Bioinformatics Advance Access published May 30, 2007

Structural Bioinformatics

StrBioLib: a Java library for development of custom computational structural biology applications

John-Marc Chandonia*

Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, and
Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA

Associate Editor: Dr. Alex Bateman

ABSTRACT

Summary: StrBioLib is a library of Java classes useful for developing software for computational structural biology research. StrBioLib contains classes to represent and manipulate protein structures, biopolymer sequences, sets of biopolymer sequences, and alignments between biopolymers based on either sequence or structure. Interfaces are provided to interact with commonly used bioinformatics applications, including (PSI)-BLAST, MODELLER, MUSCLE, and Primer3, and tools are provided to read and write many file formats used to represent bioinformatic data. The library includes a general-purpose neural network object with multiple training algorithms, the Hooke and Jeeves nonlinear optimization algorithm, and tools for efficient C-style string parsing and formatting. StrBioLib is the basis for the Pred2ary secondary structure prediction program, is used to build the ASTRAL compendium for sequence and structure analysis, and has been extensively tested through use in many smaller projects. Examples and documentation are available at the site below.

Availability: StrBioLib may be obtained under the terms of the GNU LGPL license from http://strbio.sourceforge.net/

Contact: JMChandonia@lbl.gov

Computational structural biology research often requires time-consuming development of custom software to analyze data. Development of such software is facilitated by publicly available libraries that read and write the multitude of file formats in which bioinformatic data is stored, implement commonly used algorithms, and otherwise efficiently perform common tasks (Mangalam, 2002). Object-oriented languages such as Java or C++ are particularly well suited for such libraries, as judicious choice of an object representation allows methods to be described and implemented in high level terms, thus facilitating rapid development and testing of alternative algorithms. In addition, object-oriented programming languages facilitate efficient reuse of code through extension and inheritance of existing classes. StrBioLib is a library of Java classes that represent objects, concepts, and tools useful for the development of algorithms for computational structural biology research. StrBioLib is complementary to existing libraries such as BioJava (Pocock, et al., 2000) that focus on tools for analysis of biological sequences. StrBioLib is mature, having been used by several research groups over more than 10 years; some classes predate the Java programming language and were ported from earlier C and C++ versions. A new public release of the library, version 1.1, was made available through SourceForge in January 2007. Details and applications of StrBioLib are given below.

*To whom correspondence should be addressed.

Molecular Biology Classes

The core of StrBioLib is the org.strbio.mol package, which contains classes that represent objects from the field of structural molecular biology: Atom, Molecule (composed of Atoms), Monomer (also containing Atoms), Polymer (an ordered set of Monomers and associated metadata), Residue (a type of Monomer representing an amino acid residue), Nucleotide (another subclass of Monomer), and Protein (a type of Polymer composed of Residues). The mol package also contains objects that represent groups of Polymers (PolymerSet) and Proteins (ProteinSet), including specialized groups such as a Profile (representing a set of sequences aligned to a Protein). In addition to providing methods to efficiently manipulate the objects in memory, each class also contains methods to read and write the objects from a variety of file formats, including the widely used FASTA, PDB, MSF, DSSP, HSSP, and BLAST formats. The mol package also contains an Alignment object, an efficient representation of a sequence or structure-based alignment between two Polymers.

The org.strbio.mol package is supported by classes in the org.strbio.mol.lib package, which contain objects that represent and implement more abstract concepts in structural biology, such as algorithms (e.g., secondary structure prediction, threading, sequence searching, and sequence alignment), scoring matrices, and parameter sets. Tweaking an algorithm is often a simple matter of extending a class to change functionality; this greatly simplified development of the MAKERAF software, described below.

Interfaces to Bioinformatic Tools and Databases

StrBioLib also contains classes to interact and exchange data with many commonly used bioinformatic tools and databases. Tools that must be installed locally have their corresponding classes in the org.strbio.local package, and tools that must be accessed over the Internet correspond to classes in the org.strbio.net package. A partial listing of tools and databases that StrBioLib can manipulate or interact with is given in Table I.

General Purpose Tools

StrBioLib also contains packages of tools that are useful in a wide range of applications beyond the field of structural biology. While some of the objects are now provided in current releases of the JDK, StrBioLib contains implementations that are also compatible with earlier versions of Java. The org.strbio.util package contains a neural network object that implements both traditional Steepest Descent and Scaled Conjugate Gradient (Møller, 1993) algorithms. It also contains an algorithm for nonlinear optimization using the

Table I: Tools and Databases accessible through StrBioLib 1.1

Tool	Purpose	Access
BLAST (Altschul, et al., 1990)	Search for similar sequences	L, W
CATH (Orengo, et al., 1997)	Protein domain classification	L
DSSP (Kabsch & Sander, 1983)	Secondary structure calculation	L
EXPASY (http://expasy.org)	Predicted sequence properties	W
MAXSUB (Siew, et al., 2000)	Evaluate quality of models	L
MELTING (Le Novere, 2001)	Polynucleotide Tm calculation	L
Mfold (Mathews, et al., 1999)	Polynucleotide fold prediction	L
MINAREA (Falicov & Cohen, 1996)	Structural alignment	L
MODELLER (Sali & Blundell, 1993)	Comparative modeling	L
MUSCLE (Edgar, 2004)	Multiple sequence alignment	L
NACCESS (Hubbard, unpublished)	Solvent accessibility calculation	L
PDB (Berman, et al., 2000)	Protein/nucleic acid structures	L
Primer3 (Rozen & Skaletsky, 2000)	PCR primer design	L
PSI-BLAST (Altschul, et al., 1997)	Profile-based sequence search	L, W
PubMed (http://pubmed.gov)	Literature searches	W
SCOP (Murzin, et al., 1995)	Protein domain classification	L, W
TargetDB (Chen, et al., 2004)	Structural genomics data	W

A partial listing of bioinformatics resources that can be accessed through StrBioLib objects is shown above, along with the method StrBioLib uses to access each resource. L = local installation required; W = web-based access.

direct search method of Hooke and Jeeves (1961), and a double-linked list that allows efficient random access to any element. The org.strbio.io package contains an extensive library of string functions for implementing C-style formatted I/O without the overhead of creating and destroying objects; these methods are essential for supporting the multitude of file formats used by bioinformatic programs with speed comparable to that of C code. The org.strbio.math package contains classes to support matrix algebra, as well as statistical objects that provide calculations such as Pearson and Matthews correlation coefficients (Matthews, 1975). The org.strbio.util.ui package contains classes useful for developing graphical user interfaces, and the org.strbio.util.graph package contains classes useful for graphing data.

Applications

StrBioLib has been used to develop a number of published applications, including secondary structure prediction software (Chandonia and Karplus, 1995; Chandonia and Karplus, 1996; Pred2ary, Chandonia and Karplus, 1999), threading methods (JThread, Chandonia and Cohen, 2003), and the MakeRAF software that creates mappings between the sequence and experimentally observed residues from PDB files in the ASTRAL database (Chandonia, et al., 2002). MakeRAF provides an example of how to create a customized function for scoring gaps in sequence alignments by implementing the org.strbio.mol.lib.GapModel interface. Both MakeRAF and Pred2ary are included with StrBioLib, along with instructions for stand-alone installation and testing, and sample output useful for validation.

Several additional programs are included with StrBioLib in the org.strbio.app package. These programs may be run as stand-alone utilities or used as models for further application development. ConvertProtein is an application for converting between various protein file formats and performing basic data manipulation (e.g., rotation and translation of protein structure, or elimination of particular atoms or residues). Align is a front end to the sequence alignment algorithms included in StrBioLib. FindProteins and SplitProteins are utilities to manipulate large sets of proteins; they work, respectively, by separating particular proteins from a group by name, and splitting a group into multiple subsets of equal size.

ACKNOWLEDGEMENTS

Thanks to L. Howard Holley, Jonathan D. Blake, and Marcin P. Joachimiak for discussions leading to implementation of some of the classes. This work is supported by grants from the NIH (R01-GM39900, R01-GM073109, and 1-P50-GM62412) and by the U.S. Department of Energy under contract DE-AC02-05CH11231.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool, J Mol Biol, 215, 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res, 25, 3389-3402.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank, Nucleic Acids Res, 28, 235-242.

Chandonia, J.M. and Cohen, F.E. (2003) New local potential useful for genome annotation and 3D modeling, J Mol Biol, 332, 835-850.

Chandonia, J.M. and Karplus, M. (1995) Neural networks for secondary structure and structural class predictions, Protein Sci, 4, 275-285.

Chandonia, J.M. and Karplus, M. (1996) The importance of larger data sets for protein secondary structure prediction with neural networks, Protein Sci, 5, 768-774.

Chandonia, J.M. and Karplus, M. (1999) New methods for accurate prediction of protein secondary structure, Proteins, 35, 293-306.

Chandonia, J.M., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M. and Brenner, S.E. (2002) ASTRAL compendium enhancements, Nucleic Acids Res, 30, 260-263.

Chen, L., Oughtred, R., Berman, H.M. and Westbrook, J. (2004) TargetDB: a target registration database for structural genomics projects, Bioinformatics, 20, 2860-2862.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res, 32, 1792-1797.

Falicov, A. and Cohen, F.E. (1996) A surface of minimum area metric for the structural comparison of proteins, J Mol Biol, 258, 871-892.

Hooke, R. and Jeeves, T.A. (1961) Direct Search Solution of Numerical and Statistical Problems, Journal of the ACM, 8, 212-229.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers, 22, 2577-2637.

Le Novere, N. (2001) MELTING, computing the melting temperature of nucleic acid duplex, Bioinformatics, 17, 1226-1227.

Mangalam, H. (2002) The Bio* toolkits--a brief overview, Brief Bioinform, 3, 296-302.

Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, J Mol Biol, 288, 911-940.

Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim Biophys Acta, 405, 442-451.

Møller, M.F. (1993) A scaled conjugate gradient algorithm for fast supervised learning, Neural Networks, 6, 525-533.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, J Mol Biol, 247, 536-540.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH--a hierarchic classification of protein domain structures, Structure, 5, 1093-1108.

Pocock, M.R., Down, T. and Hubbard, T. (2000) BioJava: Open Source Components for Bioinformatics, ACM SIGBIO Newsletter, 20, 10-12.

Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers, Methods Mol Biol, 132, 365-386.

Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints, J Mol Biol, 234, 779-815.

Siew, N., Elofsson, A., Rychlewski, L. and Fischer, D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality, Bioinformatics, 16, 776-785.