Virtual Network File Servers for Sequence Database Management and Bioinformatics Tool Integration

Roger A. Sayle[1] and J. Scott Dixon[2].
Metaphorics LLC, Santa Fe, New Mexico

1. roger@metaphorics.com
2. scott@metaphorics.com

Keywords: Virtual Network File Servers, Database Management, Sequence Databases, Relational Database, DBMS, CORBA.

The current trend towards storing bioinformatics data in advanced database management systems, such as relational or object-oriented databases, has created a serious data management problem for the bioinformatics community. Although biological data requires complex schemas to keep track of annotations and relationships between entities, most of the tools for searching and processing this data (for example BLAST) operate on flat files. Many applications, including BLAST, SRS and GCG, even require the flat files to be in their own file format. This situation currently requires data to be duplicated in multiple locations, with the associated administrative problems of keeping all of these formats synchronized. The data duplication problem for already large data sets is exaggerated further by the need to subset and superset such data; for example maintaining non-redundant, EST or organism specific databases.

One possible solution is the current move towards CORBA to provide a uniform interface for bioinformatics applications to access complex biological databases. However this approach requires modification of each application's source code and may negate many of the benefits of having an application specific file format or index.

The alternative solution described here makes use of virtual network file servers to integrate modern data stores with existing legacy applications. A virtual network file server emulates the behaviour of a remote file server, using a protocol such as NFS, Microsoft LanManager/SMB/CIFS or Appleshare. Much like a WWW server, a VNFS server receives requests to read or write files from a client application and generates the appropriate responses and acknowledgements. No modifications are required to the application software or even to the client machine; once a VNFS directory or disk is mounted it is indistinguishable from a real remote file system. The server, however, is free to maintain the data internally in whatever formats it chooses, rapidly generating flat-file formats such as FASTA, SWISSPROT or BLAST indices on-the-fly. This allows a VNFS server to act as a gateway to a relational database or a CORBA ORB. Indeed the performance of running BLAST unmodified on a VNFS file system can be faster than accessing a real file system.