Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 5 (2015), pp. 3665-3675

http://www.ripublication.com

Profile Hidden Markov Model For Sequence Alignment To Cancer Sequence

R.Sasikumar and V.Kalpana

Department of Statistics,
Manonmaniam Sundaranar University, Tirunelveli-627 012.
Emails:sasikumarmsu@gmail.com and kalpanaselwyn@gmail.com

Abstract

Profile Hidden Markov models (PHMMs) are remarkable types of Hidden Markov Model (HMM) and much used in biological sequence analysis. We emphasize the use of PHMM for Multiple Sequence Alignment (MSA). Modern bioinformatics systems use the effective technique of MSA in all of their applications. The Biomedical method and algorithms used in MSA have enormous importance in solving a series of biological related problems. The renowned and broadly used statistical method of characterizing the unique properties of the reaming of a genomic or proteomic pattern is the HMM approach. PHMMs have proved to provide a better solution for MSA. In this paper we have applied the PHMM in cancer genomic sequence and then we found that optimal path of sequence. Finally, we noticed that PHMM gives a better solution for cancer sequence.

Key words: Hidden Markov Model, Profile Hidden Markov models, Multiple Sequence Alignment, Cancer sequence.

1. Introduction

Sequence alignment is a method of writing one particular sequence on the top of other sequences where the remaining in one position are said to have a general evolutionary origin. If the same letter comes in both sequences then this position has been placed in evaluation. If the letters vary it is assumed that the second come from a formal letter. Some sequences have variable in length but same sequences are explained through insertions or deletions in sequences. Hence, a letter or a total of letters may be paired up dashes in the following sequence to show such an insertion or deletion. Hence insertion in a one sequence usually shown as a deletion of the other one. The Moto of MSA is to find out the similarities between many other sequences. All the similar

sequences can be compared in a single table or figure. Due to setup a coordinate system, the sequences are aligned on top of each other. In that every row is assumed as the sequence for one protein. In every sequence each column is in the same position. In order to score the sequence alignment, the MSA use the substitution matrices. To study the sequence alignment PHMMs are considered the as a remarkable statistical tool. Bioinformatics is a new scientific field and PHMMs used to analyze sequence alignment. For estimating the model parameters, PHMMs provide a more methodical approach. The main motivation for doing this comparison work is to find the evolutionary relationship between species on a molecular level.

2. Review of literature

Many more important problems in biological sequence analysis have been resolved due to the essential role of HMMs in engineering. For biological researchers, Durbin et.al [1] discussion on probabilistic models of proteins and nucleic acid is very useful. Churchill et.al [2] explained the DNA sequence as a stochastic process, the states of a Markov chain. Brown et.al [3] used HMM to derive for protein families. Asai et.al [4] introduced the prediction system of protein secondary structure by HMM. Baldi et.al [5] established the algorithm for the transition and emission parameters of HMM. Eddy [6] initiated simulated annealing method and produced multiple sequence alignments from unalign proteins or DNA sequences. Sojilander et.al [7] presented a method for detecting the weak but significance protein sequence homology. Bateman et.al [8] presented HMM to detect Fibronectin type III domins in yeast. Birney et.al [9] developed code generating language for biological sequence comparison. Barrett et.al [10] analyzed scoring methods to compare probability of sequence generated by HMM. Dalgaard et.al [11] used HMM to identify the KlbA proteins which are used in the formation of surface-associated protein complexes. Sonnhammer et.al [12] apply the domineer algorithm to cluster an align protein sequences after removing Pfam-Adomains. Francesco et.al [13] apply the HMM to alpha class proteins even when to detectiable primary amino acid sequence similarity is present. Ahola et.al [14] finding the efficient estimation of emission probabilities in profile HMM. Liang et.al [15] proposed Bayesian approach basecalling for DNA sequence analysis using HMM. Madera [16] introduced Profile comparer program for scoring and aligning profile HMM of protein families. Baldi et.al [5] first proposed the modeling the characteristics of a number of protein families such as globins, immunoglobulins and kinases. Eddy [17] introduced the simulated annealing method and produced multiple sequence alignments from unaligned protein or DNA sequences. Difranceco et.al [18] found the method for fold recognition from secondary structure predictions of proteins which is based on HMM. They used FORESST web server of the library of HMM of structural families. Wistrand and sonnhammer [19, 20] improved Profile HMM discrimination by adapting transition probabilities and also discussed Profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. Söding [21] illustrated Profile HMMs to detect the protein homology and sequence alignment of protein structure prediction, function prediction and evaluations

3. Preliminaries

3.1 Biological sequence analysis

Deoxyribonucleic Acid (DNA), Ribonucleic Acid (RNA) and Proteins are the basics of our body. These three are molecules. DNA is composed of adenine (A), cytosine (c), guanine (G), and thymine (T). Like this, RNA has the following adenine (A), Cytosine (C), guanine (G), and uracil (U). Similarly, there is only one difference between DNA and RNA is that RNA has uracil instead of thymine. Proteins are having different structure and function compared to other molecules and having 20 Amino acids which are represented as A, V, L, I, F, P, M, D, E, K, R, S, T, C, N, Q, H, Y, W, and G. The molecules are linked in a linear sequence. That sequences are known as biological sequences. This simple sequence representation of the molecules enables them to be compared in a simple way. So it is sure to match or align two sequences and to see how they match up. The main reason for doing this comparison is to find the evolutionary relationship between species on a molecular level. HMM are the most valuable methods used to model sequence alignment. In that, PHMMs has vital role to model multiple alignments.

3.2 Multiple sequence alignment

A multiple sequence alignment is a sequence of alignment comprising three or more biological sequences commonly DNA, RNA and protein. MSA means a process of aligning biological sequences. The biologically symmetrical length sequences can be hard and take more times to align by hand. MSA need more comfortable methodologies than pair-wise alignment since they are much complex to computation.

3.3 profile HMM

PHMMs are statistical tools that can model the commonalities of the amino acid sequences for a family of proteins. Considered to be more expressive than a standard consensus sequence or a regular expression, PHMMs allow position dependent insertion and deletion penalties, as well as the option to use a separate distribution for inserted portions of the amino acid sequence. Once a model is trained on a number of amino acid sequences from a given family or group, it is most commonly used for the following purposes:

1. By aligning sequence to the model, one can construct multiple alignments.
2. The model itself can offer insight into the characteristics of the family when one examines the structure and probabilities of the trained HMM
3. The model can be used to score how well a new protein sequence fits the family motif.

4. Methodology

4.1 HMM basis

HMM is a stochastic model which is not directly observable. But describes the observable events that are depends on internal factors. The observable events are represented as symbols, where the invisible factor involved in the observation is represented as a state. Among the two stochastic processes, one process is called the

hidden states and other is visible process or observable symbols. So it is also called a doubly embedded stochastic Process [1]. There are two layers, visible and invisible presented in HMM, which is very useful in many real world problems. Thus if $S=\{S_n, n=1,2,...\}$ is Markov process and $V=\{V_k, k=1,2,...\}$ is a function of S, then S is a Hidden Markov process that is observed through V and we can write $V_k = f(S_k)$ for some function f. In this way we can regard S as the state process that is hidden and V as the observation process that can be observed.

A HMM is usually defined as 5-tuple (S, V, A, B, Π) where

$S=\{S_n, n=1,2,...S_N\}$ is a finite set of N hidden states.

$V=\{v_1, v_2, ...v_M\}$ is a finite set of M possible observed states.

$A=\{a_{ij}\}$ is transition probabilities of the state, where a_{ij} is the probability that the system goes from state S_i to state S_j.

$B=\{b_i(v_k)\}$ are the observation probabilities where $b_i(v_k)$ is the probability that the symbol v_k is emitted when the system is in state S_i

$\Pi=\{\pi_i\}$ are the initial probabilities that is π_i is the probability that the system starts in state S_t

4.2 Fundamental Problems in HMM

There are three main types of problems occurrence in the use of HMMs are

4.2.1 The Evaluation Problem

Given a model $\lambda=(A, B, \Pi)$ and an observation sequence $V=v_1, v_2, ..., v_T$ of length T, to compute the probability that the model generated the observation sequence $P[V|\lambda]$.

$P[V|\lambda]$ is given by $P[V|\lambda]=\sum_U [V|U,\lambda]P[U|\lambda]$

Where, $U=u_1, u_2, u_3 \ldots u_t$ is a fixed sequence

$P[V|U,\lambda]$ is the probability of observation sequence V for the state sequence U and $P[V|\lambda]$ is the probability of the sequence U for a given model.

Assume that observations are independent; the two probabilities are given by

$$P[V|U,\lambda]=\prod_{t=1}^{T} P[V_t|U_t,\lambda]$$

We obtain

$$P[V|\lambda]=\sum_U P[V|U,\lambda]P[U|\lambda]$$

4.2.2 The Decoding Problem

Given a model $\lambda=(A, B, \Pi)$, compute the most likely sequence of hidden states that could have generated a given observation sequence.

$U^*=\{u_1^*, u_2^*, u_3^* ... u_t^*\}$ for a given observation sequence $V=\{v_1, v_2, v_3 \ldots v_t\}$

Let the function $\arg\max_{x}(y)$, $\arg\max_{u}\{P[U \mid V,\lambda]\} = \arg\max_{u}[P[U,V \mid \lambda]$

4.2.3 The Learning Problem

Given a sequence of observations, find an optimal model. The learning problem is usually solved by the Baum-Welch algorithm.

$$\lambda^{*} = \arg\max_{\lambda}\{P[V \mid \lambda]\}$$

Sometimes this problem is to find very complex then we choose the model parameter P[V|λ] is locally maximized. This method is an iterative solution called Baum-welch algorithm.

4.3 The architecture of Profile Hidden Markov Models

PHMM was introduced by Krogh (1994). PHMMs is well suited to the popular 'Profile' methods for searching databases using multiple sequence alignments instead of single query sequences. It has three types of states; Match states that are represented by square labeled M, Insert states that are represented by diamonds labeled I and Delete states that are represented by circles labeled D.

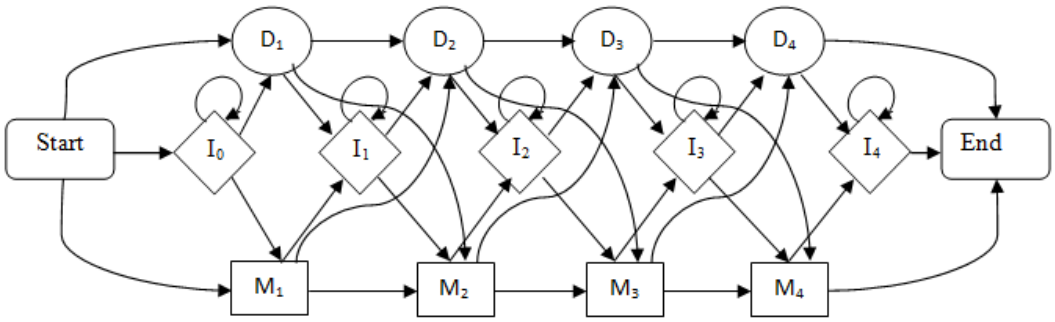

Figure 4.3.1: Architecture of Profile Hidden Markov Models

HMM is estimate the emission and transition probabilities for PHMMs, these parameters are obtained from multiple alignment sequences in a Protein, DNA, or RNA sequence family. PHMMs software is well suited for modeling a particular sequences family of interest and finding additional remote homologues in sequence data base [6]. Several available software packages are available for implementing PHMMs. There are two program packages to the academic community. HMMER developed by Sean Eddy and SAM developed by Krogh.

4.4 Parameter estimation

Maximum likelihood estimates of parameters of the PHMMs are

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}},\ \mathrm{e_k(x)} = \frac{\mathrm{E_k(x)}}{\sum_{x'} \mathrm{E_k(x')}}.$$

To implement Laplace's rule each should be increased by one. The updated counts

can be used directly to estimate the transition and emission probabilities a_{kl} and $e_k(x)$ of the PHMMs. The same estimates can be obtained using the updated equations.

$$a_{kl} = \frac{A_{kl} + 1}{\sum_{l'} A_{kl'} + M_k},\ e_k(x) = \frac{E_k(x) + 1}{\sum_{x'} E_k(x') + K'}$$

The updated equations cannot be affected the state diagram of PHMMs. The heuristic rule suggests that an alignment column with a fraction of gap symbols below 0.5 corresponds to a match state of a PHMMs. Otherwise, it corresponds to an insert state. Therefore we obtained this sequence match state M_1, M_2, and M_3 respectively. Then HMM includes a begin state B is an end state E, along with insert states I_0, I_1, I_2, I_3 and delete states D_1, D_2, D_3.

4.5 Viterbi algorithm

The Viterbi algorithm is used to find the optimal alignment of the sequence. Let $V_j^M(i)$ be the log-odds score of the highest scoring alignment of sub sequence x_1, $x_2 \ldots x_i$ with symbol x_i emitted by state M_j. Similarly $V_j^I(i)$ and $V_j^D(i)$ are the log-odds score of the highest scoring alignment of sub sequence x_1, $x_2 \ldots x_i$ with symbol x_i emitted by state I_j and D_j. The algorithm is presented as follows:

The initialization step : $V_0^m(0) = 0$, $V_0^I(0) = -\infty$, $V_0^D(0) = -\infty$

The update equations for the log-odds scores are as follows:

$$V_j^M(i) = \log \frac{e_{M_j}(x_i)}{q_{x_1}} + \max \begin{cases} V_{j-1}^M(i-1) + \log a_{M_{j-1}M_j} \\ V_{j-1}^I(i-1) + \log a_{I_{j-1}M_j} \\ V_{j-1}^D(i-1) + \log a_{D_{j-1}M_j} \end{cases}$$

$$V_j^I(i) = \log \frac{e_{I_j}(x_i)}{q_{x_1}} + \max \begin{cases} V_j^M(i-1) + \log a_{MjI_j} \\ V_{j-1}^I(i-1) + \log a_{IjI_j} \\ V_{j-1}^D(i-1) + \log a_{D_jI_j} \end{cases}$$

$$V_j^D(i) = \max \begin{cases} V_{j-1}^M(i) + \log a_{M_{j-1}D_j} \\ V_{j-1}^I(i) + \log a_{I_{j-1}D_j} \\ V_{j-1}^D(i) + \log a_{D_{j-1}D_j} \end{cases}$$

The termination step

$$V = \max \left\{ V_N^M(L) + \log a_{M_N\varepsilon}; V_N^I(L) + \log a_{I_N\varepsilon}; V_N^D(L) + \log a_{D_N\varepsilon} \right\}$$

Calculate the log-odds score V of the optimal path.

Any sequence can be represented by a path through the model. The probability of any sequence, given the model is computed by multiplying emission and transition

probabilities along the path. The probability of a sequence is used to calculate a score for the sequence. Score measures the probability that a sequence belongs to a given family, a high score implies that the sequence of interest is probably member of the class while a low score implies it is probably not a member.

5. Dataset

The dataset was used from gene bank. The family of Data set is Breast cancer for DNA sequence. To aligned the multiple sequence of this data using clustalW. The PHMMs gives the optimal path of this data.
The alignment of this data is given below,

Figure 5.1: An alignment of five breast cancer sequence. The consensus sequence below the alignment.

When identifying a new sequence as a breast cancer sequence, it would be available to concentrate on checking that these more conserved features are present. Then we have to build PHMMs for finding the optimal path of the sequence.
Estimate the parameters of PHMMs for the given alignment of breast cancer DNA sequence.

C A - C C

C A T C T

C A - - C

C C T C G

C C - G G

Figure 5.2: Five Columns From the Multiple Alignment of Five Breast Cancer Sequence.

PHMMs assigned directly position specific scores for each match state and gap penalty, for use in standard 'best match' dynamic programming. The Viterbi algorithm is used to find the optimal alignment of the sequence

Table 5.1: Estimate the Parameters of PHMMs for the Given Alignment

		0	1	2	3
Emission probabilities for match states	A	-	0.45	0.125	0.11
	C	-	0.33	0.5	0.33
	G	-	0.11	0.25	0.33
	T	-	0.11	0.25	0.23
Emission probabilities for insert states	A	0.25	0.17	0.25	0.25
	C	0.25	0.33	0.25	0.25
	G	0.25	0.17	0.25	0.25
	T	0.25	0.33	0.25	0.25

Now we apply the Viterbi algorithm to find the optimal alignment of the sequence AGGCT. This Viterbi algorithm is computed with the help of a JAVA program hmm.java. In that program, the input is given as a sequence of elements and formulated and the optimal solution is obtained.

Table 5.2: Estimate Optimal Alignment Using Viterbi Algorithm

	$X_{1=}A$	X_2=G	X_3=G	X_4=C	X_5=T
M_1	**-1.779**	-2.5918	-5.557	-3.5398	-9.3966
M_2	-2.5589	-4.6714	-4.4756	**-2.3727**	-6.7124
M_3	-6.2308	-4.0505	-4.8635	-8.2338	**-2.799**
I_0	-2.6772	-2.7414	-3.6269	-4.9119	-8.0636
I_1	-5.6527	**-2.4724**	**-2.3687**	-5.7471	-7.0339
I_2	-4.3507	-5.7146	-5.8343	-9.2049	-6.9099
I_3	-5.044	-5.7147	-6.5274	-9.8978	-7.0031
D_1	-3.7859	-3.850	-4.7337	-6.0205	-6.024
D_2	-3.8553	-4.6712	-7.636	-5.619	-11.4730
D_3	-5.2446	-6.0574	-9.0222	-7.0053	-12.859

The best path is highlighted.

6. Experimental results:

Apply the derived PHMMs into the 25 different types of gene sequence. The result which scores maximum probability will fit in the PHMMs. The PHMMs derived based on the cancer sequences. According to that, the PHMMs is only fit in the maximum score of the cancer sequence.

Table 6.1: Sequence Probabilities Fit in the PHMM

Cancer	Diabetes	Tuberculosis
0.98	0.063	0.318
0.855	0.137	0.363
0.898	0.132	0.245
0.987	0.063	0.318
0.78	0.037	0.373
0.855	0.122	0.473
0.998	0.027	0.445
0.77	0.012	0.373
0.828	0.063	0.345
0.757	0.167	0.498
0.763	0.143	0.243
0.955	0.045	0.333

Table 6.1 shows the probability values of the testing data. In fact, the cancer sequence has the maximum probability, it will fit in the PHMMs and the other sequence which has the less probability does not fit on derived PHMMs. These results show that the PHMMs succeeded to identify the cancer genotypes. To fit PHMMs using test data which consist of different types of disease sequences. The maximum probability value will only fit in the PHMMs.

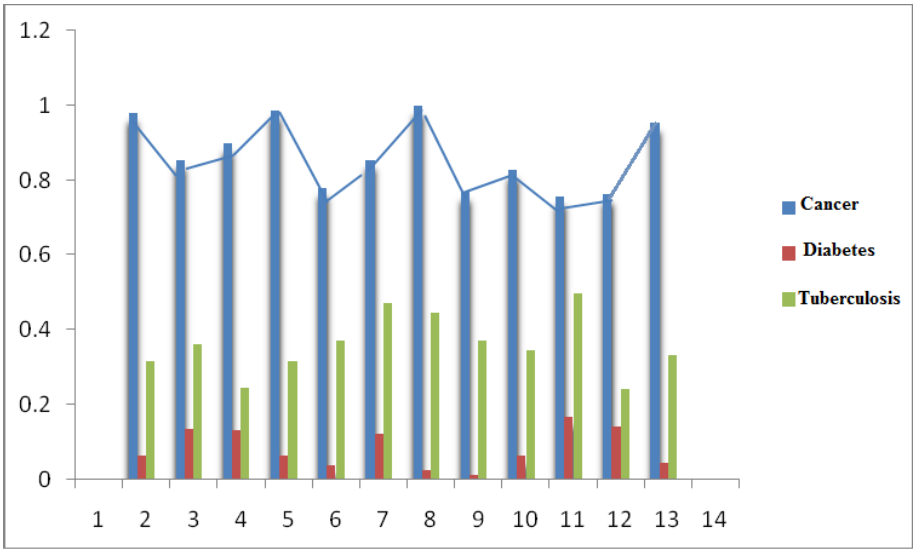

Figure 6.1: Relation Between Probability Values of Testing Data

7. Conclusion:

A sequence alignment is a key tool in bioinformatics. It is often necessary to deal with multiple sequences, including DNA and protein sequences. PHMMs are applied to the common tasks of simultaneously aligning multiple cancer sequences to each other, aligning a new sequence to an already –aligned family of sequences, and evaluating a new sequence for membership in a family of sequence. In this paper, how the PHMMs is used to identify the cancer genomic sequence is explained and essential role of PHMMs in the sequence alignment is elucidated. With the help of dataset, we achieved PHMMs is a appropriate tool to find an optimal path of cancer genomic sequence.

References

1. Durbin, R, Eddy, S, Krogh, A, Mitchison, G., 1998, "Biological Sequences Analysis: Probabilistic Models of Proteins and Nucleic Acid", Cambridge University Press, Cambridge.
2. Churchchill, G.A, 1989, Stochastic Models for Heterogeneous DNA sequences, Bull Math Biol, 51, 79-94.
3. Brown, M., Hughey, R., Krogh, A., Mian, I.S., SjSolander, K. and Haussler, D., 1993, Using Mixture Priors to Drive Hidden Markov Models for Protein Families", Proceedings of the ISMB-93, pp 47-53.
4. Asai, K., Hayamizu, S., and Handa, K., 1993, "Prediction of protein secondary structure by the Hidden Markov Model, 9(2), pp 141-146.
5. Baldi, P, Chauvin, Y., Hunkapiller, T., and McClure, M.A., 1994, "Hidden Markov Models of Biological Primary Sequence Information", Proc. Natl.Acsd.Sci. USA, 91(3), pp 1059-1063.
6. Eddy, S.R., 2004," What is Hidden Markov Model?" Nature Biotechnology, 22(10), pp 1315-1316.
7. SjSolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian I.S., and Haussler, D., "Dirichlet mixtures: A method for improving detection of weak but significant Protein Sequence Homology, Applic.Biosci, 12(4), pp 327-345.
8. Bateman,A., and Chothia, C., " Fibronectin III Domain in Yeast Detected by a Hidden Markov Model", Curr Biol, 6(2), pp 1544-1547.
9. Birney, E., Durbin, R, (1997), "Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison". ISMB proceedings.
10. Barrett, C., Hughey, R., and Karplus, K., 1997, "Scoring hidden markov models", CABIOS, 13(2), pp 191-199.
11. Dalgaard J.Z, Moser, M. J, Hughey, R., and Mian, S., 1997, "Statistical modeling, Phylogentic analysis and structure prediction of a protein splicing domain common to Inteins and Hedgehog Proteins", 4(2), pp 193-214.
12. Sonnhammer E.L.L., Eddy, S.R, and Durbin., R., 1997, "Pfam: A comprehencive Database of Protein Domain Families Based on Seed Alignments", PROTEINS: Structure, Function, and Genetics, 28, pp 405-420.

13. Di Francesco V., Garnier, J. and Munson, P.J, 1997, "Protein Topology Recognition from Secondary Structure Sequences: Application of the Hidden Markov Models to the Alpha Class Proteins", Journal of molecular biology, 267(2), pp 446-463.
14. Ahola V., Aittotallio, T., Uusipaikka, E., and Vihinen M., 2003, "Efficient estimation of emission probabilities in Profile Hidden Markov models", Bioinformatics, 19(18), pp 2359-2368.
15. Liang K.C, Wang, X, Anastassiou D, 2007, "Bayesian Basecalling for DNA Sequence Analysis Using Hidden Markov Models", Transactions on Computational Biology and Bioinformatics, 4(3), pp 430-440.
16. Madera, M., 2008, "Profile comparer: a program for scoring and aligning Profile Hidden Markov models, Bioinformatics, 24(22), pp 2630-2631.
17. Eddy, S.R., 1995, "Multiple Alignment Using Hidden Markov Models, Proc of Third Int. conf.on Intelligent Systems for Molecular Biology", 3, pp 114-120.
18. Di Francesco, V., Munson, P.J., and Garnier, J., 1999, "FORESST: fold recognition from secondary structure predictions of proteins", Bioinformatics, 15, pp 131-140.
19. Wistrand, M., and Sonnhammer, E.L., 2004, "Improving Profile HMM discrimination by adopting transition probabilities", Journal of molecular biology, 338, pp 847-854.
20. Wistrand M, and Sonnhammer E.L, 2005, "Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER", BMC Bioinformatics, 6(99).
21. Söding,.J., 2005, Protein homology detection by HMM–HMM comparison, Bioinformatics, 21(7), pp 951-960.