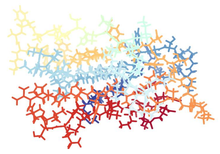
Protein side chain conformation prediction

Ayushi Sood <ayushiso@andrew>, Nikhil Bhale <nbhale@andrew>, Hongwei Ye <hongweiy@andrew>

Dataset and Task

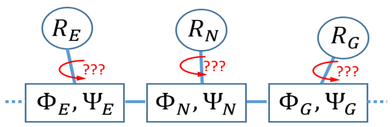
- Predicting 3D protein structure from sequence is an open problem

[PDB ID 2XKU]

...ENGKSNFLNCYVS... ??? 

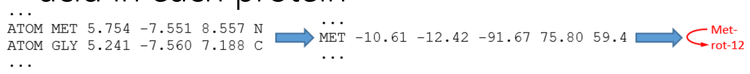
(amino acid sequence of one protein)

- We focus on a subset of the question, predicting side chain angles of amino acids given other structural and sequence information



(E, N, and G are amino acids)

- Dataset consists of 7,000 PDB (Protein Data Bank) files of experimentally determined protein structures
- Predict best rotamer from a discrete set of physically allowed "rotamers" for each amino acid in each protein



Related work

• Yedidia, Freeman and Weiss (2005) showed that **belief propagation accurately estimates thermodynamics interactions** represented by factor graphs

• Donovan-Maiye et al. (2019) show that given a fixed structure, **loopy belief propagation** on an MRF representation gives good empirical estimates of the **free energy of the structure**

• We use the above framework for prediction of side- chains angles to minimize overall energy instead of calculating the energy of a given structure (going from prediction to inference)

Methods

Data-derived vs. hybrid vs. energy-based predictions

Model 1: [Data-derived] BiLSTM

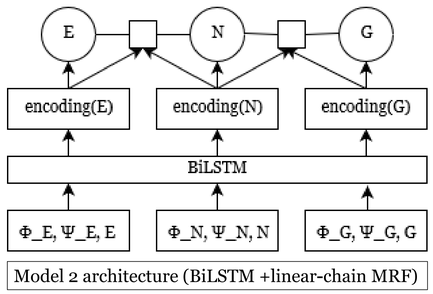
- (Φ_i, Ψ_i, i) features, trained with cross-entropy
- No knowledge encoded about spatial dependencies

Model 2: [Hybrid] BiLSTM + linear-chain MRF

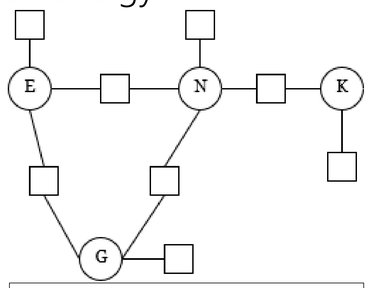
- Unary and edge potentials from BiLSTM, MRF with edges between sequential amino acids
- Get marginals and predict using belief propagation on MRF, backpropagate loss through whole network

Model 3: [Energy-based] Markov Random Field with cycles

- MRF constructed from protein spatial data (see below)
- Inference using loopy belief propagation and pick the structure which minimizes overall energy

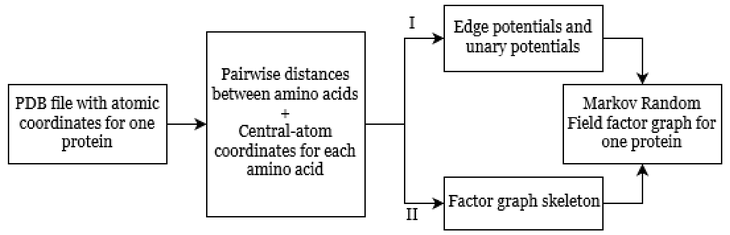


Model 2 architecture (BiLSTM +linear-chain MRF)



Model 3 architecture (MRF)

Data processing and factor graph construction



I: Edge potentials and unary potentials calculated using van der Waals energy between sidechain-sidechain (edge) and sidechain-backbone (unary)

$$E(a,b) = \begin{cases} 0 & : \quad d > R_0 \\ -k_2 \frac{d}{R_0} + k_2 & : \quad R_0 \geq d \geq k_1 R_0 \\ E_{max} & : \quad k_1 R_0 > d \end{cases}$$

$$\Psi_{ij}(r_i, r_j) = e^{-\frac{1}{T}E(r_i, r_j)}$$

$$\Psi_i(r_i) = \tilde{p}_i(r_i) e^{-\frac{1}{T}E(r_i, backbone)}$$

II: Nodes created for each amino acid, add edge between two amino acids if their distance is less than 5 Å

Results

Metric: per-tag accuracy across all proteins

Model	Train tag accuracy	Test tag accuracy
BiLSTM	56.9%	56.4%
BiLSTM + linear MRF	58.4%	57.8%
MRF	NA	

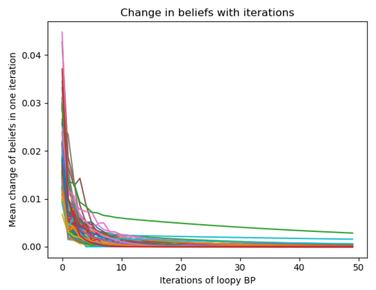


Fig1: Convergence of loopy BP on 100 proteins

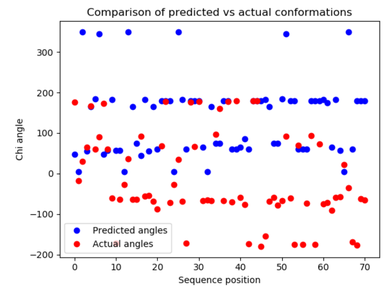


Fig2: Angles estimated by LBP vs actual angles for a representative protein

Discussion

- Clearly, conformations of amino acid sidechains are **highly dependent on spatial structure** and energy dynamics
- Only slight increase in accuracy with linear-chain MRF shows that **non-linear interactions play an important role**
- Making the graph **more connected** leads to the algorithm becoming much slower without observable gain in accuracy
- The **discretization protocol** has a dramatic effect on accuracy

Future work

- **Improving hybrid estimates** by building full MRF (model 3) on top of neural networks
- Comparing loopy belief propagation with **generalized belief propagation**
- Improving data-derived estimates through feature selection and **feature engineering**
- Moving towards **directly evaluating continuous states** rather than discretizing