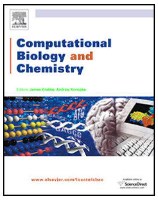
Computational Biology and Chemistry 61 (2016) 245–250

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem



A computational model for predicting fusion peptide of retroviruses

Sijia Wu[a], Jiuqiang Han[a,*], Ruiling Liu[a], Jun Liu[b], Hongqiang Lv[a]

[a] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
[b] School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China

ARTICLE INFO

Article history:
Received 30 June 2015
Received in revised form 27 January 2016
Accepted 10 February 2016
Available online 2 March 2016

Keyword:
Fusion peptide domain prediction
Hidden Markov Method
Similarity comparison

ABSTRACT

As a pivotal domain within envelope protein, fusion peptide (FP) plays a crucial role in pathogenicity and therapeutic intervention. Taken into account the limited FP annotations in NCBI database and absence of FP prediction software, it is urgent and desirable to develop a bioinformatics tool to predict new putative FPs (np-FPs) in retroviruses. In this work, a sequence-based FP model was proposed by combining Hidden Markov Method with similarity comparison. The classification accuracies are 91.97% and 92.31% corresponding to 10-fold and leave-one-out cross-validation. After scanning sequences without FP annotations, this model discovered 53,946 np-FPs. The statistical results on FPs or np-FPs reveal that FP is a conserved and hydrophobic domain. The FP software programmed for windows environment is available at https://sourceforge.net/projects/fptool/files/?source=navbar.

1. Introduction

Retroviruses are enveloped RNA-containing viruses including human endogenous retrovirus (HERV), human immunodeficiency virus (HIV), simian immunodeficiency virus (SIV), human T-cell lymphotropic virus (HTLV) and murine leukemia virus (MLV) (Coffin, 1992; Rosenberg, 2010). These viruses require membrane fusion to enter host cell cytoplasm for reverse transcription (Sieczkarski and Whittaker, 2004). The fusion process is controlled and executed by viral envelope glycoprotein (*env*) (White et al., 2008a). *Env* is composed of a surface and trans-membrane (TM) subunits respectively mediating receptor binding and virus-cell fusion (Barnard et al., 2006). Locating at the N-terminal of TM subunit, fusion peptide (FP) represents an absolute requirement for the fusogenic function of retroviruses (Apellániz et al., 2014). Upon fusion activation, FP must insert itself obliquely into target cell membrane to disorganize locally the structure of lipid bilayer (Epand, 2003). The interaction of FP with target cell causes a formation of an intermediate pre-hairpin structure which bridges and fuses the viral and host membranes together (Apellániz et al., 2014). In contrast with targeting later steps in the retrovirus life cycle, such as reverse transcription, integration and maturation of virions, targeting membrane fusion to block retrovirus into host cells has significant advantages for therapeutic intervention (Wolf-Georg et al., 2010). For example, it contributes to reducing the risk of undesired side effects, preventing the establishment or maintenance of latent viral reservoirs and so on (Wolf-Georg et al., 2010). Consequently, the dependence of retrovirus on FP during infection process and the advantages of fusion inhibitor make FP an available and promising drug target (Wolf-Georg et al., 2010; Münch et al., 2007). However, classical database search tools are inefficient to retrieve FPs in *env* sequences (Table 1) and annotated FPs in online database are not sufficient for a better understanding of FP. Thus, it is of great importance to propose a computational model to predict new putative FPs (np-FPs). This proposed FP model is helpful to accelerate identification of FPs in retroviruses by reducing the sequence dataset for biochemical experiment corroboration.

The main content of this thesis is arranged as follows. In Section 2, four procedures taken into account for the sequence-based FP model (Chou, 2011) will be sufficiently clarified. They are dataset construction, protein sample representation, FP prediction algorithm and proper evaluation methods. Then assessment results of FP model, FP software, predicted np-FPs and statistic hydrophobic properties about FP will be described in detail in Section 3. Eventually, validity of FP model, FP motif and evolutionary relationship about FP will be discussed in Section 4.

2. Material and methods

2.1. Datasets

All the retroviral protein sequences involved in this work were collected from NCBI database (Wheeler et al., 2006) and divided

[*] Corresponding author.
E-mail address: jqhan@mail.xjtu.edu.cn (J. Han).

http://dx.doi.org/10.1016/j.compbiolchem.2016.02.013
1476-9271/© 2016 Elsevier Ltd. All rights reserved.

Table 1
Comparison results between FP model and three classical database search tools.

Method	np-FPs with correct positions	np-FPs with wrong positions	np-FPs undetected	Percentage[a]
PSI-BLAST	16	50	51	13.68%
CS-BLAST	30	57	30	25.64%
HMMER3	39	47	31	33.33%
FP model	111	3	3	94.87%

[a] The percentage of np-FPs with correct positions.

into two datasets. One benchmark dataset contains *env* sequences with FP annotations so as to complete the establishment of FP prediction model, and the other dataset includes *env* sequences without FP annotations for predicting potential np-FPs.

In NCBI database, there are 124 *env* sequences with FP annotations related to HERV, HIV, SIV, HTLV and MLV. After looking over these data for reliability, 117 sequences (Table 2) meet following criteria are qualified to be included into benchmark dataset. The criteria stress not only *env* sequence to be non-repetitive but also FP annotation to be experimentally validated and non-suspicious. For each one of benchmark dataset, FP or non-FP domain was considered as positive or negative sample to train the prediction model.

Except for benchmark dataset, there are also a large amount of protein sequences downloaded from NCBI. They are *env* sequences without FP annotations (Table 2) relevant to the five retroviruses. These sequences were prepared for predicting more np-FPs with the proposed FP model.

2.2. Protein sample representation

Two straightforward sequential samples with mathematical expressions were formulated in this work. They can truly reflect the intrinsic correlations of predicted FP with inquired *env* sequence. One is the observation sequence *O*, which was expressed as

$$O = (o_1, \ldots, o_T) \qquad o_t \in \{1, \ldots, M\} t \in \{1, \ldots, T\}, \tag{1}$$

where o_t is *t*-th amino acid residue of protein *O*,*M* is the number of native amino acid types, and *T* is the length of inquired sequence. The other one is the state sequence *Q*, which was given by

$$Q = (q_1, \ldots, q_T) \qquad q_t \in \{1, \ldots, N\} t \in \{1, \ldots, T\}, \tag{2}$$

in which q_t is the state of *t*-th residue indicating FP ($q_t = 2$) or non-FP ($q_t = 1$), and *N* is the number of states.

2.3. FP prediction model

FP model predicts np-FP domain through two phases. Firstly, it adopted HMM method (Duda et al., 2001; Bonneville and Jin, 2013) to determine the existence and rough location of np-FP. Subsequently, it performed similarity comparison for a more precise np-FP. The prediction algorithm (Fig. 1) will be described in detail as follows.

2.3.1. HMM training

Three matrixes were defined and estimated to represent HMM model, which are *A*, *B* and Π. *A* stands for transition probability between states of FP and non-FP, *B* denotes emission probability of the residue under a state, and Π reflects the state distribution of initial residue. The elements of these matrixes were computed by Maximum Likelihood Estimate (Pfanzagl, 1994). According to FP and non-FP annotations in sequences of benchmark dataset, elements of *A* and *B* were respectively given by:

$$\begin{cases} a_{\mathrm{ij}} = P(q_{t+1} = j | q_t = i) = \dfrac{c(q_{t+1} = j, q_t = i)}{c(q_t = i)} \\ b_{\mathrm{jk}} = P(o_t = k | q_t = j) = \dfrac{c(o_t = k, q_t = j)}{c(q_t = j)} \end{cases} \quad i, j \in \{1, \ldots, N\}, k \in \{1, \ldots, M\}, \tag{3}$$

in which $c(x)$ is the occurrence number of event *x*. Locating in the upstream of *env* protein containing FP domain, the initial amino acid should be a residue with non-FP state. Thus, elements of Π were assumed to be:

$$\pi_i = \begin{cases} 1 if\ i = 1 \& t = 0 \\ 0 \text{ if } i = 2 \& t = 0 \end{cases}, \tag{4}$$

where $t = 0$ represents the moment before observation and $i = 1$ or $i = 2$ denotes non-FP or FP state respectively.

2.3.2. HMM decoding

Viterbi algorithm (Viterbi, 1967; David Forney, 2005) was applied to decode the most likely sequence of hidden states with trained HMM model λ. The state sequence *Q* was feasible to

Table 2
The datasets and results for FP modeling and prediction.

Groups	Train/test		Scan	
	Env number[a]	Result[d]	*Env* number[b]	np-FPs[c]
HERV	19		333	39
HIV	60	10-fold CV:	168,049	43,139
SIV	14	Acc = 91.97%, Se = 97.16%, Sp = 99.99%	18,381	9,908
HTLV	8	LOOCV:	1048	794
MLV	16	Acc = 92.31%, Se = 97.19%, Sp = 99.99%	107	66
Total	117		187,918	53,946

[a] The *env* sequences with FP annotations.
[b] The *env* sequences without FP annotations.
[c] The new putative FPs predicted by the model.
[d] The performance tested by two cross-validation methods.