Protein Secondary Structure Prediction (PSSP) Using Conditional Random Fields (CRF)

CSC 696H Fall’22 Project Proposal Presentation
(Moyeen Uddin)

Outline

1. Problem Definition
2. Background
3. Graphical Model
4. Existing Work
5. Proposal
6. Evaluation

A Similar problem: Parts of Speech Tagging

Protein Secondary Structure Prediction

Input: A sequence of Amino Acids

(eg: **N**ISQHQCVKKQCPQNSGCFRHLDEREEC…)

Output: For each position, a label (from one of 3 or 8 chars)

(eg: **H**ETHECGCE…..)

Q3: {helix (H), strand (E), and coil (C)} or

Q8: {helix (G), α-helix (H), π-helix (I), β-stand (E), bridge (B), turn (T), bend (S), and others (C)}

Parts of Speech Tagging

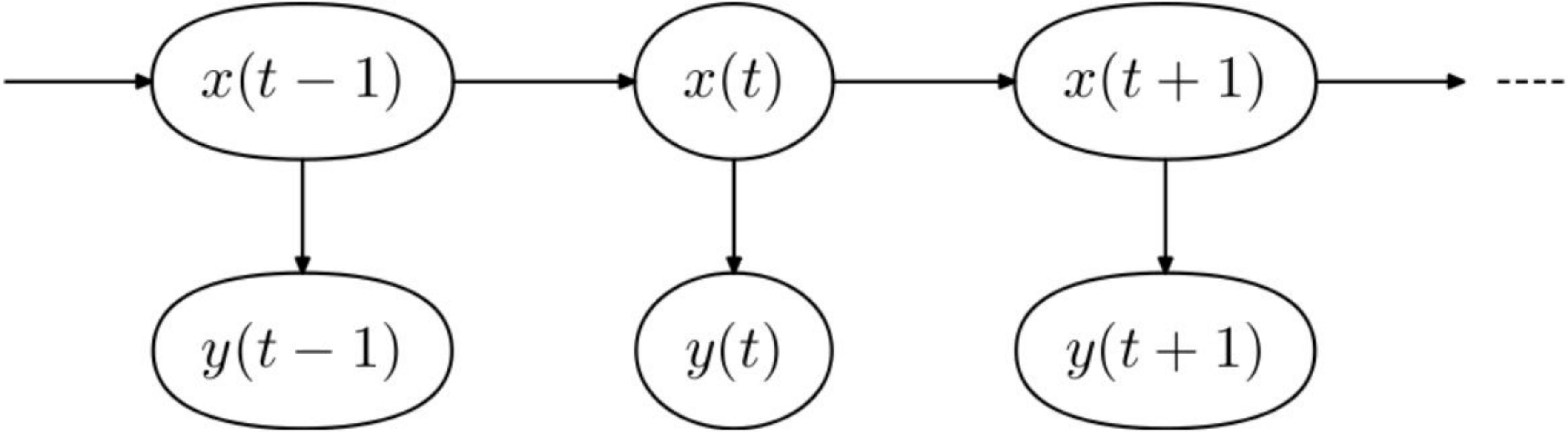
Input: Sequence of Words

(eg: "**Reality** is probabilistic…")

Output: For each word, a parts of speech tag

(eg: <**Noun**> <Verb> <Adjective> …)

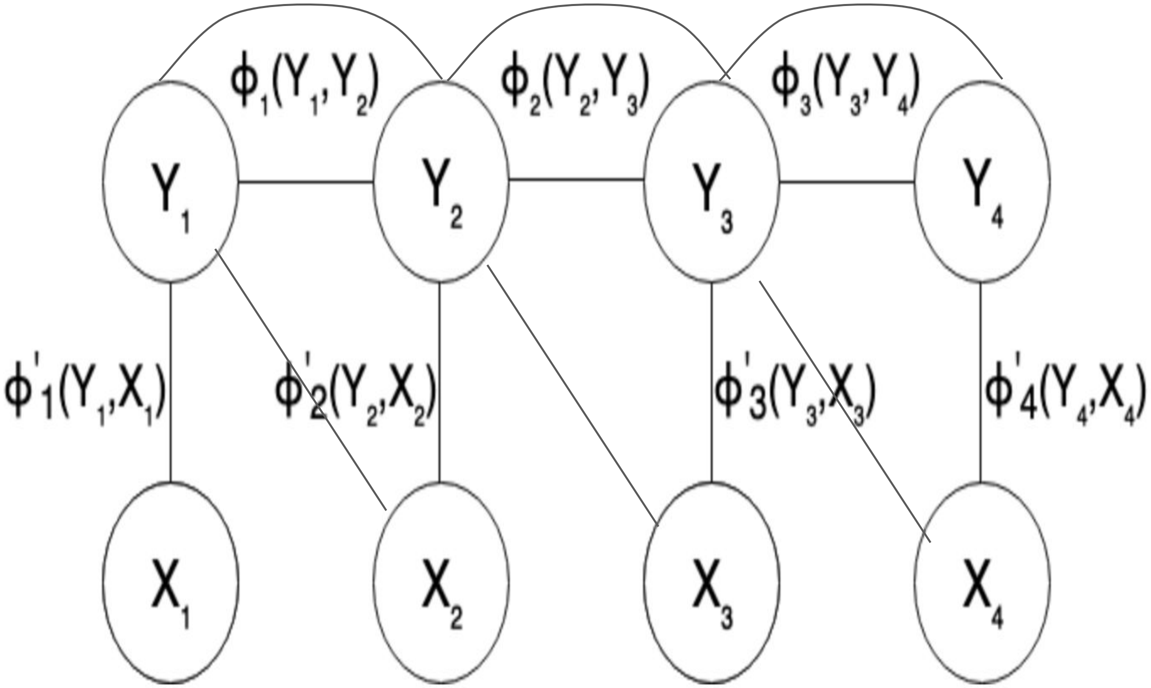
HMM



Some limitations: (1) Fixed transition and emission probabilities, (2) Emission probabilities depend only on one hidden state.

https://en.wikipedia.org/wiki/Hidden_Markov_model

Conditional Random Field (CRF) [esp: linear-chain]



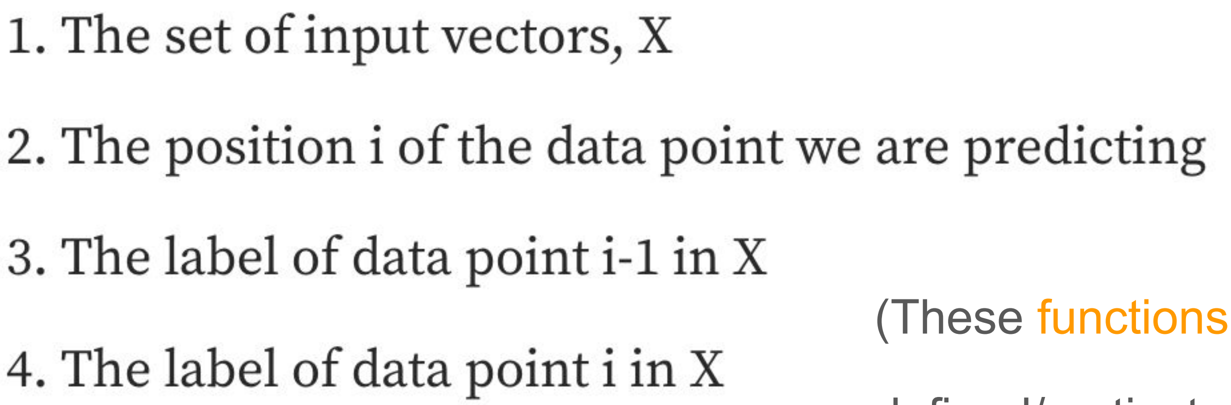
Conditional Random Field structure

1. A Markov Random Field (MRF)
   a. Hence, good to infer conditional independence structure.
   b. But complication in factorizing the joint probability distribution.
   c. Marginal (P(Y)) can be computed
2. Discriminative Model: **P(Y|X)**
3. Compared to HMM:
   a. Transition probabilities

      depend on position value: i
4. Similarities with Logistic Regression

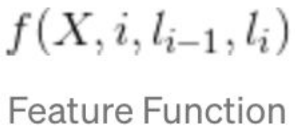
https://towardsdatascience.com/conditional-random-fields-explained-e5b8256da776

Feature Functions in a CRF



1. The set of input vectors, X
2. The position i of the data point we are predicting
3. The label of data point i-1 in X
4. The label of data point i in X

We define the feature function as:



$$f(X, i, l_{i-1}, l_i)$$

Feature Function

(These functions can be

defined/motivated from domain knowledge.

linguistic for the POST taks

(or, structural biology in the PSSP task)

$$P(y, X, \lambda) = \frac{1}{Z(X)} exp\{\sum_{i=1}^{n} \sum_{j} \lambda_j f_i(X, i, y_{i-1}, y_i)\}$$

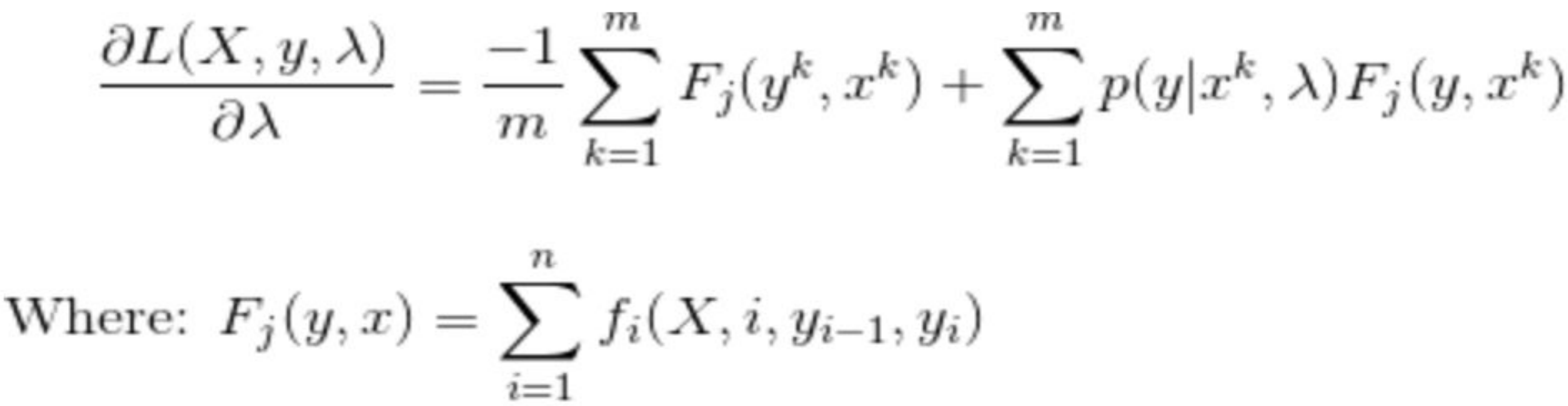
$$\text{Where: } Z(x) = \sum_{y' \in y} \sum_{i=1}^{n} \sum_{j} \lambda_j f_i(X, i, y'_{i-1}, y'_i)$$

Probability Distribution for Conditional Random Fields

$$L(y, X, \lambda) = -log\{\prod_{k=1}^{m} P(y^k | x^k, \lambda)\}$$

$$= -\sum_{k=1}^{m} log[\frac{1}{Z(x_m)} exp\{\sum_{i=1}^{n} \sum_{j} \lambda_j f_j(X^m, i, y^k_{i-1}, y^k_i)]$$

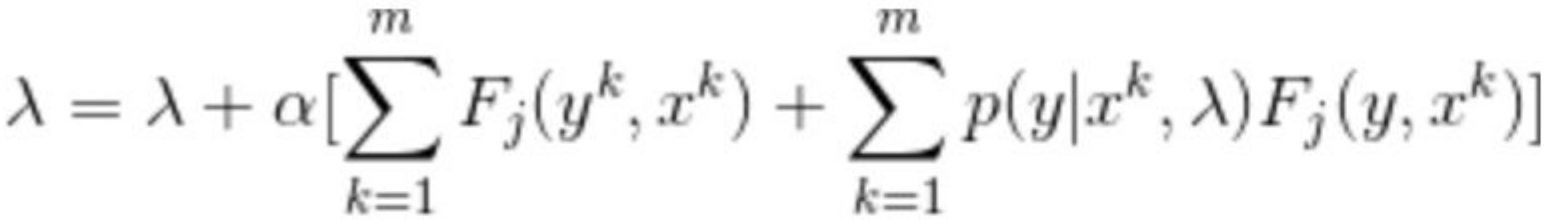
Negative Log Liklihood of the CRF Probability Distribution



$$\frac{\partial L(X,y,\lambda)}{\partial \lambda} = \frac{-1}{m}\sum_{k=1}^{m} F_j(y^k, x^k) + \sum_{k=1}^{m} p(y|x^k,\lambda)F_j(y,x^k)$$

$$\text{Where: } F_j(y,x) = \sum_{i=1}^{n} f_i(X, i, y_{i-1}, y_i)$$

Partial Derivative w.r.t. lambda



$$\lambda = \lambda + \alpha[\sum_{k=1}^{m} F_j(y^k, x^k) + \sum_{k=1}^{m} p(y|x^k,\lambda)F_j(y,x^k)]$$

Gradient Descent Update Equation for CRF

https://medium.com/ml2vec/overview-of-conditional-random-fields-68a2a20fa541

Label Prediction

1. During training, for each input point (x, y), the log-partition function Z has to be recalculated
2. During testing
   a. Global:
      i. **Most Probable Sequence:**
         1. $\text{argmax}_{y} P(Y \mid X)$ (eg: with Viterbi Algorithm)
   b. Local:
      i. **Marginal Probability:**
         1. $P(y_{i} \mid X)$: (eg: using sum-product algorithm in factor graph)

Motivation

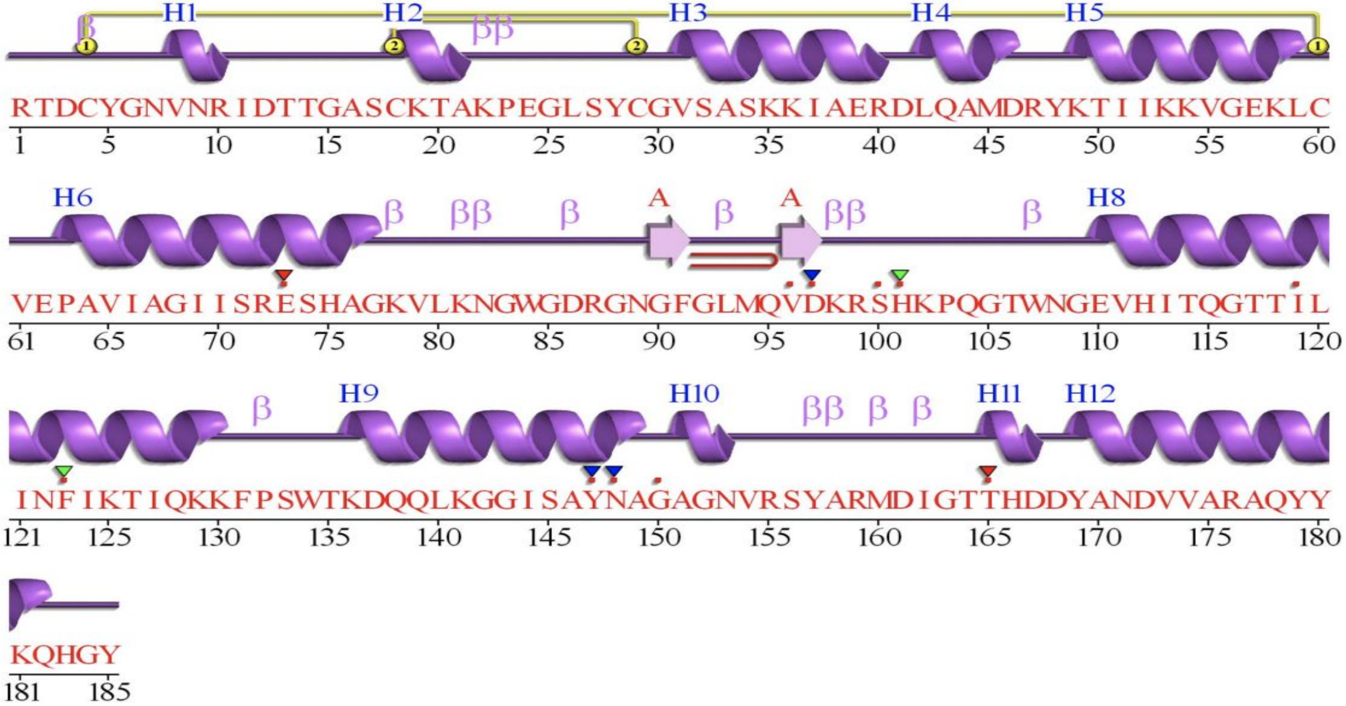


Figure 1: The amino acid sequence and its corresponding 3-state secondary structure of PDB 154L with UniProtKB accession number (P00718), which consists of 185 residues.

Existing Work

Capturing

a. Local Pattern

   i. Convolutional Architecture (CNN)

b. Global Pattern

   i. Recurrent Neural Network (RNN)

   ii. Conditional Random Field (CRF)

   iii. Or, both!

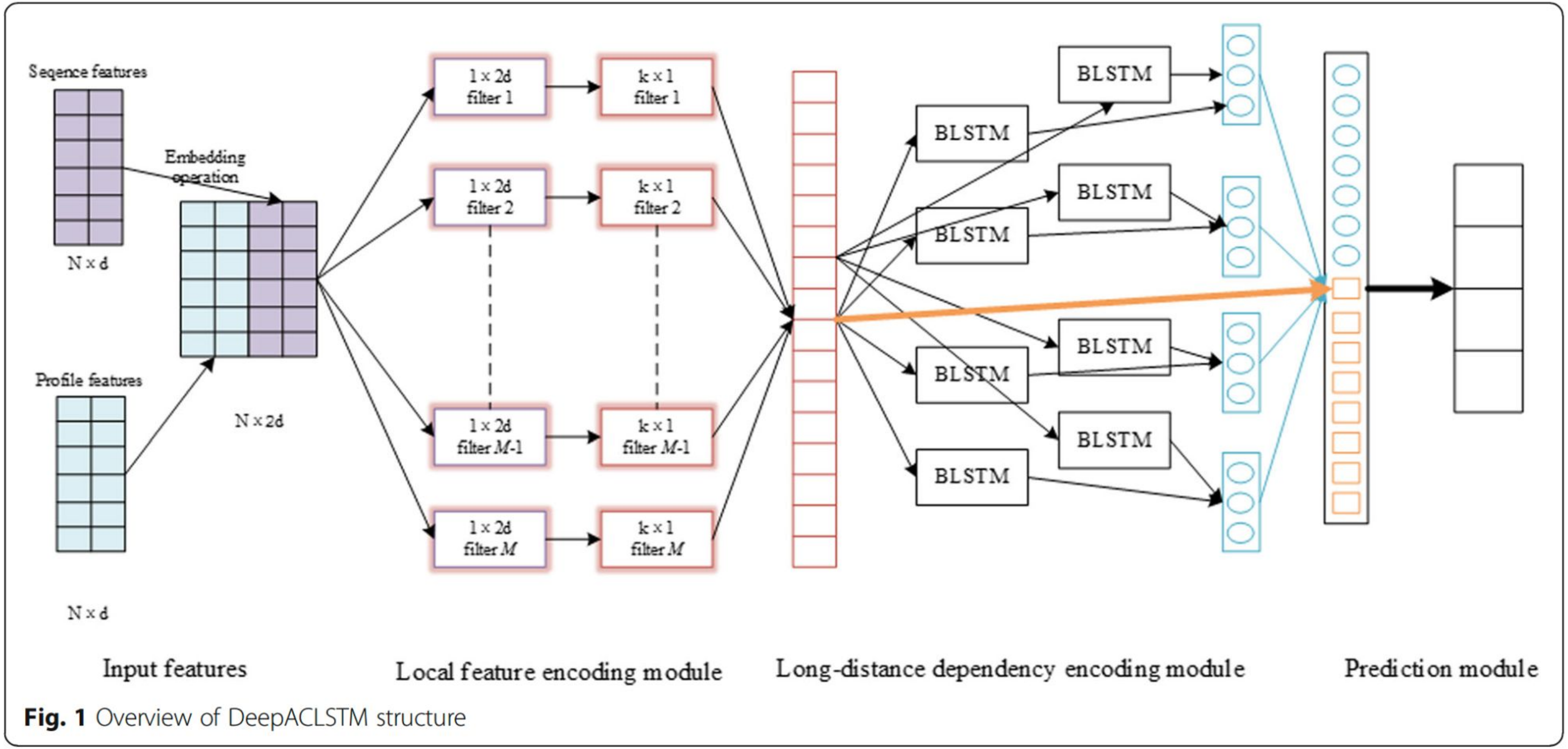
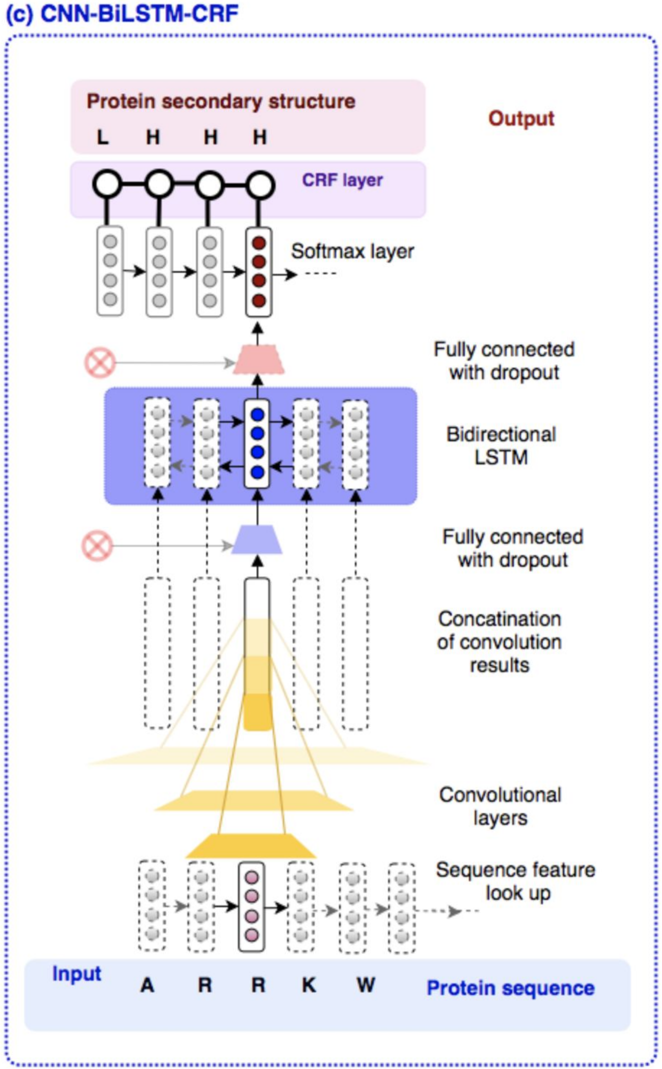


Fig. 1 Overview of DeepACLSTM structure

DeepACLSTM (Guo et al., 2019)



(Asgari et al., 2019)

Motivation

1. The input sequence are created in such a way that it encodes long range dependency relationships
2. However, the linear-chain CRF model has weaker assumptions that output at position “i” depends only on position “i-1”

Proposal

1. Data Pre-processing Focused:
    a. Using a different encoding for the inputs (based on some heuristic found in existing papers)
2. PGM focused:
    a. Incorporating more structural information into the CRF model formulation by relaxing the linear chain assumption (ie: considering long range edges)
        i. Something like **General CRF (but maybe simpler).**
    b. Evaluating how the inference complexity rises as edges are added.
    c. Finding scope for optimization in the Forward-Backward (ie: Sum-Product) Algorithm. Eg: finding out whether the existing tools doing exact or approximate computation.
    d. Being ambitious and propose a full Generative Model

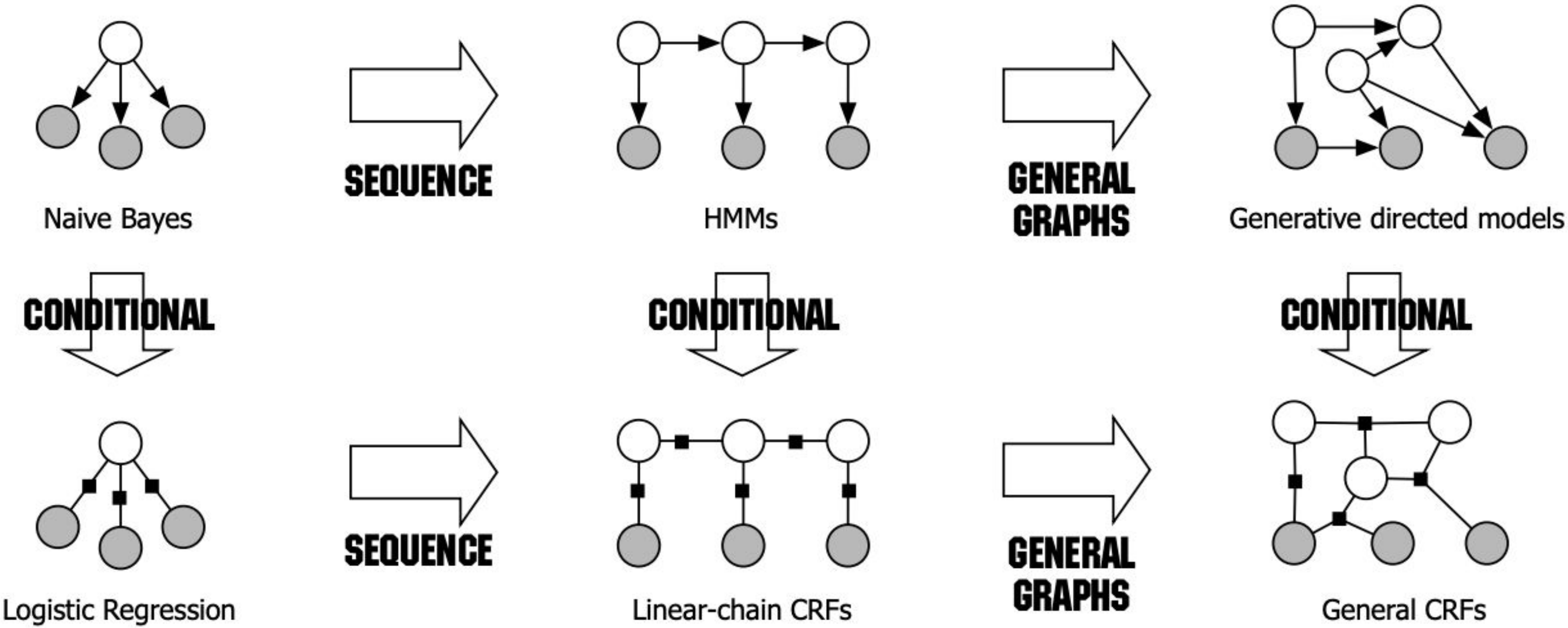


Fig. 2.3 Diagram of the relationship between naive Bayes, logistic regression, HMMs, linear-chain CRFs, generative models, and general CRFs. (Sutton & McCallum, n.d.)

Evaluation

Existing Codebases:

- Tool: biRNN-CRF
  a. https://github.com/alrojo/biRNN-CRF
- Tool: CNN+BiLSTM+CRF
  a. https://github.com/ehsanasgari/DeepPrime2Sec

Existing Tools:

- Tool: Training General CRF
  - https://mallet.cs.umass.edu/grmm/general_crfs.php

DataSet and Benchmarks:

1. Publicly available (eg: PDB (Protein Database))
2. Pre-processed Train-Test dataset from existing prediction tools.
3. Benchmark Dataset: CASP10

References:

1. Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports, 6(1), Article 1. https://doi.org/10.1038/srep18962
2. Johansen, A. R., Sønderby, C. K., Sønderby, S. K., & Winther, O. (2017). Deep Recurrent Conditional Random Field Network for Protein Secondary Prediction. Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics, 73–78. https://doi.org/10.1145/3107411.3107489
3. Asgari, E., Poerner, N., McHardy, A. C., & Mofrad, M. R. K. (2019). DeepPrime2Sec: Deep Learning for Protein Secondary Structure Prediction from the Primary Sequences (p. 705426). bioRxiv. https://doi.org/10.1101/705426
4. Sutton, C., & McCallum, A. (n.d.). An Introduction to Conditional Random Fields. 90.