Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2012, 4(12):5071-5075

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Development of python based software tool in predicting antigenicity of proteins

Polani B Ramesh Babu*, Krishnamoorthy P and Vamsi Krishna

Department of Bioinformatics, School of Bioengineering, Bharath University, Chennai, Tamil Nadu, India.

ABSTRACT

Peptide based vaccine designing and immunodiagnosis is the most important field in the diagnosis and therapy of various infectious and noninfectious diseases. It does critically require identification of regions in the pathogen native protein sequences, which are recognized by either B-cell or T-cell receptors. The antigenic regions of protein recognized by the binding sites of immunoglobulin molecules are called B-cell epitopes. The experimental identification of epitopes binding specifically to anti-peptide antibodies requires the binding assay of each peptide in an antigenic protein sequence which are very laborious and time consuming . A bioinformatics approach to predict linear B cell epitope in a protein sequence can be the best alternative to reduce the number of peptides to be synthesized for wet lab experimentation. The aim of this study is to develop a Python based software with graphical user interface for predicting the antigenic properties of protein. Hence the tool was named as Analysis of protein sequence and antigenicity prediction (ASAP). ASAP predicts the antigenicity of the protein sequence from its amino acid sequence, based on Chou Fasman turns and Antigenic index.

Key words : Epitopes, Vaccines, Antigenicity, bioinformatic tool.

INTRODUCTION

The B cell epitopes are exposed parts of pathogenic protein molecules, which are recognized by the antigen binding sites of antibodies or B cell receptors. They are consisting of atoms in a specific spatial arrangement. Protein B-cell epitopes are classified into linear and discontinuous epitopes and ~90% of epitopes in globular proteins are discontinuous [1]. Identified B cell epitopes are very useful because they can further be developed into diagnostics, therapeutics and vaccines [2]. Therefore, it's only natural that B cell epitope mapping has been a major field of immunology research. As identifying B cell epitopes experimentally is time-consuming and expensive, techniques to predict B cell epitopes have been developed for almost 30 years [3]. Most of these techniques are sliding window based sequence profiling methods. In brief, a window slides from the N-terminal to C-terminal of the query protein sequence.

Several linear B-cell epitopes in B-cell epitope databases fail to produce neutralizing antibodies (and hence fail to offer protective immunity). This has led to efforts to compile well-characterized datasets of protective linear B-cell epitopes, i.e., those that offer protective immunity [4]. A crucial step in designing of peptide vaccines involves the identification of B-cell epitopes. In past, numerous methods have been developed for predicting continuous B-cell epitopes, most of these methods are based on physico-chemical properties of amino acids [2,3]. Presently, it is difficult to say which residue property or method is better than the others because there is no independent evaluation or benchmarking of existing methods.

The inherent complexity of immune presentation and recognition processes complicates epitope prediction. Number of methods has been developed for predicting B cell epitopes, which are based on physico-chemical properties of the amino acids [5]. Classical methods of identifying potential linear B-cell epitopes from antigenic sequences typically rely on the use of amino acid propensity scales [6,7]. Several methods based on machine learning and statistical approaches have been recently proposed for predicting linear B-cell epitopes [8]. Hopps and Woods [9] used hydrophilic analysis (on twelve proteins) to investigate the possibility that at least some antigenic determinants might be associated with stretches of amino acids sequence that contain charged and polar residue and lack large hydrophobic residue. Parker et al, used the modified hydrophilic scale based on peptide retention times during high-performance liquid chromatography (HPLC) on a reversed-phase column [10]. A link between antigenicity and segmental mobility was developed for predicting mobility of protein segments on the basis of the known temperature B factors of the a-carbons of 31 proteins of known structure. They utilize the flexibility scale for predicting the B-cell epitopes. Methods were also developed for predicting epitopes based on surface accessibility of the amino acids [9,10]. Hopps and Parker derived their own scale of antigenicity based on frequency of residues in 169 experimentally known epitopes [9]. Pellequer et al., derived turn scales based on the occurrence of amino acids at each of the four positions of a turn using a structural database comprised of 87 proteins [7]. The turn scales correctly predicted 70% of the known epitopes.

The original hydrophilicity plotting procedure of Hopp and Woods remain most useful for locating the portions of protein sequences that are involved in interactions at the molecular surface [11]. The choice of an averaging group length or window is an important factor in determining any of the methods. a window of six residues performs optimally in locating protein antigenic sites [9]. Mapping B-cell epitopes plays an important role in vaccine design, immunodiagnostic tests, and antibody production. Because the experimental determination of B-cell epitopes is time-consuming and expensive, there is an urgent need for computational methods for reliable identification of putative B-cell epitopes from antigenic sequences [12].

A bioinformatics approach to predict linear B cell epitope in a protein sequence can be the best alternative to reduce the number of peptides to be synthesized for wet lab experimentation. In the past, numbers of computational methods and programs have been developed for predicting linear B-cell epitopes, which are based on hydrophilicity, accessibility, flexibility, or secondary structure propensities scales of the 20 natural amino acids [13]. In this study the performance of various residue properties commonly used in B-cell epitope prediction has been evaluated. The aim of the study is to develop a python based tool that will calculate the physiochemical properties of the protein from the primary sequence of the protein and predict the secondary elements of the protein using python, therefore identified antigenic sites in the protein could be used as a B-Cell epitope from the amino acid sequence.

EXPERIMENTAL SECTION

Python programme

Python is an easy to learn, powerful multi-paradigm programming language, which has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms. It permits several styles: object oriented and structured programming is fully supported, and there are a number of language features which support functional programming and aspect-oriented programming. It uses dynamic typing and a combination of reference counting and a cycle detecting garbage collector for memory management. An important feature of Python is dynamic name resolution (late binding), which binds method and variable names during program execution.

Molecular weight determination of proteins using Algorithm :

The molecular mass can be calculated as the sum of the individual isotopic masses of all the atoms in any molecule. The weight of a molecule is the sum of the weights of the atoms of which it is made. The unit of weight is the Dalton, one-twelfth the weight of an atom of ^{12}C. Thus the molecular weight (MW) of water is 18 Daltons. The masses for the amino acids in the protein are first summed, and then the mass of water is subtracted (for parameters see appendix).

Mw=∑ molecular weight of each amino acid – (n-1) molecular mass of water.

GRAVY refers to grand average of hydropathy. The hydropathy index of an amino acid is a number representing the hydrophobic or hydrophilic properties of its side-chain (for parameters see appendix).

$$\text{GRAVY} = \frac{\sum(\text{ hydropathy of individual amino acids})}{\text{Total number of residues}}$$

The first element of the program is a calling function, designed to count the numbers of copies of the amino acids which play a role in determining pI and to propose pH values for the other functions. The charge determination function determines the expected charge on the protein for a particular pH value. It, in turn, makes use of the charge ratio determination functions, which determine the expected proportion of charged and uncharged side chains for a particular amino acid, which are assessed from the supplied *pK values*.

Secondary structure prediction

The Chou-Fasman method of secondary structure prediction depends on assigning a set of prediction values to a residue and then applying a simple algorithm to the conformational parameters and positional frequencies. The Chou-Fasman algorithm is simple in principle, which is based on incidence of b turns. The calculation was based on a turn scale and there were three scales for describing turns instead of a single one. A window of seven residues is used for analyzing epitope region. The corresponding value of the scale was introduced for each of the seven residues and the arithmetical mean of the seven residue value is assigned to the fourth, (i+3), residue in the segment. Each property scale consists of 20 values assigned to each of the amino acid types on the basis of their relative propensity as described by the scale. In order to compare the profiles obtained by different methods, normalization of the various scales is done. We calculated the average of seven maximum and seven minimum values of a given physico-chemical scale and then calculated the difference between the two. The original values of the each scale are set between +3 to -3 by using the formulae

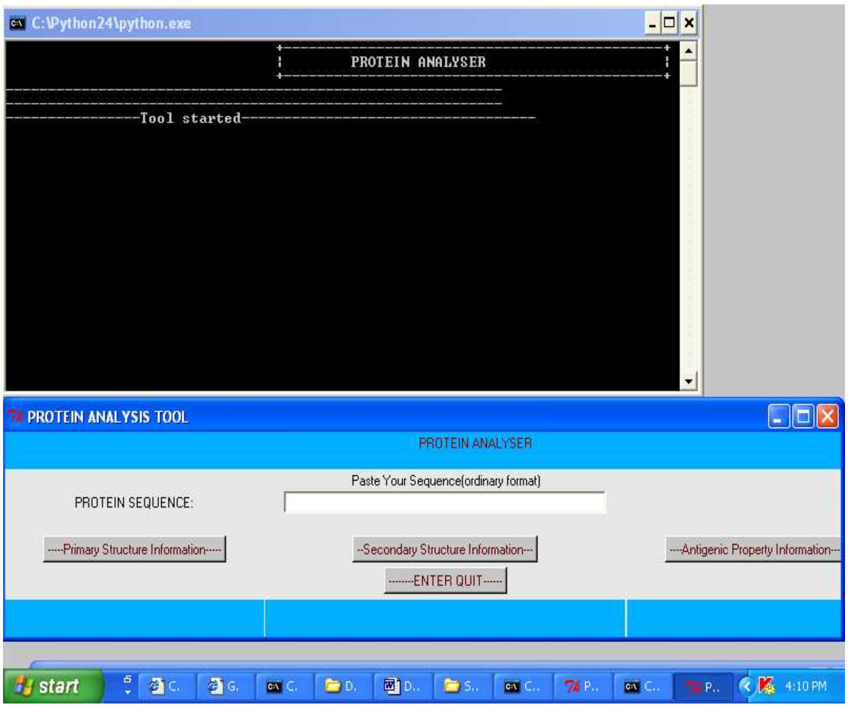
Normalization Score =AMS/DS * 6

Where AMS refer to Average of seven maximum/minimum values from the physico-chemical scale and DS refer to difference between the maximum and minimum score. Normalization score are set to +3 (Maximum) and -3 (Minimum) by subtracting or adding additional values.

RESULTS

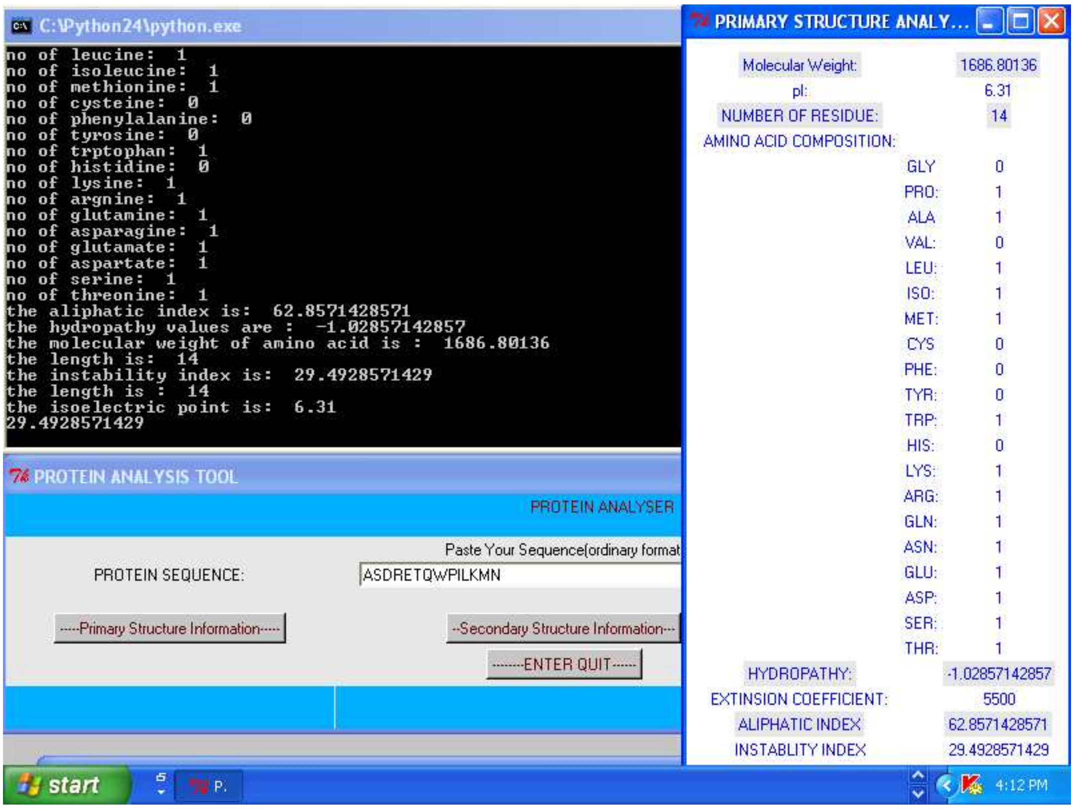
Python can be executed in two modes, interactive mode and command mode (Fig 1). The above snapshot reveals the command mode of python. Kyte and Dootile values were used to calculate the hydropathic index of the protein. The hydropathic index refers to the hydrophobicity nature of the protein. The program will assign the parameter value for each amino acid and calculates the GRAVY value of the protein. The above snapshot shows that the tool has been started and it is ready to execute all the commands in the backend. The Figure 1 shows the front end of the tool created using the kinker module, which is the one of the frequently used module in python for creating the graphical user interface.

In protein analysis tools raw protein sequence can be given as an input. The protein sequence should be in the ordinary format, which does not accept the Fasta format. Analysis of protein sequence and antigenicity prediction (ASAP) can calculate the physiochemical properties of the protein, turns prediction and antigenicity of the protein. (Fig 2). The protein sequence is pasted in the area provided by clicking the primary structure icon, the tool will calculate all the physiochemical properties of the protein and the output is displayed in the separate window. The properties include the total no of amino acids, their composition, molecular weight, isoelectric point, instability index, aliphatic index, extinction coefficient, GRAVY value. Further the program can be extended to determine the tertiary structure of the protein. As the protein sequence is submitted, the tool will predict the regions that are turns, which predicts the turns based on the Chou Fasman algorithm based on the propensity values of each amino acid. The propensity values are referred as the conformational parameter. Therefore, it is possible to predict the antigenic sites in the protein from its primary structure. There are several algorithms for identifying the antigenic sites, most of them uses the window size of seven. The regions with index value greater than one can be used as a B- cell epitope.


C:\Python24\python.exe
PROTEIN ANALYSIS TOOL
PROTEIN ANALYSER
Paste Your Sequence(ordinary format)
PROTEIN SEQUENCE:
-----Primary Structure Information-----
--Secondary Structure Information---
---Antigenic Property Information---
--------ENTER QUIT------
start
4:10 PM

```
                                     +--------------------------------------------+
                                     |           PROTEIN ANALYSER                 |
                                     +--------------------------------------------+
---------------------------------------------------------------------------------
---------------------------------------------------------------------------------
-----------------Tool started--------------------------------------------------
```

Figure 1 : Python Command Prompt


C:\Python24\python.exe
PROTEIN ANALYSIS TOOL
PROTEIN ANALYSER
Paste Your Sequence(ordinary format
-----Primary Structure Information-----
--Secondary Structure Information---
--------ENTER QUIT------
PRIMARY STRUCTURE ANALY...
start
4:12 PM

```
no of leucine:  1
no of isoleucine:  1
no of methionine:  1
no of cysteine:  0
no of phenylalanine:  0
no of tyrosine:  0
no of trptophan:  1
no of histidine:  0
no of lysine:  1
no of argnine:  1
no of glutamine:  1
no of asparagine:  1
no of glutamate:  1
no of aspartate:  1
no of serine:  1
no of threonine:  1
the aliphatic index is:  62.8571428571
the hydropathy values are :  -1.02857142857
the molecular weight of amino acid is :  1686.80136
the length is:  14
the instability index is:  29.4928571429
the length is :  14
the isoelectric point is:  6.31
29.4928571429
```

PROTEIN SEQUENCE: ASDRETQWPILKMN

Molecular Weight:	1686.80136
pI:	6.31
NUMBER OF RESIDUE:	14
AMINO ACID COMPOSITION:	
GLY	0
PRO:	1
ALA	1
VAL:	0
LEU:	1
ISO:	1
MET:	1
CYS	0
PHE:	0
TYR:	0
TRP:	1
HIS:	0
LYS:	1
ARG:	1
GLN:	1
ASN:	1
GLU:	1
ASP:	1
SER:	1
THR:	1
HYDROPATHY:	-1.02857142857
EXTINSION COEFFICIENT:	5500
ALIPHATIC INDEX	62.8571428571
INSTABLITY INDEX	29.4928571429

Figure 2 : Protein analysis data tool

DISCUSSION

Proteins are the major interest in drug designing as a target for variety of diseases. Hence, there is a need to analyze the various properties of protein. Hydropathy is a dominant force in protein folding. Complementaries in hydropathy is an important for several protein interaction [14]. Hydropathy index represent the hydrophobic or hydrophilic properties of the side chain. Larger the number is the more hydrophobic the amino acid. The most hydrophobic amino acids are isoleucine and valine [15]. It is necessary to measure the extinction co-efficient of proteins when it is purified [16]. Extinction co-efficient output in ASAP helped to calculate the absorbance of the protein. ASAP provides the pI value for the input sequence at a general pH. Most of the proteins are unstable in *in vitro* conditions. Hence before extracting a protein, the stability of the proteins must be known [17]. In ASAP a statistical procedure was applied to calculate the dipeptide instability weight which is an index for a predicting whether a particular protein would be instable or stable. In case of globular protein, thermostability is a important property [18,19,20]. Aliphatic index is directly proportional to thermostability factor [21]. In the present study a bioinformatics approach was to used to identify the antigenic sites in the protein primary structure that would reduce the experimental task. ASAP identifies the antigenic sites using three methods, hydrophilic, turns and antigenic index of the protein. To conclude, Python based ASAP will be a better proteomic tool to characterize a protein from the amino acid sequence. The tool has a user friendly interface with the convenient input format. It doesn't require fasta format and works in an offline mode. Hence, ASAP will act as good prediction tool for proteome analysis in identifying epitopes for immunogenicity in manufacturing vaccines.

Acknowledgements
The authors thank management Bharath University, Chennai. India for their support and encouragement.

REFERENCES

[1] Barlow,D.J., Edwards, M.S., and Thornton,J.M., Nature, **1986.** Vol 322, pp747-748.
[2] Meloen, R.H., Puijk, W.C., Langeveld, J.P., Langedijk, J.P. and Timmerman, P.,Menendez-Arias, L. and Rodriguez, R. *Comput Appl Biosci*, **1990.** 6(2):101-105.
[3] Kazim, A.L. and Atassi, M.Z. Bi*ochem J*, **1997**. 167(1):275-278.
[4] Sollner. J, R. Grohmann, R. Rapberger, P. Perco, A. Lukas, B. Mayer, and M. Blythe. *Immunome Res*: **2008.**1–17,
[5] Kolaskar, A.S. and Tongaonkar,P.C., FEBS, **1990**: Vol 276 , pp172-174.
[6] Odorico, M. and Pellequer, J.L. *J Mol Recognit*, **2003**: 16(1):20-22.
[7] Pellequer,J.L., Westhof,E. and Regenmortel, M.H.V. Methods in Enzymology, **1991**: 176-201.
[8] Saha. S and G. Raghava. *Proteins*, **2006.** 65:40–48.
[9] Hopp, T.P. and Woods, K.R. *Proc Natl Acad Sci U S A.* **1981.** 78(6):3824-3828.
[10] Parker, J.M., Guo, D. and Hodges, R.S. *Biochemistry*, **1986.** 25(19):5425-5432.
[11] Hopp, T.P. *Methods Enzymol*, **1989.** 178:571-585.
[12] Y. El-Manzalawy, D. Dobbs, and V. Honavar. *7th International Conference on Computational Systems Bioinformatics (CSB'08),* **2008.** pages 121–132.
[13] Pellequer. J, E. Westhof, and M. Van Regenmortel. *Immunol Lett,* **1993.** 36:83–99.
[14] Baranyi L, Campbell W, Ohshima K, Fujimoto S, Boros M, Okada H. *Nat. Med* 1, **1995**. pp. 894901.
[15] Dill K. "Dominant forces in protein folding". *Biochemistry* **1990**, 29, pp. 1337155.
[16] Lolkema J, Slotboom D. *FEMS Microbiol. Rev* **1998**. 22, pp. 305322.
[17] Goldberg,A.L. and St John.A.C. *Annu. Rev. Biochem.,* **1976.** 45, 747-803.
[18] Cenmidtal C. Mulyanto, Rosari Saleh. J Biophysical chemistry : **2011.** 2 (3) 258-267.
[19] Irini A Doytchinova[1] and Darren R Flower. BMC Bionformatics. **2007**: 8:4.
[20] Okuno,Y., Isegawa, Y., Sasao, F. and Ueda, dan S. J Virology. **1999**. 67, 2552-2558.
[21] Atsushi Ikai . *J Biochem* **1980***: 88 (6): 1895-1898.*