Overview

The RosettaCon 2012 Special Collection: Code Writ on Water, Documentation Writ in Stone

Ingemar André[1]◑, **Jacob Corn**[2]◑*

[1] Department of Biochemistry and Structural Biology, Lund University, Lund, Sweden, [2] Department of Early Discovery Biochemistry, Genentech Inc., South San Francisco, California, United States of America

Rosetta is a powerful software suite for the modeling and design of macromolecules [1]. Originally written within the laboratory of David Baker, the Rosetta developers community (RosettaCommons, https://www.rosettacommons.org/) has expanded to encompass hundreds of developers across tens of institutions. The Rosetta community, from software developers to academic and industry users, meets yearly to discuss exciting new advances, with 2012 marking the tenth anniversary RosettaCon. This 2012 Special Collection captures a selection of the scientific advancements in the two years since the last RosettaCon Special Collection [2].

Reproducibility and Computational Biology

In addition to highlighting exciting science, one of the primary goals of the RosettaCon Special Collections is to address issues of reproducibility in computational biology [3,4]. At first glance, "dry" computational biology seems inherently more reproducible than its "wet" experimental counterpart. All input to a computational experiment is precisely known and controlled, and the output is generated by a well-defined algorithm that is also under the programmer's control. In practice, reproducibility often suffers when the formatting requirements of a journal meet the massive datasets and complex workflows of modern computational biology.

Computational methods often synthesize a wide range of techniques to reach an interesting result, with multi-layered tasks that are difficult if not impossible to reproduce with a single command line argument. Yet documentation in the Methods section of a manuscript is rarely complete enough for outside experts to replicate these complex experiments, since the details of a protocol often mean the difference between success and failure. Monte Carlo algorithms can be particularly susceptible to these traps, since input data is often pre-processed, simulation output is stochastically generated, and the output is often significantly post-processed to synthesize a meaningful result. Without an accompanying test case, constant random number seed, and thorough description of the sampling necessary to obtain reasonable output, individuals attempting to replicate data may never be able to determine the root cause of "funny" results. Even in cases where core protocols are laboriously described, specialized pre- or post-processing scripts and programs written by former lab members further complicate matters and may even prevent replication within the originating lab.

Several causes may underly poor documentation and code distribution, including a reward system built upon high-profile papers as opposed to robust frameworks. But at the end of the day, the whole fields suffers as groups are forced to re-learn lessons obscured by time and poor documentation. Some projects are notable exceptions, such as bioperl/python/java and bioconductor [5–8], which freely distribute their source code under open source licenses and incorporate extensive documentation and tutorials. Not coincidentally, these projects enjoy widespread adoption and use, with tens to hundreds of thousands of downloads per year (http://www.bioperl.org/wiki/Getting_BioPerl, http://biopython.org/wiki/Download, http://www.bioconductor.org/packages/stats/).

Overview of Rosetta and the PLoS Collection

The Rosetta macromolecular modeling suite also enjoys widespread use, yet in the past has suffered from incomplete documentation, partially due to its extremely active development. Rosetta was originally developed for ab-initio protein structure prediction [9] but has evolved into a multi-purpose program that includes methods for template based modeling [10], protein-protein [11,12] and protein-DNA design [13], enzyme design [14,15], protein-protein [16] and protein-ligand docking [17], structure inference from limited experimental data [18], RNA structure prediction [19] and design, and peptidomimetic design [20].

Rosetta's rapid growth is fueled by the RosettaCommons, which is a non-profit entity that coordinates the development of the program and handles academic and commercial licensing. RosettaCommons (http://www.rosettacommons.org) is a collaboration between more than 15 research groups involved in the development of the Rosetta code base. The revenue generated by commercial licenses funds infrastructure for validation of code developments, users support, and developer meetings. The philosophy behind RosettaCommons is further described in the overview paper presenting the 2010 RosettaCon meeting [2].

In addition to addressing scientific problems via the Rosetta macromolecular modeling suite, the papers presented in this special collection tackle problems of reproducibility and documentation head-on. Publication of a paper in the collection is conditioned on the submission of an archive containing links to the exact version of the code used in the paper, all input data, links to external tools, and an example script to illustrate the use of the

Citation: André I, Corn J (2013) The RosettaCon 2012 Special Collection: Code Writ on Water, Documentation Writ in Stone. PLoS ONE 8(9): e73775. doi:10.1371/journal.pone.0073775

Editor: Vladimir N. Uversky, University of South Florida College of Medicine, United States of America

Published September 26, 2013

Funding: This work was supported by the Swedish Research Foundation (I.A.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared the following interests: Jacob Corn is an employee of Genentech. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: cornj@gene.com

◑ These authors contributed equally to this work.

code to carry out the protocol described in the paper. In addition, the paper is required to contain a detailed procedural description in the methods section. This "protocol capture" approach has also inspired a set of guidelines for how to present Rosetta computational workflows outside the PLOS collection. Importantly, the procedural description is used to audit each article, such that all protocols and documentation have been independently followed and verified to be complete by individuals outside the authors' laboratory. The large amount of testing data involved in this documentation is available via the RosettaCommons website (http://www.rosettacommons.org).

Naturally, while exhaustive documentation is necessary to recreate or modify a protocol, some users simply wish to try an established workflow on their favorite system, without spending large amounts of time deeply understanding the underlying theory or replicating test cases. However, the usage of many computational methods, including Rosetta, still requires considerable computational fluency and access to large computational resources, prohibiting wider use. This year's RosettaCon special collection addresses this need with the inclusion of ROSIE (Rosetta Online Server that Includes Everyone) [21], a general framework for the rapid development of public Rosetta web servers. Lowering the barrier to entry for the use of Rosetta protocols will hopefully democratize their use, such that the power of Rosetta becomes more accessible to a general audience.

Summary of papers

This special issue includes articles that describe a wide variety of aspects relating to the application of Rosetta in structure prediction and design. The articles can be divided into three categories: increasing the usability of Rosetta, improvements to current structure prediction methods, and completely new Rosetta procedures and applications. Each article is supplemented with full a "protocol capture," including documentation, test data, and processing scripts that have been peer reviewed by individuals outside the developers' research group. In a few cases the protocol capture is supplanted by a ROSIE web server interface to the application.

Lowering barriers to using Rosetta

Two articles in this special issue describe advancements that significantly lower the barriers for non-experts to use complex Rosetta applications. Lyskov et al. [21] introduce ROSIE (Rosetta Online Server that Includes Everyone); a framework for the serverification of Rosetta protocols. The ROSIE workflow allows Rosetta developers to rapidly convert Rosetta applications into web servers, all of which run on common hardware resources. This framework allows for the development of fully functional web servers for Rosetta applications within a few weeks. In a time scale of a few months nine servers based on the ROSIE framework have been launched, including two of the new applications described in this special issue [22,23].

Another means for lowering the barriers for non-experts is to provide a graphical user interface (GUI) to Rosetta. Adolf-Bryfogle and Dunbrack [24] describe the development of a GUI called the PyRosetta Toolkit, which allows users to to create and run common Rosetta molecular modeling and protein design tasks as well as analyze the results of Rosetta calculations. New applications can rapidly be modified to take advantage of the PyRosetta Toolkit.

Improvements to current structure prediction methods

Several articles describe improvements in Rosetta's structure prediction and design methodology. Drew et al [23] develop a framework to represent "nancanonical" peptidomimetic backbones in Rosetta, allowing the modeling and design of molecules such as peptoids and oligooxopiperazines. Notably, peptidomimetic design has already been incorporated into a ROSIE server. Alexander et al. [25] also explore the addition of new chemistries to Rosetta via improvements to RosettaEPR, a framework for using Electron Paramagnetic Resonance data to improve structure prediction. The new version of RosettaEPR includes a new rotamer library for a common spin label and more accurate reproduction of experimentally determined distance distributions.

Due to the astronomical size of protein conformation space, sampling is a long-standing bottleneck for Rosetta. Stein and Kortemme [26] find that significant improvements in loop conformational sampling can be achieved by combining several sampling strategies in the context of Rosetta. This strategy extends the KIC method [27] to yield even more accurate predictions of local conformations of proteins. Zhang and Lange [28] also tackle sampling, finding that a replica exchange approach greatly improves conformational sampling during the low resolution stage of RosettaDock. Khar et al. [29] have recently developed a ray-casting method (DARC) for small molecule docking and now demonstrate that its speed can be increased 25-fold via GPU-based computing, thereby enabling virtual screening of large compound libraries.

New Rosetta procedures and applications

New computational procedures and applications often debut at RosettaCon, and this issue contains several articles describing new Rosetta methodology. Lemmon and Meiler [30] introduce two methods for dealing with the challenging problem of performing small ligand docking with explicit interface water. Dong Nguyen et al. [31] provide a method for ligand docking into homology models of G-protein coupled receptors and present extensive benchmarking results. Although Rosetta protein design has recently achieved some landmark successes [12,13,15,32], the preparation of template "scaffold" proteins is non-trivial. Nivon et al [33]. describe a procedure to optimally pre-refine scaffold proteins prior to the computational design of functional sites. Computational design is also discussed by Der et al. [22], who explore two methods of automatically supercharging of protein surfaces to increase solubility. The authors experimentally test the performance of each method and have already made the supercharging protocol available as a ROSIE web server. Finally, Kahraman et al. [34] introduce protocols to drive both Rosetta de novo modeling and protein docking via the incorporation of experimental cross-linking data, as well as describe a structure-based crosslink database.

Conclusion

The Rosetta community has rapidly grown from a single lab to hundreds of people across many institutions, all contributing to (as of April, 2013) more than 1 million lines of code. As Rosetta expands in both users and developers, we must continually strive to keep the software readily available, transparent, and usable. This includes behind-the-scenes efforts, such as automated testing servers to ensure code robustness, as well as public outreach, such as help/announcement forums (https://www.rosettacommons.org/forum) and workshops (http://structbio.vanderbilt.edu/comp/workshops/rosetta_13/). The RosettaCon Special Collections and their associated protocol captures offer an accessible window into the fast-moving world of Rosetta development. We look forward to future Rosetta improvements to increase the availability of new Rosetta functionality, such as greatly accelerated release cycles, and hope that efforts such as the Special Collections ensure that bleeding-edge protocols are as usable as more established workflows.

Author Contributions

Wrote the paper: IA JC.

References

1. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, et al. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Meth Enzymol 487: 545–574. doi:10.1016/B978-0-12-381270-4.00019-6.
2. Renfrew PD, Campbell G, Strauss CEM, Bonneau R (2011) The 2010 Rosetta developers meeting: macromolecular prediction and design meets reproducible publishing. PLOS One 6: e22431. doi:10.1371/journal.pone.0022431.
3. Gentleman R (2005) Reproducible research: A bioinformatics case study. Statistical Applications in Genetics and Molecular....
4. Morin A, Urban J, Adams PD, Foster I, Sali A, et al. (2012) Research priorities. Shining light into black boxes. Science 336: 159–160. doi:10.1126/science.1218263.
5. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423. doi:10.1093/bioinformatics/btp163.
6. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611–1618. doi:10.1101/gr.361602.
7. Prlić A, Yates A, Bliven SE, Rose PW, Jacobsen J, et al. (2012) BioJava: an open-source framework for bioinformatics in 2012. Bioinformatics 28: 2693–2695. doi:10.1093/bioinformatics/bts494.
8. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80. doi:10.1186/gb-2004-5-10-r80.
9. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268: 209–225. doi:10.1006/jmbi.1997.0959.
10. Das R, Qian B, Raman S, Vernon R, Thompson J, et al. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 69 Suppl 8: 118–128. doi:10.1002/prot.21636.
11. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, et al. (2004) Computational redesign of protein-protein interaction specificity. Nat Struct Mol Biol 11: 371–379. doi:10.1038/nsmb749.
12. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, et al. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332: 816–821. doi:10.1126/science.1202617.
13. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, et al. (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 441: 656–659. doi:10.1038/nature04818.
14. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, et al. (2008) Kemp elimination catalysts by computational enzyme design. Nature 453: 190–195. doi:10.1038/nature06879.
15. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, et al. (2008) De novo computational design of retro-aldol enzymes. Science 319: 1387–1391. doi:10.1126/science.1152692.
16. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, et al. (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 331: 281–299.
17. Meiler J, Baker D (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins 65: 538–548. doi:10.1002/prot.21086.
18. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. Proceedings of the National Academy of Sciences 105: 4685–4690. doi:10.1073/pnas.0800256105.
19. Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. Proc Natl Acad Sci USA 104: 14664–14669. doi:10.1073/pnas.0703836104.
20. Renfrew PD, Choi EJ, Bonneau R, Kuhlman B (2012) Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design. PLOS One 7: e32637. doi:10.1371/journal.pone.0032637.
21. Lyskov S, Chou F-C, Conchúir SÓ, Der BS, Drew K, et al. (2013) Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). PLOS One 8: e63906. doi:10.1371/journal.pone.0063906.
22. Der BS, Kluwe C, Miklos AE, Jacak R, Lyskov S, et al. (2013) Alternative computational protocols for supercharging protein surfaces for reversible unfolding and retention of stability. PLOS One 8: e64363. doi:10.1371/journal.pone.0064363.
23. Drew K, Renfrew PD, Craven TW, Butterfoss GL, Chou F-C, et al. (2013) Adding diverse noncanonical backbones to rosetta: enabling peptidomimetic design. PLOS One 8: e67051. doi:10.1371/journal.pone.0067051.
24. Adolf-Bryfogle J, Dunbrack RL (2013) The PyRosetta Toolkit: A Graphical User Interface for the Rosetta Software Suite. PLOS One 8: e66856. doi:10.1371/journal.pone.0066856.
25. Alexander NS, Stein RA, Koteiche HA, Kaufmann KW, Mchaourab HS, et al. (2013) RosettaEPR: Rotamer Library for Spin Label Structure and Dynamics. PLOS One 8: e72851. doi:10.1371/journal.pone.0072851.
26. Stein A, Kortemme T (2013) Improvements to robotics-inspired conformational sampling in rosetta. PLOS One 8: e63090. doi:10.1371/journal.pone.0063090.
27. Mandell DJ, Coutsias EA, Kortemme T (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods 6: 551–552. doi:10.1038/nmeth0809-551.
28. Zhang Z, Lange OF (2013) Replica Exchange Improves Sampling in Low-Resolution Docking Stage of RosettaDock. PLOS One 8: e72096.
29. Khar KR, Goldschmidt L, Karanicolas J (2013) Fast Docking on Graphics Processing Units via Ray-Casting. PLOS One 8: e70661. doi:10.1371/journal.pone.0070661.
30. Lemmon G, Meiler J (2013) Towards ligand docking including explicit interface water molecules. PLOS One 8: e67536. doi:10.1371/journal.pone.0067536.
31. Nguyen ED, Norn C, Frimurer TM, Meiler J (2013) Assessment and challenges of ligand docking into comparative models of G-protein coupled receptors. PLOS One 8: e67302. doi:10.1371/journal.pone.0067302.
32. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, et al. (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 336: 1171–1174. doi:10.1126/science.1219364.
33. Nivón LG, Moretti R, Baker D (2013) A Pareto-optimal refinement method for protein design scaffolds. PLOS One 8: e59004. doi:10.1371/journal.pone.0059004.
34. Kahraman A, Herzog F, Leitner A, Rosenberger G, Aebersold R, et al. (2013) Cross-Link Guided Molecular Modeling with ROSETTA. PLOS One 8: e73411. doi:10.1371/journal.pone.0073411.