GeneNet: A Meta-Sequence Similarity Analysis System

Dong-Sun Park dspark@bric.postech.ac.kr
Sung Hwan Kim shkim@bric.postech.ac.kr
Tae-Jin Ahn tj@bric.postech.ac.kr
Yang-Suk Kim yskim@bric.postech.ac.kr
Hong-gil Nam hgn@bric.postech.ac.kr
Biological Research Information Center, POSTECH Pohang, 790-784, Korea

Keywords: similarity analysis, DNA, protein, meta-sequence, periodic search

1 Introduction

Sequence analysis using bioinformatics tools is an essential step in molecular biology. However, most of the web sites for analysis are dispersed in different locations, users must remember all web sites and change query in the proper format in each web site. Furthermore, databases are being updated continuously, users must repeat laborous searchs if they do not get satisfactory results from the first search. GeneNet was developed for resolving time-consuming search schemes and has three functions as a meta-search engine. First, GeneNet has a adapter for communicating sequentially with four databases (GenBank [1], PDB [2], BLOCKS [3], and KEGG [5]); therefore, users can get search results from all four databases through GeneNet's homepage. Next, GeneNet supports periodic searchs. When users perform search using the GeneNet system, GeneNet automatically stores the user's profile such as user's email address, search data, time, and query sequences; and, performs periodic searchs if the user wants. The final function is the parsing of output files. We integrated the Sequence Retrieval System (SRS) [6] to the GeneNet system for providing enriched information about selected sequences by users. The GeneNet was developed using Perl language under the Linux 6.0 environment and MySQL is used as a DBMS to store user's profile. Now, GeneNet is available at WWW(http://brac.postech.ac.kr/eng/)

2 Systems and Methods

All programs have been written in Perl for CGI, query analysis, pattern matching, connecting external web servers (Fig. 1)

Query parser and analyzer module play a role in checking errors and converting sequence in regular format (Fig. 1). Two modules check the misspelling of an input sequence, the kind of sequence, and the compatibility of the sequence and the search programs. Extra characters such as number, space bar and CF/LF are also removed by two modules.

LWP (Library for WWW access in Perl) was used for the client socket programming. The WWW interface module (Fig. 2) sends customized requests to each WWW server and receives results in HTML format.

Detailed information for searched sequences are directly provided by integrating SRS to GeneNet. If the user selects interested sequences from an output file, GeneNet automatically subtracts the primary key of the selected sequences and sends queries to SRS. The results are provided with users in table form for user's convenience. Periodic searchs are performed by using cron daemon of linux system.

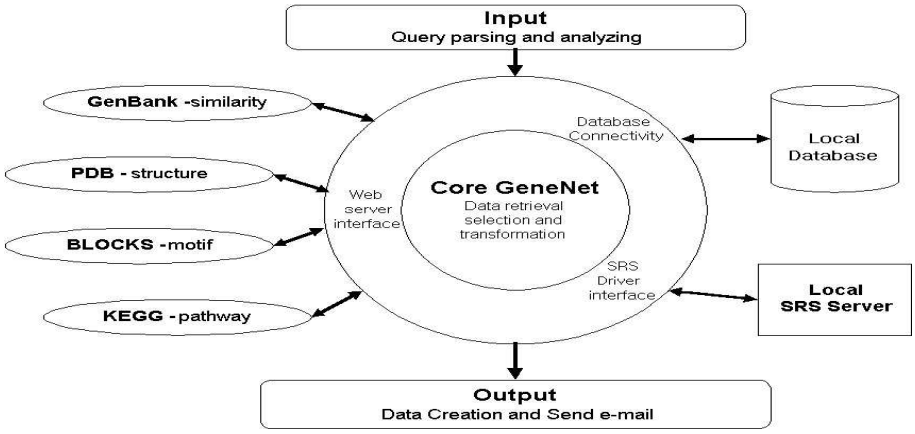

Figure 1: A schematic diagram of the GeneNet system and Web servers.

```
$ua=LWP::UserAgent->new();
$h=new HTTP::Headers(Accept => "text/html", User_Agent   => "Mozilla/1.0",
                Content_type => "application/x-www-form-urlencoded",
                Content_length => "$query_length",);
$request=new HTTP::Request("POST","http://www.ncbi.nlm.nih.gov/blast/blast.cgi",$h,$qry);
$response=$ua->request($request);
```

Figure 2: implements WWW client socket using LWP.

References

[1] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L., GenBank, *Nucleic Acids Res.*, 28(1):15–18, 2000.

[2] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., The protein data bank, *Nucleic Acids Research*, 28(1):235–242, 2000.

[3] Henikoff, J.G., Greene, E.A., Pietrokovski, S. and Henikoff, S., Increased coverage of protein families with the blocks database server, *Nucleic Acids Res.*, 28(1):228–230, 2000.

[4] Wall, L., Christiansen, T., and Schwartz, R.L., Programming Perl, *OŔeilly&Associates*, Inc, 1996.

[5] Kanehisa, M. and Goto, S., KEGG, *Nucleic Acids Res.*, 28(1):27–30, 2000.

[6] Etzold, T., Ulyanov, A., Argos, P., SRS: Information Retrieval System for Molecular Biology Data Banks, *Methods in Enzymology*, 266:114–128, 1996