CASS: Protein sequence simulation with explicit genotype-phenotype mapping

Johan A. Grahnen, David A. Liberles

Department of Molecular Biology, University of Wyoming, Laramie, WY USA

Abstract

CASS (coarse-grained artificial sequence simulator) is a software package for simulating protein sequences with an explicit genotype-to-phenotype mapping that takes protein structure and function into account. It is capable of reproducing many structure-specific properties of protein sequence evolution, most notably spatial and temporal variation in rates, and has been used to investigate several hypotheses about the influence of thermodynamics on molecular evolution. The software is implemented in object-oriented C++, is supported on Linux, and the source code is made freely available under the GPL v3 license at http://www.wyomingbioinformatics.org/LiberlesGroup/CASS/.

Report

In the field of protein evolution, simulation is an invaluable tool and simulation-based approaches have been widely used. Examples include testing the accuracy of ancestral character reconstruction,[1] evaluating the statistical power to detect selection,[2] benchmarking methods for multiple sequence alignment,[3] and phylogenetic tree reconstruction,[4] and addressing the interplay between complex evolutionary processes where biological functions and selection can be controlled.

However, previous methods have not explicitly modeled the interdependence between sites caused by selection on protein structure and function. One class of methods is strictly phenomenological, simulating the substitution process without any connection between genotype and phenotype.[5-7] Another type of approach does include a weak genotype-to-phenotype map, either site-specific and predefined or a statistically assigned random mapping.[8-12] Sequence simulation in a number of evolutionary contexts can be found in Arenas M. and Hoban *et al*.[13,14]

Although some authors have recently modeled the influence of structure on sequence evolution,[15-17] function is not considered and models are not readily available as software. A new biophysical model of sequence evolution was developed that accounts for selection to fold into a specific conformation and bind to a specific partner, the structural consequences of mutations, and the influence of population size and mutation rate on fixation probabilities.[18] This work briefly describes that model and its implementation in software, the Coarse-grained Artificial Sequence Simulator (CASS).

In CASS (Figure 1), mutations occur on the codon level and are translated to protein. Protein sequences are then threaded through a constant target protein backbone, and side chains of mutated residues are re-packed using either a coarse-grained approximation or a slower all atom approach which is subsequently converted to coarse-grained space.[19-21]

The re-packed structure is then assessed for changes in unfolding stability, misfolding stability, and stability in alternative structures. The stability scoring function takes into account proper packing, van der Waals forces, salt bridges, disulfide bonds, secondary structure, and the hydrophobic effect. The target conformation is compared to alternative conformations to ensure folding into a specified structure. This calculation uses a coarse-grained structural representation to achieve the necessary speed to evaluate the millions of sequences produced during population-scale simulations or phylogenetic inference procedures.

Function in the form of protein-ligand interaction is also modeled. Binding stability is scored using the non-bonded interaction terms of the folding model. Binding specificity is modeled by selection against deleterious interactions as well as selection for the native interaction. This approach generates functional specificity and influences the substitution rate at the binding interface.[22]

Finally, CASS allows modeling of the interplay between population size, mutation rate and selection by simulating a population of virtual organisms, each assigned a fitness value based on folding stability and binding function.[23] Organisms are propagated across generations at random, weighted by fitness, mimicking the opposing forces of genetic drift and selection. Simulation of sequences over a phylogenetic tree is possible, and heterogeneous processes are easy to accommodate. For instance, selective pressures may be altered on a clade-specific basis or mutation rates can be different for different species. Generally, any parameter can be changed on any branch of the tree, and the use of this feature is described in the accompanying documentation. With a script that describes the tree structure, the population along any phylogenetic lineage can be simulated for a specified number of generations identically to that in a single population in forward time.

As software, CASS has been implemented in object-oriented C++ with modern memory-management techniques, supported on Linux distributions with a recent compiler (for example, GCC v 4.3 or higher; the accompanying website provides help with compilation and installation). It is provided as a collection of classes for representing protein structure, folding and binding stability scoring, and simulation conditions. These classes can easily be combined to create programs to, for example, simulate sequence evolution or sample near-native protein sequences. Such modularity also ensures that it is an easy programmatic change to alter the software to fit novel research questions. The software package comes with example applications for common tasks and a variety of models of thermodynamics, descriptions of protein structure and fitness functions. An example of the application of the method can be found in a recent publication.[18]

The various components were tested to ensure that they produce the expected output. For instance, it was confirmed that simulated DNA sequences sustain the same average number of mutations as specified in the input, and that protein sequences under no selection

Correspondence: David A. Liberles, Department of Molecular Biology,University of Wyoming, 1000 E. University Ave, Laramie, WY 82071, USA.
Tel. +1.307.7665206 - Fax: +1.307.7665098.
E-mail: liberles@uwyo.edu

Acknowledgments: the authors would thank Russell Hermansen for supporting the CASS website. This work was supported by the National Institutes of Health INBRE program (grant number P20 RR016474) and DAL received additional support from the National Science Foundation (grant number DBI-0743374).

Key words: sequence-structure-function relationship, simulation, molecular evolution.

Contributions: JAG contributed to the design of the algorithms, wrote the simulation software and co-wrote the manuscript; DAL conceived of the study, contributed to the design of the algorithms and co-wrote the manuscript.

Conflict of interests: the authors report no conflict of interests.

Received for publication: 13 July 2012.
Revision received: 8 October 2012.
Accepted for publication: 8 October 2012.

Trends in Evolutionary Biology 2012; 4:e9
doi:10.4081/eb.2012.e9

diverge at the rate expected from population genetic theory.[24] Although the simulations are inherently time-consuming due to the complexity of the model, substantial efforts were made to optimize the algorithms involved. For example, caching of the structural and energetic consequences of previously encountered mutations provides a speed-up proportional to the number of simulated organisms in regimes with realistically low mutation rates. The calculation-memory tradeoffs mean that a typical simulation (sampling on the order of 10^6 mutations) will consume ~5 Gb of memory over a modest runtime of ~48 hours on a modern desktop machine.

The design of the approach was subjected to four levels of biological validation. First, native sequences that are known to fold into a given structure should be well described by the model and the model should generate sequences that show specificity for the fold they were evolved to fit. Second, evolved proteins should retain a hydrophobic core and a hydrophilic surface. Third, evolved proteins should show heterogeneity of rates across sites, with the hydrophobic core generally evolving more slowly than the hydrophilic surface. Lastly, the structure and function should impose selective constraint due to evolved metastability,[25] resulting in dN/dS ratios similar to those known from comparative genomics,[26] *i.e.* approximately 0.2. As described in Grahnen *et al.*,[18] the method can be made to show the last three levels of biological validation, but has a tendency to evolve homopolymer runs that appear not to be fold-specific. The simulation approach is still an active area of research and release of models as open source software enables users to change the force field or mutation acceptance rules. Ultimately, the software is useful currently for applications involving evolutionary dynamics and as a framework for future developments by the field.

Despite its simplicity relative to the complicated reality of protein folding and protein-protein interactions,[27,28] the model has been applied to several problems in sequence evolution. For example, this approach has shown how selection for not binding to deleterious ligands could be a major driver in restricting diversity of sequence and function at protein binding interfaces.[22] As one of the strengths of the model is reproduction of spatial and temporal variation in rates of evolution, it was recently used to study how selectively driven shifts in function induce shifts in rates (Grahnen *et al.*, submitted), and how that compares with neutral rate-shifting caused by compensatory substitutions for folding stability (dynamics described by Pollock *et al.*)[29]

The approach described above has many other potential applications. Aside from the benchmarking tasks of validating methods of alignment and phylogenetic reconstruction, one might use it to derive expectations of rates of evolution in the many possible fates of gene duplicates, test hypotheses about the effects of population size and mutation rate on the evolution of evolvability and robustness, characterize the fit of novel models of sequence evolution to data with known evolutionary history, and answer questions about the nature of the genotype-to-phenotype map for proteins. The ability to simulate sequence evolution with a high degree of biological realism, with software such as CASS, should be of considerable aid in developing novel inference methods and asking questions about evolutionary processes which cannot directly be observed.

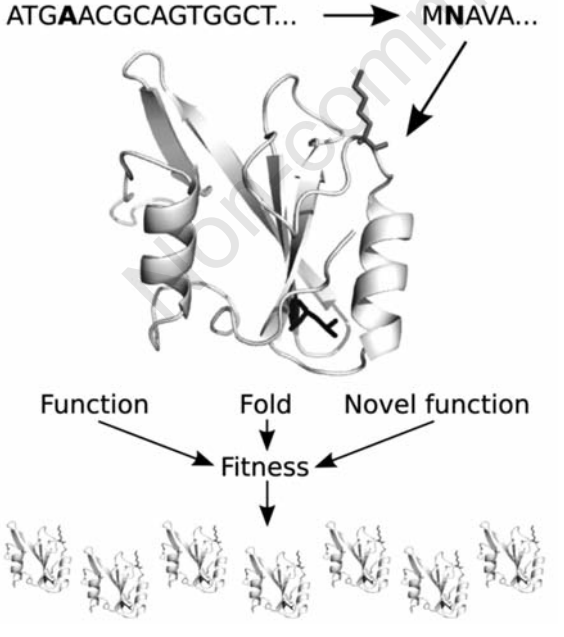

Figure 1. An overview of the method. Mutations (bold) occur in coding DNA, are translated to protein, and their consequences for local packing are modeled (grey surface mutant: unlikely to be disruptive; black buried mutant: possibly highly disruptive). The resulting structure is scored for folding stability, maintenance of function and development of novel function, and assigned a combined fitness score. The fitness of each sequence influences the frequency of the allele in the next generation of the explicitly simulated population.

References

1. Zhang J, Nei M. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol 1997;44:S139-46.
2. Anisimova M, Yang Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 2007;24:1219-28.
3. Liu K, Raghavan S, Nelesen S, et al. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 2009;324:1561-4.
4. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 2009;26:1641-50.
5. Rambaut A, Grass NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 1997;13: 235-8.
6. Stoye J, Evers D, Meyer F. Rose: generating sequence families. Bioinformatics 1998; 14: 157-63.
7. Cartwright RA. DNA assembly with gaps (Dawg): simulating sequence evolution. Bioinformatics 2005;21Suppl3:iii31–iii38.
8. Strope CL, Abel K, Scott SD, Moriyama EN. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. Mol Biol Evol 2009; 26:2581-93.
9. Sipos B, Massingham T, Jordan GE, Goldman N. PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment. BMC Bioinformatics 2011;12:104.
10. Koestler T, Von Haeseler A, Ebersberger I. REvolver: modeling sequence evolution

under domain constraints. Mol Biol Evol 2012;29:2133-45.
11. Hall BG. Simulating DNA Coding sequence evolution with EvolveAGene 3. Mol Biol Evol 2008;25:688-95.
12. Fletcher W, Yang Z. INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol 2009;26:1879-88.
13. Arenas M. Simulation of molecular data under diverse evolutionary scenarios. PLoS Comput Biol 2012;8:e1002495.
14. Hoban S, Bertorelle G, Gaggiotti OE. Computer simulations: tools for population and evolutionary genetics. Nat Rev Gen 2012;13:110.
15. Kleinman CL, Rodrigue N, Lartillot N, Philippe H. Statistical potentials for improved structurally constrained evolutionary models. Mol Biol Evol 2010;27: 1546-60.
16. Lakner C, Holder MT, Goldman N, Naylor GJP. What's in a likelihood? Simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. Syst Biol 2011;60:161-74.
17. Nasrallah CA, Mathews DH, Huelsenbeck JP. Quantifying the impact of dependent evolution among sites in phylogenetic inference. Systematic Biology 2011;60:60-73.
18. Grahnen JA, Nandakumar P, Kubelka J, Liberles DA. Biophysical and structural considerations for protein sequence evolution. BMC Evol Biol 2011;11:361.
19. Lathrop RH. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng 1994;7:1059-68.
20. Grahnen JA, Kubelka J, Liberles DA. Fast side chain replacement in proteins using a coarse-grained approach for evaluating the effects of mutation during evolution. J Mol Evol 2011;73:23-33.
21. Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009;77:778-95.
22. Liberles DA, Tisdell MDM, Grahnen JA. Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. Proceedings of the Royal Society B: Biological Sciences 2011;278:1930-5.
23. Lynch M. The origins of genome architecture. Sunderland, MA: Sinauer Associates Inc; 2007.
24. Hartl DL, Clark AG. Principles of population genetics. Sinauer Associates; 2006.
25. Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins 2002; 46:105-9.
26. Roth C, Liberles DA. A systematic search for positive selection in higher plants (Embryophytes). BMC Plant Biol 2006; 6:12.
27. Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. Annu Rev Biophys 2008;37:289-316.
28. Zacharias M. Accounting for conformational changes during protein-protein docking. Curr Opin Struct Biol 2010; 20:180-6.
29. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary stokes shift. PNAS 2012;109: E1352-9.