Machine Learning is Revolutionizing Structural Bioinformatics

Jian Wang
Nov 16, 2022
Biomolecules

Lipids  Nucleic acids  Proteins  Chemical compounds
Quantum Computation

Quantum computation supports calculating the dynamics of a small molecule

$$H(t)|\psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle$$

Schrödinger Equation
Molecular Dynamics Simulation

By using force field, molecular dynamics simulation supports the simulation of the dynamics of a protein.

Newton’s second law: $F = m \cdot a$
Molecular Dynamics Simulation

Analytical Engine

ENIAC

Modern Computer

HPC

Quantum Computer

Moore's law

Atoms

Small molecules

Proteins

Organelles

Cells
Graph

Network

Graph

G(N, E)

N: nodes

E: edges
Protein Structure
RNA
Contents

- RNA 3D Structure Prediction
- Protein-ligand Interaction Prediction
- Protein Allostery
- Applications of Structural Bioinformatics
Non-coding RNA 3D structure prediction

1969
The first manually predicted tertiary structure of tRNA was regarded as a milestone in the emergence of bioinformatics

1989
The core of group I intron was solved based on extensive sequence comparisons, secondary structures, and published mutagenesis data.

1995
31 RNA 3D structures in PDB

2005
462 RNA 3D structures in PDB

Now
Only 4924 RNA 3D structures in PDB (>165,266 proteins in PDB)
22,776,905 non-coding RNA sequences in RNAcentral

Gene

Protein encoding <3% Non-coding
Difficulties in RNA 3D structure prediction

Sequence

Multiple Conformations

AUGCGUCA⋯

Sampling

Scoring

Final Prediction
Base pairing and base stacking information is implied in distance distribution.
Backbone torsion angle distribution

Backbone conformation is implied in torsion angle distribution
Results in Test I, II, and III

\[ E_{\text{total}} = E_{\text{distance}} + \omega E_{\text{dihedral}} \]

\[ E_{\text{distance}} = -k_B T \sum_{ij} \ln \frac{f_{ab}^{OBS} (d_{ij})}{f_{ab}^{REF} (d_{ij})} \]

\[ E_{\text{dihedral}} = -k_B T \sum_{i} \ln \frac{f_{a}^{OBS} (\theta_{i})}{f_{a}^{REF} (\theta_{i})} \]
Difficulties in RNA 3D structure prediction

Sequence

AUGCGUCA⋯

Multiple Conformations

Sampling (3dRNA, iFoldRNA)

Final Prediction
Secondary Structure Tree

- L1
- L2
- L3
- L4
- L5
- L6
- L7
- L8
- L9

Secondary Structure Element

- Hairpin Loop
- Internal Loop
- Junction
Fragment-assembly

Fragment assembly is like **building blocks**.
Integrating Restraints

Restraints such as hydroxyl radical probing (HRP), cross-linking, and direct coupling analysis (DCA) can be used as restraints to improve RNA modeling.
RNA 3D Structure Optimization

Coarse-grained model

Sampling

Energy Function

\[ G = G_{vb-len} + G_{vb-ang} + G_{vb-tot} + G_{stacking} + G_{pairing} + G_{restr} \]

\[ G_{vb-len} = k_l(l - l_0)^2 \]
\[ G_{vb-ang} = k_\alpha(\alpha - \alpha_0)^2 \]
\[ G_{vb-tor} = k_t \sin^2\left(\frac{t - t_0}{2}\right) \]
\[ G_{restr} = k_r \sum_n (r^n - r_0^n)^2 \]
Test Results

Native, Optimize w/o restraints, Optimize w/ restraints
How many restraints should we impose?

When using more than 60% of the contacts as constraints, the RMSD of the predicted model gets higher.
Difficulties in RNA 3D structure prediction

By using more restraints, there may be more energy barriers in the free energy landscape.

NN: local minimum state near the native state
DN: local minimum state distant from the native state
Restraints Derivation

The larger the distance variation, the higher the importance of the corresponding restraint.

Distance Variation (DV): the difference between the minimum distance and the maximum distance between two residues

Triangle Inequality

\[ c_{\text{max}} = a_{\text{max}} + b_{\text{max}} \]

\[ c_{\text{min}} = \begin{cases} b_{\text{min}} - a_{\text{max}}, & b_{\text{min}} > a_{\text{max}} \\ a_{\text{min}} - b_{\text{max}}, & a_{\text{min}} > b_{\text{max}} \end{cases} \]

\[ DV_c = c_{\text{max}} - c_{\text{min}} \]
By sorting the constraints by the importance, we can use only the constraints that let us achieve the highest performance.
Contents

RNA 3D Structure Prediction

Protein-ligand Interaction Prediction

Protein Allostery

Applications of Structural Bioinformatics
Virtual Screening

Molecular Docking

QSAR (Quantitative structure-activity relationship)

Risk of highly inaccurate predictions of pharmacological or biological activity

Re-ranking Virtual Screening results
Compound selection & in vitro assays

Drug-like & Lead-like
Toxicity
Bad pose
PAINS
ADME

Empirical rules

QSAR (Quantitative structure-activity relationship)

Risk of highly inaccurate predictions of pharmacological or biological activity
MedusaDock

A

STROLL Generation

Clustering

Adding Constraints

N rounds of docking

Coarse Docking

Fine Docking

Clustering

Receptor

Ligand

Adding constraints

Energy (kJ)

Distance (Å)

RMSD (Å)

Number of Constraints
MedusaDock website
MedusaNet: Guiding Conventional Protein–Ligand Docking Software


Dr. Mahmut Kandemir

Output: Binding Probability

Fully-connected Layers

Convolution Layers

Dr. Mahmut Kandemir
MedusaNet: Guiding Conventional Protein–Ligand Docking Software

(a) MedusaDock run less attempts with CNN guiding.

(b) Number of proteins found good pose by each approach.

MedusaNet improves both the efficiency and accuracy of MedusaDock
NeuralDock: Rapid and conformation-agnostic docking of small molecules


(a)

(b)

(c)

Congzhou Sha

Dr. Nikolay Dokholyan
Comparison to traditional docking software

NeuralDock is comparable to traditional docking software

96000 protein-small molecule pairs
Tesla T4 GPU, training in a week
937 million ZINC compounds took 21 hours on 25 GPUs
Target identification

Virtual Screening

Drug-like & Lead-like
Toxicity
Bad pose
PAINS
ADME

Filtering tools

Empirical rules

Target Identification

Off-target

Drug

Off-target

Target
Current compound-protein interaction prediction models

DeepDTA

DeepConv-DTI
The protein features and the compound features are multiplied to evaluate the pairwise residue-atom interaction.
**Evaluation of the predictability of Yuel**

- **Davis/Davis**: Models are trained on Davis and tested on Davis
- **PDBbind/PDBbind**: Models are trained on PDBbind and tested on PDBbind
- **Davis/PDBbind**: Models are trained on Davis and tested on PDBbind

**Davis and PDBbind are dissimilar**
Testing CPIP in the datasets with the protein sequence shuffled

Shuffle the protein sequence in the test sets

When **shuffling** the protein sequences, DeepDTA and Deep-Conv-DTI still predict high affinity between the shuffled protein and the compounds.
Yuel can predict hotspot atoms and residues

Yuel can predict compound atoms that interact with the protein (hotspot atoms) as well as protein residues that interact with the compound (hotspot residues).
Contents

RNA 3D Structure Prediction

Protein-ligand Interaction Prediction

Protein Allostery

Applications of Structural Bioinformatics
Allostery in various heterogeneous materials

**Allostery in GPCR**


**Allostery in RNA**


**Allostery in mechanical networks**

Allosteric Pathways and Critical Nodes

Inactive

Active

Allosteric Site?

Allosteric Pathways

Critical nodes

Allosteric Site?
Residue Interaction Network

Social Network

Residue Interaction Network

Perturbation Propagation Algorithm

Contact Matrix (C)

\[ C_{ij} = \sum_{a,b} H \left( r_0 - |r_{ia} - r_{ib}| \right) \]

Matrix of number of contacts per atom (N)

\[ N_{ij} = \frac{C_{ij}}{C_i}, N_{ij} = \frac{C_{ij}}{C_i} \]

Perturbation Transfer Probability Matrix (P)

\[ P_{ij} = 1 - e^{-\alpha N_{ij}} \]
Perturbation Propagation Algorithm

Perturbation Transfer Probability Matrix (P)

\[
\begin{array}{cccccc}
0.00 & 0.50 & 0.10 & 0.62 & 1.00 \\
0.50 & 0.00 & 0.40 & 0.87 & 0.20 \\
0.10 & 0.40 & 0.00 & 0.98 & 0.22 \\
0.62 & 0.87 & 0.98 & 0.00 & 0.75 \\
1.00 & 0.20 & 0.22 & 0.75 & 0.00 \\
\end{array}
\]
Perturbation Propagation Algorithm

Round 1
Perturbation Propagation Algorithm

Round 2
Perturbation Propagation Algorithm

Round 3

Repeat 10000 times

Allosteric Correlation Intensity
Chemotaxis protein Y (CheY)

The phosphorylation of D57 residue of CheY can activate the binding of FliM and other flagellar motors at the distal binding surface.

“Y–T” coupling scheme

Identification of the allosteric site in CheY

Critical residues in allosteric pathways of CheY

D57

Y87

T106

Allosteric correlation intensities in mutations of CheY

Prediction of All Residue-wise Allosteric Correlations

NMR Chemical Shift Covariance Analysis (CHESCA)

OHM
Explore Allostery

Identify allostERIC sites and pathways

Predict AlloSteric Sites  Identify AlloSteric Pathways  Perturb AlloSteric Pathways  Predict AlloSteric Correlations

What is Ohm?

Allostery is a natural phenomenon in proteins whereby distal structural elements are dynamically coupled. The origins of the allosteric phenomenon are rooted in physical properties of inter-atomic interactions in heterogenous media. Protein sequences are heterogeneous and their corresponding structures represent a diverse range of forces between amino acids that shape their structures. In proteins perturbations of amino acids will propagate non-uniformly throughout the structure: a residue that is stronger coupled to a neighbor, would be more affected by the perturbation of the neighbor.

http://Ohm.Dokhlab.org
Applications of Structural Bioinformatics

- Protein Allostery
- RNA Modeling
- Compound-protein interaction

Identifying new druggable site
- Drug delivery
- Antisense drugs

Drug screening, Target identification
Contents

RNA 3D Structure Prediction

Protein-ligand Interaction Prediction

Protein Allostery

Applications of Structural Bioinformatics
RNA Modeling in Nucleic Acid Nanoparticle Design

Dr. Kirill Afonin
RNA Modeling in Nucleic Acid Nanoparticle Design
Virtual Screening

Re-ranking Virtual Screening results
Compound selection & in vitro assays

Drug-like & Lead-like
Toxicity
Bad pose
PAINS
ADME
Targeting 6- and 7-transmembrane μ-opioid receptor isoforms

Dr. Luda Diatchenko

ZINC Drug Now
~11M compounds

10 docking runs
(~350K compounds)

100 docking runs
(~5K compounds)

500 docking runs
(~1K compounds)

6TM-MOR selective
(10 compounds)

non-selective
(10 compounds)

7TM-MOR selective
(10 compounds)

Compound 5

Compound 10

Compound 11

Compound 25
Conclusion

RNA Scoring

Protein Allostery Mapping

CPI Evaluation

RNA Modeling

Understanding  Prediction  Design
Designing new biomolecules

Protein Design

Graph Design

RNA Design

Drug Design
Policy network *makes decisions* and Value network *evaluates the situation*
Questions?