BIOINFORMATION
Discovery at the interface of physical and biological sciences

BIOMEDICAL INFORMATICS

Open access

www.bioinformation.net
Volume 12(1)

Prediction model

Classification of anti hepatitis peptides using Support Vector Machine with hybrid Ant Colony Optimization

Gunjan Mishra[1#], Vivek Ananth[1#], Kalpesh Shelke[2], Deepak Sehgal[1] & Jayaraman Valadi[1*]

[1]Shiv Nadar University, Gautam Budha Nagar, U.P 201314; [2]Centre for Modeling and Simulation, Pune University, Pune 4110001; Jayaraman Valadi - Email: valadi@gmail.com; *Corresponding author; #Equal contribution

Received January 04, 2016; Accepted January 06, 2016; Published January 31, 2016

Abstract:
Hepatitis is an emerging global threat to public health due to associated mortality, morbidity, cancer and HIV co-infection. Available diagnostics and therapeutics are inadequate to intercept the course and transmission of the disease. Antimicrobial peptides (AMP) are widely studied and broad-spectrum host defense peptides are investigated as a targeted anti-viral. Therefore, it is of interest to describe the supervised identification of anti-hepatitis peptides. We used a hybrid Support Vector Machine (SVM) with Ant Colony Optimization (ACO) algorithm for simultaneous classification and domain feature selection. The described model shows a 10 fold cross-validation accuracy of 94%. This is a reliable and a useful tool for the prediction and identification of hepatitis specific drug activity.

Keyword: Antiviral peptides, Support vector machine, Hepatitis, Ant colony optimization, Feature selection

Abbreviations: AMP = Antimicrobial peptides, SVM = Support Vector Machines, HIV = Human Immunodeficiency Virus, ACO = Ant Colony Optimization, AHP = Anti Hepatitis Peptides

Background:
The global pattern for hepatitis infection shows an annual rate of 500 million. World Health Organization (WHO) observes "World Hepatitis day" on July 28 each year for emphasizing the overall implications of Hepatitis A, E and B, C, D viruses, transmitted faeco-orally or parenteral, respectively. The reported clinical manifestation are: (1) fulminant cirrhosis, (2) organ failure and death in risk groups of patients receiving organ transplant and pregnant ladies (rate ~20-30%), (3) extra-hepatic manifestations, (4) zoonotic transmission, and (5) cancer. The present therapeutics is interferon, nucleoside analogue based and manifest adverse drug reaction thus emphasizing the need for novel treatment modalities **[1].**

Antimicrobial peptides (AMP) or 'Host defense peptides' facilitate innate immunity, bind to host protein, exhibit broad spectrum activity and are undergoing clinical trials. Comprehensive information on AMP along with pattern prediction using peptide sequence, structure and physio-biochemical attributes is reported in literature **[2, 3]**. While building a prediction classifier, hybrid filter-wrapper methods are advantageous for informative domain feature selection as faster selection of highly ranked feature of filter methods is augmented by wrapper methodology for accurate prediction. The method when amalgamated with evolutionary algorithm like ant colony optimization and classifier like support vector machines helps in early and improved convergence towards best informative subset **[4, 5]**. Therefore, it is of interest to describe the identification of anti-hepatitis peptides and study the sequence based traits responsible for anti-hepatitis activity. We have used hybrid filter wrapper approach employing SVM classifier and evolutionary ant colony optimization (ACO) method for obtaining improved subset of informative descriptors towards predicting the anti-hepatitis peptides.

Methodology:

Dataset

The experimentally validated antiviral peptides and relevant information were collected from PubMed, AMP databases and UniProt **[6]** for anti-hepatitis resources**.** We obtained 501 peptides after removal of redundancy having experimentally proven anti-hepatitis activity (positive dataset) and 404 peptides not known to have any anti-hepatitis activity (negative dataset) **[7].**

Attributes calculation

Sequence based descriptors inclusive of amino-acid, dipeptide, tripeptide, pseudo-amino acid composition (PAAC) [8], amphiphilic pseudo-amino acid composition (APAAC) **[9]** and compositional triad descriptors **[10]** of 905 sequences were calculated using the ProtR web server and R based code **[11].** The input to the Hybrid ACO-SVM based algorithm consisted of a total of 1838 descriptors.

Hybrid ACO-SVM algorithm

We employed SVM, a very effective algorithm based on statistical learning theory for the purpose of classification. SVM employs a maximum margin hyperplane to separate two different classes of sequences for linear classification. SVM takes the data to a higher dimensional feature space and subsequently employs a linear hyperplane for non-linear separations. The use of appropriate kernels enables all computations in the original space **[12]**. WEKA Information gain based filter ranking was first performed on the input dataset **[13]**. It is noted that 547 descriptors had quantitative information content value more than zero. Subsequently, these 547 descriptors were used for simultaneous classification and informative feature extraction was performed with ACO based wrapper – filter algorithm in synergistic combination with SVM algorithm.

ACO **[14]** is inspired by co-operative search behavior of real life ants. The pheromone mediated search is mimicked by software ants for solving real life optimization problems. The feature selection algorithm closely follows ACO methodology for solving Travelling Salesman problem of finding the shortest route **[15, 16]**. The features are equivalent to cities in feature selection algorithm. Here, the nodes are treated as features and the links connecting the nodes are initially deposited with some amount of pheromone.

The difference is that for feature selection the ants conduct a partial tour corresponding to the most informative subset. The hybrid algorithm employs pheromone as the learning capacity to find better tours. Additionally, information gain based feature ranking is used as domain information to enhance accuracy and speed of the algorithm. A software ant starts with a random initial feature. Selection of further features is based on exploration and exploitation. Exploitation means selecting the next feature with the maximum value of product of pheromone information gain score. Otherwise the feature was selected probabilistically as shown in Equation 1,

$$f = \begin{cases} max[\tau(f_{ij})\ \eta(f_j)] \\ \tau(f_{ij})\ \eta(f_j) \\ \Sigma\tau(f_{ij})\ \eta(f_j) \end{cases} \rightarrow (1)$$

where, $\tau(f_{ij})$ and $\eta(f_j)$ are pheromone concentration in the link connecting feature i and j and WEKA information gain score, respectively. Information gain score characterizes prior domain information and pheromone concentration reflects learning capabilities of ants to identify informative features. Thus, the ant proceeds by exploration and exploitation for selection of features. It completes the tour once the predefined number of features is selected. Similarly, a predefined number of ants complete their tours. The subsets selected by every ant are evaluated by SVM 10 fold cross validation accuracy. The pheromone values of the best subset links are increased while the values are decreased for other links. The algorithm is run for several such iterations and the best subset size is found.

Results & Discussion

We employed Hybrid SVM-ACO-information-gain algorithm to find the best subset of descriptors. After calculating the info-gain guided selection of the descriptors, we obtained 547 descriptors out of 1838 features for further analysis. Our methodology followed combination of ant colony optimization, infogain and support vector machines for feature selection and classification. **Table 1** summarizes the comparative analysis of the results for the hepatitis dataset.

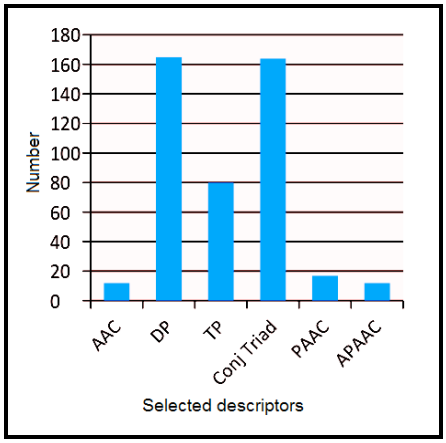

Figure 1: Plot of the respective descriptor prediction of the model.

The results show that our *Hybrid ACO-infogain algorithm* is quite effective for classifying the dataset for the hepatitis specific activity as compared with Infogain-SVM algorithm. We were successful to find the feature subset of size 450 that gave the best 10 fold CV accuracy of 94%. The best model hence obtained, with cross validation accuracy of 94% with 450 feature subset, will be helpful for the predictive modeling of the peptide specific activity analysis.

Table 1: Comparison of the results between SVM algorithms

No	Feature Subset	10 fold Cross-Validation Accuracy using SVM algorithms	
		Hybrid ACO-infogain	Infogain
1	500	93.3%	92.8%
2	450	**94.0%**	91.9%
3	400	93.5%	91.8%
4	350	93.0%	91.1%
5	300	92.9%	91.0%

We were also interested to note the predominant characteristics of the anti-hepatitis peptides that differentiate it from non anti-hepatitis peptides. The overall frequency was highest for the dipeptide and conjoint triad descriptors while there was a clear predilection for the acidic and aliphatic amino-acid residues **(Figure 1).** The selection of conjoint triad as favored descriptor has emphasized on the importance of the collective effect of hydrophobicity and polarity on the activity of anti-hepatitis peptides. The other selected Chou s' pseudo-amino acid composition descriptors are representative of the amino acid sequence with reference to its hydrophobicity and side-chain mass. The selection of the PAAC descriptors implied that the variation of the amino acid residues through composition and position have major role in the specific activity of the peptides as reported in the studies of Nanni *et al.* and Chang *et al.* **[17-18]**. Thus, we hypothesize that polarity, volume and hydrophobicity are the important features which differentiate active and inactive anti-hepatitis peptides and hence are crucial factors in designing new anti-hepatitis peptides.

Conclusion:

We performed supervised prediction of anti-hepatitis peptides employing a collection of experimentally validated positive (AHP) and negative sequences (non-AHP). Our methodology followed combination of ant colony optimization, infogain and support vector machines for feature selection and classification. Our algorithm was effective in classifying the anti-hepatitis peptides with 94% 10 fold cross-validation accuracy. Robust identification of anti Hepatitis peptides on improved representative features will not only aid in developing the disease stage specific treatment but will also lead to enhanced understanding of the characteristic of the genes and proteins, discovery of novel targets through the evolutionary pattern. This also helps in the improved understanding of the underlying mechanism of disease causation and pathogenicity empirically at level of host-pathogen interactions.

References:

[1] http://who.int/topics/hepatitis/en/
[2] Fox JL. *Nat Biotechnol*. 2013 31: 379 [PMID: 23657384]
[3] Guyon I *et al. J Mach Learn Res.* 2003 3: 1157
[4] Otero FE & Freitas AA, *Evol Comput.* 2015 1: 25 [PMID: 26066807]
[5] Tekin Erguzel T *et al. Comput Biol Med.* 2015 64: 127 [PMID: 26164033]
[6] www.uniprot.org/
[7] Li W & Godzik A, *Bioinformatics.* 2006 22: 1658 [PMID: 16731699]
[8] Chou KC *et al. Proteins: Struct. Funct. Genet*. 2001 43: 246
[9] Chou KC *et al. Bioinformatics* 2005 21: 10
[10] Shen J *et al. Proceedings of the National Academy of Sciences*, 2007 104: 4337
[11] Xiao N *et al. Bioinformatics* 2015 31: 1857 [PMID: 25619996]
[12] Chang C *et al. ACM Trans Intell Syst Technol*. 2011 2: 1
[13] Hall M *et al. ACM SIGKDD Explor Newsl*. 2009 11: 10
[14] Dorigo M *et al. Artif Life*. 1999 5: 137 [PMID: 10633574]
[15] Dorigo M & Gambardella LM. *Biosystems*. 1997 43: 73 [PMID: 9231906]
[16] Erguzel TT *et al. Psychiatry Investig*. 2014 11: 243 [PMID: 25110496]
[17] Nanni L *et al. J Theor Biol*. 2014 360: 109 [PMID: 25026218]
[18] Chang KY& Yang JR, *PLoS One*. 2013 8: e70166 [PMID: 23940542]

Edited by P. Kangueane

Citation: **Mishra *et al.*** Bioinformation 12(1): 12-14 (2016)