BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 9(7)

Artificial signal peptide prediction by a hidden markov model to improve protein secretion via Lactococcus lactis bacteria

Jafar Razmara*, Safaai B Deris, Rosli Bin Md Illias & Sepideh Parvizpour

Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia; Jafar Razmara – Email: jafar@utm.my; *Corresponding author

Received March 03, 2013; Accepted April 05, 2013; Published April 13, 2013

Abstract:
A hidden Markov model (HMM) has been utilized to predict and generate artificial secretory signal peptide sequences. The strength of signal peptides of proteins from different subcellular locations via Lactococcus lactis bacteria correlated with their HMM bit scores in the model. The results show that the HMM bit score +12 are determined as the threshold for discriminating secreteory signal sequences from the others. The model is used to generate artificial signal peptides with different bit scores for secretory proteins. The signal peptide with the maximum bit score strongly directs proteins secretion.

Keywords: Artificial signal peptide prediction, Protein secretion, Hidden markov model.

Background:

The proteins are generally directed to their subcellular destinations by different embedded signals, such as secretory signal peptides (SP), mitochondrial targeting signals (MTS), and nuclear localization signals (NLS). The majority of secreted proteins is secreted via the classic secretory pathway and is directed by secretory signal peptide sequences **[1].** The signal peptide is a short peptide chainthat binds at the N-terminus of the protein and directs translocation of the protein to certain organelles such as nucleus, endoplasmic reticulum or cell membrane. The sequence is normally started with a methionine and includes a positively charged n-region, a central characteristic hydrophobic region, and a c-region ending with a cleavage site. After translation of mRNA to the nascent protein, the signal sequence interacts with the signal recognition particle (SRP) for co-translational translocation. In the sequel, the complex binds to the SRP receptor **[2],** and signal peptide is then cleaved off from the mature protein at a specific site by signal peptidase **[3].** Finally, after undergoing further post-translational modification and folding process, the mature protein is secreted outside of the cell.

Secreted proteins play an important role in genome and have particular value in industrial and medical applications. They constitute about 10% of all proteins in genome **[4]** and represent collections of hormones, enzymes, immune-proteins, antibodies and many others. Therefore, predicting and identifying secreted proteins from their primary sequence is a major step in automated protein annotation. Several studies have been done within past decades to efficiently recognize and identify this class of proteins. The proposed methods commonly predict signal peptide sequence using machine learning techniques such as neural networks, support vector machines, and hidden Markov models **[5-7]**. The methods are used to identify targets of nascent proteins for secretion, export to the extracytoplasmic compartment, outer membrane, or tethering to the cell wall **[8].**

Hidden Markov models (HMM) as a superior to any formal statistical modeling technique provide a powerful method to profile protein domains based on a training dataset and have been widely used to model biological sequences. The key concept in this approach is a finite model that represents probability distributions over infinite numbers of possible sequences. The model is used to describe a series of observations by a hidden stochastic process which is called Markov process. An HMM constructed for proteins consists of a set of states, each with a characteristic amino acid distribution. The states are connected by transition probabilities which specify possible orders of states. The model is constructed on protein sequences with varying length and used to predict the most probable way for generation of a given sequence. Moreover, insertions and deletions of sequences are implemented as regular transition states, where it is traditionally a difficult problem in modeling. The most commonly used form of HMM in computational biology is profile HMM **[9, 10]**, with a structure based on profiles **[11]**. HMM is introduced in several literatures **[12, 13]**.

The knowledge of biological sequences is easily transformed to HMM in contrast to other machine learning techniques such as neural networks. In order to model signal peptides, HMM has the facility to design model according to three parts of these sequences. It also has the capability of extending the model by adding new modules. HMM has already used to predict and model signal peptides in various studies **[4, 13-15]** and has been proven to be the most successful technique in terms of accuracy **[16]**. Using a given sufficient dataset of signal peptides, the model can be trained to describe signal peptides, score an input sequence to predict a signal sequence, and generate artificial sequences at given bit scores.

The strength of signal peptide sequence plays an important role in secretion of a particular protein. Since the signal peptide is removed after secretion, replacing an original signal peptide with a strong artificial one does not change the structure of mature protein. Accordingly, it is highly interested to generate and replace a strong artificial signal peptide to improve protein secretion. In this study, we have developed a signal peptide model based on HMM for modeling and optimizing protein secretion via lactococcus lactis bacteria. To this end, a model was built on signal peptide sequences of a set of proteins synthesized within the bacteria and translocated to different organelles. It has been proven that the strength of secretory signal peptides depends on bit scores given by HMM, whereas the higher bit score represents the stronger secretory signal **[4]**. Thus, the model is used to evaluate bit scores of signal sequences of different organelles and determine thresholds of secretory signals. The underlying objective is to constitute an artificial strong signal peptide using the model and replace it instead of a native signal sequence of target protein in lactococcuslactis bacteria. This replacement causes to increase secretion of the target protein via the bacteria.

Methodology:

HMM Model for Signal Peptide

In order to make the model based on HMM, we used HMMER software version 2.3.2 developed by S. Eddy under Ubuntu operating system. HMMER is implemented for biological sequences analysis based on profile hidden Markov model **[9, 10]**. The HMMER software package includes a set of programs in order to manage the model. We used "hmmalign" program to train the model and "hmmemit" to emit sequences of the model of a required bit score. Therefore, the package is utilized to model and analyze signal sequences of the dataset and classify them based on the bit score.

One of scoring criteria provided by HMMER is bit score, which indicates the sequence is a better match to the profile model (positive score) or to the null model of non-homologous (negative score). The bit score is computed in log base 2 via the following formula:

$$S = \log_2 P(seq\,|\,HMM)/P(seq\,|\,null)$$

Where $P(seq\,|\,HMM)$ denotes the probability of the target sequence based on the model, and $P(seq\,|\,null)$ is the probability of the target sequence according to a "null hypothesis" model of the statistics of random sequence. The null hypothesis model in HMMER is considered as a simple one-state HMM that means that random sequences are constituted with a specific residue composition. Accordingly, a positive bit score indicates that the model gives good match for the target sequence than the null model.

Dataset

In order to collect training and test dataset, we used UniProt (http://www.ebi.ac.uk/uniprot/) from EMBL-EBI to extract proteins synthesized within lactococcuslactis bacteria. A set of 147 proteins is selected from the databases belong to different subcellular locations of lactococcus lactis bacteria. In this collection, effort was made to include only one representative from each group of closely homologous proteins. **Table 1 (see supplementary material)** shows a summary of the selected proteins from different organelles. Moreover, signal peptide database (http://www.signalpeptide.de/) was used to retrieve signal sequences of each protein in the dataset.

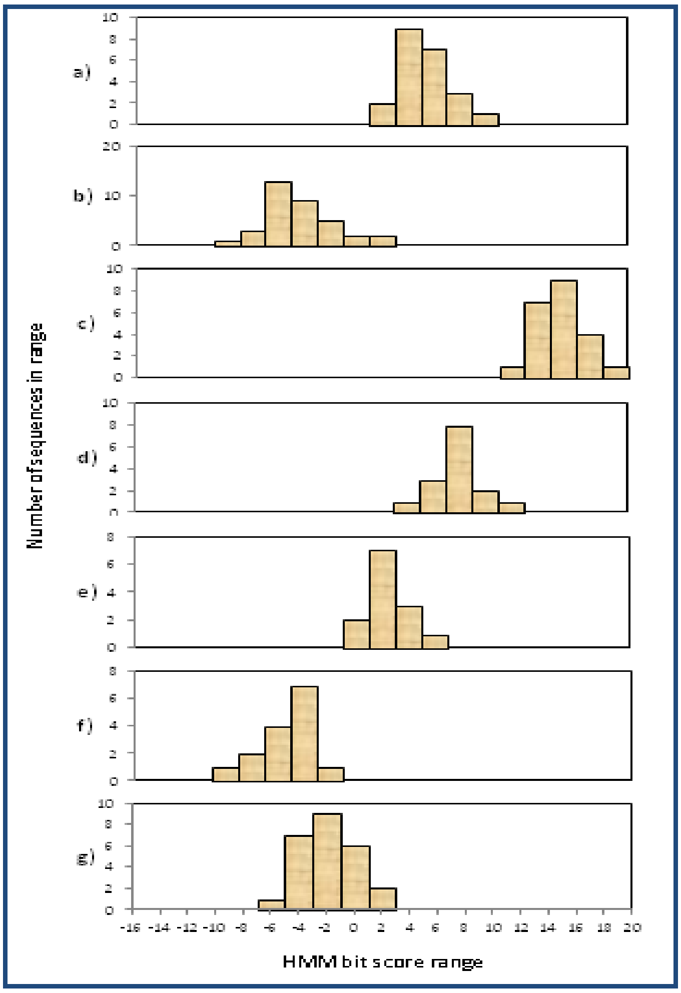

Figure 1: The bit scores histogram of signal sequences of the dataset for known subcellular locations: **a)** Cell membrane, **b)** Cytoplasmic proteins, **c)** Secretory proteins, **d)** Signal sequences, **e)** Transport proteins, **f)** Nuclear proteins, and **g)** Biosynthesis proteins.

Results & Discussion:

A hidden Markov model (HMM) was built in order to model training dataset. To this end, firstly, a multiple sequence alignment of signal peptides within the dataset was prepared using HMMER software 2.3.2. The output multiple alignment shows start methionines in the first column for all sequences, 141 out of 147 cleavage sites which are aligned in the same column, and all 147 hydrophobic cores. This resulting output demonstrates that HMM produces a high accurate model of signal peptide. Thus, the HMM model was created based on maximum likelihood parameter estimation. The background data for random model was prepared from the amino acid frequencies of the dataset.

In order to optimize the boundary of the sequences within the dataset, the resulting HMM model was re-estimated based on maximum discrimination method. Accordingly, any bias in the maximum likelihood model was removed due to repetition of the members of sub-families. This also partitions effectively sequences to the related organelles based on bit scores calculated by the method. **Figure 1** represents obtained bit scores distribution for different sets of signal sequences in the dataset. As seen in the figure, secretory proteins have been partitioned from the others at a threshold of bit score of 12.

According to the bit score boundaries obtained in the above experiment, an experiment was setup to create an artificial strong signal peptide. The "hmmemit" program of HMMER software package was used to emit sequences at HMM bit scores of +12, +14, +16, +18, +20, and +24. **Table 2 (see supplementary material)** represents the emitted sequences at different bit scores.

Conclusion:

The major objective in this study was to describe and model signal peptides by hidden Markov model to improve protein secretion via lactococcuslactis bacteria. The underlying purpose was to replace original signal peptides of secreted proteins with a strong one in protein production. Accordingly, we have developed a hidden Markov model (HMM) to analyze signal peptide sequences and generate a strong artificial signal sequence for the proteins synthesized within lactococcuslactis bacteria. Based on the correlation between strength of secretory proteins and their HMM bit scores, we experimented signal sequences of different organelles and analyzed their HMM bit score. The results shows that they are partitioned effectively according to the thresholds of HMM bit score. Therefore, a set of artificial signal sequences is generated at different bit scores for secretory proteins. In order to confirm the validity of the strength of generated artificial signal sequences, an in vitro activity has been devised by replacing original signal sequences by the computationally generated sequences. In conclusion, the study suggests a general idea to improve the strength of signal peptides to increase secretion of target proteins via bacterial cells.

Acknowledgement:

We would like to thank our research grant sponsors, Universiti Teknologi Malaysia (UTM) Research Management Center (RMC) for their support (research vote number: Q.J130000.2528.04H91).

References:

[1] Stroud RM & Walter P, *Curr Opin Struct Biol*. 1999 **9:** 754 [PMID: 10607673]
[2] Gilmore R *et al. J Cell Biol*. 1982 **95:** 470 [PMID: 6292236]
[3] Paetzel M *et al. Nature*. 1998 **396:** 186 [PMID: 9823901]
[4] Barash S *et al. Biochem Biophys Res Commun*. 2002 **294:** 835 [PMID: 12061783]
[5] Menne KML *et al. Bioinformatics*. 2000 **16**: 741 [PMID: 11099261]
[6] Nielsen H *et al. Protein Eng*. 1999 **12:** 3 [PMID: 10065704]
[7] Ladunga I, *Bioinformatics*. 1999 **15:** 1028 [PMID: 10745993]
[8] Song C *et al. Genomics Proteomics Bioinformatics*. 2009 **7**: 37 [PMID: 19591790]
[9] Krogh A *et al. J Mol Biol*. 1994 **235:** 1501 [PMID: 8107089]
[10] Eddy SR, *Curr Opin Struct Biol*. 1996 **6:** 361 [PMID: 8804822]
[11] Gribskov M *et al. Proc Natl Acad Sci USA*. 1987 **84:** 4355 [PMID: 3474607]
[12] Yoon BJ, *Curr Genomics*. 2009 **10**: 402 [PMID: 20190955]
[13] Oliver TF, *IEEE Trans on Information Technology in Biomedicine*. 2009 **13:** 740 [PMID: 19273034]
[14] Bagos PG *et al. Bioinformatics*. 2010 **26**: 2811 [PMID: 20847219]
[15] Kall L *et al. Nucleic Acids Res*. 2007 **35:** W429 [PMID: 17483518]
[16] Bendtsen JD *et al. J Mol Biol*. 2004 **340:** 783 [PMID: 15223320]

Edited by P Kangueane
Citation: **Razmara *et al.*** Bioinformation 9(7): 345-348 (2013)

Supplementary material:

Table 1: Summary of the training dataset and the sub cellular locations

Class name	Qty
Secretory proteins	22
Cytoplasmic proteins	35
Nuclear proteins	15
Cell membrane proteins	22
Transport proteins	13
Signal sequences	15
Biosynthesis proteins	25
Total	**147**

Table 2: HMM peptide sequences emitted at different bit scores

Bit score	HMM Peptide sequence
+24	MKFNKRRVAIWLLLIALLLIFVSFFFFTISTIQDNLFA
+20	MKFNKRIATFWLLLALIFVSFFFFTISSIQDNLQFNA
+18	MKFNKKRVAIATFIALIFVSFFTISSIQDNQTNA
+16	MKKIWNLAWLLLTLATLMGVSSSTAVVFA
+14	MQRKKKGLSILLAGTVVLGLLAVLPVGEIQAKA
+12	MKKILGFFWCSLGGLSATVHG