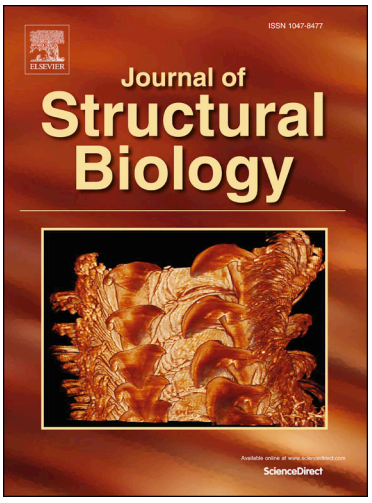
Journal Pre-proofs



The SIRAH force field: a suite for simulations of complex biological systems at the coarse-grained and multiscale levels

Florencia Klein, Martín Soñora, Lucianna Helene Santos, Ezequiel Nazareno Frigini, Andrés Ballesteros-Casallas, Matías Rodrigo Machado, Sergio Pantano

PII: S1047-8477(23)00048-5
DOI: https://doi.org/10.1016/j.jsb.2023.107985
Reference: YJSBI 107985

To appear in: *Journal of Structural Biology*

Received Date: 15 March 2023
Revised Date: 18 May 2023
Accepted Date: 13 June 2023

Please cite this article as: Klein, F., Soñora, M., Helene Santos, L., Nazareno Frigini, E., Ballesteros-Casallas, A., Rodrigo Machado, M., Pantano, S., The SIRAH force field: a suite for simulations of complex biological systems at the coarse-grained and multiscale levels, *Journal of Structural Biology* (2023), doi: https://doi.org/10.1016/j.jsb.2023.107985

The SIRAH force field: a suite for simulations of complex biological systems at the coarse-grained and multiscale levels.

Florencia Klein[1,2,†], Martín Soñora[1,†], Lucianna Helene Santos[1], Ezequiel Nazareno Frigini[3], Andrés Ballesteros-Casallas[1,4], Matías Rodrigo Machado[1], and Sergio Pantano[1,4,*]

1- Institut Pasteur de Montevideo, Montevideo Uruguay

2- Laboratoire de Biochimie Théorique, UPR9080, CNRS, Paris, France

3- Instituto Multidisciplinario de Investigaciones Biológicas de San Luis (IMIBIO-SL), Universidad Nacional de San Luis - CONICET. San Luis, Argentina

4- Facultad de Química, Universidad de la República, Uruguay.

** Correspondence to:* spantano@pasteur.edu.uy

† *equally contributed*

Abstract: The different combinations of molecular dynamics simulations with coarse-grained representations have acquired considerable popularity among the scientific community. Especially in biocomputing, the significant speedup granted by simplified molecular models opened the possibility of increasing the diversity and complexity of macromolecular systems, providing realistic insights on large assemblies for more extended time windows.

However, a holistic view of biological ensembles' structural and dynamic features requires a self-consistent force field, namely, a set of equations and parameters that describe the intra and intermolecular interactions among moieties of diverse chemical nature (i.e., nucleic and amino acids, lipids, solvent, ions, etc.). Nevertheless, examples of such force fields are scarce in the literature at the fully atomistic and coarse-grained levels. Moreover, the number of force fields capable of handling simultaneously different scales is restricted to a handful. Among those, the SIRAH force field, developed in our group, furnishes a set of topologies and tools that facilitate the setting up and running of molecular dynamics simulations at the coarse-grained and multiscale levels. SIRAH uses the same classical pairwise Hamiltonian function implemented in the most popular molecular dynamics software. In particular, it runs natively in AMBER and Gromacs engines, and porting it to other simulation packages is straightforward.

This review describes the underlying philosophy behind the development of SIRAH over the years and across families of biological molecules, discussing current limitations and future implementations.

Keywords: SIRAH, Coarse-Grained, Molecular dynamics, simulations, multiscale models.

1. INTRODUCTION

1.1 Molecular Dynamics and Coarse-Grained models

Molecular dynamics (MD) simulations have become a method of choice for studying the dynamics of biological assemblies. The high accuracy attained by state-of-the-art methods, together with the ability to simulate increasingly realistic biological conditions such as temperature, ligands, salt concentration, and molecular modifications, have led MD techniques to be considered as a computational microscope (Dror et al., 2012; Lee et al., 2009). Constant developments in software and hardware allowed us to go from simulations of small, single proteins in the timescale of the picoseconds, performed over forty years ago, to complex systems made of millions of atoms in the timescale of micro and, exceptionally, milliseconds performed nowadays. A remarkable example of application is constituted by the use of MD simulations to study the structural stability of large virus-like particles (VLPs) provided by cutting-edge CryoEM techniques. The combined use of fully atomistic and coarse-grained (CG) MD simulations of entire viruses to complement experimental techniques and improve our current knowledge of emerging pathogens constitutes remarkable examples of the utility of these methods in biomedical research (Hadden and Perilla, 2018; Jones et al., 2021; Machado and Pantano, 2021; Sztain et al., 2021; Yu et al., 2021).

Intermolecular interactions in present MD simulations are frequently calculated using a classical two-body Hamiltonian to solve Newton's equations of motion. The MD Hamiltonian represents chemical bonds with harmonic terms and periodic functions in their most common form. At the same time, non-bonded interactions are treated by Lennard-Jones (LJ) and Coulomb's potentials (Equation 1) (Bayly et al., 2002). Most popular all-atom force fields have converged over the years to a set of interaction parameters that represents the best compromise between accuracy and computational cost. However, combined with this formalism, the physicochemical nature of the systems of interest requires a relatively small integration time step to obtain a proper dynamical sampling of the fastest oscillations. In biological systems, they correspond to the oscillations of Hydrogen bound to Oxygen atoms. Therefore, integration time steps in all-atoms simulations are upper bound to up to 4 femtoseconds (Hopkins et al., 2015). This relatively small integration step constitutes one of the limiting factors for fully atomistic simulations, as it requires massive amounts of computer time to achieve biophysical/biologically meaningful results.

The considerable computational cost associated with MD simulations motivates a continuous interest in developing cost-effective approximations to increase the system's complexity and spatio-temporal sampling (Ingólfsson et al., 2014; Reith et al., 2003). This interest has driven the development of CG approximations that reduce the computational burden but still capture key

physicochemical interactions that rule biological phenomena. Since the pioneering work of Levitt and Warshel (Levitt and Warshel, 1975), a plethora of physically inspired models have been successfully used to describe intricate processes ranging from protein folding (Davtyan et al., 2012; Onuchic and Wolynes, 2004; Scheraga et al., 2007) to membrane fusion (Marrink and Mark, 2003), protein-DNA interactions (Brandner et al., 2018; Dai and Yu, 2020) and virus assembly, just to quote a few.
During the last two decades, a number of CG models that retain residue-level chemical specificity have been reported to perform MD simulations of different families of biological molecules (Derreumaux and Mousseau, 2007; Hinckley et al., 2013; Kar et al., 2013; Kenzaki et al., 2011; Marrink and Tieleman, 2013; Orsi and Essex, 2011; Pasi et al., 2013; Scheraga et al., 2007; Seo and Shinoda, 2019; Shinoda et al., 2011; Sterpone et al., 2014). Most of these initiatives have been recently reviewed (Singh and Li, 2019).

This work describes the effort devoted by the Biomolecular Simulations group at the Institut Pasteur de Montevideo to develop a residue-based, general-purpose, CG force field for biomolecular simulations named SIRAH (Southamerican Initiative for a Rapid and Accurate Hamiltonian). Our force field has now been ported to the popular MD engines, AMBER (https://ambermd.org/) and GROMACS (https://www.gromacs.org/) and has been part of the official AMBER release since 2020 (Case et al., 2020).

At the current stage, SIRAH, freely available at http://www.sirahff.com, provides plug & play parameters and CG topologies for aqueous solvent (water and electrolytes), phospholipids, DNA, metal ions, and proteins. Its latest version, SIRAH 2.0 includes: (i) a series of modifications to both bonded and non-bonded parameters of amino acids while preserving their original topologies; (ii) a description of different protonation states and post-translational modifications for protein residues; (iii) an improvement in the compatibility for mapping different force fields atom types and experimental structures; and (iv) the ability to leverage GPU acceleration in AMBER and GROMACS codes.

Moreover, a substantial effort has been dedicated to making the package easy to use. We created a collection of scripts, referred to as SIRAH Tools (Machado and Pantano, 2016), that facilitate the process of mapping all-atom files to CG representations, backmapping, visualizing, and analyzing SIRAH trajectories directly on the popular VMD program for molecular visualization. Step-by-step tutorials for setting up and simulating different systems are also available on our website. Furthermore, a web server for setting up and running MD-CG simulations in GROMACS is available at https://molsim.sci.univr.it/mermaid/sirah/sirah.php (Marchetto et al., 2020).

The following paragraphs provide an overview of SIRAH's underlying philosophy and its most salient features. Readers interested in specific and more quantitative details are referred to the original publications.

2. The SIRAH force field for CG simulations.

A fundamental step in the coarse-graining process is the mapping between the all-atoms representation and the CG one. Even though some methods for systematic coarse-graining have been presented in the literature (Dama et al., 2013; Dannenhoffer-Lafage et al., 2016; Davtyan et al., 2016), the mapping is often decided ad-hoc to address the solution to specific problems. For instance, some protein models feature only one bead per amino acid (frequently on the position of the Cα carbons) to describe protein folding processes (Bahar and Jernigan, 1997; Clementi et al., 2000).

However, this intuitive choice may not be well suited for CG models intended for general purposes. In particular, the distance between Cα carbons changes if peptides are in extended, coil, or helical configurations, making it challenging to attain an unbiased conformational description for models using only one bead in that position.

A workaround for that limitation was presented in the popular MARTINI CG model, in which the protein backbone is represented by only one bead per amino acid (Monticelli et al., 2008). In this case, specific constraints are imposed to maintain structural elements. Although this alternative is adequate for representing different backbone configurations, it precludes unbiased conformational sampling (Marrink et al., 2023; Monticelli et al., 2008). In contrast with this mapping strategy, other CG force fields use effective beads at the positions of all the backbone atoms, while the side chains are described at a much lower level of detail (Davtyan et al., 2012; Derreumaux and Mousseau, 2007; Sterpone et al., 2014).

In the case of SIRAH, the CG mapping is performed by keeping effective interaction beads in the position of a subset of atoms that are determinant for the structure or intermolecular interactions. The positions of these beads are designed to correspond with the intended interactions of the selected functional groups in terms of size and charge. Our mapping scheme results from a pragmatic combination of empirical databases, knowledge of organic chemistry (including canonical structures), and physicochemical insights. As illustrated in the following sections, this leads to a heterogeneous distribution with higher bead density in moieties that establish more diverse intermolecular interactions. The all-atoms to CG mapping of different biomolecular families are described in the next sections.

Once the CG mapping is defined, the next decision to make regards the nature of the interaction potential among the different beads. We made a couple of

strategic choices. First, we adopted the classical two-body Hamiltonian used by most MD simulation packages (Equation 1).

$$V(r^N) = \sum_{i \in \text{bonds}} k_{bi}(l_i - l_i^0)^2 + \sum_{i \in \text{angles}} k_{ai}(\theta_i - \theta_i^0)^2$$

$$+ \sum_{i \in \text{torsions}} \sum_n \frac{1}{2} V_i^n [1 + \cos(n\omega_i - \gamma_i)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

Equation 1: The classical Hamiltonian used in most popular MD simulations packages. This particular functional form corresponds to that used in AMBER (Bayly et al., 2002).

Therefore, it is immediate to use the CG force field in virtually any MD engine, profiting from decades of theoretical advancements from leading research groups globally, leveraging GPU optimizations, and constantly improving simulation algorithms. Moreover, the usual concepts of atom type/name, partial charges, etc., remain unchanged in our force field, obviating the need to learn new setups or file formats. Practically, anyone with the basic knowledge to run a standard all-atoms MD simulation in AMBER or GROMACS can run a CG simulation using SIRAH. Indeed, the MD engines "do not know" they are running a CG simulation.

However, using this classical two-body Hamiltonian requires the determination of a relatively high number of parameters. This led us to the second strategic choice. As per the predetermined mapping guidelines, CG moieties in SIRAH contain beads placed in the position of selected atoms in their fully atomistic representation. Therefore, all equilibrium distances can be directly derived from statistical data from the Protein Data Bank (PDB, https://www.rcsb.org), quantum-level calculations, or canonical conformations defined from organic chemistry rules. This mapping strategy considerably reduces the number of parameters to be determined and facilitates the implementation of a simple backmapping scheme that recovers pseudo-atomistic information at any frame of the MD simulation (Machado and Pantano, 2016).

SIRAH's first CG model was DNA. Thus, the force constants in the bonded terms, partial charges, and LJ parameters were initially derived by trial & error simulations on double-stranded dodecameric DNA segments by imposing the same value on all harmonic bonds and angles. Six effective beads were used to represent DNA's four nucleotides, with the bead sizes corresponding to their respective atomistic functional group, backbone, or base (Figure 1). Most of the DNA's parameters were then transferred and, eventually, adapted to different molecular moieties, using the general concept that similar functional groups should have analogous interaction parameters. For instance, protein residues use three beads for backbone nitrogen, Cα carbon, and carboxylic Oxygen, while side chains are depicted as beads with diverse sizes and charges based on the physicochemical properties of the amino acid, such as hydrophobicity, aromaticity, and salt bridges.

The following subsections provide a comprehensive overview of some of SIRAH's CG models.

2.1 The CG DNA model.

DNA was the first CG model developed by our group (Dans et al., 2010). The four CG nucleotides in our model are made of six effective beads that respond to the following (gross) physicochemical vision of DNA:

i) At least two backbone beads are necessary to represent the 5' – 3' prime polarity, and they should carry a -1 charge, typical of the polyanion. Therefore, we represent the DNA backbone by two beads placed at the phosphate and C5' Carbon positions, with the phosphate bead carrying a -1 charge.

ii) A-T and G-C base pairs recognize each other through electrostatic complementarity. So, we place three beads on the position of atoms on the so-called Watson-Crick edge. Partial charges on those beads ensure electrostatic recognition.

iii) The bead sizes in the Watson-Crick edge should be atomistic to allow for the correct base-base stacking, while those in the backbone can be bulkier.

iv) Finally, the details of the sugar moiety are neglected, and the 5-member ring is simply represented by only one bead in the C1' position connecting the backbone to the Watson-Crick edge.

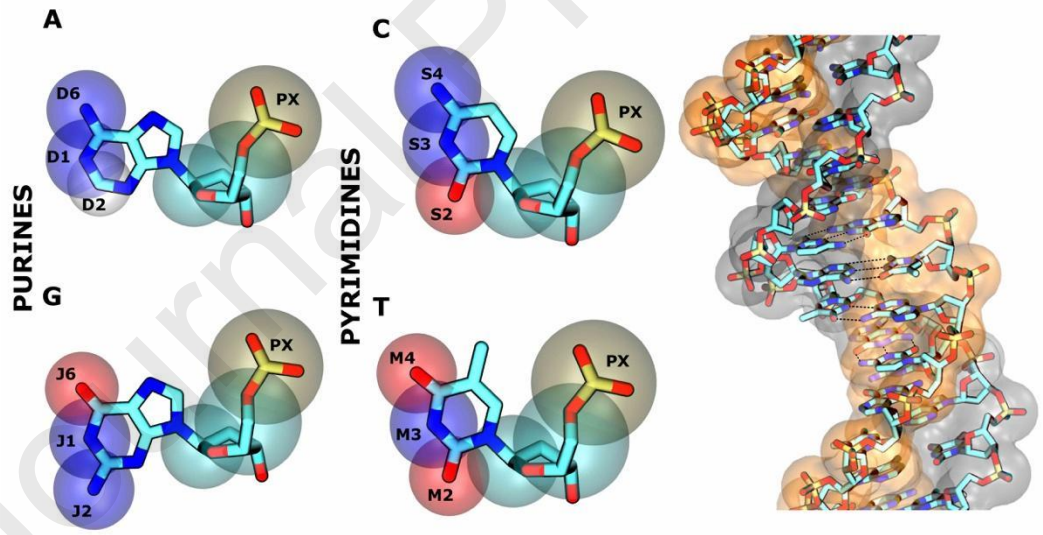


Figure 1: Mapping of the SIRAH DNA model. The heavy atoms representation is superimposed on the CG topology. Left: CG beads for the four nucleotides are drawn with their actual LJ size. The color of the CG beads corresponds to the atomic position where they are mapped (cyan: Carbon, blue: Nitrogen, red: Oxygen, brown: Phosphorus). Right: superposition of double-stranded DNA at atomistic and CG levels. Both strands are shown with semitransparent solvent-accessible surfaces at the CG level.

This mapping choice (Figure 1) allows for preserving the specific base-pair recognition of the B-form of DNA. Similarly, the distorting effect of mismatches is correctly captured, as the size and gross electrostatics signature of the four

nucleotides are correctly represented. Nevertheless, less frequent internucleotide interactions involving the sugar edge or Hoogsteen base pairs interactions are precluded from this CG mapping and have to be considered as limitations of the model.

In SIRAH, the size of the beads determined by the LJ parameters is heterogeneous (Figure 1). This characteristic was initially motivated by the necessity of maintaining the correct stacking distance between consecutive bases in a double-stranded configuration while accounting for the exposed molecular surface. Indeed, the CG model represents a compact double helix where minor and major grooves are clearly recognized (Figure 1). Therefore, effective beads representing the bases adopted the LJ size of the Barcelona force field (bsc0) implemented in the AMBER package for MD simulations (Pérez et al., 2007). Beads representing the backbone were assigned a bigger size, which was determined by a spherical approximation to the excluded volume of the corresponding functional group. Initial guesses for the depth of the LJ potential were also taken from the bsc0 force field and progressively adjusted.

Partial charges were assigned with the primary criterion of adding a negative integer to every base. To this end, a -1 charge was assigned to the phosphate beads, and partial charges adding to zero were assigned to the Watson-Crick edges to ensure electrostatic recognition in A-T and G-C base pairs. The value of the partial charges was fitted to roughly reproduce (within 10%) the electrostatic potential created by the bsc0 force field in the grooves of a double-stranded dodecahedron.

Finally, the masses were initially distributed by summing up those corresponding to the atoms included in the functional groups represented by each bead. These choices, along with a heterogeneous mass distribution, allowed running MD simulations using a timestep of 5 fs.

This first version of a CG DNA model was developed to work within the Generalized Born model for implicit solvation, as implemented in AMBER (Case et al., 2008).

This initial model was shown to reproduce the structure and dynamics of double-stranded DNA at a resolution comparable to atomistic force fields as well as breathing profiles and melting temperatures of segments of different lengths, sequences, and on a range of ionic strengths (Dans et al., 2013).

Worth mentioning, although the model was parameterized to reproduce the B form of double-stranded DNA, it reproduces the spontaneous formation of large “bubbles” within double-stranded DNA, fraying and rehybridization of terminal tails and nicely matched experimentally determined persistence lengths of single-stranded filaments (Zeida et al., 2012).

In addition, with the development of our CG water model (see next subsection), a uniform mass redistribution was implemented that boosted the performance of our CG DNA, granting a four-fold increase in the timestep.

In its final version, it achieves a speedup of 3-4 orders of magnitude using implicit solvation and two orders of magnitude in comparison with all-atoms explicit solvent simulations.

2.2 The WatFour (WT4) model for CG explicit solvent.

In parallel with the development of the DNA model, we started the development of a CG aqueous solvent that included a representation of CG water and monovalent electrolytic ions, namely sodium, potassium, and chloride. Most CG solvents lump several water molecules into a single, larger spherical particle. They interact in different ways according to the physics underlying the model. Some carry partial charges to address electrostatic screening, while others simply rely on LJ potentials (Darré et al., 2012a; Noid, 2013). As SIRAH is a structure-based force field, our goal was to replicate the configuration of an elementary water cluster. This transient 3D structure includes a central water surrounded by four identical molecules positioned at the vertices of a tetrahedron, as shown in Figure 2A. Since the central water molecule is already fulfilling its capacity to form hydrogen bonds, it can be excluded in a CG representation of a water cluster.

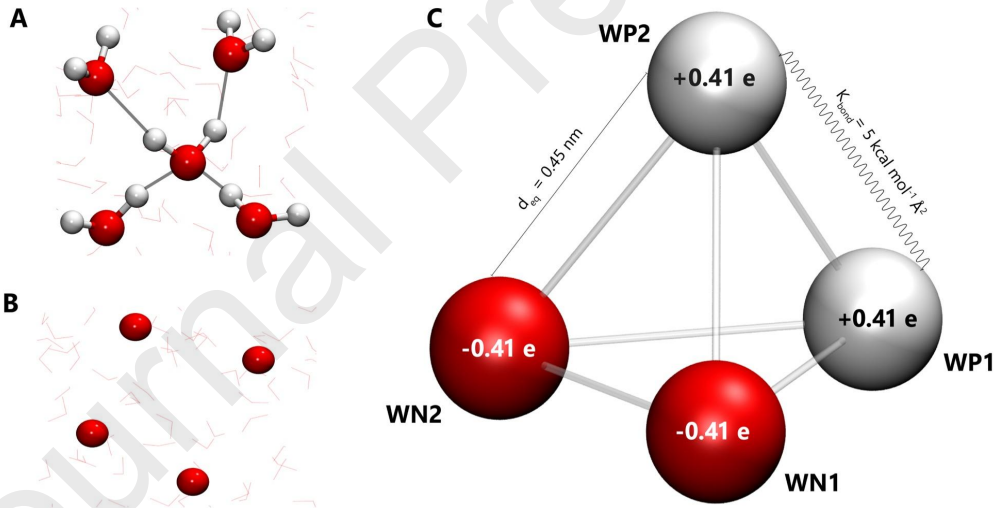


Figure 2: Conceptual derivation of the WT4 model. A) Water forms transient tetrahedral clusters held together by four Hydrogen bonds. The central water molecule in a cluster satisfies all Hydrogen bond acceptor and donor possibilities and is not accessible from the bulk. B) The WT4 model keeps the position of the four external Oxygen atoms in a cluster. C) The positions of those four Oxygen atoms are used to produce a covalently bound, tetrahedral CG water model.

Moreover, all Hydrogen atoms were removed, and only the Oxygen atoms at the tetrahedron's vertices were retained (Figure 2B), linking them with harmonic

bonds (Figure 2C). Since the geometry of the elementary water cluster is well determined by neutron scattering (Darré et al., 2010), the equilibrium distances could be set to the experimental value. Force constants were arbitrarily set to 5 kcal/mol, producing energies comparable to those associated with water-water Hydrogen bonds in fully atomistic force fields. This soft force constant resulted in a quite flexible tetrahedral structure that we called WatFour or WT4 for shortness (Darré et al., 2010).

For the model to generate its own dielectric permittivity and electrostatic screening, we added partial charges to the four beads, generating a quadrupole with two beads charged positively and two negatively. Since we wanted the model to be compatible with fully atomistic water models for multiscale simulations (see below), the numeric value of the partial charges was not adjusted to fit a target property. Instead, we adopted the partial charges of the Hydrogen atoms in the popular SPC water model (Berendsen et al., 1981). Namely, the beads in the WT4 model carry a partial charge of +/- 0.41 e.

The beads' size was set to reproduce the second solvation peak of water, while the deepness of the LJ energy well was iteratively fitted to match the experimental diffusion coefficient of pure water at 300 K.

Finally, the mass of the beads (50 a.u.) was adopted to match a density of 1kg/dl.

Considering that a water molecule is a tetrahedron with vertices at the Hydrogen atoms and Oxygen's lone pair electrons, the WT4 model can be conceived as a bulkier "water molecule". Indeed, the radial distribution function matches the second peak of the atomistic water. This feature prompted the development of monovalent ions that could modulate the ionic strength of the solution. The first version of electrolytic ions in SIRAH was represented by single beads with a net charge set to +/- 1 e. The size of ions was adjusted based on neutron scattering data to reflect the chemical identity of sodium, potassium, and chloride ions according to their second solvation shell (Darré et al., 2010). As ions are supposed to carry the first solvation shell, the depth of the LJ well was identical to that of the WT4 beads.

The availability of electrolytic ions is an asset to the model, as it is possible to set the ionic strength of the solution by modifying the amount of added salt in the simulation box. This feature, together with the permittivity created by the partial charges in WT4 and the calculation of long-range electrostatics using the Particle Mesh Ewald summation methods (Darden et al., 1993), makes the accurate description of electrostatics one of the salient features of SIRAH.

2.3 The CG protein model.

In analogy with DNA, the protein model was also derived using different bead sizes according to the nature of the interaction we intended to describe for each amino acid. The first version of the CG protein model was introduced in 2015 (Darré et al., 2015) and refined in 2019 (Machado et al., 2019a). We focus here only on the latest version, as even though the conceptual changes may appear minor, they have significantly enhanced the ability to reproduce protein structures.

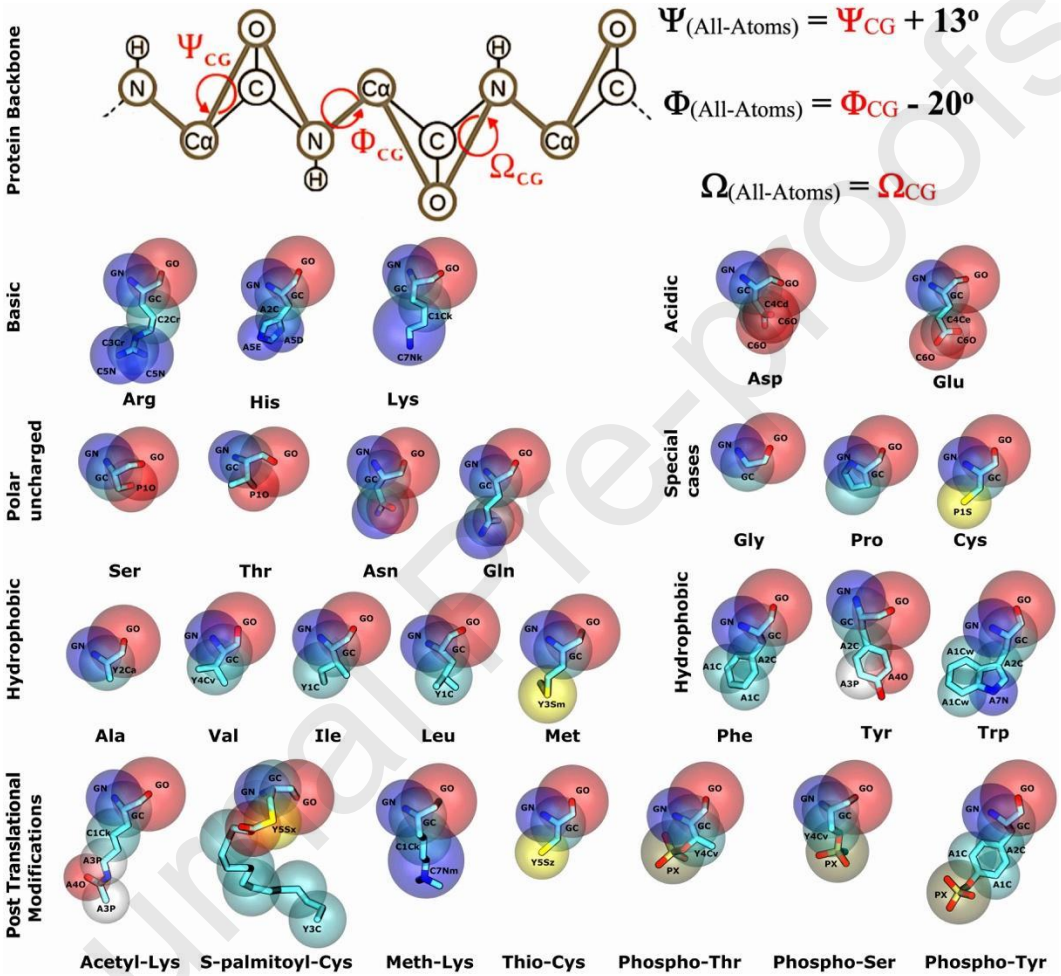


Figure 3: Mapping of the CG model for Amino Acids. Top: schematic mapping of the backbone beads (brown) on all-atoms (black) representation and mapping rule for Ramachandran dihedral angles. Bottom: CG amino acids and post-translational modifications in SIRAH. General scheme and coloring, as in Figure 1. Aspartic and Glutamic acids may exist in charged and neutral (protonated) states. Labels correspond to bead types indicated in Figure 4.

The atomistic to CG mapping of the protein's side chains follow the same philosophy of the DNA model. A comprehensive view of the CG representation of the amino acids is shown in Figure 3. Effective beads are placed on the position of selected atoms along the fully atomistic side chains with sizes and charges according to the interactions they are expected to establish (hydrophobic, aromatic,

salt bridges, etc.). Specifically, hydrophobic amino acids are represented by neutral beads at the Carbon in position β (in Valine) and γ (Leucine and Isoleucine), with an LJ diameter of 0.42nm, which roughly matches the excluded volume of the side chain.

Smaller bead sizes (0.35 nm) are used to represent aromatic amino acids (Figure 3), enabling the formation of stacking-like interactions among aromatic side chains. Partial charges are added to Tyrosine, Histidine, and Tryptophan to preserve their possibility of establishing Hydrogen bonds.

Polar amino acids in the CG representation are characterized by retaining beads in their respective functional groups, such as hydroxyl, thiol, amine, etc. Noteworthy, we kept beads at the location of Hydrogen atoms in Serine, Threonine, and Cysteine, along with partial charges that facilitate Hydrogen-bond-like interactions. Additionally, the acidic and basic amino acids possess partial charges, which add up to a net charge of +/- 1 e. It is important to note that in the cases of Cysteine, Aspartic, and Glutamic acids, negatively charged and neutral (or protonated) versions are available. Just like in fully atomistic force fields, changing the amino acid name in the initial PDB file is necessary to set their protonation state (Klein et al., 2020; M.R. Machado et al., 2019a).

The aminoacidic backbone is represented at a relatively high level of detail, with three beads placed on the Nitrogen, Cα Carbon, and carboxylic Oxygen positions. This allows a straightforward transformation from all-atoms to CG and backward, facilitating the interpretation of the conformations sampled by the CG proteins. Indeed, the so-called Ramachandran torsional angles can be obtained by a simple arithmetic rule (Figure 3).

Bonded parameters for amino acids were derived following the rules outlined for DNA, and in general, the same force constants for bond and angular stretching were used. This strategy of translating the same set of parameters, reminiscent of how early force fields were derived (Schuler et al., 2001), proved effective and time-saving. Therefore, we employed it as a general scheme to produce new topologies or, at least, to generate good initial guesses for bonded and non-bonded terms in the classical two-body Hamiltonian. For instance, phosphate parameters developed for DNA were successfully translated to phosphorylated residues in proteins and phospholipids. Worth mentioning, although the amino acid parameters were derived and fitted to reproduce the structural features of folded proteins, we recently showed that no modifications are needed to attain a good description of the conformational dynamics of intrinsically disordered proteins (He et al., 2023; Klein et al., 2021).

In version 2.0, the masses of all beads in the force field were set to 50 a.u.. Furthermore, the most common post-translational modifications in proteins are also available. They include phosphorylated Serine, Threonine, Tyrosine, Lysine in acetylated and methylated forms, and palmitoylated Cysteine (Garay et al., 2020).

Parameters for divalent ions are also available. Using a statistical analysis from the available metal-bonded structures reported in the PDB, we generated parameters for Zinc, Magnesium, and Calcium. The ions mentioned above account

for over 80% of the metal-coordinated structures. By using the statistical and structural data available, a refined set of interaction parameters from specific Lennard-Jones interactions was developed and optimized (Klein et al., 2020). These parameters were validated through multiple CG simulations of proteins and DNA systems, leading to a compelling structural reproduction, as demonstrated in Klein et al. (2020), and enabling SIRAH simulations of a wide range of metal-bound macromolecules.

2.4 The dark side of the force: overwriting combination rules.

While developing parameters for proteins, we realized that some interactions were well described in most cases but not properly weighted in others. For instance, electrolytic ions, including their first solvation shell, were well suited to neutralize DNA charges, but they remained permanently bound to charged side chains in amino acids.

To solve these issues, we introduced an essential change in how LJ interactions are calculated. In common MD packages, the potential energy contribution from LJ interactions is calculated according to $V(LJ)_{ij} = 4\varepsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/ r_{ij})^{6}]$, where ε_{ij} corresponds to the depth of the potential well and σ_{ij} indicates the distance r_{ij} at which the interaction between particles i and j reaches its energy minimum. The values for σ_{ij} and ε_{ij} are frequently calculated according to the Lorentz-Berthelot (LB) as $\varepsilon_{ij} = \sqrt{(\varepsilon_i \varepsilon_j)}$, and $\sigma_{ij}=(\sigma_i + \sigma_j)/2$.

In this way, interactions between individual atoms (or effective beads in our case) are assigned specific parameters according to their chemical nature. However, it is also possible to set specific LJ parameters, allowing certain bead pairs to establish specific interactions different from those calculated from LB combination rules. Modifying LJ interactions outside of LB combination rules offers an adaptable manner to regulate certain interactions that apply only to specific bead pairs.

This “outside-of-LB trick” provides a convenient way to fine-tune specific interactions between certain moieties so that each bead type can consistently account for the different exposed physicochemical environments. A similar strategy is also used in other CG force fields (e.g., MARTINI (Marrink and Tieleman, 2013; Marrink et al., 2023), and all-atoms force fields (e.g., CUFIX (Yoo and Aksimentiev, 2018)).

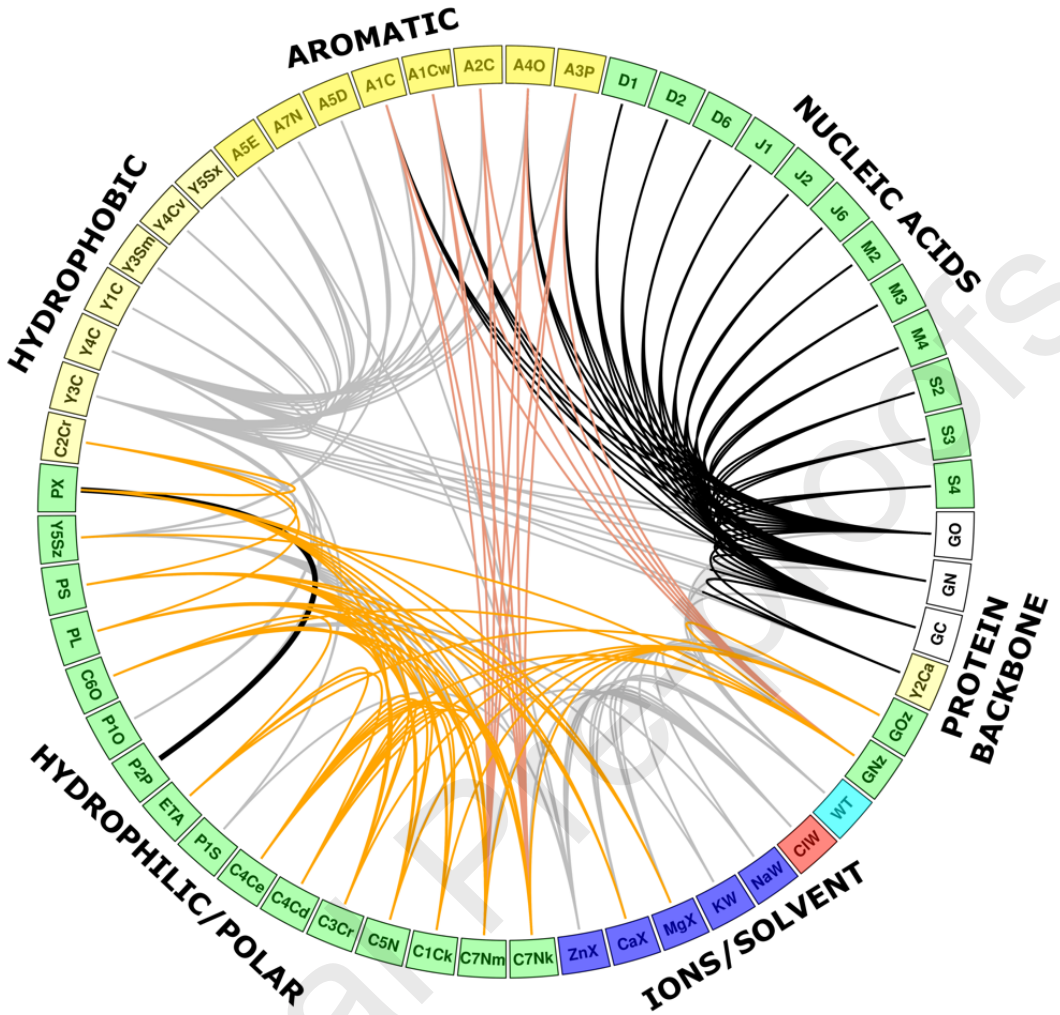


Figure 4: LJ interactions computed outside of Lorentz-Berthelot combination rules in SIRAH. Boxes at the circle rim identify bead types colored by physicochemical characteristics. Their position in different residues is indicated in Figures 1, 3, and 5.

Currently, there are 56 different beads types in SIRAH. Among a total of 1540 possible pair combinations, 197 interactions are defined outside of the LB combination rules. A pictorial representation of these interactions is provided in a chord plot format (Figure 4). To provide a brief explanation of the rationale behind these exceptions to the LB combination rules, we will shortly analyze a few of them.

Salmon lines in Figure 4 represent particular interactions between basic moieties and aromatic beads. They correspond to cation-π interactions between beads in the side chains of aromatic residues and Lysine (bead C7Nk), methylated Lysine (bead C7Nm), and zwitterionic N-terminal of all amino acids (bead GNz).

Black lines indicate which beads can interact with each other at nearly atomistic size while keeping a bigger size with the rest of the force field. This

correction is of utmost significance for backbone beads as the possibility of forming α helices depends crucially on the size of Cα carbons and the required distance to establish Hydrogen bonds between nitrogen and carboxylic Oxygen. Larger sizes than atomistic ones are incompatible with the compact structure of α helices (Maritan et al., 2000). A similar reasoning holds for the formation of Hydrogen bonds between two β strands. On the other hand, an atomistic size for the backbone beads results in over-stabilizing the interaction with certain beads, especially those of the WT4 model. Therefore, this workaround facilitates the formation of secondary structure elements, establishing at the same time adequately weighted interactions with the rest of the force field. Likewise, interactions between backbone and nucleotide beads allow for a better docking of peptides within the minor groove of double-stranded DNA (Brandner et al., 2018). The thick black line on the left of the chord plot connects the Phosphate moiety, present in DNA and phosphorylated amino acids, with the Proton-like bead (P2P) present in Serine, Threonine, and Cysteine, allowing for the formation of Hydrogen bond-like interactions.

Orange lines indicate special modifications that modulate salt bridges, avoiding an overstabilization among charged pairs.

Other out-of-LB interactions include cation-π, backbone-side chain Hbonds, and water-metal ions, among others. These are colored in gray and are not discussed here for brevity.

2.5 Fat SIRAH: CG models for phospholipids.

After completing the DNA, aqueous solvent, and protein models, we focused on incorporating an appropriate CG lipid representation to enable simulations of membrane proteins. Given the immense range of lipid species, we restricted our efforts to develop only a few prototypical phospholipids. To represent the diverse range of polar and nonpolar lipid constituents found in biological membranes, we examined three prototypical phospholipid heads: phosphatidyl-choline (PC), -ethanolamine (PE), and –serine (PS), which respectively represent big/small polar and acidic heads (see Figure 5). Additionally, we selected myristoyl (M), palmitic (P), and oleic (O) acyl chains as models of short, long, and unsaturated tails. Despite their minimalistic nature, the combinations of these lipid heads and tails enable the simulation of the most prevalent eukaryotic membrane components (Barrera et al., 2019).

To parameterize the lipid heads and tails, we capitalized on the significant number of functional groups present in the SIRAH force field, which not only reduced the workload associated with parameterization but also ensured compatibility. With only minimal parameter modifications, we successfully developed a set of phospholipids that accurately replicated the mechanical properties of lipid bilayers (Barrera et al., 2019; Capelli et al., 2021).

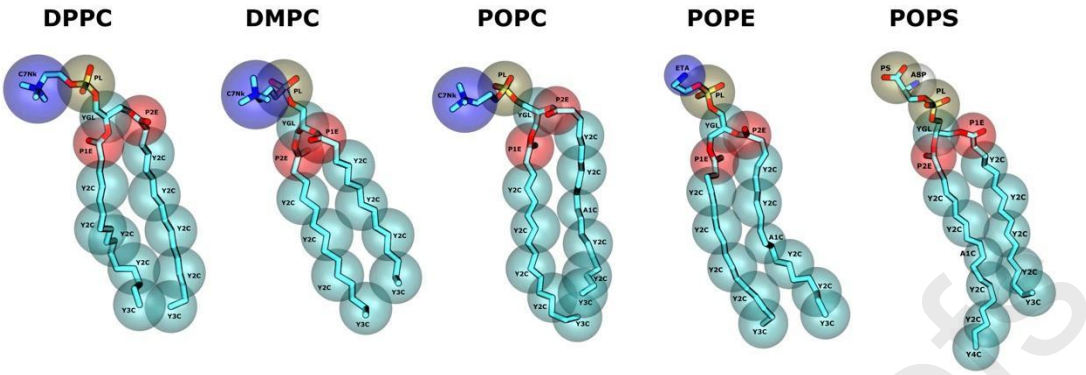


Figure 5: CG representation of phospholipids. General scheme and coloring, as in Figures 1 and 3. Labels correspond to bead types indicated in Figure 4.

In the course of test simulations of proteins embedded in a lipid bilayer, we observed some spurious insertions of acylic tails within the protein core. To address this problem, we set specific interactions between the hydrophobic side chains of the protein and the acyl chains out of the LB combination rules, rendering them equivalent to those found among Cα Carbon beads in Alanine. This modification led to an accurate depiction of the tilted orientation of the SarcoEndoplasmic Reticulum Calcium (SERCA) pump in a DMPC bilayer, which corresponds to the membrane thickness of the Endoplasmic Reticulum (Barrera et al., 2019). These parameters were further utilized to explore the electrostatics-driven opening of Connexin 26 channels, demonstrating impressive predictive power in identifying a mutation that inhibited channel opening (Zonta et al., 2018). Additionally, our group recently leveraged this cost-effective approach to perform simulations of entire viral capsids and envelopes, enabling us to construct and simulate a Zika VLP on the multi-microsecond time scale (Soñora et al., 2021).

3. Multiscale simulations

The development of the CG force field within the classical two-body Hamiltonian used by common MD simulation programs facilitated the implementation of multiscale simulations (recently reviewed by us in Soñora et al., 2021). The crosstalk between CG and atomistic regions occurs naturally within the same Hamiltonian, obviating the need to define non-Hamiltonian interaction terms. Besides the simplicity of the setup and running simulations, this also implies no loss of efficiency because of the time spent communicating between software modules. There are currently two different multiscale implementations in SIRAH. The first involves an all-atoms/CG model that covalently links both levels of resolution in one single, covalently bound nucleic acid chain (Machado et al., 2011). The combination of this all-atoms/CG model with the well-established QM/MM techniques implemented in AMBER (Walker et al., 2008) allowed for the first time to implement a QM/MM/CG scheme within the same simulation setup (Machado et al., 2019b).

The second multiscale approach within SIRAH involves a multi-resolution model for the solvent. In this approach, SIRAH can mix fully atomistic solutes with a shell of atomistic solvent surrounded by CG water (Darré et al., 2012b; Machado and Pantano, 2015). This approach is compatible with the most popular water solvent models (Gonzalez et al., 2013). Additionally, a triple solvation scheme in which water can be treated at all-atoms, CG, and supraCG levels is also available (Machado et al., 2017). These implementations are particularly well suited for highly solvated systems in which large amounts of bulk water are needed to ensure the proper solvation effects, pressure, etc. Concrete examples of such systems are viral capsids or envelopes. Because of intrinsic disorder or symmetry operations, the genetic material of viruses is difficult to resolve by experimental methods. Therefore, the reported structures of virus systems correspond actually to VLPs (Machado et al., 2017). The field of computational virology is entering a new era due to technological advances in computers and experimental techniques that enable the simulation and analysis of increasingly larger biological systems (Gonzalez-Arias et al., 2020; Jefferys and Sansom, 2019; Marzinek et al., 2020; Perilla et al., 2015). To further enhance current capabilities, multiscale strategies that combine atomic and CG resolutions are being utilized (Ayton and Voth, 2010; Liu et al., 2020; Machado et al., 2017; Yu et al., 2021). Nonetheless, constructing and configuring complex cellular systems still presents a computational obstacle. Therefore, even assembling such systems is a complex task. Recently, we developed a strategy to assemble and simulate VLP systems within the multiscale framework of the CG SIRAH force field (Soñora et al., 2021). To further reduce the computational cost of simulating VLPs, the solvent is represented at two levels of resolution using a CG water model (WT4) (Darré et al., 2010) and a supra-CG solvent (WLS) (Machado et al., 2017) to obtain an onion-shaped configuration of the system. MD simulations of entire VLPs are crucial to understanding their dynamics, providing insights only accessible to computer simulations (Jones et al., 2021; Machado and Pantano, 2021). We are currently using this approach to address the study of Flaviviruses containing the membrane and proteinaceous envelope (Soñora et al., 2022, 2021) (Figure 6).

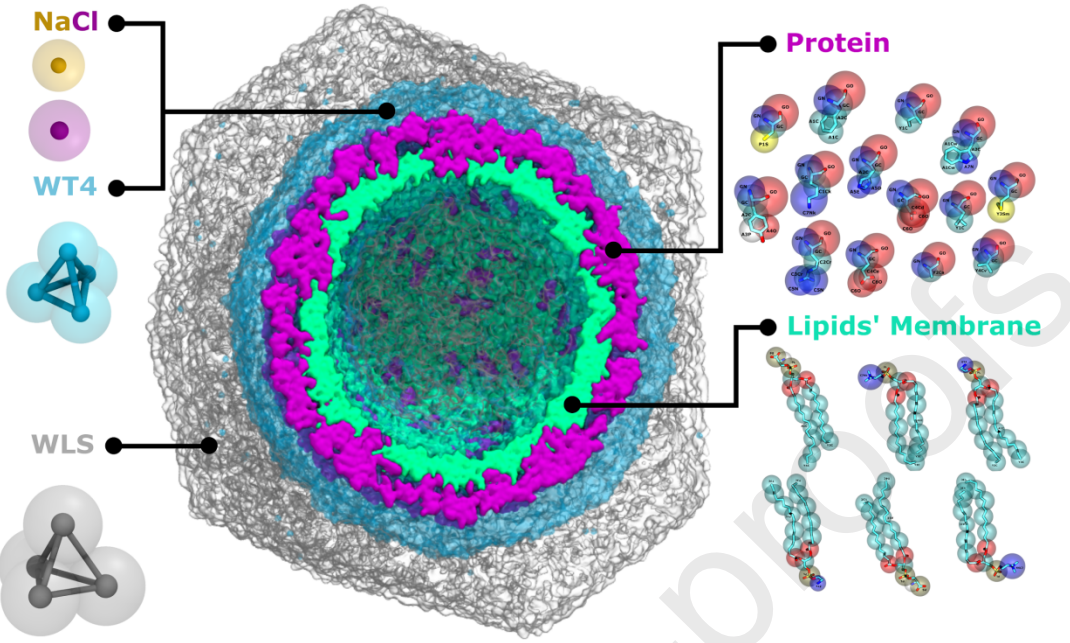


Figure 6. Simplified schematic representation of a multiscale model of the Zika virus VLP. From the outside to the inside, grey: WLS (supra CGsolvent), light blue: WT4 (CG solvent), violet: viral proteins (CG), and green: viral lipid membrane (CG). To the shake of a better visualization core of the VLP is shown empty. The complete system has a WT4 CG solvent layer internal to the lipid membrane and a supra-CG solvent layer (WLS) in the inner core (not shown for visual clarity).

4. Performance

As stated previously, the implementation of the latest version of the SIRAH force field profits from all available GROMACS and AMBER package options. The GPU implementations found in these packages facilitate CG simulations of medium-size systems at a rate of a few microseconds per day in desktop computers, while for larger-sized systems of about a million particles, the rate can reach hundreds of nanoseconds per day.

In order to provide a quantitative illustration of SIRAH performance, a comparison was made between a simulation of a system in SIRAH CG representation and the corresponding system in its atomistic model (Amber's FF14SB) with AMBER 20 using a node computer equipped with two Intel® Xeon® Gold 5317 processors (3.0 GHz x 24 cores), 252 GB RAM DDR4 memory (3,200 MHz) and accelerated with an NVIDIA GeForce RTX 3090 (24 Gb) GPU. The selected system for the study comprised the receptor binding domain (RBD) of the SARS-CoV-2 Spike protein, along with the complete human ACE2 and in the presence of the neutral amino acid transporter B0AT1 (PDB code 6M17). This system was previously simulated using SIRAH and subsequently analyzed in the work of Garay et al. (2021). Keeping similar dimensions, the use of an atomistic representation instead of a CG one resulted in a ten-fold increase in the number of

elements, with the former consisting of 110,910 beads and the latter consisting of 1,162,885 atoms, including water, electrolytes, and metal ions bound to the catalytic site. The SIRAH CG model exhibited a simulation speed of approximately 660 nanoseconds per day using a 20 fs time step. In comparison, the atomistic model showed a simulation speed of approximately 11 nanoseconds per day using a 2 fs time step. Thus, in this architecture and utilizing AMBER 20, we obtained a 60-fold speedup when using SIRAH CG representation instead of an atomistic one for the same system.

Nevertheless, it has to be kept in mind that a direct comparison between CG force fields might not always be possible, as not all share the same capabilities. Beyond simple speedup considerations, users must consider the force field's limitations.

5. Limitations

It is crucial to underline that while CG methods can enhance performance, they are not a panacea for faster results without any trade-offs. Any shortcuts taken usually come with a cost, which in this instance, could compromise structural or energetic precision or both.

Similar to other CG force fields, SIRAH has limitations in handling situations that demand atomic-level precision. These could include scenarios where single water molecules mediate interactions within macromolecules or when ligands possess a high degree of specificity within binding sites. For example, simulations of a potassium channel, which is stabilized by alternating desolvated potassium and single water molecules within the channel cavity (Díaz-Franulic et al., 2015), or aquaporins, in which individual water molecules modify their orientation while traversing the water pore (Canessa Fortuna et al., 2019). The atomistic details revealed by these two latter examples could be challenging to accomplish using CG techniques since CG models that combine multiple water molecules into a single effective bead are unable to investigate intermolecular interactions involving single water molecules. Moreover, the inherent loss of accuracy in CG components can negatively impact the faithful reproduction of native contacts in protein-protein interfaces, even when intermolecular contact surface values remain within experimental bounds (Darré et al., 2015; Machado et al., 2019a).

Despite the unbiased treatment of the backbone conformations, the use of SIRAH for protein folding simulations has not been deeply explored. Although we succeed in reproducing the spontaneous aggregation (Barrera et al., 2021a, 2021b), folding of small peptides (Klein et al., 2021), and the specific recognition of Calmodulin upon the presence of a Calmodulin binding peptide (Machado et al., 2019a), the unbiased formation of large helical segments is still challenging to our force field.

Despite incorporating parameters for the most prevalent biological molecules, the molecular diversity within biological systems is so extensive that it is virtually impossible to encompass all relevant biomolecules. In our case, it is not immediate to establish a generally valid methodology for creating arbitrary molecular topologies. As elaborated in the preceding paragraphs, converting a new topology from all-atom to CG represents a critical stage. In the case of SIRAH, this process necessitates data gleaned from experimental databases, organic chemistry (canonical structures, for instance), and physicochemical intuition.

Unlike in the fully atomistic realm, where chemical principles are well established, the CG universe can be somewhat fuzzy. Indeed, diverse topologies for the same molecule may emerge from different academic backgrounds. Hence, there is no definitive "correct" or "incorrect" CG parameterization. A "good" parameterization is one that yields an accurate portrayal of the intermolecular interactions being studied, keeps compatibility with the rest of the force field, and grants a significant acceleration in calculations (at least one or two orders of magnitude).

6. Perspectives

The development of rapid and accurate simulation potentials will continue to gain relevance in biomedical sciences, as the ever-growing computer power made of MD a well-established and complementary technique to further our understanding of complex processes and large biological systems (Stevens et al., 2023). In this context, consistent force fields, including all biological families of molecules enabling the crosstalk at different molecular resolutions, remain an important challenge.

Our perspectives for the near future include extending the SIRAH´s universe to include glycans, which will soon be available to simulate polysaccharide chains and protein glycosylation. We are also working to increase the lipid diversity. We will incorporate sphingomyelins, ceramides, and cholesterol. These molecules play an essential role in the endoplasmic reticulum membranes, which are also components of flaviviral envelopes, one of our main lines of research. Additionally, we are testing POPG parameters, as this lipid constitutes a main component of bacterial membranes; it is fundamental for realistic descriptions of the mode of action of antibiotic peptides.

In the medium term, we plan to incorporate a CG model for RNA, which are fundamental ingredients for describing viral particles and constitute a mainstream of research in our group.

ACKNOWLEDGEMENTS:

We wish to acknowledge all past and present members of the Biomolecular Simulations group at the Institut Pasteur de Montevideo and the people, especially within the South American biocomputing community, who collaborated on the development and maintenance of this initiative through direct involvement or very constructive exchange of ideas.

FUNDING:

This work was partially funded by FOCEM (MERCOSUR Structural Convergence Fund), COF 03/11. SP and MRM are members of the Uruguayan SNI. FK and MS are Ph.D. fellows of ANII, CAP, and PEDECIBA. ENF is a postdoc fellow of the Argentinean CONICET, and his participation in this manuscript was possible thanks to a scholarship from CEBEM and CZI Foundation (2021-240160(5022)). SP acknowledges support from the ICTP through the Associates Programme (2019-2024)

REFERENCES:

Ayton, G.S., Voth, G.A., 2010. Multiscale Computer Simulation of the Immature HIV-1 Virion. Biophys J 99, 2757–2765. https://doi.org/10.1016/J.BPJ.2010.08.018

Bahar, I., Jernigan, R.L., 1997. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 266, 195–214. https://doi.org/10.1006/JMBI.1996.0758

Barrera, E.E., Machado, M.R., Pantano, S., 2019. Fat SIRAH: Coarse-Grained Phospholipids to Explore Membrane-Protein Dynamics. J Chem Theory Comput 15, 5674–5688. https://doi.org/10.1021/ACS.JCTC.9B00435/SUPPL_FILE/CT9B00435_SI_002.MP4

Barrera, E.E., Pantano, S., Zonta, F., 2021a. A homogeneous dataset of polyglutamine and glutamine rich aggregating peptides simulations. Data Brief 36, 107109. https://doi.org/10.1016/J.DIB.2021.107109

Barrera, E.E., Zonta, F., Pantano, S., 2021b. Dissecting the role of glutamine in seeding peptide aggregation. Comput Struct Biotechnol J 19, 1595–1602. https://doi.org/10.1016/J.CSBJ.2021.02.014

Bayly, C.I., Merz, K.M., Ferguson, D.M., Cornell, W.D., Fox, T., Caldwell, J.W., Kollman, P.A., Cieplak, P., Gould, I.R., Spellmeyer, D.C., 2002. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem Soc 117, 5179–5197. https://doi.org/10.1021/JA00124A002

Berendsen, H.J.C., Postma, J.P.M., Gunsteren, W.F. van, Hermans, J., 1981. Interaction Models for Water in Relation to Protein Hydration 331–342. https://doi.org/10.1007/978-94-015-7658-1_21

Brandner, A., Schüller, A., Melo, F., Pantano, S., 2018. Exploring DNA dynamics within oligonucleosomes with coarse-grained simulations: SIRAH force field extension for protein-DNA complexes. Biochem Biophys Res Commun 498, 319–326. https://doi.org/10.1016/J.BBRC.2017.09.086

Canessa Fortuna, A., Zerbetto De Palma, G., Aliperti Car, L., Armentia, L., Vitali, V., Zeida, A., Estrin, D.A., Alleva, K., 2019. Gating in plant plasma membrane aquaporins: the involvement of leucine in the formation of a pore constriction in the closed state. FEBS J 286, 3473–3487. https://doi.org/10.1111/FEBS.14922

Capelli, R., Gardin, A., Empereur-Mot, C., Doni, G., Pavan, G.M., 2021. A Data-Driven Dimensionality Reduction Approach to Compare and Classify Lipid Force Fields. Journal of Physical Chemistry B 125, 7785–7796. https://doi.org/10.1021/ACS.JPCB.1C02503/SUPPL_FILE/JP1C02503_SI_001.PDF

Case, D., Darden, T., Cheatham, T., Simmerling, C., Wang, J., Duke, R., Luo, R., Crowley, M., Walker, R.C., Zhang, W., Merz, K., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K., Paesani, F., Vanícek, J., Kollman, P., 2008. AMBER 10, University of California, San Francisco.

Case, D.A., Belfon, K., Ben-Shalom, I., Brozell, S.R., Cerutti, D., Cheatha, T., Cruzeiro, V.W.D., Darden, Tom ; Duke, R.E., Giambasu, G., Gilson, M., Gohlke, H., Götz, A., Harris, R., Izadi, S., Izmailov, S.A., Kasavajhala, K., Kovalenko, A., Krasny, R., Kurtzman, T., Lee, T., LeGrand, S., Li, P., Lin, C., Liu, J., Luchko, T., Luo, R., Man, V., Merz, K.M., Miao, Y., Mikhailovskii, , O., Monard, G., Nguyen, H., Onufriev, A., Pan, F., Pantano, S., Qi, R., Roe, D.R., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shen, J., Simmerling, C., Skrynnikov, N.R., Smith, J., Swails, J., Walker, R., Wang, J., Wilson, L., Wolf, R.M., Wu, X., Xiong, Y., Xue, Y., York, D., Kollman, P.A., 2020. Amber 2020. University of California Press,.

Clementi, C., Nymeyer, H., Onuchic, J.N., 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. https://doi.org/10.1006/JMBI.2000.3693

Dai, L., Yu, J., 2020. Inchworm stepping of Myc-Max heterodimer protein diffusion along DNA. Biochem Biophys Res Commun 533, 97–103. https://doi.org/10.1016/J.BBRC.2020.08.004

Dama, J.F., Sinitskiy, A. V., McCullagh, M., Weare, J., Roux, B., Dinner, A.R., Voth, G.A., 2013. The Theory of Ultra-Coarse-Graining. 1. General Principles. J Chem Theory Comput 9, 2466–2480. https://doi.org/10.1021/CT4000444

Dannenhoffer-Lafage, T., White, A.D., Voth, G.A., 2016. A Direct Method for Incorporating Experimental Data into Multiscale Coarse-Grained Models. J Chem Theory Comput 12, 2144–2153. https://doi.org/10.1021/ACS.JCTC.6B00043

Dans, P.D., Zeida, A., Machado, M.R., Pantano, S., 2010. A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. J Chem Theory Comput 6, 1711–1725.

Dans, P.D., Darré, L., Machado, M.R., Zeida, A., Brandner, A.F., Pantano, S., 2013. Assessing the Accuracy of the SIRAH Force Field to Model DNA at Coarse Grain Level. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8213 LNBI, 71–81. https://doi.org/10.1007/978-3-319-02624-4_7 https://doi.org/10.1021/CT900653P/SUPPL_FILE/CT900653P_SI_002.AVI

Darden, T., York, D., Pedersen, L., 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J Chem Phys 98, 10089–10092. https://doi.org/10.1063/1.464397

Darré, L., Machado, M.R., Dans, P.D., Herrera, F.E., Pantano, S., 2010. Another coarse grain model for aqueous solvation: WAT FOUR? J Chem Theory Comput 6, 3793–3807. https://doi.org/10.1021/CT100379F/SUPPL_FILE/CT100379F_SI_001.PDF

Darré, L., Machado, M.R., Pantano, S., 2012a. Coarse-grained models of water. Wiley Interdiscip Rev Comput Mol Sci 2, 921–930. https://doi.org/10.1002/WCMS.1097

Darré, L., Tek, A., Baaden, M., Pantano, S., 2012b. Mixing Atomistic and Coarse Grain Solvation Models for MD Simulations: Let WT4 Handle the Bulk. J Chem Theory Comput 8, 3880–3894. https://doi.org/10.1021/CT3001816

Darré, L., Machado, M.R., Brandner, A.F., González, H.C., Ferreira, S., Pantano, S., 2015. SIRAH: A Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics. J Chem Theory Comput 11, 723–739. https://doi.org/10.1021/CT5007746

Davtyan, A., Schafer, N.P., Zheng, W., Clementi, C., Wolynes, P.G., Papoian, G.A., 2012. AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. Journal of Physical Chemistry B 116, 8494–8503. https://doi.org/10.1021/JP212541Y/SUPPL_FILE/JP212541Y_SI_001.PDF

Davtyan, A., Voth, G.A., Andersen, H.C., 2016. Dynamic force matching: Construction of dynamic coarse-grained models with realistic short time dynamics and accurate long time dynamics. J Chem Phys 145, 224107. https://doi.org/10.1063/1.4971430

Derreumaux, P., Mousseau, N., 2007. Coarse-grained protein molecular dynamics simulations. J Chem Phys 126, 025101. https://doi.org/10.1063/1.2408414

Díaz-Franulic, I., Sepúlveda, R. V., Navarro-Quezada, N., González-Nilo, F., Naranjo, D., 2015. Pore dimensions and the role of occupancy in unitary conductance of Shaker K channels. Journal of General Physiology 146, 133–146. https://doi.org/10.1085/JGP.201411353

Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H., Shaw, D.E., 2012. Biomolecular Simulation: A Computational Microscope for Molecular Biology. http://dx.doi.org/10.1146/annurev-biophys-042910-155245 41, 429–452. https://doi.org/10.1146/ANNUREV-BIOPHYS-042910-155245

Garay, P.G., Barrera, E.E., Pantano, S., 2020. Post-Translational Modifications at the Coarse-Grained Level with the SIRAH Force Field. J Chem Inf Model 60, 964–973. https://doi.org/10.1021/ACS.JCIM.9B00900/SUPPL_FILE/CI9B00900_SI_008.PDF

Gonzalez, H.C., Darré, L., Pantano, S., 2013. Transferable mixing of atomistic and coarse-grained water models. J Phys Chem B 117, 14438–14448. https://doi.org/10.1021/JP4079579

Gonzalez-Arias, F., Reddy, T., Stone, J.E., Hadden-Perilla, J.A., Perilla, J.R., 2020. Scalable Analysis of Authentic Viral Envelopes on FRONTERA. Comput Sci Eng 22, 11–20. https://doi.org/10.1109/MCSE.2020.3020508

Hadden, J.A., Perilla, J.R., 2018. All-atom virus simulations. Curr Opin Virol 31, 82–91. https://doi.org/10.1016/J.COVIRO.2018.08.007

He, X., Man, V.H., Gao, J., Wang, J., 2023. Investigation of the Structure of Full-Length Tau Proteins with Coarse-Grained and All-Atom Molecular Dynamics Simulations. ACS Chem Neurosci 14, 209–217. https://doi.org/10.1021/ACSCHEMNEURO.2C00381/SUPPL_FILE/CN2C00381_SI_003.PDB

Hinckley, D.M., Freeman, G.S., Whitmer, J.K., De Pablo, J.J., 2013. An experimentally-informed coarse-grained 3-Site-Per-Nucleotide model of DNA: structure, thermodynamics, and dynamics of hybridization. J Chem Phys 139. https://doi.org/10.1063/1.4822042

Hopkins, C.W., Le Grand, S., Walker, R.C., Roitberg, A.E., 2015. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. J Chem Theory Comput 11, 1864–1874. https://doi.org/10.1021/CT5010406

Ingólfsson, H.I., Lopez, C.A., Uusitalo, J.J., de Jong, D.H., Gopal, S.M., Periole, X., Marrink, S.J., 2014. The power of coarse graining in biomolecular simulations. Wiley Interdiscip Rev Comput Mol Sci. https://doi.org/10.1002/wcms.1169

Jefferys, E.E., Sansom, M.S.P., 2019. Computational Virology: Molecular Simulations of Virus Dynamics and Interactions. Adv Exp Med Biol 1140, 201–233. https://doi.org/10.1007/978-3-030-14741-9_10/COVER

Jones, P.E., Pérez-Segura, C., Bryer, A.J., Perilla, J.R., Hadden-Perilla, J.A., 2021. Molecular dynamics of the viral life cycle: progress and prospects. Curr Opin Virol 50, 128–138. https://doi.org/10.1016/J.COVIRO.2021.08.003

Kar, P., Gopal, S.M., Cheng, Y.M., Predeus, A., Feig, M., 2013. PRIMO: A Transferable Coarse-grained Force Field for Proteins. J Chem Theory Comput 9, 3769. https://doi.org/10.1021/CT400230Y

Kenzaki, H., Koga, N., Hori, N., Kanada, R., Li, W., Okazaki, K.I., Yao, X.Q., Takada, S., 2011. CafeMol: A Coarse-Grained Biomolecular Simulator for Simulating Proteins at Work. J Chem Theory Comput 7, 1979–1989. https://doi.org/10.1021/CT2001045

Klein, F., Cáceres, D., Carrasco, M.A., Tapia, J.C., Caballero, J., Alzate-Morales, J., Pantano, S., 2020. Coarse-Grained Parameters for Divalent Cations within the SIRAH Force Field. J Chem Inf Model 60, 3935–3943. https://doi.org/10.1021/acs.jcim.0c00160

Klein, F., Barrera, E.E., Pantano, S., 2021. Assessing SIRAH's Capability to Simulate Intrinsically Disordered Proteins and Peptides. J Chem Theory Comput 17, 599–604. https://doi.org/10.1021/ACS.JCTC.0C00948

Lee, E.H., Hsin, J., Sotomayor, M., Comellas, G., Schulten, K., 2009. Discovery through the computational microscope. Structure 17, 1295–1306. https://doi.org/10.1016/J.STR.2009.09.001

Levitt, M., Warshel, A., 1975. Computer simulation of protein folding. Nature 1975 253:5494 253, 694–698. https://doi.org/10.1038/253694a0

Liu, Y., De Vries, A.H., Barnoud, J., Pezeshkian, W., Melcr, J., Marrink, S.J., 2020. Dual Resolution Membrane Simulations Using Virtual Sites. Journal of Physical Chemistry B 124, 3944–3953. https://doi.org/10.1021/ACS.JPCB.0C01842/ASSET/IMAGES/LARGE/JP0C01842_0004.JPEG

Machado, M.R., Dans, P.D., Pantano, S., 2011. A hybrid all-atom/coarse grain model for multiscale simulations of DNA. Physical Chemistry Chemical Physics 13, 18134–18144. https://doi.org/10.1039/C1CP21248F

Machado, M.R., Pantano, S., 2015. Exploring LacI-DNA dynamics by multiscale simulations using the SIRAH force field. J Chem Theory Comput 11, 5012–5023. https://doi.org/10.1021/ACS.JCTC.5B00575

Machado, M.R., Pantano, S., 2016. SIRAH tools: Mapping, backmapping and visualization of coarse-grained models. Bioinformatics 32, 1568–1570. https://doi.org/10.1093/bioinformatics/btw020

Machado, M.R., González, H.C., Pantano, S., 2017. MD Simulations of Virus-like Particles with Supra CG Solvation Affordable to Desktop Computers. J Chem Theory Comput 13, 5106–5116. https://doi.org/10.1021/ACS.JCTC.7B00659

Machado, M.R., Barrera, E.E., Klein, F., Sónora, M., Silva, S., Pantano, S., 2019a. The SIRAH 2.0 Force Field: Altius, Fortius, Citius. J Chem Theory Comput 15. https://doi.org/10.1021/acs.jctc.9b00006

Machado, Matías R., Zeida, A., Darré, L., Pantano, S., 2019b. From quantum to subcellular scales: multiscale simulation approaches and the SIRAH force field. Interface Focus 9. https://doi.org/10.1098/RSFS.2018.0085

Machado, M.R., Pantano, S., 2021. Fighting viruses with computers, right now. Curr Opin Virol 48, 91–99. https://doi.org/10.1016/J.COVIRO.2021.04.004

Marchetto, A., Chaib, Z.S., Rossi, C.A., Ribeiro, R., Pantano, S., Rossetti, G., Giorgetti, A., 2020. CGMD Platform: Integrated Web Servers for the Preparation, Running, and Analysis of Coarse-Grained Molecular Dynamics Simulations. Molecules 2020, Vol. 25, Page 5934 25, 5934. https://doi.org/10.3390/MOLECULES25245934

Maritan, A., Micheletti, C., Trovato, A., Banavar, J.R., 2000. Optimal shapes of compact strings. Nature 406, 287–290. https://doi.org/10.1038/35018538

Marrink, S.J., Mark, A.E., 2003. The Mechanism of Vesicle Fusion as Revealed by Molecular Dynamics Simulations. J Am Chem Soc 125, 11144–11145. https://doi.org/10.1021/ja036138+

Marrink, S.J., Tieleman, D.P., 2013. Perspective on the martini model. Chem Soc Rev 42, 6801–6822. https://doi.org/10.1039/c3cs60093a

Marrink, S.J., Monticelli, L., Melo, M.N., Alessandri, R., Tieleman, D.P., Souza, P.C.T., 2023. Two decades of Martini: Better beads, broader scope. Wiley Interdiscip Rev Comput Mol Sci 13, e1620. https://doi.org/10.1002/WCMS.1620

Marzinek, J.K., Huber, R.G., Bond, P.J., 2020. Multiscale modelling and simulation of viruses. Curr Opin Struct Biol 61, 146–152. https://doi.org/10.1016/J.SBI.2019.12.019

Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G., Tieleman, D.P., Marrink, S.J., 2008. The MARTINI Coarse-Grained Force Field: Extension to Proteins. J Chem Theory Comput 4, 819–834. https://doi.org/10.1021/CT700324X

Noid, W.G., 2013. Perspective: Coarse-grained models for biomolecular systems. J Chem Phys 139, 090901. https://doi.org/10.1063/1.4818908

Onuchic, J.N., Wolynes, P.G., 2004. Theory of protein folding. Curr Opin Struct Biol 14, 70–75. https://doi.org/10.1016/J.SBI.2004.01.009

Orsi, M., Essex, J.W., 2011. The ELBA Force Field for Coarse-Grain Modeling of Lipid Membranes. PLoS One 6, e28637. https://doi.org/10.1371/JOURNAL.PONE.0028637

Pasi, M., Lavery, R., Ceres, N., 2013. PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties. J Chem Theory Comput 9, 785–793. https://doi.org/10.1021/CT3007925

Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T.E., Laughton, C.A., Orozco, M., 2007. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J 92, 3817–3829. https://doi.org/10.1529/BIOPHYSJ.106.097782

Perilla, J.R., Goh, B.C., Cassidy, C.K., Liu, B., Bernardi, R.C., Rudack, T., Yu, H., Wu, Z., Schulten, K., 2015. Molecular dynamics simulations of large macromolecular complexes. Curr Opin Struct Biol 31, 64–74. https://doi.org/10.1016/J.SBI.2015.03.007

Reith, D., Pütz, M., Müller-Plathe, F., 2003. Deriving effective mesoscale potentials from atomistic simulations. J Comput Chem 24, 1624–1636. https://doi.org/10.1002/JCC.10307

Scheraga, H.A., Khalili, M., Liwo, A., 2007. Protein-Folding Dynamics: Overview of Molecular Simulation Techniques 58, 57–83. https://doi.org/10.1146/ANNUREV.PHYSCHEM.58.032806.104614

Schuler, L.D., Daura, X., Van Gunsteren, W.F., 2001. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. J Comput Chem 22, 1205–1218. https://doi.org/10.1002/JCC.1078

Seo, S., Shinoda, W., 2019. SPICA Force Field for Lipid Membranes: Domain Formation Induced by Cholesterol. J Chem Theory Comput 15, 762–774. https://doi.org/10.1021/ACS.JCTC.8B00987/SUPPL_FILE/CT8B00987_SI_001.PDF

Shinoda, W., Devane, R., Klein, M.L., 2011. Coarse-grained force field for ionic surfactants. Soft Matter 7, 6178–6186. https://doi.org/10.1039/C1SM05173C

Singh, N., Li, W., 2019. Recent Advances in Coarse-Grained Models for Biomolecules and Their Applications. Int J Mol Sci 20. https://doi.org/10.3390/IJMS20153774

Soñora, M., Martínez, L., Pantano, S., Machado, M.R., 2021. Wrapping Up Viruses at Multiscale Resolution: Optimizing PACKMOL and SIRAH Execution for Simulating the Zika Virus. J Chem Inf Model 61, 408–422. https://doi.org/10.1021/ACS.JCIM.0C01205

Soñora, M., Barrera, E.E., Pantano, S., 2022. The stressed life of a lipid in the Zika virus membrane. Biochim Biophys Acta Biomembr 1864. https://doi.org/10.1016/J.BBAMEM.2021.183804

Sterpone, F., Melchionna, S., Tuffery, P., Pasquali, S., Mousseau, N., Cragnolini, T., Chebaro, Y., St-Pierre, J.F., Kalimeri, M., Barducci, A., Laurin, Y., Tek, A., Baaden, M., Nguyen, P.H., Derreumaux, P., 2014. The OPEP protein model: From single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. Chem Soc Rev. https://doi.org/10.1039/c4cs00048j

Stevens, J.A., Grünewald, F., van Tilburg, P.A.M., König, M., Gilbert, B.R., Brier, T.A., Thornburg, Z.R., Luthey-Schulten, Z., Marrink, S.J., 2023. Molecular dynamics simulation of an entire cell. Front Chem 11, 24. https://doi.org/10.3389/FCHEM.2023.1106495/BIBTEX

Sztain, T., Ahn, S.H., Bogetti, A.T., Casalino, L., Goldsmith, J.A., Seitz, E., McCool, R.S., Kearns, F.L., Acosta-Reyes, F., Maji, S., Mashayekhi, G., McCammon, J.A., Ourmazd, A., Frank, J., McLellan, J.S., Chong, L.T., Amaro, R.E., 2021. A glycan gate controls opening of the SARS-CoV-2 spike protein. Nature Chemistry 2021 13:10 13, 963–968. https://doi.org/10.1038/s41557-021-00758-3

Walker, R.C., Crowley, I.F., Case, D.A., 2008. The implementation of a fast and accurate QM/MM potential method in Amber. J Comput Chem 29, 1019–1031. https://doi.org/10.1002/JCC.20857

Yoo, J., Aksimentiev, A., 2018. New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. Physical Chemistry Chemical Physics 20, 8432–8449. https://doi.org/10.1039/C7CP08185E

Yu, A., Pak, A.J., He, P., Monje-Galvan, V., Casalino, L., Gaieb, Z., Dommer, A.C., Amaro, R.E., Voth, G.A., 2021. A multiscale coarse-grained model of the SARS-CoV-2 virion. Biophys J 120, 1097–1104. https://doi.org/10.1016/J.BPJ.2020.10.048

Zeida, A., MacHado, M.R., Dans, P.D., Pantano, S., 2012. Breathing, bubbling, and bending: DNA flexibility from multimicrosecond simulations. Phys Rev E Stat Nonlin Soft Matter Phys 86, 021903. https://doi.org/10.1103/PHYSREVE.86.021903/FIGURES/4/MEDIUM

Zonta, F., Buratto, D., Crispino, G., Carrer, A., Bruno, F., Yang, G., Mammano, F., Pantano, S., 2018. Cues to opening mechanisms from in silico electric field excitation of cx26 hemichannel and in vitro mutagenesis studies in HeLa transfectans. Front Mol Neurosci 11, 170. https://doi.org/10.3389/FNMOL.2018.00170/BIBTEX

- SIRAH is one of the most complete force fields for coarse-grained and multiscale simulations of complex biological systems.

- Its implementation in popular molecular dynamics simulation packages provides plug&play solution for coarse-grained simulations.

- This review provides a comprehensive outlook of the development strategies and illustrate the future of the force field.

Manuscript Number: JSB-23-34
The SIRAH force field: a suite for simulations of complex biological systems at the coarse-grained and multiscale levels

Credit Author Statement

FK, MS, and SP drafted the initial version of the text and figures. All authors participated in writing, discussing and correcting the original and revised version of the manuscript.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Graphical Abstract