Whole Cell Simulations of Processes Related to Protein Translation

Principal Investigator

Zaida Luthey-Schulten

Department of Physics, Department of Chemistry, Center for the Physics of Living Cells, Beckman Institute for Advanced Science and Technology, and Carl R. Woese Institute for Genomic Biology; University of Illinois, Urbana, IL, USA

Collaborators

Thomas Kuhlman

Department of Physics, Center for the Physics of Living Cells; University of Illinois, Urbana, IL, USA

James R. Williamson

Skaggs Institute for Chemical Biology, Department of Integrative Structural and Computational Biology, Department of Chemistry; Scripps Research Institute, La Jolla, CA, USA

Corresponding author: Zaida Luthey-Schulten <zan@illinois.edu>

Executive summary

Central to all life is the assembly of the ribosome: a process involving the association of ~50 proteins to 3 RNA molecules in a hierarchical, coordinated manner. Through the synthesis of data from many disparate sources including *in vitro* kinetics, cryo-electron tomography, and genomics data, we have developed a spatially resolved, stochastic model of the biogenesis of the ribosomal small subunit in *Escherichia coli*. Using our GPU-accelerated Lattice Microbes (LM) software on Blue Waters, we have observed biogenesis in modeled cells at timescales up to 2 hours. Key intermediates in the assembly pathway inferred from *in vitro* kinetics were investigated using molecular dynamics. We then extend this whole-cell model to include cell growth and DNA replication. To accomplish all this, high performance computing using GPU-based accelerators was necessary. Blue Waters provided the necessary computing power to complete this research as well as directed the development of LM to maximize performance. In total, we used 457,000 node hours.

Contents

1 Ribosome biogenesis in *Escherichia coli* 2

1.1 Whole-cell simulations . . . 4

1.1.1 *In vivo* ribosome biogenesis . . . 4

1.1.2 *In vivo* ribosome biogenesis with growth and DNA replication . . . 5

1.2 Molecular dynamics . . . 9

2 Elongation factor Tu 11

3 Continuing development of Lattice Microbes 13

3.1 Reaction kernel evaluation . . . 13

3.2 Increased particle density . . . 13

3.3 Multi-node parallelism through MPI . . . 14

4 Population flux balance analysis of *Saccharomyces cerevisiae* 14

1 Ribosome biogenesis in *Escherichia coli*

Ribosomes, the molecular machines responsible for protein synthesis, are fundamental to all life. The precise synthesis and assembly of a ribosome requires the association of ~50 individual proteins to the three strands of ribosomal RNA (rRNA) in order to form the two halves of the ribosome: the large- (LSU) and small- (SSU) subunits. This process is referred to as ribosomal biogenesis, and is composed of four steps (36): (1) the transcription of rRNA from the multiple copies of ribosomal DNA, which is the rate limiting step for the ribosome biogenesis *in vivo* (41); (2) the synthesis of the ribosomal proteins (r-proteins), which is regulated on the translational level based on organization of the r-protein operons in the genome; (3) post-transcriptional processing and modification of both the rRNA and r-proteins; and (4) the highly coordinated assembly of r-proteins and rRNAs towards the mature ribosomal subunits. Ribosomal assembly requires the cooperation of many molecular components. It requires that the r-proteins which are translated in different regions of the cell to find and bind with the rRNA in the correct order of assembly. In addition, approximately 20 assembly cofactors are engaged to facilitate the process at various assembly stages

The intricacy of ribosome assembly first attracted Nomura et al (44), who originally mapped out the hierarchical dependency of the r-proteins binding to the *Escherichia coli* 16S rRNA using equilibrium reconstitution experiments (Fig. 1a). Later developments boosted our understanding of the protein assisted dynamics of RNA folding (10, 31, 37), and the kinetic cooperation of protein binding (24, 28, 40) during *in vitro* ribosome self-assembly. Both aspects suggested that assembly of the *E. coli* 30S subunit proceeds through multiple parallel pathways nucleated at different positions on the 16S rRNA, and a 5′ to 3′ directionality during 30S ribosomal assembly is always present. However, conflicting reports of the order in which r-proteins associate to the nascent subunit hamper our investigation of ribosome assembly under *in vivo* conditions. Therefore, a comprehensive model that captures the topology of the protein–RNA interaction network is needed to decipher the underlying rules governing the assembly of the ribosome.

Ribosome biogenesis is a fundamental biological process present in all three domains of life. For this reason alone it is of great interest to basic science. However, a detailed understanding of the process will impact the study of other biological systems. For example, cellular metabolism is tightly coupled to ribosome biogenesis since cell growth relies on protein synthesis. The details of these interactions and their integrated effects on the organism level are not completely understood. Due to the complexity of the chemical reaction networks involved, progress can only be

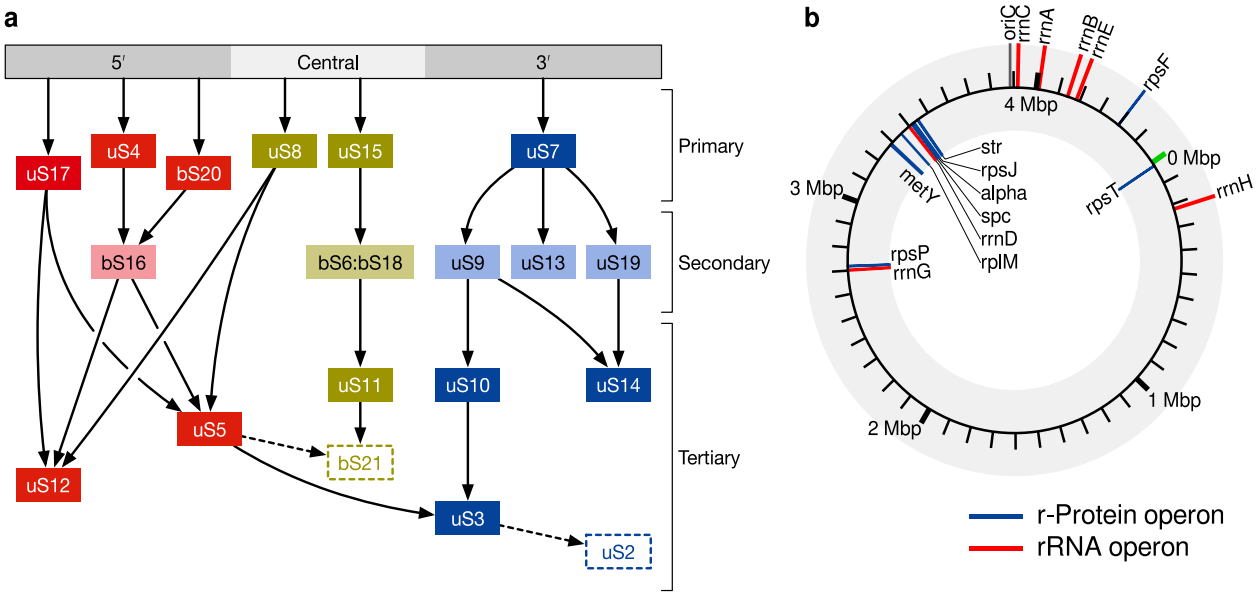

FIGURE 1 (a) Graph of thermodynamic binding dependencies of r-protein to the 16S rRNA (44). Only the major dependencies used in our models are depicted here. The arrows indicate binding dependencies, e.g. for the case of A→B, A must be bound to the intermediate before B can bind. Proteins uS2 and bS21, shown in open rectangles, are not included in our models due to the lack of available kinetic data. (b) Diagram of the *E. coli* genome showing the location of r-protein and rRNA operons associated with the assembly of the SSU. Red lines indicate the location of rRNA operons, where as blue lines indicate the location of r-protein operons. The replication of the genome begins at *oriC* (12 o'clock) and progresses on both sides of the chromosome until both replication forks reach 6 o'clock, where upon replication completes and the two daughter chromosomes separate.

made through the design and study of detailed kinetic models which capture the complex interactions between cellular processes.

Blue Waters was crucial to the investigation of this problem due to the size of our simulations. Our software suite to simulate stochastic reaction-diffusion systems, Lattice Microbes (LM), requires NVIDIA GPUs, and was readily modified to take full advantage of the available K20X accelerators. These simulations are I/O intensive, generating trajectory files approximately 10 GB in size per trajectory, of which many are needed for a proper statistical analysis of the results. Blue Waters provided all the technical resources necessary for us to study this system.

1.1 Whole-cell simulations

To study our model of ribosomal biogenesis at the whole-cell level we employ LM, a suite of highly-parallel GPU-accelerated algorithms for stochastic simulation of complex biochemical reaction networks under realistic cellular conditions (5, 8, 15, 16, 22, 27). The software includes implementations of both chemical master equation (CME) and reaction-diffusion master equation (RDME) sampling algorithms, including a unique multiple particle diffusion (MPD-RDME) (27), which factors the diffusion operator in to x-, y-, and z-components, allowing for more opportunities for parallelization. The CME describes the time-evolution of the probability of the system being in a given state—the copy number of each chemical species—under the assumption of spatial homogeneity. It can be thought of as a special case of the more general RDME which governs the time-evolution of the system's spatially resolved state—the copy numbers found in each subvolume of the lattice. By discretizing space and treating diffusion as hops between lattice sites, LM trajectories are capable of reaching hour long timescales—orders of magnitude longer than competing codes (17, 23, 33, 38).

Briefly, realistic models of crowded cellular environments based on cryo-electron tomography and proteomics data are coarse-grained to a cubic lattice, where initial species and reactions are assigned to each subvolume. The stochastic, spatially resolved trajectories of the system are implicitly sampled from the solution to the RDME using a Gillespie-type Monte Carlo approach:

$$\begin{aligned}\frac{\partial P(\boldsymbol{x},t)}{\partial t} = &\sum_{v\in V}\sum_{r=1}^{R} -a_r(x_v)P(\boldsymbol{x},t) + a_r(x_v - S_r)P(\boldsymbol{x} - S_r\mathbf{1}_v,t) \\ &+ \sum_{i\in V}\sum_{j\in V}\sum_{\alpha=1}^{N} -d_{ij}^{\alpha}x_i^{\alpha}P(\boldsymbol{x},t) + d_{ji}^{\alpha}(x_j^{\alpha}+\mathbf{1}_j^{\alpha})P(\boldsymbol{x}+\mathbf{1}_j^{\alpha}-\mathbf{1}_i^{\alpha},t) \end{aligned} \quad (1)$$

Here, x_v^{α} is the number of molecules of species α ($\alpha = 1,\cdots,N$) in subvolume v ($v \in V$). R is the number of reactions. a_r is the reaction propensity for reaction r given the state of a subvolume x_v. S is the stoichiometry matrix. d_{ij}^{α} is the diffusive propensity for one molecule of species α to diffuse from subvolume i to j. The first line on the right hand side describes chemical reactions between species in each subvolume; it is similar to the CME, in effect assuming spatial homogeneity within each subvolume during each time step. The second line on the right hand side describes diffusion of particles into and out of each subvolume. The MPD-RDME methodology assumes that diffusive events are the most frequent events occurring in a simulation, with several likely to take place roughly simultaneously in subvolumes spread across the simulation volume. By taking short time steps such that active particles are unlikely to take part in multiple reactions, the subvolumes are rendered independent, and can be calculated in parallel on many-core machines (details on methodology and implementation can be found in (8, 16, 27)).

1.1.1 *In vivo* ribosome biogenesis

We have completed the design, simulation, and analysis of a spatially resolved, stochastic model of ribosomal biogenesis in *E. coli* (2). The assembly of the small subunit requires the binding of 20 individual proteins to the 16S rRNA. If all possible binding reactions are treated fairly, a combinatorial explosion of network complexity results. This leads to on the order of 10^6 intermediates and 10^{18} reactions. We instead construct a model that explicitly uses the thermodynamic protein binding dependencies summarized in the Nomura map ((44), Fig. 1a) to decreases the size of the network to

a manageable level of 1,633 species and 7,000 reactions, taking effective r-protein binding rates published by James R. Williamson's group performed at two *in vitro* conditions: low temperature (15 °C) (24) and high temperature (40 °C) (25). The low temperature data provided us with binding rates which are dependent on the identity of the intermediate; the high temperature data provided only binding rates independent of intermediate identity, however also included predicted intermediates using cryo-electron microscopy. Though the low temperature model was of little physiological significance since the optimal growth temperature for *E. coli* is 37 °C, it possesses a clear assembly pathway and suggested key assembly intermediates to study using MD (§ 1.2).

To prepare for RDME simulation using LM, we reduced the network size of the high temperature model. Using an ODE-based representation of the reduced assembly network, we identified intermediates which were unlikely to contribute significantly to the overall assembly process. The effect of removing these species was tested by removing them sequentially in order from the network in order of lowest to highest significance, and comparing the concentration time course of the modified network to the original 1,633 species network. This process led to an assembly network consisting of 145 intermediates and 325 reactions.

From this assembly model, we constructed a model of the biogenesis of the SSU which includes the transcription of rRNA and mRNA coding for the SSU ribosomal proteins, translation of proteins, and degradation of mRNA. This model of biogenesis was then embedded in a spacial model of *E. coli*, derived from cryo-electron tomography data (22), which takes into account the geometry of the cell as well as its compartmentalization into distinct regions (Fig. 2a). The cell—4.0 μm long and 0.9 μm in diameter—was constructed in a $32 \times 32 \times 128$ lattice with a lattice spacing of 32 nm. The simulated volume is divided into four regions: extracellular, membrane, cytoplasm and nucleoid. The particles in the cytoplasm have a higher diffusion rate than in the nucleoid to account for molecular crowding arising from the chromosome. To account for the partial exclusion of translating ribosomes from the nucleoid region (11), the transition rate from nucleoid to cytoplasm was set to 0.25× the reverse transition. Species representing the ribosomal genes were placed in the nucleoid along the long axis of the cell at positions determined from the location of the gene in the *E. coli* genome (Fig. 1b). This stochastic, spatially resolved model was simulated using LM on Blue Waters. We investigated the statistics of the relatively fast process of assembly (10 min. simulated time) over 128 replicates, as well as the behavior over an entire cell cycle (120 min. simulated time) over 16 replicates. To use Blue Waters effectively, modifications to LM to improve performance were necessary (see § 3).

We discovered that a simple kinetic model of SSU assembly *in vitro* can be incorporated into a model of ribosome biogenesis with no changes to the experimentally derived association rate constants. The differences in intracellular concentrations compared to the *in vitro* experiments is enough to make up for the difference in assembly timescales (2 hours *in vitro*, compared to 90 seconds *in vivo*.) Our *in vivo* model estimated the mean SSU assembly time within a factor of two, as well as predicting that the assembly time distribution roughly follows a gamma distribution (Fig. 3d). This model predicts the major assembly pathways of the SSU in terms of protein binding order and identifies the important assembly intermediates.

Starting with cells containing no 16S-protein intermediates (i.e. only LSU, SSU, 16S rRNA, and protein), we ran 128 10-minute trajectories to gather statistics on the assembly process *in vivo*. To simplify the visualization of the intermediate distributions, we analyzed the copy number of multiple intermediate configurations clustered by their abundance versus time profiles (Fig. 3a), and plotted their spatial distribution along the xz-plane (Fig. 3b). We were able to identify spatial localization of early SSU intermediates (Fig. 3c) to the nucleoid region near the rRNA operon of origin. These intermediates are composed of states where only the primary and secondary proteins of the 5′ and central domains are bound. We also observed spatial heterogeneity of the mRNA distributions, where messenger is only found near the originating operon and outside of the nucleoid region (Fig. 3e). This appears to arise due to the imposed exclusion of ribosomes from the nucleoid region in our model. To date, we have used 9,680 node hours on these simulations.

1.1.2 *In vivo* ribosome biogenesis with growth and DNA replication

Our latest work on this model involves the addition of growth, DNA replication, and cell division (1). Using estimates of when replication begins relative to cell division and how long it takes to replicate the genome from experiments performed by our experimental collaborator Thomas Kuhlman, we extended the previous model to include gene replication and cell growth. Duplication of the chromosome affects the copy number distribution of gene products

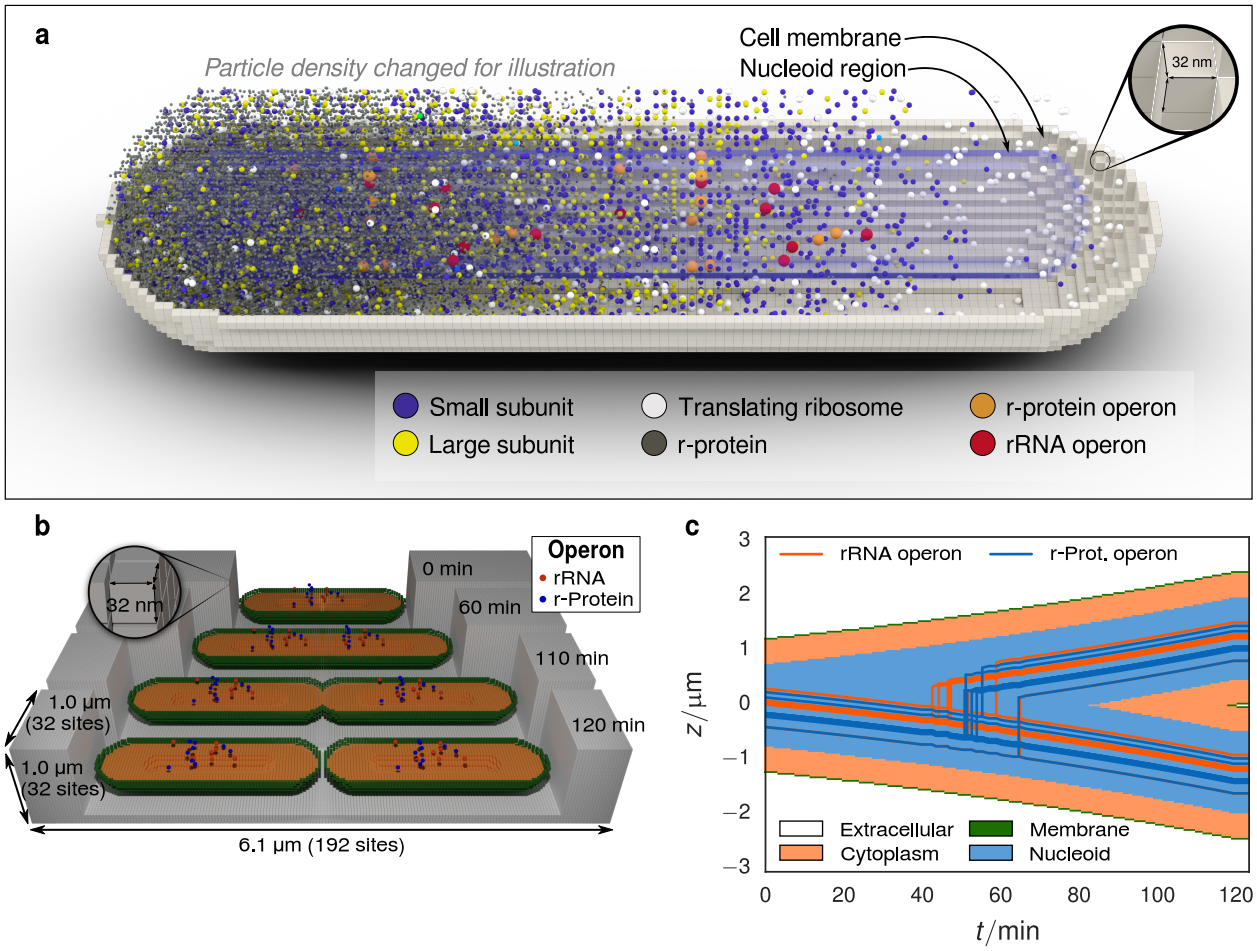

FIGURE 2 (a) Cut away of a representative cell configuration. The particle density has been modified for illustrative purposes, however the left end of the cell presents a representation of the true particle density. Ribosomal protein operons locations (orange) are fixed within the nucleoid region. Messengers (not visible) are transcribed from these sites and diffuse to find and associate with 30S particles (purple) upon which a 50S subunit (yellow) joins the complex forming a translating ribosome (white). The ribosome translates r-protein (gray) which diffuse away and assembly with nascent SSU intermediates (not visible) which are based on 16S rRNA transcribed from the rRNA operons (red). Translating 70S particles are excluded from the nucleoid region through a bias in their intercompartmental transition rates. (b) Schematic of geometry used in the replicating cell simulations. The lattice is $32 \times 32 \times 192$ sites in the x, y, and z directions respectively, with a lattice spacing of 32 nm. The simulation volume consists of 4 regions: (1) extracellular space (gray), (2) membrane (green), (3) cytoplasm (orange), and (4) nucleoid (not colored, found in center of cytoplasm). Operon species are placed within the nucleoid region based on their genomic loci and replicated at times predicted by the replication model. The position of the operon species is evolved in time such that the operons in the daughter cell are found in the same position as the operons in the mother cell. The cell volume grows constantly throughout the cell cycle at an exponential rate from 2.4 μm to 4.7 μm, where upon it divides into two daughter cells of length 2.4 μm. (c) Kymograph showing the evolution of spatial compartments and operon locations over one cell cycle. The jagged steps arise from the discreteness imposed by the 32 nm lattice.

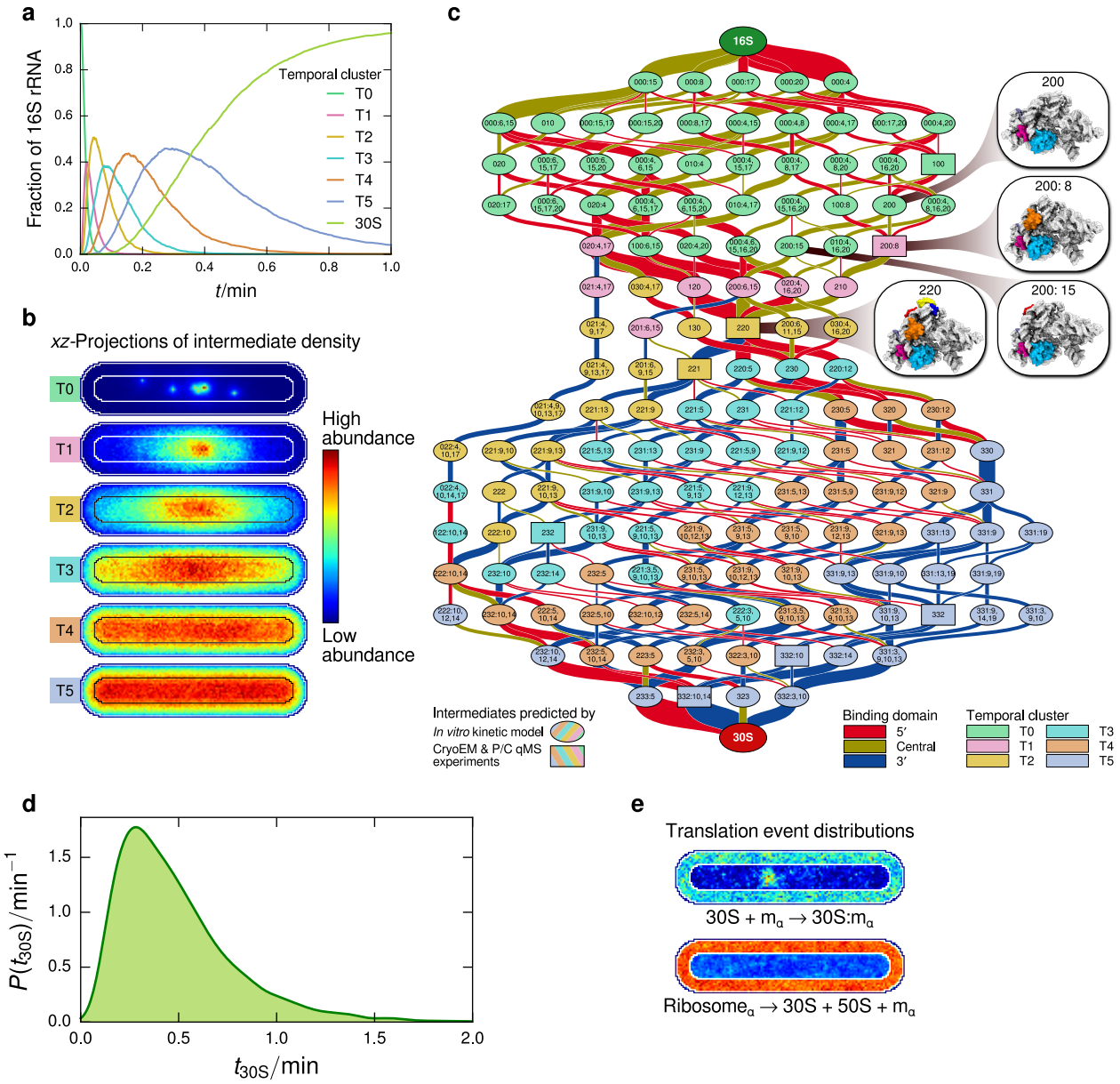

FIGURE 3 (a) Fraction of assembly intermediates as a function of time. The traces represent the collective count of intermediates which have been clustered into classes with similar abundance versus time profiles. (b) Projections of the intermediate spatial probability distributions for the 6 clusters onto the xz axis. Early intermediates (T0 and T1) are so short-lived that they remain near their originating rRNA operon. (c) Reduced network for 30S assembly at 40 °C. Each node is an assembly intermediate, labeled according to which proteins are bound. A three digit number describes the set of r-proteins bound to each domain (5′-, central-, and 3′- respectively), and all remaining r-proteins are listed after the three digit number (2). The edges connecting the intermediates represent the r-protein binding reactions. The width represents the total amount of intermediate converted by that reaction, and the color indicates the binding domain of that protein (5′–red, central–yellow, and 3′–blue.) Predicted assembly intermediates from pulse/chase qMS and cryoEM (25) are represented using rectangles. Intermediates investigated with MD simulations (§ 1.2) are indicated by the callouts. (d) Distribution of assembly times for the SSU. The shape of the distribution takes the form of a gamma distribution with a mean assembly time of 30 s—comparable to to the experimentally determined *in vivo* assembly time of 1.3–3.5 minutes. (e) Distribution of mRNA–30S association (top), and translation termination (bottom) events, showing the spatial dependence of translation. From central y slices, it is clear that the translating ribosomes and even the initial 30S/mRNA assembly complexes are excluded from the nucleoid region. Since 30S particles can enter the nucleoid, transcription initiation can occur near the DNA from which the messenger is transcribed, and in the cytoplasm from mRNA is released from ribosomes terminating translation.

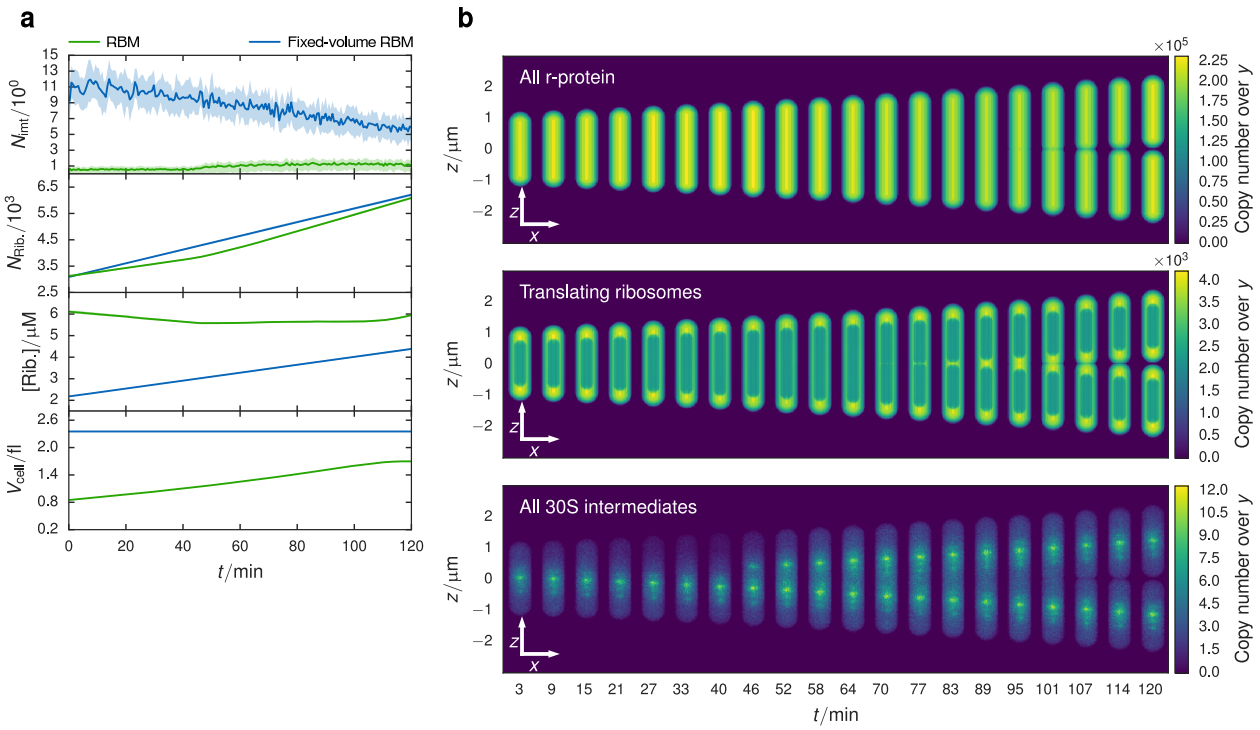

FIGURE 4 (a) Comparison between constant cell (Fixed-volume RBM, blue) and replicating cell (RBM, green) models over 16 replicates. Means are represented by solid lines and the interquartile range is given by the shaded area. There is a significantly lower average SSU intermediate count seen in the replicating cell model compared to the constant cell model (top panel), which is a result of the changing cell volume. In the last three panels are plotted the absolute count of ribosomes (translating as well as dissociated), the absolute concentration of ribosomes, and the cell volume. The replicating cell model produces ribosomes at approximately the same pace as volume expansion, leading to a constant ribosome concentration over the cell cycle. (b) xz copy number projections of cells at evenly spaced times throughout the cell cycle. The time evolution of the cell geometry is evident in this series of projections. Division begins approximately 90 minutes into the cell cycle through the constriction of the cell membrane. Ribosomal protein (top) diffuses rapidly through all compartments, leading to a distribution which mirrors the thickness of the cell at each (x, z) coordinate. The transition rates of translating ribosomes (middle) between the nucleoid and cytoplasm regions is biased to limit the number of ribosomes in the nucleoid. The most pronounced example of spacial heterogeneity in these models is due to the 30S intermediates (bottom), where the earliest intermediates which result from the binding of primary proteins are found near the rRNA operon from which the 16S rRNA was transcribed.

since the rate of transcript production effectively doubles (9), however to correctly capture the shape of the distribution the behavior over the interval where the mRNA copy number relaxes to the new steady state under the doubled gene dose must be accounted for (7). The changing cell volume also impacts the copy number dynamics due to the dependence of reaction rates on volume. The effect of these processes were investigated by deterministically modifying the cell geometry, operon locations, and copy numbers over the course of the simulated cell cycle.

To allow for a direct comparison between the original (constant) model and the augmented (replication) model, all transcription rates were scaled by a factor of 1.8 and the birth rate of LSU particles was scaled by 0.55 to ensure that the same copy numbers of r-protein and ribosomes were maintained in both models at the beginning and end of the cell cycle. Over a 120 minute cell cycle the cell grows from 2.3 μm to 4.7 μm, using a lattice size of $32 \times 32 \times 192$ with a 32 nm lattice spacing (Fig. 2b). Genes are replicated in the simulation at times determined by using the start time and duration of DNA replication, and the position of the gene in the genome (Fig. 1b). Their location within the cell is moved linearly across the z-axis such that their position in the daughter cell mirrors their original positioning (Fig. 2c). Though this is not at all correct biologically, the trajectory of gene loci in the space of the cell during gene replication is not known and this approximation is sufficient for the scope of the study.

Comparing the two models, the initial and final species counts are practically identical for all classes of particles with the exception of the 30S intermediates. We see that the initial intermediate count in the constant model is approximately 6× the count seen in the replicating cell model (Fig. 4a). The origin of this effect is due in part to the increased protein concentration at the start of the cell cycle in the replicating cell model. Though the absolute protein numbers are approximately equal, the replicating cell model volume is smaller than the constant cell volume at the start of the cell cycle. However there appear to be other effects at play since the volume difference of 1.8× is not enough to account for the total difference.

The changing volume due to cell growth causes particle concentrations to remain relatively constant throughout the cell cycle (Fig. 4a). For example the ribosome concentration in the replicating cell model spans 5.5–5.9 μM, where as in the constant model the concentration spans 2.4–4.0 μM. However in the replicating cell model the concentration tends to peak before and after cell division (Fig. 4a). In the bottom panel of Fig. 4a, the increase in volume slows down near the end of the cell cycle when the cell begins dividing. The ribosome number increases linearly over the whole cell cycle, however the growth of the cell volume can no longer keep the pace with ribosome production during this slowing, leading to an increase in ribosome concentration at the end of the cell cycle. When the cell finally divides, the protein concentration can now relax to the steady state concentration.

Though the majority of the chemical species in the RDME simulations show no spatial heterogeneity, e.g. r-protein (Fig. 4b, top), two classes of particles exhibit nonuniform distributions throughout the cell. Translating ribosomes, composed of a 30S particle, a 50S particle, and a mRNA, are excluded from the nucleoid region using biased transition rates. The transition rate from the nucleoid region to the cytoplasm is 4× greater than the reverse transition. This is to model the excluded volume effects arising from the chromosome. This effect has been observed in single particle tracking experiments, which showed that the fully associated ribosome is partially excluded from the nucleoid (11, 18) region while the individual subunits are not (11).

The other particle class exhibiting a nonuniform spatial distribution is the 30S assembly intermediates (Fig. 4b bottom). Ribosomes assemble in a well-defined binding order, where some proteins can only bind once other proteins are associated with the nascent subunit. The earliest 30S intermediates, consisting of the primary and secondary binding proteins associated with the 5′ and central domains of the 16S rRNA (2), are short lived and are found only within a few hundred nanometers of the site from which the rRNA was transcribed. Due to their short lifetime, their density tracks the position of the rRNA operon tightly. Later intermediates which are beginning to include tertiary binding proteins diffuse farther away from the originating rRNA operon until all memory of their birthplace is washed out.

The lack of checkpointing in the augmented simulations precluded utilizing Blue Waters for these simulations initially. The simulations could not be checkpointed since the state necessary to describe growth and DNA replication was not under the control of LM. We have recently developed an checkpointing mechanism independent of LM, and have begun executing production runs on Blue Waters to support the revisions to our manuscript (1). To date, we have used 17,768 node hours on the replication model simulations.

1.2 Molecular dynamics

To complement our whole-cell modeling in (2), we performed molecular dynamics simulations of crucial intermediates in the assembly map (Fig. 3c). In these models, assembly began with the binding of all of the primary and secondary r-proteins in the 5′ domain leading to state 200 (see definition of state nomenclature in the caption to Fig. 3c). State 200 is the bifurcation point in which both assembly pathways begin to branch. The majority of assembly pathways complete the central domain before starting the 3′ domain; however, a minority of the pathways involve the reverse order.

To understand the structural basis for the bifurcation in the assembly pathways, we used MD simulations to probe conformations near the state 200 bifurcation point (Table I) starting with the crystal structure of the complete *E. coli* SSU[1]. All states contain the intact 16S rRNA and are prebound with r-proteins uS4, uS17, bS20, and bS16 while states 201, 200:8, and 200:15 have in addition uS7, uS8, and uS15 bound respectively. To observe the maximum fluctuations

[1]PDB: 2I2P

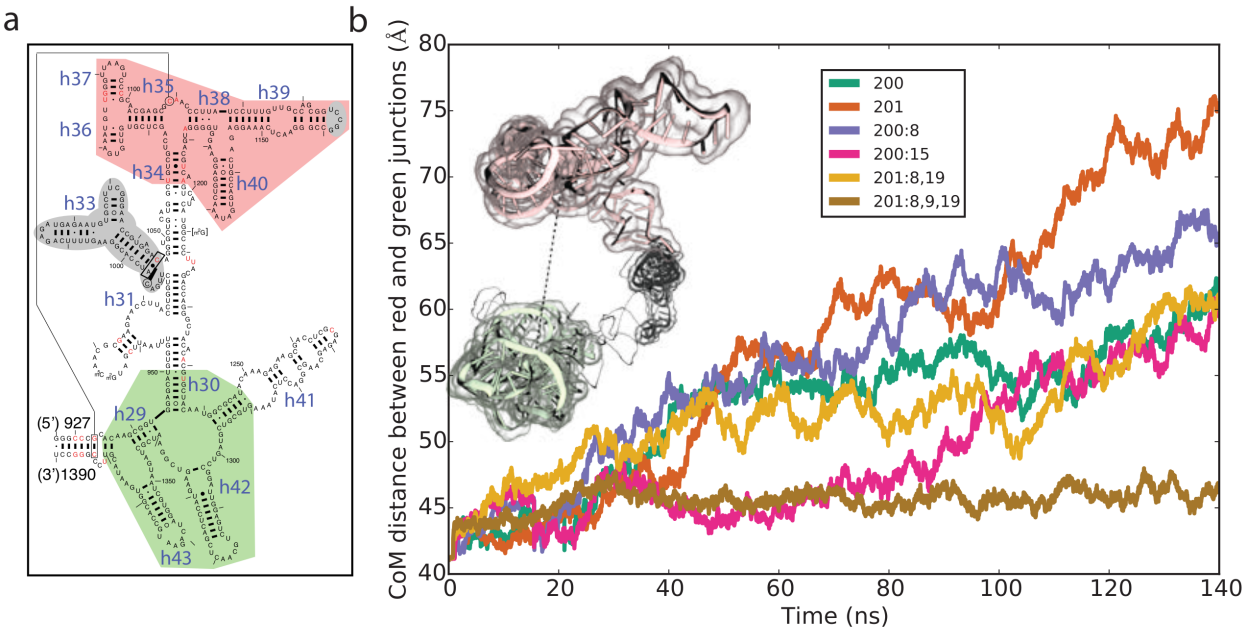

FIGURE 5 (a) Secondary structure diagram of the 3′ domain with the center of masses defined. Center of masses are computed from the lower four-way junction helices h29, h30, h41-h43 (green) and the upper three-way junction helices h34-h40 (red). The exact residues are marked on the modified secondary structure diagram (42). These centers are separated by the structural signature—marked in gray circles—h33 and numerous sequence signatures (34). (b) Time traces of center of mass distances in the 3′ domain.

in the nucleic acid conformations, we prepared the MD simulations with a neutralizing concentration of sodium ions with no magnesium ions present.

In the 3′ domain, all four simulations exhibited similar motions. These fluctuations are dominated by the partial unfolding of the 3′ domain. Helices in the lower four-way junction (h29, h30, h41-h43) separate from helices in the upper three-way junction (h34-h40) (Fig. 5a). Time traces of the center of masses for the different junctions in all four MD simulations show that the helices separate from 40 Å to over 60 Å after 140 ns (Fig. 5b). Simultaneously, the structural signature (34) h33 separates from h31 and h32 and becomes more solvent exposed. This is expected since h33 is connected to these junctions. Similar results are seen in simulations with the *Thermus thermophilus* small subunit[2], suggesting that these motions are probably common to all bacterial organisms. The fact that states 200, 201, 200:8 or 200:15 all have similar motions suggests that there is no strong bias to binding either uS7, uS8, or uS15 and that the next major assembly barrier, the opening of the 3′ domain, occurs further along in the assembly pathway.

Because the binding of uS7 and uS8 have a minimal effect globally on the structure of ribosome assembly intermediates, we probed the effect of adding two 3′ domain binding r-proteins (uS9 and uS19). In the folded ribosomal small subunit, uS9 binds to both the lower four-way- and upper three-way-junctions while uS19 binds to the structural signature h33. As the uS19 binding site is more local than uS9, we probed its binding first uS19. Adding uS19 to the simulations (moving from state 200:8 to 201:8, 19), tightens the structural signature in h33 and keeps h33 packed against h31-h32 and like the four previous simulations, state 201:8, 19 also shows similar unfolding of the 3′ domain (Fig. 5b). State 201:8, 9, 19, on the other hand, does not have the separation in the 3′ domain (Fig. 5b). Interestingly, all six MD simulations showed the 3′ domain rotating away from the five-way junction in the 5′ domain, suggesting that there is another folding barrier further along in the assembly pathway. This motion might only be arrested upon the addition of uS5.

[2]PDB: 1HR0

TABLE I Summary of MD simulations performed using Blue Waters. All systems have the following 5′ domain r-proteins prebound: uS4, uS17, bS20, and bS16. These calculations used 200,000 node-hours over our current and previous allocations.

		Bound r-proteins				
Index	States	Central domain	3′ domain domain	Number of atoms	Dimensions	Simulation time (ns)
1	200	-	-	1,046,000	$182 \times 202 \times 290$	140
2	201	-	uS7	1,041,000	$181 \times 202 \times 289$	140
3	200:8	uS8	-	1,027,000	$179 \times 201 \times 289$	140
4	200:15	uS15	-	1,031,000	$179 \times 201 \times 290$	140
5	201:8, 19	uS8	uS7,uS19	1,052,000	$180 \times 205 \times 290$	140
6	201:8, 9, 19	uS8	uS7,uS9,uS19	1,011,000	$176 \times 200 \times 290$	140

2 Elongation factor Tu

The universally conserved heterotrimeric GTPase, elongation factor Tu, (EF-Tu) is responsible for bringing charged tRNAs to the ribosome. To fulfill this function, EF-Tu cycles through a number of defined conformations. Starting from a GTP-bound conformation, EF-Tu binds to charged tRNA forming a ternary complex EF-Tu·GTP·tRNA. Upon binding to the ribosome A-site, the ternary complex becomes activated through recognition between the mRNA codon and tRNA anti-codon. Activation initiates GTP hydrolysis and causes the dissociation of the ternary complex, releasing the tRNA for accommodation. Finally, EF-Tu·GDP dissociates from the ribosome and can begin the next cycle after being recharged with GTP.

Crystal structures of the pre-[3] and post-hydrolysis[4] EF-Tu·tRNA complexes show the different rearrangements of the EF-Tu domains. Nonetheless, the atomic mechanism describing how the hydrolysis induced structural change causes the tRNA release remains unknown. Using steered molecular dynamics, well-tempered metadynamics and path collective variables (35), we sought to understand the energetics and mechanism of the EF-Tu post-hydrolysis conformational transition.

Briefly, path collective variable uses two metrics, $S(\chi)$ and $Z(\chi)$, to describe the evolution of a system along and orthogonal to a particular pathway respectively. $S(\chi)$ and $Z(\chi)$ have the following form:

$$S(\chi) = \lim_{\lambda \to \infty} \frac{\int_0^1 t e^{-\lambda(\boldsymbol{R}[\chi] - \boldsymbol{R}[\chi](t))^2} dt}{\int_0^1 e^{-\lambda(\boldsymbol{R}[\chi] - \boldsymbol{R}[\chi](t))^2} dt} \tag{2}$$

and

$$Z(\chi) = \lim_{\lambda \to \infty} -\frac{1}{\lambda} \ln \int_0^1 e^{-\lambda(\boldsymbol{R}[\chi] - \boldsymbol{R}[\chi](t))^2} dt \tag{3}$$

where χ refers to the atomic coordinates of the system and R[χ] to the RMSD between the system and the pathway. The initial path connecting the GTP to GDP form of EF-Tu is generated by picking structural snapshots, separated from each other by ~2 Å in RMSD space, from a SMOG trajectory (29). This initial pathway is subsequently refined using several rounds of steered molecular dynamics (SMD).

Using 2D umbrella sampling, we explored the free energy surface of the EF-Tu transition. Umbrellas were placed on a regular grid along S (from 1 to 24 every 0.5) and Z (from 0 to 0.6 every 0.1). Free energy calculations suggest that the first barrier (~0.5 kcal/mol) in the EF-Tu conformational change is the reorientation of the domains (Fig. 6ab) while the positioning of the switch regions (Fig. 6 top, red and orange) are separated by a larger barrier (~1 kcal/mol; Fig. 6bc). Simulations are still ongoing and have not yet converged.

In total, 271,000 node hours were used for this project. A paper describing the conformational changes of EF-Tu following GTP hydrolysis is currently in preparation (4).

[3]PDB: 1B23
[4]PDB: 1TUI

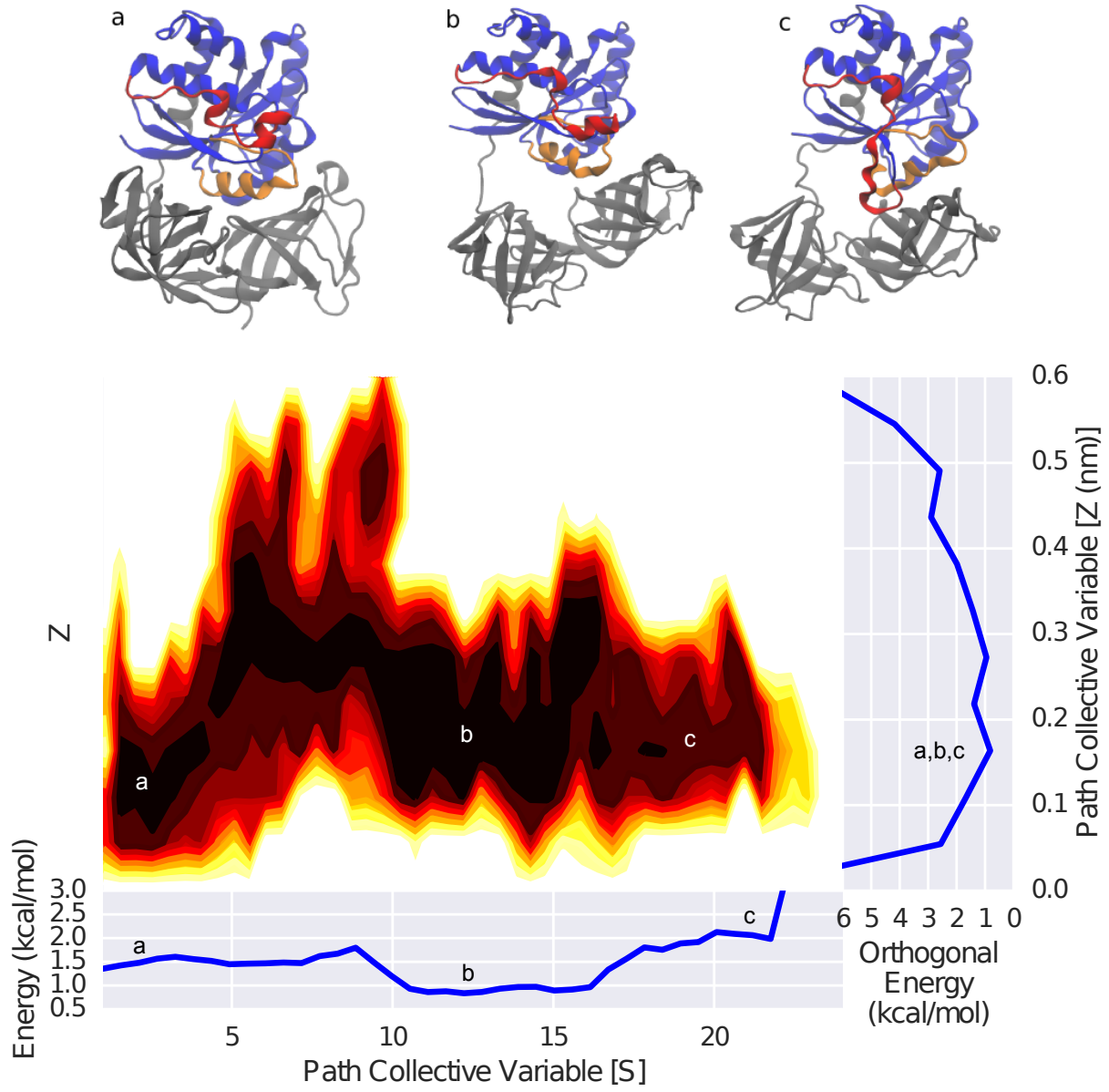

FIGURE 6 Free energy surface of EF-Tu. Example structures of EF-Tu are shown: G-domain (blue), OB-folds (gray), Switch I (red) and Switch II (orange). In (a), G-domain and OB-folds are packed together; (b) G-domain and OB fold separate; (c) Switch I and II move downward. All-atom model build using CHARMM36 (19) and run at 1 atm and 300 K with a 2 fs timestep for 80 ns. System contains 95,000 atoms and neutralized with 150 mM KCl + 4 mM $MgCl_2$. Calculations are run with GROMACS 5.0.4 (6), NAMD 2.10 (39) and Plumed 2.2 (26).

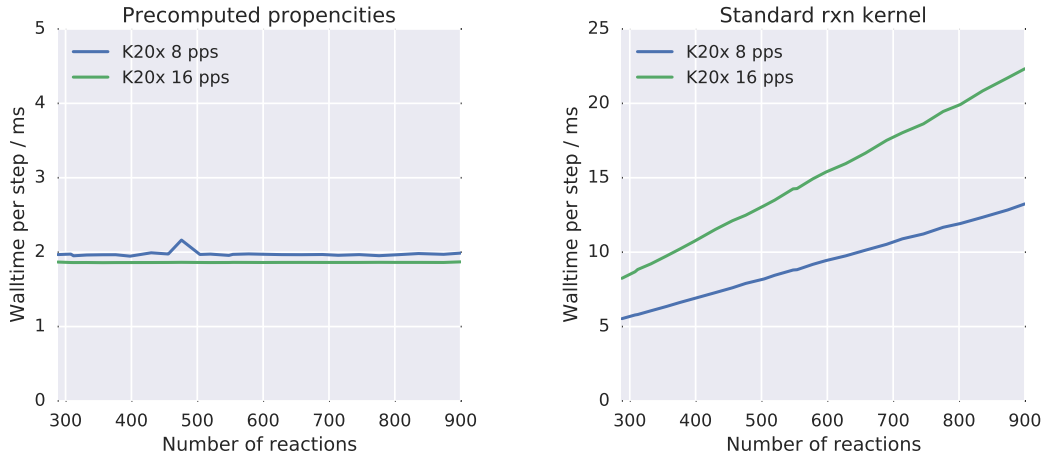

FIGURE 7 Performance comparison between reaction kernels and site capacity. In the ribosome biogenesis simulations, the new reaction kernel outperformed the original kernel due to the large number of model reactions. The simulations using a site capacity of 16 particles per site were more performant in spite of the increased memory usage due to the high particle density causing many overflows in the 8 particle per site lattice.

3 Continuing development of Lattice Microbes

3.1 Reaction kernel evaluation

The evaluation of the reaction kernel, the GPU-based computation responsible for simulating the stochastic nature of chemical reactions in each subvolume, is significantly more computationally intensive in the ribosome biogenesis models than earlier systems simulated with LM due to its sheer reaction count (1336). This model was the most complicated chemical reaction network simulated using Lattice Microbes (LM) to date. The version of Lattice Microbes used initially (LM 2.2) performed poorly due to the large number of reactions, taking 18 ms to perform a step using the K20X. At a timestep of 25 μm, a 120 minute cell cycle would take over eight weeks of computer time. To improve the performance, we employed a programmatic code generation technique to generate a bespoke reaction kernel for each model, replacing majority of the constant memory lookups with immediate value loads and unrolled all loops whose bounds were defined by the reaction network. This decreased the runtime on the K20X to 4.0 ms per time step, allowing for 2 hours of simulation time to be finished in 12 days. This work was prior to the environment upgrade to CUDA 7.0 which introduced NVRTC, enabling just-in-time (JIT) compilation support for CUDA applications. Preliminary work is under way to enable use of NVRTC in LM (5), allowing for dynamic code generation as an option to accelerate the reaction kernel for future simulations utilizing Blue Waters.

Further improvements were made resulting in LM 2.3a leading to even faster reaction kernel evaluations. This new reaction kernel dropped the evaluation time per step from 4.0 ms to 1.0 ms for the ribosomal biogenesis model by using a particle-based propensity calculation (5) (see Fig. 7 for comparison using ribosome biogenesis model.) This technique performs better on models with a large number of reactions, whereas the JIT-based acceleration is superior for smaller chemical reaction networks. The replicating cell simulations ran slower (1.7 ms/timestep) than the constant cell simulations (1.0 ms/timestep) for reasons yet to be determined. This figure is a measurement of the kernel runtime and does not include the overhead due to the lattice modifications needed for cell growth.

3.2 Increased particle density

To allow for greater particle densities, we adapted LM to support lattices with larger site occupancies. Originally LM was restricted to 8 particles per site. We have doubled this limit to allow for 16 particles per site. When more particles occupy a lattice site than capacity allows, the extra particles are said to have “overflowed” and special handling is

required to rectify the situation. A procedure on the CPU locates candidate neighboring lattice sites and moves the excess particles into them. This is costly as the lattice must be from copied to host memory and then back to the GPU after overflows are corrected. Additionally, a higher capacity lattice incurs a cost as well, since the diffusion and reaction operators must access a larger amount of memory to account for the greater number of particles. However, simulations that experience overflows on a frequent basis benefit from the greater capacity, as the cost of accessing more memory is offset by the savings gained from not needing to perform overflow handling. Due to high particle densities resulting from free r-protein and the relatively small lattice size in our ribosome biogenesis simulations, we experienced a minor speedup by using 16 particles per site in our growth simulations (Fig. 7.)

3.3 Multi-node parallelism through MPI

We have already enabled LM to use multiple GPUs attached to a single computer to accelerate time-to-solution for an MPD-RDME simulation (8). While LM can currently utilize MPI in a cluster environment in order to run multiple independent replicates of a model simultaneously, a major goal is to extend the use of MPI to allow a single MPD-RDME simulation to aggregate GPUs from multiple nodes. This will allow us to exploit multiple levels of parallelism in a single simulation: the GPU's inherent capacity for parallel execution and the coordinated use of multiple XK nodes across Blue Waters. Lattice Microbes 2.3a extends the multi-GPU algorithm (8) over multiple nodes using MPI, where a single rank on each node controls the GPU to participate in a one-dimensional spatial decomposition of the simulation domain. Future work will focus on multi-dimensional decomposition strategies to improve scalability over large numbers of GPU equipped nodes.

Preliminary work is complete on porting the MPI feature for MPD-RDME onto Blue Waters, as demonstrated by the benchmark results in figure Fig. 8. Four benchmark systems are used to study the run-time scaling over multiple nodes over a range of biologically-relevant system sizes. The *E. coli* benchmark is a $64 \times 64 \times 128$ lattice, dividing *E. coli* is a $64 \times 64 \times 256$ lattice, yeast is a 256^3 lattice, and the blood cell benchmark is a 512^3 lattice. Each system contains the same particle density, and a test network of four reactions.

Through the development of these improvements, we have consumed 63 node hours.

4 Population flux balance analysis of *Saccharomyces cerevisiae*

A cell's phenotype—its set of distinguishing observable traits—can be as much an emergent property of the cell's environment and gene expression state as it is a result of the cell's genotype. While some observables, like an organism's response to Gram staining, can be immutable and tied to specific genes, others can be more fluid, varying from cell-to-cell with the random fluctuations in each cell's molecular makeup (12, 21, 30, 32) A cell might by chance over- or under-express the enzymes involved in a given biosynthetic pathway, in which case the over- or underproduction of that pathway's end product might signify a naturally occurring phenotype. Understanding this type of phenotypic variability requires models capable of connecting comprehensive gene expression profiles with cellular function.

Population flux balance analysis (popFBA), wherein protein counts are sampled from experimental distributions in order to apply constraints to a genome-scale model of metabolism, was recently used to describe metabolic heterogeneity in *Escherichia coli* (14). Here we extend the method in order to account for correlations in protein expression arising from co-regulation and apply it to the Yeast 7.6 (13) metabolic reconstruction of *Saccharomyces cerevisiae*. We find excellent agreement between the predicted intracellular fluxes and those determined experimentally by a recent ^{13}C fluxomics experiment (20). In particular, popFBA recovers the Crabtree effect (the reliance on fermentative metabolism even in the presence of oxygen) where unconstrained FBA predicts mainly respiration. We also find a broad distribution of growth rates that match experimental observations, and we characterize several metabolic phenotypes within our modeled populations that make use of sometimes complex reaction pathways in order to grow given the internal (enzyme-associated) and external (substrate-associated) constraints imposed. Finally we investigate the degeneracy of the sets of protein-associated constraints that are necessary to give rise to growth rate distributions that agree with experiment. We find that a core set of between 30 and 55 constraints, depending on growth medium, are essential but that additional constraints are still necessary to reproduce the observed growth curves.

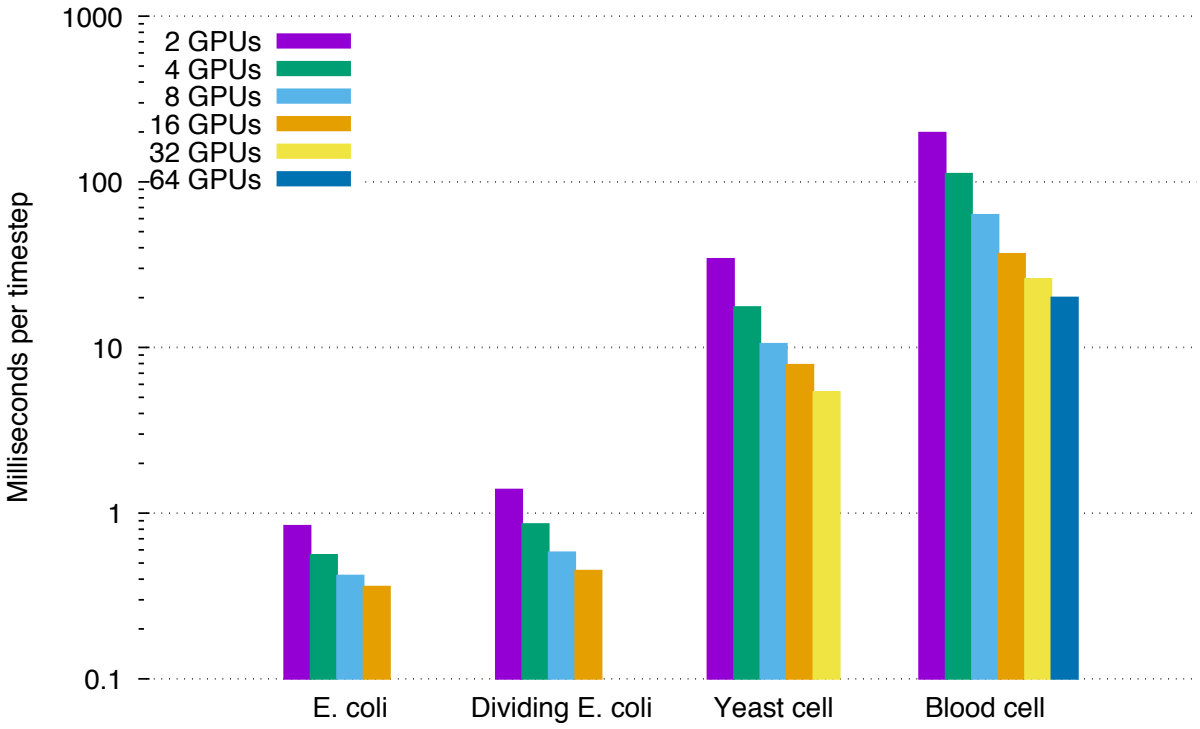

FIGURE 8 MPD-RDME over MPI Benchmark results. Four benchmark systems that are representative sizes of systems of interest— an *E. coli* cell, an *E. coli* cell undergoing cell division, a small yeast cell, and a red blood cell. Each benchmark system contains a network of four reactions, and maintain a constant density of particles. The average wall-clock time for each simulation step is shown for each of the XK node counts tested. MPI-acceleration is crucial for simulating large, eukaryotic cells.

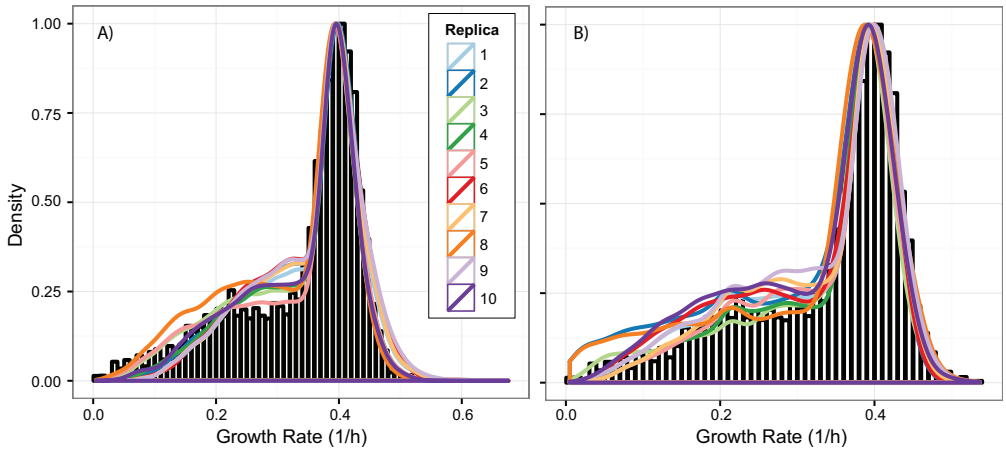

FIGURE 9 Growth Rate Distributions After GA Optimizations. The plots show a comparison between the observed growth rate distribution (20) (black bars) and the distributions obtained after 10 GA optimizations (colored lines) using: (A) The Yeast 7.6 model in SD media conditions and (B) The Yeast 7.1.1 model in ^{13}C media conditions.

One of the main achievements of popFBA is the ability to predict growth rate distributions based on initial molecular constraints imposed on the metabolic map. These growth rate distributions display diverse subpopulations, each one with characteristic metabolic activities. Since popFBA depends on an initial selection of experimentally determined constraints, namely the turnover numbers (K_{cat}) from enzymes present in the metabolic map, a new stochastic search method was developed to determine how sensitive the final growth distribution is to the initial constraint selection. Blue Waters was used to run the extensive calculations needed to randomly search the space of possible constraint sets by running a parallel implementation of a Micro Genetic Algorithm (43). After obtaining 10 different constraint sets that produce growth rate distributions which exactly match the experimentally observed distributions (Fig. 9), we could see that very few constraints are completely incompatible with the experimentally observed distribution, i.e. no solution was found which used that constraint. Moreover, no individual K_{cat} is absolutely essential to producing realistic growth rate distributions, meaning no constraint was used in all 10 optimizations.

Performing these optimizations consumed 1,400 node hours.

Publications resulting from Blue Waters usage

[1] Earnest TM, Cole JA, Peterson JR, Hallock MJ, Kuhlman TE, and Luthey-Schulten Z (2016) "Ribosome biogenesis in replicating cells: integration of experiment and theory". *Biopolymers*, 105(10): 735–751. doi:10.1002/bip.22892.

[2] Earnest TM, Lai J, Chen K, Hallock MJ, Williamson JR, and Luthey-Schulten Z (2015) "Toward a whole-cell model of ribosome biogenesis: Kinetic modeling of SSU assembly". *Biophys. J.*, 109(6): 1117–1135. doi:10.1016/j.bpj.2015.07.030.

[3] Labhsetwar P, Melo MCR, Cole J, and Luthey-Schulten Z. "Population FBA reveals metabolic phenotypes in yeast". In preparation.

[4] Lai J, Ghaemi Z, Tran T, and Luthey-Schulten Z. "Free energy calculations of the post-hydrolysis conformational transition in EF-Tu (*title TBD*)". In preparation.

References

[5] Hallock MJ and Luthey-Schulten Z (2016) "Improving reaction kernel performance in lattice microbes: particle-wise propensities and run-time generated code". In "Parallel and Distributed Processing Symposium Workshop (IPDPSW), 2016 IEEE International", Accepted.

[6] Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, and Lindahl E (2015) "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". *SoftwareX*, 1-2: 19–25. doi:10.1016/j.softx.2015.06.001.

[7] Peterson JR, Cole JA, Fei J, Ha T, and Luthey-Schulten ZA (2015) "Effects of DNA replication on mRNA noise". *Proc. Natl. Acad. Sci. USA*, 112(52): 15886–15891. doi:10.1073/pnas.1516246112.

[8] Hallock MJ, Stone JE, Roberts E, Fry C, and Luthey-Schulten Z (2014) "Simulations of reaction diffusion processes over biologically-relevant size and time scales using multi-gpu workstations". *Parallel Comput.*, 40: 86–99. doi:10.1016/j.parco.2014.03.009.

[9] Jones DL, Brewster RC, and Phillips R (2014) "Promoter architecture dictates cell-to-cell variability in gene expression". *Science*, 346(6216): 1533–1536. doi:10.1126/science.1255301.

[10] Kim H, Abeysirigunawarden SC, Chen K, Mayerle M, Ragunathan K, Luthey-Schulten Z, Ha T, and Woodson SA (2014) "Protein-guided RNA dynamics during early ribosome assembly". *Nature*, 506(7488): 334–338. doi:10.1038/nature13039.

[11] Sanamrad A, Persson F, Lundius EG, Fange D, Gynna AH, and Elf J (2014) "Single-particle tracking reveals that free ribosomal subunits are not excluded from the Escherichia coli nucleoid". *Proc. Natl. Acad. Sci. USA*, 111(31): 11413–11418. doi:10.1073/pnas.1411558111.

[12] Assaf M, Roberts E, Luthey-Schulten Z, and Goldenfeld N (2013) "Extrinsic noise driven phenotype switching in a self-regulating gene". *Phys. Rev. Lett.*, 111: 058102. doi:10.1103/PhysRevLett.111.058102.

[13] Aung HW, Henry SA, and Walker LP (2013) "Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism". *Ind. Biotech.*, 9(4): 215–228. doi:10.1089/ind.2013.0013.

[14] Labhsetwar P, Cole JA, Roberts E, Price ND, and Luthey-Schulten ZA (2013) "Heterogeneity in protein expression induces metabolic variability in a modeled Escherichia coli population". *Proc. Natl. Acad. Sci. USA*. doi:10.1073/pnas.1222569110.

[15] Peterson JR, Hallock MJ, Cole JA, and Luthey-Schulten ZA (2013) "A problem solving environment for stochastic biological simulations". In "PyHPC 2013", Supercomputing 2013. doi:10.13140/2.1.3207.7440.

[16] Roberts E, Stone JE, and Luthey-Schulten Z (2013) "Lattice Microbes: high-performace stochastic simulation method for the reaction-diffusion master equation". *J. Comp. Chem.*, 3: 245–255. doi:10.1002/jcc.23130.

[17] Schöneberg J and Noé F (2013) "ReaDDy - a software for particle-based reaction-diffusion dynamics in crowded cellular environments". *PLoS ONE*, 8(9): e74261. doi:10.1371/journal.pone.0074261.

[18] Bakshi S, Siryaporn A, Goulian M, and Weisshaar JC (2012) "Superresolution imaging of ribosomes and RNA polymerase in live Escherichia coli cells". *Molecular Microbiology*, 85(1): 21–38. doi:10.1111/j.1365-2958.2012.08081.x.

[19] Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, and Alexander D MacKerell J (2012) "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles". *J. Chem. Theor. Comput.*, 8(9): 3257–3273. doi:10.1021/ct300400x.

[20] Levy SF, Ziv N, and Siegal ML (2012) "Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant". *PLoS Syst. Biol.*, 10(5): e1001325. doi:10.1371/journal.pbio.1001325.

[21] MacNeil LT and Walhout AJ (2011) "Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression". *Genome Res.*, 21(5): 645–657. doi:10.1101/gr.097378.109.

[22] Roberts E, Magis A, Ortiz JO, Baumeister W, and Luthey-Schulten Z (2011) "Noise contributions in an inducible genetic switch: A whole-cell simulation study". *PLoS Comput. Biol.*, 7(3): e1002010. doi:10.1371/journal.pcbi.1002010.

[23] Andrews SS, Addy NJ, Brent R, and Arkin AP (2010) "Detailed simulations of cell biology with Smoldyn 2.1". *PLoS Comput. Biol.*, 6(3): e1000705. doi:10.1371/journal.pcbi.1000705.

[24] Bunner AE, Beck AH, and Williamson JR (2010) "Kinetic cooperativity in Escherichia coli 30S ribosomal subunit reconstitution reveals additional complexity in the assembly landscape". *Proc. Natl. Acad. Sci. USA*, 107(12): 5417–22. doi:10.1073/pnas.0912007107.

[25] Mulder A, Yoshioka C, Beck A, Bunner A, Milligan R, Potter C, Carragher B, and Williamson J (2010) "Visualizing ribosome biogenesis: Parallel assembly pathways for the 30S subunit". *Science*, 330(6004): 673–677. doi:10.1126/science.1193220.

[26] Bonomi M, Branduardi D, Bussi G, Camilloni C, Provasi D, Raiteri P, Donadio D, Marinelli F, Pietrucci F, Broglia RA, and Parrinello M (2009) "PLUMED: A portable plugin for free-energy calculations with molecular dynamics". *Comput. Phys. Commun.*, 180(10): 1961–1972. doi:10.1016/j.cpc.2009.05.011.

[27] Roberts E, Stone JE, Sepulveda L, Hwu WMW, and Luthey-Schulten Z (2009) "Long time-scale simulations of in vivo diffusion using GPU hardware". In "Parallel Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on", pp. 1–8. doi:10.1109/IPDPS.2009.5160930.

[28] Sykes M and Williamson J (2009) "A complex assembly landscape for the 30S ribosomal subunit". *Ann. Rev. Biochem.*, 38: 197–215. doi:10.1146/annurev.biophys.050708.133615.

[29] Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, and Onuchic JN (2009) "An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields". *Proteins: Struct., Func., Gen.*, 75(2): 430–441. doi:10.1002/prot.22253.

[30] Acar M, Mettetal JT, and van Oudenaarden A (2008) "Stochastic switching as a survival strategy in fluctuating environments". *Nat. Genet.*, 40(4): 471–475. doi:10.1038/ng.110.

[31] Adilakshmi T, Bellur DL, and Woodson Sa (2008) "Concurrent nucleation of 16S folding and induced fit in 30S ribosome assembly." *Nature*, 455(7217): 1268–72. doi:10.1038/nature07298.

[32] Choi PJ, Cai L, Frieda K, and Xie XS (2008) "A stochastic single-molecule event triggers phenotype switching of a bacterial cell". *Science*, 322(5900): 442–446. doi:10.1126/science.1161427.

[33] Kerr RA, Bartol TM, Kaminsky B, Dittrich M, Chang JCJ, Baden SB, Sejnowski TJ, and Stiles JR (2008) "Fast monte carlo simulation methods for biological reaction-diffusion systems in solution and on surfaces". *SIAM J. Sci. Comput.*, 30(6): 3126–3149. doi:10.1137/070692017.

[34] Roberts E, Sethi A, Montoya J, Woese C, and Luthey-Schulten Z (2008) "Molecular signatures of ribosomal evolution". *Proc. Natl. Acad. Sci. USA*, 105(37): 13953–13958. doi:10.1073/pnas.0804861105.

[35] Branduardi D, Gervasio FL, and Parrinello M (2007) "From A to B in free energy space". *J. Chem. Phys.*, 126(5). doi:10.1063/1.2432340.

[36] Kaczanowska M and Rydén-Aulin M (2007) "Ribosome biogenesis and the translation process in Escherichia coli". *Microbiol. Mol. Biol. Rev.*, 71(3): 477–494. doi:10.1128/MMBR.00013-07.

[37] Adilakshmi T, Ramaswamy P, and Woodson S (2005) "Protein-independent folding pathway of the 16S rRNA 5′". *J. Mol. Biol.*, 351(3): 508–519. doi:10.1016/j.jmb.2005.06.020.

[38] Hattne J, Fange D, and Elf J (2005) "Stochastic reaction-diffusion simulation with MesoRD". *Bioinformatics*, 21(12): 2923–2924. doi:10.1093/bioinformatics/bti431.

[39] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, and Schulten K (2005) "Scalable molecular dynamics with NAMD". *J. Comp. Chem.*, 26(16): 1781–1802. doi:10.1002/jcc.20289.

[40] Talkington M, Siuzdak G, and Williamson J (2005) "An assembly landscape for the 30S ribosomal subunit". *Nature*, 438(7068): 628–632. doi:10.1038/nature04261.

[41] Paul BJ, Ross W, Gaal T, and Gourse RL (2004) "rRNA transcription in Escherichia coli". *Ann. Rev. Genet.*, 38(1): 749–770. doi:10.1146/annurev.genet.38.072902.091347.

[42] Gutell R (2002). "Comparative RNA web site and project". URL http://www.rna.ccbb.utexas.edu/.

[43] Krishnakumar K (1990) "Micro-genetic algorithms for stationary and non-stationary function optimization". In "1989 Advances in Intelligent Robotics Systems Conference", pp. 289–296. International Society for Optics and Photonics. doi:10.1117/12.969927.

[44] Held W, Ballou B, Mizushima S, and Nomura M (1974) "Assembly mapping of 30S ribosomal proteins from Escherichia coli: Further studies". *J. Biol. Chem.*, 249(10): 3103–3111.