BIOINFORMATICS APPLICATIONS NOTE
Vol. 20 no. 12 2004, pages 1964–1965
doi:10.1093/bioinformatics/bth161

GenAlyzer: interactive visualization of sequence similarities between entire genomes

Jomuna V.Choudhuri[1,†], Chris Schleiermacher[2,†], Stefan Kurtz[3,*] and Robert Giegerich[1]

[1]Faculty of Technology, University of Bielefeld, P.O. Box 100131, 33501 Bielefeld, Germany, [2]Artemis Pharmaceuticals, Neurather Ring 1, 51063 Köln, Germany and [3]Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

Received on December 4, 2003; accepted and revised on January 27, 2004
Advance Access publication February 26, 2004

ABSTRACT

Summary: GenAlyzer is a software tool designed for the interactive visualization of sequence matches between DNA or protein sequences. It provides visualizations on different levels of granularity, from complete overviews via zoomed regions to alignments of particular matching substrings. GenAlyzer can efficiently handle very large datasets, allowing to display tens of thousands of matches between sequences of tens of millions of bases.

Availability: GenAlyzer is available free of charge for non-commercial research institutions. For more details, see http://www.genalyzer.de

Contact: kurtz@zbh.uni-hamburg.de

Since its first release in 1999, the REPuter software tool (Kurtz and Schleiermacher, 1999; Kurtz *et al*., 2001) has been improved in many aspects, and it is now one of the most widely used tools for detecting repeats in large sets of DNA sequences. Besides its main focus on repeat analysis, REPuter has also been used to detect similarities between two different large sequences, say S and T, thereby exploiting the fact that a similarity between S and T is a repeat in the concatenation of S and T. Such usage of REPuter has been extensively described in Kurtz *et al*. (2001). While this kind of usage is conceptually simple, the visualization component of REPuter (called REPvis) lacks explicit support for displaying large-scale matches between two different sequences. Encouraged by many users of REPuter, we have now developed a new software tool GenAlyzer, which allows for the interactive visualization of matches between different sequences, as well as repeats in a single sequence. Thus, it generalizes REPvis. Apart from this fact, GenAlyzer has an improved user interface and provides many new features not present in REPvis. In this application note, we describe the most important aspects of GenAlyzer. A recent application of GenAlyzer is reported in Eder *et al*. (2003).

GenAlyzer visualizes matches (exact or approximate) between two input sequences (DNA or protein) on different levels of granularity. Initially, the central part of the GenAlyzer-display shows the main match graph. The two input sequences are depicted as horizontal lines. A match is shown as a vertical or a diagonal colored line, connecting the start positions of the match in the input sequences. The color encodes the length of the similarity region. A color spectrum specifies which color stands for which length.

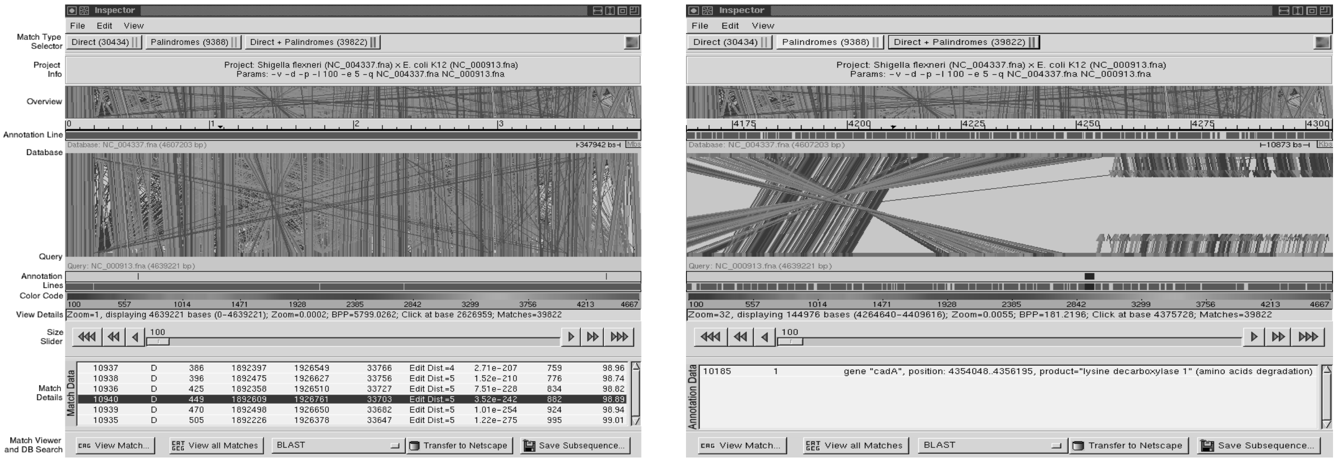
While other software tools apply similar schemes of visualization, they often produce only static views. GenAlyzer, however, is fully interactive. For example, a slider allows to select the minimum length of matches to be shown. By repeated mouse clicks on the colored lines representing the matches, the user can zoom into and out of the corresponding regions of the main match graph. On top of the main match graph, an overview of the entire match graph remains unchanged during the zooming process. To preserve orientation, the zoomed region is depicted by a red frame in the overview match graph.

Additional lines below or above the input sequences allow the display of sequence annotations using colored symbols, as specified by the user in an annotation file. On the finest level of granularity, the user selects a particular match in a zoomed region, and an additional window displays the local alignment of the sequences involved in the match. The latter can directly be transferred into a browser mask to perform a DNA or a protein sequence database search. Depending on the preferences of the user, most of the eight different subparts of the GenAlyzer display can be toggled on or off.

Written in C, GenAlyzer can efficiently handle very large datasets, allowing it to display tens of thousands of matches between sequences of tens of millions of bases. Unlike

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



Project: Shigella flexneri (NC_004337.fna) × E. coli K12 (NC_000913.fna)
Params: −v −d −p −l 100 −e 5 −q NC_004337.fna NC_000913.fna

Database: NC_004337.fna (4607203 bp) ⊢347942 bs⊣

Query: NC_000913.fna (4639221 bp)

Zoom=1, displaying 4639221 bases (0−4639221); Zoom=0.0002; BPP=5799.0262; Click at base 2626959; Matches=39822

10937	D	386	1892397	1926549	33766	Edit Dist.=4	2.71e−207	759	98.96
10938	D	396	1892475	1926627	33756	Edit Dist.=5	1.52e−210	776	98.74
10936	D	425	1892358	1926510	33727	Edit Dist.=5	7.51e−228	834	98.82
10940	D	449	1892609	1926761	33703	Edit Dist.=5	3.52e−242	882	98.89
10939	D	470	1892498	1926650	33682	Edit Dist.=5	1.01e−254	924	98.94
10935	D	505	1892226	1926378	33647	Edit Dist.=5	1.22e−275	995	99.01

(a)

Project: Shigella flexneri (NC_004337.fna) × E. coli K12 (NC_000913.fna)
Params: −v −d −p −l 100 −e 5 −q NC_004337.fna NC_000913.fna

Database: NC_004337.fna (4607203 bp) ⊢10873 bs⊣

Query: NC_000913.fna (4639221 bp)

Zoom=32, displaying 144976 bases (4264640−4409616); Zoom=0.0055; BPP=181.2196; Click at base 4375728; Matches=39822

10185 1 gene "cadA", position: 4354048..4356195, product="lysine decarboxylase 1" (amino acids degradation)

(b)

Fig. 1. GenAlyzer display of the comparison of *S.flexneri* (top) and *E.coli K-12* (bottom). (**a**) In the main match graph, matches ranging from 1400 to 3300 bp shine up in green in front of a blue background of smaller matches (around 100–900 bp). (**b**) Zooming into the match graph on the right end of the two genomes (marked in the overview graph by a red rectangle), one recognizes two large inverted regions, interrupted by a colinear segment. The inversions break the colinearity of the involved segments between both genomes. They are mostly associated with deletions, generating organism-specific DNA regions. Immediately to the right of the inverted region, *E.coli K-12* shows no matches with *S.flexneri*. Close inspection in the annotation shows that this match-free region in *E.coli K-12* holds the gene *cadA*, shown by a black box. This gene encodes a lysine decarboxylase, responsible for converting lysine into cardverine. The absence of this gene in *S.flexneri*, due to some deletion event, affects adversely its virulence, influencing (or contributing to) the pathogenicity of this enteric bacterium (see Jin *et al*., 2002).

REPvis, the input sequences for GenAlyzer are stored in indexed files. Hence no extra parsing is necessary, and the different displays of GenAlyzer pop-up virtually without any delay, thus supporting its interactive use.

The input for GenAlyzer is a match file, specifying the matches and their locations in a simple text file format, e.g. generated by the program Vmatch (http://www.vmatch.de). However, this input format can also be delivered by other programs.

To illustrate the use of GenAlyzer, we show a large-scale comparison of the *Shigella flexneri* genome (4 607 203 bp) and the *Escherichia coli K-12* genome (4 639 221 bp), (Fig. 1). In a first step, we have used Vmatch to compute all 39 822 direct or reverse complemented matches of at least 100 bp between the two genomes, allowing for up to five insertions, deletions or mismatches of single bases in each match. Computing the matches and storing them in a match file takes about 20 s on a 1 GHz Pentium III computer. Extracting the coding sequence annotation from the GenBank entries of the two genomes, and storing them in an annotation file is also done in less than a minute, using standard tools. When GenAlyzer is called to display the match file and the corresponding annotation file, it first shows the longest match between *S.flexneri* (top line) and *E. coli K-12* (bottom line). Using a slider, we bring all 39 822 matches into the display (Fig. 1). A similar visualization (static, without colors and produced by a different software) is shown in Jin *et al*. (2002).

REFERENCES

Eder,V., Ventura,M., Ianigro,M., Teti,M., Rocchi,M. and Archidiacono,N. (2003)Chromosome 6 phylogeny in primates and centromere repositioning. *Mol. Biol. Evol.*, **20**, 1506–1512.

Jin,Q., Yuan,Z., Xu,J., Wang,Y., Shen,Y., Lu,W., Wang,J., Liu,H., Yang,J. and Yang,F. *et al*. (2002) Genome sequence of *Shigella flexneri 2a*: insights into pathogenicity through comparison with genomes of *Escherichia coli K12* and *O157*. *Nucleic Acids Res.*, **30**, 4432–4441.

Kurtz,S., Choudhuri,J., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2001)REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29**, 4633–4642.

Kurtz,S. and Schleiermacher,C. (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**, 426–427.