SNAPper: gene order predicts gene function

Grigory Kolesov1,*, Hans-Werner Mewes2 and Dmitrij Frishman1

1Institute for Bioinformatics, GSF-National Research Center for Environment and Heath, Ingolstäder Landstraße 1, 85764 Neuherberg, Germany and 2Technical University of Munich, Wissenschaftszentrum Weihenstephan, D-85350 Freising-Weihenstephan, Germany

Received on October 22, 2001; revised on November 22, 2001; accepted on December 6, 2001

ABSTRACT

Summary: SNAPper is a network service for predicting gene function based on the conservation of gene order.
Availability: The SNAPper server is available at http://pedant.gsf.de/snapper. SNAPper-based functional predictions will soon be offered as part of the PEDANT genome analysis server (http://pedant.gsf.de).
Contact: g.kolesov@gsf.de

One of the major recent advances in bioinformatics has been the development of similarity-free methods to predict gene function. These methods explore genomic context, rather than sequence alignment, to establish functional links between genes. The notion of ‘genomic context’ can be interpreted as a coordinated occurrence of genes in genomes (Pellegrini et al., 1999), similarity of expression profiles (Marcotte et al., 1999), or gene fusion events (Enright et al., 1999). However, the most straightforward definition is based on the physical proximity of genes on the chromosome (Tamames et al., 1997; Dandekar et al., 1998). Neighbouring genes in bacteria are often functionally related owing to their involvement in the same operon—a group of genes transcribed and regulated as one unit. Specific, non-random gene order in prokaryotes represents an additional information signal. Automatic computational techniques have been proposed that explore cross-genome sequence comparisons to identify conserved gene strings and hence putative operon remnants (Overbeek et al., 1998, 1999; Fujubuchi et al., 2000; Wolf et al., 2001). A specialized web server, STRING (Snel et al., 2000) provides on-line analysis of such recurring instances of neighbouring genes.

The main difficulty in predicting functionally coupled genes based on their chromosome location is that collinear gene strings are poorly conserved between distantly related species (Mushegian and Koonin, 1996). We have recently proposed a more general approach to detect functionally linked genes based on the conservation of gene order (Kolesov et al., 2001). Our algorithm, Similarity-Neighbourhood Approach (SNAP) does not explicitly rely on the conservation of individual gene strings across several genomes. Instead, a Similarity-Neighbourhood Graph (SN-Graph) is built that involves chains of alternating S- and N-relationships. The former represent BLAST (Altschul et al., 1997) similarity hits between putative orthologs in different genomes while the latter involve neighbouring genes on the same genome. An SN-Graph can thus be thought of as a walk across many genomes which begins with a particular gene, say gene number 750 in genome A (denoted geneA750) and proceeds to its ortholog in genome B, say geneB750. The walk then continues to encompass n neighbours of geneB750 on each side. Assuming n = 1, geneB1334 and geneB1336 will be considered involved in an N-relationship with geneB1335. Subsequently, orthologues of geneA1334 and geneA1336 are found in other genomes, their neighbours identified, and so on. Closed paths on an SN-graph, that we call SN-cycles, are strongly non-random and have the tendency to join functionally related genes involved in the same metabolic or regulatory pathway. SN-cycles have significant prediction power in uncovering associations between genes not detectable by the analysis of either S- or N-relationships alone. Note that conventional collinear gene strings will also be found by SNAP as a special case of SN-cycles.

Here we report a web-based tool, SNAPper, which allows to conduct SNAP function predictions for query protein sequences. The service builds on the PEDANT genome analysis server (Frishman et al., 2001) which currently contains automatically generated annotation for over 500,000 genomic sequences. Each gene product is subjected to exhaustive bioinformatics analysis, including homology searches, detection of protein motifs, prediction of secondary structure and other protein features. Proteins are also automatically attributed to pre-defined functional categories and EC numbers. All PEDANT data are stored in a relational database and can be directly accessed by the SNAPper server. Using SQL queries, it is possible to correlate pre-computed properties of gene products with

*To whom correspondence should be addressed.
the results of the SNAPper analysis.

At the time of writing SNAPper utilizes a selection of 12 phylogenetically distant microbial genomes. The pre-computed SN-graph for these organisms is constantly resident in the computer memory. A SNAPper search can be initiated either with a query protein sequence submitted via a Web form, or by specifying a PEDANT gene id, if known. In the former case a BLAST search is performed to find orthologues, the number of gene neighbours considered, etc.) are set to strict values in order to reduce the number of false positives found, but can be manipulated by advanced users.

ACKNOWLEDGEMENTS

We thank Mr John Reguera and Mr David Butenhof from Compaq for their priceless hints on Alpha Tru64 threads programming.

REFERENCES


