Special Issue
"Progress of Theoretical and Computational Biophysics"

Vol. 16, pp. 485–489 (2019)
doi: 10.2142/biophysico.16.0_485

Biophysics and Physicobiology
https://www.jstage.jst.go.jp/browse/biophysico/

Review Article

A tool written in Scala for preparation and analysis in MD simulation and 3D-RISM calculation of biomolecules

Itaru Onishi, Hiroto Tsuji and Masayuki Irisa

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan

Received July 1, 2019; accepted October 1, 2019

Researchers studying biomolecules require easy-to-use, customizable tools allowing them to effectively use other molecular science packages written for molecular dynamics (MD), quantum chemistry, statistical mechanics, and molecular graphics. This paper presents a Scala tool for the computational science of biomolecules (STCSB) developed in Scala, which allows users to prepare and run MD simulations, as well as perform three-dimensional reference interaction site model (3D-RISM) calculations of biomolecules. Features of the STCSB include the following: (1) a cross-platform application running on a Java virtual machine; (2) handling hierarchical XML-based data formats such as the protein data bank markup language (PDBML); (3) prepared application programming interfaces (APIs) with both character user interface (CUI) and graphical user interface (GUI) options; (4) prepared APIs for molecular graphics; and (5) a scalable source code based on the Model-View-Controller (MVC) architectural pattern.

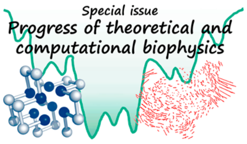
Key words: PDBML, MD, 3D-RISM, MVC architectural pattern

Availability: The STCSB is freely available from https://github.com/irisa-lab/STCSB

Corresponding author: Masayuki Irisa, Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan.
e-mail: irisa@bio.kyutech.ac.jp

Several application packages for molecular simulations, such as Amber [1], CHARMM [2], and GROMACS [3], are available for researchers studying biomolecules. In many cases, other tools should be used in conjunction with these packages for the pre- and post-processing of molecular simulations. AmberTools [1] is one of the most popular utility tools for the Amber application. Molecular graphics systems UCSF Chimera [4] and VMD [5] provide not only visualization of molecules but also function for molecular modeling and analyzing results from other molecular science application packages. MDAnalysis [6] and MD-TASK [7] are tools to analyze trajectories from molecular dynamics (MD) simulations. Often, researchers have to use multiple additional tools to complement the functionality provided by molecular science application packages. Each tool has its own advantages depending on the science or technology subject it is based on. To manage multiple tools, researchers usually prepare a shell script containing a sequence of commands, which varies depending on the purpose of the required calculations. To simplify this process, CHARMM-GUI [8], for example, composes tools for the CHARMM package as a web application. However, the menus of different tools are usually fixed, not allowing users to customize them.

An application framework [9] is a software framework consisting of libraries used to implement a standard structure of an application software. One such application framework



◀ ***Significance*** ▶

We described a tool called STCSB written in Scala for computational biophysics allowing users to analyze molecular dynamics and perform 3D-RISM theory (a statistical mechanics theory) calculations to study biomolecules. STCSB is a cross-platform application running on a Java virtual machine and has prepared application interfaces, e.g. GUI supporting 3D-RISM pre-processing, MD trajectory analyzer, validator of atoms and residues in PDB/PDBML files, and molecular graphics with a stereoscopic view. STCSB employs the MVC architectural pattern as its programming architecture, which enables the incremental addition of functions calling external application-software.

is Ruby on Rails, which is adopted in various web services such as GitHub, Airbnb, Kickstarter, and Square.

Employing an application framework using the Model-View-Controller (MVC) architectural pattern is an effective way for researchers to create their own application to manipulate multiple computational molecular-science tools. Separating the model component from other components in the MVC architectural pattern makes it easy to incorporate new tools for molecular science developed by other researchers into this model component. This is similar to web application development, where an application framework helps developers to create new web applications containing a data-handler connected to a relational database for customers (model component), a beautiful layout of web pages (view component), and functions implementing interactive responses for mouse actions in the web pages (controller component).

To summarize, researchers studying biomolecules require easy-to-use, customizable tools allowing them to effectively use other molecular science packages written for MD, quantum chemistry, statistical mechanics, and molecular graphics. This paper presents the Scala tool for the computational science of biomolecules (STCSB) developed in Scala, which allows users to prepare and run MD simulations, as well as perform three-dimensional reference interaction site model (3D-RISM) calculations of biomolecules. STCSB was originally developed for the pre- and post-processing of 3D-RISM [10] calculations.

Overview of STCSB

STCSB was made based on the philosophy of application frameworks to prepare and analyze data for MD and 3D-RISM calculations. A molecular modeling software package, Tinker [11], is used in STCSB to prepare a direct input file for MD or 3D-RISM calculations. In this aspect, STCSB is addressed as a graphical user interface (GUI) tool for Tinker.

Some parts of the Tinker source code to extract specific information required from input files have been rewritten in Scala to facilitate the preparation steps required for 3D-RISM calculation. We also supplied a simple molecular-graphics application programming interface (API) in STCSB to visually check the target molecule structure that was modeled with Tinker.

One of the strengths of STCSB is the hybrid-usage of the MD and 3D-RISM theory (a statistical mechanics theory). For example, coordinates of solvent molecules (water molecules and ions) surrounding a protein in the initial structure of MD simulation can be calculated with the 3D-RISM theory. In this example, the ambiguity of coordinates of solvent molecules in an X-ray structure is avoided by efficiently performing many MD simulations with available initial-structures, which have ions at different positions predicted by the 3D-RISM theory [12].

Features

Specific features

STCSB features include the following: (1) a cross-platform (i.e., platform-independent) application distributed as jar files (Java byte codes) running on a Java virtual machine (VM) similar to other Java applications; (2) handling hierarchical XML-based data formats such as the protein data bank markup language (PDBML); (3) prepared APIs with both character user interface (CUI) and GUI options; (4) prepared APIs for molecular graphics; and (5) a scalable source code using the MVC architectural pattern to enable users writing both a shell-script-like source code to execute external application programs in the Model modules and a CUI/GUI/molecular-graphics source code in the Controller and View modules using a single programming language (Scala in this case). A ported version of STCSB for Android is also available, because Android system has Java APIs despite the luck of Java VM. The ported version works on Android devices, smartphones/tablets, as an Android application.

Required libraries

STCSB contains the following libraries: Standard Widget Toolkit (SWT), Lightweight Java Game Library (LWJGL), scala-xml, and scala-parser-combinator (Fig. 1). These libraries are distributed with STCSB as a single jar file. Tinker is required to pre-process the MD and 3D-RISM calculations.

Prepared APIs

We prepared some APIs in Model components for frequent use when studying biomolecules using 3D-RISM or MD. Non-expert users can easily add these Model components, while expert users can implement other Model components in any programming language (for languages other than Scala, a wrapper class written in Scala would be required) to perform more complex tasks.

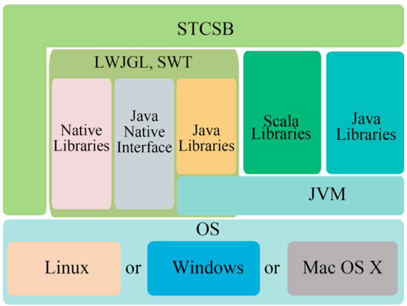


Figure 1 Schematic overview of STCSB.

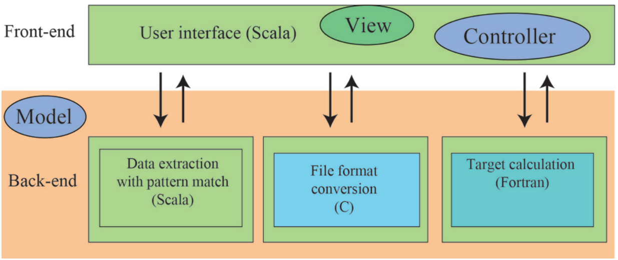
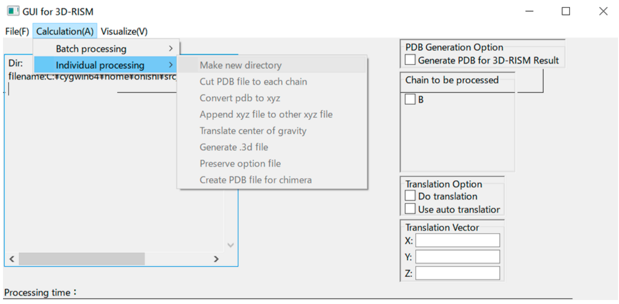

Front-end
User interface (Scala)
View
Controller
Model
Back-end
Data extraction with pattern match (Scala)
File format conversion (C)
Target calculation (Fortran)

Figure 2 Schematic drawing of STCSB following the MVC model.



Dir:
filename:C:¥cygwin64¥home¥onishi¥src

PDB Generation Option
□ Generate PDB for 3D-RISM Result

Chain to be processed
□ B

Translation Option
□ Do translation
□ Use auto translatior

Translation Vector
X:
Y:
Z:

Processing time :

Figure 3 Snapshot of the STCSB GUI with built-in menus for the pre-processing of 3D-RISM calculations.

Command line user interface to support pre- and post-processing of 3D-RISM calculations and post-processing of MD simulations in Amber

The CUI can call nearly all functions of the framework including the Model components introduced below. Following the MVC architectural pattern, the CUI consists of three components: Model, View, and Controller (Fig. 2). By editing the Controller component, users can add/modify mapping between commands in the CUI and sub-programs, actually executing calculations without modifying the Model components.

GUI supporting 3D-RISM pre-processing

A GUI supporting the pre-processing of 3D-RISM calculations is implemented in Scala using the SWT and LWJGL (Fig. 3). The GUI can execute functions implemented in other programming languages, such as Fortran, C, Python, and Java, as a different process by calling shell commands.

Programs to read and write PDBML files

Programs to read, edit, and write PDBML files are implemented in Scala. The programs allow users to prepare initial structures for computation by using PDBML files. The information on an atom is stored as an object of the case class “AtomInfo.” Other programs can use these programs as a module to handle PDBML files.

Program to validate atoms and residues in PDB/PDBML files

A validator called the PDBValidator for DNA and peptide sequences is implemented in Scala. The program validates atoms and residues from input files such as PDB or PDBML. In X-ray crystallography, the coordinates of some atoms and residues may not be assigned due to thermal fluctuations of protein structures. While PDB and PDBML files would contain the missing information in REMARK records, validating atoms and residues is important for the preparation of the initial structure.

```
---------------------------------
Validating DNA in Chain C
---------------------------------
---------------------------------
Validating DNA in Chain D
---------------------------------
---------------------------------
Validating AminoAcids in Chain A
---------------------------------
Missing atoms: CG,OD1,ND2 in 9-ASN
.
.
.
Missing atoms: CG,CD,OE1,OE2 in 57-GLU
Missing residues:  res67--67
Missing atoms: CG,CD,OE1,NE2 in 68-GLN
.
.
.
---------------------------------
Validating AminoAcids in Chain B
---------------------------------
Missing residues:  res13--18
Missing atoms: CG,CD1,CD2,CE1,CE2,CZ,OH in 12-TYR
.
.
.
Missing atoms: CG,CD,OE1,NE2 in 224-GLN
Missing atoms: CG,OD1,ND2 in 227-ASN
Missing atoms: CG,OD1,OD2 in 228-ASP
Missing atoms: CD,CE,NZ in 245-LYS
===[Result]====================
Coordinates of some atoms/residues are missing!!
```

Figure 4 Example of a PDBValidator output.

PDBValidator can be called from the CUI. The program prints results to the standard output. It shows residues, whose coordinates are not assigned after the “Missing residues:” statement, and atoms, whose coordinates are not assigned after the “Missing atoms:” statement (Fig. 4).

Placevent algorithm

We implemented the Placevent algorithm [13] in Scala. While a Python-based Placevent program is already available, it is implemented for Amber, whereas our version is for the STCSB package in Scala. Placevent algorithm is a simple algorithm for automatically predicting the explicit solvent atom distribution (coordinates) of biomolecules from the three-dimensional continuous distribution obtained by the 3D-RISM calculation.

Program for MD trajectory analysis

We implemented an MD trajectory analyzer for Amber, which can analyze the distance between two atoms, angle between three atoms, and torsion angle of four atoms. The MD trajectory analyzer can analyze root-mean-square deviation (RMSD) of the trajectory after superimposing structures over the reference structure using specified atom coordinates.

The analyzer program reads the arbitrary information

```
prmtop 1rvb_withNaCl_siteIandIVdagger_HIS2HIP.prmtop
mdcrd  1rvb_NVT_MD_withNaCl_siteIandIVdagger.mdcrd
prefix 1rvb_NVT_MD_withNaCl_siteIandIVdagger-traj

# distance
distance 355@OD1-525@O

# angle
angle 6@O3'-7@P-538@O

# torsion
torsion 6@C5'-6@C4'-6@C3'-6@O3'
```

Figure 5 Example of an input file for MD trajectory analyzer.

```
#step     #6@O3'-7@P-538@O  #17@O3'-18@P-525@O
     1           153.76208           153.37374
     2           156.94835           154.26489
     3           158.89551           154.84515
    .
    .
    .
 10998           151.61876           140.91769
 10999           152.09419           140.28708
 11000           149.96356           139.69684
```

Figure 6 Example of an output file produced by MD trajectory analyzer.

from the MD trajectory to avoid memory shortage and performance loss. The program requires an input file, as illustrated in Figure 5.

After analysis, the program outputs another file (Fig. 6) with a name specified by a "prefix" keyword, except for the file extension, which can be "angle" for angle analysis, "distance" for distance analysis, "torsion" for torsion analysis, and "rmsd" for RMSD analysis.

Molecular graphics

A simple molecular-graphics API in STCSB uses OpenGL through the LWJGL [14]. A molecule is represented as a space-filling model (Fig. 7). The API enables stereo-views of a molecule with 3D-glasses in the manner of QuadBuffered Stereo in OpenGL, for which a graphic card, NVIDIA Quadro, is required. This API can be used to check the designed molecular structures generated by Tinker [11].

Handling the newly generated 3'/5' terminus of DNA fragments using DNA hydrolysis

STCSB enables users to handle a diversity of biomolecule structures such as the newly generated 3′/5′ terminus of DNA fragments based on DNA hydrolysis. For example, PDB ID: 1rvc [15], a post-reactive X-ray crystal structure of the *Eco*RV–DNA complex, has the newly generated 5′ terminus of DNA fragments based on A-type DNA hydrolysis. The newly generated 5′ terminus should be properly handled during pre-processing to study the DNA hydrolysis reaction by *Eco*RV. However, LEaP in AmberTools and Tinker cannot handle the 5′ terminus due to the lack of topology information and hard-coded-assumption that no such structure appears, respectively. LEaP requires a modification of the DNA topology information in a parameter file that is difficult to read and edit, while Tinker requires a modification of the source code. On the other hand, STCSB can handle the 5′ terminus properly using the developed Model components.

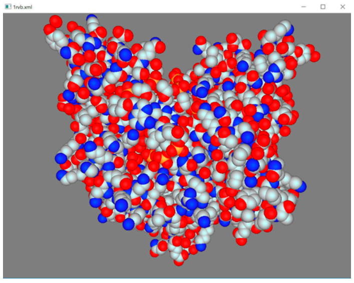


Figure 7 Snapshot of the STCSB molecular-graphics API, *Eco*RV–DNA complex (PDB ID: 1rvb).

Conclusion

In this paper, we described a tool called STCSB written in Scala for computational biophysics allowing users to analyze MD and perform 3D-RISM theory calculations to study biomolecules.

STCSB employs the MVC architectural pattern as its programming architecture. The Controller component in STCSB enables the incremental addition of functions calling external application-software that can be written in any programming language, including Fortran, C, and Java.

STCSB is highly scalable and suitable for both non-expert and expert users, allowing the latter to edit the Controller and Model components. The tool allows users to easily read, edit, and write PDBML (the next standard file-format for PDB) files. It enables users to handle diverse biomolecular structures, including the newly generated 3'/5'-terminus of DNA fragments based on DNA hydrolysis by optimizing its Model components. The tool features can facilitate collaboration between expert and non-expert users.

Acknowledgment

This work was supported by Grants-in-Aid (16041234) from MEXT, Japan.

Conflict of Interest

I. O., H. T., and M. I. declare that they have no conflict of interest.

Author Contribution

M. I. directed the entire project. M. I. and I. O. co-wrote the manuscript. H. T. made a draft version of STCSB, especially GUI part. I. O. has completed STCSB.

References

[1] Case, D. A., Betz, R., Cerutti, D., Cheatham, T., III, Darden, T., *et al.* AMBER 16. University of California, San Francisco, 2016.

[2] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1983).

[3] Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**–**2**, 19–25 (2015).

[4] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

[5] Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

[6] Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319–2327 (2011).

[7] Brown, D. K., Penkler, D. L., Sheik Amamuddy, O., Ross, C., Atilgan, A. R., Atilgan, C., *et al.* MD-TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics* **33**, 2768–2771 (2017).

[8] Jo, S., Cheng, X., Lee, J., Kim, S., Park, S. J., Patel, D. S., *et al.* CHARMM-GUI 10 years for biomolecular modeling and simulation. *J. Comput. Chem.* **38**, 1114–1124 (2017).

[9] Fayad, M. & Schmidt, D. Object-Oriented Application Frameworks. *Communications of the ACM* **40**, 32–38 (1997).

[10] Hirata, F. ed. *Molecular Theory of Solvation* (Kluwer Academic Publishers, 2003).

[11] Rackers, J. A., Wang, Z., Lu, C., Laury, M. L., Lagardère, L., Schnieders, M. J., *et al.* Tinker8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **14**, 5273–5289 (2018).

[12] Onishi, I., Sunaba, S., Yoshida, N., Hirata, F. & Irisa, M. Role of Mg^{2+} Ions in DNA Hydrolysis by *Eco*RV, Studied by the 3D-Reference Interaction Site Model and Molecular Dynamics. *J. Phys. Chem. B* **122**, 9061–9075 (2018).

[13] Sindhikara, D. J., Yoshida, N. & Hirata, F. Placevent: An Algorithm for Prediction of Explicit Solvent Atom Distribution—Application to HIV-1 Protease and F-ATP Synthase. *J. Comput. Chem.* **33**, 1536–1543 (2012).

[14] LWJGL community, LWJGL Offcial web page. http://www.lwjgl.org/.

[15] Kostrewa, D. & Winkler, F. K. Mg^{2+} Binding to the Active Site of *Eco*RV Endonuclease: A Crystallographic Study of Complexes with Substrate and Product DNA at 2 Å Resolution. *Biochemistry* **34**, 683–696 (1995).