**ABSTRACT**

Summary: In order to understand and interpret phylogenetic and functional relationships between multiple prokaryotic species, qualitative and quantitative data must be correlated and displayed. GECO allows linear visualization of multiple genomes using a client/server based approach by dynamically creating .png- or .pdf-formatted images. It is able to display ortholog relations calculated using BLASTCLUST by color coding ortholog representations. Irregularities on the genomic level can be identified by anomalous G/C composition. Thus, this software will enable researchers to detect horizontally transferred genes, pseudogenes and insertions/deletions in related microbial genomes.

Availability: [http://bioinfo.mikrobio.med.uni-giessen.de/geco2/GecoMainServlet](http://bioinfo.mikrobio.med.uni-giessen.de/geco2/GecoMainServlet)

Contact: Carsten.Kuenne@mikrobio.med.uni-giessen.de

**1 INTRODUCTION**

Genome based studies of microbial evolution have a strong focus on universally conserved genes across genera and comparisons of shared and distinct gene sets between different groups. This is especially true when comparing closely related species (frequently pathogenic and non-pathogenic). The increasing number of sequenced genomes has raised considerable interest in devising means for comparative analysis to gain further knowledge of function and phylogeny. In such large-scale genome comparisons, it is important to distinguish between conserved and potentially horizontally transferred genes, detect framshifts and insertions/deletions. As horizontally transferred genes often insert preferentially in well defined ‘hot spots’, these loci differ considerably between virulent and non-virulent bacteria. A dynamic visualization method, capable of carrying out multiple comparisons is required to observe this process of genome reduction and expansion and further understanding of disease mechanisms.

Most existing tools in this category are either commercial: GenomeSCOUT (Suter-Cazzolara et al., 2000), ERGO (Overbeek et al., 2002), not locally deployable: MBGD (Uchiyama et al., 2005), GenomeComp (Yang et al., 2003), GenomePixelizer (Kozik et al., 2002). Generic Genome Browser (Stein et al., 2002) could technically deliver on this task, but lacks visual quality and customizability in the interface, which made us implement our own solution. Currently, GECO includes all publically available fully sequenced prokaryotes (389).

**2 IMPLEMENTATION**

GECO (abbreviation of comparative genomics) is implemented in Java 1.5 on a Debian 3.1 Linux PC. It facilitates a client/server architecture using Java Servlets and Apache Tomcat for webinterface and MySQL as database backend.

**3 METHODS**

Orthologs are calculated using BLASTCLUST, which is part of the NCBI BLAST package and uses BLASTP to determine homology of two proteins. It identifies orthologs above a certain cutoff and groups them by single linkage clustering (AB + BC = ABC).

Two genes are considered orthologs if they have: (1) BLASTP E-value <1 x 10^-6, (2) protein identity >60%, (3) both coverages (alignment-protein1 and alignment-protein2) >80% (to filter single domain hits).

All proteins which are elements of the same cluster are colored identically. Introduction of new genomes to GECO is a straightforward process: (1) data in .embl-format is put into genome directory, and (2) the update function is run.

**4 FUNCTIONS**

GECO’s query form was designed for simplicity and adopts a natural workflow. At first, the base gene for the current view has to be defined by choosing a specific identifier (e.g. *dnaA*) and organism. The base track (topmost track of image) is centered on that gene, which is highlighted by a red frame (optionally). All subsequent tracks relate to it. The range of nucleotides to be displayed before and after the base gene can be varied. All genes within this range make up the base track. The number and order of genomes to be compared to the basetrack can be picked, as well as the degree of similarity deemed suitable for an ortholog (currently 60% identity, 80% coverage). Tracks can also be sorted descending by the number of orthologs shared with the base track. GECO incorporates a parameter to mask core genes (present in all selected organisms), so that differences between the tracks are easier to identify. Each gene is associated with a popup-box, displaying gene annotation and a customizable link. The width of the resulting image can be varied to compensate for different screen resolutions.

GECO has two distinct modes of operation: ‘pinned’ and ‘area’ (Fig. 1A and B).

In ‘pinned mode’, GECO will first look for the basegene-ortholog in all chosen organisms and subsequently display each track within the baserange defined. If an ortholog is not found in an organism, its track is omitted. This mode is the default one and useful for quick scanning of a region around a certain gene in related genomes. The display of GC compositional deviation from average is only available using ‘pinned mode’.
In ‘area mode’, GECO will retrieve all genes of the basetrack first. Orthologs of all these genes are looked up in the selected organisms. If an ortholog could be determined, GECO retrieves the ‘neighbourhood’ of the hit (elongation default: 3000 bp up- and downstream of the ortholog). These regions are then connected by gaps to improve readability. In this manner, all orthologs of the basetrack genes can be shown in a single image, even if they are scattered all over the selected genomes.

To improve usability of the visualization servlet, GECO can show a list of all identifiers/annotations and general genomic data (GC content, percentage of coding genes, etc.).

ACKNOWLEDGEMENTS

The authors wish to thank M. Maier for technical support. T.C. acknowledges support from the Bundesministerium für Bildung und Forschung, Germany (NGFN 01GS0401), T.H. from Pathogenomics (PTJ-Bio/03U213B), and R.G. is a member of the Graduate School ‘Biochemistry of Nucleoprotein Complexes’ supported by the Deutsche Forschungsgemeinschaft.

Conflict of Interest: none declared.

REFERENCES