Biological Sequence Analysis using Profile Hidden Markov Models

1 Introduction

The problem of sequence alignment consists in finding arrangements that aid the discovery of similarity regions in protein sequences. These similarities may be developed in such sequences because of structural or evolutionary relationships between protein families. Comparing sequence profiles can be beneficial to make assumptions about the shared evolutionary history of multiple species or predict the structure and the function of not yet discovered proteins.

In this work I present a deep learning approach for the problem of sequence alignment using a variation of Hidden Markov Models (HMMs) named Profile Hidden Markov Models (PHMMs) (Eddy, 1998). A PHMM is based on inference of knowledge and output a probabilistic model capable of classifying that a sequence belongs to a protein families or not. Algorithms for solving the problem are introduced, along with comparison between the use of computational resources, *HMMER* (pronounced *hammer*) (Finn et al., 2011) and *hmm pomegranate* an open source machine learning package for probabilistic modeling (Schreiber, 2017).

2 Hidden Markov Models

A *Hidden Markov Model* (HMM) is a type of Markov model, which is itself a subclass of *Dynamic Bayesian Networks.* Markov models are used to model randomly changing systems called Markov processes also know as Markov chains. HMMs are used to model Markov processes that we cannot observe. The hidden nature of an HMM is due to the lack of information about the value of a specific state, which is represented by a probability distribution over all possible values.

There are three main problems when analyzing and describing observations with HMMs. Firstly, one must find the probability of an observed sequence to a given model. This challenge is commonly tackled using a *Forward Algorithm* (Jurafsky and Martin, 2008), and the output can be used to select the HMM that best describes an observation. Second, one find the most likely state sequence from which an observation was drawn from. The process of discovering a sequence of hidden states given a sequence of observations is known as decoding or inference. A common approach to this process is to apply the *Viterbi algorithm* (Jurafsky and Martin, 2008). Finally, one must optimize the model, maximizing the probabilities of a given observations and learning the parameters of a HMM. The best known method to HMM learn, it is training using the Expectation Maximization algorithm (Baum, 1972).

3 Profile Hidden Markov Models

Variations of HMMs have been proposed by adding flexibility to the model, such as: Introducing new features; developing dependencies among existing feature sets; and creating additional relationships between existing features.

One of the most statistically powerful methods used to model sequence alignment is a PHMM (Russell, 2014). Well suited to the popular "profil" methods for searching databases using multiple sequence alignments instead of single query sequences. In a PHMM, each column in the alignment is represented by a frequency distribution of the symbols (called a "state"), and insertions and deletions are represented by other states. A random walk in the model follows a particular path from state to state in a Markov chain (i.e., random choice of next move), trying to match a given sequence.

A profile-HMM repetitively uses three types of hidden states, namely: *match, delete* and *insert* (Yoon, 2009; Eddy, 1998). A *match* state represents a consensus amino acid for position specific symbol frequencies in the protein family. The next matching symbol is chosen from each state, recording its probability (frequency) and the probability of transitioning to a new state. A *delete* state is a deletion of the *k-th* symbol in the original consensus sequence, representing symbols that are missing. A delete state is a non-emitting state, or a silent state, which means that it used as a placeholder that

Preprint. Under review.

interconnects the neighboring states. Finally, an *insert* state models is used to handle the symbols that are inserted between the *k-th* and the *k-th* $+1$ position in the consensus sequence (Yoon, 2009). We illustrate a consensus modeling method, from simple patterns to an HMM. Higher transitions probabilities are indicated by bolder arrows in Figure 1 (Eddy et al., 1995).

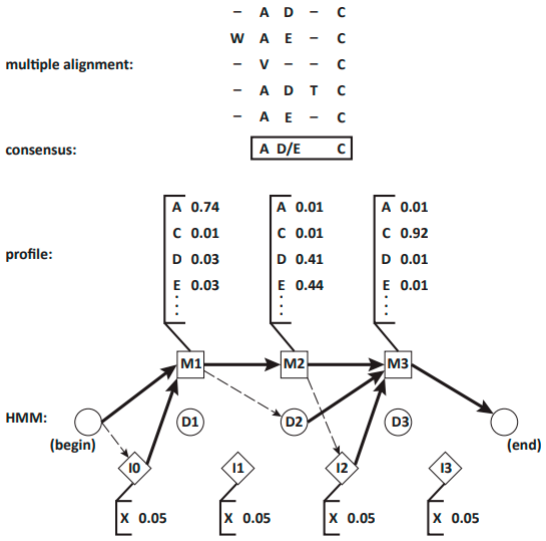

Figure 1: A prototypical profile HMM.

4 Experiments

HMM training is the estimation of the emission and transition probabilities. For PHMM, these parameters are obtained from multiple alignment sequences in a protein, DNA, or RNA sequence family. Any sequence can be represented by a path through the model. Given a PHMM, the probability of a sequence is the product of the emission and transition probabilities along the path of the sequence. The probability is then is used to calculate a *score* for the sequence, which characterizes the accuracy of the method. A score measures the probability that a sequence belongs to a given family. A high score implies that the sequence of interest is probably a member of a given class. On the other hand, a low score implies that the sequence is probably not a member of a given class.

To evaluate profile HMMs for multiple sequences alignment, the database *Pfam* (El-Gebali et al., 2018), a large collection of protein families, combined with *HMMER biosequency analysis* (Finn et al., 2011), was used and compared to the *hmm pomegranate*(Schreiber, 2017) adapted to sequence simple data, both using the Viterbi Algorithm. The FASTA (El-Gebali et al., 2018) format used, is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which a nucleotide or an amino acid is represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences, similarly exemplified in "multiple alignment", as seen in Figure 1. There are nine programs in the adaptation of the HMMER software package, one of them is *HMMerSeach*. The datasets used as a query to search a sequence database with *HMMerSeach* was *Mitoguardin (Miga) - Miga (PF10265)*. Miga Family was first identified in flies as a mitochondrial outer-membrane protein that promotes mitochondrial fusion.

5 Conclusion

The HMMER (Finn et al., 2011) learning model has proved to be more effective, and is currently one of the best software capable of performing sequence manipulation with varying parameters, that was found during this researches.

HMMs are used in a variety of applications such as speech recognition and biological sequence analysis, like DNA sequence modeling. The advance of this model can help identify families of viruses that are still being studied. We envision that this research will contribute to further biological advances in gene prediction.

References

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3(1):1–8.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763.

Eddy, S. R. et al. (1995). Multiple alignment using hidden markov models. In *Ismb*, volume 3, pages 114–120.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C., and Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl_2):W29–W37.

Jurafsky, D. and Martin, J. H. (2008). Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*.

Russell, D. J. (2014). *Multiple sequence alignment methods*. Springer.

Schreiber, J. (2017). Pomegranate: fast and flexible probabilistic modeling in python. *The Journal of Machine Learning Research*, 18(1):5992–5997.

Yoon, B.-J. (2009). Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402–415.