W512–W515 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki489

GENSTYLE: exploration and analysis of DNA sequences with genomic signature

Bernard Fertil*, Matthieu Massin, Sylvain Lespinats, Caroline Devic, Philippe Dumee and Alain Giron

INSERM U. 678, 91 boulevard de l'Hôpital, 75634 Paris, France

Received February 15, 2005; Revised March 31, 2005; Accepted April 21, 2005

ABSTRACT

GENSTYLE (http://Genstyle.imed.jussieu.fr) is a workspace designed for the characterization and classification of nucleotide sequences. Based on the genomic signature paradigm, GENSTYLE focuses on oligonucleotide frequencies in DNA sequences. Users can select sequences of interest in the GENSTYLE companion database, where the whole set of GenBank sequences is grouped per species, or upload their own sequences to work with. Tools for the exploration and analysis of signatures allow (i) identification of the origin of DNA segments (detection of rare species or species for which technical problems prevent fast characterization, such as micro-organisms with slow growth), (ii) analysis of the homogeneity of a genome and isolation of areas with novel functionality (horizontal transfers for example) – and (iii) molecular phylogeny and taxonomy.

GENOMIC SIGNATURE AND DNA STYLE

A great number of DNA sequences are now available from web-based databases. DNA samples of >140 000 named organisms can be found, for example, in GenBank. The characteristics of these sequences have been extensively studied, and extracted information is often interpreted in terms of evolution or systematic molecular biology.

Many works are devoted to the so-called metagenomic analysis of DNA sequences. One approach deals with the frequencies of short oligonucleotides. Karlin and Burge initially focused on dinucleotide relative abundance (1). It quickly became obvious that the set of oligonucleotide frequencies was species specific (2–6). The set of oligonucleotide frequencies was subsequently considered to be a genomic signature.

Studies based on genomic signature are becoming more and more popular (7,8). It has been observed that the genomic signature results from a species-specific 'writing STYLE' (4,9,10). Indeed, on the one hand, the genomic signatures of species differ from one another, and, on the other hand, the majority of genome segments within a species have comparable signatures. As a consequence, each species can be assigned a DNA style that can be derived from most of its available DNA fragments.

The methodology that we have developed thus makes it possible to study and compare a great number of sequences and species, inasmuch as the calculation of a signature on a laptop computer requires <1 s per million nucleotides. The genomic signature is visualized as a parametric image using the 'chaos game representation' algorithm (3,5,8,11–14). Our experience with genomic signatures shows that the comparison of four-letter word signatures offers a good trade-off between accuracy of classification, usual size of DNA fragments and computer load (9,15). In our hands, comparison of signatures is achieved by means of the Euclidian metric in a space with 256 dimensions (there are 256 different 4-letter words). Of course, other methods for comparison of signatures are available. They often provide slightly different results [see Refs (4,8,16,17) for some other measures of dissimilarities].

It must be pointed out that comparisons of DNA style do not require homologous sequences and almost any DNA segment is eligible (4,9). In fact, the species-specific DNA style concept motivates and justifies most of the works dealing with the genomic signature, including, for example, assignment of genomic fragments (4,18), taxonomic/phylogenetic analyses (15,17,19) and detection of horizontal transfers (HTs) (20,21). Detection of HTs is a major application of the DNA style concept. Some of the abnormal patterns in a genome may be considered to result from HTs. Numerous methods relying on a gene's nucleotide or oligonucleotide composition for the detection of HTs are available (22–32). Among them, hidden Markov models (HMMs) and wavelet transforms are two of the efficient approaches in use for detecting and characterizing original motifs and patterns. Their performances have been subjected to extensive comparisons (20,21,31,32).

Many other applications are emerging, such as the characterization of unknown sequences, the quality control of sequencing and pre-processing for homologous sequences screening.

*To whom correspondence should be addressed: Tel: +33 1 53 82 84 05; Fax: +33 1 53 82 84 46; Email: bernard.fertil@imed.jussieu.fr

A web service (http://www.megx.net/tetra) has recently been made available for the comparison of tetranucleotide usage patterns in DNA sequences (33). It comes with pre-computed tetranucleotide usage patterns for 166 prokaryote chromosomes as a source for limited data mining.

GENSTYLE

GENSTYLE is grounded in the genomic signature paradigm. It offers three sets of tools for the characterization and classification of nucleotide sequences. Parts of GENSTYLE were made accessible to the bioinformatics community through our site (http://genstyle.imed.jussieu.fr/) starting in 1999, after the publication of the seminal paper describing the concept and its usefulness (3). The current version results from a substantial redesign that developed into the GENSTYLE workspace. Three dedicated toolboxes have been implemented for collecting, selecting and processing sequences. The sequence analysis toolbox is made for

(i) Identification of the origin of short DNA fragments. Any DNA sequence is eligible for searching for its origin. This feature is useful, for example, for the recognition of rare and/or slow growth organisms (sequences usually hard to characterize).

(ii) Detection of 'atypical' areas in a genome, in particular the detection of HTs (and potential donors). The closest species (from the genomic signature point of view) of an atypical DNA segment give clues about the donor in the case of putative HTs (under implementation).

(iii) Building of taxonomic and phylogenetic trees. Distance between signatures remains to been established as a reference for phylogenetic studies, but several recent and interesting results have shown its potentially great value (15,17). In particular, our current work with corona viruses is very promising with this respect.

There is a large genomic signature database behind GENSTYLE that greatly enhances its power and scope. The full set of GENBANK sequences, stored by species, is available for signature studies. The GENSTYLE companion genomic signature database handles ~170 000 species and unspecified organisms (>2 000 000 DNA sequences). It is updated on a regular basis, using the bimonthly releases issued by GenBank.

GENSTYLE WORKSPACE

GENSTYLE tools are available from within a user workspace. This makes it possible to work online on the whole set (or part of it) of GenBank nucleotide sequences belonging to one or several species. User's sequences can also be uploaded to work with. Tools for the exploration and analysis of signatures are straightforward. They do not require much prior knowledge.

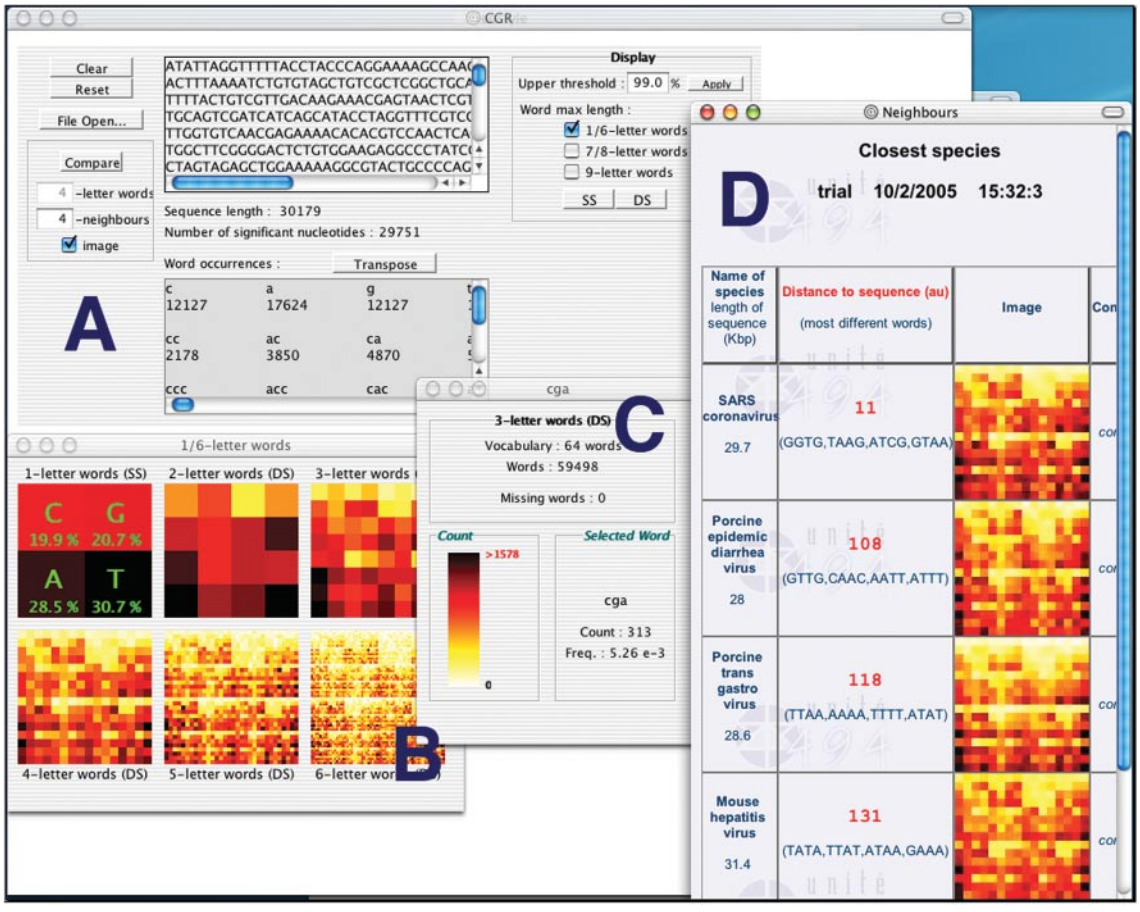

Figure 1. An example from GENSTYLE's online tutorial.

Results are displayed in specific windows with images, tables and charts. Most of the outputs can be downloaded for further processing. The user's workspace can be saved for later use.

There are three toolboxes in the GENSTYLE workspace:

(i) Sequence collector toolbox. The sequence collector allows workspace to be loaded with sequences of interest. Sequences can be selected through the GENSTYLE companion database browser. The user's sequences (FASTA format, eventually grouped into a single text file, zipped or not) can be uploaded through the uploader.

(ii) Sequence filters toolbox. Although many sequences can be uploaded to a given workspace, it may be interesting to work on selected subsets. Several tools are available for this task, including selection of DNA type and size of sequences.

(iii) Sequence analysis toolbox. The tools of this toolbox operate on the selected sequences of the workspace. They allow
   (a) Visualization of sequence signatures.
   (b) Detailed examination of signatures. If several sequences are analysed together, a distance matrix (Euclidean metric, Phylip format) is provided to build taxonomic/phylogenetic trees (8).
   (c) Searching for species with similar DNA signatures in the GENSTYLE companion database.
   (d) Observation of similarities (and differences) between signatures by principal components analysis (PCA) (34).

Online versions of additional tools already in use in our lab are currently under development. They include navigation along genomes by means of local signatures (for HT detection, for example), visualization of similarities between local signatures along several genomes and taxonomic trees.

WORKING WITH GENSTYLE

A tutorial is available online. It demonstrates how the origin of a small DNA sequence can be looked for in the GENSTYLE companion database. Briefly, the sequence of interest has to be pasted into the appropriate field of the demonstrator tool (Figure 1A). The sequence signatures for oligonucleotides (words) 1–9 nt long are subsequently calculated, oligonucleotide counts are obtained (Figure 1A) and signatures are displayed (Figure 1B). Specific word counts and frequencies are available in popup windows (Figure 1C). Species with the closest signatures are then determined (Figure 1D). Distances to the sequence of interest are expressed in an arbitrary unit (AU). It can be seen that the sequence of interest belongs to the SARS Virus ($d = 11$) and that the closest species are PEDV and PTGV corona viruses. Although this procedure seems to mimic BLAST/FASTA functions, it is quite different in nature. Similarities between sequences can be observed even when they are not homologous. As a consequence, the origin of a sequence can be obtained once the DNA material characterizing the genomic signature of the species of origin is available (typically 2000 nt). Homologous DNA counterparts are not required in the database.

ACKNOWLEDGEMENTS

This work was supported by an action Inter EPST Bioinformatique 2001 grant (No. 120910), French Research Ministry. Funding to pay the Open Access publication charges for this article was provided by Pierre & Marie Curie University.

Conflict of interest statement. None declared.

REFERENCES

1. Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**, 283–290.
2. Karlin,S., Campbell,A.M. and Mrazek,J. (1998) Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, **32**, 185–225.
3. Deschavanne,P.J., Giron,A., Vilain,J., Fagot,G. and Fertil,B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, **16**, 1391–1399.
4. Sandberg,R., Winberg,G., Branden,C.I., Kaske,A., Ernberg,I. and Coster,J. (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res.*, **11**, 1404–1409.
5. Almeida,J.S., Carrico,J.A., Maretzek,A., Noble,P.A. and Fletcher,M. (2001) Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics*, **17**, 429–437.
6. Abe,T., Kanaya,S., Kinouchi,M., Ichiba,Y., Kozuki,T. and Ikemura,T. (2002) A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency. *Genome Informatics*, **13**, 12–20.
7. Jernigan,R.W. and Baran,R.H. (2002) Pervasive properties of the genomic signature. *BMC Genomics*, **3**, 23.
8. Wang,Y., Hill,K., Singh,S. and Kari,L. (2005) The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene*, **346**, 173–185.
9. Deschavanne,P.J., Giron,A., Vilain,J., Vaury,A. and Fertil,B. (2000) Genomic signature is preserved in short DNA fragments. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB'00)*, April 8–11, Tokyo, Japan, pp. 232–233.
10. Chapus,C., Fertil,B., Edwards,S., Giron,A. and Deschavanne,P. (2003) Classification of species based on DNA style. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB'03)*, Berlin, Germany, pp. 147–148.
11. Jeffrey,H.J. (1990) Chaos game representation of gene structure. *Nucleic Acids Res.*, **18**, 2163–2170.
12. Jeffrey,H.J. (1992) Chaos game visualization of sequences. *Comput. Graph.*, **16**, 25–33.
13. Dutta,C. and Das,J. (1992) Mathematical characterization of Chaos Game Representation. New algorithms for nucleotide sequence analysis. *J. Mol. Biol.*, **228**, 715–719.
14. Oliver,J.L., Bernaola-Galvan,P., Guerrero-Garcia,J. and Roman-Roldan,R. (1993) Entropic profiles of DNA sequences through chaos-game-derived images. *J. Theor. Biol.*, **160**, 457–470.
15. Edwards,S., Fertil,B., Giron,A. and Deschavanne,P.J. (2002) A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst. Biol.*, **51**, 599–613.
16. Yu,Z. and Jiang,P. (2001) Distance, correlation and mutual information among portraits of organisms based on complete genomes. *Phys. Lett. A*, **286**, 34–46.
17. Yap,Y.L., Zhang,X.W. and Danchin,A. (2003) Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. *BMC Bioinformatics*, **4**, 43.
18. Teeling,H., Meyerdierks,A., Bauer,M., Amann,R. and Glockner,F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, **6**, 938–947.
19. Pride,D.T., Meinersmann,R.J., Wassenaar,T.M. and Blaser,M.J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.*, **13**, 145–158.
20. Dufraigne,C., Fertil,B., Lespinats,S., Giron,A. and Deschavanne,P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.*, **33**, e6.

21. Tsirigos,A. and Rigoutsos,I. (2005) A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.*, **33**, 922–933.
22. Hayes,W.S. and Borodovsky,M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res.*, **8**, 1154–1171.
23. Nakamura,Y., Itoh,T., Matsuda,H. and Gojobori,T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genet.*, **36**, 760–766.
24. Schbath,S., Prum,B. and de Turckheim,E. (1995) Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol.*, **2**, 417–437.
25. Phillips,G.J., Arnold,J. and Ivarie,R. (1987) Mono- through hexanucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. *Nucleic Acids Res.*, **15**, 2611–2626.
26. Reinert,G., Schbath,S. and Waterman,M.S. (2000) Probabilistic and statistical properties of words: an overview. *J. Comput. Biol.*, **7**, 1–46.
27. Arneodo,A., d'Aubenton-Carafa,Y., Audit,B., Bacry,E., Muzy,J.F. and Thermes,C. (1998) What can we learn with wavelets about DNA sequences? *Physica A*, **249**, 439.
28. Nicolas,P., Bize,L., Muri,F., Hoebeke,M., Rodolphe,F., Ehrlich,S.D., Prum,B. and Bessieres,P. (2002) Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res.*, **30**, 1418–1426.
29. Campbell,A., Mrazek,J. and Karlin,S. (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl Acad. Sci. USA*, **96**, 9184–9189.
30. Garcia-Vallve,S., Romeu,A. and Palau,J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.*, **10**, 1719–1725.
31. Ragan,M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.*, **201**, 187–191.
32. Lawrence,J.G. and Ochman,H. (2002) Reconciling the many faces of lateral gene transfer. *Trends Microbiol.*, **10**, 1–4.
33. Teeling,H., Waldmann,J., Lombardot,T., Bauer,M. and Glockner,F.O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, **5**, 163.
34. Flury,B. and Riedwyl,H. (1988) *Multivariate Statistics*. Chapman and Hall, London.