Machine Learning in Structural Biology: Interpreting 3D Protein Images

By

ALOYSIOUS KIKOMEKO – 2019/HD07/24856U

CHISOM SOREMEKUN - 2020/HD07/24246U

Outline

- Introduction
- Background
- Literature review
- Methodology
- Results
- Conclusion
- references.

Introduction

- This chapter discusses an important problem that arises in structural biology: given an electron density map, a three-dimensional "image" of a protein produced from crystallography.
- Traditionally, a human performs this interpretation, perhaps aided by a graphics terminal.
- However, over the past 15 years, a number of research groups have used machine learning to automate density map interpretation.

Introduction ...

- Early methods had much success, saving thousands of crystallographer-hours, but required extremely high-quality density maps to work.

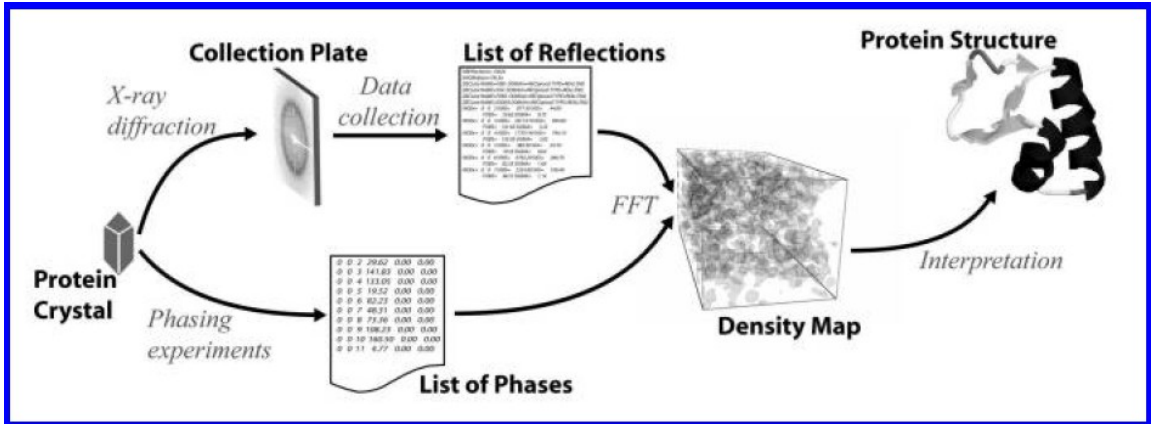
- Newer methods aim to automatically interpret poorer and poorer quality maps, using state-of-the-art machine learning and computer vision algorithms.

Background

- A protein's structure provides great insight into the mechanisms by which a protein acts.
- knowing these mechanisms helps increase our basic understanding of the underlying biology.
- No existing computer algorithm can accurately map sequence to 3D structure.
- However, several experimental "wet lab" techniques exist for determining macro molecular structure.

Background

- The most commonly used method, employed for about 80% of structures currently known, is x-ray crystallography .



An overview of the crystallographic process.

Literature review

Algorithms

- Algorithms for automatically interpreting electron density maps draw heavily from the machine learning and statistics communities.
- These communities have developed powerful frameworks for modeling uncertainty.
- Reasoning from prior examples and statistically modeling data, all of which have been used by researchers in crystallography. (S. Russell and P. Norvig; 1995).

Probabilistic models

- A model here refers to a system that simulates a real-world event or process.
- Probabilistic models simulate uncertainty by producing different outcomes with different probabilities. These models include;
- The unconditional or prior probability ie. P (A).
- The conditional or posterior probability is used when other, previously unknown, information becomes available. I.e. P(A|B).
- The joint probability of two or more events is the probability of both events occurring. P(A,B) and is read "the probability of A and B."
- Finally, Bayes' rule allows one to reverse the direction of a conditional:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Case-Based Reasoning

- Broadly defined, case-based reasoning (CBR) attempts to solve a new problem by using solutions to similar past problems.

- Algorithms for case-based reasoning require a database of previously solved problem instances, and some distance function to calculate how "different" two problem instances are.

- There are two key aspects of CBR systems.

1; Learning in such systems is lazy: the models only generalize to unseen instances when presented with such anew instance.

2; They only use instances "close" to the unseen instance when categorizing it.

Case-Based Reasoning algorithms

1. **k-nearest neighbor (kNN).**

In kNN, problem instances are feature vectors, that is, points in some n -dimensional space.

2. **Neural networks**

An artificial neural network (ANN) is a nonlinear function estimator used to approximate real or discrete target functions. Inspired by neurons in the brain.

Approaches to automatic density map interpretation

- A number of research groups have investigated automating the interpretation of electron-density maps.

There are four algorithmic for used in 3D protein interaction.

1. **The ARP/WARP (Automated Refinement Procedure)**
2. **Resolve's method**
3. **Textal method**
4. **ACMI (Automatic Crystallographic Map Interpreter)**

Density Map Interpretation

- By far the most commonly used method for Automatic Density Map Interpretation is ARP/WARP. R. Morris, A. Perrakis, and V. Lamzin (2002)

- This atombased method heuristically places atoms in the map, connects them and refines their positions.

- To handle poor phasing, ARP/WARP uses an iterative algorithm, consisting of alternating phases in which;

    (a) a model is built from a density map

Density Map Interpretation ...

(b) The density map's phases are improved using the constructed model.

•This algorithm is widely used, but has one drawback: fairly high resolution data, about 2.3A or better, is needed.

•Given this high resolution data, the method is extremely accurate; however, many protein crystals fail to diffract to this extent.

•Several approaches represent the density map in an alternate form, in the process making the map more easily interpretable for both manual and automated approaches.

Density Map Interpretation ...

- One of the earliest such methods, skeletonization , was proposed for use in protein crystallography by Greer's bones algorithm. J. Greer (1974).

- Skeletonization, similar to the medial-axis transformation in computer vision, gradually thins the density map until it is a narrow ribbon approximately tracing the protein's backbone and sidechains.

Density Map Interpretation ...

- Cowtan's FFFEAR efficiently locates rigid templates in the density map K. Cowtan (2001).
- It uses Fast Fourier Transforms (FFTs) to quickly match a learned template over all locations in a density map.
- Unfortunately, the technique can only locate large rigid templates.
- However, a number of methods use FFFEAR as a template-matching subroutine in a larger interpretation algorithm.

Density Map Interpretation ….

- T. Oldfield (2001) noted that, X-autofit, part of the quanta package, uses a gradient refinement algorithm to place and refine the protein's backbone.
- Their refinement takes into account the density map as well as bond geometry constraints.
- They report successful application of the method at resolutions ranging from 0.8 to 3.0A.

Density Map Interpretation ...

- Terwilliger's resolve contains an automated model-building routine that uses a hierarchical procedure in which helices and strands are located and fitted, then are extended in an iterative fashion, using a library of tripeptides.

- Levitt's maid approaches map interpretation "as a human would," by first finding the major secondary structures, alpha helices and beta sheets, connecting the secondary-structure elements, and mapping this fit to the provided sequence.

Density Map Interpretation ...

- Ioerger's TEXTAL [T. Ioerger (2005)] attempts to interpret poor-resolution (2.2 to 3.0A) density maps using ideas from pattern recognition. Ioerger constructs a set of 15 rotation-invariant density features.
- Finally, ACMI takes a probabilistic approach to density map interpretation [F. DiMaio, J. Shavlik, and G. Phillips (2006)].
- Residues of the protein are modelled as nodes in a graph, while edges model pairwise structural interactions arising from chemistry.

Methodology

Practical application: Machine Learning in Structural Biology: Interpreting 3D Protein Images

Data;

Standardized datasets were obtained from the CASP competition storage to test machine learning algorithms. The CASP competition datasets were well organized and published in deep learning / machine learning friendly manner through a project call ProteinNet.

The dataset selected form CASP11 dataset of [ProteinNet](https://github.com/aqlaboratory/proteinnet) project. Random set of proteins which have length between 64 and 128 amino acids for this project. Original dataset was in text and TF record format, was converted to __*npz*__ before to make the code.

Libraries used

- numpy
- matplotlib
- Tensorflow v1.15 (keras)

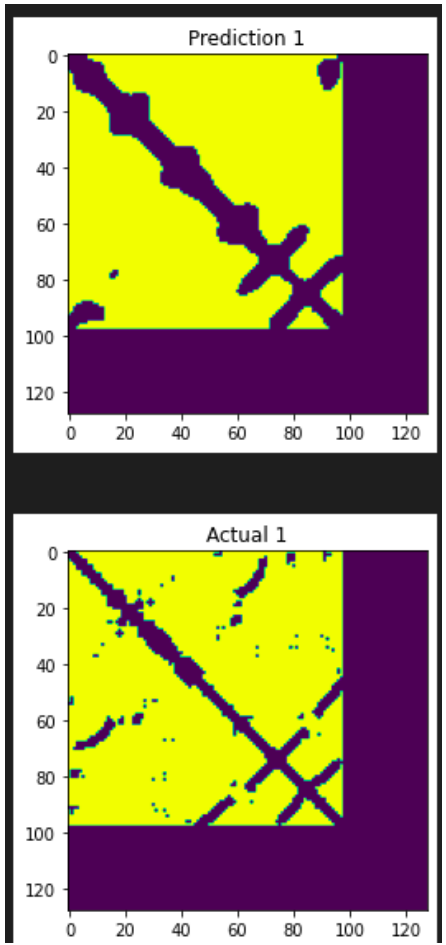
Methods:

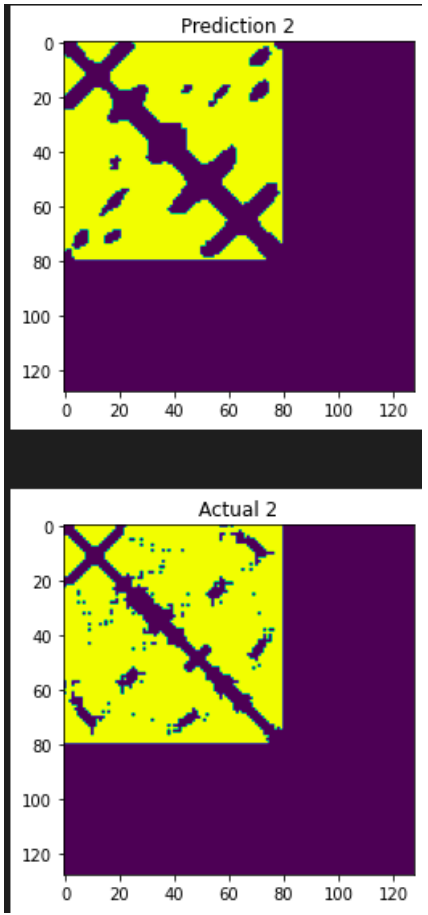
Contact maps were defined by relative distance between pair of each _*C_Alpha*_ atoms of each amino acids using the 9 angstrom distance threshold, to create binary contact maps. Protein contact maps were calculated for training and test datasets. Neural network design were then used to capture transformation within the network. The Embedding technique was employed to convert encoded integers to continuous vectors. Then we used Mean Square Error (MSE) as the loss function, Adam-optimizer as the optimizing algorithm.

- Training and prediction was done by the model-fit and validation techiniques.

- Classification of data is also done using supervised machine learning algorithms and these are Support vector Machines(SVM), K-Nearest neighbors classifier.

Results





Although high similarity was seen, Predicction 1 was less noisey than the actaul state.

Prediction2, seems to have almost the same trends like the actual situation 2

Plot couple of predicted contact maps vs actual contact maps to see model prediction.

Conclusion

- Bioinformatics involves generation of big data and this requires machine learning to ease its analysis.
- Although, Machine learning ideas have already been employed in implementation of some Bioinformatics programs, interpretations of clinical results, conclusions for diagnosis and prescription of drugs is still needs human support.
- Machine Learning in Structural Biology is of great importance especially for predicting and Interpreting 3D Protein Images.

References

Sushmita Mitra at el (2008), Introduction to Machine Learning and Bioinformatics. Â© 2008 by Taylor & Francis Group, LLC

References (for sample data used in exercises)

Dataset for this pet project is obtained from ProteinNet. And GitHub repository of ProteinNet dataset can be accessed here, ProteinNet Dataset: (https://github.com/aqlaboratory/proteinnet)

Copyright © 2018 AlQuraishi Laboratory

AlQuraishi, M. ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinformatics 20, 311 (2019). https://doi.org/10.1186/s12859-019-2932-0.

This paper is about Google Deepmind's neural network called "alphafold". These 2 videos also nicely explained "alphafold" in detail,

(AlphaFold: Improved protein structure prediction and DeepMind AlphaFold)

Senior, A.W., Evans, R., Jumper, J. et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710 (2020). https://doi.org/10.1038/s41586-019-1923-7