Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Views & Challenges**

An automated protein annotation filter for integrating web-based annotation tools

Vijayakumar Saravanan and Primanayagam Shanmughavel*

DBT Bioinformatics Facility, Department of Bioinformatics, Bharathiar University, Coimbatore - 641046, Tamil Nadu, India;
Primanayagam Shanmughavel* - E-mail: shanvel_99@yahoo.com; *Corresponding author
received May 11, 2007; revised July 28, 2007; accepted October 12, 2007; published online December 15, 2007

Abstract:
A wide range of web based prediction and annotation tools are frequently used for determining protein function from sequence. However, parallel processing of sequences for annotation through web tools is not possible due to several constraints in functional programming for multiple queries. Here, we propose the development of APAF as an automated protein annotation filter to overcome some of these difficulties through an integrated approach.

Keywords: protein annotation; integrated tool

Background:
There is a persistent need to apply a wide range of prediction and annotation tools to one or more sequences. The use of numerous tools in a discovery environment for routine analysis of sequences is often time consuming and laborious. During the discovery process scientists generally search SWISS-PROT **[1]** and NCBI **[2]** databases for specific queries. **[3]** The database search seldom generates useful search results for newly generated sequences. In such cases, several prediction tools are applied for potential functional inferences. A number of biological prediction tools are available over the web for annotation. Computational prediction of protein function by various approaches is important in assigning function to hypothetical proteins for the probable identification of drug targets. **[4]** However, these tools generally do not allow for parallel processing. In addition, processing time over the network is yet another parameter requiring attention. Here, we describe the development of an automated protein annotation filter named APAF to circumvent some of these difficulties through an integrated approach.

Model development
The overall architecture of APAF is shown in (Figure 1). The system is implemented in VISUAL BASIC using GUI interface. The APAF system consists of three main components: (1) module for submission of large number of sequences (maximum 50,000) to different servers; (2) module for analyzing, filtering and collating the result; and (3) display module for producing result in HTML format for comparing. The most multifaceted part of the system is the display module that produces the result from various annotation/prediction services in a uniform concerted manner. The analyzing module filters the result from the prediction/annotation services based on a E-value (expect value) cut-off value and thus by omitting hits for unknown function (based on inbuilt database of anonymous entry). Each module is developed to run independently and hence the master code will simply execute the entire module for each input. Our system allows single sequences to be sent to each prediction/annotation server multiple times, and therefore it prevents the server side restriction for multiple sequence submission. The tools included in the system are BLAST **[5]**, InterPro Scan **[6]**, COGnitor-prokaryotic, KOGnitor-eukaryotic **[7]**, CELLO **[8]**, and Pfam HMM **[9]**. The system is implemented in a WINDOWS operating system through an internet facility.

Input
The tool accepts protein sequences as input. A raw formatted text file should be used for single sequence input. In the case of multiple sequence submission, each sequence should be enclosed between "<" (less than symbol) and ">" (greater than symbol) within a text file with the sequence ID placed before the "<" symbol. The maxiumum allowed limit for input is 50000 sequences. On submission the user is prompted with options for prokaryotic and eukaryotic search.

Output
The APAF system produces a single result file in HTML format with filtered annotation data. Results from each tool are presented in separate column for easy comparison. In addition to the filtered data, results for individual sequences can be accessed through the corresponding hyperlinks provided in the output. The result for different tools is presented in a consistent format and hence the data can be aligned or arranged in a desired format for inter-operability.

Caveats and future development
APAF is designed in Visual Basic 6.0 and framed in a way to perform a simple but repetitive task in an ambiguity free manner. This application tool is tested with different sets of sequences for efficiently check. A huge data set can be annotated in a single run through this application tool by saving time. We plan to incorporate GO **[10]** and data from PUBMED into the system.

Availability: The set-up files are available for free from the authors upon request.

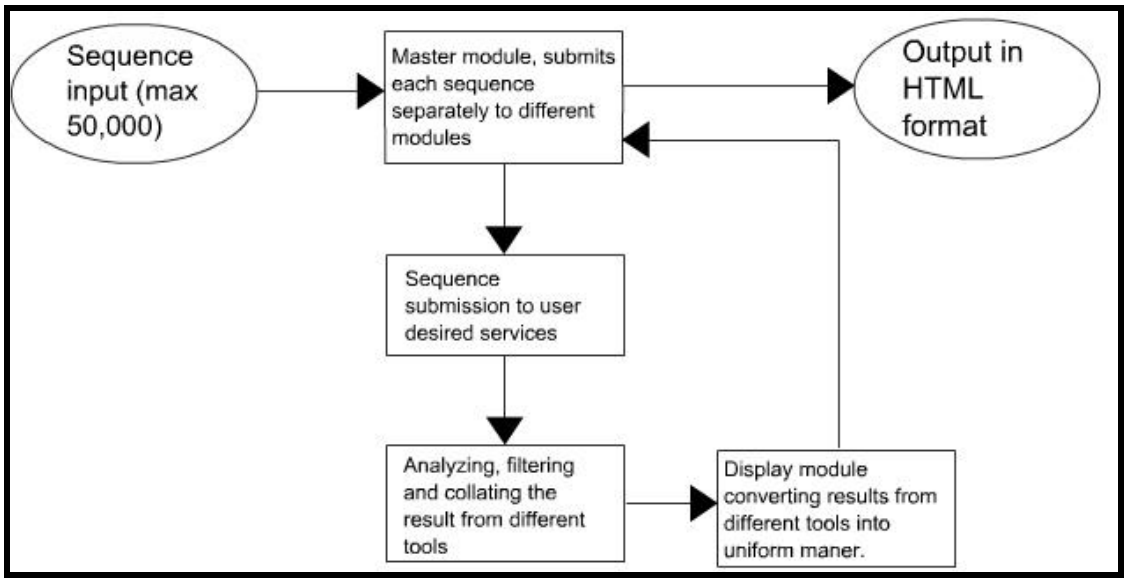

Figure 1: Architecture of APAF. It consists of (1) submission module; (2) analysis module; (3) display module

References

[01] A. Bairoch & R. Apweiler, *Nucleic Acid Research*, 28: 45 (2000) [PMID:10592178]

[02] www.ncbi.nlm.nih.gov

[03] B. Boeckmann, *et al.*, *Nucleic Acid Research*, 31: 365 (2003) [PMID: 12520024]

[04] Y. Orfan, *et al.*, *Drug Discov Today*, 10: 1475 (2005) [PMID: 16243268]

[05] S. F. Altschul, *et al.*, *J Mol Biol.*, 215: 403 (1990) [PMID: 2231712]

[06] E. Quevillon, *et al.*, *Nucleic Acids Res.*, 33: 116 (2005) [PMID: 15980438]

[07] R. L. Tatusov, *et al.*, *Science*, 278: 631 (1997) [PMID: 9381173]

[08] C. S. Yu, *et al.*, *Protein Sci.*, 13: 1402 (2004) [PMID: 15096640]

[09] E. L. Sonnhammer *et al.*, *Proteins*, 28: 405 (1997) [PMID: 9223186]

[10] M. Ashburner *et al.*, *Nature Genet.*, 25: 25 (2000) [PMID: 10802651]

Edited by P. Kangueane

Citation: Saravanan & Shanmughavel, Bioinformation 2(2): 76-77 (2007)