PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

SMOG 2: A Versatile Software Package for Generating Structure-Based Models

Jeffrey K. Noel[1,2]*, Mariana Levi[3], Mohit Raghunathan[1], Heiko Lammert[1], Ryan L. Hayes[1], José N. Onuchic[1]*, Paul C. Whitford[3]*

1 Center for Theoretical Biological Physics, Rice University, Houston, Texas, United States of America, **2** Kristallografie, Max Delbrück Center for Molecular Medicine, Berlin, Germany, **3** Department of Physics, Northeastern University, Boston, Massachusetts, United States of America

* jeffrey.noel@mdc-berlin.de (JKN); jonuchic@rice.edu (JNO); p.whitford@neu.edu (PCW)

OPEN ACCESS

Citation: Noel JK, Levi M, Raghunathan M, Lammert H, Hayes RL, Onuchic JN, et al. (2016) SMOG 2: A Versatile Software Package for Generating Structure-Based Models. PLoS Comput Biol 12(3): e1004794. doi:10.1371/journal.pcbi.1004794

Editor: Andreas Prlic, UCSD, UNITED STATES

Received: August 31, 2015

Accepted: February 7, 2016

Published: March 10, 2016

Data Availability Statement: Software is GPL licensed and is publicly available at https://bitbucket.org/smog-server

Funding: Work at the Center for Theoretical Biological Physics was sponsored by the NSF (Grants No. PHY- 1427654 and No. MCB-1214457) and by the Welch Foundation (Grant No. C-1792). PCW was supported by a NSF CAREER Award (Grant No. MCB-1350312). JKN is an Alexander von Humboldt Postdoctoral Fellow. JNO acknowledges support as a CPRIT Scholar in Cancer Research sponsored by the Cancer Prevention and Research Institute of Texas. Computing resources were supported in part by the Cyberinfrastructure for

Abstract

Molecular dynamics simulations with coarse-grained or simplified Hamiltonians have proven to be an effective means of capturing the functionally important long-time and large-length scale motions of proteins and RNAs. Originally developed in the context of protein folding, structure-based models (SBMs) have since been extended to probe a diverse range of biomolecular processes, spanning from protein and RNA folding to functional transitions in molecular machines. The hallmark feature of a structure-based model is that part, or all, of the potential energy function is defined by a known structure. Within this general class of models, there exist many possible variations in resolution and energetic composition. SMOG 2 is a downloadable software package that reads user-designated structural information and user-defined energy definitions, in order to produce the files necessary to use SBMs with high performance molecular dynamics packages: GROMACS and NAMD. SMOG 2 is bundled with XML-formatted template files that define commonly used SBMs, and it can process template files that are altered according to the needs of each user. This computational infrastructure also allows for experimental or bioinformatics-derived restraints or novel structural features to be included, e.g. novel ligands, prosthetic groups and post-translational/transcriptional modifications. The code and user guide can be downloaded at http://smog-server.org/smog2.

This is a *PLOS Computational Biology* Software Article.

Introduction

The study of biomolecular folding has produced theoretical concepts that are generalizable to many processes, such as conformational rearrangements in proteins and the functional dynamics of molecular assemblies. In particular, the principle of minimal frustration [1] and the

Computational Research funded by NSF under Grant No. CNS-0821727. Additionally, computing was supported by the National Science Foundation through XSEDE resources provided by the Texas Advanced Computing Center under grant Nos. TG-MCB110021 and MCB140274, the C3DDB Cluster, and the Northeastern University Discovery Cluster.

Competing Interests: The authors have declared that no competing interests exist.

folding funnel concept [2, 3] describe an energy landscape where the native interactions (i.e. the molecular interactions present in low free-energy configurations of folded proteins and RNAs) are on average more stabilizing than non-native interactions. Thus, the effective energetics of a biomolecule can be well described by a set of stabilizing native interactions, along with excluded volume to prevent chain crossing. Potential energy functions of this type are known as "structure-based models," (SBMs) and they are powerful tools for probing the relationship between structure, folding and function in biomolecular systems. The simplified character of the potential energy function allows for reduced computational requirements, and the explicitly-encoded native interactions provide a baseline model for molecular modeling, or for studying physical perturbations. For a detailed discussion of the theoretical foundation and applications of SBMs, the reader is referred to the following reviews [4, 5] and the references therein.

SBMs were first extensively used to explore the predictions of energy landscape theory in the context of protein folding [6–15]. These studies showed that minimally-frustrated protein models reproduce many thermodynamic features of real proteins, and the predicted transition state ensembles are frequently consistent with experimental findings [8, 9, 16–18]. In addition to folding, studies used SBMs to show that protein binding could be understood within a common theoretical framework [19, 20]. Since protein function is governed by the same energy landscape that determines folding dynamics [21], these models have also been used to study the conformational dynamics involved in macromolecular function, e.g. adenylate kinase [22], kinesin [23, 24], and the ribosome [25]. These models have structural resolutions that vary from a single bead per residue [10], to all heavy atoms being explicitly represented [26], and their energetic complexity varies from "perfectly-funneled" landscapes, to Hamiltonians that include various flavors of non-native interactions [27–29]. Recently, SBMs have found utility in molecular modeling applications. For example, MDfit combines SBMs and cryogenic electron microscopy data to create atomically-grained structural models that are consistent with experimental electron densities [30]. Another example is SBM+DCA, where SBMs include co-evolutionary residue-residue interactions to predict difficult-to-crystallize oligomers [31, 32]. Together, SBMs (sometimes called "Go-models" [33]) have a thirty year history that spans countless applications, where the common feature is that biomolecular contacts present in high-resolution structures are given stabilizing energetics.

With the versatility of SBMs, investigators often apply customizations that are tailored to address specific physical questions. This contrasts with the more linear development of empirical explicit-solvent potentials, which is driven by the reproduction of experimental observables for model systems. As a result, SBM development has been decentralized, which has limited the portability and transferability of the models. Web servers [34, 35] that produce output for running SBMs on modern molecular dynamics (MD) packages have been very popular, and have provided some degree of standardization. However, since these web servers only provide the specific variations of the models that the developers decide to support, modifications made by the general community are typically unavailable to other researchers.

SMOG 2 is intended to facilitate SBM development by allowing modifications and extensions to be easily shared by the research community. In SMOG 2, an SBM potential is translated into a template format, allowing forcefields to be easily disseminated and modified. SMOG 2 processes user-designated structural information provided in standard Protein Data Bank (PDB) format and a SMOG 2 template, in order to generate the forcefield files required to perform simulations with MD platforms. Two of the most widely used MD platforms, GROMACS [36] and NAMD [37], support SMOG 2 output files. SMOG 2 is licensed under the GNU GPL and the source code is publicly available. See http://smog-server.org/smog2 for details and the user guide.

Design and Implementation

Template-based design

Many functional biological macromolecules are polymers of amino acids or nucleic acids, the building blocks of proteins, RNA and DNA. Each residue has a unique set of atoms, called the side chain in proteins (or base in nucleic acids), and a common set of atoms that constitutes the polymer backbone. Thus, an intuitive approach for defining the covalent connectivity within a biomolecule is to predefine the covalent structure of each possible residue and then map these interactions on a per-residue basis. Conditions must also be provided that ensure adjacent residues are covalently linked. In addition, generic non-bonded (non-covalent) interactions between atoms can be defined by assigning each atom a chemical "type" and then specifying the functional forms of the interactions between all possible combinations of types. This approach is sufficient to describe any polymer sequence composed of the predefined building blocks. SMOG 2 adopts this strategy for defining SBMs, which is consistent with the organization used for semi-empirical models, such as AMBER [38], CHARMM [39], and GROMOS [40]. This consistency in the construction of the models also allows the interactions defined in semi-empirical models to be mapped to SMOG 2 syntax in order to construct hybrid-variants of these models.

SMOG 2 templates are written in XML (eXtensible Markup Language) for readability and standardization. A SMOG 2 template, which defines a specific forcefield, is comprised of four files with the following suffixes:

- **.bif**: Defines the atoms and bonds in each residue and their connectivity. Any atom names may be used, though the naming between the .bif and input PDB file must be consistent.
- **.sif**: Defines the available functional forms for interaction potentials and system-wide energetic settings.
- **.b**: Sets the specific functional forms to be applied for bonded interactions between atom types.
- **.nb**: Sets the specific functional forms to be applied for non-bonded interactions between atom types.

The included templates (see section **Included templates**) follow standard PDB nomenclature for simplicity. Internally, the code makes no assumptions about the molecular structure corresponding to specific residue names or the interactions associated with specific atom names. Thus, adding new ligands and residue types involves defining the consituent atom names and their covalent bonds in the .bif. Each atom has three associated "types" that can be used to control the interactions between atoms: bonded-type, non-bonded-type, and pair-type. These parameters define how to map the bonded interactions, non-native non-bonded interactions, and native contact interactions, respectively. It should be pointed out that irregular molecular chains (i.e. without a common backbone) such as polysaccharides cannot be automatically handled. To accommodate for these types of irregular chains, the inter-residue bonds must be explicitly defined in the PDB file, as described in the SMOG 2 manual.

SBM Hamiltonians are defined by the input structure. The main purpose of SMOG 2 is to facilitate the creation of input files for simulations that contain structure-based interactions. For our purposes, a "structure-based interaction" is an interaction that is parameterized by the atomic coordinates of a known, low-free-energy configuration (e.g., an X-ray crystallographic structure). In a "pure" SBM, the global minimum of the Hamiltonian is encoded as the configuration of the input (native) structure by explicitly defining the native value of each interaction

to be the potential energy minimum. For illustration, consider the Hamiltonian of a commonly used coarse-grained SBM [7], where each protein residue is represented by a bead at the position of the C_α atom:

$$\begin{aligned} H_{C\alpha}(\vec{\mathbf{x}},\vec{\mathbf{x}}^0) &= \sum_{ij\in\text{bonds}} \frac{\epsilon_b}{2}(r_{ij}-r_{ij}^0)^2 + \sum_{ijk\in\text{angles}} \frac{\epsilon_\theta}{2}(\theta_{ijk}-\theta_{ijk}^0)^2 + \sum_{ijkl\in\text{dihedrals}} \epsilon_D F_D(\varphi_{ijkl}-\varphi_{ijkl}^0) \\ &+ \sum_{ij\in\text{contacts}} \epsilon_C\left[5\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij}^0}{r_{ij}}\right)^{10}\right] + \sum_{ij\notin\text{contacts}} \epsilon_{NC}\left(\frac{\sigma_{NC}}{r_{ij}}\right)^{12}. \end{aligned} \quad (1)$$

The dihedral potential F_D is

$$F_D(\delta\varphi) = [1-\cos(\delta\varphi)] + \frac{1}{2}[1-\cos(3\delta\varphi)].$$

The backbone structure is maintained by harmonic bonds and angles, the secondary and tertiary structure is stabilized by dihedral and short-range contact potentials, and all beads interact through an excluded volume interaction. Contacts are defined as being between residue pairs that are in spatial proximity in the native structure [41]. The superscript 0 denotes that a parameter is calculated from the input structure, which is used to explicitly set the global minimum of the potential to the input configuration.

Including native structural information and coarse-graining. SMOG 2 was written specifically for Hamiltonians of the general form shown in Eq 1. In the SMOG 2 templates a question mark is used to indicate that a parameter should be calculated from the native structural information. For example, in the .b file for the Hamiltonian in Eq 1, the bond function would be declared as:

```
<bond func="bond_harmonic(?,20000)">
  <bType>*</bType>
  <bType>*</bType>
</bond>
```

This specifies that a harmonic bond potential with $\epsilon_b = 20000$ be given between an atom pair *ij* of the indicated bonded types (bType). In this case, the asterisks stipulate that this function be applied to all bType combinations that are not explicitly defined elsewhere in the .b file. Note that the units should be consistent with how GROMACS implements reduced units; for a more detailed discussion, see the user manual. As noted above, a unique feature of SMOG 2 is the question mark special character. In this example, the question mark specifies that the native distance r_{ij}^0 should be used to define the minimum of the harmonic potential. This question mark syntax can be similarly used for any interaction term in Eq 1. To provide a parameter that is independent of structure, such as σ_{NC}, which defines the excluded volume between the beads, a numerical value should be provided in place of a question mark.

SMOG 2 implements automatic coarse-graining by using two templates internally, one atomistic template that is consistent with the input PDB structure, and one coarse-grained template. The coarse-grained template specifies one atom within each residue to map interactions and include in the simulation model. This feature is useful for creating single-bead models of proteins, such as the commonly-used C_α-model of Clementi, Nymeyer and Onuchic [7]. Coarse-grained geometries differing from a single-bead-per-residue representation can be implemented by creating a template consistent with a preprocessed PDB structure containing only the coarse-grained atoms. Note that the Shadow.jar contact map generation tool is only intended for use with a structure containing all the heavy atoms. Thus, for general coarse-

Table 1. Description of the SMOG 2 templates included in the distribution. Except where noted, the native contact map is generated by the Shadow algorithm [41] using an input all-atom PDB structure. The elastic network model is in the same spirit as Tirion's [46], but the contact map is different and the spring stiffness is system independent.

Template	Ref.	Description
`SBM_AA`	[26]	All heavy atoms explicitly represented, Lennard-Jones potentials for native atomic contacts, handles RNA/DNA/protein/ligands
`SBM_AA+gaussian`	[42, 43]	`SBM_AA` with Gaussian potentials for native atomic contacts
`SBM_AA_charged`	[44]	`SBM_AA` with charged ARG, LYS, GLU, ASP, N/C-terminal
`SBM_CA`	[7]	Single C_α bead per residue, Lennard-Jones potentials for native residue contacts, developed for proteins
`SBM_CA+gaussian`	[45]	`SBM_CA` with Gaussian potentials for native residue contacts
`ENM`		All-atom elastic network model, harmonic potentials for native atomic contacts, 6 Å cutoff determines native contact map

doi:10.1371/journal.pcbi.1004794.t001

graining, the native contacts will have to be mapped onto the coarse-grained atoms by the user and be provided to SMOG 2 as input.

Included templates. SMOG 2 is packaged with templates for some commonly used structure-based Hamiltonians [7, 26, 42–46] (Table 1). These templates can be used as-is or modified to create new SBM variants. Users that generate new templates are encouraged to make them publicly available through the SMOG webpage. This can help provide transparency and encourage collaboration.

Code implementation

SMOG 2 is written in the Perl programming language, and it uses the Perl Data Language (PDL) for its primary data structures. PDL extends the native Perl data structures by allowing for large multidimensional arrays that can be manipulated through vector-based operations. PDL arrays are more compact, and can be manipulated faster than native Perl arrays. This is important for the most computationally intensive task performed by SMOG 2, which is to dynamically calculate all angles and dihedrals that can exist in a molecule based on the bonded geometry. The Perl implementation has a few dependencies: String::Util, XML::Simple, Exporter, and XML::Validator::Schema. Additionally, the Java Runtime Environment (JRE 1.7 or greater) is necessary for SMOG 2 to call the included Shadow.jar contact map tool [41].

In order to ensure that SMOG 2 is properly configured, test modules are available to the user as a separate download (smog-check) from the SMOG website. The smog-check bundle contains two test programs. One is a basic check that ensures that the local installation reproduces benchmark output files (.top, .gro). The second testing suite is a rigorous test-driven-development package that inspects the output of SMOG 2 for accuracy after code modifications. SMOG 2 has been extensively beta tested and exception-driven-development (i.e. checking for previously encountered errors and providing feedback to the user on how to correct the errors) has been implemented throughout the code.

Workflow

SMOG 2 is invoked from the command line. The two necessary inputs are 1) a biomolecular structure in PDB format and 2) a directory name containing the set of SMOG 2 templates. Users are encouraged to use the included tool `smog_adjustPDB`, which resolves common formatting/naming inconsistencies between standard PDB format and the default templates.

The templates define the general form and parameters of the Hamiltonian. SMOG 2 can process default templates that are included within the package (Table 1), as well as user-generated templates. The native contact map can be either automatically generated or provided as input. Running SMOG 2 generates output files that are formatted for input to MD software packages (in GROMACS format, which can also be read by NAMD). At a minimum, two of the generated files are required in order to run a simulation:

- **.gro**: The coordinates of the input PDB structure in GRO format. This is often used as the initial configuration for MD simulations.
- **.top**: The topology file specifies the Hamiltonian by listing all atomic interactions.

The user's manual and the SMOG web server [34] both provide tutorials for using the generated files to perform MD simulations. While SMOG 2 can be used to generate a wide range of possible models, for some extended SBM variants, it will be necessary to further process the topology files. For example, combining multiple SBMs into a multi-basin landscape is a commonly used technique that is not automatically handled by SMOG 2. This task and other useful post-processing of topology files can be performed with the Python-based eSBMTools [47].

Results and Discussion

Protein folding with the default models

As an illustration of the types of SBM variants that can be explored with SMOG 2, folding simulations of the well studied protein chymotrypsin inhibitor 2 (CI2) [48] are considered. The results for two different models are shown, a single-bead-per-residue graining [7] and an all-heavy-atom graining [26] (SBM_CA and SBM_AA in Table 1, respectively), using the input run parameters suggested in the user's manual (Fig 1). A standard reaction coordinate for the analysis of biomolecular folding is the fraction of native structure formed, often called Q [7, 49, 50]. The SMOG-enhanced version of GROMACS v4.5 available on the SMOG website contains the tool "g_kuh," which analyzes trajectories using native structural measures, including Q. Consistent with the experimentally-observed two-state folding dynamics of CI2 [51], plotting the free energy as a function of Q shows two basins at the folding temperature (T_F). There is a folded basin at high Q and an unfolded basin at low Q, which are separated by a free-energy barrier (Fig 1C). Here, Q is defined as the fraction of natively-contacting residue pairs that are within 1.5 times their native distance.

Using SMOG 2 to explore multiple levels of structural detail

In addition to models with C_α or all-atom resolution, SMOG 2 templates can be modified to describe any level of structural detail. For example, included in the distribution is a template that accommodates the explicit representation of hydrogens (Fig 2). This template, called "SBM_AA+hydrogen", uses heterogeneous atomic radii modeled from the vdW parameters in the Amber99sb forcefield [38]. In contrast to the other included templates, there are multiple non-bonded types and associated changes, which can serve as an example of how to manipulate the SMOG 2 template syntax.

Here, we use the SBM_AA+hydrogen template to study protein folding with models that have identical native contact potentials, but differing levels of geometric detail. This provides a baseline test of the effects of atom size and molecular geometry on the folding landscape. The free-energy profiles along Q are shown for three SBMs of CI2, two with uniformly-sized heavy atoms of diameters 1.7 Å and 2.5 Å (parameter σ_{NC} in Eq 1), and one using the SBM_AA+-hydrogen template (Fig 2A). These three models are denoted $M^{1.7}$, $M^{2.5}$, and M^{+H}, respectively.

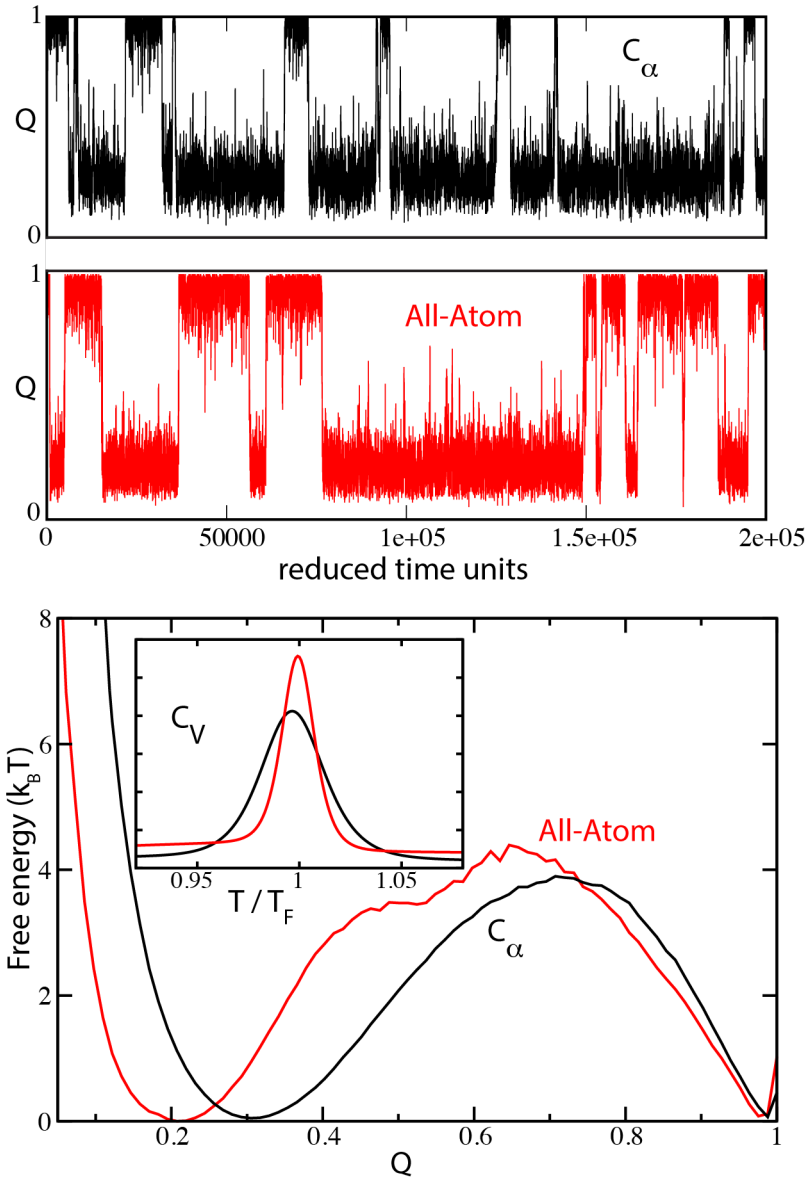

Fig 1. Protein folding simulations with the default C_α and all-atom models of the 64 residue chymotrypsin inhibitor 2 (PDB code: 1YPA). Top: Folding trajectories near folding temperature (T_F) of the C_α (black) and all-atom (red) models. Bottom: Free energy as a function of $Q_{C\alpha}$, the number of native C_α pairs within 1.5 times their native distance. The same coordinate is used to describe both models. Inset: Specific heat for the two models (normalized to have equal area). T_F in reduced units for the all-atom model is 0.97 and for the C_α model is 1.17 (117 and 140 in the GROMACS .mdp file, respectively).

doi:10.1371/journal.pcbi.1004794.g001

Notably, increasing the excluded volume raises the folding barrier and lowers the folding temperature (T_F) [41]. $M^{2.5}$ has a folding barrier of the same height as M^{+H}, though M^{+H} has a single flat barrier shape and $M^{2.5}$ has a significant shoulder. The overall character of CI2 folding is consistent between the three models: two-state kinetics with a barrier centered around $Q = 0.4$. Differences in folding mechanism can be discerned by comparing the average contact formation in the barrier region. Visual inspection of average contact maps in the upper triangles of Fig 2 (panels B-D) shows that the transition state ensemble (TSE) is highly similar between the models. The detailed differences between M^{+H} and the models with uniform atom sizes are

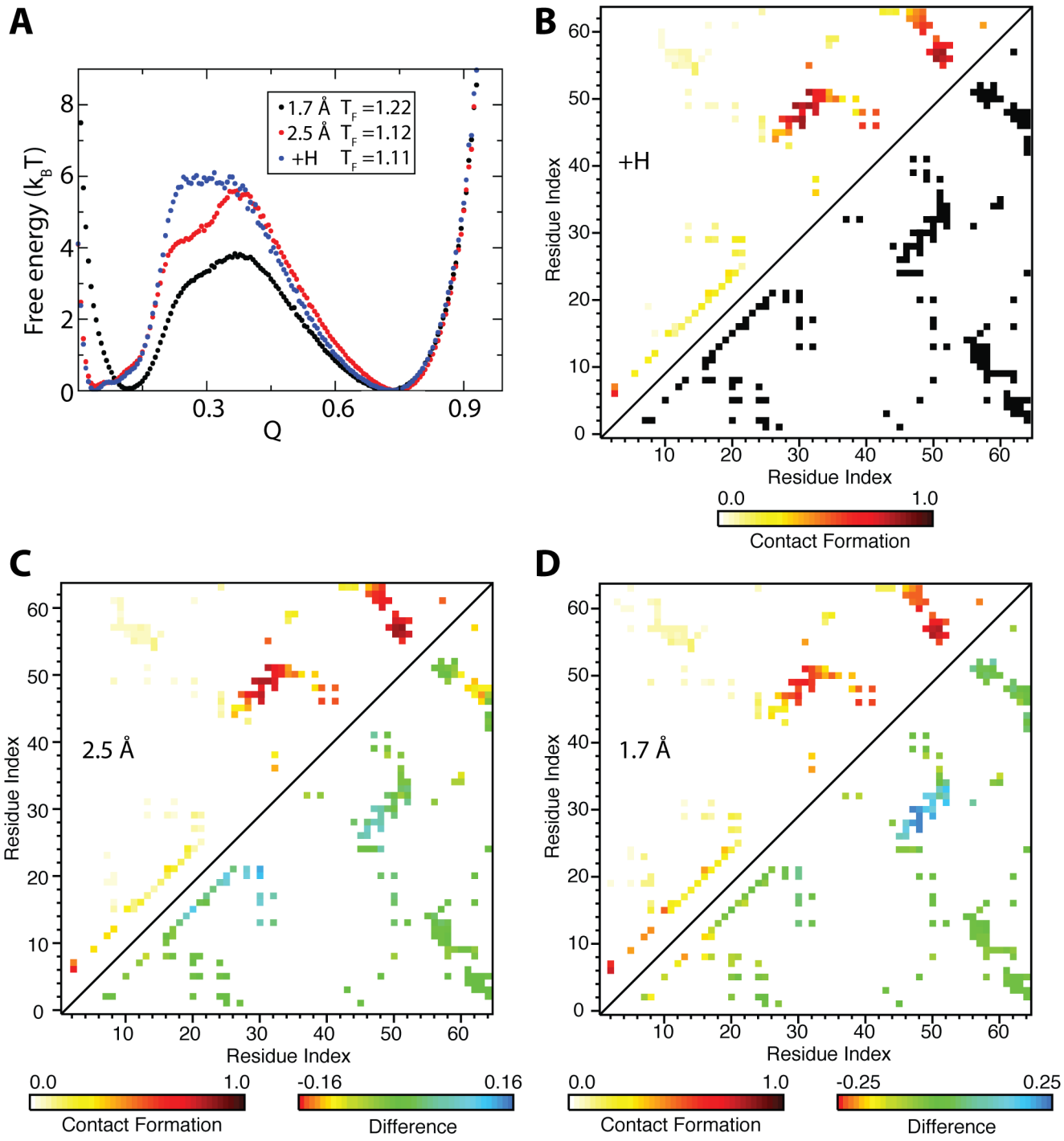

Fig 2. Composition of the folding TSE is robust to variations in the structural resolution of CI2. A) Free energy profiles as a function of the number of native atom-atom contacts *Q*, for three atomic geometries: uniform heavy atom diameter of 1.7 Å ($M^{1.7}$, black), uniform heavy atom diameter of 2.5 Å ($M^{2.5}$, red), and heterogeneous heavy atom sizes with hydrogen excluded volume, +H (M^{+H}, blue). The similar barrier height between $M^{2.5}$ and M^{+H} suggests that the excluded volume in Amber99sb is roughly equivalent to an average of 2.5 Å diameter for heavy atom beads. Increasing the excluded volume raises the folding barrier and lowers T_F [41]. Note that the profile in Fig 1 is different because it was generated using the SBM_AA default of 2.1 Å diameter and a cutoff of 1.5 times the native distance to define a formed native contact, whereas a cutoff of 1.2 is used here. B) CI2 native contact map (lower triangle) and average contact formation at the unfolding side of the free-energy barrier at $Q = 0.30$ for M^{+H} (upper triangle). C) Comparison at $Q = 0.30$ of M^{+H} and $M^{2.5}$. Average contact formation of $M^{2.5}$ (upper triangle) and difference for each contact with positive values indicating higher formation in M^{+H} (lower triangle). D) Comparison at $Q = 0.30$ of M^{+H} and $M^{1.7}$. Average contact formation of $M^{1.7}$ (upper triangle) and difference for each contact with positive values indicating higher formation in M^{+H} (lower triangle).

doi:10.1371/journal.pcbi.1004794.g002

highlighted in the lower triangles of Fig 2 (panels C and D). M^{+H} versus $M^{2.5}$ shows early formation of some secondary structure and delayed formation around ARG48 and ARG62. M^{+H} versus $M^{1.7}$ mainly shows early formation of the parallel β-strand. Overall, while this analysis indicates that two-state folding and the character of the TSE are insensitive to the details of the atomic geometry in CI2, there are subtle effects on secondary structure formation and the shape of the free-energy barrier.

Applications of SMOG 2 to large systems

To demonstrate of capacity of SMOG 2 to study systems of increased size, we used it to prepare simulations of the HIV-1 capsid shell. The HIV-1 capsid shell is composed of 1356 p24 proteins, which form hexameric and pentameric subunits. As noted in the original manuscript [52], the inherent plasticity of the p24 motif enables the formation of this heterogeneous assembly. Together, there are 216 hexameric units and 12 pentameric units that coincide with vertices of the assembly, which together form a "fullerene cone" shape. In total, there are 2.4 million non-hydrogen atoms in this system, making it the largest asymmetric structure available in the PDB. Previously, using explicit-solvent simulation of the full complex, it was found that the structural model maintained its structural integrity on the timescale of 100 ns [52]. This observation lends support to the details of the structural model, thereby implicating the formation of specific stabilizing interprotein interactions.

To elucidate the global motions of the mature HIV-1 capsid, we prepared a structure-based model with SMOG 2. Due to large number of chains and atoms, the web-based smog-server is not capable of processing this system. Since this system lacks global symmetry, it is important to simulate the full assembly in order to probe the dynamics. This is in contrast to more symmetric viral systems, where it may be possible to reduce the computational requirement by utilizing knowledge of the symmetry. From our simulation of the full complex, we performed principle component analysis (PCA) to identify the global modes of motion. Specifically, we calculated the center of mass of each domain of p24 (in total 2712 pseudoparticles) as a function of time and then evaluated the PCAs of the motions of the centers of mass. We find that the first two PCAs provide dominant contributions to the overall fluctuations of the complex, where the five largest eigenvalues were 7.7, 3.5, 2.9, 2.4, 1.7 nm^2. Visualization of the first PCA (Fig 3) shows that the capsid exhibits an overall breathing-like motion. That is, there is correlated expansion and contraction of opposing sides of the capsid. With regards to the second mode, there was not a visible pattern in the direction of motion of the atoms. Nonetheless, when comparing the relative mobility of each domain we find that the largest fluctuations associated with this mode are centered around a specific hexamer (chain 1218 in the PDB file). Since subsequent conformational changes and disassembly are involved in HIV infection, this elevated degree of mobility suggests that this region may facilitate functional processes (e.g. recognition, or rupture propagation).

Computational performance

For modestly sized systems (<20,000 atoms) the SMOG 2 program is lightweight and runs in under a minute on a desktop computer. The numbers quoted here use the template SBM_AA and are performed on a single core of a 2.30 GHz Intel Xeon E5-2630 CPU. For example, creating a topology file for adenylate kinase (1 chain, containing 1656 heavy atoms) takes 7 seconds and 124 MB of memory. While the largest systems considered in this manuscript take significantly more resources, topologies can easily be generated on modern desktop computers. SMOG 2 for the 70S ribosome (150 thousand atoms) takes 12 minutes and 3.1 GB of memory, and the HIV-I capsid (2.4 million atoms in 1356 chains, [52]) takes 89 minutes and 13.9 GB of memory.

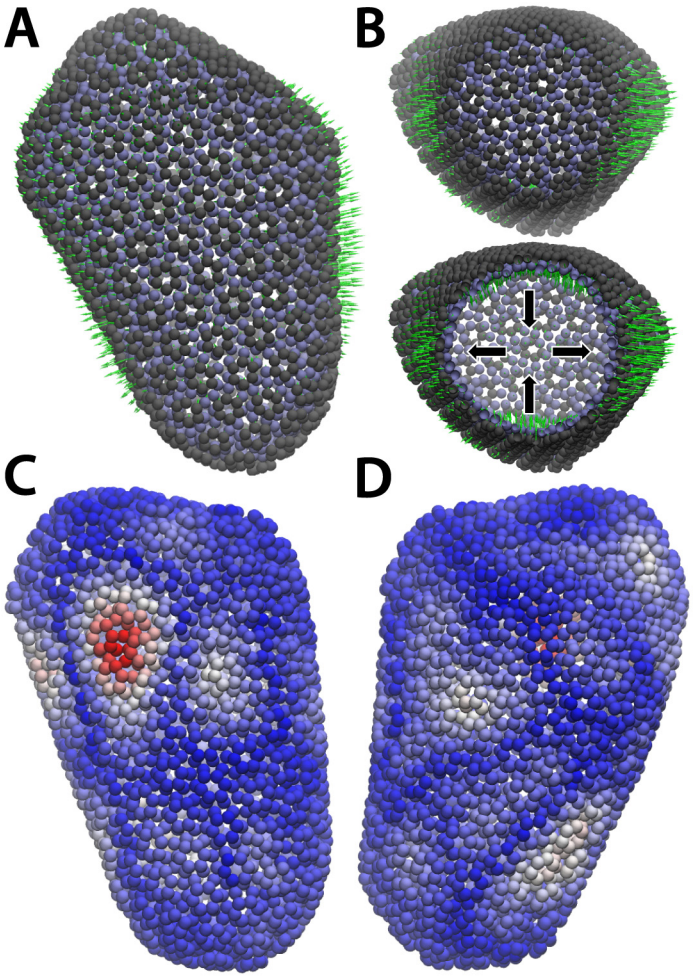

Fig 3. Correlated fluctuations are observed in all-atom simulations of the HIV-1 capsid. A) Side view of HIV-1 capsid with the center of mass of each domain shown as a grey (N-terminal) and ice blue (C-terminal) sphere. The first principal component is shown with green arrows (length of the arrows is not to scale). B) Same as panel (A), rotated 90°. The bottom panel shows the same complex with part of the system hidden. This reveals that, while a large number of domains move outwards (A), others move inward, resulting in a concerted breathing-like motion. C) Capsid shown with centers of mass colored by the scale of the motion in the second mode (blue: small, red: large). The largest fluctuations are centered around hexamer 1218. D) Rotated view of (C).

doi:10.1371/journal.pcbi.1004794.g003

Regarding the performance of MD simulations, SBMs exhibit strong scaling with parallelization on modern computing architectures. With GROMACS (v4 or v5), smaller simulations (<2000 atoms) can typically scale up to the number of processors available on a single motherboard, and larger simulations can significantly benefit from the combination of multiple compute nodes. The ribosome has previously been studied using SBMs [25], and it scales to ∼1000 cores on modern supercomputers (Fig 4). SBMs also exhibit weak scaling, which can be seen with the 2.5M atom EF-G lattice scaling to ∼2000 cores.

Future Directions

There are many exciting applications for exploring the dynamics of biomolecules and molecular modeling that can be incorporated into the SMOG 2 infrastructure. Investigators are currently studying the entropic effects of post-translational modifications such as glycosylation

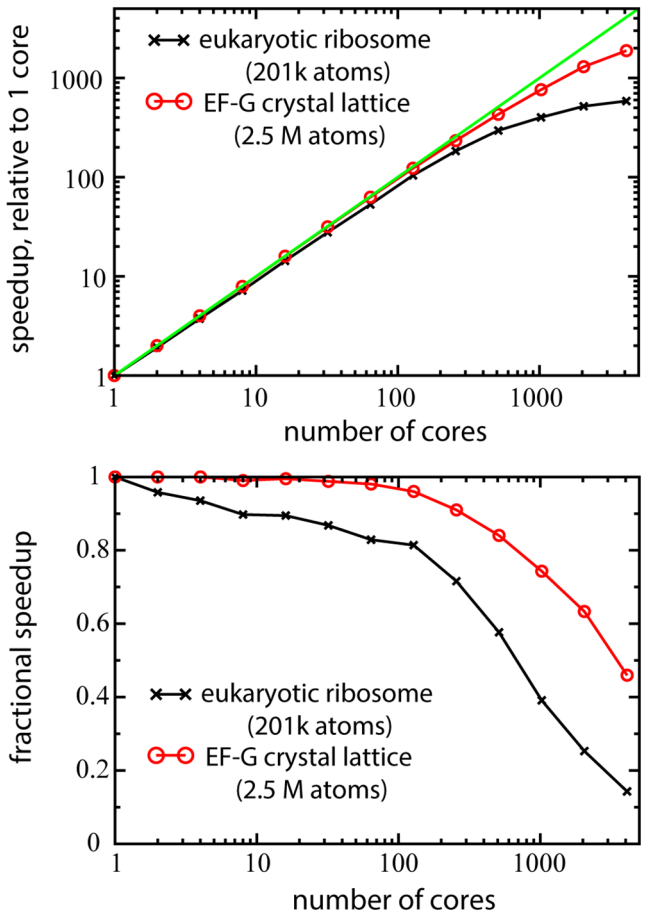

Fig 4. Large structure-based simulations are highly scalable. The two test systems are the eukaryotic ribosome containing 201 thousand atoms and a modeled 8x8x8 crystal lattice of EF-G containing 2.5 million atoms. Simulations used the SBM_AA template with GROMACS v4.6.3 and ran on the Stampede supercomputer located at the TACC.

doi:10.1371/journal.pcbi.1004794.g004

[53] and the energetic effects of electrostatic interactions between nucleic acids and proteins [44, 54]. Another interesting development has been the integration of residue-level co-evolutionary information into structure-based potentials [55, 56]. Co-evolutionary information has a similar theoretical basis to SBMs in the "principle of minimal frustration" [1, 57], and they can help extend SBMs beyond the single-minimum paradigm [58]. With these new directions in mind, it is our intention that SMOG 2 will support the development of diverse applications of SBMs, by establishing a common framework that facilitates portability and collaboration.

Author Contributions

Analyzed the data: JKN HL PCW RLH. Wrote the paper: JKN PCW HL RLH JNO. Conceived and designed the software: JKN JNO PCW. Wrote the software: MR ML JKN PCW.

References

1. Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. Proc Nat Acad Sci USA 84: 7524. doi: 10.1073/pnas.84.21.7524 PMID: 3478708

2. Leopold PE, Montal M, Onuchic JN (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc Nat Acad Sci USA 89: 8721–8725. doi: 10.1073/pnas.89.18.8721 PMID: 1528885
3. Onuchic JN, Wolynes PG (2004) Theory of protein folding. Curr Opin Struct Biol 14: 70–75. doi: 10.1016/j.sbi.2004.01.009 PMID: 15102452
4. Noel JK, Onuchic JN (2012) The many faces of structure-based potentials: From protein folding landscapes to structural characterization of complex biomolecules. Computational Modeling of Biological Systems, Springer US: 31–54. doi: 10.1007/978-1-4614-2146-7_2
5. Whitford PC, Sanbonmatsu KY, Onuchic JN (2012) Biomolecular dynamics: order-disorder transitions and energy landscapes. Rep Prog Phys 75: 076601. doi: 10.1088/0034-4885/75/7/076601 PMID: 22790780
6. Shea J, Onuchic J, CB III (1999) Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment b of protein a. Proc Nat Acad Sci USA 96: 12512–12517. doi: 10.1073/pnas.96.22.12512 PMID: 10535953
7. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. J Mol Biol 298: 937–953. doi: 10.1006/jmbi.2000.3693 PMID: 10801360
8. Kaya H, Chan HS (2000) polymer principles of protein calorimetric two-state cooperativity. Proteins 40: 637–661. doi: 10.1002/1097-0134(20000901)40:4%3C637::AID-PROT80%3E3.3.CO;2-W PMID: 10899787
9. Kaya H, Chan HS (2003) Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? J Mol Biol 326: 911–931. doi: 10.1016/S0022-2836(02)01434-1 PMID: 12581650
10. Clementi C (2008). Coarse-grained models of protein folding: toy models or predictive tools? Curr Opin Struct Biol 18: 10–15. doi: 10.1016/j.sbi.2007.10.005 PMID: 18160277
11. Cho SS, Levy Y, Wolynes PG (2009) Quantitative criteria for native energetic heterogeneity influences in the prediction of protein folding kinetics. Proc Nat Acad Sci USA 106: 434–439. doi: 10.1073/pnas.0810218105 PMID: 19075236
12. Oliveira RJ, Whitford PC, Chahine J, Wang J, Onuchic JN, et al. (2010) The origin of nonmonotonic complex behavior and the effects of nonnative interactions on the diffusive properties of protein folding. Biophys J 99: 600–608. doi: 10.1016/j.bpj.2010.04.041 PMID: 20643080
13. Noel JK, Sulkowska J, Onuchic JN (2010) Slipknotting upon native-like loop formation in a trefoil knot protein. Proc Nat Acad Sci USA 107: 15403–15408. doi: 10.1073/pnas.1009522107 PMID: 20702769
14. Best RB, Hummer G (2010) Coordinate-dependent diffusion in protein folding. Proc Nat Acad Sci USA 107: 1088–1093. doi: 10.1073/pnas.0910390107 PMID: 20080558
15. Chan HS, Zhang Z, Wallin S, Liu Z (2011) Cooperativity, local-nonlocal coupling, and nonnative interactions: principles of protein folding from coarse-grained models. Annu Rev Phys Chem 62: 301–326. doi: 10.1146/annurev-physchem-032210-103405 PMID: 21453060
16. Weinkam P, Zong C, Wolynes PG (2005) A funneled energy landscape for cytochrome c directly predicts the sequential folding route inferred from hydrogen exchange experiments. Proc Nat Acad Sci USA 102: 12401–12406. doi: 10.1073/pnas.0505274102 PMID: 16116080
17. Best RB, Paci E, Hummer G, Dudko OK (2008) Pulling direction as a reaction coordinate for the mechanical unfolding of single molecules. J Phys Chem B 112: 5968–5976. doi: 10.1021/jp075955j PMID: 18251532
18. Pincus DL, Cho SS, Hyeon C, Thirumalai D (2008). Minimal models for proteins and RNA: From folding to function. Prog Mol Biol Transl Sci 84: 203–250. doi: 10.1016/S0079-6603(08)00406-6 PMID: 19121703
19. Levy Y, Wolynes PG, Onuchic JN (2004) Protein topology determines binding mechanism. Proc Nat Acad Sci USA 101: 511–516. doi: 10.1073/pnas.2534828100 PMID: 14694192
20. Ramírez-Sarmiento CA, Noel JK, Valenzuela SL, Artsimovitch I (2015) Interdomain contacts control native state switching of rfah on a dual-funneled landscape. PLOS Comput Biol 11: e1004379. doi: 10.1371/journal.pcbi.1004379 PMID: 26230837
21. Okazaki K, Koga N, Takada S, Onuchic JN, Wolynes PG (2006) Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. Proc Nat Acad Sci USA 103: 11844–11849. doi: 10.1073/pnas.0604375103 PMID: 16877541
22. Whitford PC, Miyashita O, Levy Y, Onuchic JN (2007) Conformational transitions of adenylate kinase: switching by cracking. J Mol Biol 366: 1661–1671. doi: 10.1016/j.jmb.2006.11.085 PMID: 17217965

23. Hyeon C, Onuchic JN (2007) Internal strain regulates the nucleotide binding site of the kinesin leading head. Proc Nat Acad Sci USA 104: 2175–2180. doi: 10.1073/pnas.0610939104 PMID: 17287347

24. Jana B, Hyeon C, Onuchic JN (2012) The origin of minus-end directionality and mechanochemistry of ncd motors. PLoS Comput Biol 8: e1002783. doi: 10.1371/journal.pcbi.1002783 PMID: 23166486

25. Noel JK, Chahine J, Leite VBP, Whitford PC (2014) Capturing transition paths and transition states for conformational rearrangements in the ribosome. Biophys J 107: 2881–2890. doi: 10.1016/j.bpj.2014.10.022 PMID: 25517153

26. Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, et al. (2009) An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. Proteins 75: 430–441. doi: 10.1002/prot.22253 PMID: 18837035

27. Clementi C, Plotkin SS (2004) The effects of nonnative interactions on protein folding rates: theory and simulation. Protein Sci 13: 1750–1766. doi: 10.1110/ps.03580104 PMID: 15215519

28. Zhang Z, Chan HS (2009) Native topology of the designed protein Top7 is not conducive to cooperative folding. Biophys J 96: L25–7. doi: 10.1016/j.bpj.2008.11.004 PMID: 19186118

29. Li W, Wolynes PG, Takada S (2011) Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins. Proc Nat Acad Sci USA 108: 3504–3509. doi: 10.1073/pnas.1018983108 PMID: 21307307

30. Whitford PC, Ahmed A, Yu Y, Hennelly SP, Tama F, et al. (2011) Excited states of ribosome translocation revealed through integrative molecular modeling. Proc Nat Acad Sci USA 108: 18943–18948. doi: 10.1073/pnas.1108363108 PMID: 22080606

31. Cheng RR, Morcos F, Levine H, Onuchic JN (2014) Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Proc Nat Acad Sci USA 111: E563–71. doi: 10.1073/pnas.1323734111 PMID: 24449878

32. dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN (2015) Dimeric interactions and complex formation using direct coevolutionary couplings. Sci Rep 5: 13652. doi: 10.1038/srep13652 PMID: 26338201

33. Taketomi H, Ueda Y, Gō N (1975) Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. Int J Pept Prot Res 7: 445–459. doi: 10.1111/j.1399-3011.1975.tb02465.x

34. Noel JK, Whitford PC, Sanbonmatsu KY, Onuchic JN (2010) SMOG@ctbp: simplified deployment of structure-based models in GROMACS. Nucleic Acids Res 38: W657–61. doi: 10.1093/nar/gkq498 PMID: 20525782

35. Karanicolas J, Brooks CL (2003) Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. J Mol Biol 334: 309–325. doi: 10.1016/j.jmb.2003.09.047 PMID: 14607121

36. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, et al. (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29: 845–854. doi: 10.1093/bioinformatics/btt055 PMID: 23407358

37. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. (2005) Scalable molecular dynamics with namd. J Comput Chem 26: 1781–1802. doi: 10.1002/jcc.20289 PMID: 16222654

38. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple AMBER force fields and development of improved protein backbone parameters. Proteins 65: 712–725. doi: 10.1002/prot.21123 PMID: 16981200

39. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, et al. (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30: 1545–1614. doi: 10.1002/jcc.21287 PMID: 19444816

40. Schmid N, Eichenberger AP, Choutko A, Riniker S, Winger M, et al. (2011) Definition and testing of the gromos force-field versions 54a7 and 54b7. Eur Biophys J 40: 843–856.

41. Noel JK, Whitford PC, Onuchic JN (2012) The Shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. J Phys Chem B 116: 8692–8702. doi: 10.1021/jp300852d PMID: 22536820

42. Lammert H, Wolynes PG, Onuchic JN (2012) The role of atomic level steric effects and attractive forces in protein folding. Proteins 80: 362–373. doi: 10.1002/prot.23187 PMID: 22081451

43. Noel JK, Schug A, Verma A, Wenzel W, Garcia AE, et al. (2012) Mirror images as naturally competing conformations in protein folding. J Phys Chem B 116: 6880–6888. doi: 10.1021/jp212623d PMID: 22497217

44. Azia A, Levy Y (2009) Nonnative electrostatic interactions can modulate protein folding: Molecular dynamics with a grain of salt. J Mol Biol 393: 527–542. doi: 10.1016/j.jmb.2009.08.010 PMID: 19683007

45. Lammert H, Schug A, Onuchic JN (2009) Robustness and generalization of structure-based models for protein folding and function. Proteins 77: 881–891. doi: 10.1002/prot.22511 PMID: 19626713

46. Tirion M (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 77: 1905–1908. doi: 10.1103/PhysRevLett.77.1905 PMID: 10063201

47. Lutz B, Sinner C, Heuermann G, Verma A, Schug A (2013) eSBMtools 1.0: enhanced native structure-based modeling tools. Bioinformatics 29: 2795–2796. doi: 10.1093/bioinformatics/btt478 PMID: 24021379

48. Harpaz Y, Elmasry N, Fersht AR, Henrick K (1994) Direct observation of better hydration at the n terminus of an α-helix with glycine rather than alanine as the n-cap residue. Proc Nat Acad Sci USA 91: 311–315. doi: 10.1073/pnas.91.1.311 PMID: 8278384

49. Cho S, Levy Y, Wolynes PG (2006) P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. Proc Nat Acad Sci USA 103: 586–591. doi: 10.1073/pnas.0509768103 PMID: 16407126

50. Best RB, Hummer G, Eaton WA (2013) Native contacts determine protein folding mechanisms in atomistic simulations. Proc Nat Acad Sci USA 110: 17874–17879. doi: 10.1073/pnas.1311599110 PMID: 24128758

51. Jackson SE, Fersht AR (1991) Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. Biochemistry 30: 10428–10435. PMID: 1931967

52. Zhao G, Perilla JR, Yufenyuy EL, Meng X, Chen B, et al. (2013) Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature 497: 643–646. doi: 10.1038/nature12162 PMID: 23719463

53. Shental-Bechor D, Levy Y (2008) Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. Proc Nat Acad Sci USA 105: 8256–8261. doi: 10.1073/pnas.0801340105 PMID: 18550810

54. Hayes RL, Noel JK, Mandic A, Whitford PC, Sanbonmatsu KY, et al. (2015) Generalized manning condensation model captures the rna ion atmosphere. Phys Rev Lett 114: 258105. doi: 10.1103/PhysRevLett.114.258105 PMID: 26197147

55. Jana B, Morcos F, Onuchic JN (2014) From structure to function: the convergence of structure based models and co-evolutionary information. Phys Chem Chem Phys 16: 6496–6507. doi: 10.1039/c3cp55275f PMID: 24603809

56. Cheng RR, Raghunathan M, Noel JK, Onuchic JN (2015) Constructing sequence-dependent protein models using coevolutionary information. Protein Sci 25: 111–122. doi: 10.1002/pro.2758 PMID: 26223372.

57. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG (2014) Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc Nat Acad Sci USA 111: 12408–12413. doi: 10.1073/pnas.1413575111 PMID: 25114242

58. Noel JK, Morcos F, Onuchic JN (2016) Sequence co-evolutionary information is a natural partner to minimally-frustrated models of biomolecular dynamics. F1000Res 5: 1–7. doi: 10.12688/f1000research.7186.1