Jurnal TIMES , Vol. V No 2 : 55-57 , 2016
ISSN : 2337 - 3601

USING PROFILE HIDDEN MARKOV MODELS IN MULTIPLE SEQUENCE ALIGNMENT

Edi
Department of Information System
STMIK TIME
Jalan Merbabu No. 32 AA-BB, Medan, North Sumatra 20212, Indonesia
e-mail: edi_foe@yahoo.com

Abstract

Sequence alignment is a method of arranging sequences of DNA, RNA, or protein to identify similar regions between the sequences. Profile HMM is a type of HMM that can be used to construct family structure of multiple sequence alignment so that it can be used as a database for aligning or finding other sequences in that family. Literature review used as a method in this paper. Softwares BioEdit and HMMER are introduced.

Keyword: *profile Hidden Markov Model, Multiple Sequence Alignment, computational biology*

1. Introduction

A Hidden Markov Model (HMM) is a statistical model that has been used for speech recognition since the early 1970s [1]. HMMs have been successfully implemented in computational biology field such as multiple sequence alignment, genetic mapping, secondary structure protein prediction, gene finding, signal peptide prediction, transmembrane protein prediction, epitope prediction, RNA secondary structure prediction, and phylogenetic analysis [2]. T. Nguyen et al. [3] used HMMs for cancer classification using gene expression profiles. The authors compared HMMs with six other classifiers: k-nearest neighbors (kNN), probabilistic neural network (PNN), support vector machine (SVM), multilayer perceptron (MLP), fuzzy ARTMAP (FARTMAP), and ensemble learning AdaBoost. They found that HMMs are the most robust method among seven other classifiers. They also found that by combining modified analytic hierarchy process (AHP) and HMMs can improve cancer classification performance by individually increasing the efficiency of gene selection (using modified AHP) and also the classifier (HMM).

The most well-known use of HMM in molecular biology is a profile HMM, a 'probabilistic profile' of DNA/protein family [1]. A profile HMM of DNA/protein family can be useful for searching a database of other members of the family [1]. It uses the sequence family to build a profile based on position-specific probabilities of variation in nucleotides or amino acids, including insertions and deletions [a]. Profile HMMs show good performance in sequence alignment because they contain more information about sequence family than a single sequence [a].

2. Method

Literature survey is used as research method in this paper. In section 2, we will introduce multiple sequence alignment, profile HMMs, and implementing profile HMMs in multiple sequence alignment.

2.1 Multiple Sequence Alignment

Sequence alignment is a way of arranging sequences of DNA, RNA, or protein to identify similar regions that could be results of functional, structural, or evolutionary relationships between the sequences [b]. There are two types of sequence alignments namely pair-wise sequence alignment and multiple sequence alignment. A pair-wise sequence alignment is comparing two sequences of DNA, RNA, or protein whereas a multiple sequence alignment (MSA) is comparing more than two sequences. A MSA is obtained by inserting gaps (-) into the sequences so that all sequences have same length N and can be arranged into K rows and N columns in which each column represents homologous position [b].

There are few main applications of MSA [b]:

a. Extrapolation

A good MSA can help convincing that an uncharacterized sequence is really a member of a protein family.

b. Phylogenetic analysis

MSA can be used to infer a evolutionary tree based on amino or nucleic acid sequences.

c. Pattern identification

By discovering very conserved positions, we can identify a region that is characteristic of a function.

d. Domain identification

It is possible to turn a MSA into a profile that describes a protein family or a protein domain. This profile can be used to scan databases to find new members of family.

e. DNA regulatory elements

We can turn a DNA MSA of a binding site into a weight matrix and scan other DNA sequences for potential similar binding sites.

f. Structure prediction

A good MSA can give an almost perfect prediction of secondary structure for proteins or RNA. It can also help building a 3D model.

MSA is generally a global multiple sequence alignment [c]. Several tools available for MSA such

as MUSCLE, T-Coffee, MAFFT, and CLUSTALW [c].

Figure 1. Global alignment in MSA [c]

Figure 2 shows a MSA in sequences that assumed to be homologous (i.e. having a same ancestor) [d]. Homologous nucleotides are put in the same column of a MSA. Phylogenetic tree is shown at the left side.

Figure 2. Phylogenetic tree in MSA [d]

2.2 Profile Hidden Markov Models

A profile HMM is a type of HMM that allows position dependent gap penalties [1]. Figure below shows the structure of profile HMM.

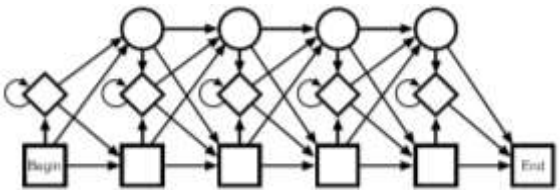

Figure 3. The structure of profile HMM [1]

In figure 3, the first bottom row "square shape" is called as match states. They represents the columns of multiple alignment. The probability distribution in these states is the frequency of amino acids or nucleotides of protein or DNA [1]. The second row "diamond shape" is insertion states. They are used to represent highly variable regions in the alignment [1]. The top row "circular shape" is delete states. In these states, they do not match any residues and the function is to skip over another column of the alignment.

Besides three types of states in profile HMM architecture, there are two sets of parameters: transition probabilities and emission probabilities [4]. Emission probabilities show the possibility of match states in generating one residue of nucleotides or amino acids. Transition probabilites show the transitions from one match state to another match state (from left state to right state) in the model.

Table 1 shows several HMM software packages along with internet sources. One important difference among theses packages is in model architecture.

Table 1. profile HMM software

Software	URL
SAM	https://compbio.soe.ucsc.edu/sam2src/
HMMER	http://hmmer.org/
PFTOOLS	http://web.expasy.org/pftools/
GENEWISE	http://www.ebi.ac.uk/Tools/psa/genewise/
META-MEME	http://meme-suite.org/
Pfam	http://pfam.xfam.org/
PSI-BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE=Proteins&PROGRAM=blastp&RUN_PSIBLAST=on

Figure 4 shows the model architecture in profile HMM software packages. In figure [5] from top to bottom: META-MEME shows multiple motif model; the original profile HMM of Krogh [1]; the 'Plan 7' architecture of HMMER2, representative of the new generation of profile HMM software in SAM, HMMER and PFTOOLS.

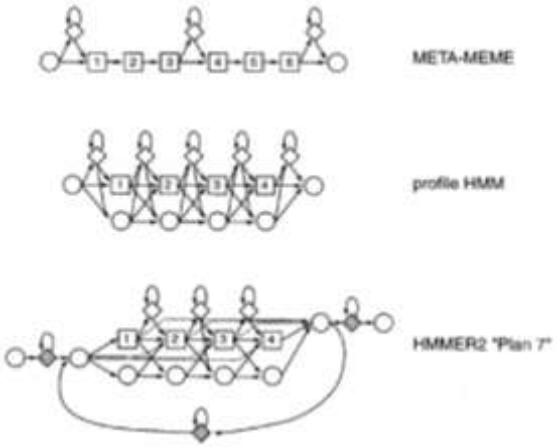

Figure 4. Different model architectures [5]

2.3 Implementing Profile HMM in MSA

Below is the example of MSA of nucleotides, taken from [1].

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

Figure 5. DNA motif [1]

To build profile HMM, each column of MSA is regarded as match state except gaps. The first match state will generate emission probabilities 0.8 for residue 'A' (there are four 'A' so 4/5 = 0.8) and 0.2 for residue 'T'. Transition probabilities from first match state to second match state will be 1.0 since there is no gap. From third match state (third column) to fourth column, there are two possibilites, since there are gaps in fourth column. Transition probabilites from third match state to another match state (top) will be 0.6 (gaps are ignored). In top match

state, there is insertion (emission probabilities is 0.4) because there are residues in fifth and sixth column in second sequence. Transition probabilities from third match state to another match state (bottom) that skip over gaps is 0.4. Transition probabilites that come out from match state must be sum to 1. This also applies to emission probabilities. The result of profile HMM structure based on DNA motif above is shown in figure 6.

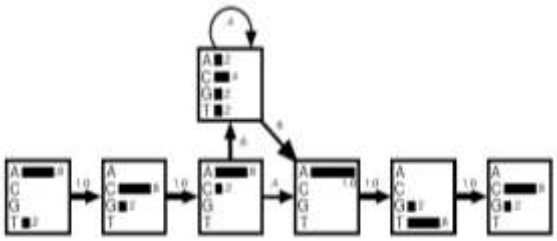

Figure 6. Profile HMM derived from figure 5 [1]

We also used software HMMER3.1 to generate profile HMM based on DNA motif above. First, we created MSA using BioEdit Sequence Alignment Editor and save as FASTA format. After that we use command hmmbuild in HMMER3.1 to generate profile HMM. Figure 7 shows the result.

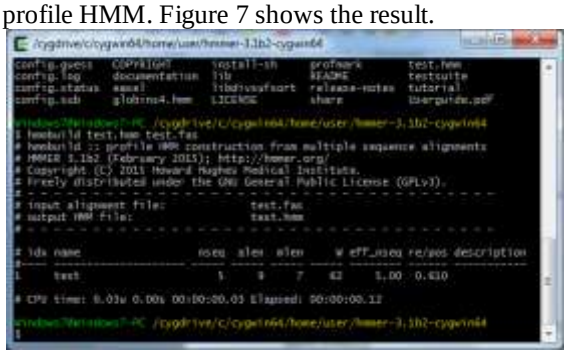

Figure 7. Profile HMM using HMMER3.1

The result shows that filename test.fas that contains DNA motif consists of five sequences with 9 aligned columns. HMMER formed profile HMM of 7 consensus positions (look at figure 6: seven match states) which means there are 2 gap-containing alignment columns to be insertions relative to consensus.

3. Discussion

Method of literature review uses sources from journals, conferences, and websites. We also used softwares BioEdit to write down MSA and HMMER3.1 to build profile HMM based on MSA.

4. Result

The purpose of this paper is to introduce MSA and profile HMM in computational biology field. Profile HMM is useful to build family structure of DNA or proteins based on MSA. We also introduce softwares such as BioEdit as a sequence editor and HMMER to build profile HMM.

In future, we will introduce sequence alignment search based on structure of profile HMM that has been built.

5. Daftar Pustaka

[1] A. Krogh, *Computational Methods in Molecular Biology edited by S.L.Salzberg, D. B. Searls and S. Kasif*, pp.45-63, Elsevier, 1998.

[2] V.D.Fonzo, F. Aluffi-Pentini,and V.Parisi, "Hidden Markov Models in Bioinformatics", *Current Bioinformatics*, vol. 2, no. 1, pp. 49-61, 2007.

[3] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi. "Hidden Markov models for cancer classification using gene expression profiles", *Information Sciences*, vol. 316, pp. 293-307, 2015.

[4] K. H. Choo, J. C. Tong, and L. Zhang, "Recent Applications of Hidden Markov Models in Computational Biology", Geno. Prot. Bioinfo., vol. 2, no. 2, pp. 84-96, 2004.

[5] S. R. Eddy, "Profile hidden Markov models", Bioinformatics Review, vol.14, no.9, pp. 755-763, 1998.

Websites:

[a] N. Mimouni, G. Lunter, C. Deane, Hidden Markov Models for Protein Sequence Alignment. https://www.stats.ox.ac.uk/__data/assets/file/0012/3360/naila_report.pdf.

[b] Anonymous, Multiple Sequence Alignment, www.srmuniv.ac.in/sites/default/files/files/1(7).pdf.

[c] Anonymous, Difference between Pairwise and Multiple Sequence Alignment. http://www.majordifferences.com/2016/05/difference-between-pairwise-and-multiple-sequence-alignment.html#.V9ZpTaJUXCs.

[d] Clemens Gropl et al., Multiple Sequence Alignment, www.mi.fu-berlin.de/.../MultipleAlignmentWS12/multal.pdf, 2012.