate null models of eukaryotic DNA in computational biology. For example, the score statistics of sequence alignment and the performance of motif finding algorithms are significantly affected by the presence of genomic long-range correlations. We use an expansion-randomization dynamics to efficiently generate the correlated random sequences. The server is available at http://corgen.molgen.mpg.de

INTRODUCTION

Eukaryotic genomes reveal a multitude of statistical features distinguishing genomic DNA from random sequences. They range from the base composition to more complex features like periodicities, correlations, information content or isochore structure. A widespread feature among most eukaryotic genomes are long-range correlations in base composition (1–6), characterized by an asymptotic power-law decay $C(r) \propto \mathrm{r}^{-\alpha}$ of the correlation function

$$C(r) \equiv \sum_{n \in \{A,C,T,G\}} [\mathrm{Prob}(a_i = a_{i+r} = n) - \mathrm{Prob}(a_i = n)^2] \quad \mathbf{1}$$

along the DNA sequence $\vec{a} = a_1, \ldots, a_N$. See the top part of Figure 1 for an example. Amplitudes and decay exponents differ considerably between different species and even between different genomic regions of the same species (6). Often the correlations are restricted to specific distance intervals $\mathrm{r}_{\min} < \mathrm{r} < \mathrm{r}_{\max}$.

The widespread presence of long-range correlations raises the question if they need to be incorporated into an accurate null model of eukaryotic DNA, reflecting our assumptions about the 'background' statistical features of the sequence under consideration (7). The need for a realistic null model arises from the fact that the statistical significance of a computational prediction derived by bioinformatics methods is often characterized by a *P*-value, which specifies the likelihood that the prediction could have arisen by chance. Popular null models are random sequences with letters drawn independently from an identical distribution, or *k*th order Markov models specifying the transition probabilities $P(a_{i+1}|a_{i-\mathrm{k}+1}, \ldots, a_i)$ in a genomic sequence (8). However, both models are incapable of incorporating long-range correlations in the sequence composition. In CorGen we use a dynamical model that was found to efficiently generate such long-range correlated sequences (9). Recent findings already demonstrated that long-range correlations have strong influence on significance values for several bioinformatics analysis tools. For instance, they substantially change the *P*-values of sequence alignment similarity scores (10) and contribute to the problem that computational tools for the identification of transcription factor binding sites perform more poorly on real genomic data compared to independent random sequences (11).

In this paper we present CorGen, a web server that measures long-range correlations in DNA sequences and can generate random sequences with the same (or user-specified) correlation and composition parameters. These sequences can be used to test computational tools for changes in prediction upon the incorporation of genomic correlations into the null model.

ALGORITHM

Several techniques for the generation of long-range correlated sequences have been proposed so far (12–14). Here, we use a simple dynamical method based on single site duplication and mutation processes (15). This dynamics is an instance of a, so called, expansion-randomization system, which recently have

*To whom correspondence should be addressed. Tel: +49 30 8413 1161; Fax: +49 30 8413 1152; Email: philipp.messer@molgen.mpg.de

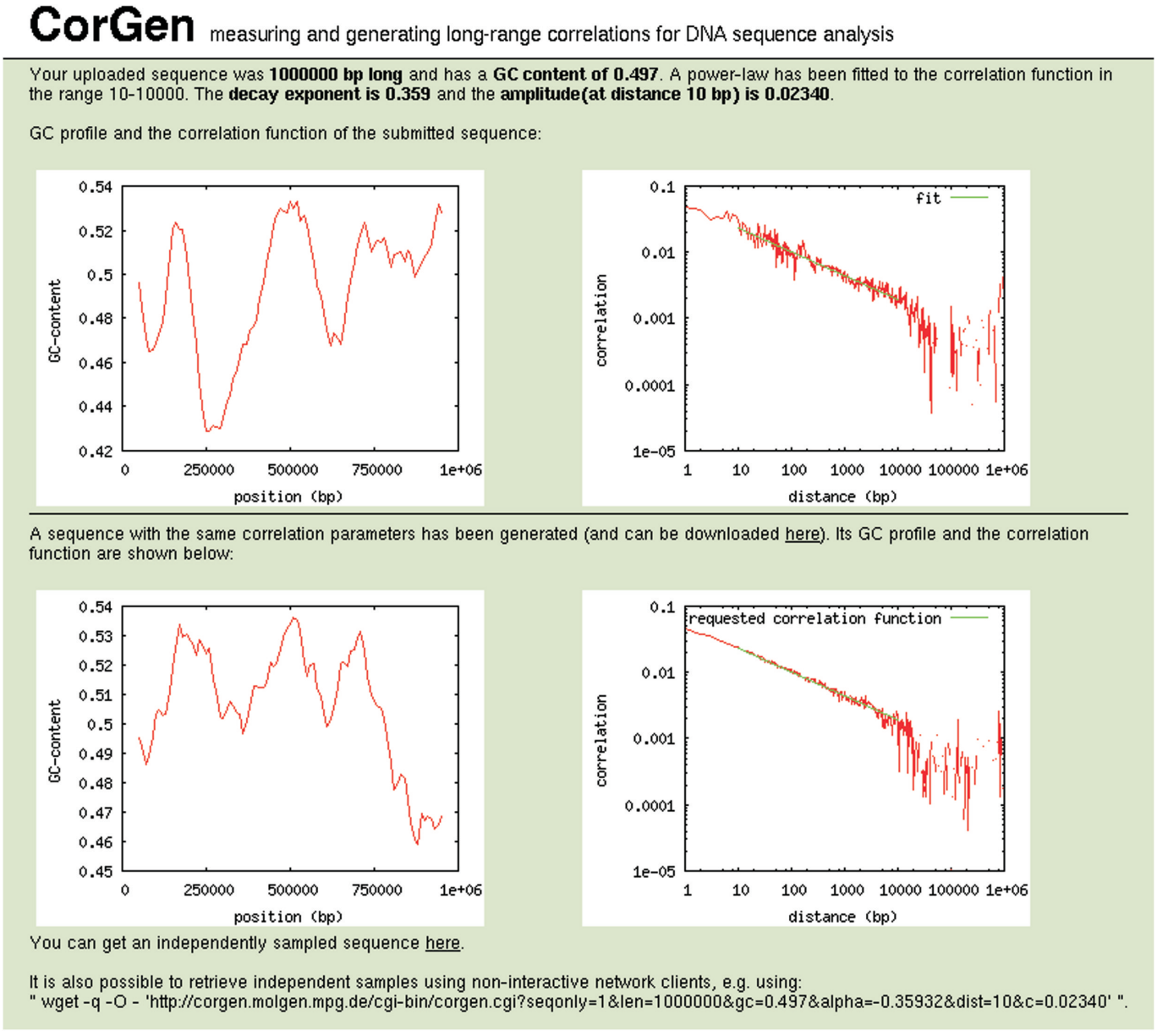

Figure 1. CorGen analysis of a 1 Mb region on human chromosome 22. The two plots in the top part show the measured GC-profile (left) and correlation function (right) of the chromosomal region. In the double-logarithmic correlation graph, power-law correlations $C(r) \propto r^{-\alpha}$ show up as a straight line with slope α. The fitting has been performed in the range $10 < r < 10\,000$, and the obtained parameters are $\alpha = 0.359$ and $C(10) = 0.0234$ (green line). A corresponding random sequence of length 1 Mb with the measured long-range correlation parameters and average GC-content of the query sequence has been generated and can be downloaded by the user. Its composition profile and correlation function are shown in the two plots at the bottom.

been shown to constitute a universality class of dynamical systems with generic long-range correlations (9,16). In contrast to any of the methods (12–14), the duplication-mutation model combines all of the following advantages: (i) exact analytic results for the correlation function of the generated sequences have been derived; (ii) the method allows to generate sequences with any user-defined value of the decay exponent $\alpha > 0$, desired GC-content g, and length N; (iii) the correlation amplitude is high enough to keep up with strong genomic correlations and can easily be reduced to any user-specified value; (iv) the dynamics can be implemented by a simple algorithm with runtime $O(N)$; (v) the duplication and mutation processes are well known processes of molecular evolution.

In CorGen the single site duplication mutation dynamics is implemented by the following Monte Carlo algorithm. We start with a short sequence of random nucleotides ($N_o = 12$). The dynamics of the model is then defined by the following update rules:

(i) A random position j of the sequence is drawn.
(ii) The nucleotide a_j is either mutated with probability P_{mut}, or otherwise duplicated, i.e. a copy of a_j is inserted at position $j + 1$ thereby increasing the sequence length by one.

If the site $a_j = X$ has been chosen to mutate, it is replaced by a nucleotide Y with probability

$$\text{Prob}(X \to Y) = \begin{cases} (1-g)/2 & Y = A, T \\ g/2 & Y = C, G. \end{cases}$$

This assures a stationary GC-content g. Extending the results derived in (16) it can analytically be shown that the correlation function of sequences generated by this dynamics is a Euler beta function with $C(r) \propto r^{-\alpha}$ in the large r limit. By varying the mutation probability P_{mut}, the decay exponent α of the long-range correlations can be tuned to any desired positive value, as it is determined by $\alpha = 2P_{\text{mut}}/(1-P_{\text{mut}})$. The correlations $C(r)$ of the generated sequences define the maximal amplitude obtainable by our dynamics for the specific settings of α and g. However, this amplitude can easily be decreased by the following procedure: after the sequence has reached its desired length, the duplication process is stopped. Subsequent mutation of M randomly drawn sites using the transition probabilities defined in (2) will uniformly decrease the correlation amplitude to $C^*(r) = C(r)\exp(-2M/N)$ without changing the exponent α and the GC-content g (9).

We use a queue data structure to store the sequences, since this allows for a fast implementation of a nucleotide duplication in runtime $O(1)$. The complexity of the algorithm therefore is of the order $O(N + M)$. The software is implemented in C++. Sources are available upon request from the corresponding author.

THE WEB SERVER CorGen

The web server CorGen offers three different types of services: (i) measuring long-range correlations of a given DNA sequence, (ii) generating long-range correlated random sequences with the same statistical parameters as the query sequence and (iii) generating sequences with specific user-defined long-range correlations. The first two tasks require the user to upload a query DNA sequence in FASTA or EMBL format. For long-range correlations to be detectable, the sequences need to be sufficiently long (we recommend at least 1000 bp). The distance interval where a power-law is fitted to the measured correlation function can be specified by the user.

Upon submission of a query DNA sequence, CorGen will generate plots with the measured GC-profile and correlation function, as defined by Equation 1. Unsequenced or ambiguous sites are thereby excluded from the analysis. The user can specify a distance interval where a power-law should be fitted to the measured correlation function. The obtained values for the decay exponent α and the correlation amplitude will be reported by CorGen. If a long-range correlated random sequence with the same statistical features in the specified fitting interval has been requested, its corresponding composition and correlation plots will also be shown. See Figure 1, for an example output page. The generated random sequences can be downloaded by the user. If large ensembles of the generated sequences are needed, independent realizations of the sequences can directly be obtained via non-interactive network clients, e.g. wget. Corresponding samples are given on the relevant pages.

CorGen can also be used to generate long-range correlated random sequences with specific user-defined correlation parameters. In this case, the user needs to specify the decay exponent α, the correlation amplitude $C(r^*)$ at a reference distance r^*, the desired GC-content g and the sequence length. Notice that there is a generic limit for the correlation amplitude depending on the values of α and g. As a typical example, the measurement of $C(r)$ for human chromosome 22 takes $\sim$65 s, while a random sequence of length 1 Mb with the same correlation parameters can be generated in <5 s.

ASSESSING SEQUENCE ALIGNMENT SIGNIFICANCE SCORES

In the following, we want to exemplify a possible application of CorGen related to the problem that long-range correlations significantly affect the score distribution of sequence alignment (10). Imagine one aligns a 100 bp long query sequence to a 1 Mb region on human chromosome 22 in order to detect regions of distant evolutionary relationship. The alignment algorithm reports a poorly conserved hit with a P-value of 10^{-2} calculated from the standard null model of a random sequence with independent nucleotides. However, the user does not trust this hit and wants to test whether it might be an artifact of long-range correlations in human chromosome 22. As a first step, the correlation analysis service provided by CorGen is used to assess whether such correlations are actually present in the chromosomal region of interest. It turns out that a clear power-law with $\alpha = 0.359$ can be fitted to $C(r)$, as is shown in the top part of Figure 1. The next step is to retrieve an ensemble of random sequences generated by CorGen with the same correlation and composition parameters as the 1 Mb region of chromosome 22 (large ensembles can also be retrieved by non-interactive network clients). For one such realization the measured GC-profile and correlation function are shown in the bottom part of Figure 1. The 100 bp query sequence is then aligned against each realization of the ensemble in order to obtain the by chance expected distribution of alignment scores under the more sophisticated null model incorporating the genomic long-range correlations. As has been shown in (10), for the measured correlation parameters this can increase the P-value of a randomly predicted (false-positive) hit by more than one order of magnitude. In conclusion, the hit might be rejected as a true orthologous region. CorGen can therefore help to reduce the often encountered high false-positive rate of bioinformatics analysis tools.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the Max-Planck Institute for Molecular Genetics.

Conflict of interest statement. None declared.

REFERENCES

1. Peng,C.-K., Buldyrev,S.V., Goldberger,A.L., Havlin,S., Sciortino,F., Simons,M. and Stanley,H.E. (1992) Long-range correlations in nucleotide sequences. *Nature*, **356**, 168.

2. Li,W. and Kaneko,K. (1992) Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Lett*., **17**, 655.
3. Voss,R.F. (1992) Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. *Phys. Rev. Lett*., **68**, 3805.
4. Arneodo,A., Bacry,E., Graves,P.V. and Muzy,J.F. (1995) Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett*., **74**, 3293.
5. Bernaola-Galvan,P., Carpena,P., Roman-Roldan,R. and Oliver,J.L. (2002) Study of statistical correlations in DNA sequences. *Gene*, **300**, 105.
6. Li,W. and Holste,D. (2005) Universal 1/*f* noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome. *Phys. Rev. E*, **71**, 041910.
7. Clay,O. and Bernardi,G. (2001) Compositional heterogeneity within and among isochores in mammalian genomes: II. Some general comments. *Gene*, **276**, 25.
8. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, England ISBN: 0–521–62971–3.
9. Messer,P.W., Arndt,P.F. and Lässig,M. (2005) Solvable sequence evolution models and genomic correlations. *Phys. Rev. Lett*., **94**, 138103.
10. Messer,P.W., Bundschuh,R., Vingron,M. and Arndt,P.F. (2006) Alignment statistics for long-range correlated genomic sequences. In Apostolico, A., Guerra, C., Istrail, S., Pevzner, P.A. and Waterman, M.S. (eds), *Proceedings of the Tenth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2006).* Springer, Venice, Italy, pp. 426–440.
11. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y. and Kent,W.J. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol*., **23**, 137.
12. Makse,H.A., Havlin,S., Schwartz,M. and Stanley,H.E. (1996) Method for generating long-range correlations for large systems. *Phys. Rev. E*, **53**, 5445.
13. Wang,X.J. (1989) Statistical physics of temporal intermittency. *Phys. Rev. A*, **40**, 6647.
14. Clegg,R.G. and Dodson,M. (2005) Markov chain-based method for generating long-range dependence. *Phys. Rev. E*., **72**, 026118.
15. Li,W. (1991) Expansion-modification systems: A model for spatial 1/*f* spectra. *Phys. Rev. A*, **43**, 5240.
16. Messer,P.W., Lässig,M. and Arndt,P.F. (2005) Universality of long-range correlations in expansion–randomization systems. *J. Stat. Mech*, P10004.