Multiscale Molecular Dynamics of Protein Aggregation

César L. Ávila[1,#], Nils J. D. Drechsel[2,#], Raúl Alcántara[2] and Jordi Villà–Freixa[2,*]

[1]*Departamento Bioquímica de la Nutrición, Instituto Superior de Investigaciones Biológicas (CONICET-UNT), Chacabuco 461 (4000), Tucumán, Argentina;* [2]*Computational Biochemistry and Biophysics Laboratory, Research Unit on Biomedical Informatics, IMIM Hospital del Mar-Universitat and Universitat Pompeu Fabra, C/Doctor Aiguader, 88, (08003) Barcelona, Catalunya, Spain*

Abstract: The 60's gave birth to the practical implementation of classical mechanics to unravel the dynamics and energetics of biomolecules. In the 70's the use of generalized force fields and more advanced integrative solutions to the microscopic understanding of nature (like hybrid QM/MM) were introduced. During the 80's, algorithms to obtain free energy values were further developed and in the 90's practical integration schemes of molecular mechanics force fields with other levels of detail (QM on one extreme and advances in implicit solvation on the other) were implemented in widely spread software. In the first decade of the XXIst century a considerable effort has been put in two seemingly discordant models for the simulation of biomolecules. On the one hand, extraordinary advances in computing technologies (both in terms of processor power and of new efficient parallel and distributed computing schemas) have allowed researchers to deal with bigger systems and longer simulations, reaching molecular processes including millions of particles or lying in the milisecond scale. On the other hand, the realization that the relevant answers to many biomolecular problems are not homogeneously distributed through the molecular structure, something already envisioned by the QM/MM pioneers more than three decades ago, has led researchers to find smart ways of putting different emphases on different ranges of the spatial or system time scale. In this context, e.g., molecular aggregation represents a paradigm for multiscalability, as molecular recognition can be understood with simple (semi-)macroscopic electrostatic terms when the two fragments are far apart, while the atomic interactions need to be considered in full detail upon close distances. In this manuscript the current status of the techniques that use multiple scale representations of biomolecules are reviewed, and the findings are synthesized in a modular schema that can be extensively used when studying aggregation processes. It is shown that a smart alternative to brute force and massive computation of uninteresting regions in the all atom potential energy surface is the consideration of a simplified reference potential, explored thoroughly in the relevant regions, combined with a free energy perturbation approach that transforms this simple representation to a full atom representation.

Keywords: Multiscale stimulations, free energy perturbations, coarse grain.

1. INTRODUCTION

The last decade has framed yet another step on Moore's law to produce more powerful computer processors. The arrival of the Cell processor was quickly surpassed by the advent of the new GPU programming paradigm [1-2]. In addition, new schemas for distributing data and calculations are also available with low-cost implementation [3-5]. Molecular simulations, one of the leading disciplines in terms of requested computer power, could not be alien to such a revolution. Faced with big supercomputer centers with restrictive (and often capricious) mechanisms to evaluate requests for computing time, researchers have discovered they are able to run previously unimaginable computations on a desktop PC equipped with several graphical cards.

As usual, the availability of new resources has boosted scientific initiatives, and the community has moved into a gold rush of increasingly larger examples to apply their molecular simulation algorithms. Such rush, however, leads sometimes to confusing scientific questions with technological challenges, although the solution of the former are often intimately linked to the uncovering of the latter.

Parallel to the technological explosion, another, only apparently orthogonal, line of research has focused on understanding the interplay between different layers of a given scientific problem. This approach is challenging both in terms of a) its mathematical formulation [6] and of b) disciplines intercommunication. Initiatives like the Virtual Physiological Human (VPH) in Europe [7] the Biomedical Information Science and Technology Initiative (BISTI) in the USA [8] or the Systems Biology Institute (SBI) in Japan [9], among others, have emerged from the need to solve the two questions above. Following this, the term "multiscalability" has become an item of intense research, recognizing the need to tackle problems in an integral way, considering the detail that is critical at each level while avoiding expensive exploration of uninteresting regions of the multiple level phase space. The multiscale view even holds true within the realm of biomolecular simulations [10], in which interests range from the quantum mechanical description of enzymatic reactivity [11-14] to, say, the mesoscopic description of protein membrane insertion [15-16] and function [17-18]. Thus,

*Address correspondence to this author at the Research Group on Biomedical Informatics (GRIB) - IMIM/UPF, C/Doctor Aiguader, 88, 08003 Barcelona, Catalunya, Spain; Tel: +34 93 316 0504; Fax: +34 93 316 0550;
E-mail: jordi.villa@upf.edu

[#]*These authors contributed equally to the article*

1389-2037/11 $58.00+.00

one can distinguish between several time scales of interest when studying proteins. For example, side chain movement can occur at the scale of ps, while loop closure or helix formation (≈1ns-1μs), β-hairpins folding (≈1μs-1ms), domain folding (≈1min-1h) or protein aggregation (≈1h-years) represent totally different scales and, accordingly, biological problems [19].

Protein-protein interactions themselves represent a paradigm for the integration of different levels of granularity. When proteins and/or metabolites are considered far apart and immersed in a continuous medium, macroscopic simulations of concentration as a function of time can be handled by continuum modelling of biochemical kinetics. Thus, deterministic differential equations can be used to understand processes taking place at this high granularity level [9]. At a closer look, stochastic simulations are needed to understand discrete transformations of molecular populations [20-22]. The propensities to interact used in the previous approach can be related through kinetics theory to higher granularity approaches that solely consider the Brownian motion of particles representing the interacting molecules [23], guided by random friction of the surrounding water environment and by electrostatic interactions. An even closer analysis leads to a visualization of proteins as a collection of secondary structure moieties and individual residue interactions, which further reduces to an atomic level description when the two proteins reach contact distance [24-25].

The different levels described above are part of disciplines as far apart as systems biology and quantum chemistry, and integrating them is a formidable challenge. In this review the study of the formation of peptide aggregates by means of multiscale simulation approaches is put into focus, expanding the view that has been built in the last decades within the protein folding community. Thus, the scope is narrowed to the level in which two approximating molecules switch the "perception" they have of each other from a blurry and fuzzy electrostatic potential to a detailed collection of precise hydrophobic patches and directed interactions of the hydrophilic groups (see Fig. (**1**)). Despite the fact that on-lattice methods have made extraordinary contributions to the understanding of the basic principles underlying protein folding and aggregation, this review will focus solely on off-lattice methods and their use in molecular simulations as an ultimate tool directly applicable to problems of biomedical interest.

In particular, because of its biomedical relevance, protein or peptide aggregation has been one of the initial systems researchers have explored using multiscale schemas to shed a light onto disease related peptide oligomerization. This includes transtyretine, amyloid peptides, α-Synuclein, prions or tau protein aggregations, among others [26-27]. Although the location and hydrophobic character of specific protein segments is key for aggregation, experimental studies have determined that the propensity to acquire pathogenic conformations is also critical [28]. Thus, both exploration of the folding landscape and understanding of the energetics of molecular interaction have to be taken into account when exploring the aggregation propensities of peptides. In this paper we are not concerned about the detailed description of the different cases but rather on the multiscale simulation methods used to understand the oligomerization process, and thus, different examples will appear when needed to illustrate particular methodological developments.

The paper is organized as follows. The current literature on the use of multiscale simulations in peptide aggregation is first reviewed, building on top of solutions found in related fields like protein folding simulation or protein docking. After that a short paragraph on the analysis of the methods is provided, and these are dissected into small pieces of a common machinery, trying to make sense of the way researchers tackle the multiscale problem. Such analysis leads to a synthetic effort in the section *A general protocol for MS of protein aggregation*, in which an integrative method proposed by Warshel and collaborators is explained in detail and combined with new developments on simplified solvent modelling. A discussion of the possibilities of the proposed approach is also presented. Finally, a discussion on the challenges multiscale methods are facing is included, with a reflection on the way science should take profit of both computer power and improved algorithms in the upcoming years, especially in the dawn of the VPH and related initiatives.

2. STATE OF THE ART

While protein aggregation plays a key role in many naturally occurring processes, like the assembly of the cytoskeleton, in other cases it leads to a form that does not perform any function or even appears to disrupt natural processes. For example, amyloid fibrils have been implicated in several diseases, as Alzheimer's and Huntington's diseases. Although fibril formation can be described as a nucleation-growth process, the molecular details of the steps leading to the formation of amyloid fibrils are still unknown and are the subject of intense research. Simultaneously, it is obvious that protein aggregation is intimately related to protein folding, except for the non-negligible extra complexity due to the unconstrained combination of peptide fragments in the former process. Thus, one may take profit of the work done on protein folding in the last decades in order to extract hints on how peptide fragments will aggregate and, indeed, many researchers have been bridging the two research subfields. Thus, it is a good strategy to start this review of the state of the art by summarizing our current understanding of the process of protein folding.

Protein function is intimately related to its structure, which is determined by the interactions among amino acid residues. The interplay between hydrophobic and hydrophilic interactions (including charge-charge and hydrogen bond interactions) determines the overall shape of a protein. Despite the huge number of degrees of freedom of a peptide chain, it is generally accepted that proteins have been selected through evolution to quickly fold into one equilibrium structure, the so-called native state. This overall process has been explained by the folding funnel hypothesis, that summarizes the exploration of the complex potential energy surface into the elegant proposal that the native state corresponds to a global minimum in the free energy landscape (FEL) of a single protein in solution [29-30]. Proteins tend to be relatively stable to mutations and changes in the environment near physiological conditions [29]. Nevertheless, an appropriate change in temperature, pH, osmotic conditions,

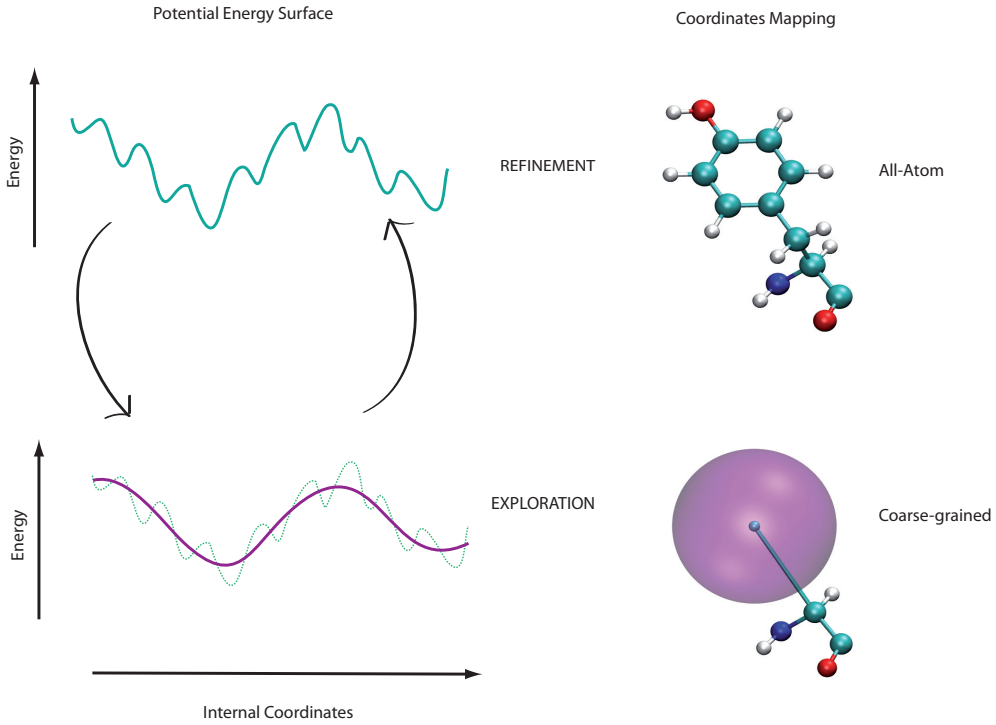

Fig. (1). A multiscale approach for molecular simulations consists of a minimum of two phases. The first (coarse grain) provides a fast exploration of the potential energy surface (PES) and the objective is to sample its relevant regions in an approximate way. In a second phase, one transforms the coarse grain representation to a more detailed one. Now the PES has a more complex structure which forces the run to be done with smaller step sizes and more expensive potentials. Thus, the critical aspects of a multiscale run can be summarized in the need for a reasonable and fast exploration of the relevant regions at the low level description and a detailed potential for the high level description that can give realistic behavior of the system.

redox state, mutations, etc., modifies the FEL with the concomitant appearance of new misfolded structures that can lead to, *e.g.*, protein aggregation. This means that studying the conformational space of proteins and peptides can yield hints on their propensity to achieve a given secondary, supersecondary or tertiary structure, and that understanding the effect of the above mentioned conditions on the free energy of the explored conformations can help rationalize the way protein fragments interact. The determination of the FEL requires an accurate description of the energy function, the potential energy surface (PES), along with a statistically robust sampling of the phase space, which is difficult to achieve due to the size and the complexity of the biomolecular system.

In what follows, our discussion of the literature focuses on three main problems for multiscale molecular simulations in general and on protein aggregation models in particular: the definition of both proper coarse grain and explicit potentials, the connection between the potential energy surfaces at the different levels of detail, and the use of a sampling technique that ensures correct exploration of the relevant regions of both the coarse grain and all atom PES.

2.1. Looking for a Proper but Reasonably Cheap Description of the Energy

The central issue to solve when dealing with protein aggregation is finding a potential function that properly describes the physico-chemistry of molecular interactions. A number of complications arise when one attempts to build such a function. Among others, the proper treatment of interaction energies, the need for simplification in the description of the constitutive elements in the peptide chain, and the introduction of a bias to emphasize native from non-native contacts.

At the highest level of resolution, quantum mechanical (QM) methods have to be chosen in order to describe the potential energy surface of a given atomic system, from the Born-Oppenheimer approximation to the time independent Schrödinger equation (see, *e.g.*, [31]). Unfortunately, this approach is absolutely unpractical when evaluating the energy of systems bigger than around a few hundred atoms and is out of the question if one aims at understanding their dynamics. In order to study larger systems, some empirical description of the PES must be adopted, namely molecular mechanics (MM), the classical counterpart of choice to QM methods. MM allows the simulation of fully hydrated proteins at all-atom (AA) detail. The proper treatment of electrostatic effects is one of the main problems to overcome, as occurs in receptor-ligand binding or enzyme reactivity simulations. Despite efforts to produce proper methods [32], these also need to be cheap to allow a thorough exploration of the conformational space of the two interacting peptides. In particular, when dealing with water properties and degrees of freedom in a simulation, one can include water explicitly, implicitly, or, taking the best of both extreme models, somewhere in the middle [33]. Although this review is not concerned with solvation methods, the extensive recent literature on microscopic, semi-macroscopic and macroscopic treatment of water as a solvent in molecular simulations

demonstrates that there is still no ideal solution in terms of the quality of the potential and speed of exploration (see, *e.g.*, [32-40]).

In order to explore the phase space of systems larger than small peptides one should decrease the complexity of the model through the elimination of some degrees of freedom. This reductionist approach, known as coarse-graining (CG), condenses groups of atoms into single interacting sites. An important effort has been made over the last years to develop cheaper but still physically realistic models for the description of proteins with different scopes on their application [41-43]. Several models using a different number and definition of beads to provide a coarse grain description of the peptide fragments have been proposed [42-43].

Many approaches have been considered to construct CG models. One possibility is the use of all-atom simulations in order to derive parameters for CG models that are able to span larger lengths and time scales. In this way, the multiscale coarse graining (MS-CG) methodology is a bottom-up approach in which atomistic MS simulation data can be directly incorporated into the CG force field. In a typical example of the application of the MS-CG approach in a dynamical context, Zhou *et al.* describe the equilibrium properties of two simple peptides, the Ala_{15} helix and the V_5PGV_5 β hairpin [44]. Two-bead and four-bead per residue mapping schemes were tested on their ability to reproduce structural features of the all-atom simulation. The sidechain is represented with one bead and the backbone is modeled using either a single bead or three beads, i.e., one for the NH, one for the HCα, and one for the CO. Each water molecule is represented explicitly by one bead. The effective FF interactions are divided into bonded and non-bonded terms. The former involves sites separated by 1, 2 or 3 bonds, defined based on chemical connectivity and described by employing standard functional forms (harmonic terms for bonds and angles, cosine series for dihedral angles) that are parametrized to fit probability distribution functions from all-atom trajectories. Nonbonded interactions were evaluated using the MS-CG force-matching method. As a consequence all nonbonded (electrostatic and van der Waals) interactions are subsumed into a short-range effective FF, which are computed and tabulated (they are not fitted to classical analytical forms) to be used in CG simulations. While the two-bead model was able to capture the behavior of the Ala_{15} helix it was insufficient to mimic the asymmetries inherent to the β hairpin. The resolution of this model is also too low to differentiate between stereoisomers [45]. On the other hand, the four-bead model reproduced the structural features of both peptides correctly even over-stabilizing the secondary structural elements. All these models with simplified representations have been found useful when analyzing physical based hypotheses like the principle of minimal frustration, but are problematic when trying to reproduce detailed mutational experiments by the nature itself of the coarse graining itself. The method also has poor transferability, so it may need to be determined separately for different proteins and even for different thermodynamic states.

On the contrary, the Martini CG FF [46] is a top-down approach where the parameters are fitted to thermodynamic data, in particular oil/water partitioning coefficients, thus providing a transferable CG FF. This forcefield is well suited for processes such as protein-protein recognition which depend critically on partition coefficients between polar and nonpolar environments. On the negative side, the need to apply constraints to preserve the protein structure stability during simulation prevents the FF from being applied to the exploration of conformational changes and protein folding. Slightly separate from this classification are normal mode analysis and related approaches, which can be referred to as examples of harmonic or quasi-harmonic static descriptions of the PES by bead-based representations of protein structures [47].

Despite their limitations, these methods become powerful when combined with some sort of knowledge-based potential. The OPEP (Optimized Potential for Efficient prediction of Protein Structure) [48] is a 6 particle-based coarse grained model inspired by Levitt and Warshel's early work [49]. The backbone is represented explicitly, while sidechains are accounted for by specific beads, and the energy function is adjusted to discriminate native from nonnative structures. It is of interest that the OPEP-generated free energy landscapes of small proteins are fully consistent with experimental data, providing further support of the capability of this forcefield to also reproduce thermodynamics. In a similar manner, Messer *et al.*[41] have recently proposed an update of the original 1975 CG FF [49] (see Fig. (**2**)) that is the basis for the discussion in the synthesis section below. Typically, though, approaches based on statistical potentials lack the powerful decomposition schemas which form the basis of common molecular mechanics potentials, although some efforts to rationalize such decompositions are already available[50].

In CABS, the coarse-grained forcefield developed by Kolinski in 2004 [51], the atomic structure of a protein is radically changed and replaced by a number of interaction centers. In particular, the aminoacid side-chains are removed and two interaction centers are placed at the former alpha and beta carbons, which gives the forcefield its name (carbon-α carbon-β side chain). The forcefield uses heuristic potential functions which are derived from a statistical analysis of structural regularities in folded proteins. The alpha carbons are restricted to a position on a sufficiently fine grained lattice so as to prevent any strong lattice artifacts, while side chains are allowed to move off-lattice. Computations of micromodifications of the structure are intrinsically faster than in similar coarse-grained forcefields, as they merely consist of a few references to precomputed tables of allowed conformational transitions. The same group studied amyloid protein aggregation [52] using a topologically similar, but methodologically different representation called REFINER [53]. It differs from CABS by not employing lattice restrictions, but it also simplifies the protein backbones and side chains with a similar amount of beads. By using this type of representation and a replica exchange protocol, they studied the amyloid aggregation of a selection of artificial peptides, which in their native basin all adopt helix structures, but also adopt beta-form structures in the higher energy misfolded and metastable basin. Both basins are separated by a conformational energy barrier. The results of these simulations showed that under normal conditions the most stable structure always remains a two-helix bundle, while

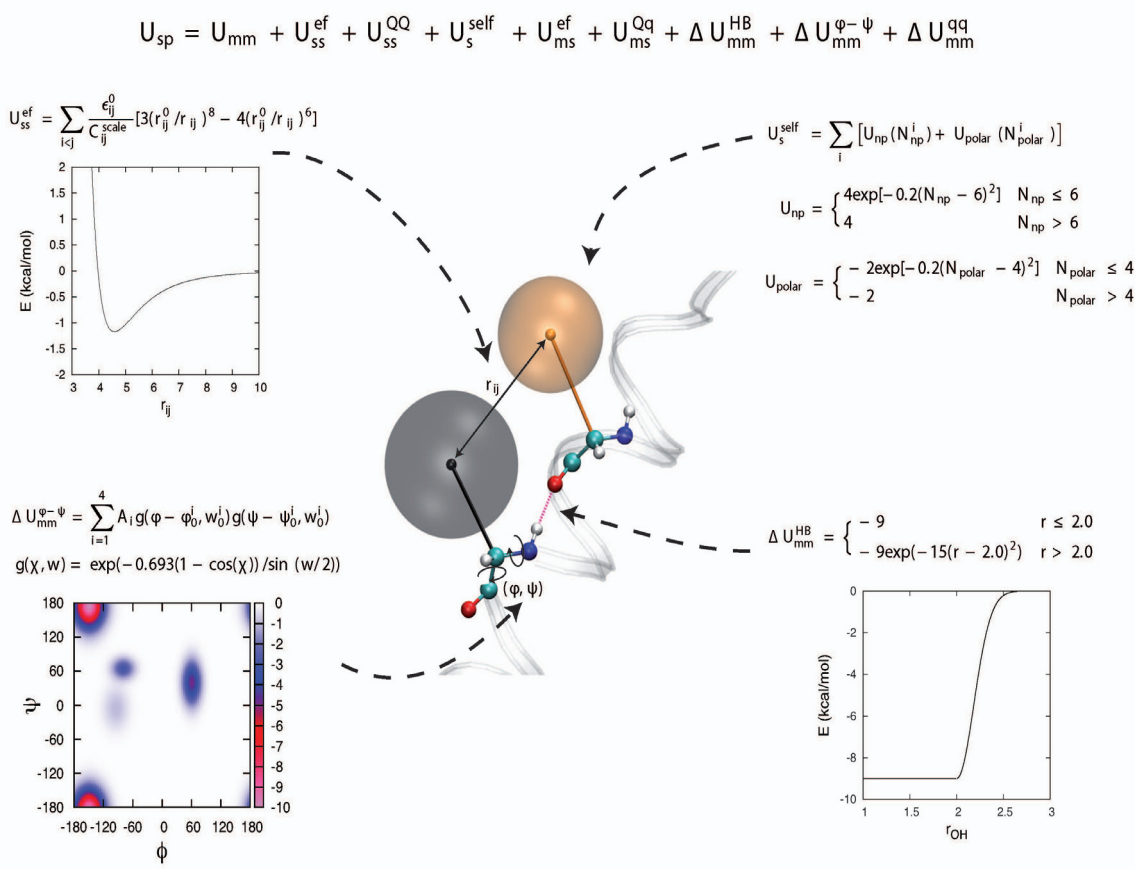

Fig. (2). The coarse grain model proposed by Warshel and collaborators[41]. Its interest lies in the proper combination of native-like conformations by forcing the Ramachandran plot to be repro- duced and in the detailed inclusion of the permanent dipoles of the protein backbone, critical for secondary structure formation.

dimerization simulations with one of the peptides frozen in the already misfolded beta-form structure causes the other peptide to also misfold in the majority of cases.

Our understanding of how proteins fold is in part derived from the so-called "principle of minimal frustration", by which protein molecules reduce, through evolution, the prevalence of conflicting interactions that would lead to alternate configurations [54]. Thus, native contacts are cooperatively responsible for the minimum free energy folded state and, because of this, using them to bias the potential energy surface towards native like conformations may help accelerate folding or aggregation simulations. This leads to the design of so-called Gō-like models, in which the simulation is biased towards the native structure. Of course, again, such bias can be achieved by the use of statistical potentials, obtained from the analysis of a huge number of structures from the protein data bank (PDB), although unfortunately typical Gō-like models are specifically parametrized for the protein being studied and thus lack generality. Some CG models consider only the native interaction energies as the driving force in the folding problem, neglecting non-native interactions. Although this approach has been successful in qualitatively reproducing the general features of the transition state structure for small proteins, there are two problems. First, Gō-like models typically consider all native contacts to have the same importance (although B-factors or the protein contact number [55] can be used to weight them); and second, they do not consider the energetic contribution of non-native contacts, which is essential for understanding misfolding and aggregation. More precisely, non-native interactions may play an important role in shaping the early stages of the folding landscape, in the formation of on-pathway intermediates, in restricting the accessible configurational space in the unfolded state and in the formation of the transition state.

The work of Clementi and collaborators nicely illustrates the change in the paradigm from folding to misfolding and aggregation studies, and the importance of incorporating non-native interactions. Matysiak *et al.* [56] proposed a Gō-like model that only takes into account native contacts to represent the protein, with one bead per amino acid. Such models have been previously used with good qualitative results for two-state folding proteins. In order to get better quantitative agreement with experiments, they tuned the weight of the different native interactions to reproduce experimental data for kinetic folding/unfolding. The procedure was inspired by the reverse Monte Carlo method originally proposed by Lyubartsev & Laaksonen [57]. In [56], the one bead CG force field was parametrized with experimental data on S6 (wild type -WT- and P13-14 permutation mutant), a 97-residue ribosomal protein consisting of a four-stranded B-sheet and two alpha helices. Folding/unfolding of S6 can be shown to occur by two-state kinetics, and their model was able to reproduce experimental data on a different permutant,

P68-69. Next, Das *et al.* [58] modeled the protein as a sum of pairwise interactions between 20 aminoacid "colors", not removing a priori non native interactions. The parameters were optimized to get a minimally frustrated Hamiltonian driving the model to fold into a known protein structure, resembling pure Gō-like approaches. The effective residue-residue interactions were modulated by "colors" representing the amino acid's chemical and physical diversity, and the geometry of different amino acids was implicitly considered by using equilibrium distances per amino acid pair (taken from distributions observed in the PDB) rather than a constant value (as was used before). As recognized by the authors, the incorporation of sequence details and energetic frustration into an overall minimally frustrated protein landscape opens the possibility of a realistic study of the delicate balance between energetic and topological factors in shaping the folding free-energy landscape. With this model, Das *et al.* were able to correctly reproduce experimental data on protein folding for SH3, a protein showing two-state kinetics and for which the folding mechanism can be qualitatively mimicked by using completely unfrustrated models. Nevertheless, several studies had highlighted the important role played by non native interactions in shaping the folding landscape. The incorporation of non-native interactions into the model allows for this hypothesis to be tested, and indeed they found that taking them into account results in faster folding rates (lower energy for the transition state). Extending the work in [58], Matysiak and Clementi [59] studied the effect of multiple mutations on the folding mechanism of S6. This time, the inclusion of nonnative interactions in the CG model allowed it to be used not just for folding but also for misfolding and aggregation studies. They found perturbations in the free energy landscape of S6 upon the mutation of 4 gatekeeper residues, providing an interpretation for the increased propensity to form protein aggregates: the mutation of these residues exposes the protein to partial misfolding. Non-native interactions play an important role in destabilizing the native state while stabilizing misfolded traps. Interestingly, the simulation in ref. [59] of multiple copies of the WT protein does not produce any changes on the free energy landscape, while running the same simulations on the mutant leads to the formation of stable aggregates (corresponding to new minima in the free energy landscape), including one form that is reminiscent of an amyloid protofibril.

2.2. Sampling the Potential Energy Surface

As explained above, multiscale methods aim at providing efficient tools for sampling the explicit potential phase space through serial or parallel integration of information at several levels. For example, classical MD approaches employed to study the correlation between the states sampled by small peptide monomers with their tendency to form amyloid structures [60-63] are no longer valid when dealing with more complex systems. Coarse grain potentials help in producing a better exploration of the potential energy surface in two ways: they reduce the size of the system and thus the degrees of freedom to explore and the computing time for the forces at each time step; and consequently they smoothen the potential energy surface. Both effects help in accelerating the system dynamics.

Multiscalability is also reflected in the way the sampling of the PES is carried out. For example, in [64] two different methods developed for protein folding and docking are melted into a more general procedure that permits the blind prediction of intertwined complexes of proteins at near atomic resolution. Starting from an extended conformation, in a first stage a set of moves that perturb the backbone conformation and moves that perturb the symmetric docking arrangement of the monomers are performed. The conformational energy is determined by the low-resolution energy function and each move is accepted or rejected based on the standard Metropolis criterion. In a second stage, the sidechains are reconstructed within a context of a high-resolution energy function. The lowest energy models are clustered and the five most populous are selected. The methodology works properly for small peptides. However, for systems with a larger degree of freedom (> 60 residues), the protocol fails due to insufficient conformational sampling, a common problem encountered with classical Molecular Dynamics (MD) and Metropolis-Monte Carlo (MC) approaches.

In order to improve the exploration of the PES, several methods have been devised. Although the vast amount of enhanced sampling techniques available exceeds the scope of this review, some interesting tools that have been lately proposed are worth mentioning.

The ART (Activation-Relaxation Technique) [65] is a method developed to identify transition states without prior knowledge of the final state. ART consists of four steps: 1) starting from a local minimum, the system is pushed slowly in a random direction, and then the direction of the lowest curvature is evaluated, until it becomes convex (negative eigenvalue of the PES at that configuration); 2) the configuration is displaced along the direction of the negative eigenvalue while the energy is minimized in the orthogonal directions; 3) when the total force reaches zero, one has reached the saddle point; 4) move over the saddle point and minimize to the new minimum. The technique resembles other methods of uphill exploration of the PES. Although the method is able to generate a fully connected walk through the energy landscape, it should be noticed that such trajectories do not belong to a well-defined thermodynamical ensemble, and as a consequence an exact weighting of the various conformations is not possible. An interesting research line to explore is the use of this method within the context of the variational transition state theory [66] and the use of arbitrary reaction coordinates to describe the free energy profile [67]. In spite of its limitations, the combination of ART-OPEP was able to provide a quantitative match to experiments while considering protein folding and protein aggregation of amyloid-forming proteins. It has proved to be especially useful in identifying the richness in the structure of small aggregates of short chains. The simulations also indicate that there is an important difference in the early steps of aggregation between short and long peptides. While the former visit fibril-competent structures very often, the latter populate structures that are far from the amyloid fibril structure.

Lyman *et al.* [68] presented two methodologies to increase the sampling in molecular dynamics simulations: resolution exchange (ResEx) and PseudoExchange (PsEx). The key idea behind ResEx is that one can swap a subset of

configurational coordinates. A well-chosen subset of coordinates of a detailed model can make up the full set of coordinates for a CG model. For instance, the backbone of the AA protein model can represent the full set of coordinates for the CG model. The process involves running two independent simulations of the protein and trying to swap the subset of coordinates common to both representations, according to Metropolis criteria. PseudoExchange, in turn, is a serial process where one first generates a well-sampled ensemble at the coarse-grained level and randomly reorders this trajectory. While keeping the original distribution of states, it shows a key feature: extremely rapid barrier hops. One then performs a fine-grained simulation and exchanges are attempted with the shuffled coarse-grained trajectory following the same Metropolis criteria. The method has two limitations. First, it does not enable exchange between continuum and explicit solvent representations. Secondly, in order to be exchangeable, the two models must be sufficiently similar and there should exist overlap between low-energy coarse variable conformations. The usefulness of these methodologies is demonstrated in the sampling of butane and a dileucine peptide represented in all-atom and united atom models. When trying to apply the ResEx procedure to systems bearing larger degrees of freedom, a new problem arises as rejection rate increases [69]. In order to increase acceptance rates one may minimize the high resolution trial before checking acceptance. As shown in their work, such solution violates the detailed balance condition by biasing the generating probability while not introducing any compensating correction to the acceptance criterion. A workaround is suggested by the authors where a ladder of incremental models at intermediate resolutions is employed that allows the acceptance rates to be tuned to reasonable values (thus, not requiring a minization). This ladder is constructed as follows: starting from an all-atom model, in the next level only one residue is coarse-grained, next two residues, and so on and so forth until the whole chain is represented with a coarse-grained model. The implementation of such an incremental ladder requires the construction of a potential function with mixed potentials between CG and AA representations. Again, a top-down exchange scheme is used, i.e., first a coarse-grained model is used to achieve a converged trajectory, then configurations from this trajectory are exchanged with the trajectory from a lower level (more detailed) in a serial approach until the all atom level is reached. A potential weakness of the ResEx method is that regions which are not sampled by the top level will be difficult to sample in any other level.

2.3. Attempts for Integration of Different Levels

Multiscale simulations allow one to split the treatment of the hugely complex aggregation process between at least two levels, but this also includes the difficulty of linking them. This is a well known problem in enzyme reactivity and QM/MM approaches were developed as protocols to link quantum and molecular mechanics descriptions of different regions of the system [70-72], in a first example of a parallel multiscale approach.

Ayton *et al.* [15] extensively reviewed different multiscale methods implemented so far. They noted that multiscale methods can be broadly classified in two categories, serial and parallel approaches [15], according to the way that the information is transferred across different resolutions. In the former, the different models are used in sequence with no direct interaction between them. On the other hand, in parallel approaches the different representations are treated at the same time, with information being transferred among them. They also noted that most methods at that time put emphasis on obtaining CG models (either from AA simulation or thermodynamic data) and not on the integration workflows.

We have seen in the previous section how coarse grain potentials are built to allow for an extensive exploration of the conformational space of the interacting protein fragments. The possibility of extensively exploring such simplified potentials is at the basis of multiscale modelling, but a connection to the explicit potential (typically an all atom potential) is needed. The underlying assumption is that the CG potential allows for the sampling of relevant conformations of the protein. The reintroduction of atomic resolution into the model is desirable in order to zoom into the finer details of the process and to correct for deficiencies in the original CG model.

In this way, the MS-CG was further applied to the study of the peptide folding landscapes of the Ala_{15} α-helix and the V_5PGV_5 β-hairpin [45]. Here, MS-CG peptides exhibited a preference for folded regions of configuration space as a consequence of the knowledge based potential employed. As a result, effective interactions in the CG systems are to some degree incompatible with regions of configuration space which differ from those typically explored by folded peptides. This effect biases sampled CG configurations towards the original domain of all-atom integration by facilitating fluctuations that convey the system down the relevant free energy gradients. As a result, unfolded configurations are disfavored and progress towards the folded basin, resembling a Gō-like model (see below). In contrast though, the MS-CG representations were not explicitly constructed to be smoothly funneled towards the native state. It is thus surprising to find that these models are still able to reproduce characteristics of partially unfolded regions on the free-energy landscape that lie outside of the original basin corresponding to the native structure used for deriving the CG parameters. This shows that the effective interactions are, to some degree, transferable between configuration spaces.

Additionally, the authors also explored the ability to reconstruct missing all-atom detail from CG trajectories. It was observed that the free energy landscapes generated in this way exhibit reasonable stability in the vicinity of the minima originally observed in the MS-CG simulations. This demonstrates that MS-CG represent refolding landscapes, providing further validation to the model. However, it should be noted that the reconstructed all-atom simulations were still unable to surmount the barrier to folding. As a conclusion, MS-CG can be used to explore the configuration space efficiently and the trajectories used to reconstruct all atom ensembles in the regions of interest, correcting deficiencies inherent to the original CG model, but still the study of folding and aggregation barriers needs a more precise simplified potential.

A multiscale approach to the study of protein folding and misfolding was also considered by Heath *et al.* [73-74]. They introduced the RACOGS method to generate all-atom

representations from C-alpha models and applied it to reconstruct folding trajectories generated in previous studies on SH3 and S6 [58-59]. Their reconstruction procedure is able to efficiently produce relatively low-energy (i.e. statistically significant when Boltzmann-weighted) AA structures from most CG structures. The reconstructed structures are validated by their energy in a standard forcefield (AMBER99) with a GB/SA implicit solvent model. The final energy of the structure is used as input to the weighted histogram analysis method (WHAM) to calculate the free energy of the all-atom model. Although very different energy functions are associated with the CG and AA models, the free energy landscape obtained from both trajectories remain remarkably similar, proving that it is possible to use a good CG model as a robust starting point for an extensive sampling of protein complex landscapes at an all-atom resolution. Upon addition of AA detail to the FEL of S6 Alz, it still shows a bulge in the landscape that is associated with a population of partially misfolded structures not present in the FEL of the wt protein [74]. Moreover, the population associated with misfolded structures becomes larger and more distinct, signaling that the misfolded states are partially stabilized in the AA structures and, thus, again shows the need for the inclusion of non-native interactions. A closer look at the misfolded structures in [74] yields information on the misfolding mechanism. While β-strands 1 and 2 do not interact in the correctly folded structures, these strands pack against each other and are stabilized by the formation of multiple interactions between their side-chains in the overall repacking of the four β-strands in the misfolded structure.

Urbanc *et al.* [27] used coarse-grained discrete molecular dynamics simulations to predict the aggregated dimerical structure of two *A*β peptide monomers. In a multiscale approach these predictions were retransformed into an all-atom representation and free energy differences between the monomers and the dimer were calculated. The coarse-grained representation involved the modeling of the amino acids as 4 beads that correspond to the amide nitrogen, alpha, prime, and beta carbons. Coarse grained monomers were placed 40 *A* apart and simulations were performed under different temperature levels. The simulations resulted in ten dimer structures which were frequently reoccurring. These ten structures were tested for stability by performing free energy difference calculations between the dimers and corresponding monomers with the CEDAR all-atom forcefield and the ES/IS method [75]. This methodology gave insights into *A*β monomer aggregation by showing that thermally induced conformational changes lead to dimer formation, corresponding to one of the ten dimer structures. The probability of which dimer is formed is dependent on the temperature, with each dimer structure varying from parallel or anti-parallel beta sheets to nested structures.

Kolinski's lab employed the CABS forcefield within a replica exchange Monte Carlo simulation protocol to study protein-protein docking [52, 76]. Their protocol included the positioning of both monomer structures in a random orientation relative to each other, each with its own replica, and a posterior multiscale analysis of the resulting dimer structures. Their multiscale approach included the clustering of discovered dimers and the selection of cluster centroids, as well as the best structure in terms of coarse-grained energy. These structures were subsequently converted into all-atom representations by firstly reconstructing the backbone atoms with BBQ [77], secondly, the assembly of side-chains via a rotamer library and thirdly, all-atom minimization using an AMBER forcefield with implicit solvent. They observed that the all-atom energies correlated much better with the RMSD than the CABS energy which subsequently increased the accuracy of the prediction of the dimer. From 11 protein-protein or protein-peptide dimerization experiments, all predictions were qualitatively correct with the percentage of native contacts each greater than fifty.

The explicit solvent / implicit solvent (ES/IS) [75, 78] methodology is a tool to calculate conformational free energies of proteins. It is a multiscale approach in which, during a short molecular dynamics simulation in explicit solvent, microstates are sampled for which free energies are computed by using implicit solvent calculations. In an ES/IS simulation, the conformational free energy is compartmentalized into terms describing the intra-protein bonded and non-bonded interactions as well as its entropy, and terms describing the free energies of creating an empty cavity in the solvent, putting a protein into the cavity, and charging the protein, which also results in the polarization of the water and vice versa. While the conformational entropy can be easily estimated by a quasi-harmonic analysis of the covariance matrix of the protein's fluctuating movements, and the intra-protein enthalpic energies by averaging over microstates, the same is not true for the free energy of solvation. This is due to the need of approximating the partition functions of the protein in water, as well as the water itself, and a free energy perturbation from which the difference in free energy can be obtained. However, the perturbation is extremely costly due to long relaxation cycles of the water and, as such, it is not a feasible approach. By using the Poisson Boltzmann equation and calculating free energies implicitly with numerical Poisson Boltzmann solutions, Generalized Born [79-81] or Semi-Explicit Assemblies [35, 82], free energies of solvation are cheaply available. Vorobjev and coworkers employed this multiscale methodology to calculate free energy differences between correctly folded and incorrectly folded proteins in order to determine if such a method could reproduce the fact that native conformations are energetically more favorable. Their experiments resulted in successful predictions of the native states of the proteins in their test set, showing that conformational entropy is responsible only for a small part of the overall free energy of the system.

3. DISSECTING THE PIECES

From the previous section an overall schema for multiscale simulations of proteins in general and of protein aggregation in particular can be extracted. The first step, the choice of a CG representation, requires taking into account several factors: good representation of the driving interactions for the system under study, speed of the force calculations and ease of going back to full atomistic details. In the case of proteins, typical choices span the spectrum from coarse graining at the residue level all the way down to the coarse graining of non-polar hydrogens only.

Next, the CG model needs to be parametrized. Parameters can be derived from several sources: from experimental

data (PDB structures, NMR constraints, thermodynamic data) that allow construction of a statistical potential; from parameter sweeps that reproduce in the CG representation some observed behavior of the system (*e.g.* folding); from full atomistic simulations that compute potentials of mean force or other thermodynamic quantities; directly from atomistic potentials; and potentially using combinations of any of these.

Actual simulations with the CG model can benefit from standard simulation techniques. Depending on the process under consideration methods such as thermodynamic integration, umbrella sampling, and replica exchange can help overcome energy barriers and improve convergence of the CG simulation. In case better sampling of the CG potential is needed, other more complex approaches can be used as shown in the previous section.

The last step is to map the CG model back to a full atomic description. However, by the very nature of of the CG model there is not always a one to one correspondence between the CG model and the atomic model, nevertheless incorporating atomic details back into a CG model can guide the sampling of a CG simulation [68-69, 83] or help building a thermodynamic cycle to describe the FEL of the system [41, 84].

In the next section, a complete protocol for multiscale simulations of protein aggregation and folding as being implemented in Adun [4, 85] is described in more detail.

4. A GENERAL PROTOCOL FOR MS OF PROTEIN AGGREGATION

4.1. Thermodynamic Cycle for Protein Aggregation

We follow here the approach by Warshel and coworkers [41, 84] in which a basic thermodynamic cycle is exploited to describe a general multiscale framework that is able to calculate differences in free energy for an arbitrary process. The thermodynamic cycle as visualized in Fig. (**3**) describes a simulation from the reactant state to the product state either using an explicit all-atom representation of the involved molecules or a coarse-grained representation. Differences in free energy are meaningful statistics for the description of biological processes, however their evaluation in an explicit potential is very costly. With the rise of coarse-grained architectures, a way has emerged to speed-up the evaluation enormously. As shown above, refinement of the CG energies is a key issue to be addressed by any useful method. In our implementation, following [41], we combine a coarse-grained simulation to explore the main features of the complete conformational space of the protein with a free energy perturbation protocol (FEP) in which the partition functions of small windows of the coarse-grained and explicit landscape are calculated.

As described in detail above, the question remaining is how to utilize these coarse-grain models in a multiscale framework in the most advantageous way.

4.2. Idea of a Simplistic Forcefield

The continuum of forcefields that are nowadays available and which have been in part presented in this review share all some common characteristics. First, computational demand is greatly reduced by a massive reduction of degrees of freedom. In addition, the majority of small vibrational and rotational modes is eliminated and as such the energy landscape is smoothened. The latter removes local minima in which the simulation is not trapped, but unnecessarily stalled. The thermodynamic cycle asserts that for every state of the reaction coordinate a back-transformation from the coarse-grained representation to the explicit is possible. This property restricts the type of force field to be used. Heath and coworkers, for example, defined a general solution to the back-transformation problem that can be employed in any forcefield as long as the alpha carbons of the backbone are retrievable [74]. Although the idea of using the above thermodynamic cycle is independent of the coarse-grained and explicit forcefields, we will herein present in some detail the coarse-grained forcefield from Warshel and coworkers, which is based on the forcefield of Levitt and Warshel [49]. The details of this force field can be found in [41] and are shown in Fig. (**2**). However, we emphasize here some key aspects of the potential.

In contrast to most of the aforementioned CG models, the forcefield from Warshel and coworkers does not rely on a priori knowledge of native structures to perform molecular simulations. Besides the replacement of the solvent with a continuum dielectric and the coarse-graining of its side-chains to spherical structural units, the atomic structure is not further modified and most notably, the backbone is left intact. This has two implications. Firstly, computational demand is greatly reduced due to the massive loss of atomic detail typical to all CG force fields. Secondly, important main-chain interactions i.e. hydrogen bonds can still be completely described with as much detail as their explicit counterparts. Side chains are represented by a single bead whose center overlapping with the former geometrical center of the heavy atoms, or in the case of ionized residues, with a point shifted to the charged center of the residue. At the position of the former beta carbon, a dummy atom is placed which can be used in a back-transformation step in which the coarse-grained protein is converted into a full-atomic representation. As mentioned earlier and in contrast to other coarse-grained forcefields, *e.g.* MARTINI, the solvent is completely removed. Solvation effects are instead described by three correction terms which are added to the forcefield function. Briefly, the force field function consists of mainly four sections. One corresponds to the aforementioned solvation terms. The other three are concerned with energies due to interactions of the coarse-grained side-chains with other side-chains as well as the main-chain, and also explicit interactions involving the full-atomic interactions of the main-chain. These explicit interactions are modeled completely in agreement with the ENZYMIX forcefield. This basically means the employment of bond, angle, torsional van der Waals and electrostatic terms. For the side-chain potentials only interactions regarding van der Waals and electrostatics are taken into account. Both are calculated for the side chains and their interactions with the main chain. Van der Waals energies are calculated with a modified 8-6 Lennard Jones potential, with parameters being optimized to match in silico values for atomic distances and protein sizes of a number of different proteins. Electrostatic energies are computed with a

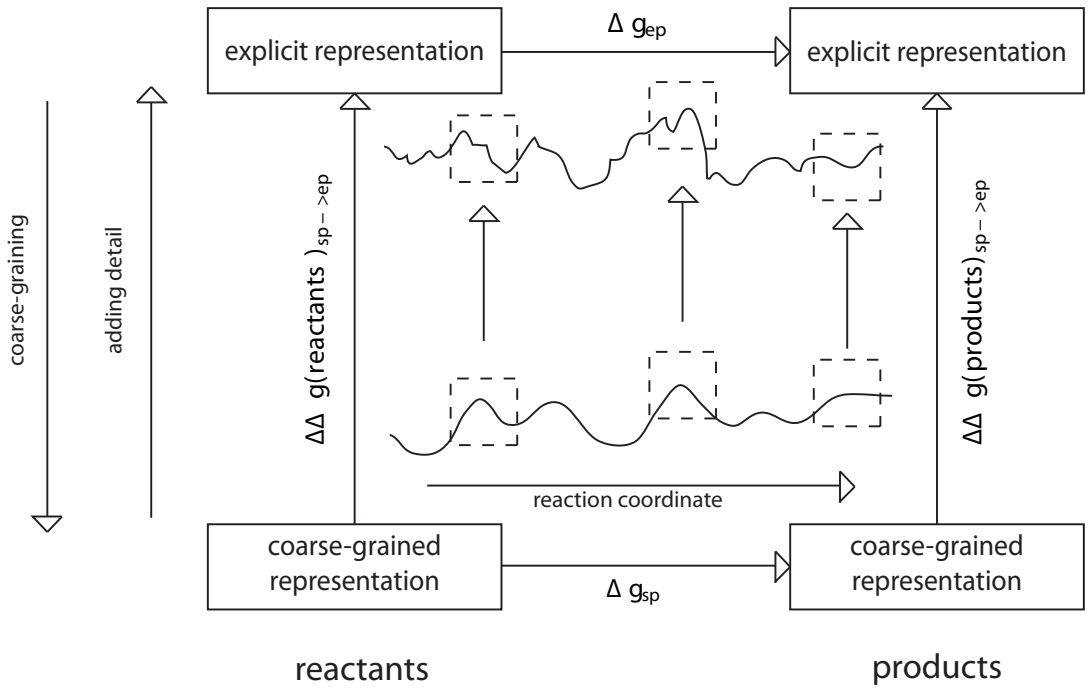

Fig. (3). Thermodynamic cycle employed to calculate free energy differences in Warshel's schema [41, 84]. The simulation on the coarse-grained potential (see Figure 2) is corrected, with the cost of moving from the explicit to the coarse-grained representation made using a standard free energy perturbation approach. The key issue lies on exploring the relevant regions already in the CG potential.

standard Coulomb potential. Charges used for this potential consist either of main-chain partial charges or residual charges which correspond to a residue's net charge. Interactions between side chains are treated with a high dielectric constant such as 40, while interactions between side chains and main chain are treated with a value of 10, and a value of 4 used for interactions among main chain atoms only. These dielectric constants were chosen to include some of the solvent's effects into the forcefield. Furthermore the aforementioned correction terms were incorporated into the forcefield equation to emphasize secondary structures and explain interactions with the environment. This includes a non-polar self energy term for transferring a side chain to the protein, a torsional potential to preferably form helices and beta sheets if φ and ψ torsional angles are already in certain regions of the Ramachandran diagram, and a hydrogen bonding potential to stabilize those secondary structures. The solvation correction terms were tuned with the inclusion of several potential wells into their potential functions. A very detailed reproduction of, in this case, the coarse-grained solvation free energy landscape in consensus with the explicit counterpart was possible.

4.3. Obtaining Free Energy Differences

The forcefield can be employed to calculate differences in free energy Δg for an arbitrary process in which the system evolves from the reactant state to the product state. Due to the coarse-grained composition of the potential functions, the energies cannot be as detailed as in a fully explicit representation. Another *e.g.* fully explicit forcefield is cooperatively used to counteract the loss of detail while maintaining the level of computational demand. Their liaison is visualized as a thermodynamic cycle in Fig. (**3**) in which the process is partitioned into four different sections. In the first, differences in free energy Δg_{ep} are directly evaluated with the explicit potential. In the other three, differences in free energy are evaluated by the employment of the coarse-grained potential Δg_{sp} and the involved energies of changing from one forcefield to the other $\Delta\Delta g_{sp\rightarrow ep}$.

While differences of free energy for the coarse-grained potential are obtained directly from the simulation itself, the same is not true for the energies of moving between the forcefields, but these can be derived from the ratio of the partition functions of the explicit and coarse-grained potentials. This ratio is given by:

$$exp\left[-\Delta G_{ep\rightarrow sp}\beta\right]=\frac{Q_{sp}}{Q_{ep}} \quad (\mathbf{1})$$

β being the inverse of Boltzmann's constant multiplied by the temperature of the system and Q_{sp},Q_{ep} the partition functions of the coarse-grained (simplistic) and explicit representations. This formula can be transformed to express the ratio related to the difference in energies of all sampled states in the coarse-grained potential and their counterparts in the explicit potential.

$$\frac{Q_{sp}}{Q_{ep}}=\frac{\int dRdr\ exp\left[-U_{sp}(R)\beta\right]}{\int dRdr\ exp\left[-U_{ep}(r,R)\beta\right]} \quad (\mathbf{2})$$

Note that the partition functions are expressed by using energies derived from conformations represented by the simplistic coordinate set R and the explicit coordinate set r,R.

$$\frac{Q_{sp}}{Q_{ep}}=\frac{\int dRdr\ exp\left[-\left(U_{sp}(R)-U_{ep}(r,R)\right)\beta\right]exp\left[-U_{ep}(r,R)\beta\right]}{\int dRdr\ exp\left[-U_{ep}(r,R)\beta\right]} \quad (\mathbf{3})$$

Which is:

$$\frac{Q_{sp}}{Q_{ep}} = exp\left[-\Delta G_{ep\to sp}\beta\right] \quad (4)$$

This equation reflects the average over all differences in potential energy.

$$\frac{Q_{sp}}{Q_{ep}} = \langle exp\left[-\left(U_{sp}(R) - U_{ep}(r,R)\right)\beta\right]\rangle_{V_{ep}} \quad (5)$$

This equation can now be sampled with the employment of a free energy perturbation method that uses the following mapping potential:

$$U_m = U_{ep}(r,R)(1-\lambda_m) + U_{sp}(R)\lambda \quad (6)$$

so that:

$$exp\left[-\delta\Delta G_{m\to m+1}\beta\right] = \langle exp\left[-\left(U_m - U_{m+1}\right)\beta\right]\rangle_{V_m} \quad (7)$$

and finally:

$$\Delta G_{m\to m+1} = \sum_{m=1}^{n+1}\delta\Delta G_{m\to m+1} \quad (8)$$

4.4. Going From Coarse-Grained to Explicit

From equation 3 we can conclude that for every sampled conformation *R* in the coarse-grained potential, its corresponding energy in the explicit potential has to be calculated. In order to do so, the coarse-grained representation has to be retransformed into an explicit representation. As the explicit representation is more detailled, this transformation is not unambiguous and therefore no unique explicit solution exists. However, it is possible to find one that is energetically sound. The procedure for the forcefield of Warshel and coworkers starts by replacing the dummy atom with the beta carbon of the corresponding explicit side-chain [41]. The new side-chain should comply to two requirements. Firstly, the displacement between the centers of the coarse-grained and the explicit side-chains should be minimal. Secondly, overall potential energy should be minimal. The side-chain rotamer configuration that accomplishes both goals with the employment of a random search followed by a steepest decent torsional minimization is selected and has to undergo side-chain only all-atom minimization. The resulting all-atom representation is then ready to be used.

If a different forcefield is used, it is still possible to achieve a back-transformation with the utilization of a general approach like *e.g.* RACOGS [74]. The only information this algorithm needs and uses is the position of the alpha carbons of the protein backbone. The procedure starts by filling up the missing heavy atoms of the backbone using a statistical method based on the average distances of the heavy atoms to the alpha carbons. Systematically, side-chains are attached to the growing structure by choosing the rotamer that minimizes the energy between the backbone and the other side-chains. After the side-chains have been assembled, unnecessarily high energy configurations are eliminated by finding the side-chains that cause them and performing a side-chain only all-atom minimization with the rest of the protein held fixed. In its final step, the algorithm performs a short all-atom minimization for the whole protein in which the structure moves into the nearest local minimum.

So far we have introduced a method for all-atom reconstruction of the solute, but have not mentioned what to do about the solvent. The previously presented forcefield approximates solvation effects as correction factors based on a continuous dielectric model. Its solvation technique is therefore classified as an implicit method. In order to restore the atomic representation of the solvent, water molecules have to be arranged around the solute in a physically sound way, which is firstly a very problematic task and secondly its subsequent simulation is very time consuming, and as such, not native to our seamless approach to multiscale molecular dynamics. However, by not going completely "all-atom", we open our protocol to a variety of macroscopic or semi-macroscopic solvation models for otherwise explicit forcefields which are nowadays excessively available [33, 79-81]. Focussing on generalized Born methods, the free energy of solvation is compartmentalized into two terms: polar and nonpolar. The nonpolar term is typically approximated by a value which is proportional to the solvent accessible surface area times a tension factor (γA), while the polar free energy of solvation is mostly determined by the buriedness of an atom, which describes the degree of electrostatic screening that atom is exposed to by the solvent. One major drawback of the GB-based methods is that the geometry of macromolecules is largely neglected. Unfavorable volume to surface ratios give nonphysically free solvation energies. In protein aggregation as well as protein folding, single water molecules can lead to unexpected but potentially essential effects that cannot be captured by a continuous model. Semi-explicit assembly (SEA) is a new implicit solvation method which explicitly takes into account the geometry of macromolecules and was developed recently to approximate the behavior of explicit solvation [35, 82]. Like the previously mentioned γA terms for the nonpolar part of solvation free energy, SEA uses the solvent accessible surface area for the same purpose, but employs it to weight previously calculated free energy values obtained from model atoms that are completely exposed to the solvent. These nonpolar free energy precalculations were performed for a number of Lennard Jones spheres with varied well depths and equilibrium distances by solvating them in TIP3 water. SEA calculates the polar part of solvation free energy as the sum of the solute's interaction with the first water shell and its bulk interaction with anything that goes beyond the first shell. In that sense, it is very different from GB methods as the discreteness of water and geometrical properties are directly taken into account in first shell interactions. With the computation of the electric field around a single atom, a dipole can be placed accordingly and the location of temporary solvent molecules established. Electrostatic effects are then accumulated for all solvent molecules and all atoms in the macromolecule and added to the overall free energy.

CONCLUSION

Multiscale simulations of molecular aggregation are a promising field of research. Computational experiments like the critical assessment of predicted protein interactions (CAPRI), despite the advances made in recent years [86], demonstrate the difficulty in reaching quantitative protein-

protein interaction predictions from unbound proteins. The conformational flexibility of proteins, demonstrated by their ability to explore quasi-stable states with different free energies in the bound and unbound states see Fig. (**1**) in ref. [24]) precludes rigid body docking schemas from producing quantitative predictions of protein-protein interactions. Thus, although a variety of primary amino acid sequences of polypeptides are compatible with amyloid formation, some basic properties such as conformational exploration, solubility, charge state, or the ability to pack in amyloid fibers, differently affect their propensity to aggregate [87]. This means that both dynamics and energetics need to be properly explored in molecular simulations of peptide aggregation, in the same way this is critical for enzyme reactivity analysis [12, 14]. In this sense it is interesting to recall here the difference between transient and permanent protein-protein interactions. The former, because of their lability, are produced by a delicate balance between protein-protein and protein-water interactions [88], and in many cases are mediated by water bridges. In the case of permanent interactions, because of their more hydrophobic nature [89], unbound proteins suffer from stronger strain energies and are thus expected to undergo wider conformational changes upon binding. Thus, in both situations, docking protocols that consider rigid fragments or implicit solvation at most, are expected to yield incorrect predictions of protein complex structures. It is, accordingly, necessary to turn the attention towards molecular simulations with some finer granularity and consideration of explicit water molecules or, at least, their directionality, entropic effects and correct electrostatics [35, 90].

Performing standard molecular dynamics simulations on protein aggregation is a challenge that, despite the amazing advances in computer architecture, is far from generating quantitative results comparable with experiments. Thus, as occurred in the field of enzymatic reactivity, smart algorithms for the integration of several layers of detail are needed. In the case of the computational evaluation of free energy profiles for enzymes, a central region of the problem should be treated with quantum detail while the rest of the system can be taken into account with mean field effects and different degrees of accuracy [72]. It would correspond to what Ayton *et al.* identifies as "serial multiscale simulations" [15]. When dealing with protein-protein interactions, we may think in a "parallel multiscale simulations" schema [15] in which different resolution replicas are taken into account simultaneously, or we can decide using a progressive focus during the simulation, starting with a coarse grain description of the two approaching moieties that is then transformed into an all-atom description when the objects get closer.

The contents of this review suggest that both approaches are going to be successful in the upcoming years, and we predict that the convergence of powerful machines and simulation algorithms, smart protocols for multiscaling and, although sometimes obscure by the technology but especially important, a deep understanding of the physico-chemical properties of protein and peptide interfaces [91] are going to produce accurate predictions of complex formation. In this direction, computational tools that are able to integrate different algorithms and facilitate the researcher's work are going to become central, as well as formats and protocols to share methods, systems and simulations. Basic science and applied technology are becoming obligated partners in other disciplines, and obviously molecular simulations will benefit from the same synergy in the years to come as well as they have been doing in the past.

ACKNOWLEDGMENTS

The authors acknowledge the continuous help of Michael A. Johnston in the development of Adun, which serves as a benchmark for tests of new algorithms in the lab, and to James Dalton for careful proof reading of the manuscript. This work has been partially funded by the virtual physiological human (VPH) NoE (FP7-ICT-2007-2-223920), the Spanish Ministry of Industry, Tourism and Commerce (TSI-020110-2009-431), the Spanish Ministry of Science and Innovation (BIO2008-04469-E; CTQ2008-00755; BFU2006-28430-E, and RETIC COMBIOMED RD07/0067/0001). CLA acknowledges the receipt of a fellowship by the CONICET and ND a partial fellowship by the Elstatik Foundation.

REFERENCES

[1] Dematte, L.; Prandi, D. GPU computing for systems biology. *Brief. Bioinform.,* **2010**, *11* (3), 323-333.

[2] Harvey, M.J.; Giupponi, G.; De Fabritiis, G. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.,* **2009**, *5* (6), 1632-1639.

[3] Buch, I.; Harvey, M.J.; Giorgino, T.; Anderson, D.P.; De Fabritiis, G. High-throughput all-atom molecular dynamics simulations using distributed computing. *J. Chem. Inf. Model.,* **2010**, *50* (3), 397-403.

[4] Johnston, M.A.; Villà-Freixa, J. Enabling data sharing and collaboration in complex systems applications. *Lect. Notes Bioinform.,* **2007**, *4360*, 124-140.

[5] Kurowski, K.; de Back, W.; Dubitzky, W.; Gulyás, L.; Kampis, G.; Mamonski, M.; Szemes, G.; Swain, M. In *Computational Science - ICCS 2009*, Computational Science-ICCS 2009, Heidelberg, Germany, Springer: Heidelberg, Germany, 2009; pp 387-396.

[6] MacNamara, S.; Burrage, K.; Sidje, R.B. Multiscale modeling of chemical kinetics via the master equation. *Multiscale Model. Simul.,* **2008**, *6* (4), 1146-1168.

[7] Cooper, J.; Cervenansky, F.; de Fabritiis, G.; Fenner, J.; Friboulet, D.; Giorgino, T.; Manos, S.; Martelli, Y.; Villà-Freixa, J.; Zasada, S.; Lloyd, S.; McCormack, K.; Coveney, P.V. The Virtual Physiological Human Toolkit. *Phil. Trans. R. Soc. A,* **2010**, *368*, 3925-3936.

[8] Miller, K. Bringing the fruits of computation to bear on human health: Its a tough job, but the NIH has to do it. *Biomed. Comput. Rev.,* **2009**, *5* (2), 18-28.

[9] Kitano, H. Computational systems biology. *Nature,* **2002**, *420* (6912), 206-10.

[10] Zhang, J.; Li, W.; Wang, J.; Qin, M.; Wu, L.; Yan, Z.; Xu, W.; Zuo, G.; Wang, W. Protein folding simulations: from coarse-grained model to all-atom model. *IUBMB Life,* **2009**, *61* (6), 627-43.

[11] Guallar, V.; Wallrapp, F.H. QM/MM methods: Looking inside heme proteins biochemisty. *Biophys. Chem.,* **2010**, *149* (1-2), 1 - 11.

[12] Kamerlin, S.C.L.; Warshel, A. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins: Struct., Funct., Bioinf.,* **2010**, *78* (6), 1339-1375.

[13] Truhlar, D.G.; Gao, J.; Alhambra, C.; Garcia-Viloca, M.; Corchado, J.e.; Sánchez, M.L.; Villà, J. The incorporation of quantum effects in enzyme kinetics modeling. *Acc. Chem. Res.,* **2002**, *35* (6), 341-349.

[14] Villà, J.; Warshel, A. Energetics and dynamics of enzymatic reactions. *J. Phys. Chem. B,* **2001**, *105* (33), 7887-7907.

[15] Ayton, G.S.; Noid, W.G.; Voth, G.A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr. Opin. Struct. Biol.,* **2007**, *17* (2), 192-8.

[16] Ayton, G.S.; Voth, G.A. Systematic multiscale simulation of membrane protein systems. *Curr. Opin. Struct. Biol.,* **2009**, *19* (2), 138-44.

[17] Burykin, A.; Kato, M.; Warshel, A. Exploring the origin of the ion selectivity of the KcsA potassium channel. *Proteins: Structure, Function, and Bioinformatics,* **2003**, *52* (3), 412-426.

[18] Olsson, M.H.M.; Warshel, A. Monte Carlo simulations of proton pumps: On the working principles of the biological valve that controls proton pumping in cytochrome c oxidase. *Proceedings of the National Academy of Sciences,* **2006**, *103* (17), 6500-6505.

[19] Serohijos, A.W.R.; Tsygankov, D.; Liu, S.; Elston, T.C.; Dokholyan, N.V. Multiscale approaches for studying energy transduction in dynein. *Phys. Chem. Chem. Phys.,* **2009**, *11* (24), 4840-50.

[20] Gillespie, D.T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.,* **1977**, *81* (25), 2340-2361.

[21] Gillespie, D.T. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.,* **2001**, *115* (4), 1716.

[22] Rué, P.; Villà-Freixa, J.; Burrage, K. Simulation Methods with Extended Stability for Stiff Biochemical Kinetics. *BMC Syst. Biol.,* **2010**, *4*, 110.

[23] Bicout, D.J.; Field, M.J. Stochastic dynamics simulations of macromolecular diffusion in a model of the cytoplasm of Escherichia coli. *J. Phys. Chem.,* **1996**, *100* (7), 2489-2497.

[24] Bonet, J.; Caltabiano, G.; Khan, A.K.; Johnston, M.A.; Corbi, C.; Gomez, A.; Rovira, X.; Teyra, J.; Villà-Freixa, J. The role of residue stability in transient protein-protein interactions involved in enzymatic phosphate hydrolysis. A computational study. *Proteins: Struct., Funct., Bioinf.,* **2006**, *63* (1), 65-77.

[25] Scheper, J.; Oliva, B.; Villà-Freixa, J.; Thomson, T.M. Analysis of electrostatic contributions to the selectivity of interactions between RING-finger domains and ubiquitin-conjugating enzymes. *Proteins,* **2009**, *74* (1), 92-103.

[26] Hall, D.; Hirota, N. Multi-scale modelling of amyloid formation from unfolded proteins using a set of theory derived rate constants. *Biophys. Chem.,* **2009**, *140* (1-3), 122 - 128.

[27] Urbanc, B.; Cruz, L.; Ding, F.; Sammond, D.; Khare, S.; Buldyrev, S.V.; Stanley, H.E.; Dokholyan, N.V. Molecular dynamics simulation of Amyloid β dimer formation. *Biophys. J.,* **2004**, *87* (4), 2310 - 2321.

[28] Waxman, E.A.; Mazzulli, J.R.; Giasson, B.I. Characterization of hydrophobic residue requirements for alpha-synuclein fibrillization. *Biochemistry,* **2009**, *48* (40), 9427-36.

[29] Dill, K.A.; Ozkan, S.B.; Shell, M.S.; Weikl, T.R. The protein folding problem. *Annu. Rev. Biophys.,* **2008**, *37* (1), 289-316.

[30] Mousseau, N.; Derreumaux, P. Exploring energy landscapes of protein folding and aggregation. *Front. Biosci.,* **2008**, *13*, 4495-516.

[31] Leach, A.R. *Molecular modelling: principles and applications*. Addison-Wesley Longman Ltd: 2001.

[32] Warshel, A.; Sharma, P.K.; Kato, M.; Parson, W.W. Modeling electrostatic effects in proteins. *Biochim. Biophys. Acta,* **2006**, *1764* (11), 1647-1676.

[33] Warshel, A.; Papazyan, A. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol.,* **1998**, *8* (2), 211-217.

[34] Chen, J.; Brooks, C.L. Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys. Chem. Chem. Phys.,* **2008**, *10* (4), 471-81.

[35] Fennell, C.J.; Dill, K.A. Oil/Water transfer is partly driven by molecular shape, not just size. *J. Am. Chem. Soc.,* **2010**, *132* (1), 234-240.

[36] Gilson, M.K.; Zhou, H.-X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.,* **2007**, *36*, 21-42.

[37] Grochowski, P.l.; Trylska, J. Continuum molecular electrostatics, salt effects, and counterion binding-a review of the Poisson-Boltzmann theory and its modifications. *Biopolymers,* **2008**, *89* (2), 93-113.

[38] Lange, A.W.; Herbert, J.M. Polarizable continuum reaction-field solvation models affording smooth potential energy surfaces. *J. Phys. Chem. Lett.,* **2010**, *1* (2), 556-561.

[39] Schutz, C.N.; Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins: Struct., Funct., Bioinf.,* **2001**, *44* (4), 400-417.

[40] Sham, Y.Y.; Chu, Z.T.; Warshel, A. Consistent calculations of pKa's of ionizable residues in proteins: semi-microscopic and microscopic approaches. *J. Phys. Chem. B,* **1997**, *101* (22), 4458-4472.

[41] Messer, B.M.; Roca, M.; Chu, Z.T.; Vicatos, S.; Vardi-Kilshtain, A.; Warshel, A. Multiscale simulations of protein landscapes: using coarse-grained models as reference potentials to full explicit models. *Proteins: Struct., Funct., Bioinf.,* **2010**, *78* (5), 1212-27.

[42] Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.,* **2005**, *15* (2), 144-150.

[43] Tozzini, V. Multiscale modeling of proteins. *Acc. Chem. Res.,* **2010**, *43* (2), 220-230.

[44] Zhou, J.; Thorpe, I.F.; Izvekov, S.; Voth, G.A. Coarse-grained peptide modeling using a systematic multiscale approach. *Biophys. J.,* **2007**, *92* (12), 4289-4303.

[45] Thorpe, I.F.; Zhou, J.; Voth, G.A. Peptide folding using multiscale coarse-grained models. *J. Phys. Chem. B,* **2008**, *112* (41), 13079-90.

[46] Marrink, S.J.; Risselada, H.J.; Yefimov, S.; Tieleman, D.P.; De Vries, A.H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B,* **2007**, *111* (27), 7812-24.

[47] Cui, Q.; Bahar, I. Normal Mode Analysis: Theory and applications to biological and chemical systems. CRC press: Dordrecht, Netherlands, 2006.

[48] Maupetit, J.; Tuffery, P.; Derreumaux, P. A coarse-grained protein force field for folding and structure prediction. *Proteins: Struct., Funct., Bioinf.,* **2007**, *69* (2), 394-408.

[49] Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature,* **1975**, *253* (5494), 694-698.

[50] Aloy, P.; Oliva, B. Splitting Statistical Potentials into meaningful scoring functions: Testing the prediction of near-native structures from decoy conformations. *BMC Struct. Biol.,* **2009**, *9* (1), 71.

[51] Kolinski, A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.,* **2004**, *51* (2), 349-371.

[52] Malolepsza, E.; Boniecki, M.; Kolinski, A.; Piela, L. Theoretical model of prion propagation: A misfolded protein induces misfolding. *Proc. Natl. Acad. Sci. USA.,* **2005**, *102* (22), 7835-7840.

[53] Boniecki, M.; Rotkiewicz, P.; Skolnick, J.; Kolinski, A. Protein fragment reconstruction using various modeling techniques. *J. Comput. Aided Mol. Des.,* **2003**, *17* (11), 725-738.

[54] Wolynes, P.G. Recent successes of the energy landscape theory of protein folding and function. *Q. Rev. Biophys.,* **2005**, *38* (4), 405-10.

[55] Huang, S.-W.; Yu, S.-H.; Guan, H.-W.; Shih, C.-H.; Huang, T.-T.; Hwang, J.-K. On the relationship between catalytic residues and their protein contact number. 2010.

[56] Matysiak, S.; Clementi, C. Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: how far can a minimalist model go? *J. Mol. Biol.,* **2004**, *343* (1), 235-248.

[57] Lyubartsev, A.P.; Laaksonen, A. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys. Rev. E,* **1995**, *52* (4), 3730-3737.

[58] Das, P.; Matysiak, S.; Clementi, C. Balancing energy and entropy: a minimalist model for the characterization of protein folding landscapes. *Proc. Natl. Acad. Sci. USA,* **2005**, *102* (29), 10141-10146.

[59] Matysiak, S.; Clementi, C., Minimalist protein model as a diagnostic tool for misfolding and aggregation. *J. Mol. Biol.,* **2006**, *363* (1), 297-308.

[60] Baumketner, A.; Shea, J.-E. The Structure of the Alzheimer Amyloid β 10-35 peptide probed through replica-exchange molecular dynamics simulations in explicit solvent. *J. Mol. Biol.,* **2007**, *366* (1), 275-285.

[61] Li, M.S.; Klimov, D.K.; Straub, J.E.; Thirumalai, D. Probing the mechanisms of fibril formation using lattice models. *J. Chem. Phys.,* **2008**, *129* (17), 175101.

[62] Ma, B.; Nussinov, R. The stability of monomeric intermediates controls amyloid formation: Aβ 25-35 and its N27Q mutant. *Biophys. J.,* **2006**, *90* (10), 3365 - 3374.

[63] Tarus, B.; Straub, J.E.; Thirumalai, D. Dynamics of Asp23-Lys28 salt-bridge formation in Aβ 10-35 monomers. *J. Am. Chem. Soc.,* **2006**, *128* (50), 16159-16168.

[64] Das, R.; André, I.; Shen, Y.; Wu, Y.; Lemak, A.; Bansal, S.; Arrowsmith, C.H.; Szyperski, T.; Baker, D. Simultaneous prediction of protein folding and docking at high resolution. *Proceedings of the National Academy of Sciences,* **2009**, *106* (45), 18978-18983.

[65] Malek, R.; Mousseau, N. Dynamics of Lennard-Jones clusters: A characterization of the activation-relaxation technique. *Phys. Rev. E,* **2000**, *62* (6), 7723-7728.

[66] Truhlar, D.G.; Garrett, B.C.; Klippenstein, S.J. Current status of transition-state theory. *J. Phys. Chem,* **1996**, *100* (31), 12771-12800.

[67] Villà, J.; Truhlar, D.G. Variational transition state theory without the minimum energy path. *Theor. Chem. Acc.,* **1997**, *1-4*, 317-323.

[68] Lyman, E.; Ytreberg, F.M.; Zuckerman, D.M. Resolution exchange simulation. *Phys. Rev. Lett.,* **2006**, *96* (2), 028105.

[69] Lyman, E.; Zuckerman, D.M. Resolution exchange simulation with incremental coarsening. *J. Chem. Theory Comput.,* **2006**, *2* (3), 656-666.

[70] Hu, H.; Yang, W. Free energies of chemical reactions in solution and in enzymes with ab initio quantum mechanics/molecular mechanics methods. *Annu. Rev. Phys. Chem.,* **2008**, *59* (1), 573-601.

[71] Sherwood, P.; Brooks, B.R.; Sansom, M.S.P. Multiscale methods for macromolecular simulations. *Curr. Opin. Struct. Biol.,* **2008**, *18* (5), 630-40.

[72] Warshel, A.; Levitt, M. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.,* **1976**, *103* (2), 227-249.

[73] Heath, A.P. Towards multiscale protein simulations: Moving from coarse-grain to all-atom models. Rice University, 2006.

[74] Heath, A.P.; Kavraki, L.E.; Clementi, C. From coarse-grain to all-atom: toward multiscale analysis of protein landscapes. *Proteins: Struct., Funct., Bioinf.,* **2007**, *68* (3), 646-61.

[75] Vorobjev, Y.N.; Hermans, J. ES/IS: Estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. *Biophysical Chemistry,* **1999**, *78* (1-2), 195 - 205.

[76] Kurcinski, M.; Kolinski, A. Hierarchical modeling of protein interactions. *J. Mol. Model.,* **2007**, *13* (6), 691-698.

[77] Gront, D.; Kmiecik, S.; Kolinski, A. Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J. Comput. Chem.,* **2007**, *28* (9), 1593-1597.

[78] Vorobjev, Y.N.; Almagro, J.C.; Hermans, J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. *Proteins: Struct., Funct., Genet.,* **1998**, (4), 399-413.

[79] Hawkins, G.D.; Cramer, C.J.; Truhlar, D.G. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.,* **1996**, *100* (51), 19824-19839.

[80] Im, W.; Lee, M.S.; Brooks, C.L. Generalized Born model with a simple smoothing function. *J. Comput. Chem.,* **2003**, *24* (14), 1691-1702.

[81] Onufriev, A.; Bashford, D.; Case, D.A. Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins: Struct., Funct., Bioinf.,* **2004**, *55* (2), 383-394.

[82] Fennell, C.J.; Kehoe, C.; Dill, K.A. SEA water: modeling molecular solvation using semi-explicit first-shell water.(Manuscript under preparation)

[83] Liu, P.; Voth, G.A., Smart resolution replica exchange: An efficient algorithm for exploring complex energy landscapes. *J. Chem. Phys.,* **2007**, *126* (4), 045106.

[84] Fan, Z.Z.; Hwang, J.K.; Warshel, A. Using simplified protein representation as a reference potential for all-atom calculations of folding free energy. *Theor. Chim. Acta,* **1999**, *103* (1), 77-80.

[85] Johnston, M.A.; Galván, I.F.; Villà-Freixa, J. Framework-based design of a new all-purpose molecular simulation application: the Adun simulator. *J. Comput. Chem.,* **2005**, *26* (15), 1647-1659.

[86] Wodak, S.J. From the Mediterranean coast to the shores of Lake Ontario: CAPRI's premiere on the American continent. *Proteins: Struct. Funct. Bioinf.,* **2007**, *69* (4), 697-698.

[87] Pawar, A.; Dubay, K.; Zurdo, J.; Chiti, F.; Vendruscolo, M.; Dobson, C. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.,* **2005**, *350* (2), 379-392.

[88] Shoichet, B.K. No free energy lunch. *Nat. Biotechnol.,* **2007**, *25* (10), 1109-1110.

[89] Nooren, I.M.; Thornton, J.M. Diversity of protein-protein interactions. *EMBO J.,* **2003**, *22* (14), 3486-3492.

[90] Singh, N.; Warshel, A. Absolute binding free energy calculations: On the accuracy of computational scoring of protein-ligand interactions. *Proteins: Struct., Funct., Bioinf.,* **2010**, *78* (7), 1705-1723.

[91] Barbany, M.; Gutiérrez-de-Terán, H.; Sanz, F.; Villà-Freixa, J. Towards a MIP-based alignment and docking in computer-aided drug design. *Proteins,* **2004**, *56* (3), 585-594.

Received: December 24, 2010 **Accepted: January 24, 2011**