

REVIEW

Published 27 Sep 2011 | DOI: 10.1038/ncomms1481

# Capturing the essence of folding and functions of biomolecules using coarse-grained models

Changbong Hyeon<sup>1</sup> & D. Thirumalai<sup>2</sup>

The distances over which biological molecules and their complexes can function range from a few nanometres, in the case of folded structures, to millimetres, for example, during chromosome organization. Describing phenomena that cover such diverse length, and also time, scales requires models that capture the underlying physics for the particular length scale of interest. Theoretical ideas, in particular, concepts from polymer physics, have guided the development of coarse-grained models to study folding of DNA, RNA and proteins. More recently, such models and their variants have been applied to the functions of biological nanomachines. Simulations using coarse-grained models are now poised to address a wide range of problems in biology.

**M**inimal models that capture the essence of complex phenomena has a rich history in the natural sciences. In condensed matter physics, insights into many phenomena have emerged from analytic theories of models, which use effective many-body Hamiltonians that succinctly capture the essence of the problems<sup>1</sup>. Examples include phase transitions, superfluidity and superconductivity. However, complex problems, such as spin glasses<sup>2</sup>, structural glasses<sup>3,4</sup> and a host of problems in biology like protein and RNA folding and functions of macromolecules, have resisted solutions using purely theoretical methods. These and other problems in material science, in which a wide range of time, energy, and length scales are intertwined, require well-designed computer simulations, which capture the essential features of the systems. Although the temptation to use detailed atomic simulations in protein folding and more complicated problems is hard to resist, such an approach has given us only limited insights. In contrast, since the first classical molecular dynamics simulation that reported phase transition in hard-sphere systems<sup>5</sup>, it has been clear that coarse-grained (CG) models are often the only way to describe phenomena that involve an interplay of multiple energy and time scales. Nowhere is the need for CG models greater than in biology in which self-assembly of macromolecules and their functions, which involve multiple partners, occur on time and length scales that cover many orders of magnitude. In the context of protein and RNA folding, simulations using CG models, guided by theoretical concepts<sup>6–10</sup>, have unearthed the principles of self-assembly. More recently, models, which were introduced to describe folding of isolated proteins<sup>11–15</sup> and RNA<sup>16</sup>, have also been adopted and extended in novel ways to predict functions of large complexes such as ribosomes<sup>17</sup>, molecular chaperones<sup>18</sup>, enzyme catalysis<sup>19</sup>, protein insertion into membranes<sup>20</sup> and a number of motors<sup>21–28</sup>. These developments have resulted in a quiet revolution, which has provided molecular insights into a variety of biological processes.

In the last two decades, fundamental breakthroughs in structural organization and dynamics of proteins, RNA and DNA have been achieved using theoretical concepts from polymer physics<sup>29</sup> and CG simulations. Here we describe how simulations using a variety of CG models have been

<sup>1</sup> School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Republic of Korea. <sup>2</sup> Biophysics Program, Institute For Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA. Correspondence and requests for materials should be addressed to D.T. (email: thirum@umd.edu).

successful in describing dynamical processes in biology, spanning a wide range of length scales. These achievements have been further extended to probe folding under cellular conditions<sup>30–32</sup>, and more recently to describe functional dynamics of biological nanomachines<sup>18,22–24,26</sup>. The use of CG models and simple theoretical ideas have also found fruitful applications in many other areas in biology such as gene networks, systems biology and analysis of complex metabolic pathways.

### Length scales determine extent of coarse-graining

Description of reality using models requires a level of abstraction, which depends on the phenomenon of interest. For example, near a critical point, exponents that describe the vanishing of order parameter or divergence of correlation length are universal, depending only on the dimensionality ( $d$ ), and are impervious to atomic details. These findings, which are rooted in the concepts of universality and renormalization group<sup>33</sup>, are also applicable to the properties of polymers<sup>29</sup>. For example, the size of a long homopolymer and its distribution of end-to-end distance depend only on the solvent quality, the degree of polymerization, and  $d$ , but not on the details of monomer structure<sup>29</sup>. However, on length scales that are on the order of a few nm, one has to contend with chemical properties of the monomer.

In the absence of rigorous theoretical underpinnings, intuitive arguments and phenomenology are involved in modelling complex biological processes (Box 1). Here also the level of description depends on length scales. In nucleic acids, at short length scales ( $l \leq 5 \text{ \AA}$ ), detailed chemical environment determines the basic forces (hydrogen bonds and dispersion forces) between two nucleotides. On the scale  $l \sim (1\text{--}3) \text{ nm}$ , interactions between two bases, base stacks and grooves of the nucleic acids become relevant. Understanding how RNA folds ( $l \sim (1\text{--}3) \text{ nm}$ ) requires energy functions that provide at least a CG description of nucleotides, and interactions between them in the native state and excitations around the folded structure. On the persistence length scale  $l_p \approx 150 \text{ bp} \approx 50 \text{ nm}$  (ref. 34) and beyond it suffices to treat dsDNA as a stiff elastic filament without explicitly modelling the base-pairs. If  $l \sim \mathcal{O}(1) \mu\text{m}$  dsDNA behaves like a self-avoiding polymer<sup>35</sup>. On the scale of chromosomes ( $l \sim \text{mm}$ ), a much coarser description suffices. Thus, models for DNA, RNA and proteins vary because the scale of structural organization changes from nearly mm in chromatin to several nm in the folded states of RNA and proteins.

### Polymer models for dsDNA and chromosome structure

Length,  $L$ , of double-stranded DNA (dsDNA) exceeds a few  $\mu\text{m}$  with persistence length,  $l_p \approx 50 \text{ nm}$ . On these scales global properties of dsDNA, such as the end-to-end distance and the dependence of  $l_p$  on salt concentration, are not greatly affected by fluctuations of individual base pairs. Consequently, dsDNA can be treated as a fluctuating elastic material, for which the worm-like chain (WLC) is a suitable polymer model. On much longer scales ( $L \sim 1 \text{ mm}$ ), which is relevant to chromosome, the genomic material can be described as a flexible polymer. Using these scale-dependent models, a number of predictions for DNA organization and dynamics can be made.

**Looping dynamics.** Loop formation in biopolymers is an elementary process in the self-assembly of DNA, RNA and proteins. However, understanding cyclization kinetics is complicated, because multiple length scales and internal chain modes are intertwined in bringing distant parts of DNA into proximity. For a short chain, the cyclization time,  $\tau_c$ , scales as  $L^{3/2}$  while  $\tau_c \sim L^2$  when  $L$  increases<sup>36,37</sup>. The problem of cyclization becomes more challenging in the looping dynamics of dsDNA, an elementary process that is relevant in controlling gene expression and DNA condensation. In the CG model, a single pitch of a double helix, formed by 10.5 base pairs, represents one interaction centre (Fig. 1a). Thus,  $l_p$  encompasses (14–15) CG

interaction centres ( $l_p \approx 150 \text{ bp}$ ). The parameters for bond and bending potentials along the chain, consisting of multiple CG centres, are selected to reproduce the persistence length of dsDNA<sup>38,39</sup>, allowing us to study various dynamics of dsDNA, stretching, looping or supercoiling from the perspective of polymer physics. The ease of loop formation and the associated kinetics is characterized by  $L/l_p$ . For  $L/l_p \sim \mathcal{O}(1)$ , energy required to bend dsDNA makes the cyclization difficult for short chains. In contrast, when  $L/l_p \gg 1$  the cyclization between two ends gets harder because of loss of chain entropy. Theory and simulations using CG model showed that  $\tau_c$  is the shortest when  $L/l_p \approx 2\text{--}3$  (refs 40–42) (Fig. 1a). Interestingly, in looping of dsDNA responsible for gene regulation in prokaryotes  $L \approx 100 \text{ bp}$  ( $L/l_p \approx 0.7$ ). For such a dsDNA with  $L \approx 100 \text{ bp}$  sequence effects are also relevant<sup>43–45</sup>.

**Stretching dsDNA.** Fluctuations of dsDNA on scales comparable to  $l_p$  can be described using WLC model, which parameterizes dsDNA as a polymer that resists bending on scale  $\sim l_p$ . Smith *et al.* measured the response of a 97 kbp dsDNA (countour length  $L \approx 33.0 \mu\text{m}$ ) from  $\lambda$ -phage to a stretching force,  $f$  (refs 34,46) (Fig. 1b). In the absence of  $f$ ,  $\lambda$ -DNA conformations are determined by thermal fluctuations, whereas loss in chain entropy must be overcome to stretch dsDNA  $f \neq 0$ . The free energy of stretching of a semiflexible chain under tension is equivalent to a quantum mechanical problem of a dipolar rotor with moment of inertia  $I_p$  in an electric field  $f$ . An extrapolated formula is used to accurately describe the measured force as a function of extension (Fig. 1b). Fits to experimental data yield  $L$  of  $\lambda$ -DNA ( $32.80 \pm 0.10$ )  $\mu\text{m}$  and  $l_p \approx (53.4 \pm 2.3) \text{ nm}$ , thus confirming most directly that dsDNA is a semiflexible chain.

**Confined polymers and bacterial chromosome segregation.** Replication and passage of genetic information to daughter cells are major events in cell reproduction. These complex events are remarkably accurate even in simple organisms. Although chromosome segregation is likely to be complex and well orchestrated, it has recently been proposed that confinement-induced entropic forces due to restrictions in cellular space is sufficient to drive chromosome segregation in bacteria<sup>47,48</sup> (Fig. 1c). This proposal was formulated using molecular simulations of tightly confined self-avoiding polymers chains in cylindrical space, which show that the chains segregate and become spatially organized reminiscent of that observed in bacteria. In such highly confined spaces polymer conformations are determined by  $\xi$ , the size of a renormalized structural unit, the Flory radius  $R_F$  in the absence of confinement, and the length ( $P$ ) and diameter  $D$  of the cylinder. In *Escherichia coli* the values are  $\xi = 87 \text{ nm}$ ,  $R_F = 3.3 \mu\text{m}$ ,  $D$  and  $P$  are  $0.24 \mu\text{m}$  and  $1.3 \mu\text{m}$ , respectively. Armed with the results for confined polymers, a concentric shell model for bacterial chromosome was proposed<sup>47,48</sup> in which the nucleoid was modelled as an inner and outer cylinder. The unreplicated ‘mother’ strand, a self-avoiding chain, is restricted to the inner compartment whereas the ‘daughter’ chain (obtained in simulations by adding monomers at a set time in the Monte Carlo simulations) are free to explore the entire nucleoid volume. The results of the simulations show that the newly added (or replicated) chain segregates to the periphery of the nucleoid, driven by gain in entropy, and become spatially organized as they are synthesized (Fig. 1c). CG modelling combined with polymer theory leads to the discovery that entropic forces alone are sufficient to drive chromosome segregation in bacteria, with proteins perhaps having a secondary role in poising the state of the chromosome for enabling the entropy-driven mechanism.

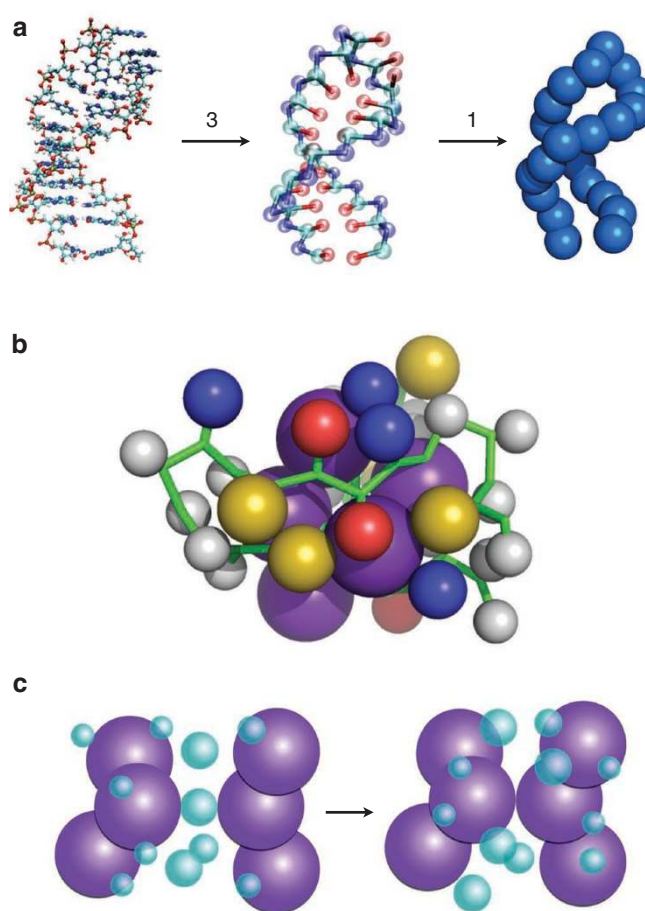
**Chromosome folding.** In eukaryotic cells chromosomes fold into globules that spatially occupy well-defined volumes known as chromosome territories<sup>49</sup>. In this process, widely separated genetic regions are brought into close proximity. Knowledge of the spatial arrangement of chromosomes is important in describing gene

## BOX 1 GENRE OF CG MODELS

Formally one can think of coarse-graining as a process by which effective energy functions are generated from microscopic Hamiltonian by integrating over irrelevant degrees of freedom. The block-spin renormalization procedure in Ising model shows that, as the degrees of freedom are thinned, multiparticle interactions are generated. Similar ideas can be used to construct effective Hamiltonian by insisting that the partition function for the CG and the microscopic Hamiltonian be the same. However, given the large inaccuracies in force fields, intuitive and physical consideration have proved far more profitable in guiding the development of CG models. The CG strategy is successful, because the characteristic time scales at each length scale are well separated from each other.

In response to the challenge of describing biological processes that span several orders of magnitude in time and length scales, a variety of CG models for DNA, RNA and proteins have been proposed. Although CG models have been prevalent in the polymer literature for over fifty years, their use in proteins began in earnest with the pioneering work of Levitt and Warshel<sup>11</sup>. The efficacy of off-lattice models for protein-folding kinetics was first demonstrated by Honeycutt and Thirumalai<sup>12</sup>. In all the CG models, polypeptide chains and nucleic acids are represented using a reduced description. The accompanying figure shows a few examples. **(a)** Three interaction site (TIS) models for a RNA hairpin obtained by representing each nucleotide by three sites one each for phosphate, ribose and base. **(b)**  $C_{\alpha}$ -SCM for WW domain obtained by replacing each amino acid by two sites one centred on the  $\alpha$ -carbon and the other at the centre of the side chain. **(c)** Water-mediated interactions can also be captured using effective potentials<sup>93</sup>. Using these representations and their variations, a number of types of CG models have been developed. The common unifying aspect of all these models is that the nucleotides and amino acids are represented by only a few interaction sites. However, these models vary in details, especially in the number of interaction centres per nucleotide or amino acid needed to encode the folded structure.

A major advantage of CG models is that their conformational space can be exhaustively sampled. However, even with simplification, accurate results for thermodynamics can only be obtained using enhanced sampling methods. Towards this end, simulation of CG models have used replica exchange methods and multicanonical methods. In addition, low-friction Langevin dynamics has also been used to efficiently sample conformational space. These methods are necessary, especially in simulating proteins with complex topology. To obtain kinetic information for folding or transition times between allosteric states typically Brownian dynamics (BD) simulations are performed. In typical BD simulations the Brownian time is  $\tau_H \approx \zeta_H a^2 / k_B T_s$  where  $\zeta_H$  is the friction constant,  $a$  is roughly the size of a CG bead, and  $T_s$  is the simulation temperature. Estimate of these quantities<sup>94</sup> have been used to map simulation times to real times in a number of applications.



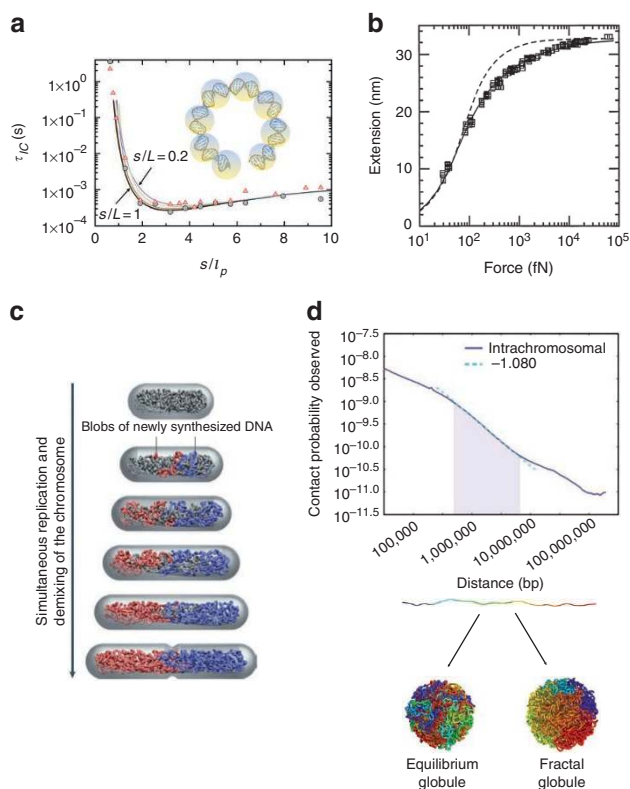
activity and the state of the cell. Polymer physics concepts have been used to describe the structures of folded chromosome using constraints derived from experiments. These calculations have provided considerable insights into their compartmentalization in the nucleus<sup>50</sup>. A number of models, such as the random walk model, and models that connect mega-based size domains by chromatin loops have been used to describe higher structures of chromatin. The experimental resolution is roughly 1Mb ( $\sim 340 \mu\text{m}$ ), and consequently coarse-graining in this context must be on length scales on the order of a  $\mu\text{m}$ . Recently folding principles for human genome were proposed using data for long-range contacts between distinct loci as constraints<sup>51</sup>. Experiments showed that contact probability,  $I(s)$ , between loci in a chromosome, which is separated by genomic distance  $s$  (measured in units of bp) exhibits a power law decay in the range  $\sim 500 \text{ kb}$  to  $\sim 7 \text{ Mb}$ . The observed dependence  $I(s) \sim s^{-1}$  can be rationalized using polymer models (Fig. 1d) introduced a number of years ago in describing collapse of homopolymers<sup>52</sup>. If chromosome folds up into an equilibrium globule (polymer in a

poor solvent) then  $I(s) \sim s^{-1.5}$ , which cannot account for the experimental observations. An alternate model suggests that interface DNA can organize itself into a fractal globule, which is compact and not entangled as an equilibrium globule would be. Monte Carlo simulations of a polymer with 4,000 beads (1 bead = 1,200 bps  $\sim 0.4 \mu\text{m}$ ) were used to generate conformations of fractal and equilibrium globules. The power law decay of  $I(s)$ , with exponent  $\sim -1$ , is consistent with measurements (Fig. 1d). More importantly, the unknotted fractal globules loci that are close in genomic sequence are also in proximity in three-dimensional spatial arrangement, which clearly is relevant for gene activity.

### RNA folding

Since the discovery that RNA can serve as enzymes there has been great impetus to describe their folding in quantitative terms. RNA folding landscape is rugged because of interplay of several competing factors. First, phosphate groups are negatively charged, which implies that polyelectrolyte effects oppose folding. Valence, size and





**Figure 1 | DNA applications.** (a) Loop formation times between two regions in dsDNA separated by  $s$  along the contour from simulations using CG model that represents a single pitch of DNA helix as a monomer unit. Lines are theoretical results. (b) Extension as a function of mechanical force for 97 kb  $\lambda$ -DNA. Symbols are experimental results and the dashed line is the fit using WLC model. (c) Model of bacterial chromosomal separation from simulations of tightly confined polymer chain. The newly synthesized DNA (blue and red) is extruded to the periphery of the unreplicated nucleoid (grey) and the two strings of blobs drift apart and segregate owing to the excluded-volume interactions and conformational entropy. (d) Top figure shows scaling law  $P(s) \sim s^{-1.08}$  where  $P(s)$  is the contact probability for a given genomic distance  $s$ , measured by Hi-C, a method that probes the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing. The exponent in the power-law decay is distinct from  $s^{-1.5}$  for an equilibrated globule (bottom left) whereas  $s^{-1.08}$  scaling (dashed lines from CG simulations in the top figure) is explained using a fractal globule (bottom right), a knot-free, polymer conformation, which enables reversible folding and unfolding at any genomic locus. Panels a–d were adapted from refs 42 (reproduced with permission, © American Institute of Physics), 34 (reproduced with permission from AAAS), 48 and 51 (reproduced with permission from AAAS), respectively.

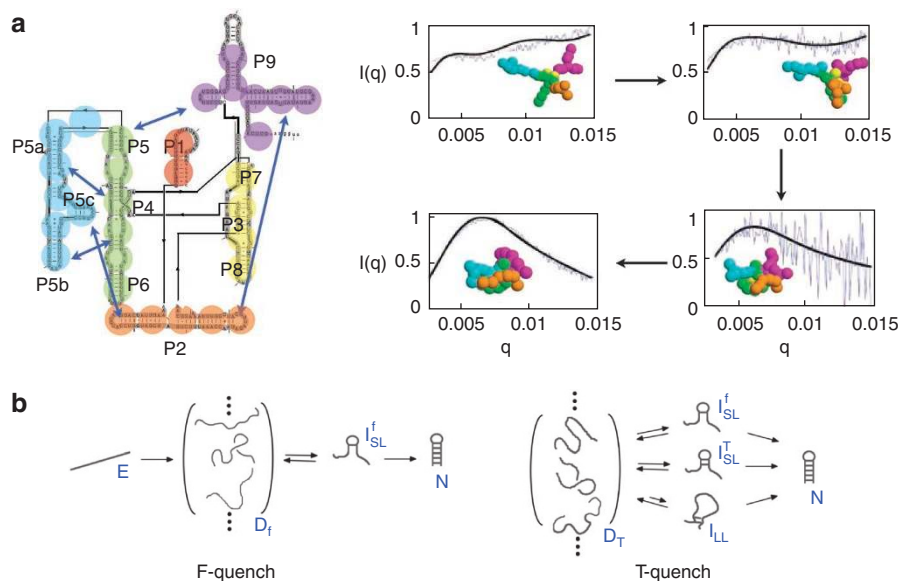
shape of counterions, necessary to induce compaction and folding<sup>53</sup>, can dramatically alter the thermodynamics and kinetics of RNA folding. Second, the nucleotides purine and pyrimidine bases have different sizes but are chemically similar. Third, only ~54% of bases form canonical Watson–Crick base pairs whereas the remaining nucleotides are in non-pairing regions<sup>54</sup>. Fourth, the lack of chemical diversity in the bases results in RNA, easily adopting alternate misfolded conformations, which means that the stability gap between the folded and misfolded structures is not too large. Thus, the homopolymer nature of the RNA monomers, the critical role of counterions in shaping the folding landscape, and the presence of low-energy excitations around the folded state make RNA folding a challenging problem<sup>6</sup>.

**Polyelectrolyte effects.** To fold, RNA must overcome the large electrostatic repulsion between the negatively charged phosphate groups. Polyelectrolyte (PE) based theory shows that multivalent cations ( $Z > 1$ ) are more efficient in neutralizing the backbone charges than monovalent ions—a prediction that is borne out in experiments. The midpoint of the folding transition  $C_m$ , the ion-concentration at which the populations of the folded and unfolded states are equal, for *Tetrahymena* ribozyme is  $\sim 3 \times 10^6$ -fold greater in  $\text{Na}^+$  than in cobalt-hexamine ( $Z = 3$ )! The nature of compact structures depends on  $Z$  with the radius of gyration scaling as  $R_g \propto 1/Z^2$ , which implies compact intermediates have larger free energy as  $Z$  increases. Thus, folding rates should decrease as  $Z$  increases, which also accords well with experiments<sup>55</sup>. Polyelectrolyte theory also shows that counterion charge density  $\zeta = Ze/V$  should control RNA stability. As  $\zeta$  increases, RNA stability should increase—a prediction that was validated using a combination of PE-based simulations and experiments. The changes in stability of in *Tetrahymena* ribozyme in various Group II metal ions ( $\text{Mg}^{2+}$ ,  $\text{Ca}^{2+}$ ,  $\text{Ba}^{2+}$ , and  $\text{Sr}^{2+}$ ) showed a remarkable linear variation with  $\zeta$  (ref. 56). The extent of stability is largest for ions with largest  $\zeta$  (smallest  $V$ ). Brownian dynamics simulations showed that this effect could be captured solely by non-specific ion-RNA interactions<sup>56</sup>. These findings and similar variations of stability in different-sized diamines show that (i) the bulk of the stability arises from nonspecific association of ions with RNA, and (ii) stability can be greatly altered by valence, shape and size of the counterions.

**Structures of RNA intermediates.** The complete characterization of counterion-mediated RNA folding requires structural description of the unfolded ( $U$ ), intermediate ( $I$ ), and the folded states. Structures of the folded states can be obtained using crystallography or NMR. However, it is difficult to characterize the ensemble of structures populated at low ( $U$ ) and moderate ( $I$ ) ion concentrations (C). To obtain the ensemble of  $I$  structures from time-resolved small-angle X-ray scattering data, a CG model for *Tetrahymena* group I was constructed by representing (5–6) nucleotide pairs by a single sphere (Fig. 2a)<sup>57</sup>. The salient findings are: (i) At times before global collapse, the domains of the ribozymes are extended because PE effects dominate. (ii) On time scales that are much less than the overall folding time, there is a drastic reduction in the size of RNA. The folding intermediates are fluid-like and must be a mixture of species that contain specifically collapsed structures (large degree of native-like order) and nonspecifically collapsed conformations (low degree of native-like order).

**Complexity of hairpin formation.** When viewed on length scales that span several bps folding of a small RNA (or DNA) hairpin is remarkably simple. However, when probed on short times (ns– $\mu$ s range) the formation of a small hairpin involving turn formation and base-stacking is remarkably complex. Recent experiments show that the kinetics of hairpin formation in RNA (or ssDNA) deviates from the classical two-state kinetics and is best described as a multi-step process<sup>58</sup>. Extra facets of hairpin formation have been revealed in single-molecule experiments that use mechanical force ( $f$ ). These experiments prompted simulations that vary both  $T$  and  $f$ . The equilibrium phase diagram showed two basins of attraction (folded and unfolded) at the locus of critical points ( $T_m$ ,  $f_m$ ). At  $T_m$  and  $f_m$  the probability of being unfolded and folded is the same. The free energy surface obtained from simulations explained the sharp bimodal transition between the folded and unfolded state when the RNA hairpin is subject to  $f$  (refs 16, 59). Thus, from thermodynamic considerations, hairpin formation can be described as a two-state system.

On temperature quench, the formation of hairpins occurs via multiple steps<sup>39</sup>, as observed in the recent kinetic experiments. Folding pathways between  $T$ -quench and  $f$ -quench refolding are different



**Figure 2 | Ribozyme to RNA hairpin folding.** (a) Left is the secondary structure map of *Tetrahymena* group I intron where the circles show that 5–6 base pairs are used to represent one interaction centre. Simulations of the CG model are used to obtain best agreement with time-dependent small-angle X-ray scattering signals as the ribozyme folds. Representative structures that produce best agreement with experiments are shown<sup>57</sup>. (b) Refolding pathways of a RNA hairpin on quenching the force from a high-to-low value (left) and obtained from temperature quench (right) using SOP model. On force-quench folding commences from an extended (*E*) state by forming the turn, which nucleated the hairpin formation (left).

(Fig. 2b). The initial conformations generated by forced unfolding are fully extended. They are structurally homogeneous. The first event in folding on *f*-quench is loop formation, which is a slow nucleation process (Fig. 2b). Zipping of the remaining base pairs leads to rapid hairpin formation. Refolding on *T*-quench commences from a structurally broad ensemble of unfolded conformations. Therefore, nucleation can originate from many regions in the molecule (Fig. 2b). The simulations showed that the complexity of the folding landscape observed in ribozyme experiments was already reflected in the formation of simple RNA hairpin<sup>16,60</sup> just as  $\beta$ -hairpin formation captures much of the complexity of protein folding<sup>61</sup>.

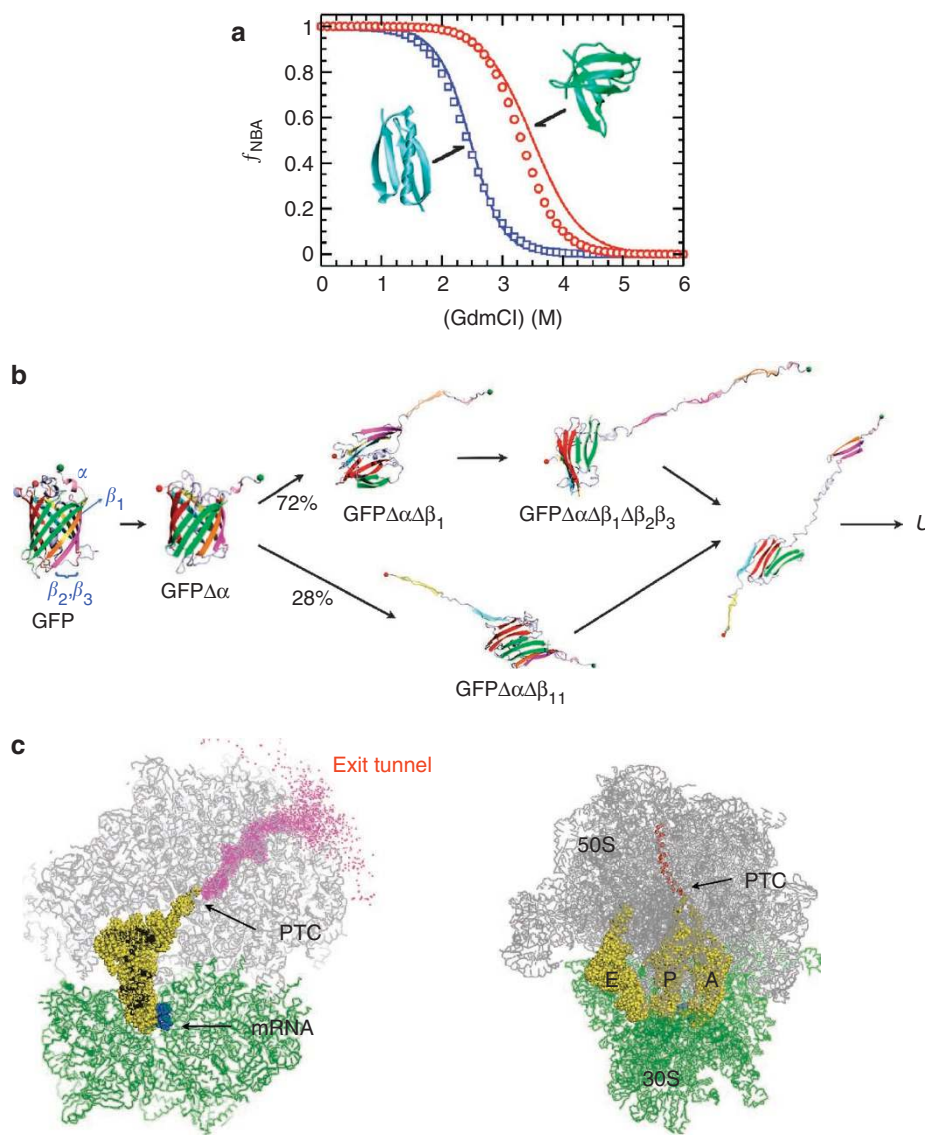
### Protein folding

The impetus to understand the mechanisms of protein folding comes from a number of different sources. First, there is increasing need to produce models that can predict folding thermodynamics and kinetics at conditions used in experiments. Second, it is urgent to describe the biophysical basis of misfolding and the link to neurodegenerative diseases. Third, as we move towards a system level description of cellular processes, it is important to develop theoretical models for describing folding in crowded solutions as well as folding of proteins as they are synthesized by the ribosome.

**Molecular transfer model.** The validity of models can only be assessed by comparing simulation results (obtained under conditions used in experiments) to experiments. Majority of computational studies use temperature to trigger folding and unfolding whereas a substantial number of experiments use denaturants for the same purpose, thus making it difficult to validate the models. This difficulty has been overcome with the introduction of a phenomenological molecular transfer mode (MTM)<sup>62,63</sup>, which combines simulations performed in condition A (for example fixed temperature,  $T_1$  and zero denaturant concentration), and the sampled conformations are assigned appropriate Boltzmann weight such that the behaviour in solution condition B ( $T$  and non-zero denaturant concentration for example) can be accurately predicted without running further

simulations. The MTM theory shows that this procedure is exact provided the conformations of the protein are exhaustively sampled in condition A, and is only limited by the accuracies of the force fields. Applications of MTM requires the free energy cost of transferring a given protein conformation from A  $\rightarrow$  B, which were taken from experimentally measured transfer-free energies for the peptide backbone and each amino acid. The MTM simulations for protein L (an  $\alpha/\beta$  protein) and the nearly all  $\beta$ -sheet cold shock protein quantitatively reproduced measured values of the dependence of the population of the folded state as a function of denaturants (Fig. 3a). Surprisingly, MTM-based simulations also accurately predicted denaturant-dependent measurements in single-molecule experiments.

**Mechanical force in protein folding.** A number of single-molecule experiments, which use mechanical force ( $f$ ) in various modes (force ramp, force quench, and constant force) to initiate folding from arbitrary regions in the energy landscape, have given a new perspective on protein (and RNA) folding<sup>64,65</sup>. These experiments, which monitor time-dependent changes in the extension,  $x(t)$ , of the protein of interest showed that folding occurs in multiple stages on force quench. The power of these experiments are fully realized only by combining them with theory<sup>66–69</sup> and simulations. Such an approach was used to construct the folding landscape of the nearly 250-residue green fluorescent protein (GFP), which has a barrel-shaped structure consisting of 11  $\beta$ -strands with one  $\alpha$ -helix in the carboxy terminus. Using simulations with self-organized polymer (SOP) representation<sup>70</sup> of GFP at the loading rate used in experiments, a rich and complex folding landscape was predicted (Fig. 3b). Unfolding of the native (*N*) began with rupture of the  $\alpha$ -helix leading to  $[GFP\Delta\alpha]$  intermediate. Subsequently, there was a bifurcation in the unfolding pathways. In most cases, the route to the unfolded (*U*) involved population of two extra intermediates,  $[GFP\Delta\alpha\Delta\beta_1]$  ( $\Delta\beta_1$  represents forced-rupture of amino-terminal  $\beta$ -strand) and  $[GFP\Delta\alpha\Delta\beta_1\Delta\beta_2\beta_3]$ . The most striking prediction of the simulations was that the minor pathway had only one intermediate  $[GFP\Delta\alpha\Delta\beta_{11}]$  besides  $[GFP\Delta\alpha]$ . The predictions using SOP



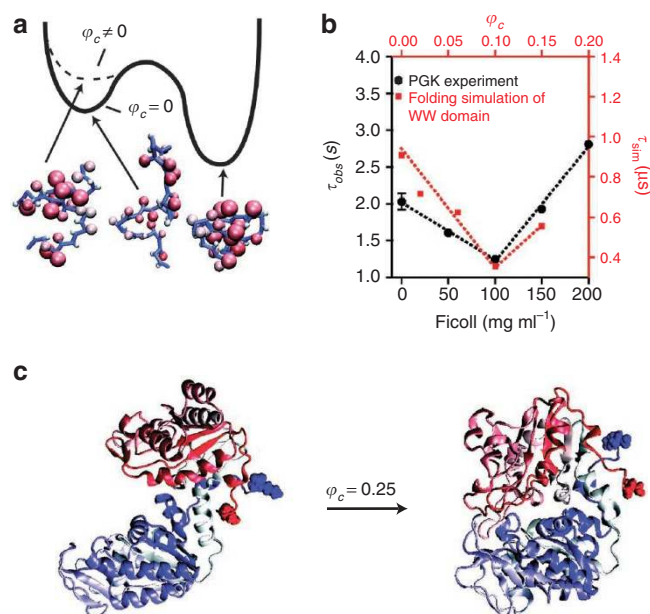
**Figure 3 | Protein folding.** (a) Dependence of fraction of molecules in the native basin of attraction as function of guanidinium chloride concentration for protein L (blue) and Cold shock protein (red) where the symbols are data from experiments, and the lines are results from Molecular transfer model simulations. (b) Unfolding of GFP, an ~250-residue 11-stranded  $\beta$ -barrel protein (left). Forced unfolding obtained from CG simulations on the right shows pathway bifurcation. The structures from the simulations at various stages of unfolding are also shown. (c) The volume of exit tunnel (left) and a helix (right) are shown in the ribosome structure. Co-translational folding of a protein occurs as it is synthesized by the ribosome.

simulations of GFP were quantitatively validated by single-molecule experiments<sup>71</sup>.

**Co-translational folding.** With the determination of the ribosome structures<sup>72,73</sup>, there is great interest in the folding of proteins as they are synthesized. On synthesis, which occurs at the rate of about 20 amino acids per second in *E. coli*, the polypeptide chain traverses a roughly cylindrical tunnel whose lining changes from the peptidyl transfer centre to the exit that is ~10 nm from peptidyl transfer centre (Fig. 3c). Experiments have shown that it is likely that certain regions can accommodate  $\alpha$ -helices depending on the sequence, which is of particular interest for transmembrane helices that can be directly inserted into the membrane by the translocon. Inspired by these experiments, theory and simulations were used to show that the extent of helix formation does depend on the sequence<sup>74</sup>, the diameter of the tunnel, and potential interactions between the nucleotides and residues that line the tunnel and the polypeptide chain.

More recently, several experiments have probed the possibility of tertiary structure formation especially in the vestibule near the exit tunnel, whose volume is large enough for tertiary structure formation of the N-terminal region of the protein. CG simulations, which use either  $C_\alpha$  model<sup>75</sup> or  $C_\alpha$ -SCM<sup>76</sup> and all atom representation of RNA or TIS or four-site model for RNA, have been used to interrogate coupled-synthesis and folding (Fig. 3c). Some general results were found in these simulations: Polypeptide synthesis and folding are not coupled for single-domain proteins, which require the synthesis of complete protein for folding to commence; However, co-translational folding is prevalent in multi-domain proteins in which the N-terminus region is likely to fold as it exits the tunnel. In this case, the *in vivo* folding pathway is expected to be different than *in vitro*; Simulations also suggest that interaction with the ribosome surface decreases folding-pathway diversity and results in a more compact transition state structure<sup>76</sup>.





**Figure 4 | Crowding effects on folding.** (a) Crowding-induced entropic stabilization of the folded states of proteins. Restriction of the extended denatured state ensemble because of volume occupied by crowding raises its free energy to a greater-than-the-folded state. The structures are from  $C_\alpha$ -SCM simulations of a three-stranded  $\beta$ -sheet protein, WW domain. (b) Folding time as a function of the concentration of the crowding agent (black lines) for PGK. The red curve shows folding of WW domain as a function of  $\varphi_c$ . The comparison is meant to illustrate that CG simulations qualitatively explain the measurements. (c) Structure of PGK in the absence of crowding agent (left) and at  $\varphi_c = 0.25$  on the right. The distance between the N and C terminus lobes has been dramatically reduced by crowding. The functional implications are given in the text and in ref. 79. Reproduced with permission from refs 32, 79 (© National Academy of Sciences, USA).

### Towards folding under cellular conditions

Cellular interior is replete with a host of macromolecules, which can alter all processes ranging from transcription to folding. For example, in *E. coli* ribosome (~20.8 nm), polymerases and other protein complexes occupy merely 22% of total volume and small complexes, and other small complexes constitute about 8% of the total volume. Thus, unlike *in vitro* experiments, where folding is studied in an aqueous solution corresponding to infinite dilution conditions, crowding effects have to be taken when describing their behaviour *in vivo*. A simple calculation shows that average spacing between cytoplasmic proteins is ~4 nm (ref. 77) (or  $\rho \geq (50\text{--}400) \text{ mg ml}^{-1}$  (ref. 78)). Given that the diameter of a typical proteins (~300 aa) is ~4 nm, the cell is an extremely crowded place, which severely inhibits conformational fluctuations that are easily realized in typical *in vitro* experiments.

An approximate mimic of the cellular environment can be realized by adding high concentrations of natural or synthetic macromolecules. Consider the simplest case of a crowding agent with radius  $R_c$  (Ficoll 70 for example) that is inert towards the protein or RNA. The volume fraction  $\varphi_c = \rho v$ , where  $v = (4\pi R_c^3/3)$ , the volume of the crowding particle, can be altered by  $R_c$  even with  $\rho$  fixed. The first crowding simulation used a  $C_\alpha$ -SCM of a  $\beta$ -sheet protein in the presence of spherical crowding particles with  $0 \leq \varphi_c \leq 0.25$  (ref. 32). The CG simulations showed that when only excluded volume interactions dominate, stabilities of globular proteins relative to  $\varphi_c = 0$  (ref. 32) are enhanced (Fig. 4a). The extent of stability change, measured using  $\Delta T_m = T_m(\varphi_c) - T_m(\varphi_c = 0)$ , showed that  $\Delta T_m \sim \varphi_c^{1/3v}$

where the  $v (= 3/5)$  is the Flory exponent that characterizes the size of the unfolded states of proteins. The scaling of  $\Delta T_m(\varphi_c)$  with  $\varphi_c$  has been confirmed in recent experiments<sup>79</sup>. Simulations also showed that the folding rate  $k_f(\varphi_c)$ , is also affected when  $\varphi_c \neq 0$ . Rate,  $k_f(\varphi_c)$ , increases monotonically till an optimum value, and subsequently decreases (Fig. 4b). Interestingly, identical behaviour was observed in the dependence of the relaxation rate of phosphoglycerate kinase (PGK) as a function of Ficoll concentration. The simulation results on crowding-induced effects on the smaller  $\beta$ -sheet WW domain explains several aspects of folding of PGK (Fig. 4c).

In an insightful application of simulations, it was recently shown that crowding can alter catalytic activity of kinase (Fig. 4c)<sup>79</sup>. As is common in many kinases, PGKs that transfers phosphate group from diphosphoglycerate to ADP, has a catalytic site between the N- and C-lobes connected by flexible hinge. To perform kinase activity, PGK must undergo a large-scale structural movement that reduces the distance between N- and C-lobes. It was found that PGK activity is increased over 15-fold in  $200 \text{ mg mol}^{-1}$  ( $\varphi_c \approx 0.2$ ) Ficoll 70. The enhancement in activity was attributed to crowding-induced shape change that brings the N- and C-lobes in proximity.

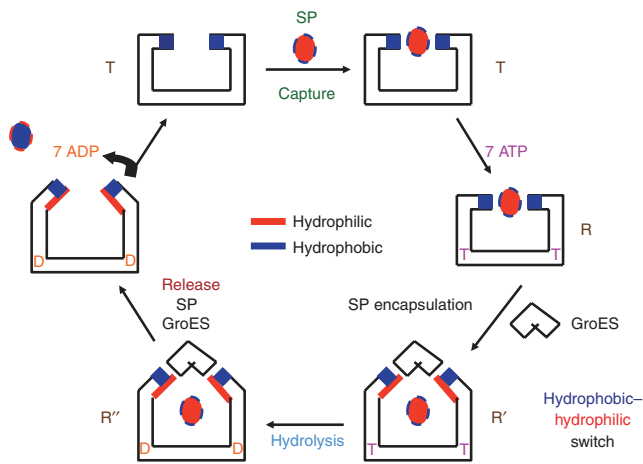
### Biological nanomachines

Biological machines are typically multi-subunit constructs that carry out myriads of functions by interacting with a range of proteins and RNA. Examples of such machines include molecular motors (kinesin, myosin and dynein), *E. coli* chaperonin GroEL,  $F_0F_1$ -ATPase, ribosomes and helicases. A common theme in the function of these systems is they consume energy and, in the process, undergo a reaction cycle that dictates their function. Free energy transduction from chemical energy to mechanical work via a series of conformational switches is the hallmark of biological nanomachines<sup>80</sup>.

**Chaperonin GroEL.** Most of the proteins in cells fold spontaneously. However, molecular chaperones have evolved to rescue a small fraction of proteins, which do not reach their native states easily and hence are destined to aggregate. In *E. coli*, it is estimated that only about 5–10% of the proteins<sup>81</sup> require assistance from the chaperonin GroEL, which has been extensively characterized using experiments and simulations<sup>82</sup>.

GroEL has two heptameric rings that are stacked back-to-back<sup>83</sup>, with each subunit consisting of apical (A), intermediate (I) and equatorial (E) domains. During the reaction cycle, GroEL undergoes a series of a structural (allosteric) transitions on binding of substrate protein (SP), ATP and the co-chaperonin GroES (Fig. 5). In the *T* state, the hydrophobic patches in the A-domain recognize the exposed hydrophobic residues of the misfolded SPs. ATP-binding triggers dramatic domain movements in GroEL resulting in the catalytic sites moving apart, which in turn imparts a stretching force to partially unfold the captured SP. This step is followed by GroES binding, which results in the encapsulation of the SP in the central cavity. The extent of structural changes at the molecular level in each subunit of GroEL (each ring has ~3,850 residues) during the reaction cycle, ( $T \rightarrow R \rightarrow R'' \rightarrow T$ ) was revealed only through CG simulations<sup>18</sup>.

Simulations using the SOP model of the entire heptameric GroEL particle vividly illustrated the conformational changes of GroEL triggered by ATP binding ( $T \rightarrow R$ ) and ATP hydrolysis ( $R \rightarrow R''$ ). Multiple simulation trajectories revealed an unprecedented view of the key interactions that drive the allosteric transitions<sup>18</sup>: (i) A domains rotate counterclockwise in the  $T \rightarrow R$  transition and clockwise in  $R \rightarrow R''$  transition. (ii) Global  $T \rightarrow R$  and  $R \rightarrow R''$  transitions follow two-state kinetics whereas the formation and kinetics of disruption of residue pairs encompass a broad range of time scales. There is an underlying kinetic hierarchy of internal dynamics that govern global transitions. (iii) For both  $T \rightarrow R$  and  $R \rightarrow R''$  transitions,



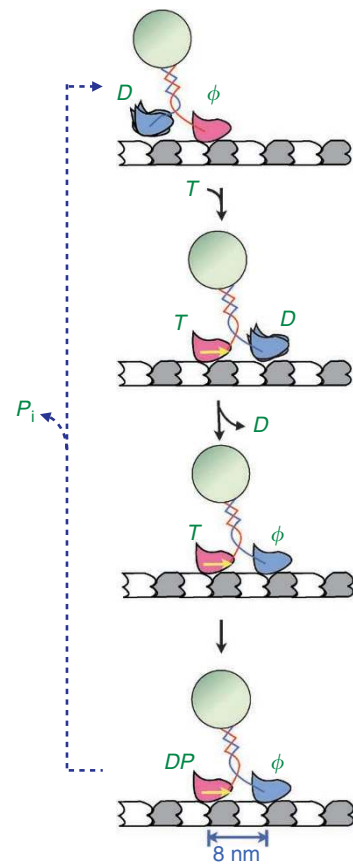
**Figure 5 | Reaction cycle and GroEL function.** The hemicycle of GroEL reaction cycle, which shows that a misfolded SP is captured by GroEL (*E. coli* chaperonin) in the *T* state. This step is followed by reversible ATP-driven transition to the *R* state to which the co-chaperonin GroES can bind to form the *R'* complex, which also results in the SP being encapsulated in the GroEL cavity. The SP can fold by the kinetic partitioning mechanism (KPM). Hydrolysis of ATP, which results in *R''* formation, is followed by an allosteric signal from the bottom ring leads to release of ADP, GroES, and SP (folded or not), thus resetting the top ring to the *T* state.

disruption and formation of salt-bridges are coordinated at multiple sites, which mediate the communication between two neighbouring subunits and synchronize the dynamics of the heptameric ring. (iv) There is a spectacular outside-in movement of two helices accompanied by interdomain salt-bridge formation, which are both solvent exposed in the *R* state. As a result the microenvironment of the SP, which is predominantly hydrophobic in the *T* state becomes progressively hydrophilic as the reaction cycle proceeds.

These large-scale conformational changes are linked to function. As long as misfolded proteins, which typically have exposed hydrophobic regions, are presented to GroEL, they are captured. In the transitions (*T*→*R*), and even more dramatically in *R*→*R''*, the structural changes in the GroEL particle results in the interactions between SP and GroEL from being favourable in the *T* state to unfavourable in the *R''* state. Changes in the microenvironment results in the SP being placed in different part of the folding landscape from which it can fold with some probability during the life time of the *R* and *R''* states (Fig. 5). If the cycle is iterated multiple times, sufficient yield of the SP can be obtained as anticipated by the iterative annealing mechanism<sup>84,85</sup>. It is amusing to note that the mechanism of GroEL function is hauntingly similar to the simulated annealing protocol<sup>86</sup> used in the context of NP hard problems. Not surprisingly, nature has stumbled on it apparently millions of years earlier.

**Kinesins.** Kinesins are motors that transport cellular organelles along the network of cytoskeletal filaments<sup>80,87,88</sup>. Made of two identical motor domains linked by a coiled-coil stalk, kinesins exploit the free energy generated from binding and hydrolysis of ATP to produce the characteristic hand-over-hand stepping motion. A number of SM experiments show that kinesin takes roughly 8 nm step along the polar microtubule (MT) track as it strides towards the (+) end consuming one ATP per step.

Because of the fundamental limitations in experimental resolution, it is difficult to provide molecular explanations of many intriguing observations related to kinesin motility such as the free energy transduction, out-of-phase coordination of the processes occurring at the two-motor domains, and the role of kinesin-MT interactions. For both heads to associate with the MT binding sites, internal ten-

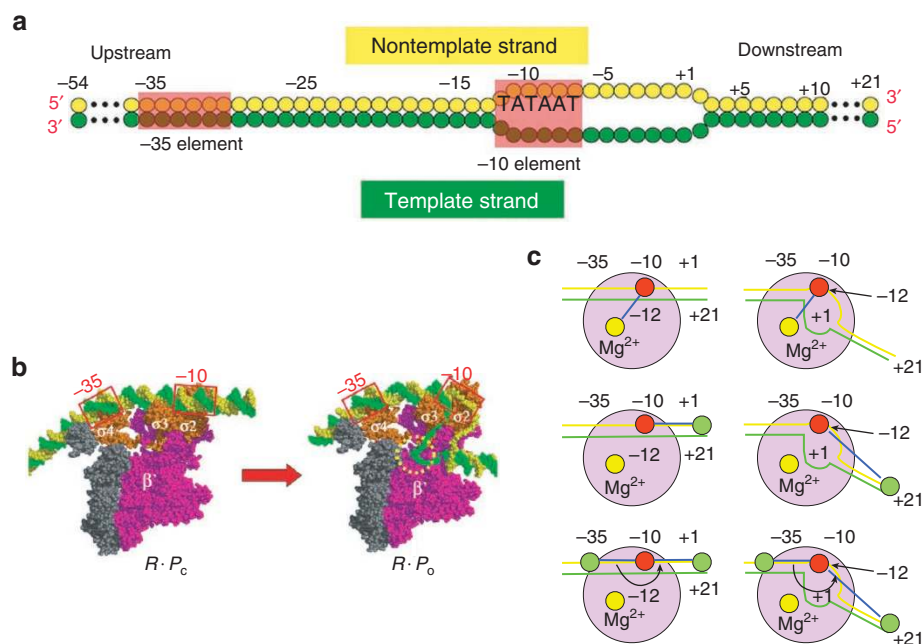


**Figure 6 | Mechanochemical cycle of the conventional kinesin.** The diagram depicts the enzymatic cycle of a dimeric kinesin that generates a single 8-nm step on MT track. The head-to-head regulation via neck-linker results in the out-of-phase coordination of catalytic cycle. *T*, *DP*, *D* and  $\phi$  denote ATP, ADP-Pi, ADP, and nucleotide-free state of the catalytic site. The yellow arrow represents the ordered state of neck-linker.

sion (~(8–15) pN) exerted through the neck-linker deforms the catalytic site from its native-like configuration<sup>23</sup>, thus inhibiting the premature binding of ATP to the nucleotide free leading head. The ATP inhibited state is maintained as long as the two heads remain bound. The deformed leading head catalytic site is restored only after the inorganic phosphate ( $P_i$ ) is released, which changes the trailing head from a strong to a weak binding state. Thus, processivity of kinesin is regulated by strain in the leading head that can be linked to the topology of the kinesin-MT complex. Simplified molecular simulations combined with theoretical ideas have also shed light on the vexing question of whether kinesin takes substeps (Fig. 6)<sup>24</sup>.

**Transcription initiation by bacterial RNA polymerase.** The synthesis of RNA, carried out by DNA-dependent RNA polymerase (RNAP) in a process referred to as transcription, involves several stages. The highly regulated transcription process in eukaryotes is extraordinarily complicated involving a whole zoo of transcription factors which interact with polymerase as it reads the codes on the template strand of DNA to make RNA (Fig. 7). Transcription in bacteria also involves a number of steps. The DNA-dependent RNAP, whose sequence, structures, and global functions are universally conserved from bacteria to man, is the key enzyme in the transcription of genetic information in all organisms. The three major stages in the transcription cycle, which first involves binding of initiation-specific transcription factors to the catalytically competent core of





**Figure 7 | Promoter melting induced by bacterial RNA polymerase.** (a) Schematics of the base pairing between the template and non-template strands of the promoter. Nucleotide positions are numbered relative to the transcription start site, +1. DNA segments that interact with RNAP, -35 and -10 elements, are shaded red. (b) Structural models correspond to  $R \cdot P_c$  (left) and  $R \cdot P_o$  (right). Transcription bubble structure is on the right (DNA non-template strand (yellow) and template strand (green)). (c) Sequence of events in transcription bubble formation, melting, scrunching, and bending process from top to bottom extracted from simulations. Blue lines show the changes in the conformation of the template strand in three major stages in the formation the transcription bubble. Yellow circles embedded within the polymerase represent the position of  $Mg^{2+}$ . Red and green circles mark the positions of the nucleotides, and highlight the processes of melting, scrunching, and bending.

RNAP, to form a holoenzyme are: (i) Initiation, during which initiation-specific  $\sigma$  factor binds to the catalytically competent RNAP to form the holoenzyme. This step is followed by recognition of the promoter DNA to form the closed ( $R \cdot P_c$ ) complex and subsequent transition to the open ( $R \cdot P_o$ ) structure. (ii) Elongation of the transcript by nucleotide addition. (iii) Termination involving cessation of transcription and disassembly of the RNAP elongation complex.

Recently the dynamics of structural transitions that occur during  $R \cdot P_c \rightleftharpoons R \cdot P_o$  transition, which leads to melting of 12 base pairs in the promoter region resulting in the formation of transcription bubble (Fig. 7b)<sup>89</sup> were probed using CG simulations<sup>89</sup>. To perform these simulations, CG model for the 3,122 residue RNAP–DNA complex (15 nm long and 11 nm wide) that is identical to those used to describe GroEL and kinesin dynamics, was used. For DNA, each strand was represented using a single site located at the centre of the nucleotide. Transcription bubble forms in three steps (Fig. 7c): (i) Melting of -10 element on the promoter region. (ii) Scrunching of promoter DNA into RNAP active channel, followed by the formation of bubble; the accommodation of dsDNA into the channel involves an internal RNAP dynamics of a transient expansion of key structural motifs in the  $\beta$  subunit. (iii) Bending of downstream DNA after the unwinding of the dsDNA. Simulation results revealed that internal RNAP dynamics resulting in transient buildup of strain is needed to fully accommodate dsDNA to gain access to the active site. The simulations (for animation, see <http://www.youtube.com/watch?v=Q6QoyDI3TCw>) also make several testable predictions to probe the relationship between RNAP motion and transcription bubble formation.

### Future perspectives

Given that biological problems are complex, it is inevitable that CG models should have a key role in informing experiments. Although not reviewed here, there are a number of areas such as protein aggregation, membrane structure and dynamics<sup>90</sup>, and lipid-membrane

interactions where such simulations have already been profitable<sup>91</sup>. Experimental constraints and theory have been the guiding factors in constructing length-scale-dependent CG models, as a few examples here illustrate. Current methods can be used to provide insights into a number of important biological problems. On a few nm-length scale, corresponding to folding problems, there is a need to study proteins in excess of 200 residues. Modelling counter ion effects to describe RNA folding can be achieved by integrating theoretical ideas from polymer physics and suitable CG models. Although electrostatic interactions have been approximately modelled in simulations of biological machines<sup>18</sup>, further refinements might be needed for more accurate simulations<sup>92</sup>. On longer length scales, there are a number of problems which could profit from CG simulations. Description of motor-driven polymerization and depolymerization kinetics of microtubule and protein-induced polymerization of actin are two examples.

The demand to develop CG models will continue to grow, because there is an appetite to understand the workings of a cell. The increasing attention paid to obtain real-time measurements on how the workers (enzymes, ribozymes, ribosomes, genomes, lipids, membranes and so on) cooperate to execute the demands on the cell is sure to spur interests in models and theories. From a modelling perspective, it is neither possible nor desirable to devise microscopic models when considering events on long length and time scales. In constructing whole-cell models, it may be sufficient to model the various workers as quasi-particles, which interact with each other through connected networks that are dynamically changing depending on the cell status and external stimuli. Such a viewpoint is already being used in systems biology. The lesson from theoretical approaches to problems in condensed matter and material science is that phenomena at different length and time scales require different levels of description. Such a perspective, which also applies to biological problems, will surely spur us on to develop suitable CG models and theories that capture the essence of the problem at hand without being encumbered by unnecessary details.

## References

- Anderson, P. W. *Basic Notions of Condensed Matter Physics* (Westview, 1997).
- This classic monograph describes powerful strategies for tackling the many-body problem by using effective Hamiltonians.**
- Mézard, M., Parisi, G. & Virasoro, M. *Spin Glass Theory and Beyond* (World Scientific, 1988).
- Kirkpatrick, T. & Thirumalai, D. Random solution from regular density functional Hamiltonian—a static and dynamical theory for the structural glass transition. *J. Phys. A* **22**, L149–L155 (1989).
- Lubchenko, V. & Wolynes, P. Theory of structural glasses and supercooled liquids. *Annu. Rev. Phys. Chem.* **58**, 235–266 (2007).
- Alder, B. J. & Wainwright, T. E. Phase transition for a hard sphere system. *J. Chem. Phys.* **27**, 1208–1209 (1957).
- Thirumalai, D. & Hyeon, C. RNA and protein folding: common themes and variations. *Biochemistry* **44**, 4957–4970 (2005).
- Onuchic, J. N. & Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70–75 (2004).
- Shakhnovich, E. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev.* **106**, 1559–1588 (2006).
- Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316 (2008).
- Thirumalai, D., O'Brien, E. P., Morrison, G. & Hyeon, C. Theoretical perspectives on protein folding. *Annu. Rev. Biophys.* **39**, 159–183 (2010).
- Levitt, M. & Warshel, A. Computer simulation of protein folding. *Nature* **253**, 694–698 (1975).
- Honeycutt, J. D. & Thirumalai, D. Metastability of the folded states of globular proteins. *Proc. Natl Acad. Sci. USA* **87**, 3526–3529 (1990).
- Clementi, C., Nymeyer, H. & Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular protein. *J. Mol. Biol.* **298**, 937–953 (2000).
- Karanicolas, J. & Brooks, C. III The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* **11**, 2351–2361 (2002).
- Weinkam, P., Zimmermann, J., Romesberg, F. & Wolynes, P. The folding energy landscape and free energy excitations of cytochrome c. *Acc. Chem. Res.* **43**, 652–660 (2010).
- Hyeon, C. & Thirumalai, D. Mechanical unfolding of RNA hairpins. *Proc. Natl Acad. Sci. USA* **102**, 6789–6794 (2005).
- Whitford, P. C., Geggler, P., Altman, R. B., Blanchard, S. C., Onuchic, J. N. & Sanbonmatsu, K. Y. Accommodation of aminoacyl-tRNA into the ribosome involves reversible excursions along multiple pathways. *RNA* **16**, 1196–1204 (2010).
- Hyeon, C., Lorimer, G. H. & Thirumalai, D. Dynamics of allosteric transition in GroEL. *Proc. Natl Acad. Sci. USA* **103**, 18939–18944 (2006).
- Coarse-grained SOP model was used to study the molecular details in the conformational changes that occur during the reaction cycle of the annealing machine GroEL.**
- Pisliakov, A., Cao, J., Kamerlin, S. & Warshel, A. Enzyme millisecond conformational dynamics do not catalyze the chemical step. *Proc. Natl Acad. Sci. USA* **106**, 17359–17364 (2009).
- Rychkova, A., Vicatos, S. & Warshel, A. On the energetics of translocon-assisted insertion of charged transmembrane helices into membranes. *Proc. Natl Acad. Sci. USA* **107**, 17598–17603 (2010).
- Strajbl, M., Shurki, A. & Warshel, A. Converting conformational changes to electrostatic energy in molecular motors: the energetics of ATP synthase. *Proc. Natl Acad. Sci. USA* **100**, 14834–14839 (2003).
- Koga, N. & Takada, S. Folding-based molecular simulations reveal mechanisms of the rotary motor F<sub>1</sub>-ATPase. *Proc. Natl Acad. Sci. USA* **103**, 5367–5372 (2006).
- Hyeon, C. & Onuchic, J. N. Internal strain regulates the nucleotide binding site of the kinesin leading head. *Proc. Natl Acad. Sci. USA* **104**, 2175–2180 (2007).
- Hyeon, C. & Onuchic, J. N. Mechanical control of the directional stepping dynamics of the kinesin motor. *Proc. Natl Acad. Sci. USA* **104**, 17382–17387 (2007).
- Employing multi-scale CG simulations, this work studied the complex stepping dynamics resulting from the interplay between neck-linker docking and diffusive search dynamics of a kinesin head on the microtubule surface.**
- Takano, M., Terada, T. & Sasai, M. Unidirectional Brownian motion observed in an in silico single molecule experiment of an actomyosin motor. *Proc. Natl Acad. Sci. USA* **107**, 7769–7774 (2010).
- Tehver, R. & Thirumalai, D. Rigor to post-rigor transition in myosin V: link between the dynamics and the supporting architecture. *Structure* **18**, 471–481 (2010).
- Liu, H., Shi, Y., Chen, X. & Warshel, A. Simulating the electrostatic guidance of the vectorial translocations in hexameric helicases and translocases. *Proc. Natl Acad. Sci. USA* **106**, 7449–7454 (2009).
- Kamerlin, S., Vicatos, S., Dryga, A. & Warshel, A. Coarse-grained (multiscale) simulations in studies of biophysical and chemical systems. *Annu. Rev. Phys. Chem.* **62**, 41–64 (2011).
- de Gennes, P. G. *Scaling Concepts in Polymer Physics* (Cornell University Press, 1979).
- Zhou, H. X., Rivas, G. & Minton, A. P. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* **37**, 375–397 (2008).
- Elcock, A. Models of macromolecular crowding effects and the need for quantitative comparisons with experiment. *Curr. Opin. Struct. Biol.* **20**, 196–206 (2010).
- Cheung, M. S., Klimov, D. & Thirumalai, D. Molecular crowding enhances native state stability and refolding rates of globular proteins. *Proc. Natl Acad. Sci. USA* **102**, 4753–4758 (2005).
- Fisher, M. E. Renormalization group in theory of critical behavior. *Rev. Mod. Phys.* **46**, 597–616 (1974).
- Bustamante, C., Marko, J. F., Siggia, E. D. & Smith, S. Entropic elasticity of  $\lambda$ -phase DNA. *Science* **265**, 1599–1600 (1994).
- Valle, F., Favre, M., de Los Rios, P., Rosa, A. & Dietler, G. Scaling exponents and probability distribution of DNA end-to-end distance. *Phys. Rev. Lett.* **95**, 158105 (2005).
- Pastor, R. W., Zwanzig, R. & Szabo, A. Diffusion limited first contact of the ends of a polymer: comparison of theory with simulation. *J. Chem. Phys.* **105**, 3878–3882 (1996).
- Toan, N., Morrison, G., Hyeon, C. & Thirumalai, D. Kinetics of loop formation in polymer chains. *J. Phys. Chem. B* **112**, 6094–6106 (2008).
- Podtelezhnikov, A. A. & Vologodskii, A. V. Dynamics of small loops in DNA molecules. *Macromolecules* **33**, 2767–2771 (2000).
- Vologodskii, A., Levene, S., Klenin, K., Frank-Kamenetskii, M. & Cozzarelli, N. Conformational and thermodynamic properties of supercoiled DNA. *J. Mol. Biol.* **227**, 1224–1243 (1992).
- Podtelezhnikov, A. & Vologodskii, A. Simulations of polymer cyclization by Brownian dynamics. *Macromolecules* **30**, 6668–6673 (1997).
- Jun, S., Bechhoefer, J. & Ha, B.-Y. Diffusion-limited loop formation of semi-flexible polymers: Kramers theory and the intertwined time scales of chain relaxation and closing. *Europhys. Lett.* **64**, 420–426 (2003).
- Hyeon, C. & Thirumalai, D. Kinetics of interior loop formation in semiflexible chains. *J. Chem. Phys.* **124**, 104905 (2006).
- Cloutier, T. & Widom, J. Spontaneous sharp bending of double-stranded DNA. *Mol. Cell* **14**, 355–362 (2004).
- Du, Q., Smith, C., Shiffeldrim, N., Vologodskaya, M. & Vologodskii, A. Cyclization of short DNA fragments and bending fluctuations of the double helix. *Proc. Natl Acad. Sci. USA* **102**, 5397–5402 (2005).
- Savelyev, A. & Papoian, G. Chemically accurate coarse graining of double-stranded DNA. *Proc. Natl Acad. Sci. USA* **107**, 20340–20345 (2010).
- Smith, S. B., Finzi, L. & Bustamante, C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**, 1122–1126 (1992).
- Jun, S. & Mulder, B. Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome. *Proc. Natl Acad. Sci. USA* **103**, 12388 (2006).
- Using CG models and polymer physics concepts an entropy-driven mechanism for DNA segregation in bacterial cells is proposed.**
- Jun, S. & Wright, A. Entropy as the driver of chromosome segregation. *Nat. Rev. Microbiol.* **8**, 600–607 (2010).
- Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2**, 292–301 (2001).
- Grosberg, A., Rabin, Y., Havlin, S. & Neer, A. Crumpled globule model of the three-dimensional structure of DNA. *Europhys. Lett.* **23**, 373 (1993).
- Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289 (2009).
- A polymer model that represents ~1Mb DNA as a CG interaction center was used to explain the interphase DNA organization of the human genome, which is postulated to be fractal globule.**
- Grosberg, A., Nechaev, S. & Shakhnovich, E. The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys.* **49**, 2095–2100 (1988).
- Thirumalai, D., Lee, N., Woodson, S. A. & Klimov, D. K. Early events in RNA folding. *Annu. Rev. Phys. Chem.* **52**, 751–762 (2001).
- Dima, R. I., Hyeon, C. & Thirumalai, D. Extracting stacking interaction parameters for RNA from the data set of native structures. *J. Mol. Biol.* **347**, 53–69 (2005).
- Moghaddam, S. *et al.* Metal ion dependence of cooperative collapse transitions in RNA. *J. Mol. Biol.* **393**, 753–764 (2009).

56. Koculi, E., Hyeon, C., Thirumalai, D. & Woodson, S. A. Charge density of divalent metal cations determines RNA stability. *J. Am. Chem. Soc.* **129**, 2676–2682 (2007).
57. Russell, R. *et al.* Rapid compaction during RNA folding. *Proc. Natl Acad. Sci. USA* **99**, 4266–4271 (2002).  
**This work used a CG model to infer the structures of T. ribozyme at each stage of collapse dynamics by selecting a conformation that best explain the SAXS profile.**
58. Ma, H., Proctor, D. J., Kierzek, E., Kierzek, R., Bevilacqua, P. C. & Gruebele, M. Exploring the energy landscape of a small RNA hairpin. *J. Am. Chem. Soc.* **128**, 1523–1530 (2006).
59. Hyeon, C. & Thirumalai, D. Multiple probes are required to explore and control the rugged energy landscape of RNA hairpins. *J. Am. Chem. Soc.* **130**, 1538–1539 (2008).
60. Chen, S. J. & Dill, K. A. RNA folding energy landscapes. *Proc. Natl Acad. Sci. USA* **97**, 646–651 (2000).
61. Thirumalai, D. & Klimov, D. K. Deciphering the time scales and mechanisms of protein folding using minimal off-lattice models. *Curr. Opin. Struct. Biol.* **9**, 197–207 (1999).
62. O'Brien, E., Ziv, G., Haran, G., Brooks, B. & Thirumalai, D. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc. Natl Acad. Sci. USA* **105**, 13403 (2008).
63. Liu, Z., Reddy, G., O'Brien, E. & Thirumalai, D. Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. *Proc. Natl Acad. Sci. USA* **108**, 7787 (2011).  
**The first demonstration that coarse-grained models can be used to characterize the folding thermodynamics and kinetics of a globular protein in the presence of denaturants.**
64. Tinoco, I. Jr., Li, P. T. & Bustamante, C. Determination of thermodynamics and kinetics of RNA reactions by force. *Q. Rev. Biophys.* **39**, 325–360 (2006).
65. Fernandez, J. M. & Li, H. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science* **303**, 1674–1678 (2004).
66. Hyeon, C., Morrison, G. & Thirumalai, D. Force dependent hopping rates of RNA hairpins can be estimated from accurate measurement of the folding landscapes. *Proc. Natl Acad. Sci. USA* **105**, 9604–9606 (2008).
67. Hyeon, C., Morrison, G., Pincus, D. L. & Thirumalai, D. Refolding dynamics of stretched biopolymers upon force-quench. *Proc. Natl Acad. Sci. USA* **106**, 20288–20293 (2009).
68. Dudko, O. K., Hummer, G. & Szabo, A. Intrinsic rates and activation free energies from single-molecule pulling experiments. *Phys. Rev. Lett.* **96**, 108101 (2006).
69. Hyeon, C. & Thirumalai, D. Can energy landscape roughness of proteins and RNA be measured by using mechanical unfolding experiments? *Proc. Natl Acad. Sci. USA* **100**, 10249–10253 (2003).
70. Hyeon, C., Dima, R. I. & Thirumalai, D. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure* **14**, 1633–1645 (2006).
71. Mickler, M., Dima, R. I., Dietz, H., Hyeon, C., Thirumalai, D. & Rief, M. Revealing the bifurcation in the unfolding pathways of GFP by using single-molecule experiments and simulations. *Proc. Natl Acad. Sci. USA* **104**, 20268–20273 (2007).
72. Harms, J. *et al.* High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell* **107**, 679–688 (2001).
73. Moore, P. B. & Steitz, T. A. The structural basis of large ribosomal subunit function. *Annu. Rev. Biochem.* **72**, 813–850 (2003).
74. Ziv, G., Haran, G. & Thirumalai, D. Ribosome exit tunnel can entropically stabilize  $\alpha$ -helices. *Proc. Natl Acad. Sci. USA* **102**, 18956–18961 (2005).
75. Elcock, A. Molecular simulations of cotranslational protein folding: fragment stabilities, folding cooperativity, and trapping in the ribosome. *PLoS Comput. Biol.* **2**, 824–841 (2006).
76. O'Brien, E. P., Hsu, S. T. D., Christodoulou, J., Vendruscolo, M. & Dobson, C. M. Transient tertiary structure formation within the ribosome exit port. *J. Am. Chem. Soc.* **132**, 16928–16937 (2010).
77. Phillips, R., Kondev, J. & Theriot, J. *Physical Biology of the Cell* (2009).
78. Minton, A. The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. *J. Biol. Chem.* **276**, 10577 (2001).
79. Dhar, A. *et al.* Structure, function, and folding of phosphoglycerate kinase are strongly perturbed by macromolecular crowding. *Proc. Natl Acad. Sci. USA* **107**, 17586–17591 (2010).  
**This work shows that the presence of crowders is critical in shaping the functional state of phosphoglycerate kinase.**
80. Vale, R. D. & Milligan, R. A. The way things move: looking under the hood of molecular motor proteins. *Science* **288**, 88–95 (2000).
81. Lorimer, G. A quantitative assessment of the role of the chaperonin proteins in protein folding *in vivo*. *FASEB J.* **10**, 5–9 (1996).
82. Thirumalai, D. & Lorimer, G. H. Chaperonin-mediated protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 245–269 (2001).
83. Xu, Z., Horwich, A. L. & Sigler, P. B. The crystal structure of the asymmetric GroEL-GroES-(ADP), chaperonin complex. *Nature* **388**, 741–750 (1997).
84. Todd, M. J., Lorimer, G. H. & Thirumalai, D. Chaperonin-facilitated protein folding: optimization of rate and yield by an iterative annealing mechanism. *Proc. Natl Acad. Sci. USA* **93**, 4030–4035 (1996).
85. Tehver, R. & Thirumalai, D. Kinetic model for the coupling between allosteric transitions in groel and substrate protein folding and aggregation. *J. Mol. Biol.* **4**, 1279–1295 (2008).
86. Kirkpatrick, S., Gelatt, D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
87. Vale, R. D., Reese, T. S. & Sheetz, M. P. Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. *Cell* **42**, 39–50 (1985).
88. Hirokawa, N. Kinesin and dynein superfamily proteins and the mechanism of organelle transport. *Science* **279**, 519–526 (1998).
89. Chen, J., Darst, S. A. & Thirumalai, D. Promoter melting triggered by bacterial RNA polymerase occurs in three steps. *Proc. Natl Acad. Sci. USA* **107**, 12523–12528 (2010).  
**SOP model of RNA polymerase and dsDNA was used to reveal the molecular mechanism of melting of promoter region upon polymerase binding, which is the first step in transcription initiation.**
90. 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* **111**, 7812–7824 (2007).
91. Straub, J. & Thirumalai, D. Principles governing oligomer formation in amyloidogenic peptides. *Curr. Opin. Struct. Biol.* **20**, 187–195 (2010).
92. Messer, B., Roca, M., Chu, Z., Vicatos, S., Kilshtain, A. & Warshel, A. Multiscale simulations of protein landscapes: using coarse-grained models as reference potentials to full explicit models. *Proteins: Struct. Funct. Bioinf.* **78**, 1212–1227 (2010).  
**The study demonstrates how different simulation strategies can be used to obtain insights into dynamics of complex biological systems.**
93. Papoian, G., Ulander, J. & Wolynes, P. Role of water mediated interactions in protein-protein recognition landscapes. *J. Am. Chem. Soc.* **125**, 9170–9178 (2003).
94. Veitshans, T., Klimov, D. & Thirumalai, D. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold. Des.* **2**, 1–22 (1997).

## Acknowledgements

This work was supported by a grant from Korea Research Foundation grants (2010-0000602 and KRF-C00180) (C.H.) and the National Science Foundation (CHE09-14033) (D.T.). The list of references is not exhaustive and should be viewed as a guide to the literature.

## Additional information

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**How to cite this article:** Hyeon, C. & Thirumalai, D. Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun.* **2**:487 doi: 10.1038/ncomms1481 (2011).