Journal of Bioinformatics and Sequence Analysis Vol. 1(3), pp. 046-049, October, 2009
Available online at http://www.academicjournals.org/jbsa

Full Length Research Paper

Bioinfotracker: A novel system for advanced genome functional insight

Gopal Ramesh Kumar[1]*, Ganesan Aravindhan[1], Thankaswamy Kosalai Subazini[1] and Radhakrishnan Sathish Kumar[2]

[1]Bioinformatics Lab, AU-KBC Research Centre, MIT Campus, Anna University, Chennai-600 044, India.
[2]NRCFOSS, AU-KBC Research Centre, MIT Campus, Anna University, Chennai-600 044, India.

Accepted 15 July, 2009

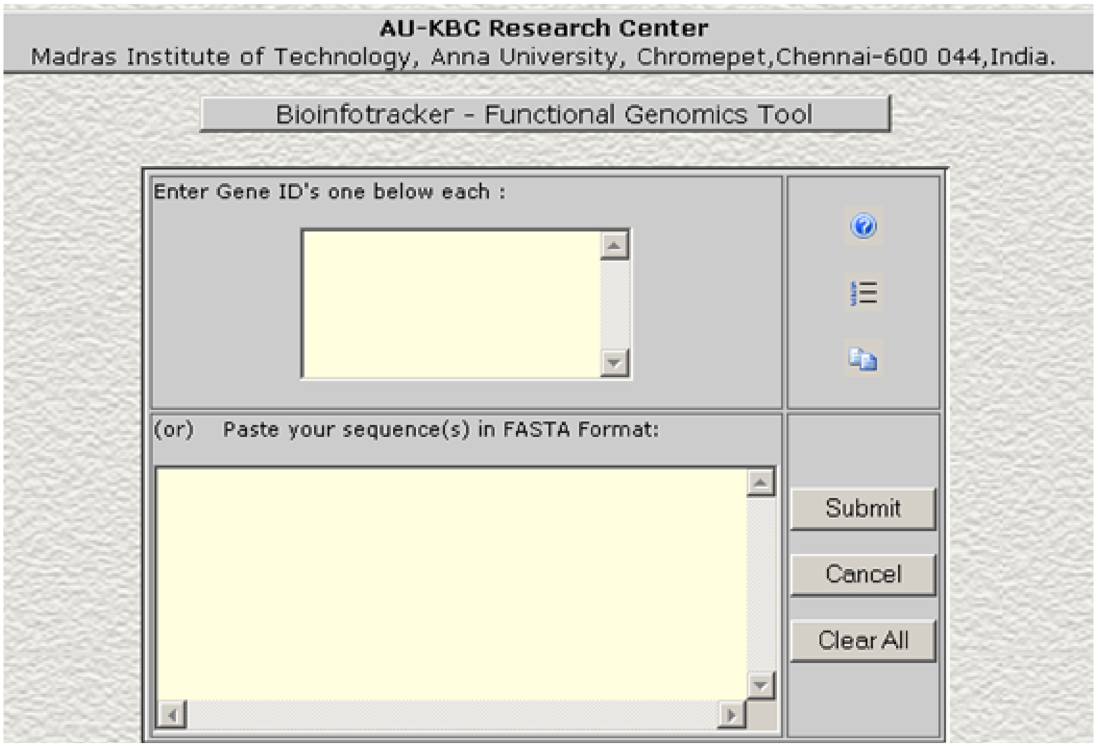
With the accelerated accumulation of genomic sequence data in the World Wide Web, it has become highly essential to understand the role of these sequences in the biological systems by incorporating various advanced research archetypes. The intricacy of handling such a huge dataset manually has increased the need to develop automated methods that can analyze enormous numbers of biological sequences and produce efficient results. This being the objective, a novel computational system, Bioinfotracker, has been developed for the purpose of carrying out large-scale protein annotations. Different online tools operating on different strategies have been integrated in Bioinfotracker so as reduce the overall processing time of these tools individually. Further, Bioinfotracker facilitates automatic parsing of the results from all the tools and produce them in an easily interpretable table format. This facility will, therefore, greatly lessen the burden of hectic human parsing. Moreover, AJAX (Asynchronous JavaScript and XML) is used as an interface within this tool that will greatly control the unwanted page refresh menace and bandwidth consumption. Thus, Bioinfotracker remains a well structured, species-independent, flexible and highly controlled functional analysis system for the protein sequences of any organism. The software is freely available at: http://biotool.nrcfosshelpline.in/.

Key words: Bioinfotracker, ajax, functional genomics, annotation, bandwidth, fasta.

INTRODUCTION

During the past decade, huge volumes of biological data have been generated and are deposited in the online repositories (Kim et al., 2003; Sasson et al., 2006). With this largely mounted data, it has become the most vital challenge for the research community to investigate these raw sequences and reveal their functions. Delineating the functions of the genome will facilitate a better insight into the biological systems (Rentzsch et al., 2009). In spite of various strategies for identifying the protein functions were carried out earlier, only 50 - 60% of genes have been identified with known functions in most of the completely sequenced genomes (Sivashankari et al., 2003). Therefore, the determination of protein functions has become the most focused research area of the post-genome era. The classical approaches for the functional genomics use different types of high-throughput techniques to characterize the actual gene products. Though these traditional biochemical/molecular experiments can assign accurate functions for the genes, they consume a lot of chemicals, reagents and other materials and thus making them more cost ineffective (Diana, 2003). Above all, these methodologies involve much of the manpower and the man-hours. This demands the use of Bioinformatics automated systems to carry out sequence analysis with the perspective of functional prediction. Recent years have seen tremendous growth in the Bioinformatics tools and approaches in genome analysis. They help in investigating the large quantity of data available and propose biologically meaningful patterns for the genes. The general Bioinformatics-led approach for functional characterization of proteins involves the comparison of the unknown sequences against the known sequences in various databases using a variety of tools. These tools are supported by a number of algorithms and statistical

*Corresponding author. E-mail: gramesh@au-kbc.org. Tel: 91-44-2223 2711/6959. Fax: 91-44-2223 1034.


AU-KBC Research Center
Madras Institute of Technology, Anna University, Chromepet,Chennai-600 044,India.

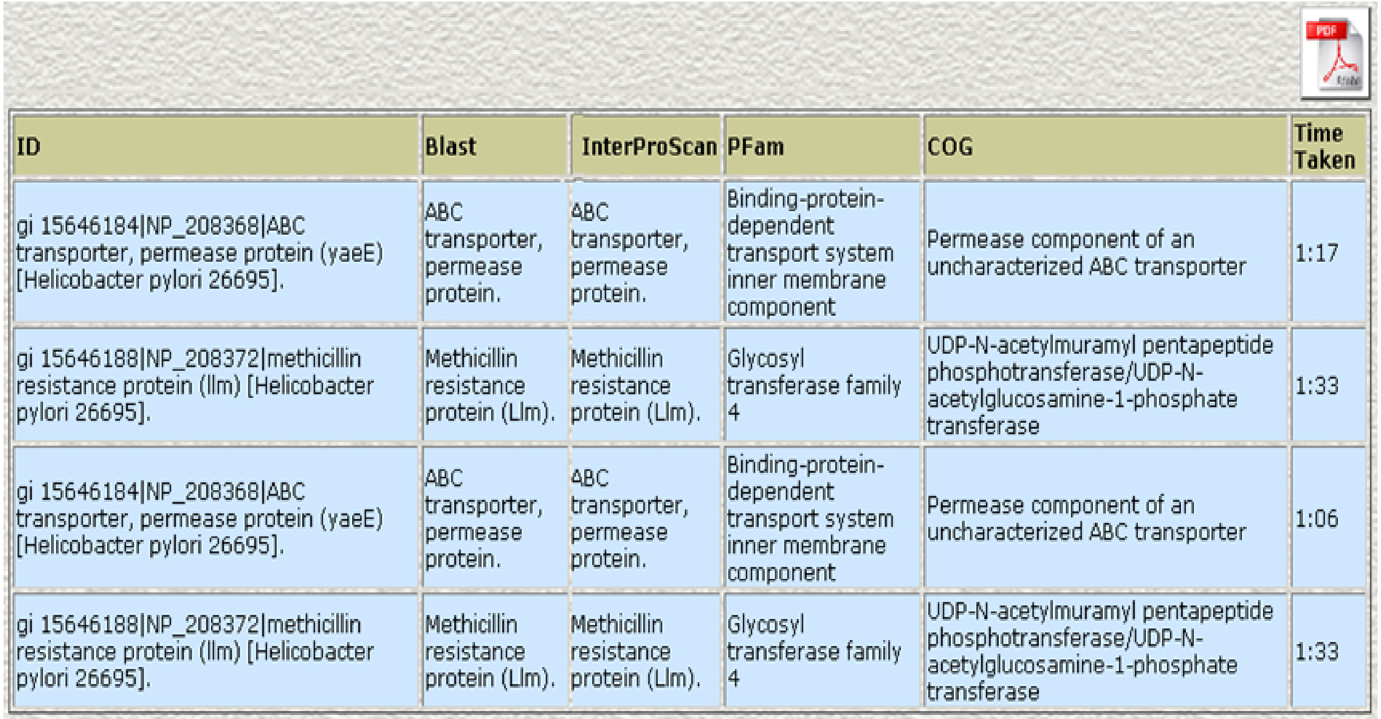
Bioinfotracker - Functional Genomics Tool
Enter Gene ID's one below each :
(or) Paste your sequence(s) in FASTA Format:
Submit
Cancel
Clear All

Figure 1. Homepage of Bioinfotracker which shows fields for entering GI number and sequences.

theories and predict the appropriate functions. In several cases, such predictions are proved to be efficient and this has led to the development of diverse Insilico protocols for the functional annotations of the proteins. Of various Insilico strategies, functional predictions using the tools that operate on the classification of proteins provide promising results. Presently, a number of different classification systems have been developed and deployed to categorize the functional annotations (Stuart et al., 2000) that include (i) BLAST (Altschul et al., 1990), a tool that helps in the sequence similarity searches (ii) Pfam (Bateman et al., 2004), a tool that is based on protein families (iii) COG (Tatusov et al., 2003) that represents phylogenetic classification of proteins (iv) Prodom, a tool that assists in protein domain searches (v) InterPro (Mulder et al., 2007), a tool integrated with different family classifications. Though functional predictions through computational programs have led to many scientific discoveries (Cathy et al., 2003), they tend to be complex as these applications are computationally intensive and time consuming. Moreover, a lot of human interventions are needed to carry out the analysis with these tools and to manually curate the results to identify potential functions. Although, there are a few tools like AIM-BLAST, Ajax Interfaced Multiple Sequence - BLAST (Aravindhan et al., 2009), that allows the users to analyse multiple sequences at an instance, their functional prediction is based on only one strategy. Hence, there is a pressing need to develop an advanced computational method that will balance these limitations and handle functional annotations better.

METHODS

Progress in computational power and the advancements in Bioinformatics research permit the integration of the available information from various sources into single qualitative models, thus making the analysis simple (Ruepp et al., 2004; Lobley et al., 2008). Here, we have developed a simple and efficient tool, Bioinfotracker (Figure 1), for carrying out large-scale protein annotation. Bioinfotracker is a system that was developed by integrating different Bioinformatics tools such as Pfam (http://pfam.jouy.inra.fr/), BLAST (http://www.ebi.ac.uk/Tools/BLAST/), COG (http://www.ncbi.nlm.nih.gov/COG/) and InterproScan (http://www.ebi.ac.uk/Tools/InterProScan/). The front end of the tool was written using HTML/Java scripts whereas the server end of the tool is coded using Perl scripts. Moreover, AJAX (Paulson, 2005) was used as an interface in Bioinfotracker that will greatly reduce the unpleasing page refresh issue that is very common in other bioinformatics tools. Hence, Bioinfotracker will consume very low bandwidth but still performs effectively. This tool makes it possible to perform the annotation of an entire genome using four different annotation strategies with only a single submit. The input for this tool can either be protein sequences in FASTA format or GI numbers of the sequences. If the GI numbers are submitted, the tool will automatically search in the NCBI database (http://www.ncbi.nlm.nih.gov) and fetch the sequences corresponding to the GI numbers submitted and then starts the analysis. If sequences are submitted, then the processing starts immediately without any delay in time. The Sequences are individually submitted to different servers, such as BLAST, Pfam, COG and InterProScan, and the analyses are carried out in the respective servers. Once the



ID	Blast	InterProScan	PFam	COG	Time Taken
gi 15646184|NP_208368|ABC transporter, permease protein (yaeE) [Helicobacter pylori 26695].	ABC transporter, permease protein.	ABC transporter, permease protein.	Binding-protein-dependent transport system inner membrane component	Permease component of an uncharacterized ABC transporter	1:17
gi 15646188|NP_208372|methicillin resistance protein (llm) [Helicobacter pylori 26695].	Methicillin resistance protein (Llm).	Methicillin resistance protein (Llm).	Glycosyl transferase family 4	UDP-N-acetylmuramyl pentapeptide phosphotransferase/UDP-N-acetylglucosamine-1-phosphate transferase	1:33
gi 15646184|NP_208368|ABC transporter, permease protein (yaeE) [Helicobacter pylori 26695].	ABC transporter, permease protein.	ABC transporter, permease protein.	Binding-protein-dependent transport system inner membrane component	Permease component of an uncharacterized ABC transporter	1:06
gi 15646188|NP_208372|methicillin resistance protein (llm) [Helicobacter pylori 26695].	Methicillin resistance protein (Llm).	Methicillin resistance protein (Llm).	Glycosyl transferase family 4	UDP-N-acetylmuramyl pentapeptide phosphotransferase/UDP-N-acetylglucosamine-1-phosphate transferase	1:33

Figure 2. Sample result table in Bioinfotracker of BLAST, COG, Pfam and InterProScan output with time taken to get the result.

results of the analyses are available, Bioinfotracker will automatically parse them and filter out the appropriate function from each server and display them in a simple table, where the automatic parsing is based on the technical filtering process carried out by the tool which is explained in the efficiency part. Bioinfotracker utilizes the SOAP (Pillai et al., 2005) web services of EMBL-EBI, (European Molecular Biology Laboratory-European Bioinformatics Institute) to fetch the results from the BLAST server and the InterProScan Server. Whereas, LWP::Simple and HTML:: TreeBuilder:: Xpath modules are used to fetch the results from the Pfam server and COG Server. Thus in this tool, the results of the analyses will be produced in a simple and easily interpretable table format that displays the ID of the sequence submitted, the results from Pfam, COG, InterProScan , BLAST and the time taken for each analysis (Figure 2). There is also an option that comes with the tool to save the results of the analysis in the PDF format. With all these features, Bioinfotracker remains user-friendly.

Efficiency of bioinfotracker

Except for the BLAST program, parsing the output of all other tools is straightforward and simple. Although Bioinfotracker is found to be efficient in carrying out searches in all the tools integrated within the program, it is worth elaborating its strength in handling the BLAST output parsing. Since, searching a single sequence against a regular BLAST program (http://www.ebi.ac.uk/Tools/BLAST/), will itself generate large amount of results in terms of hits accompanied with varied parameters such as E-value, Percentage of Identity, Percentage of Similarity, BLAST score and sequence length. Interpreting, analyzing and filtering such a voluminous BLAST textual output manually to select an appropriate hit, remains a great problem with the scientific community (Aravindhan et al., 2009a, b).

To bypass such difficulties, Bioinfotracker is incorporated with some special filtering processes that can expertly handle the voluminous BLAST results of the sequences and select one best hit for one sequence. The filtering process is performed in two parts. The first part of filtering is carried out to choose the BLAST hits that satisfy the values of all the parameters including BLAST score, the length and orientation of the hits, the percentage identity, percentage similarity and E-values. The second part of the process involves the further cleaning of the functions with any negative terms, functions that do not have any clear scientific evidence, such as predicted, putative, probable, hypothetical, conserved hypothetical and unknown. This filtering process of results, in Bioinfotracker, remains a powerful means of reducing the possibility of errors while choosing a single suitable function from mass of BLAST hits.

RESULTS AND DISCUSSION

The performance of the tool has been compared with the regular online tools using the Firefox Web browser. A sample set of sequences of varying length from *E. coli* were simultaneously submitted to Bioinfotracker and the four different tools Pfam, COG, BLAST and InterProScan. HttpFox, (https://addons.mozzilla.org/en-US /firefox/addon/6647), a Firefox add-on is operated at the backend to measure the loads of bytes transferred during the analyses. The amount of bytes sent and received for each sequence in Bioinfotracker and other tools is tabulated for comparison (Table 1). The results show that the online tools, in overall, consumed 103.87 kb of data transfer. Bioinfotracker, on the other hand, consumed only 7.38 kb of data transfer. Above all, the unwanted page refresh nuisance was completely absent when using the Bioinfotracker. Further, the results in this tool are displayed in a simple table thereby reducing human

Table 1. Comparison of bandwidth consumption between Bioinfotracker and other online tools.

Bioinfotracker		Online tools- individual analysis							
		COG		Pfam		BLAST		InterproScan	
Sent	Received	Sent	Received	Sent	Received	Sent	Received	Sent	Received
(In Bytes)		(In Bytes)		(In Bytes)		(In Bytes)		(In Bytes)	
1488	206	2195	453	2486	618	18135	176863	2195	732
850	206	2060	18564	1871	618	14742	176031	2060	453
1514	206	1221	453	2526	618	18737	176866	2526	618
1499	206	2739	732	2917	897	24206	178536	2917	618
997	206	1680	453	2409	897	16039	176532	2739	732
6348	1030	9895	20655	12209	3648	91859	884828	12437	3153
7378 [In Bytes]		1038684 [In Bytes]							

parsing. Hence, this tool prevails to be a novel system for the functional genomics research.

Conclusion

We present Bioinfotracker as one of the most appropriate and coordinated programs for performing functional annotation of the genes from any organism and for elucidating functions for unknown or hypothetical proteins. Henceforth, Bioinfotracker will be a useful tool for genomic research in the future.

ACKNOWLEDGEMENT

Ms. T.K. Subazini and Mr. G. Aravindhan, Junior Researchers were supported by the fundings provided by DIT.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215 (3): 403–410.

Aravindhan G, Kumar GR, Kumar RS, Subha K (2009a). AIM-BLAST-AJAX Interfaced Multisequence Blast. Proteomics Insights 2:: 1-7.

Aravindhan G, Kumar RS, Subha K, Subazini TK, Dey A, Kant K, Kumar GR (2009b). Proteomics Insights 2: 9-13.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004). The Pfam protein families database Nucleic Acids Res.1: 32 D138-141.

Cathy HW, Hongzhan H, Lai-Su LY, Winona CB (2003). Protein family classification and functional annotation. Comput. Biol. Chem. 27(1): 37-47.

Diana MD (2003). Genomics and Bacterial Metabolism Curr. Issues Mol. Biol. 5(1): 17-25

Kim C, Matthew B (2003). MASV—Multiple (BLAST) Annotation System Viewer Bioinformatics 19 (17): 2313 –2315.

Lobley E, Nugent T, Orengo CA, Jones DT (2008). FFPred: an integrated feature-based function prediction server for vertebrate proteomes Nucleic Acids Res. 1-6.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database Nucleic Acids Res. 35: D224-8.

Paulson LD (2005). Building rich web applications with AJAX IEEE. 38: 14-17.

Pillai S, Silventoinen V, Kallio K, Senger M, Sobhany S, Tate J, Velankar S, Golovin A, Henrick K, Rice P, Stoehr P, Lopez R (2005) SOAP-based services provided by the European Bioinformatics Institute. Nucleic Acids Res. 33: W25-W28.

Rentzsch R, Orengo CA (2009). Protein function prediction - the power of multiplicity Trends Biotechnol. 27(4): 210.

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Güldener U, Mannhaupt G, Münsterkötter M, Mewes HW (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes Nucleic Acids Res. 14:32(18):5539-5545

Sasson O, Noam K, Michal L (2006) Functional annotation prediction: All for one and one for all. Protein Sci. 15(6): 1557.

Sivashankari S, Shanmughavel P (2003) Functional annotation of hypothetical proteins - Rev. Bioinformation 291(8): 335-358.

Stuart CG, Rison T, Charles H' Janet M, Thornton J (2000) Comparison of functional annotation schemes for genomes, Funct. Integr. Genomics 1: 56–69.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes BMC Bioinformatics 11: 4-41