Integrative Modeling of Complex Biological Assemblies

Proposal for Computing Resources

Submitted to the

XSEDE Resource Allocations Committee (XRAC)

April. 1, 2019 – March. 31, 2020

Ivaylo Ivanov, Georgia State University

SUMMARY OF PROPOSED RESEARCH

This proposal details our request for supercomputing resources to support ongoing efforts in the area of molecular simulation of protein and nucleic acid assemblies engaged in DNA replication. Over the next year we have planned a research project to investigate the functional dynamics of a critical intermediate in the initiation of DNA replication - the Orc-Cdc6-Cdt1-MCM (OCCM) complex. The project involves large-scale molecular dynamics simulations with the state-of-the-art parallel code NAMD, ideally suited for the TACC Stampede 2 platform. Due to the simulation systems' sizes and the need for extensive phase space sampling we request an allocation of 233,200 node-hours on TACC Stampede 2.

INTRODUCTION

Our research is funded by a National Institutes of Health grant R01GM110387 "Integrative Modeling of Assemblies Engaged in Genome Duplication and Repair".

Our previous allocation resulted in significant research accomplishments (additional details are provided in the enclosed three-page Progress Report). To summarize, 6 publications have been completed using XSEDE resources. One additional manuscript is currently in review. Importantly, the XSEDE allocation has not only allowed us to advance the state-of-the-art in the field of molecular simulation of protein/nucleic acids complexes but also allowed rapid progress on collaborative experimental projects. Here we provide just examples of high-impact papers in the areas of DNA replication, DNA damage response pathways and gene regulation:

1. Dodd, T.; Yan, C.; Kossmann, B.R.; Martin, K.; & **Ivanov I.*** Uncovering universal rules governing the selectivity of the archetypal DNA glycosylase TDG *Proceedings of the National Academy of Sciences USA* (2018) **115**, 5974-5979 doi:10.1073/pnas.1803323115
2. Han, Y.; Yan, C.; Fishbain, S.; **Ivanov, I.** & He, Y. Structural visualization of RNA polymerase III transcription machineries. *Cell Discovery* (2018) **4**, 40 doi:10.1038/s41421-018-0044-z

The previous XSEDE allocation has also helped our continuing efforts to involve students in high performance computing (HPC) research. All group members (2 postdoctoral associates, four graduate students and several undergraduates) have been given access to and have actively used XSEDE resources. Group members were trained in the concepts, algorithms and tools of HPC and data-driven computational science, and shown how to apply these tools to biologically significant problems. Additionally, group members were given ample opportunities to highlight their work at national venues and to acquire valuable presentation skills. NSF supported cyberinfrastructure through XSEDE has been acknowledged in all our publications. Continued support from XSEDE is essential to extend and expand these training and educational activities.

BACKGROUND, SIGNIFICANCE AND SPECIFIC AIMS

Our research is broadly focused on the mechanisms and molecular machines responsible for genome duplication, maintenance and gene regulation. Understanding how DNA is copied and repaired could open new doors for disease treatment (e.g. cancer and congenital genetic diseases). Many individual components of the cell replication apparatus have been solved by protein crystallography. How these components come together to form functioning molecular machines has, however, remained elusive. Our projects are designed to integrate structural knowledge from a variety of experimental techniques (e.g. electron microscopy, SAXS, FRET, cross-linking) into efficient computational modeling protocols. This integrative approach allows us to uncover the conformational transitions in large macromolecular complexes, describe important interactions and motions and compute the associated free energies of key biological processes.

Such analyses are often intractable by purely experimental methods and could lead to practical advances in medicine and biotechnology.

In 2019 we received an INCITE award from the DOE Office of Science. While the INCITE project has no scientific overlap with the current XSEDE proposal, we expect significant efforts associated with setting up systems, submitting and monitoring jobs for INCITE. Therefore, we have scaled down our XSEDE request to a single HPC project to be carried out at TACC. At the same time, we request that our storage allocation and access to archival storage space remain comparable to what we had in previous years. The reason is that we still have data from last year's allocation that have not been fully analyzed. We do not have enough local storage capacity at GSU to archive this raw trajectory data. Additionally, uninterrupted access to our previous data would allow us to quickly ramp up our future computational efforts with XSEDE once the INCITE allocation is over.

SPECIFIC SIMULATION SYSTEMS AND EXPECTED OUTCOMES

AIM1 To deliniate the functional dynamics of a critical intermediate in DNA replication initiation - the Orc-Cdc6-Cdt1-MCM complex – we plan to carry out a series of large-scale MD simulations. We will model the OCCM in four distinct conformational states. We will also employ the string method with swarms of trajectories to model the process of loading of the Mcm2-7 hexameric helicase onto origin DNA.

Rationale

DNA replication is a tightly regulated process that involves large and dynamic macromolecular assembles. Replication is initiated at specific locations in the genome called replication origins. Eukaryotic organisms have hundreds to thousands of replication origins in each chromosome. The proteins principally involved in DNA replication initiation are the origin recognition complex (ORC), cell division control protein 6 (Cdc6), mini-chromosome maintenance proteins 2-7 (Mcm2-7), Cdc10 dependent transcript 1 (Cdt1), cell division control protein 45 (Cdc45), and the GINS complex.

The first step in the complex assembly process that initiates DNA replication is the ATP-dependent recruitment of ORC to origin DNA. Next, Cdc6 associates to ORC/DNA helping to recognize the correct DNA template. Subsequently, the ORC/Cdc6, Cdt1[1-3], and the hexameric helicase Mcm2-7 come together to form the Orc-Cdc6-Cdt1-MCM (OCCM) complex (Figure 1) [4-6]. OCCM constitutes a critical intermediate on the path toward opening duplex DNA and triggering the formation of a functional DNA replication fork. Another pre-recognition complex intermediate is the OCM complex - an assembly analogous to the OCCM wherein Cdt1 has been released. From this intermediate, a second copy of Mcm2-7 is recruited to form a Mcm2-7 double hexamer structure loaded onto DNA. The process of loading the Mcm2-7 double hexamer onto DNA occurs in the G1 phase of the cell cycle. Activation of this double hexamer occurs in S phase, resulting in the formation of the replicative CMG helicase containing Cdc45, Mcm2-7, and the GINS complex.

While the individual protein recruitment and assembly steps in replication initiation have been firmly established by decades-long biochemistry studies, the dynamic nature of the pertinent protein complexes had precluded detailed structural knowledge. Thus, the molecular architectures of the pre-recognition complex intermediates and their associated functional dynamics remain incompletely understood. With the "resolution revolution" in cryo-electron microscopy, structures of these macromolecular assemblies have recently come into view (Figure 1). Specifically, cryo-EM studies achieved near atomic visualization of the following complexes: 1) ORC/Cdc6/DNA; 2) OCCM early intermediate with an open Mcm2-7 hexamer and bent DNA; 3) OCCM with open Mcm2-7 helicase and duplex DNA threaded through the hexamer; 4) OCCM with planar closed Mcm2-7 ring [1,7,8] and DNA threaded through the Mcm2-7 central cavity.

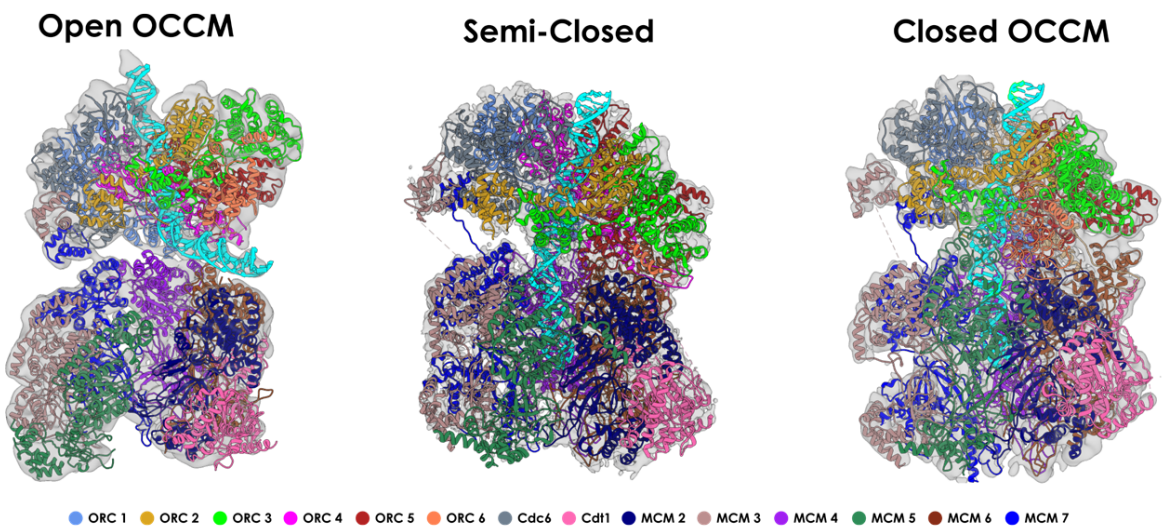


Figure 1. Cryo-EM structures and atomistic models of the OCCM in states 2-4 (A); Open OCCM conformation with spiral arrangement of the Mcm2-7 subunits and bent DNA duplex; (B) Semi closed OCCM conformation with DNA threaded through the Mcm gate; (C) Closed OCCM complex with planar Mcm2-7 hexamer.

Surprisingly, the structures imply a mechanism for OCCM loading involving threading of origin DNA through a gate formed by the Mcm2 and Mcm5 subunits[4,5]. Gate opening and closing involves a large rearrangement of the Mcm2-7 hexamer from an open helical into a closed planar conformation. In order to delineate the detailed mechanism of this conformational transition and characterize the gating dynamics we propose to carry out extensive molecular dynamics simulation of OCCM intermediates 1-4 and connect all the conformational states by computing a minimal free energy path from the open OCCM conformation to the semi-closed OCCM and finally to closed OCCM using the string method with swarms of trajectories.

Approach

To date, we have performed extensive structural comparisons of the available EM densities to create atomistic models of states 1-4. We have already performed initial equilibration of the models using supercomputing resources from our previous allocation. To summarize the simulation protocol, models for the OCCM were constructed and the system topologies were generated using the AMBER ff14SB parameters. The models were solvated using the TIP3P water model with a 10-Å margin from the protein surface to the edge of the simulation boxes. This resulted in systems comprised of **~800,000 atoms.** Counterions Na^+ and Cl^- were added to neutralize the overall protein-nucleic acid charge and achieve 150 mM salt concentration. System minimization was carried out over 5000 steps using the conjugate gradient algorithm. The systems were then slowly heated to 300 K over the course of 1 ns in the NVT ensemble with a 1-fs timestep. Positional restraints were imposed on all heavy atoms using a force constant of 5 kcal mol^{-1} $Å^{-2}$. The systems were then equilibrated over 5 ns in the NPT ensemble while gradually reducing the imposed restraints. During heating and equilibration we employed molecular dynamics flexible fitting (MDFF) with a scaling factor of 0.1. This was done to ensure that the starting configuration for the subsequent production runs remained similar to the observed EM density. Production runs were then carried out for 100 ns in the NPT ensemble. All simulations employed the SHAKE algorithm permitting the use of a **2-fs timestep**. Nonbonded interaction cutoff distance was set to **10 Å** and a switching distance at 8.5 Å. Simulations were performed using the NAMD-2.12 code. With our new XSEDE allocation we plan to extend these unbiased MD simulations to ~1 microsecond and the new trajectories will form a basis for the structural comparison of the four conformers of the OCCM.

We will also employ the string method with swarms of trajectories[9,10] to compute minimal energy paths (MEPs) connecting the experimentally observed functional states and determine the corresponding free energy profiles. This set of simulations will allow us to quantitatively examine the translocation of origin DNA through the Mcm2-7 gate and delineate the DNA loading mechanism. Since the string method is central to the proposed work, here we briefly summarize the most relevant aspects of the method. The MEP path connecting two or more functional states of the protein complex (also denoted as string) is represented by a series of replicas of the simulation system. Path optimization is typically carried out in the space of two or more collective variables (CVs) chosen to smoothly transition between the states. The replicas are initially constrained in collective variable space and a set of free unbiased MD runs (~10) are launched starting from each replica. This process allows us to compute an average drift of the replicas in CV space. The replicas are then moved in the direction of the drift. Subsequently, the simulation alternates between constrained and free MD runs, which results in evolving replica positions until convergence is reached. An interpolation step is introduced between string updates to ensure smoothness of the path and even spacing among the replicas.

We propose to carry out MEP optimization on a path comprised of 64 replicas. This choice is based on the RMSD differences among the OCCM functional states and the need to ensure appropriate spacing between adjacent replicas. MEP optimization will be followed by free energy simulations with umbrella sampling. The umbrella sampling relies on the data harvested from a collection of copies (replicas) of the simulation system. Such multiple copy algorithms (MCAs) offer a general and powerful strategy to enhance the sampling efficiency of conventional MD simulations. The computed paths will reveal the conformational plasticity of OCCM and the conformational rearrangements leading to OCCM loading onto origin DNA. Next, we will launch ensemble MD runs along the two optimal MEP paths. Such extensive unbiased sampling is required for the construction of Markov state models (MSM)[11,12] in order to reliably address the kinetics of the respective conformational transitions.

COMPUTATIONAL METHODS

Classical molecular dynamics. The majority of the computationally intensive classical molecular dynamics runs would be performed using the state-of-the-art parallel program NAMD[13,14] using the AMBER ff14SB [15] parameter set. NAMD is a well-known parallel molecular dynamics code for the simulation of large biomolecular systems designed to scale on many hundreds of cores on high-end parallel platforms. It provides efficient numerical integration of the Newtonian equations of motion, rigorous statistical mechanics methods for temperature and pressure control, algorithms for efficient evaluation of electrostatic forces, efficient parallelization through Charm++ parallel objects, dynamic load balancing and methods for computing alchemical and conformational free energies among many other features. We envision most of the production runs performed in the isothermal-isobaric ensemble (1 atm and 300 K), employing smooth particle mesh Ewald (SPME) electrostatics[16], 10 Å cut-off for short-range non-bonded interactions and multiple time-stepping with the r-RESPA method[17] (2-fs time step for bonded and short-range non-bonded interactions, 4-fs for long-range electrostatics).

Distributed memory parallelization in NAMD is achieved on most platforms through MPI or hybrid MPI/OpenMP parallelization. The limits of parallel scalability of NAMD are determined primarily by the total number of atoms in the simulated system. A conservative estimate for good efficiency on recent platforms such as the Stampede machine at TACC is 500 atoms distributed per process. Increased non-bonded interaction distance cutoff generally results in additional work to distribute and can, therefore, lead to better scaling. Benchmark data for our proposed systems shows that NAMD 2.12 is scalable for our systems up to ~10 nodes of Stampede 2. Each Skylake node features 48 hardware cores. In preparation for proposal submission we tested the optimal distribution of MPI tasks to threads and found that running 4 MPI tasks with 11 threads per task offered the best performance. It is also necessary to leave a few cores (e.g. 4 cores) available for

communication. 10 KNL nodes x 48 cores results in 480 total independent threads of execution. These thread counts are well within the scalability limits of NAMD 2.12 and are sufficient to achieve high throughput on Stampede with multiple replica runs. We plan to carry out both the string method optimization and the independent MD sampling runs in replica mode, further increasing the number nodes that could be requested in a single submission.

Storage requirements. The project is expected to generate considerable amounts of data, which falls into two general categories: (i) the raw simulation results in the form of molecular dynamics trajectories and (ii) derivative data resulting from the trajectory post-processing, analysis and visualization. Raw trajectory data (comprised of atomic coordinates and velocities) is produced directly by the simulation software and is stored in compressed format. Trajectory data varies greatly in size depending on the nature of the system, simulation length and frequency of output (normally occurring every 2 ps). We anticipate creating between 4-6 Terabytes of raw trajectory data that will need to be stored on temporary scratch space for analysis and then archived on Ranch. This estimate is based on space usage from our previous allocation. The data in the scratch space will be archived as soon as the run is complete to prevent data loss. The derivative data is much smaller in size and will include results from various analysis codes and scripts etc.

As described in the preamble, we also request additional disk and archival storage space in order to analyze and store existing data from our previous allocation. We intend to consolidate all of our existing data (currently spread between SDSC and TACC) and transfer it to Ranch via Globus. A conservative estimate of all data to be transferred in 60 TB. We do not have enough local storage capacity at GSU to archive all this raw trajectory data. Additionally, uninterrupted access to our previous data would allow us to quickly ramp up our future computational efforts with XSEDE once our INCITE allocation is over. In summary, we request **6 TB** of temporary disk storage and **65 TB** of archival storage at TACC.

LOCAL COMPUTING ENVIRONMENT

The PI's has a 20-node (240 core) parallel computing cluster with INTEL Xeon processors connected by a fast QDR Infiniband interconnect network. The cluster was purchased in 2011 and is used primarily for docking or for medium-sized classical simulations that do not scale optimally on many hundreds of cores. None of the projects listed in the proposal could be carried out or even benefit to an appreciable degree from having this local cluster. The PI's computing facilities include ten workstations (Dell Precision T7500n and T3500n models) equipped with two Intel Xeon quad-core processors and two laptop computers, used for code development, testing, and data analysis.

OTHER SUPERCOMPUTING SUPPORT

We received an INCITE Award in 2019 to carry out a project at the Oak Ridge Leadership Computing Facility. Our INCITE project aims to provide new insights into transcription initiation and the essential regulatory mechanisms controlling gene expression. The research will take advantage of new cryo-electron microscopy data and combine it with advanced computational modeling on the Summit machine at ORNL to elucidate the mechanisms of transcription initiation assemblies. There is no overlap between the projects supported by INCITE and the project described in this proposal. Our INCITE allocation runs from Jan. 1st to Dec. 31st 2019.

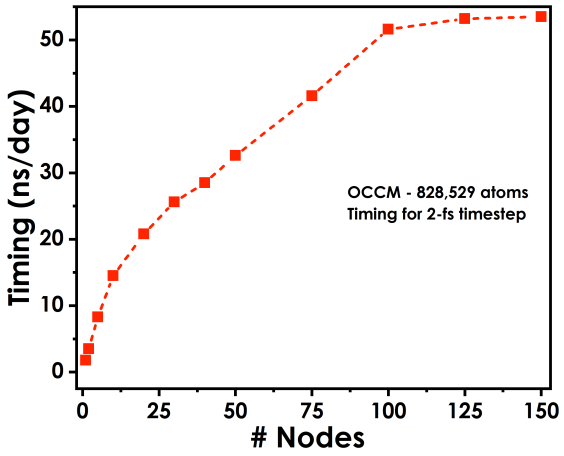
PROJECT TEAM QUALIFICATIONS

The PI has more than 16 years experience in high performance computing (HPC) and biomolecular modeling. Our XSEDE allocation TG-CHE110042 has been renewed continuously from 2010 to 2018. The group's research has resulted in over 50 publications. The majority of these studies was carried out at the national supercomputing centers and acknowledges XSEDE support. Details are provided in the attached CV.

The following individuals will be directly involved in carrying out the proposed computations: Dr. Chunli Yan (research scientist), Dr. Ashutosh Shandilya (postdoctoral scholar), Kathleen Carter (graduate student), Thomas Dodd (graduate student), Zhenyu Wang (graduate student), Jina Yu (graduate student), Kurt Martin (graduate student). The postdocs, and graduate students have extensive experience with computational science, HPC and modeling of biomolecular assemblies.

JUSTIFICATION OF REQUESTED RESOURCES

Scaling Performance



Exploratory runs (Figure 2) performed on the Skylake nodes of TACC Stampede 2 show that the OCCM complexes simulations scale up to **~10 nodes** (480 independent threads) with **14 ns/day** performance (81% parallel efficiency). More details are provided in the attached Code Performance and Scaling Report.

Resource Justification

In Aim1 we have proposed 4 OCCM systems to be simulated with free unbiased MD for ~1000 ns (~800,000 atoms system size). Required SUs to optimize and simulate the four PIC complexes:

4 systems x 1000 ns / 14 ns/day x 24 h/day x 10 nodes = **68,600 SUs**

Additionally, we proposed to simulate the conformational transitions connecting these states and the loading of OCCM onto origin DNA. Required SUs to sample the MEP (64 replicas; 50 ns of string method optimization followed by 100 ns of MD sampling per replica):

64 replica runs x 150 ns / 14 ns/day x 24 h/day x 10 nodes = **164,600 SUs**

Note: On Stampede II SUs are calculated as node-hours rather than core-hours.

In summary, the proposal requests an allocation of supercomputing resources on TACC Stampede 2 in order to carry out large-scale molecular dynamics studies relevant to the biology of DNA replication. In total the project will require **233,200 SUs.**

We thank the reviewers for their time and careful consideration of the proposal.

REFERENCES

1 Frigola, J. *et al.* Cdt1 stabilizes an open MCM ring for helicase loading. *Nat Commun* **8**, 15720, doi:10.1038/ncomms15720 (2017).

2 Ticau, S. *et al.* Mechanism and timing of Mcm2-7 ring closure during DNA replication origin licensing. *Nat Struct Mol Biol* **24**, 309-315, doi:10.1038/nsmb.3375 (2017).

3 Samel, S. A. *et al.* A unique DNA entry gate serves for regulated loading of the eukaryotic replicative helicase MCM2-7 onto DNA. *Genes Dev* **28**, 1653-1666, doi:10.1101/gad.242404.114 (2014).

4 Yuan, Z. *et al.* Structural basis of Mcm2-7 replicative helicase loading by ORC-Cdc6 and Cdt1. *Nat Struct Mol Biol* **24**, 316-324, doi:10.1038/nsmb.3372 (2017).

5 Sun, J. *et al.* Cryo-EM structure of a helicase loading intermediate containing ORC-Cdc6-Cdt1-MCM2-7 bound to DNA. *Nat Struct Mol Biol* **20**, 944-951, doi:10.1038/nsmb.2629 (2013).

6 Zhai, Y. *et al.* Open-ringed structure of the Cdt1-Mcm2-7 complex as a precursor of the MCM double hexamer. *Nat Struct Mol Biol* **24**, 300-308, doi:10.1038/nsmb.3374 (2017).

7 Lyubimov, A. Y., Costa, A., Bleichert, F., Botchan, M. R. & Berger, J. M. ATP-dependent conformational dynamics underlie the functional asymmetry of the replicative helicase from a minimalist eukaryote. *Proc Natl Acad Sci U S A* **109**, 11999-12004, doi:10.1073/pnas.1209406109 (2012).

8 Costa, A. *et al.* The structural basis for MCM2-7 helicase activation by GINS and Cdc45. *Nat Struct Mol Biol* **18**, 471-477, doi:10.1038/nsmb.2004 (2011).

9 Pan, A. C., Sezer, D. & Roux, B. Finding transition pathways using the string method with swarms of trajectories. *Journal of Physical Chemistry B* **112**, 3432-3440, doi:10.1021/jp0777059 (2008).

10 Maragliano, L., Fischer, A., Vanden-Eijnden, E. & Ciccotti, G. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J Chem Phys* **125**, 24106, doi:10.1063/1.2212942 (2006).

11 Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of chemical theory and computation* **11**, 5525-5542, doi:10.1021/acs.jctc.5b00743 (2015).

12 Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences* **106**, 19011-19016 (2009).

13 Kale, L. *et al.* NAMD2: Greater scalability for parallel molecular dynamics. *J Comput Phys* **151**, 283-312 (1999).

14 Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J Comput Chem* **26**, 1781-1802, doi:Doi 10.1002/Jcc.20289 (2005).

15 Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696-3713, doi:10.1021/acs.jctc.5b00255 (2015).

16 Essmann, U. *et al.* A Smooth Particle Mesh Ewald Method. *J Chem Phys* **103**, 8577-8593, doi:Doi 10.1063/1.470117 (1995).

17 Tuckerman, M., Berne, B. J. & Martyna, G. J. Reversible Multiple Time Scale Molecular-Dynamics. *J Chem Phys* **97**, 1990-2001 (1992).

CODE PERFORMANCE AND SCALING REPORT

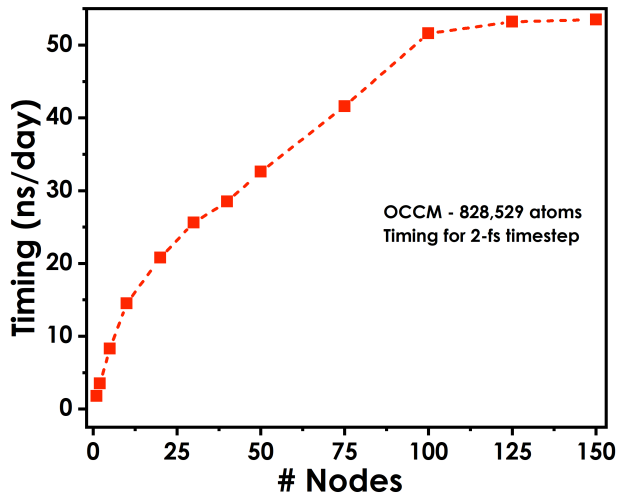
Scaling Performance

The benchmark data presented below is from exploratory runs performed on the Skylake nodes of Stampede 2. We ran the benchmarks with 48 independent processes per node and no hyperthreading. Simulation parameters that could substantively affect the timings include the cutoff for non-bonded interactions (set to 10 Angstroms), integration timestep (2-fs), use of RESPA multiple time-stepping algorithm, the size and dimensions of the grid for SPME (grid-spacing of ~1 Angstrom) and, most importantly, the number of atoms comprising the simulation system (828,529).

The scaling data demonstrate the scaling behavior on Stampede 2 showing that the OCCM systems scale up to **10 Skylake nodes** (480 independent threads).

unbiased MD (NAMD2.12)

Stampede2				
MPI tasks	Skylake nodes	NP/MPI task	*t*(ns/day)	P
48	1	17,261	1.8	1.00
96	2	8630	3.5	0.97
240	5	3452	8.3	0.93
480	10	1726	14.5	0.81
960	20	863	20.8	0.57
1440	30	576	25.6	0.47



RESEARCH ACCOMPLISHMENTS

Results from the previous XSEDE allocation

The previously awarded XSEDE allocation resulted in significant research accomplishments over the last reporting period. To summarize, 6 papers have been published and one submitted for review. These publications have appeared in high-profile journals (*PNAS, eLife, ACS J. Med. Chem.*) and exemplify the advantages of integrative modeling methods to delineate the structures and mechanisms of protein/DNA assemblies of great importance in biology. All of these projects would have been impossible without XSEDE supercomputing resources. Below are highlights from the two most significant completed projects.

Project 1 Modeling RNA Polymerase II transcription initiation assemblies

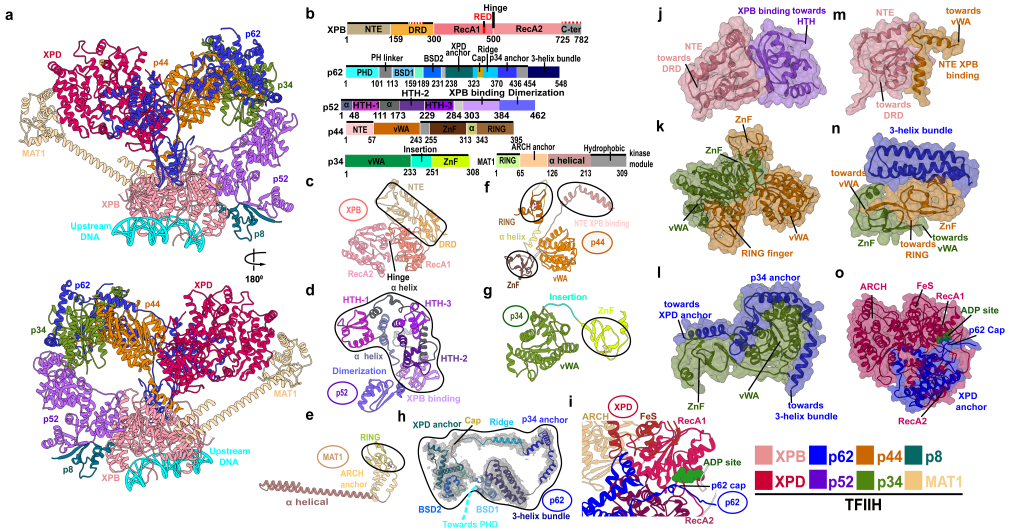


Figure 1. TFIIH integrative structure model based on comparative analysis of cryo-EM densities. Anterior and posterior cartoon views of the TFIIH GTF where missing regions or entire domains and proteins were built for XPB, p62, p52, p44, p34 and MAT1 subunits.

Complexes of RNA Polymerase II (Pol II) are foundational for transcription since all mRNA in eukaryotic cells originates from Pol II synthesis. Additionally, Pol II transcribes most small regulatory non-coding RNAs controlling gene expression levels and acting in gene silencing. As transcription regulation governs all fundamental aspects of cell biology loss of transcriptional control is a hallmark of many autoimmune disorders, cancers, neurological, metabolic and cardiovascular diseases. To begin transcription, Pol II depends on key general transcription factors (GTFs: TFIIA, TFIIB, TFIID, TFIIF, TFIIS, TFIIE and TFIIH) that recognize promoter DNA and assemble with the polymerase into a pre-initiation complex (PIC). After PIC assembly, the initial closed promoter complex (CC) transitions into an open complex (OC), in which the melted single-stranded DNA template is inserted into the Pol II active site. This transient OC is then converted into an initial transcribing complex (ITC), competent to synthesize mRNA. When the nascent RNA chain grows to a critical length, Pol II clears the promoter and a stable elongation complex (EC) ensues. Formation of PIC and its conversion into a productive elongation complex are key for transcription regulation. Yet, the molecular architecture of the PIC and its associated functional dynamics remain incompletely understood.

Importantly, with the “resolution revolution” in cryo-electron microscopy, structures of these molecular machines have recently come into view. Two recent cryo-EM studies achieved near atomic visualization of core Pol II PICs (excluding mobile TFIIH GTF) in multiple states (CC, OC and ITC) and enabled side-by-side comparison of the conformational states leading to a competent elongation complex. Two subsequent studies showed TFIIH structure both in the absence (apo-TFIIH) and in the presence of core PIC (holo-PIC). These breakthrough studies elucidated eukaryotic pre-initiation complex architectures; yet, the respective models were incomplete (>20% of residues unassigned in sequence or not modelled) and, therefore, unsuitable as starting points for detailed molecular dynamics simulations and analysis of the dynamic PIC molecular machine.

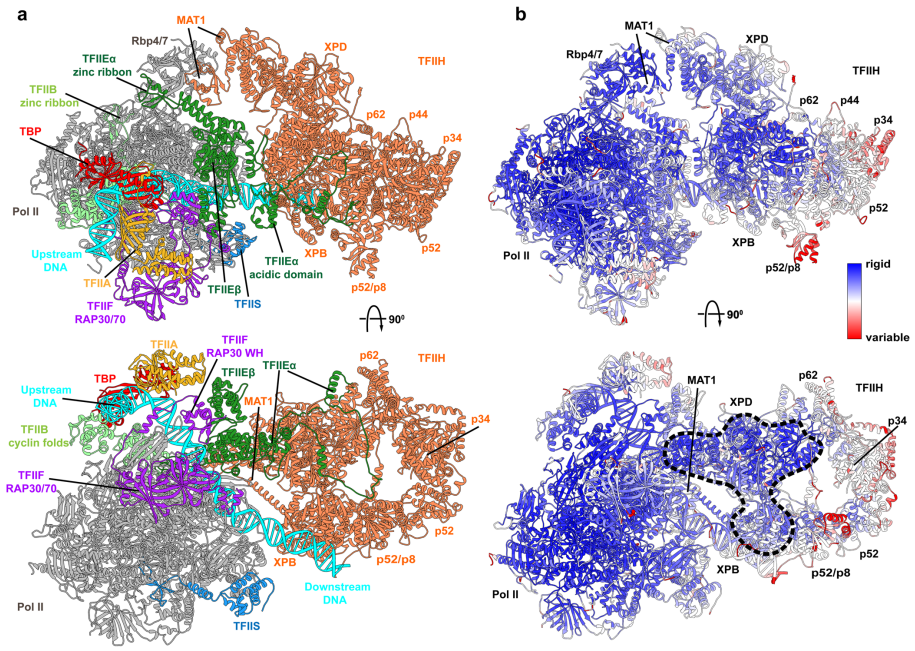


Figure 2. TFIIE links human PIC to TFIIH. a, Cartoon of human PIC structure including TFIIH. Most striking is TFIIE which crosses over a quarter of the TFIIH GTF. Model is based on integration and comparative analysis of cryo-EM densities for human apo-TFIIH, human closed-state holo-PIC density, and yeast core-PIC/TFIIH/DNA b, Computed B-factors mapped onto the PIC-TFIIH structural model with values colored from high (red) to low (blue) reveal a network of stable interactions. Black dashed outline highlights an unexpected rigid anchor region between TFIIH and PIC.

Here we synthesized all available EM data to produce the most complete atomistic model of the human PIC to date. All previously omitted/unassigned regions have now been modelled into the corresponding EM densities (Figure 1), including the ten-subunit TFIIH GTF and its flexible kinase (CAK) module. The overall assembly conformation was flexibly fitted into EM density of the CC holo-PIC (Figure 2). The quality of the new models makes them entirely suitable for molecular dynamics (MD) simulations on massively parallel computing platforms. Thus, we employed extensive MD simulations to unveil the functional dynamics of Pol II holo- and core-PICs. To our knowledge, these are the first molecular simulations of the human transcription initiation machinery. Modeling the above systems (comprised of >1,000,000 atoms) used resources of the Texas Advanced Computing Center. Importantly, these analyses revealed the hierarchical organization of the PIC machinery into dynamic communities and unveil how its interwoven structural elements function together to remodel the DNA substrate and facilitate promoter opening. Strikingly, a mapping of patient-derived TFIIH mutations onto the newly discovered dynamic communities showed that mutations were clustered at critical junctures in the TFIIH dynamic network. Thus, the results provide a new level of understanding into PIC molecular mechanism and the etiology of three devastating genetic disorders - xeroderma pigmentosum (XP, cancer), trichothiodystrophy (TTD, aging) and Cockayne syndrome (CS, development). Importantly, our methods and models provide a roadmap for future structural, biochemical and mutational experiments to understand the interplay between TFIIH structural disruption and the complex XP, CS and TTD disease phenotypes. A a manuscript describing our results has been submitted:

Chunli Yan, Thomas Dodd, Yuan He, John A. Tainer, Susan E. Tsutakawa, and Ivaylo Ivanov Transcription initiation machinery functional dynamics and genetic disease 2018 (submitted).

Project 2 Modeling the origins of selectivity of the archetypal DNA glycosylase TDG

The most prominent epigenetic modification in mammalian genomes is cytosine methylation at position 5 on the pyrimidine ring. Thymine DNA glycosylase (TDG) is a pivotal enzyme with dual roles in both genome maintenance and epigenetic regulation. Thymine DNA glycosylase plays a central role in the pathways for 5-methyl cytosine removal and, thus, influences gene silencing, stem cell differentiation and alterations in normal development. Additionally, methylation abnormalities in DNA are often observed in diseases, specifically cancer. Here we examine the mechanisms by which TDG detects, extrudes and excises modified bases in DNA. Using novel path sampling methodologies, we computed minimum free energy paths for TDG base extrusion. The computed paths reveal a novel mechanism underpinning TDG

selectivity for DNA lesions or modified bases, which involves DNA sculpting, global protein dynamics, conformational gating and specific protein-nucleic acid interactions.

Specifically, we used molecular modeling to delineate the lesion search and DNA base interrogation mechanisms of TDG. First, we examined the capacity of TDG to interrogate not only DNA substrates with 5-carboxyl cytosine modifications but also G:T mismatches and non-mismatched (A:T) base pairs using classical and accelerated molecular dynamics. To determine the kinetics, we constructed Markov State Models (MSM) based on extensive unbiased MD sampling, which required XSEDE supercomputing resources. Base interrogation was found to be highly stochastic and proceeded through insertion of an arginine-containing loop into the DNA minor groove to transiently disrupt Watson-Crick pairing. Next, we employed novel path sampling methodologies (partial nudged elastic band and string method with swarms of trajectories) to compute minimum free energy paths for TDG base extrusion. We identified the key intermediates imparting selectivity and determined effective free energy profiles for the lesion search and base extrusion into the TDG active site. Our results show that DNA sculpting, dynamic glycosylase interactions and stabilizing contacts collectively provide a powerful mechanism for the detection and discrimination of modified bases and epigenetic marks in DNA. Our results were published in the following manuscript:

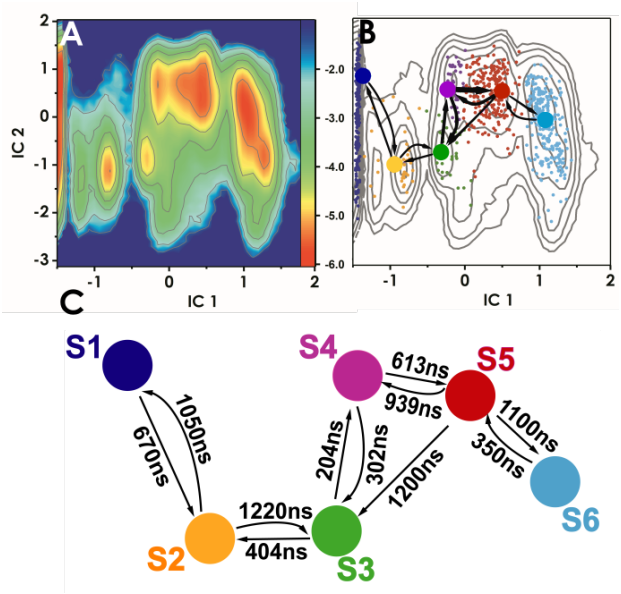


Figure 3. Conformational Dynamics of TDG/5caC-DNA complex during base eversion. (A) Free energy profile projected onto the first two ICs. Color bar inset denotes ΔG scale in kcal/mol. (B) Results from MSM analysis. Microstates (dots) are colored by macrostate they belong to. Probability fluxes between macrostates from transition path theory are shown by arrows. (C) Calculated macrostate transition timescales.

1. Dodd, T.; Yan, C.; Kossmann, B.R.; Martin, K.; & **Ivanov I.*** Uncovering universal rules governing the selectivity of the archetypal DNA glycosylase TDG *Proceedings of the National Academy of Sciences USA* (2018) doi:10.1073/pnas.1803323115