TN

IMPROVING THE AGREEMENT OF AMBER SIMULATION OF CRYSTALS OF NUCLEIC ACID BASES WITH EXPERIMENTAL DATA

Allocation: Innovation and Exploration/100 Knh
PI: Victor Anisimov[1]
Co-PIs: Valery Poltev[2], Thomas E. Cheatham III[3], Jerry Bernholc[4]
Collaborator: Rodrigo Galindo–Nurillo[3]

[1]University of Illinois at Urbana–Champaign
[2]Autonomous University of Puebla
[3]University of Utah
[4]North Carolina State University

EXECUTIVE SUMMARY

Classical molecular dynamics is a ubiquitous tool in the biopharmaceutical, chemical, and material sciences. Thousands of research teams nationwide and across the globe depend on the AMBER force field to conduct computer simulations of important practical applications. The present work extends the AMBER parameter optimization protocol to reproduce structure and thermodynamic properties of 12 crystals of five nucleobases. The target training data include enthalpies of sublimation at various temperatures, crystal volume, position of the atoms in the crystal, and key intermolecular distances. Further extending that data set, the project employs an unprecedented database of 161 base–base interaction energies to fine-tune the parameters. Based on experimental geometry of the clusters of bases, the database comprises interaction energies computed using the CCSD(T)/6-311++G** level of theory. Multiconformational charge fitting to the electrostatic potential of the clusters extracted from the crystal data produces atomic point charges that closely represent the average electrostatic potential in the crystals.

RESEARCH CHALLENGE

The utility of molecular dynamics simulations depends on the accuracy of their underlying parameters. Improving the predictive ability of molecular dynamics requires training the method against a large number of experimental data.

METHODS & CODES

The computational methodology involves running molecular dynamics simulations of crystals of nucleobases using the CHARMM package and performing grid searches for the optimal value of parameters to improve the agreement with experimental data.

RESULTS & IMPACT

The optimized parameters improved the agreement with the reference data set. On that set, the derived parameters scored above the existing CHARMM and AMBER force fields for nucleic acid bases.

WHY BLUE WATERS

The ability to run hundreds of small jobs that span to a thousand of nodes in the aggregate node allocation in backfill with high job turnaround is unique to Blue Waters.

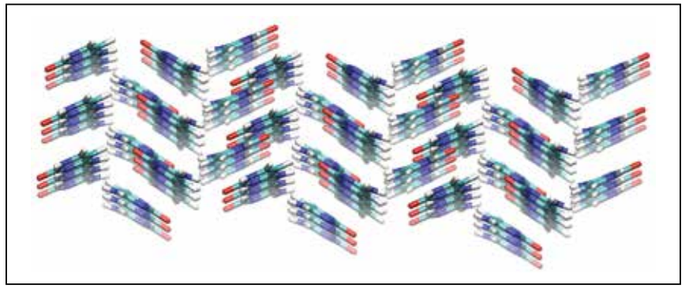

Figure 1: Packing in cytosine crystal.

ALGORITHMS FOR CANCER PHYLOGENETICS

Allocation: Director Discretionary/50 Knh
PI: Mohammed El–Kebir[1]

[1]University of Illinois at Urbana–Champaign

EXECUTIVE SUMMARY

Cancer is a genetic disease characterized by intratumor heterogeneity, or the presence of multiple cellular populations with different sets of mutations. Cellular heterogeneity gives cancer the ability to resist treatment, and quantifying the extent of heterogeneity is key to improving our understanding of tumorigenesis (the production or formation of a tumor or tumors).

In this project, the research team developed and employed novel phylogenetic techniques (the use of information on the historical relationships of lineages to test evolutionary hypotheses) to reconstruct the evolutionary histories of individual tumors from DNA and RNA sequencing data. Typically, these data are obtained from shotgun sequencing of tumor biopsies using bulk sequencing technology. Highlights of this project's research outcomes include a novel method to jointly infer a phylogeny and to estimate clone-specific expression profiles from matched RNA and DNA bulk sequencing samples.

RESEARCH CHALLENGE

A tumor is a heterogeneous population composed of clones with distinct sets of genetic mutations that result from an evolutionary process. To understand and treat cancer, we must view it through the lens of evolution. In particular, we must understand how mutations that drive cancer progression achieve their function by dysregulating (disrupting normal function of a regulatory mechanism) gene expression. Distinct clones may exhibit distinct expression profiles owing to differences in somatic mutations. To elucidate cancer evolution and the functional effect of somatic mutations, we need to infer a phylogeny (evolutionary history and relationship among organisms) that describes the clonal composition of the tumor as well as to identify an expression profile for each clone.

This is a very challenging task because, in practice, the RNA and DNA of tumor biopsies are only sequenced in bulk, resulting in a mixture of sequence reads from a large population of heterogeneous cells. The tumor phylogeny estimation problem from bulk DNA sequencing data exhibits nonuniqueness of solutions, with single-sample DNA data always supporting a linear phylogeny. These challenges have prevented the inference of clone-specific expression profiles.

METHODS & CODES

This project for the first time used matched RNA and DNA bulk sequencing samples from the same biopsy to jointly infer a phylogeny and to estimate clone-specific expression profiles. To do so, the research team leveraged a previously described pan-cancer analysis of the *cis*- and *trans*-effects of somatic mutations in The Cancer Genome Atlas (TCGA) [1].

The team formulated an optimization problem to jointly deconvolve and estimate phylogenies from matched RNA and DNA tumor samples. They solved this problem using mixed-integer linear programming. More specifically, the researchers' method used the RNA data to assign a likelihood to each phylogeny inferred from DNA data, prioritizing solutions that are most consistent with both types of data.

The method is available at https://github.com/elkebir-group/PBDRJB.

BI

RESULTS & IMPACT

Simulation experiments showed that this method is capable of prioritizing the true underlying phylogeny with high accuracy. The method can also accurately deconvolve the *cis*- and *trans*-effects of gene regulation at the clonal level. The research team analyzed matched single-sample DNA and RNA breast cancer data from TCGA and found that the method is able to differentiate between linear and branching phylogenies.

WHY BLUE WATERS

Blue Waters was essential to the research outcome of this project because it involved extensive benchmarking and validation using simulated data. The computational resources of Blue Waters allowed the research team to perform these experiments at scale, enabling the study of the performance of the algorithms and the underlying problem statements in many different experimental settings. This is something that would not have been possible on other platforms.

PUBLICATIONS & DATA SETS

Y. Luo, J. Peng, and M. El–Kebir, "Phylogenetic inference of clone-specific expression and mutation profiles from matched RNA and DNA tumor samples," submitted, 2019.