Integration of Biological Features of Multiple Genomes - The Construction of "Genome Indices" database -

Shujiro Okuda okuda@kuicr.kyoto-u.ac.jp
Akiyasu C. Yoshizawa acyshzw@kuicr.kyoto-u.ac.jp
Yuki Moriya moriya@kuicr.kyoto-u.ac.jp
Masumi Itoh itoh@kuicr.kyoto-u.ac.jp
Susumu Goto goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan.

Keywords: genome, database, GC content, exon, intron, intergenic region, codon bias

1 Introduction

The increasing availability of complete and draft genomic sequences enables us to perform the large scale comparison of the genomes. From the viewpoint of comparative genomics, most of the interests have been focused on the coding regions, passing over the non-coding regions or the whole genomic sequences. This is because that functions of genes or proteins are conserved through the evolution on the nucleic and amino acid sequences. However, it has been reported that in the non-coding regions, intron size may be correlated to genome size within a taxon [1, 3, 4]. This suggests that the history of genome evolution is not necessarily dominated by changes of compositions of coding sequences [1]. Thus, the evolutionary relations between genomes and the elements comprising them have been argued for a long time. In order to consider the genome evolution, it is very important to organize the primary information about a variety of the elements in genomes.

Hence, we have collected genome sequence data and calculated the indicators of the genomic features. In the current version of our database, named "Genome Indices", we have measured basic statistical quantities on the biological elements of the multiple genomes.

This database is available at the web site: http://gi.kuicr.kyoto-u.ac.jp/

2 Method and Results

The sequences of genes and genomes were obtained from the KEGG database at GenomeNet [2]. The species we used were comprised of 17 eukaryotes, 18 archaea and 143 bacteria. As basic biological elements to characterize them, we focused on the genome, chromosome, gene, exon, intron, intergene and RNAs. Table 1 shows an example of the contents for *Arabidopsis thaliana*. In each element, we measured several statistical quantities; the number and the maximum, minimum, average and median lengths and GC contents. In addition, we also calculated the gene density and codon bias of each genome.

We have also implemented the data retrieval system, with which the users can choose the species by specifying the maximum or minimum value through the multiple genomes. For example, when we search the genome whose GC content is maximum of all species, we can find *Streptomyces coelicolor* whose GC content is 72.4 %. On the other hand, in the case of the minimum GC content, we can see it is 23.8 % of *Plasmodium falciparum*.

Table 1: An example of the contents for *Arabidopsis thaliana* in Genome Indices.

	number	maximum	minimum	average	median	GC content
Chromosome	5	30494425	18585042	23849671.8	23470805	36.0
CDS	28293	15468	60	1246.9	1062	44.2
exon	144245	7713	1	244.6	136	44.2
intron	115952	5026	1	162.0	99	32.6
intergene	28016	194151	6	3843.8	2764	35.0
RNA	618	1808	67	81.7	73	55.5
rRNA	4	1808	164	986.0	1808	49.6
tRNA	614	90	67	75.9	73	56.0

3 Discussion

We have developed a database storing a variety of biological features included in multiple genomes. In the current version, only the information of some features derived from their nucleic acid sequences is stored. From this point of view, our database is not enough amount of data to discuss the genome evolution. However, this is a novel database in that we can observe various biological features from coding sequences to non-coding sequences across multiple species. In the near future, we will add other features about protein domains, hyper secondary structure, morphology, pathways or orthology into this database. By performing the cluster analyses of these data, we will find the indices that support the topology of the universal phylogenetic tree. These analyses will play the important roles for our understanding of an evolutionary relationship between a genome and individual molecules coded in it.

4 Acknowledgments

This work was supported by 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] Gregory, T. R. Insertion-deletion biases and the evolution of genome size *Gene*, 324:15-34, 2004

[2] Kanehisa, M., Goto, S., Kawashima, S,. Okuno, Y,. and Hattori, M. The KEGG resource for deciphering the genome, *Nucleic. Acid. Res.*, 32:D277-280, 2004.

[3] Neafsey, D. E., and Palumbi, S., R. Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodonid pufferfish genomes *Genome Res.*, 13:821-830, 2003

[4] Wendel, J. F., Cronn, R. C., Alvarez, I., Small, R. L., and Senchina, D. S. *Mol. Biol. Evol.*, 19:2346-2352, 2002