Hessisches Kompetenzzentrum
für Hochleistungsrechnen

From RNA Sequence to Coarse-Grained Molecular Dynamics

Project Manager
Maximilian Dombrowsky

Researchers
Benjamin Mayer

Principal Investigator
Prof. Dr. Heribert Warzecha

Project Term
2020 - 2021

Clusters
Lichtenberg Cluster Darmstadt

Software
Python

Additional Software
Prody, ScikitLearn, Eigen3, Biotite, MPI4Py

Institute
Computational Biology and Simulation

University
Technische Universität Darmstadt

Introduction

The dynamics of large biomolecules such as proteins or RNAs provide an important insight into their function and properties. While traditional models for dynamics prediction are based on the 3D structure of the biomolecules, there are still no models available to determine these using the 2D structure. However, since the 2D structure is much easier to predict than the 3D structure, such a model would save an enormous amount of computing time. The aim of this project was to investigate models to go from 2D to the 3D dynamics for RNA. We used gaussian network models (GNM) that approximate atom interactions as springs and are computationally light when compared to methods based on more complex interaction terms such as molecular dynamics. To improve prediction results, we optimized the GNMs spring weights using genetic algorithms and a training set of 998 RNA 3D structures. The Lichtenberg High Performance Cluster provided us with the necessary computational power to screen up to 7 different parameters in one day.

Methods

We used gaussian netwerk models to generate covariance matrices and eigen pairs for 998 RNA structures collected from the RCSB PDB. All structures were smaller than 500 bases. Only C1’ atoms were used when building the contact maps derived from 3D Structures. Deybe-Waller factors of C1’ atoms were used to compare the calculated square fluctuations. Eigenvalues were directly compared using the root mean error function. *Eigen3* was used to implement the matrix related algorithms

such as eigen solving and calculations of the pseudo inverse. We used OpenMP to parallelize the evaluation of all datasets and MPI4Py[2] with OpenMPI to parallelize the parameter screening over multiple compute nodes.

Results

We evaluated three different scoring schemes: Scoring only on eigenvalues, scoring only on Deybe-Waller factors and scoring on a combination of both. In our parameter screening we quickly realized, that only eigenvalue scoring resulted in substantial improvements and that scores based on both metrics were completely dominated by the eigenvalue metric. We were able to address this by changing the range of the calculated square fluctuations. However, the scores were now completely dominated by the Deybe-Waller factor RMSE. An extensive literatur review gave hints that the magnitude of b factors is highly depended on experimental conditions. Subsequently, we conducted an analysis of 30,000 protein 3D structures in combination with their calculated square fluctuations by *Prody*[1]. The atom wise Spearman correlation was reported as 0.35 while the molecule wise Spearman correlations had a median of about 0.32 and a mean of 0.31 in a range between -0.7 to 1. A random forest regression model trained on the resolution, water content, box volume and atom wise solvent accesible surface area using *ScikitLearn*[5] was able to achive an atom wise spearman correlation of 0.8 and molecule wise correlation between -0.5 and 1 with the median and mean at 0.55.

Discussion

The weak correlation between experimental b factors and square fluctuations predicted by GNMs hints, that experimental b factors have to be reweighted by several experimental parameters before one can use them for fitting or training. Alternatively, square fluctuations from experimental methods that result in structural ensembles like NMR, should be used to lessen the influence of crystalline properties. As most NMR studies on biomolecules are conducted in water solutions, the influence of crystallization itself can also be eliminated. Therefor, we will attempt to train our model on square fluctuations from NMR structures and try to reweight the Deybe-Waller factors to enable training on crystal structures.

Reference

Bakan, A.; Meireles, L.M.; Bahar, I.: ProDy: Protein Dynamics Inferred from Theory and Experiments. In: Bioinformatics 27.11, pp. 1575–1577, Apr. 2011 https://doi.org/10.1093/bioinformatics/btr168

Dalcın, L.; Paz, R.; Storti, M.: MPI for Python". In: Journal of Parallel and Distributed Computing 65.9, pp. 1108–1115, 2005

Guennebaud, G.; Jacob, B.; et al. Eigen v3, 2010 http://eigen.tuxfamily.org

Kunzmann, P.; Hamacher, K.: Biotite: a unifying open source computational biology framework in Python. In: BMC Bioinformatics 19.1, p. 346. ISSN: 1471-2105, Oct. 2018 https://doi.org/10.1186/s12859-018-2367-z

Pedregosa, F. et al.: Scikit-learn: Machine Learning in Python. In: Journal of Machine Learning Research 12, pp. 2825–2830, 2011

Last Update: 2021-12-21 10:22