Predicting and Representing the Structure of an Unstructured Protein through an Ensemble of Low-Energy Conformations

Md Wasi Ul Kabir (mkabir3@uno.edu), Md Tamjidul Hoque (thoque@uno.edu)
Department of Computer Science, University of New Orleans, New Orleans, LA, USA

Introduction

- Numerous proteins lack a stable 3D shape under physiological conditions, earning them the designation of **intrinsically disordered proteins** (IDPs) or **disordered regions** in proteins (IDRs).
- These proteins, alternatively known as flexible proteins (FPs), lack a discernible structure and exhibit variable conformations owing to the **flexibility of their backbone atom coordinates**.
- IDPs play a **pivotal role** in **functional** proteomics, with their Molecular Recognition Functions (MoRF) regions overseeing critical biological processes like signaling, recognition, regulation, and cell division.

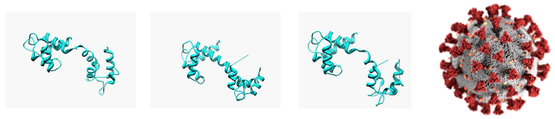


Figure 1. Three conformations of a disordered protein and SARS-CoV-2 spike protein.

https://theconversation.com/how-the-leading-coronavirus-vaccines-work-146969

Motivation

- The structural heterogeneity of IDPs is closely related to **critical human diseases** such as Parkinson's disease, Alzheimer's disease, COVID-19, type II diabetes, and cancer.
- Disordered Proteins are highly **abundant** in nature and their functional repertoire **complements** the functions of ordered proteins.
- To understand the function of flexible proteins, it is important to know the possible conformations that they can be found in. Thus, the generation of **possible conformational ensembles** of flexible proteins is important.
- Additionally, an **energy function** applicable to intrinsically disordered proteins is necessary to **rank** and **differentiate** low-energy conformations from the large ensemble of conformations generated during the search process.

Dataset

Training, Optimization and Independent Test Datasets

- Due to their structural heterogeneity, IDPs are best described by an ensemble of structures representing the conformations.
- We create a new dataset from **Protein Data Bank** (PDB).

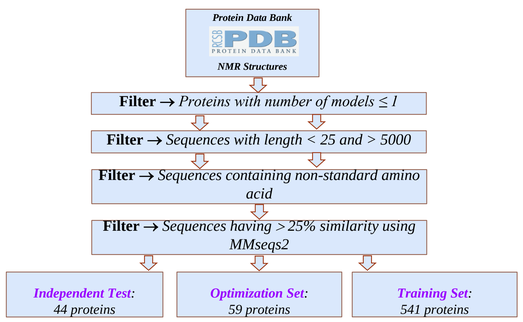
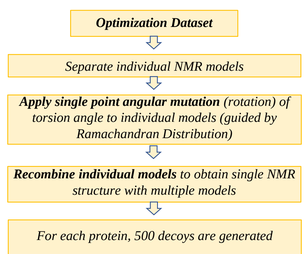


Figure 3. Illustrate the steps to create the dataset for training, optimization, and testing.

Construction of Optimization and Independent Test Decoys



Similar approach is used to obtain 500 decoys per protein *for* ***Independent test dataset***.

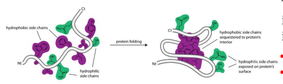
Figure 4. Generating decoy set by applying single point angular mutation.

Flexible Energy Function

Hydrophobic-Hydrophilic Properties

Hypothesis 1: ***Hydrophobic and hydrophilic interaction*** plays vital role in ***determining the conformation of the protein***

- ***Hydrophobic*** amino acid form the ***inner core*** of the protein
- ***Hydrophilic*** amino acid form the ***outer surface*** of the protein
- There exist a ***mixed hydrophobic-hydrophilic interaction layer*** in between inner core and outer surface



Therefore, to individually capture these interactions, we created three different interaction libraries:

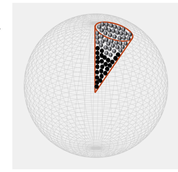
- Hydrophobic – Hydrophilic Interaction Library
- Hydrophobic – Hydrophobic Interaction Library
- Hydrophilic – Hydrophilic Interaction Library

Mishra, A. and M. T. Hoque (2017). "Three-Dimensional Ideal Gas Reference State Based Energy Function." Current Bioinformatics 12(2): 171-180. https://www.labxchange.org/library/items/lb:LabXchange:a9290de2:html:1

Ideal gas reference state

Hypothesis 2: Extends the concept of finite ideal gas reference state.

- *The well known "**finite ideal gas reference state**" **considers protein as a sphere and assumes that the expected number of atoms at distant bin "d" will increase in d^{α}*** where "α" is a constant determined through experiments.
- *Besides, we assume that the **expected number of atoms in the spherical environment increase gradually as we move from the center of the sphere to the surface.***
- Therefore, we derive ***three different reference state*** for three different interaction types (HP, HH and PP). Each reference state has its ***individual alphas (α_{hp}, α_{hh}, and α_{pp}).***
- Variable alphas is optimized using GA.



Hydrophilic Amino Acid
Hydrophobic Amino Acid

Zhou, H. and Y. Zhou (2002). "Distance-scaled, Finite Ideal-gas Reference State Improves Structure-derived Potentials of Mean Force for Structure Selection and Stability Prediction." Protein Sci. 11: 2714–2726.

Mishra, A. and M. T. Hoque (2017). "Three-Dimensional Ideal Gas Reference State Based Energy Function." Current Bioinformatics 12(2): 171-180.

Ab Initio Conformational Ensemble Generator

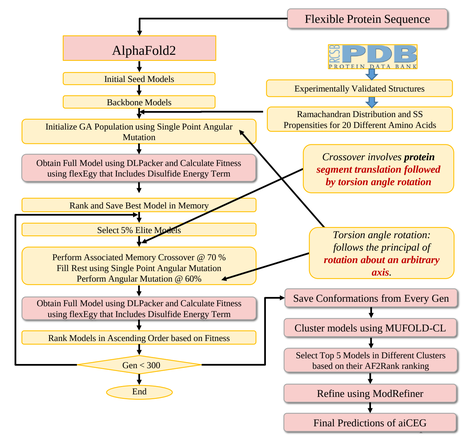


Figure 7. Conformational Ensembles Generator of Disordered Proteins using Genetic Algorithm.

Molecular Dynamics

- In the realm of Molecular Dynamics (MD) applied to comprehend the behavior of specific disordered proteins oscillating between disordered and ordered states, MD is a computer simulation methodology for scrutinizing the dynamic movements of atoms and molecules.
- The paths traced by these particles are dictated by the numerical resolution of Newton's equations of motion within a system of interacting particles.
- Our objective is to refine our predictive capabilities by strategically utilizing MD simulations to identify oscillating ones.

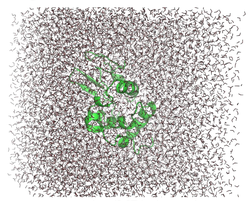


Figure 8. Molecular dynamics (MD) simulation on protein ID 1AKI.

Conformations of Disordered Proteins

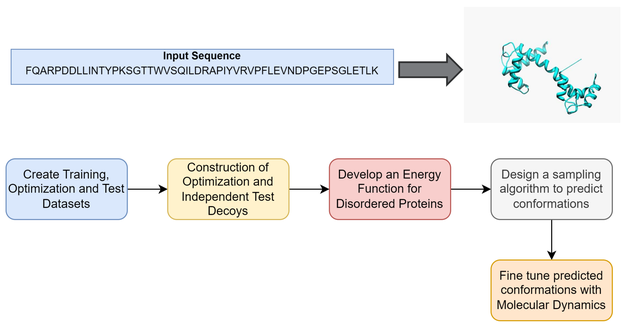


Figure 2. Pipeline to generate the conformations of Disordered Proteins.

Construction of 3D HP Libraries

Collect pair-wise distances of every two atoms from NMR models in Training Dataset

Construct three different Frequency Distribution Table considering HP, HH and PP Type Interaction (14028 rows and 30 columns)

*Convert Frequency Distribution → Energy Library (applying **three-dimensional ideal gas reference state**)*

"α" is a parameter that represents atomic distribution in HP/HH/PP interaction layer

Frequency of atom pair x, y at the last bin of HP, HH and PP libraries

$$N_{x,y}^{exp-\oplus}(d) = \left(\frac{d}{d_{cut}}\right)^{\alpha_{\oplus}} \left(N_{obs-hp}(x,y,d_{cut}) + N_{obs-hh}(x,y,d_{cut}) + N_{obs-pp}(x,y,d_{cut})\right)$$

"*d*" is cut of distance, and "d_{cut}" is the maximum cut of distance

Figure 5. Construction of energy library. $N_{x,y}^{exp-\oplus}(d)$ represents the expected number of atom pairs at distance d for $\oplus$ group, $N_{obs-\oplus}(x,y,d_{cut})$ represents the number of observation of atom pairs xth and yth observed at cutoff distance obtained from $\oplus$ library (specifically, the $\oplus$ represents HP or, HH or, PP) and $\alpha_{\oplus}$, is the parameter that belongs to $\oplus$ group, which is optimized by GA.

Scoring Technique

*The **total energy of the flexible protein** is the weighted sum of the energies obtained from the atom pairs participating in HP/HH/PP type interaction.*

Weights are optimized using GA

Energy of atom pairs that participate in HH and PP type interaction are computed through similar approach.

$$TE_{flex} = \beta_{hp-flex}E_{hp-flex} + \beta_{hh-flex}E_{hh-flex} + \beta_{pp-flex}E_{pp-flex}$$

Sum of the energies of all the atom pairs that participate in HP/HH/PP type interaction.

$$E_{\oplus-flex} = \sum_{x,y,d} ES_{x,y,d}^{\oplus}$$

Energy of an atom pair that participate in HP/HH/PP type interaction.

$$ES_{x,y,d}^{\oplus} = -ln\left(N_{obs-\oplus}(x,y,d)/N_{x,y}^{exp-\oplus}(d)\right)$$

Figure 6. Total energy score of flexible proteins.

Conclusions and Future Plans

- In this investigation, we introduce a versatile energy function and an ab initio conformational ensemble generator designed to forecast disordered proteins' conformational ensembles.
- The implementation of our proposed method is currently underway, with a primary challenge being the substantial computational resources required for the ensemble generator pipeline.
- Our strategy involves enhancing the flexible energy function by predicting disulfide bonds.
- Additionally, we aim to integrate the Dispredict3.0 disordered predictor tool into the Genetic Algorithm, utilizing crossover and mutation techniques in disordered regions.
- Furthermore, we plan to explore Molecular Dynamics (MD) simulations tailored for oscillating disordered proteins.

Acknowledgements

We gratefully acknowledge LBRN's support in providing the computational resources.