

APPLICATION OF PROFILE HIDDEN MARKOV MODEL-BASED
APPROACH USING HMMER3 ON MAJOR HISTOCOMPATIBILITY
COMPLEX (MHC) CLASS II BINDING PEPTIDES

MAZLINA ISMAIL

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2012

APPLICATION OF PROFILE HIDDEN MARKOV MODEL-BASED APPROACH
USING HMMER3 ON MAJOR HISTOCOMPATIBILITY COMPLEX (MHC)
CLASS II BINDING PEPTIDES

MAZLINA ISMAIL

(SGJ 100011)

SUBMITTED TO
INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF BIOINFORMATICS

2012

Contents

1	Introduction	1
1.1	Introduction	1
1.1.1	Overview	1
1.1.2	Objectives of the study	2
1.1.3	Organization of this report	3
2	Literature Review	4
3	Method	12
3.1	Obtaining dataset and preprocessing of dataset	12
3.2	Multiple sequence alignment and building of profile HMM	16
3.3	Evaluation of profile HMM using specificity study	16
4	Results	18
5	Discussion	32
6	Conclusion	34

List of Figures

2.1	Diagram depicting the humoral response of the immune system	5
2.2	Basic Markov model	8
2.3	A B cell HMM model with two hidden states, epitope and non-epitope . .	9
2.4	Profile hidden Markov model consisting of three states; match, insert and delete	10
4.1	A section of the multiple sequence alignment for DR1-od	19
4.2	A section of the multiple sequence alignment for DR1-90	20
4.3	A section of the multiple sequence alignment for DR4-od	21
4.4	A section of the multiple sequence alignment for DR4-90	22
4.5	Profile HMM and sequence logo of DR1-od	26
4.6	Profile HMM and sequence logo of DR1-90	27
4.7	Profile HMM and sequence logo of DR4-od	28
4.8	Profile HMM and sequence logo of DR4-90	29

List of Tables

3.1	Dataset of peptides with positive binding to MHC Class II obtained from MHCPEP.	13
3.2	Preprocessing of dataset	15
4.1	Summary of profile HMM for all seven datasets	24
4.2	Consensus sequence derived from the profile HMMs	25
4.3	Evaluation of profile HMM using hmmsearch	31

Acknowledgements

First and foremost, I would like to thank Dr. Saharuddin Mohamad for supervising my work this past semester. I would also like to thank Dr. Khang Tsung Fei for throwing some ideas in my way at the beginning stage of the research work; those nuggets of information helped me a long way down the road. Last but not least, to my family and friends for their never-ending support and encouragement.

Abstract

Major histocompatibility complex (MHC) Class II molecules play an important role in the immune system where it presents the antigenic peptides on the cell's surface for the next chain of immune response. Therefore, the process of identifying these peptidic fragments are of interest in the pipeline of vaccine design. Several probabilistic methods have been proposed for the prediction of MHC-binding peptides and one of them is the hidden Markov model. Profile hidden Markov models is a profiling technique which applies the statistical method to estimate the true frequency of a residue at a given position within the alignment. This study applies profile hidden Markov model-based approach using the HMMER3 package to train and build the profile based on the MHC-binding dataset. Evaluation of the profile HMM using a dummy dataset reveals that the profile HMM is able to differentiate the true positive data. This shows that the profile hidden Markov model approach can be a potential supporting method to identify peptides that bind to the MHC Class II.

Abstrak

“Major histocompatibility complex Class II” atau MHC Kelas II merupakan molekul yang memainkan peranan penting dalam sistem keimunan di mana ia memaparkan peptida antigenik di atas permukaan sel untuk tindakan keimunan yang seterusnya. Oleh itu, proses pengenalpastian peptida antigenik tersebut merupakan kaedah yang boleh disertakan dalam rekabentuk vaksin. Beberapa kaedah kebarangkalian telah dicadang untuk peramalan peptida yang akan mengikat pada kompleks MHC dan salah satu kaedah tersebut adalah “hidden Markov model”. “Profile hidden Markov model” merupakan kaedah pemprofilan yang menggunakan kaedah statistik untuk menganggar frekuensi sebenar sesuatu residu pada kedudukannya di dalam penjajaran. Kajian ini mengaplikasikan kaedah “profile hidden Markov model” menggunakan pakej HMMER3. Pakej ini digunakan untuk melatih dan membina profil berdasarkan data yang mengandungi peptida yang mengikat pada molekul MHC. Penilaian ke atas profil menggunakan set data menunjukkan bahawa profil tersebut dapat mengenalpasti set data yang “true positive”. Ini menunjukkan bahawa kaedah “profile hidden Markov model” boleh dijadikan sebagai salah satu kaedah untuk menyokong pengenalpastian peptida yang mengikat pada kompleks MHC kelas II.