BMC Bioinformatics

Database

FlyPhy: a phylogenomic analysis platform for *Drosophila* genes and gene families

Jinyu Wu†1, Xiang Xu†2, Jian Xiao2, Long Xu2, Huiguang Yi1, Shengjie Gao1, Jing Liu1, Qiyu Bao*1, Fangqing Zhao*3 and Xiaokun Li*2

Address: 1Institute of Biomedical Informatics/Zhejiang Provincial Key Laboratory of Medical Genetics, Wenzhou Medical College, Wenzhou 325000, PR China, 2School of Pharmaceutical Science/Zhejiang Provincial Key Laboratory of Biotechnology Pharmaceutical Engineering, Wenzhou Medical College, Wenzhou 325035, PR China and 3Department of Biochemistry and Molecular Biology, Pennsylvania State University, Pennsylvania 16802, USA

Email: Jinyu Wu - iamwujy@yahoo.com.cn; Xiang Xu - xxiang0577@126.com; Jian Xiao - xfxj2000@126.com; Long Xu - xulonggood@163.com; Huiguang Yi - yihg926@126.com; Shengjie Gao - gsjluck@163.com; Jing Liu - liuj0715ql@yahoo.com.cn; Qiyu Bao* - baoqywzmc@yahoo.com.cn; Fangqing Zhao* - fuz3@psu.edu; Xiaokun Li* - xiaokunli@163.net

* Corresponding authors †Equal contributors

Published: 25 April 2009

BMC Bioinformatics 2009, **10**:123 doi:10.1186/1471-2105-10-123

Received: 5 January 2009
Accepted: 25 April 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/123

Abstract

Background: The availability of 12 fully sequenced *Drosophila* species genomes provides an excellent opportunity to explore the evolutionary mechanism, structure and function of gene families in *Drosophila*. Currently, several important resources, such as FlyBase, FlyMine and DroSpeGe, have been devoted to integrating genetic, genomic, and functional data of *Drosophila* into a well-organized form. However, all of these resources are gene-centric and lack the information of the gene families in *Drosophila*.

Description: FlyPhy is a comprehensive phylogenomic analysis platform devoted to analyzing the genes and gene families in *Drosophila*. Genes were classified into families using a graph-based Markov Clustering algorithm and extensively annotated by a number of bioinformatic tools, such as basic sequence features, functional category, gene ontology terms, domain organization and sequence homolog to other databases. FlyPhy provides a simple and user-friendly web interface to allow users to browse and retrieve the information at multiple levels. An outstanding feature of the FlyPhy is that all the retrieved results can be added to a workset for further data manipulation. For the data stored in the workset, multiple sequence alignment, phylogenetic tree construction and visualization can be easily performed to investigate the sequence variation of each given family and to explore its evolutionary mechanism.

Conclusion: With the above functionalities, FlyPhy will be a useful resource and convenient platform for the *Drosophila* research community. The FlyPhy is available at http://bioinformatics.zj.cn/fly/.

Background

Fruit flies have been studied for many years and one species of them, in particular *D. melanogaster*, is a very important model organism for understanding genetic, developmental, cellular, ecological, and evolutionary processes. The sequencing of *D. melanogaster* and *D. pseu-*

doobscura genome, first resealed in 2000 and 2005, respectively, provide significant contributions to the fruit fly biology and genome research [1,2]. With ever-developing large-scale sequencing technologies, 12 *Drosophila* genomes are available and accessible online now [3,4]. The availability of these *Drosophila* genomes offers an unprecedented opportunity to explore the evolution of *Drosophila* gene families, which can serve as a significant base for functional genomics and provide an important advance for understanding sequence-structure-function relationships of *Drosophila* genes among different species. For example, comparative genomics analysis revealed that there was a high-frequency occurrence of gene gain and loss in *Drosophila* gene families, even among closely related *Drosophila* species [5]. Genome-wide comparison of immune-system genes in *Drosophila* revealed that, in contrast to signaling proteins, effector proteins are much more likely to vary in copy number across different *Drosophila* species [6]. Based on phylogenomic approaches, the possible reason for the origin of new genes and subsequent lineage-specific evolution at different time nodes in the *Drosophila* is well revealed from a genome-wide level [7].

Development of effective and integrated bioinformatics databases and tools is an important work for facilitating more rapid progress in *Drosophila* research, which will provide a convenient aid to *Drosophila* research communities. In support of this, several databases have been devoted to *Drosophila* in a well-organized form. FlyBase is a premier public database with integrated genetic, genomic, and functional data of *Drosophila*[8]. FlyMine is a comprehensive database with gene expression data of *Drosophila* [9] and DroSpeGe is a genome database with comparative annotations of 12 *Drosophila* species [3]. Other resources such as Berkeley Drosophila Genome Project [10] and AAA [11] are also useful resources for *Drosophila* biologists. However, all of these available resources are gene-centric. The Dfam database contains descriptions of the families, alignments, gene trees [5], but there is no integrated database to provide comprehensive information on gene families of *Drosophila*. Comparative genomics and molecular evolution analysis of *Drosophila* gene families has been demonstrated to be a powerful approach to study their evolution, structure and function. In this study, we applied a graph-based Markov Clustering algorithm to classify all the *Drosophila* proteins into families. Thereby, a comprehensive platform containing putative protein families with extensive annotation information of 12 fruit fly species was developed. Users can easily interact with the protein families of their interest and other relevant detail annotations of genes by browsing, keyword searching or BLASTing. Through the workset, the retrieved data can be well integrated for phylogenomic analysis.

Construction and content

Protein family clustering and detailed annotation

Protein families can be expressed as a group of proteins that share significant similarity in sequence and have a common evolutionary history. In comparison with other tools for grouping different proteins into putative families, such as BLASTClust (from the NCBI BLAST suite) and cd-hit program [12], the TribeMCL [13] has been proved to be a good alternative choice for clustering of divergent proteins and thus has wide applications [14-16]. The TribeMCL program first calculates the pair-wise distances between all genes from the genomes and then uses a graph-based Markov Clustering algorithm to generate clusters. With such algorithm, it can effectively break the barriers during the clustering process, such as multi-domains, fragments of proteins and promiscuous domains in the alignment. The main parameter that influences the number and size of clusters in the TribeMCL program is the inflation value, which defines the tightness of the clustering results [13]. In this study, we used the TribeMCL program to generate protein families from 181, 780 protein sequences of 12 *Drosophila* species [17]. An all-against-all BLASTP was performed for the protein sequences using the BLAST program with an E-value of 1e-5. Finally, putative protein families were generated using the TribeMCL program with multiple inflation values (1.5, 2.5, 3.0, 4.0 and 5.0 respectively). We found that different inflation parameters produced similar results and the cluster size varied greatly in different families (Figure 1), indicating the frequent gene gain-and-loss in *Drosophila*. As expected, a large number of clusters with a family size of 12 were observed and the majority of them represented families with only one gene in each species. For example, among the 3,454 families (family size = 12) obtained from the inflation value 3, a total of 3,055 families had a one-to-one orthologous relationship. It is suggested that these families may represent a core set of genes and universally present in different *Drosophila* species.

Then these genes and gene families were extensively annotated based on a number of bioinformatic tools and databases. In particular, the PepStat program implemented in the EMBOSS package was used to predict the molecular weight and isoelectric point of a given protein [18]. The InterProScan program was used to assign gene domain architectures against integrated databases, including PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIRSF and SUPERFAMILY. The InterProScan results were mapped to Gene Ontology terms, including cellular component, biological process and molecular function, using the InterPro2Go [19]. BLAST searches were performed against several major databases, such as PDB (collected on 18 December 2008), Uniprot (release 14.6) and Refseq (release 32). The best hit from the Uniprot database was used as a controlled vocabulary for the description of gene

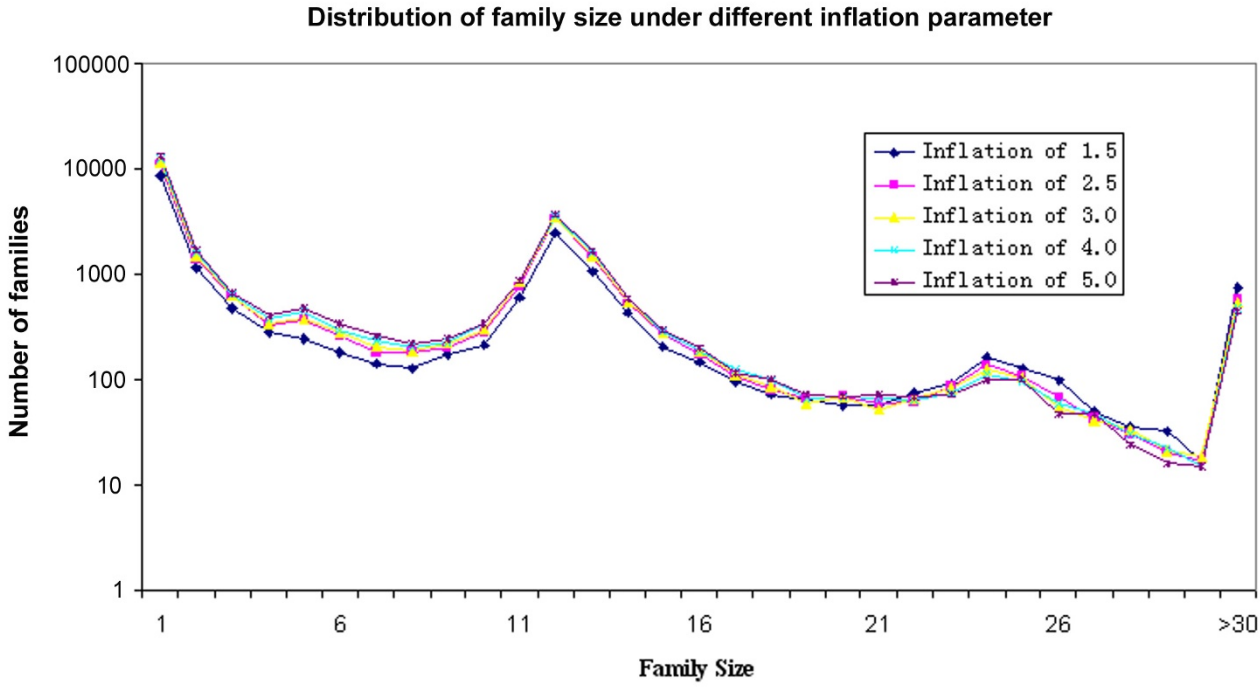

Figure 1
The distribution of family size under different inflation parameter of 1.5, 2.5, 3.0, 4.0 and 5.0 in the TribeMCL program.

function. In addition, the functional categorization of all genes and protein families was carried out by BLASTing to the COG and KEGG databases with an E-value of 1e-5. In order to annotate each cluster, we obtained the annotation information in KEGG pathway related to each member in the clusters, and then manually curated the most common description of the members in each cluster and assign them to the clusters.

Database construction

The design scheme of FlyPhy is similar to our previous integrated pipeline of ArchaeaTF [20] and, PlasmoGF [15], which is constructed based on open source software, including Apache, MySQL, PHP and Perl, etc. The curated data of gene families, as well as various annotation information, are stored in a MySQL database system and can be accessed using Structure Query Language (SQL). The web platform is base on Apache HTTP server and its pages are generated via a combination of PHP language and Perl CGI scripts. Meanwhile, the BioPerl modules are applied to manipulate data and convert different data formats. All the procedures above are executed on the Linux operating system.

Utility and discussion

Data retrieving

FlyPhy provides a simple and user-friendly interface for researchers to access gene and gene family data (Figure 2). Users can browse and search (both keyword search and sequence similarity search) all the data at different levels by the integrated functions.

Browse

all the genes have been organized into different functional categories according to the COG and KEGG database. Individual COG category can be browsed easily in the COG browser as well as the list of genes classified under each specific COG category. The KEGG pathway can be explored in the same way in the KEGG browser. Clicking on the gene ID will show its detailed annotation information, such as basic sequence features, Gene Ontology terms, gene domain organization and sequence homolog to other relevant databases.

Keyword search

FlyPhy provides a powerful multi-layered query system. Firstly, in the search page, users can search genes by keywords (gene ID or gene definition), or by clusters under different inflation parameters (eg: cluster ID, cluster size and cluster definition). Meanwhile, the functional catego-

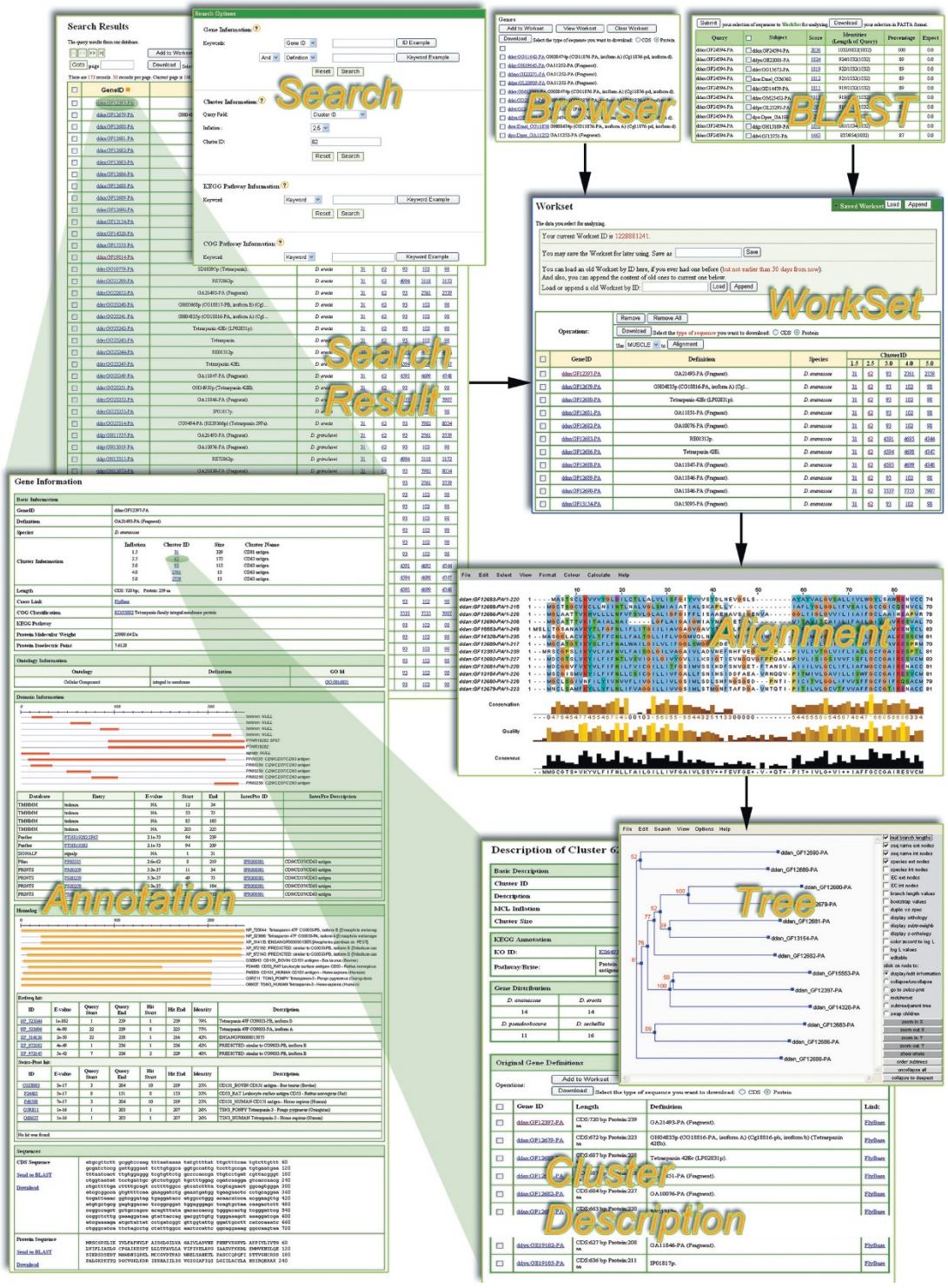

Figure 2

The snapshot indicating the interrelationship of FlyPhy data and tools. The annotated information can be retrieved at multiple levels: keyword search, browse and BLASTing. The retrieved results will be shown in a table with the basic information of each gene or gene family matching the query and further linked to the detailed annotation of the gene and gene family also is also provided. Meanwhile, users can construct their own workset to investigate the sequence variation of each given family and explore its evolutionary mechanism.

ries of different gene family can also be retrieved by KEGG or COG ID and keyword. The search results will be shown in table with the basic information of each gene matching the query. In the table, the IDs of each gene and its clusters are linked to the detailed annotation of the gene and clusters.

Sequence similarity search

We have generated genes and protein databases for all the families, and enabled protein or nucleotide BLAST in FlyPhy. This facility will help users query and verify the members of a specific gene family based on their own sequences. Further, with the implement of the ViroBLAST program [21], the sequence similarity search of FlyPhy provides many advanced options to allow users to easily parse and manipulate the search results.

Workset-centric data manipulation and phylogenomic analysis

An important functionality of FlyPhy is that it adopts the workset to organize the genes and protein families (Figure 2). All the retrieved results can be added to the workset for further data manipulation and phylogenomic analysis. Each workset is assigned with a specific ID either generated by the server randomly or saved as a user's own favorite name. All the data in the workset can be customized through appending or deleting the items. More importantly, FlyPhy allows users to load an old workset by ID if they ever had established before to avoid generating the same workset from scratch.

To explore the sequence conservation of the data stored in the workset, multiple sequence alignment can be performed using either the ClustalW [22] or MUSCLE program [23] with user-definable parameters at amino acid or DNA level. Visualization of the aligned results is conducted by the Jalview program [24], which is based on Java Applet (prior to use this program, a Java Runtime Environment is needed on the local computer). To investigate the evolutionary relationship of genes or proteins stored in the workset, users can directly use the QuickTree program to construct a phylogenetic tree [25], which is based on the neighbor-joining algorithm. The reliability of the tree can be evaluated with different replicates of bootstrapping test. The graphical representation of the inferred tree is carried out using the ATV program [26].

Conclusion

The purpose of constructing FlyPhy is to develop a comprehensive platform on which users can access detailed annotation information, investigate the sequence variation and explore the evolutionary mechanism of each family in *Drosophila*. Through the form of workset, the retrieved data are well integrated, and phylogenomic analysis can be easily performed. In the future, we will continue to collect relevant fly data once fully sequenced genome of other *Drosophila* species is available. More phylogenomic tools will be incorporated to help users reconstruct the evolutionary history of gene families, such as the TREE-PUZZLE program for phylogenetic tree construction based on maximum likelihood algorithm and the PAML program for evolutionary rate inference. We believe that FlyPhy will serve as a useful platform for *Drosophila* biologists and relevant researchers to study the comparative genomics and phylogenomics of fly gene families efficiently and conveniently.

Availability and requirements

Project name: FlyPhy: a Phylogenomic Analysis Platform for *Drosophila* Genes and Gene Families

Project home page: http://bioinformatics.zj.cn/fly/

Operating system(s)

For user: Standard WWW browser, such as Firefox3.0, Internet Explorer7.0 and Safari3.1

For server: Linux

Programming language: PHP, MySQL, Perl and BioPerl

License: GNU GPL

Any restrictions to use by non-academics: None

Authors' contributions

JW and XX performed bioinformatics analysis, constructed the database, developed the web interface, and wrote the draft manuscript. JX provided scientific suggestions and criticisms for improving the manuscript and website. LX, HY, SG and JL participated in the data analysis and the update of the database. XL, FZ and QB participated in its design, helped write the manuscript and supervised the whole project. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (30800643) and Foundation of Zhejiang Provincial Top Key Discipline of Laboratory Medicine, China.

References

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of Drosophila melanogaster.** *Science* 2000, **287(5461):**2185-2195.
2. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, *et al.*: **Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution.** *Genome Res* 2005, **15(1):**1-18.
3. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, *et al.*: **Evolution of**

genes and genomes on the Drosophila phylogeny. *Nature* 2007, **450(7167):**203-218.
4. Ledford H: **Attack of the genomes.** *Nature* 2007, **450(7167):**142-143.
5. Hahn MW, Han MV, Han SG: **Gene family evolution across 12 Drosophila genomes.** *PLoS Genet* 2007, **3(11):**e197.
6. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG: **Dynamic evolution of the innate immune system in Drosophila.** *Nat Genet* 2007, **39(12):**1461-1468.
7. Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W: **On the origin of new genes in Drosophila.** *Genome Res* 2008, **18(9):**1446-1455.
8. Drysdale R: **FlyBase: a database for the Drosophila research community.** *Methods Mol Biol* 2008, **420:**45-59.
9. Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guillier F, Janssens H, Ji W, McLaren P, North P, *et al.*: **FlyMine: an integrated database for Drosophila and Anopheles genomics.** *Genome Biol* 2007, **8(7):**R129.
10. **Berkeley Drosophila Genome Project** [http://www.fruitfly.org/]
11. **AAA** [http://rana.lbl.gov/drosophila/]
12. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22(13):**1658-1659.
13. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13(9):**2178-2189.
14. Conte MG, Gaillard S, Lanau N, Rouard M, Perin C: **GreenPhylDB: a database for plant comparative genomics.** *Nucleic Acids Res* 2007:D991-998.
15. Xu X, Wu J, Xiao J, Tan Y, Bao Q, Zhao F, Li X: **PlasmoGF: an integrated system for comparative genomics and phylogenetic analysis of Plasmodium gene families.** *Bioinformatics* 2008, **24(9):**1217-1220.
16. Wall PK, Leebens-Mack J, Muller KF, Field D, Altman NS, dePamphilis CW: **PlantTribes: a gene and gene family resource for comparative genomics in plants.** *Nucleic Acids Res* 2008:D970-976.
17. **12 Drosophila species** [ftp://ftp.genome.jp/pub/kegg/genes/organisms]
18. Olson SA: **EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite.** *Brief Bioinform* 2002, **3(1):**87-91.
19. **InterPro2Go** [http://www.geneontology.org/external2go/interpro2go]
20. Wu J, Wang S, Bai J, Shi L, Li D, Xu Z, Niu Y, Lu J, Bao Q: **ArchaeaTF: an integrated database of putative transcription factors in Archaea.** *Genomics* 2008, **91(1):**102-107.
21. Deng W, Nickle DC, Learn GH, Maust B, Mullins JI: **ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets.** *Bioinformatics* 2007, **23(17):**2334-2336.
22. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22):**4673-4680.
23. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5):**1792-1797.
24. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor.** *Bioinformatics* 2004, **20(3):**426-427.
25. Howe K, Bateman A, Durbin R: **QuickTree: building huge Neighbour-Joining trees of protein sequences.** *Bioinformatics* 2002, **18(11):**1546-1547.
26. Zmasek CM, Eddy SR: **ATV: display and manipulation of annotated phylogenetic trees.** *Bioinformatics* 2001, **17(4):**383-384.