SSDB: Sequence Similarity Database in KEGG

Yoko Sato[1] ysatoh@fqs.fujitsu.com
Akihiro Nakaya[2] nakaya@kuicr.kyoto-u.ac.jp
Kotaro Shiraishi[1] kshirais@scl.kyoto-u.ac.jp
Shuichi Kawashima[2] shuichi@kuicr.kyoto-u.ac.jp
Susumu Goto[2] goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa[2] kanehisa@kuicr.kyoto-u.ac.jp

[1] Fujitsu Kyushu System Engineering Limited, 2-2-1 Momochihama, Sawara-ku, Fukuoka 814-8589, Japan
[2] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan

Keywords: sequence similarity, complete genome, bidirectional best hit, protein universe

1 Introduction

Availability of a large number of complete genomes enables us to compare several genomes and to search common and different features between genomes in terms of protein sequence similarities, which we call comparative genomics. It produces information about proteins useful for the assignment of the function to genes and for the research on the evolution of the genome. The large number of genes accumulated in the databases of complete genomes, however, has become a bottleneck, because the computation of the sequence similarity of all pairs of proteins is time consuming even if we use a supercomputer. Therefore precomputed sequence similarities of completely sequenced organisms are indispensable for comparative genomics.

SSDB (Sequence Similarity Database) is a new addition to the KEGG suite of databases [3] and contains the information about amino acid sequence similarities among all protein-coding genes in the complete genomes, together with the information about best hits and bidirectional best hits (best-best hits). The relation of gene x in genome A and gene y in genome B is called bidirectional best hits, when x is the best hit of query y against all genes in A and vice versa, and it is often used as an operational definition of ortholog. We report here the system design and simple search capabilities of SSDB.

2 System Design

The similarity scores and information about the alignment are computationally generated from the GENES database in KEGG. We perform all possible pairwise genome comparisons by using the SSEARCH program in the FASTA package [4], and the gene pairs with the Smith-Waterman similarity score [5] of 100 or more are entered in SSDB. We also compute best hits and bidirectional best hits after the SSEARCH computation of all possible gene pairs from two complete genomes is finished.

The results are stored in the relational database management system PostgreSQL, together with the description and the amino acid sequence length of each gene. Currently, about 40 million pairs from 67 organisms are included. We have designed the view of each organism for efficient search of similar genes to a specified gene.

For end users, a web-based system has been developed to explore the universe of protein-coding genes in SSDB (http://ssdb.genome.ad.jp/). Besides the simple search of various relations shown in the next section, it enables users to search common sequence motifs, to display multiple alignments, and to display the dendrogram calculated by the Smith-Waterman scores among a set of selected genes, which are usually the search result from a query gene.

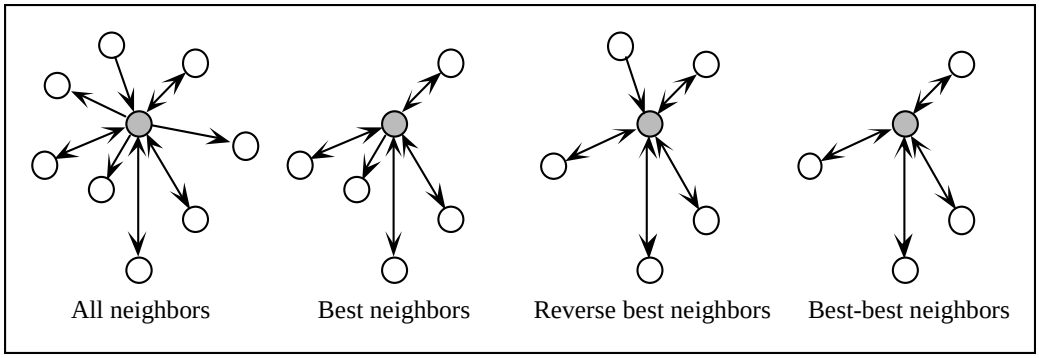

Figure 1: Different types of neighbors for a given query sequence (shaded).

3 Queries Against SSDB

SSDB is a huge graph consisting of protein-coding genes as its nodes and sequence similarities as its edges. We call this graph the protein gene universe, or simply the protein universe. At the moment, we provide the following search capabilities for the SSDB graph features (See Fig. 1).

1) All neighbors (equivalent to usual homology searches)
2) Best neighbors (functional links)
3) Reverse best neighbors (functional links)
4) Best-best neighbors (strong functional links)

Namely, in addition to the standard homology searches for all neighbors, SSDB queries can be limited to best neighbors or best-best neighbors, which correspond to functionally better-defined sets of sequence similarities. Furthermore, incorporation of the edges that represent adjacent genes on the chromosome makes it possible to identify contiguous sets of best-best neighbors. This means that conserved gene clusters are automatically computed and can be used for the construction of the KEGG ortholog group tables.

Other edges, which may be added to the SSDB graph to identify various functional links, include common sequence motifs and common folds in the 3D structures. We will implement more features in conjunction with the BRITE database (http://www.genome.ad.jp/brite/). For the moment the sequence motifs in PROSITE [2] and Pfam [1], which are precomputed, can be searched as part of SSDB for all proteins in the KEGG/GENES database.

Acknowledgements

This work was supported by the grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L., The Pfam protein families database, *Nucleic Acids Res.*, 28(1):263–266, 2000.

[2] Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A., The PROSITE database, its status in 1999, *Nucleic Acids Res.*, 27(1):215–219, 1999.

[3] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya A., The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 2002. (in press).

[4] Pearson, W.R., Effective protein sequence comparison, *Methods. Enzymol.*, 266:227–258, 1996.

[5] Smith, T.F. and Waterman, M.S., Identification of common molecular subsequences, *J. Mol. Biol.*, 147(1):195–197, 1981.