HAPPY: Hypothetical and Putative Protein Database System

Takuya Murakami takuzou@gen-info.osaka-u.ac.jp
Masatomo Najima najima@gen-info.osaka-u.ac.jp
Michihiro Ogawa michi@gen-info.osaka-u.ac.jp
Ken Kurokawa ken@gen-info.osaka-u.ac.jp
Teruo Yasunaga yasunaga@gen-info.osaka-u.ac.jp

Genome Information Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan

Keywords: putative protein, hypothetical protein, ortholog, paralog, database

1 Introduction

More than 100 bacterial complete genome sequences were determined to date, and organisms with related species increased. Under such the situation, comparative genome analysis came to reveal various things. Nevertheless, there are still many genes which are annotated as unknown function, and these consist of about 40% of all genes. However, even if they are annotated as unknown function, it turns out that there are many genes conserved over the various species from the comparative genome analysis. Although such highly conserved hypothetical genes are annotated as unknown function, we can presume that they should play very important roles in organisms. Therefore, we performed functional presumption of hypothetical gene, and in order to provide the useful data to a researcher, we have established Hypothetical And Putative Protein database sYstem (HAPPY) (`http://www.happy.gen-info.osaka-u.ac.jp/`). HAPPY resembles GeneQuiz [1] that offers large-scale biological sequence analysis using the various analyzing methods and the searching method. In marked contrast to GeneQuiz, HAPPY is specialized in treating hypothetical gene and use not only the above powerful tools but also paralog and ortholog analyses for presuming functions of hypothetical genes.

2 Method

We collected only bacterial uncharacterized genes that were annotated as hypothetical, uncharacterized and putative protein from RefSeq[4]. We performed homology search, domain predicting, and localization predicting for hypothetical genes obtained by these methods described in table 1. In addition, we performed ortholog analysis using COGs and MBGD [2] data sets and provided the result of gene order near the gene in progress and a phylogenetic tree of the genes in the orthologous group. For paralog analysis, we used the result of ParalogCluster [3]. Like ortholog analysis, we provided the gene order near the gene in progress and a phylogenetic tree of the genes in the paralogous group. Furthermore, we introduced a index called HAPPY score that was the integration of scores of tools used, and tried to anotate hypothetical genes (Fig. 1).

3 Results and Discussion

In HAPPY, it is possible to carry out presumption of function of hypothetical genes with powerful tools and the legible interface. Furthermore, in both the paralog analysis and the ortholog analysis

Table 1: Tools used in HAPPY.

Homology Search		Domain Prediction	Localization Predicting	Phylogenetic Tree and Alignment	Ortholog analysis	Pralog analysis
BLASTP	FASTA	InterPro	SOSUI	Clustalw	COGs	PralogCluster
TBLASTN	SSearch	Pfam HMM	PSORT-B	Treeview	MBGD	
PSI-BLAST			SignalP			

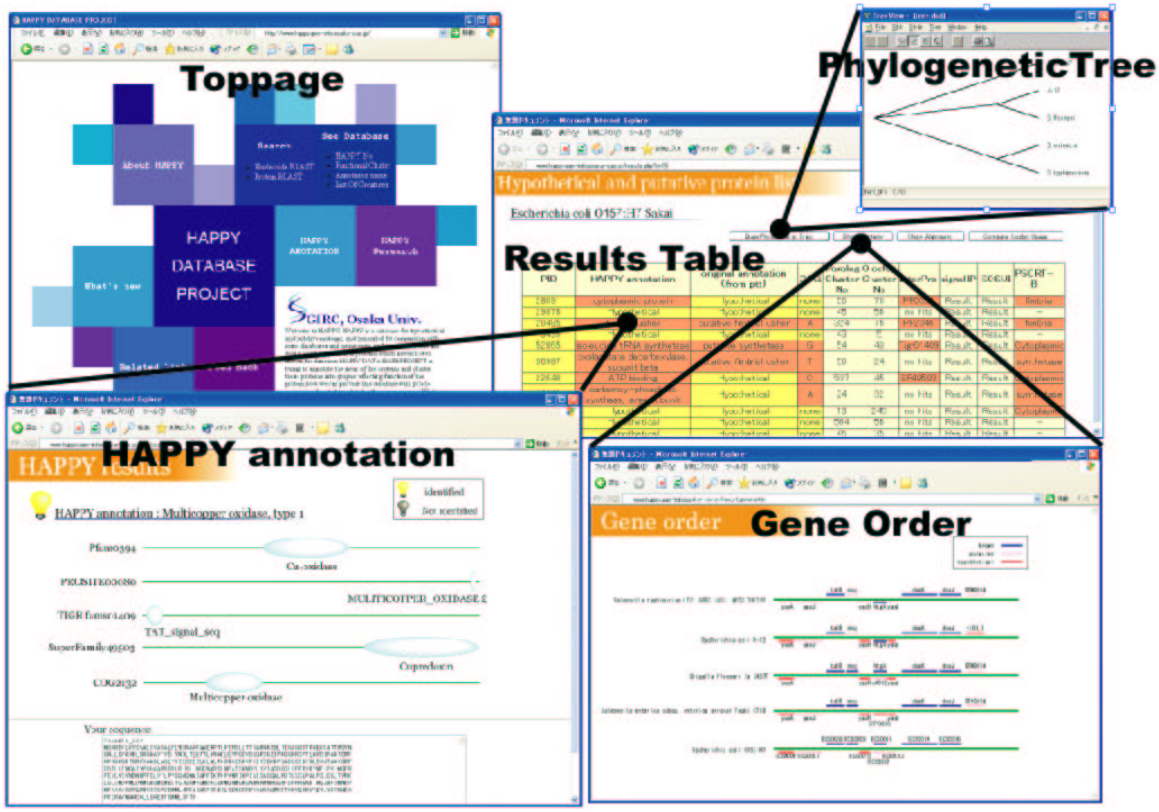

Figure 1: Screenshots of HAPPY graphical user interface.

of a hypothetical gene, we can get information about whether the gene performs essential function for organisms or not. However, a large number of genes were not annotated even in HAPPY at present. So, improvement in the further accuracy should be needed by adoption of a new tool and reexamination of a scoring. HAPPY has another feature that is analysis of a new gene which is not stored in HAPPY nor other databases. This feature is realized by immediate annotating and BLAST search for database of HAPPY. In order to meet the demands performing high load bioinformatics tools in the future, we will develop HAPPY system on grid computing.

References

[1] Hoersch, S., Leroy, C., Brown, N.P., Andrade, M.A., and Sander, C., The GeneQuiz web server: protein functional analysis through the Web, *Trends Biochem. Sci.*, 25(1):33–5, 2000.

[2] Uchiyama, I., MBGD: microbial genome database for comparative analysis, *Nucleic Acids Res.*, 31(1):58–62, 2003.

[3] Yamazaki, K., Ohnishi, M., Kurokawa, K., and Yasunaga, T., ParalogCluster: classifying paralogs in a genome into paralogous groups, *Genome Informatics*, 12:409–410, 2001.

[4] `http://www.ncbi.nlm.nih.gov/RefSeq/`