There's plenty of room in the middle: Multiscale Modelling of Biological Systems

Marc Baaden, Richard Lavery

THERE'S PLENTY OF ROOM IN THE MIDDLE: MULTI-SCALE MODELLING OF BIOLOGIGAL SYSTEMS

Marc Baaden and Richard Lavery[#]

Laboratoire de Biochimie Théorique, CNRS UPR 9080,
Institut de Biologie Physico-Chimique,
13, rue Pierre et Marie Curie,
F-75005 Paris,
France

[#]To whom correspondence should be addressed

E-mail: rlavery@ibpc.fr
Tel: +33-15841-5016
Fax: +33-15841-5026

(to appear in *Recent Advances in Protein engineering*)

Running Title: Multi-scale modelling

ABSTRACT
Understanding proteins well enough to rationally modify their biological function requires understanding how these biomolecules behave in their natural cellular, or extra-cellular, environments. This, in turn, implies understanding their interactions with a wide variety of other species within a dense and heterogeneous medium. Molecular modelling and simulation can certainly contribute to improving our understanding in this area, however the range of molecules and processes involved in the biological systems implies that a range of modelling techniques will have to be applied, balancing the requirements for accuracy and precision against the constraints imposed by the size, time and energy scales involved. This article attempts to summarize the various representations, methodologies and target functions presently available to molecular modellers, discusses how different combinations of these basic features can be combined to attack different problems and then considers the role of hybrid methods, an area where there is still much scope for development.

INTRODUCTION

Proteins have evolved to function within living organisms, which probably represent the most heterogeneous and complex dynamical systems known to mankind. Beyond the astonishing range of structures exhibited by proteins themselves, these systems are composed of a broad range of other molecules which, according to their roles within the system, bind to one another to form functional assemblies, interact transiently to pass messages inside and outside the cell, store and convert energy, or catalyze chemical reactions with remarkable specificity.

All of these processes are characterized by three features which make them very challenging to understand. First, although simple molecular species, such as water molecules and ions play important roles in biological systems, most processes also involve macromolecules or macromolecular assemblies which can contain hundreds of thousands of atoms. Second, biological systems are by definition dynamic, with important time scales ranging from femtoseconds, the characteristic vibrational frequency of C-H bonds, up to seconds or minutes, for processes as varied as folding individual proteins or accomplishing major cellular rearrangements. Third, biological systems are "soft matter", where conformation and binding result from non-covalent interactions (van der Waals, hydrogen bonds, ...) and are generally characterized by free energy changes which amount to only a few kcal mol^{-1} (although the compensating enthalpic and entropic components which underlie these changes can amount to hundreds of kcal mol^{-1}).

Experimental methods to analyze cellular processes at a molecular level are progressing very rapidly and have resulted not only in fully sequenced genomes for a wide range of organisms (1), but also in detailed cellular level data ranging from gene expression patterns to protein interaction networks. Systems biology approaches are already attempting to use such data to build mathematical models of cellular processes, and the notion of being able to create and understand a "minimal" cell has moved from the realm of science fiction to that of a scientific project (2, 3). Despite this progress, the last half century of molecular biology has clearly demonstrated that a full understanding of biological processes also requires an understanding of the structures and interactions of the molecules on which these processes depend. This is particularly true in the case of proteins where properties such as binding selectivity, thermodynamic efficiency, thermal and mechanical stability all depend on detailed structure and dynamics. This implies that both structural biology and modelling will continue to play a significant role in the current biological revolution. However, the new data available also implies that we need to think more in terms of assemblies and interacting systems than in terms of individual macromolecules, if we want to modulate a biological

function. Modelling must therefore rise to the challenge of treating much larger and more complex problems.

Given the characteristics of the systems described above, there can clearly be no unique way to model all biological objects or processes. This fact is illustrated by the way different communities look at the same biological entity. If we put aside proteins for a moment and consider the example of the iconic DNA double helix, a geneticist or bioinformatician will often be interested only in its base sequence; a physicist may ignore its sequence and concentrate only on its macroscopic elastic properties; a molecular modeller may study only its microscopic structure and dynamics and a systems biologist may see it as a set of interaction nodes in a coupled reaction scheme. Similar examples could be found for other components of the cellular machinery.

Even for the molecular modeller, the task is not simple. Despite progress in modelling algorithms and remarkable increases in computer power, it remains impossible to model all molecular processes with a single approach. Even if quantum mechanics is clearly the appropriate tool for studying atomic-level processes, this approach has to be abandoned in favour of classical mechanics when the systems studied become too large or the time scales too long. Similarly, atomic detail has to be sacrificed when we pass to even larger systems, in favour of coarse-grain representations or even continuum models. These problems become still more pressing when our attention passes from single macromolecules to macromolecular interactions, and then to the formation of multi-macromolecular assemblies or to coupled sets of macromolecular interactions. Modelling biological systems is therefore an intrinsically multi-scale task.

A rapid survey of molecular modelling and simulation over the last few decades demonstrates that these difficulties have led to the development of a wide range of approaches based on different molecular representations and using different modelling algorithms. Each approach, like a biological species, has its own niche, defined in terms of the characteristics discussed above (size, time and energy scales) where it usefully can serve to push forward our understanding. Like biological species, modelling approaches should not be dragged too far away from their natural environments if they are to perform well, but equally, they can sometimes thrive by interactions which extend their range.

We would like to take this opportunity to attempt to summarize the present state of affairs in biological modelling in the hope that this can help those interested in better understanding the function of biological molecules, and, notably, in re-engineering protein function, to take a broader view of the field and the tools available. We will begin by summarizing what is available today in terms of representations (R) and modelling algorithms (M) in the form of comparative tables. This classification is necessarily somewhat subjective and involves considerable simplifications, but it gives an idea of the range of approaches that can be called upon and of their strengths and weaknesses. In many cases, these tables group together a family of related approaches within a single column, although this may conceal a good deal of research and a number of significant refinements. Our apologies to the authors concerned.

Before discussing how representations and methodologies can be put together, we briefly discuss the target functions used with various modelling approaches. Having set out the basic choices available, we then turn to the question of how different approaches can be combined to hybrid models. This last analysis leads us to conclude that there are still plenty of possibilities for developing new hybrid approaches, and explains the first part of the title of this review, a misquote of Richard Feynman's famous 1959 APS lecture heralding nanotechnology. Lastly, it should be noted that, given the broad scope of this review, we have only been able to include a few references to illustrate each section. The reader is referred to more specialized articles to complete this bibliography.

REPRESENTATIONS

In order to simplify comparisons between the different possible representations used for modelling biological systems, we will consider applications to a standard test case consisting of a single macromolecule containing S monomers and N atoms. For proteins, S will typically range from a 100 to a 1000 and N is roughly 10S. For a nucleic acid, S will typically range from 10-10000, although chromosomal DNA's can be much longer, and N is roughly 30S for a single-stranded fragment. For membrane lipids, a single molecule contains roughly 100 to 300 atoms, and a 25 $Å^2$ fragment of a bilayer membrane will contain as many atoms as a reasonably sized protein. Depending on the representation these biological systems are modelled using a varying number of interaction points P.

Each representation occupies one column of Table 1, while its characteristics are given in the rows of the table. These include a brief description of the representation; the corresponding number of conformational degrees of freedom (DOF) for our standard test case; the principle advantages and limitations of the representation and a selection of keywords. The keywords provide a link between commonly used expressions in the literature and the classifications adopted in this article. Note that in many cases the keywords refer to the combinations of representations and methodologies which are discussed in the following section of this article.

Table 1	R1 Quantum mechanical molecule	R2 All-atom models	R3 United atom models
Description	Nuclei and electrons are treated explicitly	Each atom of the system is represented by a single interaction point	While basically maintaining an atomic representation, certain atoms are grouped into single pseudoatoms (notably groups carrying non-polar hydrogens such as CH, CH_2 and CH_3)
DOF	3N-6 for the nuclei, plus the electronic degrees of freedom	3N-6 in Cartesian coordinates, roughly 10x less in internal coordinates, if bond lengths are frozen	Up to 2-3x less than all-atom models
Advantages	Explicit treatment of electrons enables both ground and excited electronic states to be described, allows studies of chemical reactions (bond making and breaking, electron transfer, etc.), and incorporates effects such as polarization and charge transfer	The most detailed representation in classical molecular mechanics, corresponding to the resolution of most experimental biomolecular structures	Reduced number of interaction points
Limitations	Computational cost. Electronic correlation required for weak van der Waals interactions. Level of theory (basis set, ..) needs to be adapted to the problem.	Continuous electron density is partitioned onto the nuclear coordinates leading to partial atomic charges. No chemistry. No notion of electronic excited states. Polarization and charge transfer require special treatments	Loss of resolution. Time gain depends strongly on system
Keywords	*Ab initio*, Hartree-Fock, Density Functional Theory, Semi-empirical methods	All-atom model, explicit atomic representation	United atom model

Table 1	R4 Bead models	R5 Lattice models	R6 Jointed chain models
Description	These models go farther than united atom models by replacing entire groups, subunits or multiple subunits with single beads. Beads may move in Cartesian space or in partially constrained internal coordinate space	Lattice models are related to polymeric bead models, but each bead occupies a node within a regular 2D or 3D lattice. No two beads can occupy the same positions and chain crossings are forbidden	Includes polymers modelled as jointed segments of fixed length and discrete versions of elastic rod models with deformation nodes controlled by elastic constants
DOF	*O*(S) to *O*(3S)	Depending on the lattice each new bead can be added in a finite number of positions (3 in a 2D cubic lattice and 5 in a 3D cubic lattice).	*O*(S)
Advantages	Significant reduction in number of interaction points	The finite nature of the lattice makes it possible to enumerate and study all possible conformations of the polymer	Rapid calculations and, for the simplest versions, analytical solutions for properties such as persistence length, radius of gyration, etc
Limitations	Loss of resolution and reduced internal flexibility. Specific interactions (e.g. hydrogen bonds) must be treated implicitly	"Toy" representation of biopolymers in a highly simplified conformational space	Low resolution. Chain crossing can occur unless excluded volume effects are included
Keywords	Coarse-grained representation, pseudo-atoms, super-atoms, elastic network model, Debye sphere model, hydrodynamic bead model	Go model, Hydrophobic-Polar (HP) potentials	Freely-jointed chain (FJC) model

Table 1	R7 Elastic rods and sheets	R8 Surface / volume representations	R9 Implicit solvent / environment models
Description	Continuum elastic models in 1D (rods) or 2D (sheets). Models are mostly isotropic, but rods with laterally anisotropic bending and anisotropic membranes have been studied. Such models are used for semi-rigid polymers like dsDNA and for modelling lipid membranes	Complex macromolecular shapes can be reduced to tessellated surfaces (e.g. with Voronoi polyhedra) or represented by a set of monocentric or multicentric functions such as spherical harmonics	Solvent effects on a solute can be modelled using mobile or fixed polarizable particles (typically dipoles) or as a continuous dielectric medium surrounding a cavity containing the solute. Similar approaches can be applied to membrane environments
DOF	Objects are continuously deformable. Elastic constants can be varied	6 per object (overall translation and rotation)	Simple physico-chemical parameters + grid/particle density and dipole magnitude and orientation for discrete representations
Advantages	Analytical solutions are possible. Excluded volume and electrostatic interactions can be added	Relatively refined surface/volume representations can be obtained with few parameters compared to atomic representations	Avoids treating molecular environments (solvent, lipid bilayer, ...) explicitly. Cavity terms can treat surface tension effects
Limitations	Interactions are only local. Elastic properties can be modified by local or global structural transitions. Heterogeneity along rods or within sheets cannot be treated analytically	Reduced resolution. No flexibility	Does not account for specific solute interactions with the environment
Keywords	Freely-rotating chain (FRC), worm-like chain (WLC), continuum rod model, continuum sheet model	Spherical harmonics, Voronoi polyhedra	Poisson-Boltzmann, generalized Born, reaction Field, PCM model, COSMO model, Langevin dipole model

METHODOLOGIES

Each methodology is presented within the columns of Table 2, while the rows of the table provide a brief description of the methodology; the nature of the input and output data required; its principal advantages and limitations and related keywords.

Table 2	M1 Surface complementarity search	M2 Discrete conformational search	M3 Energy minimization
Description	Take two molecules – often a small ligand and a macromolecular receptor – and efficiently search for a putative interaction site based on conformational complementarity	Generate and score an ensemble of conformations of a molecular system in the space defined by its DOF. Conformations can be generated stochastically or via a grid search	Optimize the conformation of a molecular system to find a local energy minimum in the space defined by its DOF
Input	A conformation for each molecule, a procedure to search the six-dimensional space spanned by their relative positions and a way to score the resulting conformations	A starting conformation, a conformational score, optional restraints	A starting conformation, the conformational energy and at least the first derivatives of this energy with respect to the DOF
Output	Docked conformations and their complementarity scores	Conformations and their scores	A single conformation and its corresponding energy
Advantages	Fast calculation, particular with algorithms based on Fourier transformation techniques	Fast calculations, but restricted to a limited number of DOF	Relatively fast calculations. The quality of the minimum (judged by the magnitude of the final gradients) can be adjusted as required
Limitations	Internal flexibility is difficult to treat, and results depend critically on the chosen conformations of the interacting molecules	An exhaustive search is impossible in most practical cases. In a regular grid search is limited by the grid spacing.	Will locate a local minimum close to the starting conformation. (How close depends on the representation of the molecular system). Reduced coordinates (R3-R4) lead to smoother energy hypersurfaces and generally allow larger conformational changes
Keywords	Rigid-body docking	Conformational search, Grid search, Concoord	Simplex, Steepest Descent, Conjugate Gradient (Broyden-Fletcher-Goldfarb-Shanno, Fletcher-Reeves, Polak-Ribiere), Newton-Raphson

Table 2	M4 Activated / stochastic optimization	M5 Normal mode analysis	M6 Molecular Dynamics
Description	Advanced optimization methods can overcome the problems of getting trapped in local minima. Both stochastic and analytic approaches exist	Determine the harmonic vibrational modes and associated frequencies of a molecular system	Integrate Newton's classical equations of motion to calculate the trajectory of a molecular system in phase space (coordinates and velocities)
Input	Starting conformation, an optimization scheme and related parameters, energy and, in some cases, energy derivatives	An energy minimized molecular conformation, the ability to calculate the mass-weighted Hessian matrix (second derivatives of the conformational energy with respect to the DOF)	A molecular starting conformation, the ability to calculate its conformational energy and the first derivatives with respect to the DOF. Fix boundary and thermodynamic conditions
Output	A ranked population of optimized conformations	(3N-6) eigenvectors and the corresponding eigenvalues	Time series of coordinates, velocities and energies
Advantages	Escape from local minima	Rapidity. Low-frequency modes are potentially related to biological function. Vibrational entropy can be calculated. For large systems full diagonalization can be avoided	Kinetic energy allows overcoming barriers. Ergodicity implies the trajectory can be analyzed using statistical mechanics. Time series can yield kinetic data
Limitations	Appropriate choice of evolving variables and algorithmic parameters is important. More costly than simple minimizations	Requires a high quality minimum to avoid imaginary frequencies. The harmonic approximation is valid only for relatively small motions	Adequate sampling is difficult. The fastest molecular motions limit the integration timestep to the order of a few femtoseconds
Keywords	Genetic algorithms, simulated annealing, Activation-Relaxation Technique (ART)	Vibrational analysis, Gaussian Network model (GNM), Lanczos algorithms	Molecular Dynamics (MD), Essential Dynamics, Replica Exchange, Car-Parrinello, Targeted MD

Table 2	M7 Stochastic dynamics	M8 Monte Carlo	M9 Analytic / specific numerical solutions
Description	Brownian, dissipative and hydrodynamic forces are taken into account using stochastic force terms. Can involve modified soft-repulsive interactions	Sample an ensemble of molecular conformations satisfying Boltzmann statistics	Use an analytic or numerical approach to solve a mathematical description of a model. This often involves solving differential equations
Input	A molecular starting conformation, the ability to calculate its conformational energy (and possibly energy derivatives). Mean field environmental characteristics (friction, viscosity,..)	A molecular starting conformation, the ability to calculate its conformational energy. An efficient sampling scheme such as the Metropolis algorithm (at a defined temperature)	Boundary conditions, desired level of accuracy, specific parameters
Output	Time series of coordinates, velocities and energies	An ensemble of conformations and the associated energies	A solution to the mathematical problem
Advantages	Explicit sampling of the medium is either not necessary as it is represented via the friction and random terms or significantly sped-up by the use of large integration timesteps. Motions requiring large solvent rearrangements are better sampled	Well-devised moves allow efficient sampling and convergence. In contrast to molecular dynamics, forces are not usually required	No inherent limitations of the underlying models. Direct relation between parameters and results
Limitations	With additional force terms explicit interactions between the medium and the molecular system are lost. When the medium is represented by a PMF it is static and does not evolve; flexibility is difficult to include. Using exclusively “soft” forces, information on conformational changes is lost.	Moves can be difficult to devise and are performed sequentially. Motions requiring two or more moves to occur simultaneously are impossible. An ensemble of conformations, not a consecutive trajectory is obtained	Mathematical complexity. Specific problems may require new models. The validity of the model has to be assumed
Keywords	Langevin Dynamics, Brownian Dynamics, Dissipative Particle Dynamics	Monte Carlo, Parallel Tempering, Replica Exchange MC	Finite-difference solutions like in Poisson-Boltzmann calculations, elastic rod models, Helfrich (and other) membrane models

TARGET FUNCTIONS

In order to carry out calculations with a chosen representation and methodology we must also specify the target function that is going to be calculated. This function is easiest to define, but not necessarily to calculate, in the case of the quantum mechanical representation for which it is possible to obtain the formation energy of a molecule or molecular system. If enough computer resources are available, and if the system is not too large, it may be possible to achieve experimental precision for both ground and excited state properties. Although it is possible to take electron correlation into effect, this is often prohibitively expensive for large systems. Density functional approaches, which are often used in such cases, only partially treat correlation and thus underestimate the van der Waals interactions which play a significant role in stabilizing biological systems. Quantum mechanical approaches are also limited by the quality of the basis set or density functional used.

For systems containing more than a few tens of heavy (non-hydrogen) atoms it generally becomes necessary to introduce further approximations. These can either involve dividing the system into overlapping fragments (4-6), leading to energy calculations which scale linearly with the number of electrons in the system (rather than at least the cube of this number), or neglecting certain categories of electronic integrals. The latter approaches are termed semiempirical methods (CNDO, INDO, MINDO, AM1, PM3, ...), and require varying degrees of parameterization. Although they often perform well for a variety of molecules, they can also fail badly for specific cases and they remain computationally expensive for large molecular systems. It should be added that, unlike the approaches discussed below, quantum mechanical approaches can provide not only energies, but also a whole range of other properties such as electronic densities, bond orders, polarizabilities, electron affinities, ionization potentials, NMR shielding constants, etc.

The next stage of simplification leads us to so-called force fields, which are the most common solution for all-atom representations of large systems. Force fields are loosely based on a perturbation analysis of quantum mechanical energies (7, 8), they however calculate only conformational energies and not formation energies. Energy values can therefore only be compared for chemically identical systems (having the same number of atoms, bonded in the same manner). This also implies that using force fields excludes all aspects of chemistry, including making or breaking bonds, transferring electrons or protons, or even creating excited states. Force fields are based on the Born-Oppenheimer approximation, which assumes that, since electrons move much faster than atomic nuclei, this motion can be averaged out and the nuclei can be assumed to move in an averaged electronic density. Most force fields go further in assuming that the density distribution can be divided into atomic contributions, contracted onto the nuclei and summed with the nuclear charges to yield partial atomic charges. It is also assumed that the conformational energy can be broken down into a series of additive terms. These terms typically involve so-called bonded interactions, representing the energy penalties linked to deforming bond lengths, valence angles or torsion angles, and non-bonded interactions, covering electrostatic interactions, short-range exchange repulsion and attractive van der Waals dispersion. Non-bonded interactions are generally dealt with using pairwise additive terms, Coulomb's law being used for interactions between atomic partial charges and the so-called Lennard-Jones term being used for repulsion-dispersion interactions. In addition to calculating atomic partial charges, generally from quantum mechanical calculations on representative molecular fragments, force fields require the determination of a very large number of parameters. This number is related to the number of atomic "classes" which are defined. (Classes allow the chemical environment of given atom types to be taken into account, e.g. trigonal versus tetrahedral carbon atoms). This choice is linked to the accuracy required and the chemical variety of molecules that are to be treated.

For biological molecules, a great deal of effort on the part of the force field builders, and feedback from the user community, have led to good overall results in treating the conformations and interactions of the most common biological molecules, although defining error bars for specific calculations remains difficult. Refinements generally require an overhaul of the full set of parameters, despite the apparent independence of the energy terms (9). For the same reason, parameter sets from different force fields cannot generally be mixed. Major improvements being worked on today include more accurate representations of the electronic density and treatment of electronic polarization. The computational cost of force field evaluation is mainly linked to the non-bonded terms, whose number scales as N^2, although methods for dealing with long-range electrostatic interactions can reduce this dependence to N.log(N) (10).

Despite their apparently physical basis, force fields do not perform well in all cases. This is notably the case when computational restrictions prevent environmental effects (solvent and counter ions) and/or entropic contributions from being taken into account accurately. Simple force field energies are thus not a good guide to the stability of folded proteins. This difficulty can be overcome by incorporating the missing factors in effective potentials derived from experimental data. This is the approach adopted by so-called knowledge-based potentials commonly used for identifying the most likely folds of polypeptide chains (11) or, more recently, for predicting the stability of biomolecular interactions (12, 13). These potentials are generally limited to residue-residue interactions, although they can be formulated on an atom or atomic-group basis if enough data is available for parameterization. This approach assumes that the database of experimental structures represents an equilibrium Boltzmann distribution of conformations (although this is difficult to prove) and can therefore yield effective mean force potentials (14). Knowledge-based potentials are rapid to evaluate, but since they generally do not have analytic forms, energy derivatives must be obtained numerically. Derivatives are however unnecessary for comparing different static molecular conformations as in so-called threading approaches (see below).

When less experimental data is available, or when data from different sources needs to be treated together, it is possible to formulate *ad hoc* potentials which can be used to evaluate the quality of structural models of the corresponding system. This is typically done using quadratic penalties whose force constants reflect the accuracy or reliability of the data (15, 16).

Still simpler target functions are used to score predetermined molecular conformations. Scoring functions are widely used in computer-aided drug design and are often based on simplified statistical models parameterized to reproduce experimental binding free energies (17). Scoring functions allow rapid characterizations of very large numbers of molecules and/or conformations as required in high-throughput virtual screening. However, the calibration of these functions is often complex and time-consuming and it is difficult to devise generic functions applicable to a wide range of systems.

Simple physical models are also associated with simplified target functions. One common example in this category is the so-called Gaussian network approach (18). This is generally used with bead representations and involves harmonic spring interactions between beads falling below a given cut-off distance. Springs may have common or distance-related force constants. The native conformation, for which the springs are constructed, naturally becomes the energy minimum of the system. Although this approach is simple, it has proved useful for studying the local flexibility and the low frequency normal modes of macromolecular systems (19). A still simpler, but related approach, involves so-called Go potentials which are step functions, generally between residues, indicating whether or not the interaction distances present within a reference (native) conformation are reproduced. Like

Gaussian network springs, these potentials lead to optimal energies for the native conformation of the system studied. They are frequently used in studying protein folding (20). The last member of this category of target functions are HP potentials (where H stands for hydrophobic and P for polar), which are again used in protein folding studies, generally with lattice models. These potentials limit interaction energy contributions to favourable interactions between hydrophobic beads on neighbouring lattice points and favourable contributions from polar residues exposed on surface lattice points. A number of variations are possible for these functions, which capture some of the basic physics behind the stabilization of native conformations.

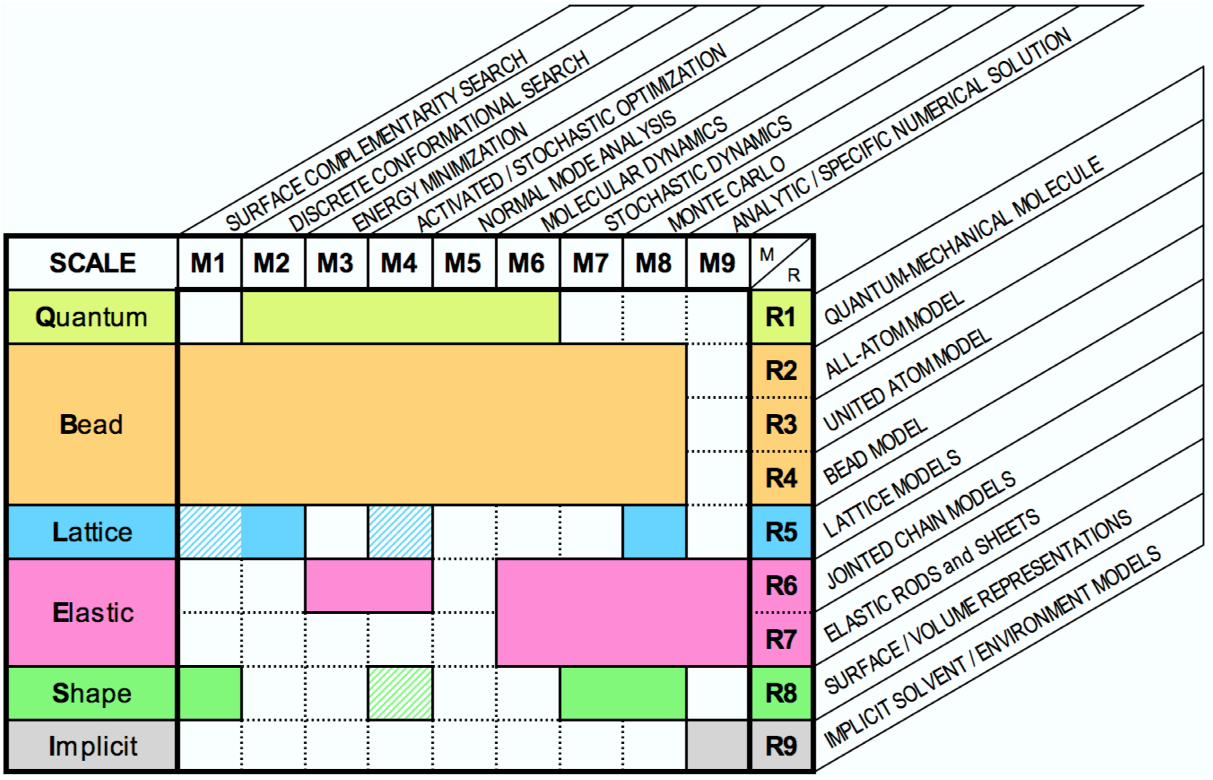
For the simplest representations, such as polymer or sheet models, only simple physical constants are generally required, such as the chain lengths in jointed-chain polymers or the elastic constants in continuum models. It is however also possible to extend these models to account for effects such as excluded volume or electrostatic interactions.

The above discussion shows that target functions are often related to the molecular representation employed, but some flexibility remains. All-atom representations, or united-atom approximations, are required for using classical force fields, but such representations can also be treated with knowledge-based potentials. Bead models can be treated within the Gaussian network approximation, with knowledge-based potentials or with simplified terms derived from classical force fields (21). It should also be noted that the boundaries between the different target functions can become blurred. Thus, it is possible to complete DFT calculations using classical dispersion energy terms (22), or to add physically-based solvent terms to knowledge-based potentials (23, 24). Equally, it is possible to drift away from classical force fields by adding adjustable weighting factors to each term in order to better reproduce experimental data (25). Many other examples of this sort of "boundary crossing" exist.

PUTTING IT ALL TOGETHER

Having summarized the different representations, modelling methodologies and target functions, we can now try and bring these elements together into an overall picture. As seen in the above discussion, the choice of target function is largely subservient to the choice of representation. We can therefore reasonably limit our overview to combinations of representations and methodologies. This picture is presented in Table 3, where shaded squares indicate combinations that are in use, or are at least potentially interesting. Certain combinations, although physically possible, have been left blank because they have little practical interest (e.g. energy minimization for elastic rods for which analytical solutions exist). Other combinations are cross-hatched, indicating that there are currently only a few examples of their use. As a reader, you are encouraged to tell us if you think this table needs correcting or extending.

Table 3. Representation-methodology combinations



The first point to note from Table 3 is that the shaded squares form horizontal blocks, showing that, as with target functions, the choice of representation again dominates the overall picture. We can also see from the structure of the table that it is possible to group certain representations together into scales for which similar methodological approaches are applicable. Although this step is somewhat subjective, it helps to simplify the picture by reducing nine representations to six broader scales. Let's now look briefly at what goes on at each of these scales.

Quantum-Scale - quantum chemistry
Using a quantum representation, as discussed above, offers significant advantages not only in terms of accuracy, but also in opening up the route for studying chemical reactions, interactions with radiation, calculating a wide variety of observable properties (polarizabilities, electron affinities, ...) and so on. However, the computational cost of such representations explains why most studies involve one-off calculations or limited geometry optimization. Linear scaling methods are improving, but not revolutionizing, this situation (26). Dynamic behaviour is accessible thanks to the Car-Parrinello approach, but in comparison to classical MD simulations, it remains limited to small systems (tens of heavy atoms) and to short time scales (a few ps) (27).

Bead-Scale - all-atom and reduced models
This scale contains the broadest and most adaptable class of representations for studying biological systems. Virtually all the numerical methodologies can be put to use, from scoring for high-throughput studies of drug candidates or simplified models of macromolecular docking, to Newtonian or stochastic dynamic simulations covering time scales ranging from picoseconds to milliseconds. When dynamic properties are not indispensable, a wide variety of Monte Carlo approaches can also be used, notably for investigating the conformational

space of individual macromolecules. A great deal of effort has gone into refining simulation methodologies for this scale. Newtonian molecular dynamics has been extended in many ways (essential dynamics, targeted dynamics, parallel tempering ...) enabling larger and slower conformational changes to be studied. In amenable cases, even relatively precise free energy profiles can be obtained. For still slower processes, various stochastic adaptations ranging from Brownian dynamics and Langevin dynamics to the dissipative particle approach can be used, notably for studying protein-protein interactions. For complex conformational changes, multi-copy approaches, which introduce a mean field strategy to enable simultaneous studies of partial replicas of the system, have also been shown to be effective. Although all-atom representations are the most accurate at this scale, reduced models are finding more and more applications, in the areas of folding (e.g. threading approaches to fold prediction, one of the varieties of the discrete conformational search methodologies), structural flexibility (e.g. with Gaussian network models), recognition (many docking strategies being limited to, or starting from, reduced representations and simplified target functions) and molecular assembly (e.g. peptide aggregation in lipid bilayers). Structure refinement is another area of application where bead-scale techniques are routinely used, ranging from high-resolution atomic (X-ray crystallography, NMR spectroscopy) to coarse-grain models (electron microscopy and tomography). In this context, the target functions can be modified to include experimentally derived restraints.

Lattice scale - lattice models

Lattice models, originally derived from physical studies of spin-glasses, have mainly been used to investigate the underlying principles of protein folding. By dramatically reducing the number of conformational states available to a linear polymer chain they enable exhaustive studies of its conformational energy space and of the pathway describing the passage from unfolded to folded states (28). Lattices started largely as "toy" models of folding, but with various refinements they can play a useful part in predicting the optimal folds of structurally uncharacterized amino acid sequences (29). This type of model is also beginning to be used in the field of macromolecular interactions (30).

Elastic scale – elastic polymer and sheet models

Discrete and continuous approaches belonging to this category are widely used in polymer physics. In the biological field, they are largely restricted to modelling macromolecules that behave like long-chain polymers. These include nucleic acids, cytoskeletal filaments (actin, microtubules), certain muscle proteins and polysaccharides. Single molecule experiments studying DNA deformability (31) and models of packaging viral genomes inside capsids (32) are two areas where applications have been found. Analytical solutions exist for various properties of both continuous and the simplest discrete polymer models, but more complex discrete models require numerical approaches. 2D continuum models of thin sheets with elastic free energy have applications ranging from models of lipid membrane patches to cellular deformations (33). The dynamics of undulating fluid bilayers can also be captured for example using Brownian dynamics approaches (34).

Shape scale - surface/volume models

Such models, while common in descriptive studies (overall conformation, binding pockets, etc), have not yet been widely used for modelling biological systems. They can however be useful as confinement restraints in modelling large molecular assemblies (15) and in studying the effects of molecular crowding (35, 36). They would however appear to be very promising for large scale simulations which require a reasonably accurate representation of the space occupied by a macromolecule, but do not need to deal with all the atoms which comprise this

space. Rigid-body models are also suitable for docking studies and can be used in conjunction with experimental X-ray or neutron small-angle scattering data (37).

Implicit scale - simplified molecular environment models

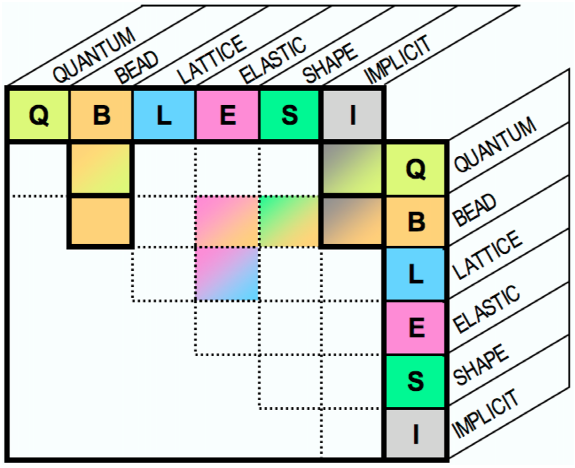
In quantum-scale studies, while gas phase calculations are not representative of the biological environment, the computational cost of treating an explicit solvent environment is often prohibitive. An implicit solvent representation offers a computationally cheaper alternative. Similarly, for bead-scale simulations of biological macromolecules including explicit solvent molecules, the vast majority of the computational effort is linked to treating the solvent. It is thus again very advantageous to avoid explicit representations of solvent molecules. Similar arguments hold for the treatment of lipid bilayers. Simplified solvent models can be discrete or continuous. Polarizable Langevin dipoles (38) are an example of a discrete model, while continuum models solve either the Poisson or the Poisson-Boltzmann equations for a solute/solvent interface (39). Even with continuum representations, analytic solutions are generally impossible, given the complex shapes of biomolecules. The cost of standard numerical solutions favour the introduction of further simplifications, such as the Generalized Born approach (40, 41), although this also implies introducing adjustable parameters.

CONCURRENT MULTI-SCALING

As we discussed in the introduction, the nature of biological systems implies that no single representation, methodology or target function can be appropriate for solving all biological modelling problems. Different system sizes, different time scales, different processes and different requirements for accuracy all suggest that the various strategies discussed above will all be able to play useful roles in solving specific problems. In some cases, a succession of different strategies can be used to solve a single problem, with a passage of information from one level to the next. A simple example of this is using more accurate methods on small systems to obtain parameters for more approximate treatments of larger systems (e.g. getting force field parameters from quantum mechanical calculations). However, a closer coupling can be obtained by using different strategies simultaneously to solve a single problem.

In Table 4 we try to summarize the state of affairs in what we can term "concurrent multi-scaling", starting from the six modelling scales described in the previous section. The best established combinations are outlined in black. Again, the reader is encouraged to point out any improvements that could be made to this table.

Table 4. Concurrent multi-scaling combinations



Although this type of concurrent multi-scaling is now frequently discussed and is the subject of many workshops, there are still relatively few examples of its application in the biological field. One that has been around for some time is the combination of quantum mechanics and classical force fields (so-called QM/MM methods), which are typically used to study the influence of a macromolecular environment on a chemical process. This involves representing a limited part of the system, for example the active site residues and the substrate of an enzyme, quantum mechanically, while treating the rest of the system as a perturbation. What communication exists between the two components of the system, and how the quantum-classical boundary is treated, depends on the level of integration of the hybrid model. This type of approach is commonly coupled with energy minimization (42) or limited dynamics (27).

Moving down the scale of accuracy, it is possible to combine all-atom force field approaches, or quantum-scale models, with an implicit treatment of the solvent and counterion environment. As discussed above, this represents a considerable time gain, particularly in molecular dynamics approaches where explicit water molecules often represent the major computational expense. Since there is generally a geometrically complex interface between the solute and the solvent, numerical Poisson-Boltzmann solutions are necessary. Until recently these were too slow for dynamic simulations and also posed problems for the calculation of derivatives with respect to atomic displacements. This situation is now evolving (43). Alternatively, approximate methods such as the Generalized Born approach can be used (44). In either case, it must be assumed that the interactions of individual water molecules at the solute-solvent interface do not play a major role in determining the behaviour of the system. Implicit environment approaches are also useful in the field of biological membranes, notably coupled to bead-scale modelling (45).

These first examples correspond to Q-B coupling for QM/MM methods and to Q-I or B-I coupling for the combination of explicit solute with implicit solvent (or bilayer) methods. Although a lot of research is still going on to improve these multi-scale methods, they have already reached a certain level of maturity.

A more recent area of development involves multi-scale approaches incorporating elastic models. Biological membranes are an important area of application, where it is possible, for example, to combine lattice models of protein diffusion within a fluctuating elastic bilayer model (46). Coupling elastic and bead-scale models is also possible as shown by a study of the phase separation of mixed lipid bilayers (47). In this case, all components of the system were treated at both B- and E-scales using a feedback mechanism to couple the

models. Another example coupling the bead and elastic scales, this time for protein-DNA interactions, can be found in the modelling of the Lac repressor/operator complex carried out in the group of Klaus Schulten (48), where an elastic model of a DNA loop is combined with explicit DNA fragments bound to protein domains.

Another example of inter-scale hybrid methods involves using geometrically defined objects to study crowding effects on macromolecules (35) or, alternatively, to confine macromolecular assemblies (15). This corresponds to B-S coupling in Table 4. Other hybrid methods more conservatively mix different representations within a given scale, such as bead models containing regions with different resolutions (49), or all-atom and bead representations, as in (50) and in work underway in our laboratory.

An alternative approach to combining existing scales into a hybrid model involves specific methods which incorporate only desired properties of a given scale into another one. As an example, this approach was used to model proton translocation and chemical reactions, typically studied at the quantum-scale, within a bead-scale representation (51-54).

Looking again at Table 4, we can see that concurrent multi-scale methods have so far exploited only five of 15 potential inter-scale combinations of the broad classifications used in constructing this table and only one of six possible intra-scale combinations. Although some of the total of 21 potential combinations are clearly unpromising, there still seems to be considerable scope for developing new approaches.

CONCLUSIONS

The aim of this review was to summarize the range of modelling strategies which are applicable to biological systems in general and proteins in particular. We have tried, as far as possible, to adopt a systematic approach. The results, summarized in the four tables contained in the review, show that although designing a modelling strategy requires choosing a molecular representation, a modelling methodology and a target function, the most important choice is that of the representation, since this sharply restricts the remaining options. This overview also suggests that while approaches are available to treat systems of widely varying size and time scales, with widely varying levels of accuracy, there is still much work to be done in combining different approaches, both in consecutive and in concurrent hybrid strategies. Such approaches nevertheless seem to be indispensable if we want our models to incorporate more of the complexity of biological systems. To return to the title of this review, there is still plenty of room in the middle.

REFERENCES

1. GenomeNewsNetwork, http://www.genomenewsnetwork.org/.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, R. D., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A., 3rd, and Venter, J. C., The minimal gene complement of Mycoplasma genitalium, *Science, 270*, 397 (1995).
3. Takahashi, K., Ishikawa, N., Sadamoto, Y., Sasamoto, H., Ohta, S., Shiozawa, A., Miyoshi, F., Naito, Y., Nakayama, Y., and Tomita, M., E-Cell 2: multi-platform E-Cell simulation system, *Bioinformatics, 19*, 1727 (2003).
4. Goedecker, S., and Scuseria, G. E., Linear scaling electronic structure methods in chemistry and physics, *Comp Sci Eng, 5*, 14 (2003).

5. Dixon, S. L., and Merz, K. M., Jr., Semiempirical molecular orbital calculations with linear system size scaling, *J Chem Phys, 104*, 6643 (1996).
6. Yang, W., and Lee, T.-S., A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules, *J Chem Phys, 103*, 5674 (1995).
7. Daudey, J. P., Claverie, P., and Malrieu, J. P., Perturbative *ab initio* calculations of intermolecular energies. I. Method, *Int J Quantum Chem, 8*, 1 (1974).
8. Gresh, N., Claverie, P., and Pullman, A., Theoretical studies of molecular conformation. Derivation of an additive procedure for the computation of intramolecular interaction energies. Comparison with *ab initio* SCF computations, *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta), 66*, 1 (1984).
9. Mackerell, A. D., Jr., Empirical force fields for biological macromolecules: overview and issues, *J Comput Chem, 25*, 1584 (2004).
10. Darden, T., York, D., and Pedersen, L., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems, *J Chem Phys, 98*, 10089 (1993).
11. Sippl, M. J., Knowledge-based potentials for proteins, *Curr Opin Struct Biol, 5*, 229 (1995).
12. Jiang, L., Gao, Y., Mao, F., Liu, Z., and Lai, L., Potential of mean force for protein-protein interaction studies, *Proteins, 46*, 190 (2002).
13. Gromiha, M. M., Siebers, J. G., Selvaraj, S., Kono, H., and Sarai, A., Intermolecular and intramolecular readout mechanisms in protein-DNA recognition, *J Mol Biol, 337*, 285 (2004).
14. Russ, W. P., and Ranganathan, R., Knowledge-based potential functions in protein design, *Curr Opin Struct Biol, 12*, 447 (2002).
15. Malhotra, A., and Harvey, S. C., A quantitative model of the *Escherichia coli* 16 S RNA in the 30 S ribosomal subunit, *J Mol Biol, 240*, 308 (1994).
16. Alber, F., Kim, M. F., and Sali, A., Structural characterization of assemblies from overall shape and subcomplex compositions, *Structure, 13*, 435 (2005).
17. Gohlke, H., and Klebe, G., Statistical potentials and scoring functions applied to protein-ligand binding, *Curr Opin Struct Biol, 11*, 231 (2001).
18. Chennubhotla, C., Rader, A. J., Yang, L. W., and Bahar, I., Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies, *Phys Biol, 2*, S173 (2005).
19. Bahar, I., and Rader, A. J., Coarse-grained normal mode analysis in structural biology, *Curr Opin Struct Biol, 15*, 586 (2005).
20. Shea, J. E., and Brooks, C. L., 3rd, From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding, *Annu Rev Phys Chem, 52*, 499 (2001).
21. Zacharias, M., Protein-protein docking with a reduced protein model accounting for side-chain flexibility, *Protein Sci, 12*, 1271 (2003).
22. von Lilienfeld, O. A., Tavernelli, I., Rothlisberger, U., and Sebastiani, D., Optimization of effective atom centered potentials for London dispersion forces in density functional theory, *Phys Rev Lett, 93*, 153004 (2004).
23. Jones, D. T., Taylor, W. R., and Thornton, J. M., A new approach to protein fold recognition, *Nature, 358*, 86 (1992).
24. Sippl, M. J., Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures, *J Comput Aided Mol Des, 7*, 473 (1993).
25. Havranek, J. J., Duarte, C. M., and Baker, D., A simple physical model for the prediction and design of protein-DNA interactions, *J Mol Biol, 344*, 59 (2004).

26. Morokuma, K., New challenges in quantum chemistry: Quests for accurate calculations for large molecular systems, *Philos Transact A Math Phys Eng Sci, 360*, 1149 (2002).
27. Carloni, P., Rothlisberger, U., and Parrinello, M., The role and perspective of *ab initio* molecular dynamics in the study of biological systems, *Acc Chem Res, 35*, 455 (2002).
28. Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S., Principles of protein folding--a perspective from simple exact models, *Protein Sci, 4*, 561 (1995).
29. Zhang, Y., Kolinski, A., and Skolnick, J., TOUCHSTONE II: a new approach to *ab initio* protein structure prediction, *Biophys J, 85*, 1145 (2003).
30. Dima, R. I., and Thirumalai, D., Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics, *Protein Sci, 11*, 1036 (2002).
31. Lavery, R., Lebrun, A., Allemand, J.-F., Bensimon, D., and Croquette, V., Structure and mechanics of single biomolecules: experiment and simulation, *J Phys (Cond. Mat.), 14*, R383 (2002).
32. LaMarque, J. C., Le, T. V., and Harvey, S. C., Packaging double-helical DNA into viral capsids, *Biopolymers, 73*, 348 (2004).
33. Deuling, H. J., and Helfrich, W., Red blood cell shapes as explained on the basis of curvature elasticity, *Biophys J, 16*, 861 (1976).
34. Lin, L. C., and Brown, F. L., Brownian dynamics in Fourier space: membrane simulations over long length and time scales, *Phys Rev Lett, 93*, 256001 (2004).
35. Takahashi, K., Arjunan, S. N., and Tomita, M., Space in systems biology of signaling pathways--towards intracellular molecular crowding *in silico*, *FEBS Lett, 579*, 1783 (2005).
36. Lago, S., Cuetos, A., Martinez-Haya, B., and Rull, L. F., Crowding effects in binary mixtures of rod-like and spherical particles, *J Mol Recognit, 17*, 417 (2004).
37. Petoukhov, M. V., and Svergun, D. I., Global rigid body modeling of macromolecular complexes against small-angle scattering data, *Biophys J, 89*, 1237 (2005).
38. Warshel, A., and Russell, S. T., Calculations of electrostatic interactions in biological systems and in solutions, *Q Rev Biophys, 17*, 283 (1984).
39. Feig, M., and Brooks, C. L., 3rd, Recent advances in the development and application of implicit solvent models in biomolecule simulations, *Curr Opin Struct Biol, 14*, 217 (2004).
40. Hawkins, G., Cramer, C., and Truhlar, D., Pairwise solute descreening of solute charges from a dielectric medium, *Chem Phys Lett, 246*, 122 (1995).
41. Tsui, V., and Case, D. A., Theory and applications of the Generalized Born solvation model in macromolecular simulations, *Biopolymers, 56*, 275 (2000).
42. Dinner, A. R., Blackburn, G. M., and Karplus, M., Uracil-DNA glycosylase acts by substrate autocatalysis, *Nature, 413*, 752 (2001).
43. Prabhu, N. V., Zhu, P., and Sharp, K. A., Implementation and testing of stable, fast implicit solvation in molecular dynamics using the smooth-permittivity finite difference Poisson-Boltzmann method, *J Comput Chem, 25*, 2049 (2004).
44. Bashford, D., and Case, D. A., Generalized Born models of macromolecular solvation effects, *Annu Rev Phys Chem, 51*, 129 (2000).
45. Tanizaki, S., and Feig, M., A Generalized Born formalism for heterogeneous dielectric environments: application to the implicit modeling of biological membranes, *J Chem Phys, 122*, 124706 (2005).
46. Brown, F. L., Regulation of protein mobility via thermal membrane undulations, *Biophys J, 84*, 842 (2003).

47. Shi, Q., and Voth, G. A., Multi-scale modeling of phase separation in mixed lipid bilayers, *Biophys J, 89*, 2385 (2005).
48. Villa, E., Balaeff, A., and Schulten, K., Structural dynamics of the lac repressor-DNA complex revealed by a multiscale simulation, *Proc Natl Acad Sci U S A, 102*, 6783 (2005).
49. Doruker, P., Jernigan, R. L., and Bahar, I., Dynamics of large proteins through hierarchical levels of coarse-grained structures, *J Comput Chem, 23*, 119 (2002).
50. Neri, M., Anselmi, C., Cascella, M., Maritan, A., and Carloni, P., Coarse-grained model of proteins incorporating atomistic detail of the active site, *Phys Rev Lett, 95*, 218102 (2005).
51. Xu, J., and Voth, G. A., Computer simulation of explicit proton translocation in cytochrome c oxidase: the D-pathway, *Proc Natl Acad Sci U S A, 102*, 6795 (2005).
52. Braun-Sand, S., Strajbl, M., and Warshel, A., Studies of proton translocations in biological systems: simulating proton transport in carbonic anhydrase by EVB-based models, *Biophys J, 87*, 2221 (2004).
53. Mongan, J., and Case, D. A., Biomolecular simulations at constant pH, *Curr Opin Struct Biol, 15*, 157 (2005).
54. Florian, J., Goodman, M. F., and Warshel, A., Computer simulation of the chemical catalysis of DNA polymerases: discriminating between alternative nucleotide insertion mechanisms for T7 DNA polymerase, *J Am Chem Soc, 125*, 8163 (2003).