Coarse-Grained Molecular Modeling of the Solution Structure Ensemble of Dengue Virus Nonstructural Protein 5 with Small-Angle X-ray Scattering Intensity

Zhu, Guanhua; Saw, Wuan Geok; Nalaparaju, Anjaiah; Grüber, Gerhard; Lu, Lanyuan

2017

Zhu, G., Saw, W. G., Nalaparaju, A., Grüber, G., & Lu, L. (2017). Coarse-Grained Molecular Modeling of the Solution Structure Ensemble of Dengue Virus Nonstructural Protein 5 with Small-Angle X-ray Scattering Intensity. *The Journal of Physical Chemistry B*, 121(10), 2252-2264.

<https://hdl.handle.net/10356/85140>

<https://doi.org/10.1021/acs.jpcb.7b00051>

© 2017 American Chemical Society. This is the author created version of a work that has been peer reviewed and accepted for publication by *The Journal of Physical Chemistry B*, American Chemical Society. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at:
[<http://dx.doi.org/10.1021/acs.jpcb.7b00051>].

**Coarse-Grained Molecular Modeling of Solution Structure Ensemble of Dengue Virus
Non-Structural Protein 5 with Small-Angle X-Ray Scattering Intensity**

Guanhua Zhu, Wuan Geok Saw, Anjaiah Nalaparaju, Gerhard Grüber, Lanyuan Lu*

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, 637551,
Singapore

Abstract

An ensemble modeling scheme incorporating coarse-grained simulations with experimental small-angle X-ray scattering (SAXS) data is applied to the dengue virus 2 (DENV2) non-structural protein 5 (NS5). NS5 serves a key role in viral replication through its two domains that are connected by a ten-residue polypeptide segment. A set of representative structures are generated from a simulated structure pool using SAXS data fitting by the non-negativity least squares (NNLS) or the standard Ensemble Optimization Method (EOM) based on a genetic algorithm (GA). It is found that a proper low-energy threshold of the structure pool is necessary to produce a conformational ensemble of two representative structures by both NNLS and GA that agrees well with the experimental SAXS profile. The stability of the constructed ensemble is validated also by molecular dynamics simulations with an all-atom force field. The constructed ensemble successfully revealed the domain-domain orientation and the domain contacting interface of DENV2 NS5. Using experimental data fitting and additional investigations with synthesized data, it is found that the energy restraint on the conformational pool is necessary to avoid the over-interpretation of experimental data by spurious conformational representations.

Introduction

Dengue virus (DENV) has four antigenically different serotypes, DENV-1 to DENV-4, and causes life-threatening viral diseases. DENV RNA replication occurs in an ER-derived replication complex consisting of host proteins and non-structural (NS) viral proteins.¹⁻³ The non-structural protein 5 (NS5) is a key enzyme in viral replication, associated in diverse activities required for genome replication, genome capping and host immune response modulation,⁴⁻⁹ thus has been considered as a promising target for the development of antiviral inhibitors.¹⁰ NS5 is a multi-domain protein with an N-terminal methyltransferase (MTase) domain (residue 1-263) and a C-terminal RNA-dependent RNA polymerase (RdRp) domain (residue 274-900) connected by a ten-residue linker. Several works have been published on structural determination of full-length NS5. The crystal structure of NS5 from Japanese encephalitis virus (JEV) (PDB entry 4K6M) was determined and its MTase-RdRp interface is majorly the formation of hydrophobic interactions.¹¹ Interestingly, the recently determined crystal structure of DENV-3 NS5 (PDB entry 4V0Q) adopted a similar overall shape as JEV NS5, but a novel MTase-RdRp binding interface that is mostly driven by polar residues.¹² The two distinct MTase-RdRp binding patterns found in crystallographic structures suggest the possibility of the existence of multiple conformations of NS5. The genome replication and capping require NS5 to interact with other viral proteins, viral RNA genome and host proteins involved in the processes dynamically, which also indicates the conformational flexibility of NS5 in solution.

The solution structure of multi-protein complexes can be characterized by small-angle X-ray scattering (SAXS) that yields the shapes and dimensions of the macromolecules. Compared with solution nuclear magnetic resonance, SAXS takes the advantage of handling large molecules like

NS5 of 900 residues. However, the known crystal structures of NS5 fail to explain the solution SAXS observables. One previous SAXS experiment indicated that DENV3 NS5 adopts multiple conformations in solution, ranging from compact to loose forms.¹³ In the loose conformations, no interaction was present between the MTase and RdRp domains. A more recent SAXS study¹⁴ offered an ensemble representation of DENV3 NS5 but without loose conformations as proposed in the previous work.¹³ It was also revealed that NS5 of all DENV serotypes form elongated shapes in solution compared to the crystal structure, and DENV-4 NS5 has a relatively more compact conformation than other serotypes.¹⁴ The structure ensemble determination of flexible biomolecules is essential for understanding their functions, but it remains a challenge for conventional experimental methods solely. Thus, computational methods serve as an important complementary tool for the structural characterization of flexible biomolecules.

The Ensemble Optimization Method (EOM)¹⁵ is the first approach for ensemble fitting of SAXS measurements. EOM adopts a genetic algorithm (GA) for ensemble selection from a pre-generated pool of random structures and its latest version EOM 2.0¹⁶ improves the ensemble selection procedure to refine the ensemble size. Similar to EOM, many approaches have also adopted a sampling-and-screening optimization strategy for ensemble construction, including ASTEROIDS^{17, 18} and Minimal Ensemble Search (MES)¹⁹ with a genetic algorithm, Sparse Ensemble Selection (SES)²⁰ with a Multi-Orthogonal Matching Pursuit algorithm, Basis-Set Supported SAXS (BSS-SAXS)²¹ approach with a Bayesian MC algorithm, Sample and Select (SAS)²² with simulated annealing, and other methods.²³⁻²⁵ Many of the methods above^{19, 20, 22, 24} tend to recover the experimental observables with a small-sized ensemble to prevent data over-interpretation. Besides, some other approaches like ENSEMBLE²⁶ and EROS²⁷ adopted an

alternative regularization method, maximum entropy²⁸ that drives the solution to avoid spurious biases and has been applied to ensemble construction with SAXS data.^{27, 29}

Molecular dynamics (MD) simulations are widely employed combined with experimental restraints to model flexible biomolecules like protein-protein complexes, multi-domain proteins and intrinsically disordered proteins. Previous studies³⁰ have incorporated SAXS data in MD simulations to drive the biased simulation ensemble more consistent with experimental SAXS profile. This data-driven simulation suffices for the cases when only one predominant state exists, but not for the structure ensemble of relatively rugged energy landscape with multiple competing minima. Therefore, many computational investigations adopted MD simulations with sampling methods for flexible biomolecules and subsequently constructed a structure ensemble from the generated conformation pool. MD simulation is superior to artificial approaches in terms of structure sampling, because force fields prevent the unphysical conformations from the candidate pool. However, conventional all-atom simulations encounter the bottleneck of a short timescale, especially for large biomolecular systems. Coarse-grained (CG) models, in contrast, reduce the degrees of freedom of the system, smoothen free energy landscape and thus accelerate molecular dynamics simulations.^{31, 32} CG models have been successfully applied to the studies of multi-domain proteins and multi-protein complexes,^{23, 33, 34} including the cases incorporating experimental SAXS data.^{21, 27, 35}

In this study, we adopted a CG model to study the multi-domain protein DENV2 NS5 on the basis of previously published experimental SAXS data.¹⁴ CG modeling is implemented here because i) it is difficult to model such large proteins like NS5 in all-atom simulations considering the expensive computational burden, and ii) SAXS offers structural information at medium or low resolution instead of atomistic level. Inspired by the CG approaches for modeling of protein

complexes,^{33, 34} we employed a structure-based potential³⁶⁻³⁸ to model the stable intra-domain structure and a CG force field³⁴ to characterize the inter-domain interactions. The experiment of swapping the ten-residue linker between DENV-3 NS5 and DENV-4 NS5 illustrated the linker may serve a crucial role in domain-domain associations of NS5 in solution.¹⁴ Therefore, the linker and flexible loops are treated as flexible parts that interact with both domains and themselves via the inter-domain CG force field. MD simulations are performed on DENV2 NS5, followed by the ensemble construction with respect to experimental SAXS profiles and the structure characterization in an energy perspective. Here, energy is proposed as a physics-based regularization restraint to prevent ensemble recovery from data overfitting. We show that the exclusion of energy restraint would cause noisy and ambiguous interpretation of experimental data in the ensemble construction, and the effect of the structure pool was further investigated by studying a synthesized structure ensemble.

Methods

CG model

DENV2 NS5 has a MTase domain and a RdRp domain that are connected by a ten-residue linker (residues 264-273). The protein is represented by CG beads that are positioned at C_α of each amino acid. The overall potential energy function consists of intra-domain interaction V_{intra} and inter-domain interaction V_{inter} . The intra-domain interaction of individual domains V_{intra} is formulated as follows:

$$\begin{aligned}
V_{\text{int ra}} = & \sum_{\text{bond}} k_b (b - b_0)^2 \\
& + \sum_{\text{angle}} k_\theta (\theta - \theta_0)^2 \\
& + \sum_{\text{dihedral}} k_\phi^{n=1,3} \left[1 + \cos(n \times (\phi - \phi_0)) \right] \\
& + \sum_{\substack{i>j-3 \\ \text{native}}} \epsilon_1 \left[5 \left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{0ij}}{r_{ij}} \right)^{10} \right] \\
& + \sum_{\substack{i>j-3 \\ \text{non-native}}} \epsilon_2 \left(\frac{C}{r_{ij}} \right)^{12}
\end{aligned} \tag{1}$$

The MTase and RdRp domains are simulated in a Gō-type model³⁶⁻³⁸ individually with their native structures as global energy minima, where force constants for bond, angle, and dihedral terms are $k_b=100$ kcal/(mol·Å²), $k_\theta=20$ kcal/(mol·rad²), $k_\phi^1=1$ kcal/mol and $k_\phi^3 = 0.5$ kcal/mol, respectively. The native-pair term assigns a Lennard-Jones 10-12 potential over a native pair between residue i and j , with an interaction strength $\epsilon_1=4$ kcal/mol and their equilibrium position r_{0ij} derived from the native structures. The "in contact" native pairs are determined by the method of shadow map.³⁹ The non-native pairs are interacted as a non-specific repulsion with $\epsilon_2=1$ kcal/mol and $C=4$ Å. The native structure of MTase domain was adopted from crystallographic structure of DENV2 NS5 (PDB entry 1L9K⁶) and the RdRp domain structure was built homologously⁴⁰ from the crystal structure of RdRp domain of DENV3 NS5 (PDB entry 4V0R¹²). The flexible linker and the missing loops in crystal structures including loop-1 (residue 407-418), loop-2 (residues 456-469) and two terminal loops, were treated as a chain of harmonic springs with a force constant of 100 kcal/mol.

The MTase-RdRp interactions are modeled via a nonbonded CG force field.^{34, 41} Besides, the linker and flexible loops are also subjected to the same nonbonded CG force field to characterize

the interaction between both domains and themselves. The inter-domain interaction V_{inter} is defined as follows:

$$V_{inter} = V_{ele} + V_H \quad (2)$$

where V_{ele} and V_H are electrostatic and hydrophobic energies respectively. A Debye–Hückel-type potential, $V_{ele} = \sum_{ij} q_i q_j \exp(-r_{ij}/\xi) / 4\pi r_{ij} D$ is used for charge-charge interactions between residue i and j separated by distance r_{ij} , where residue charges $q = +e$ for Lys and Arg, and $-e$ for Asp and Glu at pH 7.5. The Debye screening length ξ is set to 5.6 Å to mimic the screening effect of salt at 300 mM NaCl in the experimental condition. D is the dielectric constant for inter-domain electrostatics, parameterized as 10 in this CG model.³⁴ The hydrophobic term V_H is derived originally from Miyazawa-Jernigan potential⁴² and defined here either as an attraction or as a repulsion formulation, depending on the types of residue pairs. Specifically, $V_H(i, j) = |\epsilon_{ij}| \left[5(\sigma_{ij}/r_{ij})^{12} - 6(\sigma_{ij}/r_{ij})^{10} \right]$ is used for attractive pairs ($\epsilon_{ij} < 0$) and $V_H(i, j) = \epsilon_{ij} \left[5(\sigma_{ij}/r_{ij})^{12} \left(1 - \exp\left(-((r_{ij} - \sigma_{ij})/d)^2\right) \right) \right]$ (where $d = 3.8$ Å) is used for repulsive pairs ($\epsilon_{ij} \geq 0$). The residue-pair parameters ϵ_{ij} and σ_{ij} are taken from literature.⁴¹ It should be noted that for MTase and RdRp domains, the inter-domain interaction is only subjected to domain surface residues whose solvent accessible surface area (SASA) is more than 10 Å². Residue SASA was computed using a probe of the radius 1.4 Å rolling on each atomistic domain structure. The CG energy calculated using the method described above was used for GG simulations, and the inter-domain part of CG energy was adopted for the energy-based pool structure selection described later in this article.

Simulation details

The program RANCH¹⁵ was employed to generate 10000 random structures with the restraint of linker, followed by structure clustering with a RMSD cutoff of 9 Å. The resultant 102 structures were subjected to energy minimization and then taken as initial structures in independent MD simulations with GROMACS 4.6.⁴³ Each MD simulation was carried out for 100 ns at 300 K and a time step of 0.01 ps. The total simulation time was about 10.2 μ s, generating a pool of 204000 structures.

Structure characterization and ensemble construction

To handle the large amount of simulated structures from CG MD simulations, a two-step clustering procedure was performed to group similar structures. Firstly, all the structures from simulation trajectories were grouped according to domain-domain orientations that are defined as φ and θ angles in the spherical coordinates (Figure 1), and each protein structure was put into one of the θ - φ groups corresponding to the two-dimensional cells with $\Delta\theta = 9^\circ$ and $\Delta\varphi = 18^\circ$. In the second step, structure clustering⁴⁴ was carried out within each orientation (θ - φ) group with an RMSD cutoff of 4 Å. The structure that had the most neighbors within the RMSD cutoff was identified as one cluster and removed from the structure pool with all its neighbor structures. We iteratively repeated this procedure for the remaining structures in the pool until all structures are assigned to one of the clusters. In each cluster, the structure of the lowest inter-domain energy V_{inter} defined in Equation (2) was extracted as the representative conformation and subsequently taken for ensemble construction.

Figure 1 Domain-domain orientation described by angles φ and θ . RdRp domain (green) is connected with the MTase domain (purple) via a ten-residue linker (orange).

After structure clustering, the resultant 1015 representative conformations were subjected to the theoretical calculation of SAXS profile using the program CRY SOL⁴⁵ after reconstructing their atomistic structure using the program PULCHRA.⁴⁶ Although CG energy values are used in this study, the conversion from CG to atomistic structures is necessary as CRY SOL is based on atomistic structures. Ensemble construction from the representative structure pool was performed against the experimental SAXS profile by the method of non-negativity least squares (NNLS)⁴⁷ with the constraint that the total population is 100%. The representative structures were characterized in an energy perspective to understand the effect of energy-based regularization in ensemble construction. The energy for characterizing the pool candidate structures was the inter-domain energy based on the CG force field. The low-energy conformations identified in the CG model were further validated by a conventional atomistic force field. The ensemble selected by EOM with a genetic algorithm was also illustrated for comparison. The fitting quality was evaluated by χ^2 given in the following equation,

$$\chi^2 = \frac{1}{K-1} \sum_{j=1}^K \left[\frac{\mu I(s_j) - I_{\text{exp}}(s_j)}{\sigma(s_j)} \right]^2 \quad (3)$$

where K is the number of data points, $I(s)$ is the theoretical scattering, $I_{\text{exp}}(s)$ is the experimental scattering, $\sigma(s)$ is the standard deviation and μ is a scaling factor¹⁵. The experimental scattering profile was taken from the previous study,¹⁴ where the detailed description of the SAXS experiment can be found. The number of experimental SAXS data points is 813, up to s value of $0.35 \text{ \AA}^{-1}$. Note that in addition to the most commonly used χ^2 in Equation (3) that is employed by EOM, there are other formulations⁴⁸⁻⁵⁰ of χ^2 to evaluate the quality of fit between the calculated and experimental SAXS profiles. A recent study⁵¹ compared these functions that lead to a similar fitting result with respect to the experimental profile.

In this work both NNLS and EOM were implemented for the purpose of fitting to the SAXS intensity, which allows us to investigate the effect of data fitting algorithm in ensemble construction. The NNLS approach guarantees the optimal fitting for the linear least square problem, while the number of structures in the ensemble is not constrained. By contrast, EOM applies a limit on the total number of conformations in the solution. It is shown in the result section that the two methods can produce consistent results by applying a proper energy threshold to the conformation pool.

Results and Discussions

CG simulation of DENV2 NS5

The conformations of DENV2 NS5 were simulated with a hybrid CG model in which the MTase and RdRp domains were modeled using a structure-based potential and the inter-domain interaction was characterized with a CG force field of electrostatics and hydrophobic terms.^{33, 34} In this model, each residue was coarse-grained as one bead centered at its C α position to accelerate the simulation of experimental SAXS data at medium/low resolution. Some reported CG models of multi-domain proteins or well-structured units connected by intrinsically disordered segments either treated the linkers as polymers without interactions with other structure units,⁴¹ or simplified the interactions between disordered segments and other structure units as only repulsion and electrostatics³⁵. Without prior knowledge of the conformations of the flexible structures, we employed a physically reasonable treatment to simulate their interactions with electrostatic and hydrophobic energies.³³

The structure-based potential imposed on MTase and RdRp domains produced a simulated structure ensemble of multiple trajectories, where the structures preserved their domain conformations with the RMSD values below 1 Å with respect to their native structure. The representative conformations were extracted as the candidate structure pool for ensemble construction and structure analysis using the two-step clustering described in the previous section. To characterize these representative structures and cross-link their conformational features with energetic profiles, an energy landscape plot was used to depict the energy-favored conformational space. For the biomolecular system of multi-domain proteins, domain-domain orientation corresponding to domain contacting interface is an important conformational dimension to demonstrate the global structural arrangement. Here, the RdRp domain was fixed and the MTase domain could roll towards all possible faces of RdRp, defined by orientations of φ and θ angles in the spherical coordinates (Figure 1). The representative conformations were

assigned to the grid size $12^\circ \times 6^\circ$ on the orientation map on the basis of their φ and θ angles, followed by extracting the structure of lowest energy *a.k.a.* the structure with the highest probability, within each grid to represent the energy level of that grid. Because of the length restraint of the ten-residues linker, large areas on the domain-domain orientation map could not be accessed physically (Figure 2a). The energy landscape in Figure 2b shows the domain-domain orientations from an energy viewpoint. The low-energy orientation region is identified as a contiguous area with an explicit boundary from other orientations whose energy values are significantly higher. In an ensemble construction by fitting to the experimental SAXS profile, a number of structures representing different areas on the energy landscape are selected, and the weights of the selected structures are adjusted to reproduce the target SAXS profile, on the basis of the computed profiles for the individual structures. To deal with the ill-posed feature of the fitting process and avoid over-fitting, it is rational to exclude some high-energy structures from our structure pool, because the weights of those structures in an ensemble should be negligible according to the Boltzmann distribution. In this work, Figure 2b identifies some high-energy regions of domain-domain orientations on the energy landscape, and the corresponding high-energy structures may be excluded depending on the energy threshold. It is noted that the inaccessible domain-domain orientation space caused by the linker restraint is located at the high-energy area in the energy landscape.

Figure 2 Structure pool of representative conformations and energy landscape. (a) Domain-domain orientations sampled by the 1015 representative conformations (red points) of DENV2 NS5 from molecular modeling. (b) Energy landscape of inter-domain interactions of NS5. The energy of each domain-domain orientation is represented in a $12^\circ \times 6^\circ$ grid and its value is indicated in colors as defined in the color bar (kcal/mol).

Ensemble construction

The theoretical SAXS profile of each representative conformation was calculated by CRY SOL⁴⁵ after mapping their CG structures to atomistic models by PULCHRA.⁴⁶ The possible

error in the CG to all-atom conversion should not be critical in our study because the low-resolution SAXS profile is not sensitive to small changes of side chain conformations. NNLS was employed to select the structure ensemble from the candidate structure pool. To systematically investigate the ensembles constructed by the structures of different energy levels, low-energy structures were taken from the pool of representative conformations, constituting as low-energy structure pools, e.g., the structures ranked within the lowest 1% in terms of energy composing a top 1% low-energy structure pool. The ensemble construction from top 1% low-energy structure pool produced a χ^2 of 0.58, while the use of top 5% low-energy structure pool improved the fitting quality to $\chi^2=0.39$ (Table 1). As shown in Figure 3, a looser requirement of energy from 5% low-energy onward failed to contribute to a significantly higher consistency with the experimental SAXS profile, but incorporated more higher-energy structures in data fitting, consequently undermining the physical reliability of the constructed ensemble. Without the energy control, the entire structure pool achieved a χ^2 of 0.34, expected as a pure numerical fitting with high-energy structures in its reweighted ensemble. We suggest that an optimal ensemble is constructed from the energy-favored structure pool meanwhile of a good agreement with experimental measurements. In our case the L-shape-like pattern of χ^2 evolution in Figure 3 recommended the use of 5% low-energy ensemble as representative ensemble, whose calculated SAXS data reproduces the experimental profile well (Figure 3, inset). The radius of gyration (R_g) of the reweighted 5% low-energy ensemble is 3.49 nm and consistent with the experimental R_g value 3.56 ± 0.05 nm derived from the Guinier approximation. One difficulty of reproducing an equal value of the experimental R_g might be resulted from the terminal loops that probably adopt multiple conformations and could not be fully explained by 2-5 structures of the constructed ensemble.

ensemble	candidates	χ^2	conformations
1%	10	0.58	2
5%	51	0.39	2
10%	102	0.39	4
15%	152	0.37	3
20%	203	0.37	4
25%	254	0.37	5
30%	305	0.37	5
40%	406	0.35	4
50%	508	0.35	4
100%	1015	0.34	8

Table 1. Constructed ensembles reweighted from the structure pools of different energy levels. Columns correspond to low-energy pools, the number of candidate conformations, chi-square and the number of conformations in the constructed ensemble, respectively.

Figure 3 The fitting quality χ^2 produced by reweighting top 1%, top 5%, top 10%, top 15%, top 20%, top 25%, top 30%, top 40%, top 50% of low-energy ensembles and the entire 100% ensembles. Inset: a comparison between experimental SAXS profile and the calculated scattering profile of constructed 5% low-energy ensemble.

For a direct comparison between the ensembles constructed with experimental SAXS profile and simulated energetic data, no other restraints such as ensemble size were imposed in the NNLS fitting scheme. However, we found that the reweighted ensembles of different low-energy levels consisted of a small number (2~5) of conformations with a good fitting to the experiment (Table 1), suggesting DENV2 NS5 may not be a highly flexible system ranging from compact to extended shape that requires a large number of conformations to explain the experimental observables. Interestingly, despite of the enlarging candidate pool from 5% to 30%, their reweighted ensembles show a conserved conformational pattern in terms of the domain-domain orientation and domain contacting interface (Figure 4). In the reweighted 5% low-energy ensemble, two conformations are identified sharing a similar domain-domain orientation ($\varphi \approx -3^\circ$, $\theta \approx 65^\circ$ & $\varphi \approx -1^\circ$, $\theta \approx 60^\circ$) located in the low-energy area of the energy landscape. It is noted that the two conformations with almost equal domain-domain orientation are from different conformational clusters because of their different self-rotations of the MTase domain. The conformations in other constructed low-energy ensembles, like 10%, 15%, 20%, 25% and 30%, showed a similar domain-domain orientation area as the one constructed from the 5% (Figure 4a). To target the structural elements in NS5 corresponding to the conserved orientation space, we identified the contacting residues on the RdRp domain that form pair-wised contacts with the MTase residues with the C α distance smaller than 10 Å. The contacting residues of constructed ensembles from top 5% to 30% low-energy ensemble are highly conserved, majorly consisting of two segments (one loop with partial β -strand (residues 294-307) and another loop region (residues 362-368)) and the cases of top 5%, 15% and 30% low-energy ensembles are shown in Figure 4b. Therefore, the optimal ensemble constructed from the top 5% low-energy structure pool identifies the domain-domain orientation and domain contacting interface, which

reproduces experimental SAXS profile well and agrees best with the energy landscape. Selecting structures from a given over-sampled pool is a common strategy to interpret experimental observables, while the energetic levels of different pool conformations were rarely taken into consideration in the data fitting procedure.

The pre-generated pools from computational approaches based on either molecular topology or physical interactions encounter the problem of energetic reasonableness of a great amount of candidate conformations. Replica exchange molecular dynamics (REMD) simulation is popular in enhancing MD sampling, and some works employed REMD to produce simulated structures for subsequent comparison with experimental measurements^{35, 52}. We also implemented REMD simulations with 24 replicas for 400 ns each on the DENV2 NS5 system and produced a pool of 1000 structures from the MD trajectory at 300 K. The REMD-based pool showed a limited sampling of domain-domain orientation (Figure S1) that was trapped around its initial structures, despite of a comparable total simulation time with regard to our simulations. Considering the limited sampling space, the large energy contrast between different structures in this REMD-based pool has also been observed (Table S1) and is comparable to the counterpart energy landscape in our ensemble modeling scheme, emphasizing the importance of energy examination of pool structures in ensemble construction.

Figure 4 Domain-domain orientations and domain contacting structure segments of the constructed ensembles. (a) Domain-domain orientation distribution of the conformations from reweighted 5%, 15%, 30% low-energy and 100% ensembles, shown as blue triangle dots. Energy contrast is represented by colors as same as Figure 2b. (b) Ribbon representation of RdRp domain with the contacting residues (red) highlighted. Two conserved structure segments dominate the domain-domain interaction interface are suggested by the constructed low-energy ensembles, while the ensemble constructed without energy restraint shows a much more dispersed pattern of contacting interfaces.

Figure 5 reveals the two conformations in the optimal 5% low-energy ensemble with the two contacting structural segments in the RdRp domain highlighted. The two conformations project the same contacting patches in the RdRp domain towards the MTase domain with the linker arranged as an in-between conformation mediating the two domains. The conformation of the ten-residue linker is important for the orientation arrangement between the MTase and RdRp domains, as suggested in the linker-swapping experiment between DENV-3 NS5 and DENV-4 NS5.¹⁴ Among the four serotypes (DENV1 to DENV4), NS5 shares a conserved overall sequence but surprisingly low protein sequence conservation in their linker region. However,

(only) two linker residues E268 and E270 were found conserved among all the serotypes, indicating that they possibly have a crucial role in inter-domain interactions. In the two conformations of the constructed 5% low-energy ensemble, conserved charge-charge contacts between E268/E270 in the linker and R595/R596 in RdRp domain are identified, especially the close contact between E270 and R596 (Figure 5). This specific electrostatic contact has also been found in the crystal structure of DENV3 NS5¹² where E270 interacts with K596 that is a conserved basic counterpart of R596 in DENV 2 (residues numbered on the basis of DENV2 NS5 sequence).

Figure 5 Two conformations of the constructed top 5% low-energy ensemble with aligned RdRp domain (green). Linker is colored in orange and the contacting structure segments in RdRp domain are colored in red. The zoom-in figure shows a conserved electrostatic contact between E268/E270 (blue) in the linker and R595/R596 (gray) in RdRp domain, which has also been found in the crystal structure of DENV3 NS5.

The ensemble of flexible biomolecules can be derived directly from a simulated structure ensemble,³⁵ but the scenario that a good agreement is observed between experiment and a plain simulation does hardly happen on large biomolecules due to many reasons.^{53, 54} Most investigations generated a pool of protein conformations either from molecular simulations^{23, 27}

or from topology-based sampling techniques,^{15, 20, 55} and subsequently adopted the structure pool for ensemble construction and refinement with experimental measurements. In some cases, the optimization procedure was carried out with the structures of good docking scores.⁵⁵ Molecular simulations take the advantage of the exclusion of unphysical conformations in contrast with a pure random structure pool, while encountering the problem of inadequate sampling. To subdue this issue, in our study, we performed multiple independent MD simulations to enhance sampling, and characterized the fully sampled conformation pool in an energy perspective.

A direct fitting by NNLS using all sampled conformations produced an ensemble consisting of 8 structures without a more significant improvement in fitting quality than the low-energy ensembles (Figure 3). However, its conformational space and domain contacting interface appear a more arbitrary pattern compared with its low-energy counterparts (Figure 4). In the reweighted entire ensemble, most of the conformations are located far away from energy minima, and even in high-energy area in the energy landscape of domain-domain orientation. For example, as shown in Figure 6a, three conformations with a similar domain-domain orientation ($\varphi \approx -100^\circ$, $\theta \approx 85^\circ$) are of very unfavorable energies ranked at worse than 90%, whose inter-domain interactions are strongly rejected by both hydrophobic and electrostatic energies. In some cases, one component of the inter-domain interactions repulses the conformational arrangement for which the structure is unstabilized by a higher electrostatic energy ($\varphi \approx 27^\circ$, $\theta \approx 73^\circ$) as shown in Figure 6a. Thus, this ensemble recovered without energy regularization contributes to an average SAXS intensity that numerically agrees well with the experimental quantities, but it might fail to reflect the real conformational space and molecular interactions.

Figure 6 The constructed ensembles from NNLS and GA-based optimizations without energy regularization using the same structure pool from MD simulations. (a) Conformations in the NNLS-reweighted entire (100%) ensemble and their energetic profiles. ΔE (kcal/mol) is the difference between the conformation energy and the average energy over all 1015 representative conformations. ΔE_H and ΔE_{ele} are the hydrophobic and electrostatic components of the inter-domain energy difference. RdRp domains are aligned to one conformation (green), with MTase domain (purple) connected via the linker (orange). (b) Representative conformations in the ensemble constructed with the GA-based optimization method without energy regularization.

Incorporation with Ensemble Optimization Method

The EOM^{15, 16} adopts a GA-based method⁵⁶ for ensemble construction with SAXS data. In order to examine the effect of energy regularization, we input the 1015 representative conformations from molecular modeling as the candidate structure pool and implemented the ensemble construction with EOM. This optimization produced five representative conformations with a χ^2 of 0.35, including some structures of unfavorable energy values whose energy profiles are shown in Figure 6b. Three of them with a total population of 68% have adjacent domain-domain orientations ($-17^\circ < \phi < 13^\circ$, $55^\circ < \theta < 73^\circ$) as the proposed RdRp-binding interface, while the remaining two structures show a very different RdRp-binding site. The five

representative conformations (Figure 6b) showed a partial conserved conformational pattern with the ensemble selected by NNLS (Figure 6a). Three out of five conformations ($\varphi \approx -92^\circ$, $\theta \approx 36^\circ$; $\varphi \approx -105^\circ$, $\theta \approx 79^\circ$ and $\varphi \approx -17^\circ$, $\theta \approx 73^\circ$) are conserved in the NNLS ensemble, whereas the GA-selected ensemble does not incorporate the conformational space of large θ . The two conformations with $\theta > 100^\circ$ selected by NNLS in Figure 6a contributed a total population of 29% in the constructed ensemble. Therefore, despite of the good values of χ^2 , these two optimization methods failed to construct ensembles of a consistent conformational space from the same structure pool, which strongly suggests the importance of extra regularizations in ensemble construction of flexible biomolecules.

Energy as a complementary parameter can be employed in this GA-based optimization scheme. The simulated structures were subjected to the ensemble construction with energy regularization using the EOM approach. Importantly, Figure 7 demonstrates the GA-based optimization with low-energy structure pools produced highly conserved structure ensembles (Figure 7), compared with the low-energy NNLS ensembles in Figure 4. Specifically, top 5% low-energy structures were extracted as a candidate pool, and the GA-based optimization against experimental data constructed a two-structure ensemble of the identical two conformations. With the same 5% low-energy structure pool, GA and NNLS produced the exactly same result ensemble, consisting of the identical two conformation components shown in Figure 4a and Figure 5. The identical result from different ensemble-recovery approaches indicates the high consistency between this proposed low-energy ensemble and experimental observables. The GA-based ensemble constructed from top 15% low-energy structure pool has two structures, both of which have also been identified in the 15% NNLS ensemble. This conserved pattern is the same for the 30% low-energy case where all the four structures in the GA-based ensemble exist in its

NNLS counterpart. Moreover, all these conformational representatives recovered from low-energy structure pools either by NNLS or GA-based approaches (data not shown) target the same protein orientation and domain contacting interface. This physics-based restraint drives NNLS and GA to produce the same (for the 5% ensemble) or very similar (for the 15% and 30% ensembles) result in the NS5 ensemble construction, avoiding the spurious representation of conformational space caused by high-energy structures.

Figure 7 Domain-domain orientations of the constructed ensembles from Ensemble Optimization Method (EOM) with the simulated structure pool. Domain-domain orientation distribution of the conformations from the constructed 5%, 15%, 30% low-energy and 100% ensembles, shown as blue triangle dots. Energy contrast is represented by colors as same as Figure 2b.

Taken the NNLS and EOM results together, the two representative structures from the 5% ensemble are suggested by both optimization methods. Additionally, the conformational space occupied by these two structures is also supported by the 15% and 30% ensemble results of NNLS/EOM, revealing the inter-domain interaction described in Figure 4 and 5. The “true” structure ensemble of a protein in solution should consist of numerous structures according to the basic principle of statistical thermodynamics. Software programs such as EOM define a small

number of representative structures for the explanation of experimental SAXS data. In this work, the small ensemble is constructed from selected low-energy structures from MD simulations to eliminate the instability of the optimization result.

For comparison, the standard EOM computation (without inputting our simulated structure pool) was performed on the ensemble construction using the experimental SAXS profile and the structures of individual domains. The generated pool of 10000 structures was subjected to data fitting with a best fit of $\chi^2 = 0.36$, where the four representative conformations are identified. The conformational space represented by the two most populated conformations is similar to the optimal conformational space above that is suggested by the 5% ensemble. The MTase domain of the two most populated conformations interacts with the same patch on the RdRp domain as the two low-energy structures constructed from top 5% low-energy structure pool (Figure 8a). The two major conformations are slightly different from the two of the 5% ensemble in terms of the domain orientations as seen in Figure 8b, while all these structures are within or close to the same energy basin.

The remaining two minor structures from standard EOM show very different binding sites on the RdRp domain deviating from the 5% low-energy ensemble. Therefore, the standard EOM optimization is able to recover the previously discovered major conformational space of DENV2 NS5 with a different source of candidate structures, but significant discrepancy was observed for the minor states. Interestingly, the pattern of the two minor structures cannot be reproduced in multiple attempts by EOM using the 100% simulated structure pool (Figure 7), and these structures correspond to high energy regions on the energy landscape. Both the lack of numerical reproducibility and the energetic information suggest that the two minor states from the standard EOM approach may originate from data overfitting. Thus, our energy regularization approach

can be considered as a “filter” to remove some questionable structures that may lack a physical ground. It should be noted that the previous work¹⁴ performed EOM optimization employing domain structures derived from the crystal structure of DENV3 NS5,¹² whose domain sequences are highly conserved with DENV2 NS5. Our study adopted a crystallographic MTase structure of DENV2 NS5 and a DENV2 RdRp structure built homologously from DENV3 NS5 (see Methods), because the DENV2 NS5 structure model in this study would be subjected to MD simulations that require DENV2 NS5 sequence instead of DENV3 to correctly model the residue-based physical interactions of DENV2 NS5. The ensemble models in the previous work can be consistently reproduced by using the same domain structures of DENV3 NS5 as input (data not shown).

Figure 8 (a) The standard EOM solution plotted on the energy landscape. The ensemble components are shown in blue triangle dots. The structure pool was generated by the EOM software. (b) Representative conformations of DENV2 NS5 from standard EOM approach with its 10000-structure pool. The four representative conformations have their RdRp domains aligned to a same pose (translucent green envelop) for a direct comparison of domain-domain orientations and MTase domains shown as ribbon representation. The most (deep pink) and second (yellow) most populated conformations share a conserved domain-domain orientation as the two low-energy structures (green and cyan) in the constructed top 5% low-energy ensemble. The other two (minor) ensemble components (blue and purple) from the standard EOM approach have novel domain-domain binding sites.

Over-interpretation of experimental data is a common problem in ensemble recovery, thus extra regularization is usually required in the optimization procedure to drive the solution to desirable properties. Molecular dynamics simulations are increasingly popular in structure modeling of flexible biomolecules with the EOM approach.^{57, 58} It is found in this study that using an MD trajectory as the conformation pool cannot eliminate possible overfitting as high energy states from a simulation can be selected as representative structures with considerable populations. In other words, our results reveal that the SAXS intensity fitting procedure is highly ill-posed, and a well-defined solution can only be guaranteed if the structure pool consists of a small number of physically favorable structures. Thus, the energy-based pool-construction can play an essential role to prevent the noisy conformational interpretation from high-energy structures, such as in the ensemble formation of NS5. It should be noted that the energy gap in our simulations are comparable to those in conventional REMD simulations (Table S1). Therefore, the pool-construction strategy in this work may be applied to SAXS intensity fitting using energy or free energy data generated by other simulation protocols.

Validation of low-energy conformations in all-atom simulations

The last decade has witnessed a significantly growing use of CG models, especially on large biomolecular systems beyond the range of atomistic-scale simulations. CG models take the advantage of speed over atomically detailed models, and its greater efficiency allows a more complete sampling search and a better-converged energy profile.⁵⁹ A good CG model should capture the basis of physical interaction of biomolecules derived from experiments or more detailed models, despite of the tradeoff of CG accuracy considering its simplification nature

averaging out atomistic details.⁶⁰⁻⁶² Our CG model successfully identified the low-energy conformations that reproduced the experimental observables very well, and this optimal ensemble was further validated in a conventional all-atom force field. CG conformations were recovered back to atomistic structures with the program PULCHRA⁴⁶ and then subjected to all-atom MD simulations as initial structures with TIP3P water in a cubic box with GROMACS 4.6.⁴³ A time step of 0.002 ps was used for 20 ns at 300 K under the Amber99SB force field⁶³. The moderate simulation length is adopted to sample the local minimum instead of the global lowest free energy. In order to focus on the inter-domain behavior, we imposed a structure-based LJ-like potential over the native pairs of heavy atoms³⁹ in individual domains.

The two structures in the constructed top 5% low-energy ensemble maintained their stable conformations in atomistic MD simulations (Figure 9a and b). The structure clustering with a backbone RMSD cutoff 5 Å was performed on their respective equilibrium MD trajectory, producing one cluster for each case. The middle structure within the cluster was extracted as the representative for a structure comparison. The representative structure from all-atom simulations in Figure 9a has a backbone RMSD of 4.9 Å compared with its initial conformation ($\varphi \approx -3^\circ$, $\theta \approx 65^\circ$ in Figure 4a). Although a slight domain self-rotation occurred in atomistic simulations, the domain-domain orientation and the basis of interaction pattern were preserved for this conformation of low-energy CG structure model. Besides, the other proposed low-energy conformation ($\varphi \approx -1^\circ$, $\theta \approx 60^\circ$ in Figure 4a) stayed almost unchanged in all-atom MD simulations, whose representative structure deviates from the initial conformation with a backbone RMSD of only 2.8 Å (Figure 9b, inset). Taken together, the two conformations identified as low-energy in CG energy function are stable in the conventional all-atom force field, suggesting this CG model can reflect the physical behavior of domain-domain association even

though it is structurally simplified. In contrast, two high-energy conformations proposed by both the GA-based ensemble and the NNLS ensemble without energy regularization were also investigated in all-atom MD simulations with explicit solvent. Figure 9c and d reveal the large variations of their conformational flexibility deviating away from the initial conformations ($\varphi \approx -105^\circ$, $\theta \approx 79^\circ$ and $\varphi \approx -92^\circ$, $\theta \approx 36^\circ$ respectively in Figure 6) throughout the 20 ns simulations. Conformational change occurred from the start of simulations and produced very large RMSD values against initial structures, demonstrating the two conformations of unfavorable CG energy turn out to be also rejected in the all-atom force field. Since their MD trajectories both showed a continuous change of conformations, the final frames of structures in their trajectories were taken to compare with the initial conformations. As indicated in Figure 9c and d inset, the backbone RMSD values rose to 11.4 Å and 10.4 Å, respectively, with novel domain-domain contacting interfaces.

Figure 9 Backbone RMSD between all-atom MD simulation structures and the initial conformation for NS5 complex (red line), MTase (blue line) domain and RpRd domain (orange line). (a) and (b) are MD simulations starting from two low-energy conformations in the constructed top 5% low-energy ensemble. The representative structure from MD trajectory (red ribbon) is compared with the initial conformation (green ribbon) with their RpRd domains (on the left) aligned. (c) and (d) are for two high-energy conformations proposed in both GA-based

and NNLS ensembles without energy regularization, with the last frame of structure in MD trajectory aligned with the initial conformation.

Theoretical investigation of intensity fitting with noise

To examine the source of the noisy conformational space in the ensembles constructed without energy restraint, we synthesized a pseudo-experimental data as the fitting target. The pseudo-experimental ensemble was constituted by the top ten low-energy structures from the simulated pool, with an equal population of 10% each. Because solution SAXS experiments produce scattering signal with noise, previous SAXS computational studies added noise in the simulated scattering data.⁶⁴ In our study, 5% random Gaussian noise was added to the synthetic data generated by CRY SOL, because this noise level would lead to a comparable fitting quality as that from the experimental SAXS profile. The ensemble construction scheme was implemented in the GA-based EOM approach with the input candidate pools of various energy thresholds. A good representative ensemble derived from experimental data should be able to correctly describe the real conformational space with some representative conformations, and each conformation reflects a cluster of numerous structures in its corresponding free energy basin. As shown in Figure 10 for this sythetic case, the ensembles recovered with low-energy restraints represent a consistent conformational space with the synthetic “solution” ensemble, with χ^2 values of ~0.86-0.87 for each. By contrast, the use of the entire simulated pool fits the synthetic data with a trivial improvement of fit of $\chi^2=0.84$, but this seven-component ensemble identifies three component structures that deviate away from the correct protein orientations. Especially, one ensemble component structure ($\varphi \approx 45^\circ$, $\theta \approx 104^\circ$) significantly deviates away from the correct conformational space and causes a wrong ensemble representation to interpret the pseudo-experimental data. Furthermore, this incorrect representative structure was excluded

from the low-energy pools, which suggests the absence of physics-based regularization would interfere ensemble construction with high-energy noise and cause a spurious conformational representation. Taken together, a raw numerical fitting to experimental SAXS profiles from a random pool exposes the ensemble recovery to the risk of data over-interpretation, and physics-based pool selection is emphasized as a complementary use of current ensemble-construction methods such as EOM.

Figure 10 The constructed ensembles recovered with energy-based pool structures and a synthetic pseudo-experimental data. The χ^2 values are 0.87, 0.86, 0.86 and 0.84 for the low-energy 5%, 15%, 30% and the 100% ensembles, respectively.

Figure 10 also suggests an appropriate and general interpretation of the ensemble fitting results. As none of the computed ensembles in Figure 10 can precisely reproduce the original solution with the small amount of noise, an ensemble solution should only provide us the information of major conformational space or energy basin(s). For multi-domain proteins such as NS5, the inter-domain orientation and the binding interface can be identified, as shown in the fitting of experimental data in this work. For the synthesized data case an energy threshold of 5%-30% can produce reasonable representative structures that are close to the solution conformations. In applications where the true solution is unknown, it is suggested to test a few

threshold values that will not significantly lower the fitting accuracy, and discard any ensemble solution with a significant population of high-energy conformations. The spurious minor states from intensity fitting could be further examined and excluded by MD simulations.

Conclusion

Small-angle X-ray scattering (SAXS) characterizes macromolecular systems at medium to low resolution and its insensitivity to atomistic structural details permits SAXS profile to be modeled on the coarse-grained level. We used an established CG model³⁴ of multi-domain protein and multi-protein complex to simulate structure ensemble of DENV2 NS5, with a modification of modeling flexible structures with electrostatic and hydrophobic energies. The flexible linkers can control the relative locations of different structure units and the flexible loops commonly found on protein surface are involved in protein-protein associations, which is crucial for protein conformation and function regulation.⁶⁵⁻⁶⁹ This physical treatment of flexible structures enabled us to identify the specific electrostatic contact between E268/E270 in the linker and R595/R596 in RdRp domain that is a conserved charge-charge interaction identified in the crystal structure of NS5.¹² This modeling scheme of protein linker can be extended to all systems containing intrinsically disordered regions (IDRs) that usually act as a key role in many molecular activities such as molecular recognition and signaling.⁷⁰

Ensemble construction from a pre-generated structure pool always faces the over-fitting plight for flexible biomolecular systems. Here, we did not limit ensemble size in the non-negativity least squares (NNLS) optimization, because the algorithm enables a direct comparison between the recovered ensembles and simulated energetic data, and a small amount of conformations is

able to reproduce SAXS profile well in the NS5 case. Additionally, two optimization methods, NNLS and genetic algorithm, were performed individually with or without the energy restraint on the same structure pool characterized from MD simulations. Importantly, the complementary use of energy regularization drives the two methods to recover the experimental observables to the same structure ensemble by precluding the noise from energy-unfavorable conformations. The exclusion of energy restraint causes inconsistent conformational representations, which suggests a raw numerical fitting would cause an ambiguous interpretation of experimental observables. It is also revealed by a theoretical investigation that a low level of experimental error can generate the incorrect high-energy ensemble components. We found that the ensemble construction with low-energy regularization targeted the conformational ensemble with both favorable energy and high consistency with experiment, successfully identifying the domain-domain orientation and domain contacting interface of DENV2 NS5. The solution-regularization strategy in this work is general, as it can be implemented with different fitting algorithms and various simulation approaches.

Author Information

Corresponding Author

*E-mail: LYL@ntu.edu.sg; Phone: +65 63162866

Author Contributions

Conceived and designed the experiments: GZ GG LL. Performed the experiments: GZ AN LL. Analyzed the data: GZ WGS AN GG LL. Wrote the paper: GZ GG LL.

Notes

The author declares no competing financial interest.

Supporting Information

The Supporting Information (Figure S1 and Table S1) is available.

Acknowledgements

This research is supported by the Tier 2 grant (MOE2014-T2-1-065) and the Tier 3 grant (MOE2012-T3-1-008) from the Ministry of Education, Singapore.

Reference

1. Mackenzie, J. Wrapping things up about virus RNA replication. *Traffic* **2005**, *6* (11), 967-977.
2. Welsch, S.; Miller, S.; Romero-Brey, I.; Merz, A.; Bleck, C. K.; Walther, P.; Fuller, S. D.; Antony, C.; Krijnse-Locker, J.; Bartenschlager, R. Composition and three-dimensional architecture of the dengue virus replication and assembly sites. *Cell Host Microbe* **2009**, *5* (4), 365-375.
3. Fernandez-Garcia, M. D.; Meertens, L.; Bonazzi, M.; Cossart, P.; Arenzana-Seisdedos, F.; Amara, A. Appraising the roles of CBLL1 and the ubiquitin/proteasome system for flavivirus entry and replication. *J. Virol.* **2011**, *85* (6), 2980-2989.
4. Perera, R.; Kuhn, R. J. Structural proteomics of dengue virus. *Curr. Opin. Microbiol.* **2008**, *11* (4), 369-377.
5. Bollati, M.; Alvarez, K.; Assenberg, R.; Baronti, C.; Canard, B.; Cook, S.; Coutard, B.; Decroly, E.; de Lamballerie, X.; Gould, E. A., et al. Structure and functionality in flavivirus NS-proteins: Perspectives for drug design. *Antiviral Res.* **2010**, *87* (2), 125-148.
6. Egloff, M. P.; Benarroch, D.; Selisko, B.; Romette, J. L.; Canard, B. An RNA cap (nucleoside-2'-O-)-methyltransferase in the flavivirus RNA polymerase NS5: Crystal structure and functional characterization. *EMBO J.* **2002**, *21* (11), 2757-2768.
7. Issur, M.; Geiss, B. J.; Bougie, I.; Picard-Jean, F.; Despins, S.; Mayette, J.; Hobdey, S. E.; Bisailon, M. The flavivirus NS5 protein is a true RNA guanylyltransferase that catalyzes a two-step reaction to form the RNA cap structure. *RNA* **2009**, *15* (12), 2340-2350.
8. Hannemann, H.; Sung, P. Y.; Chiu, H. C.; Yousuf, A.; Bird, J.; Lim, S. P.; Davidson, A. D. Serotype-specific differences in dengue virus non-structural protein 5 nuclear localization. *J. Biol. Chem.* **2013**, *288* (31), 22621-22635.
9. Tay, M. Y.; Fraser, J. E.; Chan, W. K.; Moreland, N. J.; Rathore, A. P.; Wang, C.; Vasudevan, S. G.; Jans, D. A. Nuclear localization of dengue virus (DENV) 1-4 non-structural protein 5; protection against all 4 DENV serotypes by the inhibitor Ivermectin. *Antiviral Res.* **2013**, *99* (3), 301-306.
10. Lim, S. P.; Wang, Q. Y.; Noble, C. G.; Chen, Y. L.; Dong, H.; Zou, B.; Yokokawa, F.; Nilar, S.; Smith, P.; Beer, D., et al. Ten years of dengue drug discovery: Progress and prospects. *Antiviral Res.* **2013**, *100* (2), 500-519.
11. Lu, G.; Gong, P. Crystal structure of the full-length Japanese encephalitis virus NS5 reveals a conserved methyltransferase-polymerase interface. *PLoS Pathog.* **2013**, *9* (8), e1003549.
12. Zhao, Y.; Soh, T. S.; Zheng, J.; Chan, K. W.; Phoo, W. W.; Lee, C. C.; Tay, M. Y.; Swaminathan, K.; Cornvik, T. C.; Lim, S. P., et al. A crystal structure of the Dengue virus NS5 protein reveals a novel inter-domain interface essential for protein flexibility and virus replication. *PLoS Pathog.* **2015**, *11* (3), e1004682.
13. Bussetta, C.; Choi, K. H. Dengue virus nonstructural protein 5 adopts multiple conformations in solution. *Biochemistry* **2012**, *51* (30), 5921-5931.
14. Saw, W. G.; Tria, G.; Grüber, A.; Subramanian Manimekalai, M. S.; Zhao, Y.; Chandramohan, A.; Srinivasan Anand, G.; Matsui, T.; Weiss, T. M.; Vasudevan, S. G., et al. Structural insight and flexible features of NS5 proteins from all four serotypes of Dengue virus in solution. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2015**, *71*, 2309-2327.
15. Bernado, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **2007**, *129* (17), 5656-5664.
16. Tria, G.; Mertens, H. D.; Kachala, M.; Svergun, D. I. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2015**, *2*, 207-217.

17. Salmon, L.; Nodet, G.; Ozenne, V.; Yin, G.; Jensen, M. R.; Zweckstetter, M.; Blackledge, M. NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.* **2010**, *132* (24), 8407-8418.
18. Guerry, P.; Salmon, L.; Mollica, L.; Ortega Roldan, J. L.; Markwick, P.; van Nuland, N. A.; McCammon, J. A.; Blackledge, M. Mapping the population of protein conformational energy sub-states from NMR dipolar couplings. *Angew. Chem. Int. Ed.* **2013**, *52* (11), 3181-3185.
19. Pelikan, M.; Hura, G. L.; Hammel, M. Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys.* **2009**, *28* (2), 174-189.
20. Berlin, K.; Castaneda, C. A.; Schneidman-Duhovny, D.; Sali, A.; Nava-Tudela, A.; Fushman, D. Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *J. Am. Chem. Soc.* **2013**, *135* (44), 16595-16609.
21. Yang, S.; Blachowicz, L.; Makowski, L.; Roux, B. Multidomain assembled states of Hck tyrosine kinase in solution. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (36), 15757-15762.
22. Chen, Y.; Campbell, S. L.; Dokholyan, N. V. Deciphering protein dynamics from NMR data using explicit structure sampling and selection. *Biophys. J.* **2007**, *93* (7), 2300-2306.
23. Kim, Y. C.; Tang, C.; Clore, G. M.; Hummer, G. Replica exchange simulations of transient encounter complexes in protein-protein association. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (35), 12855-12860.
24. Huang, J.-R.; Grzesiek, S. Ensemble calculations of unstructured proteins constrained by RDC and PRE data: A case study of urea-denatured ubiquitin. *J. Am. Chem. Soc.* **2009**, *132* (2), 694-705.
25. Fisher, C. K.; Ullman, O.; Stultz, C. M. Comparative studies of disordered proteins with similar sequences: Application to A β 40 and A β 42. *Biophys. J.* **2013**, *104* (7), 1546-1555.
26. Choy, W. Y.; Forman-Kay, J. D. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **2001**, *308* (5), 1011-1032.
27. Rozycki, B.; Kim, Y. C.; Hummer, G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* **2011**, *19* (1), 109-116.
28. Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106* (4), 620.
29. Francis, D. M.; Rózycki, B.; Koveal, D.; Hummer, G.; Page, R.; Peti, W. Structural basis of p38 α regulation by hematopoietic tyrosine phosphatase. *Nat. Chem. Biol.* **2011**, *7* (12), 916-924.
30. Zheng, W.; Tekpinar, M. Accurate flexible fitting of high-resolution protein structures to small-angle x-ray scattering data using a coarse-grained model with implicit hydration shell. *Biophys. J.* **2011**, *101* (12), 2981-2991.
31. Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **2005**, *109* (7), 2469-2473.
32. Lu, L.; Voth, G. A. The multiscale coarse-graining method. *Adv. Chem. Phys.* **2012**, *149*, 47-81.
33. Kim, Y. C.; Hummer, G. Coarse-grained models for simulations of multiprotein complexes: Application to ubiquitin binding. *J. Mol. Biol.* **2008**, *375* (5), 1416-1433.
34. Ravikumar, K. M.; Huang, W.; Yang, S. Coarse-grained simulations of protein-protein association: An energy landscape perspective. *Biophys. J.* **2012**, *103* (4), 837-845.
35. Terakawa, T.; Higo, J.; Takada, S. Multi-scale ensemble modeling of modular proteins with intrinsically disordered linker regions: Application to p53. *Biophys. J.* **2014**, *107* (3), 721-729.
36. Shea, J. E.; Onuchic, J. N.; Brooks, C. L. Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (22), 12512-12517.
37. Clementi, C.; Nymeyer, H.; Onuchic, J. N. 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.* **2000**, *298* (5), 937-953.

38. Koga, N.; Takada, S. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Go-like model. *J. Mol. Biol.* **2001**, *313* (1), 171-180.
39. Noel, J. K.; Whitford, P. C.; Onuchic, J. N. The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function. *J. Phys. Chem. B* **2012**, *116* (29), 8692-8702.
40. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M. Y.; Pieper, U.; Sali, A. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics* **2006**, Chapter 5, Unit 5.6.
41. Huang, W.; Ravikumar, K. M.; Yang, S. A newfound cancer-activating mutation reshapes the energy landscape of estrogen-binding domain. *J. Chem. Theory Comput.* **2014**, *10* (8), 2897-2900.
42. Miyazawa, S.; Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **1996**, *256* (3), 623-644.
43. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435-447.
44. Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed.* **1999**, *38* (1-2), 236-240.
45. Svergun, D.; Barberato, C.; Koch, M. H. J. CRY SOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **1995**, *28* (6), 768-773.
46. Rotkiewicz, P.; Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **2008**, *29* (9), 1460-1465.
47. Press, W. H. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge University Press: Cambridge, U.K., 2007.
48. Yang, S.; Parisien, M.; Major, F.; Roux, B. RNA structure determination using SAXS data. *J. Phys. Chem. B* **2010**, *114* (31), 10039-10048.
49. Schneidman-Duhovny, D.; Hammel, M.; Tainer, J. A.; Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **2013**, *105* (4), 962-974.
50. Schneidman-Duhovny, D.; Hammel, M.; Sali, A. FoXS: A web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* **2010**, *38*, W540-544.
51. Pham, G. H.; Rana, A. S.; Korkmaz, E. N.; Trang, V. H.; Cui, Q.; Strieter, E. R. Comparison of native and non-native ubiquitin oligomers reveals analogous structures and reactivities. *Protein Sci.* **2016**, *25* (2), 456-471.
52. Metskas, L. A.; Rhoades, E. Conformation and dynamics of the troponin I C-terminal domain: Combining single-molecule and computational approaches for a disordered protein region. *J. Am. Chem. Soc.* **2015**, *137* (37), 11962-11969.
53. van Gunsteren, W. F.; Dolenc, J.; Mark, A. E. Molecular simulation as an aid to experimentalists. *Curr. Opin. Struct. Biol.* **2008**, *18* (2), 149-153.
54. van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glattli, A.; Hunenberger, P. H., et al. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem. Int. Ed. Engl.* **2006**, *45* (25), 4064-4092.
55. Kozakov, D.; Li, K.; Hall, D. R.; Beglov, D.; Zheng, J.; Vakili, P.; Schueler-Furman, O.; Paschalidis, I.; Clore, G. M.; Vajda, S. Encounter complexes and dimensionality reduction in protein-protein association. *Elife* **2014**, *3*, e01370.
56. Jones, G. *Genetic and evolutionary algorithms*. John Wiley and Sons: Chichester, U.K., 1998.
57. Zhang, Y. H.; Wen, B.; Peng, J. H.; Zuo, X. B.; Gong, Q. G.; Zhang, Z. Y. Determining structural ensembles of flexible multi-domain proteins using small-angle X-ray scattering and molecular dynamics simulations. *Protein Cell* **2015**, *6* (8), 619-623.

58. Wen, B.; Peng, J. H.; Zuo, X. B.; Gong, Q. G.; Zhang, Z. Y. Characterization of protein flexibility using small-angle X-ray scattering and amplified collective motion simulations. *Biophys. J.* **2014**, *107* (4), 956-964.
59. Takada, S. Coarse-grained molecular simulations of large biomolecules. *Curr. Opin. Struct. Biol.* **2012**, *22* (2), 130-137.
60. Rader, A. J. Coarse-grained models: Getting more with less. *Curr. Opin. Pharmacol.* **2010**, *10* (6), 753-759.
61. Riniker, S.; Allison, J. R.; van Gunsteren, W. F. On developing coarse-grained models for biomolecular simulation: A review. *Phys. Chem. Chem. Phys.* **2012**, *14* (36), 12423-12430.
62. Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **2013**, *139* (9), 090901.
63. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65* (3), 712-725.
64. Konarev, P. V.; Svergun, D. I. A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. *IUCr* **2015**, *2*, 352-360.
65. Ritchie, D. W. Recent progress and future directions in protein-protein docking. *Curr. Protein Pept. Sci.* **2008**, *9* (1), 1-15.
66. Swain, J. F.; Dinler, G.; Sivendran, R.; Montgomery, D. L.; Stotz, M.; Gierasch, L. M. Hsp70 chaperone ligands control domain association via an allosteric mechanism mediated by the interdomain linker. *Mol. Cell* **2007**, *26* (1), 27-39.
67. Chong, P. A.; Lin, H.; Wrana, J. L.; Forman-Kay, J. D. Coupling of tandem Smad ubiquitination regulatory factor (Smurf) WW domains modulates target specificity. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (43), 18404-18409.
68. Smagghe, B. J.; Huang, P. S.; Ban, Y. E.; Baker, D.; Springer, T. A. Modulation of integrin activation by an entropic spring in the {beta}-knee. *J. Biol. Chem.* **2010**, *285* (43), 32954-32966.
69. Manimekalai, S.; Saw, W. G.; Pan, A.; Grüber, A.; Grüber, G. Identification of the critical linker residues conferring differences in the compactness of NS5 from Dengue virus serotype 4 and NS5 from Dengue virus serotypes 1-3. *Acta Crystallogr. D* **2016**, *72* (6), 795-807.
70. Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. Intrinsic disorder and protein function. *Biochemistry* **2002**, *41* (21), 6573-6582.

TOC graphic