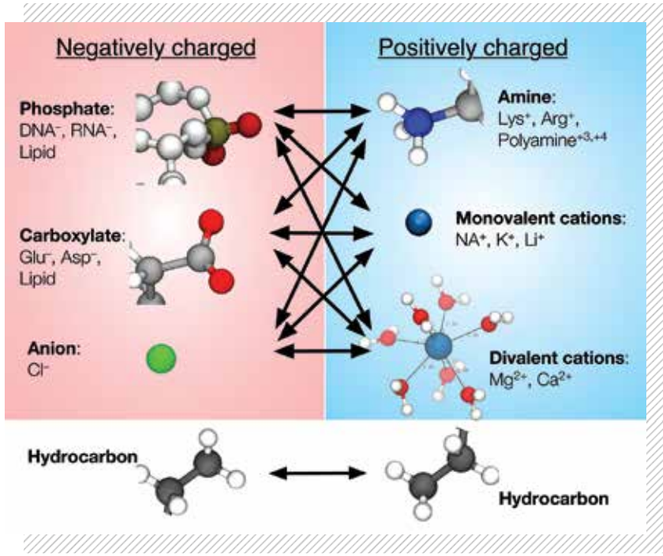
Our findings suggest a **tantalizing** possibility that polyamine-mediated inter-DNA attraction can play a major role in high-level chromatin folding.

Our inquiry into the molecular mechanism of DNA—DNA interactions found that the standard parameterization of nonbonded interactions in popular MD force fields, CHARMM and AMBER, are not accurate enough to characterize the dependence of the DNA—DNA interactions on the DNA sequence and its modifications. For example, the effective force between two DNA molecules in a 100-mM di-lysine (Lys2+) solution is attractive in both standard CHARMM and AMBER models, whereas the DNA molecules are experimentally known to repel one another at identical conditions. To improve the accuracy of the MD force fields, we reparameterized the strength of amine–phosphate and amine–carboxylate interactions against independent sets of osmotic pressure data. Our extensive validation simulations performed on Blue Waters have shown that our improved parameter set can **significantly enhance** the realism of MD simulations for a broad class of biomolecular systems, including protein folding, protein–DNA interactions, and DNA condensation.

FIGURE 3: Refinement of non-bonded interactions for accurate simulations of inter-molecular forces. The interactions indicated by arrows were reparameterized to reproduce the experimental osmotic pressure data.



WHY BLUE WATERS

Extensive sampling of biomolecular conformations was essential for characterization of DNA flexibility and calculation of inter-DNA forces. Using Blue Waters was essential to achieve the unprecedented accuracy of our simulations that matched and sometimes exceeded state-of-the-art experimental techniques.

NEXT GENERATION WORK

Using Blue Waters, we plan to create a genome-wide map of DNA flexibility that will elucidate the effect of DNA sequence on gene regulation.

PUBLICATIONS AND DATA SETS

Yoo, J., H. Kim, A. Aksimentiev, and T. Ha. Direct evidence for sequence-dependent attraction between double-stranded DNA controlled by methylation, *Nat. Commun.*, 7:11045, 2016. DOI:10.1038/ncomms11045

Ngo T., et al., Effect of cytosine modifications on DNA flexibility and nucleosome mechanical stability, *Nat. Commun.*, 7:10813, 2016. DOI:10.1038/ncomms10813

Yoo, J. and A. Aksimentiev. The structure and intermolecular forces of DNA condensate, *Nucleic Acids Res.*, 44:2036–2046, 2016. DOI:10.1093/nar/gkw081

Yoo, J. and A., Aksimentiev. Improved parameterization of amine–carboxylate and amine–phosphate interactions for molecular dynamics simulations using the CHARMM and AMBER force fields, *J. Chem. Theory Comput.*, 12:430–443, 2016. DOI:10.1021/acs.jctc.5b00967

Yoo, J., J. Wilson, and A. Aksimentiev. Improved model of hydrated calcium ion for molecular dynamics simulations using classical biomolecular force fields, *Biopolymers*, DOI:10.1002/bip.22868

Carson, S., et al., Hydroxymethyluracil Modifications Enhance the Flexibility and Hydrophilicity of Double-Stranded DNA. Nucleic Acids Res., 44: 2085–2092 (2016). DOI: 10.1093/nar/gkv1199

SEQUENCE SIMILARITY NETWORKS FOR THE PROTEIN "UNIVERSE"

Allocation: Illinois/744 Knh
PI: John A. Gerlt[1]

[1]University of Illinois at Urbana-Champaign

EXECUTIVE SUMMARY

We are devising strategies and tools to facilitate prediction of the *in vitro* activities and *in vivo* metabolic functions of uncharacterized enzymes discovered in genome projects. We used Blue Waters to establish a protocol for generating a library of sequence similarity networks (SSNs) for all Pfam protein families in the UniProt protein sequence databases for dissemination to the scientific community. We have calculated 1) all-by-all Basic Local Alignment Search Tool (BLAST) sequence relationships, 2) statistical analyses of the BLAST results; and 3) merged sets of input sequences based on sequence identity. Based on our experiences, we have defined protocols for the regular (every eight weeks) generation of the library of sequence similarity networks.

INTRODUCTION

The current UniProtKB database contains more than 60M nonredundant sequences. The functions for less than 1% of the entries have been manually curated; the functional annotations for the remaining entries are assigned by automated procedures. As a result, the conservative estimate is that the annotations for at least 50% of the entries are uncertain or incorrect. The majority of the entries are obtained from genome sequencing projects, the rationale being that knowledge of the complete complement of proteins and enzymes encoded by an organism will allow its biological and physiological capabilities to be understood. However, if at least 50% of the proteins and enzymes have uncertain or unknown functions, the considerable investments in genome projects cannot be realized. Because of the very large number of proteins and enzymes for which sequences have or will become available, strategies for predicting their functions must be high throughput and large scale, i.e., computation based.

METHODS & RESULTS

During the past year, we have continued to develop strategies to maximize the usage of RAM to enable the all-by-all sequence comparison using BLAST for the largest Pfam protein families. This has been problematic due to 1) wall time restrictions of 24 hours that recently were increased to 48 hours, and 2) the limited amount of RAM that is available for the input and output (64 GB/node) that makes node usage "inefficient."

We now are exploring the use of DIAMOND, a recently developed alternative to BLAST, for the all-by-all sequence comparisons. We have observed that DIAMOND provides a greater than or equal to 10-fold increase in the rate of all-by-all sequence comparisons relative to BLAST for almost all Pfam protein families. This significant decrease in time "solves" the wall time problem for all Pfam protein families and allows a greater number of nodes to be assigned to the largest Pfam protein families, thereby allowing the all-by-all sequence comparisons to be accomplished. We expect to be able to begin the production phase of this project in which the library of SSNs for all 16,295 Pfam protein families can be updated every eight weeks (with each update of the InterPro protein sequence analysis and classification database).

WHY BLUE WATERS

The project uses an embarrassingly parallel computing model to perform the all-by-all sequence comparison and, in principle, could be run on any cluster of sufficient size. However, because of 1) the scale of the computation (number and sizes of Pfam protein families) and 2) the time sensitivity of the production of the output relative to IntroPro database updates, only a resource at the scale of Blue Waters can perform the job in a reasonable time frame.

NEXT GENERATION WORK

We hope to more efficiently (with respect to RAM and node usage) perform the all-by-all sequence comparisons for all Pfam protein families, and eventually larger clans so that these can be disseminated to the community with each update of the InterPro database.