A Proposal for Integration of BLAST-Related Software Tools Using CORBA

Pethuru Raj peter@ics.nitech.ac.jp
Naohiro Ishii ishii@ics.nitech.ac.jp

Department of Intelligence and Computer Science, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya 466 - 8555, Japan

Keywords: GenBank, homology searching, BLAST, CORBA

1 Introduction

As information systems have come to play an increasingly important role in molecular biology area, the need has emerged for a more flexible, efficient and cost-effective application architecture-one that integrates existing technology with new and integration of software tools and applications. These requirements have been the driving forces behind the rapid adoption of the distributed computing model, **CORBA** (Common Object Request Broker Architecture) [2]. Legacy applications are now assets. Now, with CORBA as the key, it is relatively easy to unlock the important functionalities trapped within these applications. In that direction, we present here a viable proposal for integrating BLAST-related tools using CORBA.

2 Integration of BLAST-related Tools

The biological databases like GenBank contain all known nucleotide and protein sequences with supporting bibliographic and biological annotations. There are a number of software tools available for gathering, storing, managing, accessing, retrieving, and extracting the biological data being unleashed by various sequencing projects all over the world. The most frequent type of analysis performed using GenBank is the search for sequences similar to a query sequence for inferring the functionalities of different genes trapped in the DNA sequences. NCBI offers the **BLAST** family of search programs to locate good alignments between a query sequence and database sequences. Also there are enhancements to BLAST for pre-processing the query sequences and post-processing the BLAST results like **BEAUTY**, **PowerBLAST** and **VisualBLAST**.

BEAUTY (BLAST Enhanced Alignment UtiliTY) is an enhanced version of BLAST that facilitates identification of the functions of matched sequences.

VisualBLAST can facilitate and accelerate the interactive analysis of full BLAST output files containing sequence alignments.

PowerBLAST includes a number of options for masking repetitive elements and low complexity subsequences. It also has the capacity to restrict the search to any level of NCBI's taxonomy index, thus supporting "comparative genomics" applications.

These software products are loaded in different server machines and they work independently fulfilling the different needs of a variety of users. Thus there is a strong need for a synergistic integration of these software tools for coordinating the different functionalities with a common front-end view so that any user especially molecular biologist, can get all the functionalities accomplished with ease using the new distributed object computing paradigm.

To build up a CORBA compliant application, one usually requires a software package which provides an ORB and an IDL compiler. The ORB supplies a framework which facilitates object communications and the IDL compiler takes the IDL input and generates appropriate source code in a given programming languages such as Java. The details will be supplied in final version.

3 IDL Interfaces for the Integration

The separation of client and server communicating through an agreed interface is the cornerstone of CORBA distributed software design allowing concurrent use of different languages and operating systems and still allowing both clients and servers to improve implementations and add new features independently of each other. Here is the IDL file for wrapping the legacy applications BLAST, PowerBLAST, BEAUTY.

```
module GenBank {
module HomologySearching {

    typedef sequence < octet > fileFlow; //File Bytes
    exception InvalidID string reason;;

    interface BLAST {
    attribute string accessionNumber;
    typedef sequence < string > accessionNumbers;
    readonly attribute string description;
    boolean exists(in string ID);
    accessionNumbers getaccessionNumbers(in string DNAQuery);
    string getBases(in string ID) raises (InvalidID);
    fileFlow getGZIPFile(in string ID) raises (InvalidID); }

    interface BEAUTY {
    - - - -

    }

    interface PowerBLAST {
    - - - -

    }
}
}
```

Here we have given an idea of integrating various homology searching software tools for accessing GenBank entries and an IDL file as a wrapper that can be extended further according to the requirements. In the final version, we will supply the relevant details, full interfaces for individual DNA sequences, a GenBank entry, the references for each homology alignment, GenBank Factory, etc. defined using OMG's IDL (Interface Definition Language).

References

[1] Pethuru, R. and Ishii, N., Interoperability of biological databases by CORBA, *Proc. International Conference on Information, Intelligence, and Systems*, Washington, 16–24, 1999.

[2] Vogel, A. and Duddy, K., Java Programming with CORBA, John Wiley & Sons, Inc., New York.