Propositionalisation of Multiple Sequence Alignments using Probabilistic Models

Stefan Mutter
mutter@cs.waikato.ac.nz

Bernhard Pfahringer
bernhard@cs.waikato.ac.nz

Geoff Holmes
geoff@cs.waikato.ac.nz

Department of Computer Science
The University of Waikato
Hamilton, New Zealand

ABSTRACT

Multiple sequence alignments play a central role in Bioinformatics. Most alignment representations are designed to facilitate knowledge extraction by human experts. Additionally statistical models like Profile Hidden Markov Models are used as representations. They offer the advantage to provide sound, probabilistic scores. The basic idea we present in this paper is to use the structure of a Profile Hidden Markov Model for propositionalisation. This way we get a simple, extendable representation of multiple sequence alignments which facilitates further analysis by Machine Learning algorithms.

Categories and Subject Descriptors

I.2.6.g. [**Artificial Intelligence**]: Learning—*Machine Learning*; G.3.e. [**Probability and Statistics**]: Markov processes; J.3.a. [**Life and Medical Sciences**]: Biology and genetics

Keywords

Multiple Sequence Alignment, Representation, Hidden Markov Model, Propositionalisation

1. INTRODUCTION

Sequences are omnipresent in Biology. Their analysis is the key to knowledge discovery in molecular and related sciences. From the earliest stages of research in Bioinformatics, sequence alignment techniques have played a central role and have become essential for molecular research in Life Sciences.

Ultimately, biological diversity between organisms can be explained by differences in DNA, RNA and consequently amino acid sequences. Detecting differences and similarities in these sequences can be seen as string parsing, calculating edit distances and judging similarities between strings based on some metric. However, the sequences under consideration evolve in a complex molecular world. An alignment of sequences is an explicit mapping between the positions in the sequences or respectively strings [11]. As the title of this research paper suggests, we focus on alignments. An example for a simple alignment is given in Figure 1. Alignments are used to identify common, conserved regions of interest. The significance of an alignment is given by a score. A high scoring conserved sequence region in different organisms may indicate a common functional purpose of that region. Because of the complexity of sequence evolution, scoring an alignment is not an easy task. A common example of usage of an alignment is the identification of new members of a protein family. The protein family under consideration is represented as an alignment and the sequence of interest will be aligned to it. The score indicates whether or not the sequence belongs to the specific protein family.
Sequence alignments allow biologist in general, along with other applications, to determine and understand sequence family membership, phylogenetic, functional or structural relationships [8]. They represent a compressed form of knowledge that needs to be further evaluated. Currently, this task is mostly performed by biological experts. Therefore, there exist representations that facilitate knowledge extraction by human experts. Our objective is a simple and flexible representation for multiple sequence alignments that can be easily used as input for an arbitrary, propositional machine learning algorithm. By regularising the input we can gain from many recent advances that have been made in propositional machine learning.

This paper is structured in the following way. In the next section we introduce multiple sequence alignments. Section 3 presents representations beginning with human readable ones and introducing Profile Hidden Markov Models [3, 9, 5, 4] as a representation for multiple sequence alignments. They offer the advantage of a sound, probabilistic score for the alignment. The following section introduces the concept of propositionalisation and explains how we use it together with Profile Hidden Markov Models to build a new simple representation for a multiple sequence alignment. The last section summarises the paper and gives an outlook of the research project.

2. MULTIPLE SEQUENCE ALIGNMENT

In this paper we are investigating global, multiple sequence alignments (MSA). An MSA consists of at least three sequences. In this research project we align sequences globally, meaning that the whole length of a sequence gets aligned.

```
EGALLIVEHTNINAMrevetahwYLNEHTMI
EGAKHIVEHTNINAM........YLNEHTMI
-------EHTNINAM........YLNEHTMI
```

Figure 1: A global multiple sequence alignment for three amino acid sequences.

```
MMFFADDAAA
MMFFARRNSSTNNRREDPFMLW
MMFFA
```

Figure 2: Three short, unaligned amino acid sequences.

Figure 1 shows an example of a global MSA for three amino acid sequences. An MSA reflects the underlying sequence evolutionary process. A sequence can evolve in three different ways:

- A specific position in the sequence can be altered. In Biology this phenomena is called a point mutation. From a Computer Science viewpoint this is just a change of a single character in a string.
- Parts of the sequence can get deleted.
- Parts of the sequence can be newly inserted.

If there is no difference in a specific region of a sequence, this region is labelled as conserved. Generally a conserved region may indicate a conserved function or structure. In an alignment these conserved sequence positions are called matches. Therefore in an alignment we can find matches, insertions and deletions. By convention, matches are displayed as upper case letters, whereas inserts are represented as lower case letters. A deletion is symbolised by a dash. Alignments are always displayed with equal length so that conserved regions are one below another. That is why in all sequence strings (except the longest ones) dots are inserted. This representation makes it easier to human expert to analyse the alignment.

3. REPRESENTATIONS FOR A MULTIPLE SEQUENCE ALIGNMENT

For the remainder of this paper we use a simple MSA of three amino acid sequences to illustrate our approach. Figure 2 shows the unaligned sequences.

A major problem in sequence alignment construction is how to deal with point mutations. Mutations vary greatly in their effect. A change at a sequence position doesn't necessarily mean that there is a functional, structural or any other consequence. This happens when a sequence element mutates into a different one with very similar chemical properties. In order to allow a biological expert to further investigate alignment colour coding schemes are used. Each colour

```
MMFFA..........DDAAA--
MMFFArrnsstnnrrEDPFMLW
MMFFA..........-------
```

Figure 3: An MSA enhanced with colour codes for amino acid properties. Blue represents acidic amino acids, red small and hydrophobic ones. The amino acids with one letter code R, H and K are coloured magenta, whereas green is used for S, T, Y, H, C, N, G and Q.

Figure 4: A sequence logo representation for our simple example MSA. The logo is produced using the WebLogo software [2].

represents a chemical property. We would expect match regions to have the same colour. Figure 3 shows a colour coded MSA for our example sequences.

An alignment contains information about the distribution of DNA or RNA nucleotides or amino acids at each position. So far this information isn't clearly represented. A sequence logo [10] as shown in Figure 4 gives a good visual overview of the distribution of nucleotides or amino acids in each sequence position. The height of a letter indicates its frequency at a given sequence position. In addition colour codes for properties are used.

It is a challenging and problem domain dependent task to score alignments in a biologically meaningful way. A common approach uses a probabilistic measure. Profile Hidden Markov Models are a widely used representation for biological sequence alignments. They are probabilistic graphical models. Thus, they allow the calculation a probability for each sequence. We will introduce Profile Hidden Markov Models in the next section.

3.1 Profile Hidden Markov Models

A Profile Hidden Markov Model [3, 9, 5, 4] is a probabilistic graphical model. It consists of three different sets of states: match, insert and delete states. Consequently, a match states models a match in an alignment position. In the same way delete and insert states model deletions and insertions. In each alignment position a sequence can either match, or there is a deletion or an insertion. This is why the structure of Profile Hidden Markov Models is organised in columns. There is a column for each position in the alignment and therefore, each column consists of exactly one match, insert and delete state.

In practice slightly different structures for Profile Hidden

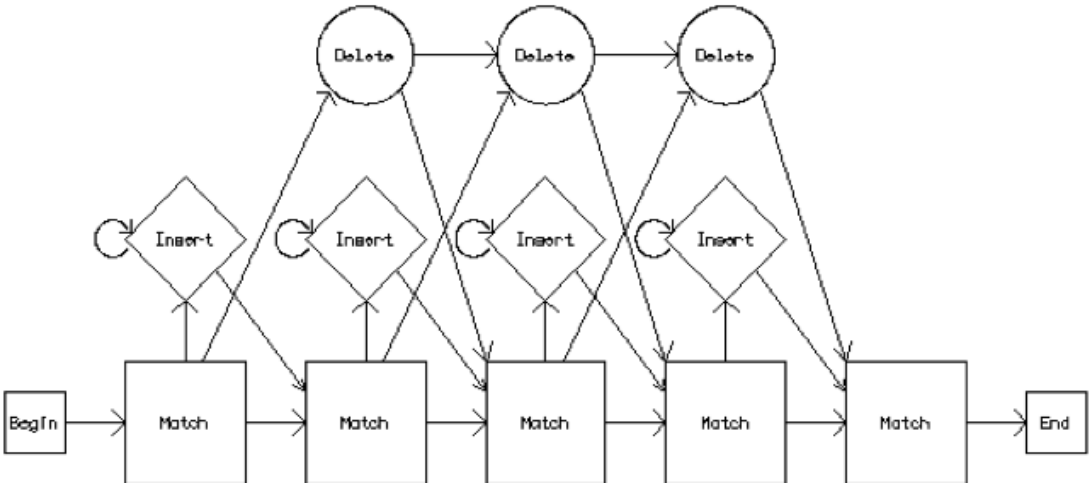

Figure 5: The general structure of a Profile Hidden Markov Model. This Figure is taken from [1].

Markov Models exist. An example is the architecture of HMMER [4], the most prominent Profile Hidden Markov Model implementation. In our implementation the first and last column are exceptions to the general rule (see Figure 5). Depending on the implementation there might also be slight differences concerning the state transitions. A discussion on this topic can be found in Durbin et al. [3]. The structure of the Profile Hidden Markov Model that we used in our experiments is shown in Figure 5.

Each match and insert state can emit sequence symbols. The set of all possible symbols of a sequence is called an alphabet. There are probabilities for each transition and for each emission in a match and insert state. A delete state does not emit a symbol.

There are different learning algorithm for the transition and emission probabilities of a Profile Hidden Markov Model. We use he most common method, the Baum-Welch algorithm [3]. It is an EM derived algorithm and guaranties convergence to a local optima. This trained Profile Hidden Markov Model represents an MSA.

Traditionally, Profile Hidden Markov Models are used to construct and represent MSAs. In this paper we also use the Profile Hidden Markov Model for propositionalisation. The next section introduces the basic idea.

The Viterbi algorithm [3] allows to efficiently calculate the most probable path for a sequence to be aligned through a trained Profile Hidden Markov Model. This probability indicates whether or not the sequence belongs to the existing MSA.

4. PROPOSITIONALISATION

Propositionalisation is an approach widely used in the Machine Learning community [7]. Basically it is nothing more than a change of representation. It refers to the transformation of relational data into a feature-based, attribute value representation [6]. This representation is then consequently called a propositional one. Traditionally, Machine Learning algorithms like decision trees and support vector machines work on propositionalised data. Therefore, propositionalisation leads to a wide range of learning algorithm that can be applied in a consecutive step. It decouples the feature construction from the model construction [6]. In many problem settings propositionalisation leads to equivalent or better results [7, 6].

In this research project we change the representation of an MSA as Profile Hidden Markov Model to an attribute-value representation. This step is called the propositionalisation using a Profile Hidden Markov Model and will be explained in the following section.

4.1 Propositionalisation using a Profile Hidden Markov Model

The propositionalisation is guided by the structure of the Profile Hidden Markov Model. Each column of the Profile Hidden Markov Model is transformed into two features or attributes. The first attribute is nominal and represents the match and delete state of the column under consideration. Thus its values can either be any letter from the sequence alphabet or the deletion symbol. The type of the second attribute is numerical and counts how many times the column's insert state was visited.

As a first step an MSA is built and represented using a Profile Hidden Markov Model. To propositionalise a new sequence we simply calculate the Viterbi path for this sequence.

Figure 6 shows the propositional representation for our simple example MSA.

Therefore we use the Profile Hidden Markov Model to construct features for our new representation. This propositionalised data is then consequently used for model construction in the learning step.

This is a very simple representation that can be used as input

```
M,0,M,0,F,0,F,0,A,0 ,D,0,D,0,A,0,A,0,A,0,-,0,-
M,0,M,0,F,0,F,0,A,10,E,0,D,0,P,0,F,0,M,0,L,0,W
M,0,M,0,F,0,F,0,A,0 ,-,0,-,0,-,0,-,0,-,0,-,0,-
```

Figure 6: The propositional representation of the simple example MSA

for propositional machine learning algorithms. In addition this representation can be extended. The Profile Hidden Markov Model provides a probability for the Viterbi path. This value can be added as an extra attribute. Furthermore emission and transition probabilities can be used as features as well.

5. PROJECT STATUS AND OUTLOOK

This paper presents parts of the first author's PhD project. We outline a technique to propositionalise a Profile Hidden Markov Model that represents an MSA. This representation is simple and easy to use as input for learning algorithms. Therefore it is useful in further analysing in silico the information stored in an MSA.

In this paper we use the Profile Hidden Markov Model in three ways. It constructs, represents and propositionalises an MSA. The advantage of this approach is that a Profile Hidden Markov Model allows the probabilistic scoring of aligning a new sequence. We can extend our simple representation to incorporate this score. In addition we can also take advantage of having emission and transition probabilities and include them as a feature in our representation. Generally, any MSA, not only the ones represented as Profile Hidden Markov Model, can be transformed into our basic simple representation without probabilistic extensions.

Preliminary results show that this approach leads to a better predictive accuracy than a pure Profile Hidden Markov Model approach.

The next stage of the project includes intense, experimental testing to verify whether or not propositionalisation enhances the models and therefore in silico analysis of MSAs. The experimental results will be presented at the conference.

In the future we plan to apply semi-supervised learning to this experimental setting to determine to what extent this technique can benefit from propositionalisation.

Acknowledgements
This work has been partially funded by the Marsden Grant UOW-306 of the Royal Society of New Zealand.

6. REFERENCES

[1] http://helix.nih.gov/docs/gcg/hmmanalysis.html, February 2001.

[2] G. Crooks, G. Hon, J. Chandonia, and S. Brenner. Weblogo: a sequence logo generator. *Genome Research*, 14:1188–1190, 2004.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. 1998.

[4] S. Eddy. Hmmer user's guide: biological sequence analysis using profile hidden markov models. http://hmmer.wustl.edu/, 1998.

[5] S. Eddy. Profile hidden markov models (review). *Bioinformatics*, 14(9):755–763, 1998.

[6] S. Kramer, N. Lavrač, and P. Flach. Propositionalization approaches to relational data mining. In *Relational Data Mining*, pages 262–286. 2000.

[7] M. Krogel, S. Rawles, F. Zelezny, P. Flach, N. Lavrac, and S. Wrobel. Comparative evaluation of approaches to propositionalization. In *Proceedings of the 13th Int. Conference on Inductive Logic Programming*, pages 197–214, 2003.

[8] C. Notredame. Recent progresses in multiple sequence alignment: a survey. *Pharmacogenomics*, 3:131–144, 2002.

[9] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[10] T. Schneider and R. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18:6097–6100, 1990.

[11] G. Schuler. Sequence alignment and database searching. In A. Baxevanis and B. Ouellette, editors, *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Wiley-Liss, Inc., 2001.