2006－BIO－5　(19)
2006／6／16

マルチドメインを考慮したオーソログ遺伝子対の検出アルゴリズムの開発：植物と細菌の遺伝子の対応付けを目指す

新井美紗子，真保陽子，Md. Altaf-Ul-Amin，金谷重彦，黒川顕
奈良先端科学技術大学院大学
情報科学研究科　情報生命科学専攻
比較ゲノム学講座

様々な生物のゲノム全配列が解析され、異なる生物種間で全遺伝子を比較解析することでオーソログ遺伝子対を検出することが可能となった。しかしながら、遺伝子は進化の過程でドメインが融合したマルチドメイン構造を形成する場合があるため、ホモロジー解析で得られた遺伝子の類似箇所の情報を考慮していない単純なクラスタリングに基づく解析手法では、遺伝子の相同関係が一対多や多対多になってしまい、オーソログ遺伝子を決定する際に問題が生じる。この問題を解消するため、本研究では、遺伝子の相同関係をドメイン単位で検出し、オーソログ遺伝子をドメイン単位で解析することを可能にする、マルチドメインの関係性を可視化するソフトウェアの開発を行った。また、本ソフトウェアを用いた、植物と細菌の遺伝子のゲノムレベルでの進化解析を行った。

Developing algorithm for detecting Orthologue gene pairs with Multi-domain: for elucidating evolution from Bacteria to Plants

Misako Arai, Yoko Shinbo, Md. Altaf-Ul-Amin, Shigehiko Kanaya, Ken Kurokawa
Laboratory of Comparative Genomics, Graduate School of Information Science,
Department of Bioinformatics and Genomics, Nara Institute of Science and Technology

Recently, genome sequences have been determined for a wide variety of organisms. Orthologous genes can be detected by comparing the genomes of different organisms. Orthologous gene pairs have high sequence similarity to each other because they are vertically connected in the evolutionary tree. However, if orthologous gene is composed of multiple domains, the smallest unit of functional split in a gene, it is wise to try to understand gene functional relationships not only based on orthologous gene relation but also by giving emphasis to domain similarity. In the present study, we developed new algorithm focused on domain combination, and a tool for visualizing gene clusters based on domain similarity. We demonstrate performance of the algorithm and the tool using plant and bacteria genes.

Introduction

Recently, genome sequences have been determined for a wide variety of organisms. Comparison between different genomes, comparative genomics, is useful to predict the functions of unknown genes and proteins and to elucidate the evolutional history. During evolution, genes go through various changes to produce more complex proteins by nucleotide substitution, duplication, recombination, and so on.

The evolutionary units of genes such as domains are rearranged in recombination process and genes with similar domains are produced by the fusion and fission events. Existence of similar domains in different genes of different or same species can be described as one long composite gene in an organism or multiple splits of genes in another organism. The genes of higher organisms tended to have multi-domains and shuffled the domains during evolution, so we can find lots of one-to-many or many-to-many homologous genes including orthologous genes based on a simple homology search analysis.

It is difficult to understand the relationships between sequence similarity and gene function by the divergence of multi-domain genes. Therefore, we focus on the orthologous and paralogous relationships of not only genes but domains for understanding gene functions and elucidating gene evolution. The purpose of the present study is to solve the above one-to-many and many-to-many problems, because domains are the smallest units which are related to functions. There have been several tools and databases that detect and visualize the orthologous and paralogous relationships of multi-domain genes of bacteria [1-5], but these tools do not focus on plant-to-plant or plant-to-bacteria. Because the sequence homology between plant and bacteria genes is very low, it is difficult to detect the orthology among them. We developed a tool for visualizing multi-domain relationship between gene pairs in different or same species. In the present tool, input data are domain clusters that are clustered using the results of bi-directional BLAST [6] search with hit positions, and outputs are visualized multi-domain relationships based on the domain cluster. Using this tool, by dissolving many-to-many domain relationships to each one-to-many relationship, we can elucidate the all gene relationships under the one-to-one domain conditions. The rest of this paper is organized as follows. Section 1 explains the algorithm based on which we developed the tool and the genomic data we used in the present work. Section 2 demonstrates and discusses the results obtained by applying our tool to genomic data. Section 3 concludes the paper.

1. Algorithm

Figure 1 shows the procedure of the present algorithm. First, the homologous gene pairs are detected by bi-directional search of BLAST. Second, according to the result of bi-directional BLAST search, when the homologous regions of the high-scoring pairs (HSPs) are almost the same between each directional search, the HSPs are extracted as the paralogous or orthologous gene. Moreover, if the other region that is overlapped with the HSPs region also shows high homology, we automatically extended the homologous region. Third, the subject genes are clustered with the query gene based on domain sequence similarity. Finally, the cluster of genes is visualized.

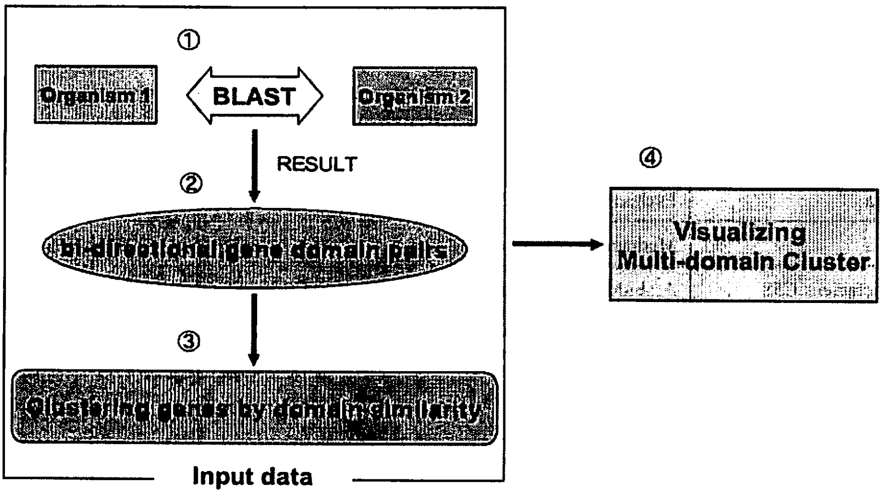


Figure 1: The procedure of the present algorithm

2. Demonstration

1. Data Set

The amino acid sequences of various organisms, whose genome sequences are completely known, are retrieved from NCBI RefSeq database [7]. We tried to demonstrate the performance of our tool by applying it to the genomes of the plant *Arabidopsis thaliana,* and bacteria *Agrobacterium tumefaciens* C58 Cereon, *Agrobacterium tumefaciens* C58 UWash, *Bacillus subtilis, Chlorobium chlorochromatii* CaD3, *Chlorobium tepidum* TLS, *Cyanobacteria* bacterium Yellowstone A-Prime, *Cyanobacteria* bacterium Yellowstone B-Prime, *Escherichia coli* K-12, *Methanococcus jannaschii.* **Table 1** describes the hit number of orthologue genes in one organism.

Table 1: The numbers and rates of orthologue genes in an organism by bi-directional BLAST search

Query	Subject	Query Gene Number	Hit Gene Number	Hit Rate (%)
Arabidopsis thaliana	Arabidopsis thaliana	26,536	*19,554	*73.68
Arabidopsis thaliana	Agrobacterium tumefaciens C58 Cereon	26,536	4,213	15.87
Arabidopsis thaliana	Agrobacterium tumefaciens C58 UWash	26,536	4,240	15.97
Arabidopsis thaliana	Bacillus subtilis	26,536	5,210	19.63
Arabidopsis thaliana	Chlorobium chlorochromatii CaD3	26,536	3,426	12.91
Arabidopsis thaliana	Chlorobium tepidum TLS	26,536	3,345	12.6
Arabidopsis thaliana	Cyanobacteria bacterium Yellowstone A-Prime	26,536	5,064	19.08
Arabidopsis thaliana	Cyanobacteria bacterium Yellowstone B-Prime	26,536	5,039	18.98
Arabidopsis thaliana	Escherichia coli K12	26,536	3,836	14.45
Arabidopsis thaliana	Methanococcus jannaschii	26,536	2,228	8.39
Agrobacterium tumefaciens C58 Cereon	Arabidopsis thaliana	5,288	1,810	34.22
Agrobacterium tumefaciens C58 UWash	Arabidopsis thaliana	5,402	1,816	33.61
Bacillus subtilis	Arabidopsis thaliana	4,105	1,466	35.71
Chlorobium chlorochromatii CaD3	Arabidopsis thaliana	2,002	789	39.41
Chlorobium tepidum TLS	Arabidopsis thaliana	2,252	858	38.09
Cyanobacteria bacterium Yellowstone A-Prime	Arabidopsis thaliana	2,760	1,170	42.39
Cyanobacteria bacterium Yellowstone B-Prime	Arabidopsis thaliana	2,862	1,182	41.29
Escherichia coli K12	Arabidopsis thaliana	4,237	1,449	34.19
Methanococcus jannaschii	Arabidopsis thaliana	1,786	588	32.92

* In the case of *Arabidopsis thaliana* vs. *Arabidopsis thaliana*, the number and the rate of hit genes are not included the results of self-self gene BLAST comparisons.

2. Demonstration

We describe the aspects of our tool, with an example. Let the query is *Agrobacterium tumefaciens* C58 Cereon and subject is *Arabidopsis thaliana.* **Figure 2** shows the gene clusters based on domain similarity and concerning statistical analysis. The input data is the domain based gene clusters determined by bi-directional BLAST search. The plot of Fig.2 (1) illustrates the density of matching of a query gene with the subject genes. Density is defined as the ratio of the number of one-to-one hit genes of the subject to the number of all subject genes. In the case of AGR_C_1966 of *Agrobacterium tumefaciens* C58 Cereon vs. *Arabidopsis thaliana*, there are 26,536 genes in *Arabidopsis thaliana*, and AGR_C_1966 matches with 107 genes of *Arabidopsis thaliana* as one-to-one relationship and therefore the density is 0.004032. The horizontal axis in Fig.2 (1) corresponds to the genes, according to the order of protein table file (*.ptt) from NCBI-RefSeq and the vertical axis represents the density. A red point in Fig. 2(1) implies that at least one subject gene shows 100% hit with the query gene while for a green point no such subject gene exists. Fig.2 (2) shows the distribution of red points of Fig.2 (1) with respect to density. Fig.2 (3) shows the rate of conservation determined by hit position of amino acid among the genes. Therefore, if certain parts of all one-to-one hit genes of the subject are matched with certain parts of the query gene, these parts are 100% conserved area. The relationships of query gene (red) and subject genes (blue) in the context of hit position are visualized in Fig.2 (4).

We should note that the highest conserved domain area is useful to predict domain function and elucidate gene evolution in different genomes. **Figure 3** shows the result of gene, AGR_C_1966 of *Agrobacterium tumefaciens* C58 Cereon vs. *Arabidopsis thaliana.* The length of AGR_C_1966 is 629 amino acids. The highest conserved area is 391~545aa which has ~100% matching with one-to-one hit subject genes. We confirmed that this area is a functional domain by Position-specific iterated and pattern-hit initiated BLAST [8] (PSI-BLAST [9] and PHI-BLAST [10]) as shown in **Figure 4**, and by InterProScan [11-12] as shown in **Figure 5**. This domain has functions related to ABC transporter and AAA ATPase.

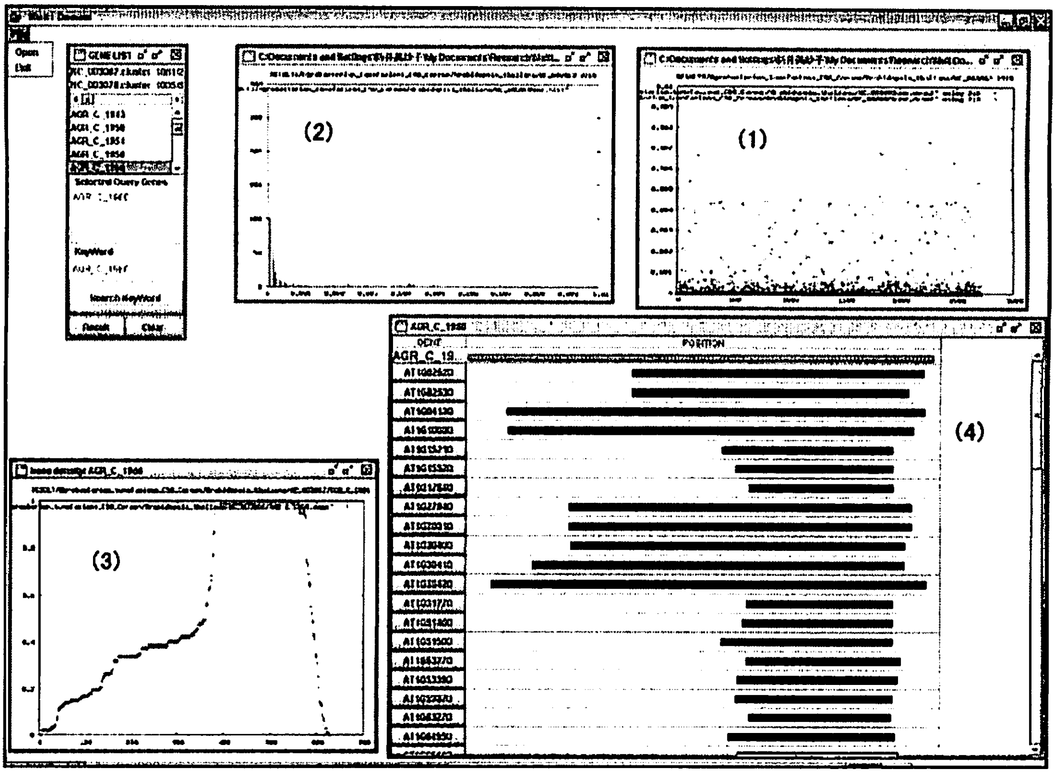


Figure 2: The tool of visualizing the relationships of multi-domain, and statistics graphs in one genome and in one query gene.

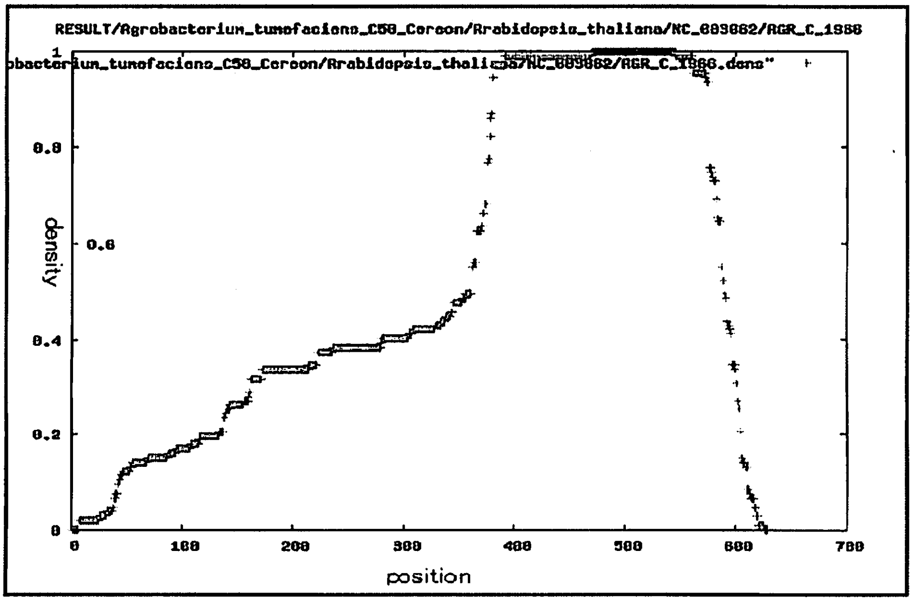
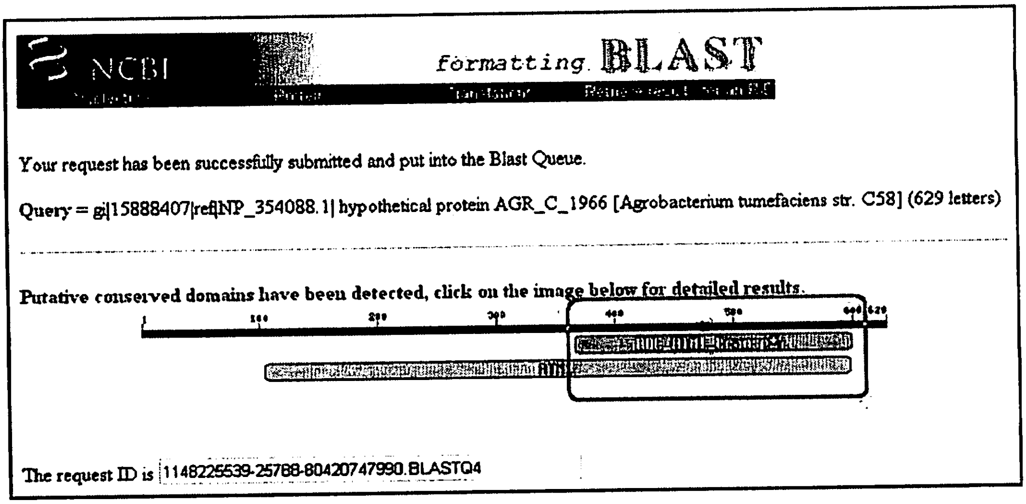


Figure 3: The conserved density graph of AGR_C_1966.



Your request has been successfully submitted and put into the Blast Queue.

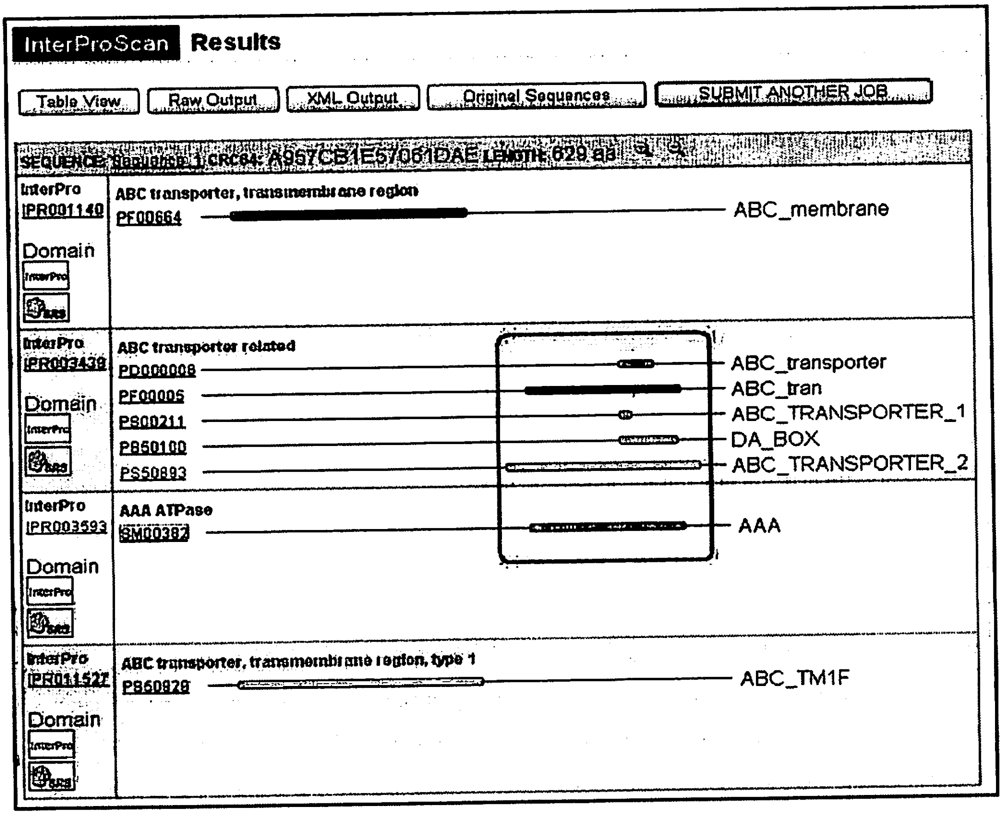
Query = gi|15888407|ref|NP_354088.1| hypothetical protein AGR_C_1966 [Agrobacterium tumefaciens str. C58] (629 letters)

Putative conserved domains have been detected, click on the image below for detailed results.

The request ID is 1148225539-25788-80420747990.BLASTQ4

Figure 4: The result of PSI-BLAST and PHI-BLAST of AGR_C_1966.

http://www.ncbi.nlm.nih.gov/BLAST/



SEQUENCE: Sequence_1 CRC64: A957CB1E57061DAE LENGTH: 629 aa

InterPro IPR001140 Domain
ABC transporter, transmembrane region
PF00664 ABC_membrane

InterPro IPR003439 Domain
ABC transporter related
PD000006 ABC_transporter
PF00005 ABC_tran
PS00211 ABC_TRANSPORTER_1
PS50100 DA_BOX
PS50893 ABC_TRANSPORTER_2

InterPro IPR003593 Domain
AAA ATPase
SM00382 AAA

InterPro IPR011527 Domain
ABC transporter, transmembrane region, type 1
PS50929 ABC_TM1F

Figure 5: The result of InterProScan of AGR_C_1966.

http://www.ebi.ac.uk/InterProScan/

3. Conclusion

We developed the algorithm for detecting similar domain area between genes by bi-directional BLAST search in the context of hit position and no overlapping, and the software tool we developed makes it possible to visualizing gene clusters based on domain similarity. This tool is available for different species, for example, plants vs. bacteria. We demonstrated the performance of this tool by applying it to *Arabidopsis thaliana vs. Agrobacterium tumefaciens* C58 Cereon. Conserved area obtained by the present system is consistent with those in multi-domain database for bacteria such as PSI-BLAST and PHI-BLAST and InterProScan. In future, we plan to create database for multi-domain, and connect it to our tool.

References:

[1] Tatusov, R. L., Koonin, E.V. and Lipman, D. J. (1997) A Genomic Perspective on Protein Families. *Science*, **278**, 631-637.

[2] Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B.S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. and Koonin, E.V. (2001), The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 41-43.

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

[4] Enright, A. J., Kunin, V. and Ouzounis, C. A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.*, **31**, 4632-4638.

[5] Suhre, K. and Claverie, J. M. (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res.*, **32**, D273-D276.

[6] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol.* **215**, 403-410.

[7] NCBI from ftp://ftp.ncbi.nih.gov/refseq/ on May 9, 2006

[8] PSI-BLAST and PHI-BLAST http://www.ncbi.nlm.nih.gov/BLAST/

[9] Altschhul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J. Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

[10] Zhang, Z., Schäffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. and Altschul, S. F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986-3990.

[11] InterProScan http://www.ebi.ac.uk/InterProScan/

[12] Zdobnov, E. V. and Apweiler, R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847-848.