Identification of Defensins Employing Recurrence Quantification Analysis and Random Forest Classifiers

Shreyas Karnik[1,3], Ajay Prasad[1], Alok Diwevedi[1], V. Sundararajan[2,⋆], and V.K. Jayaraman[2,⋆]

[1] Chemical Engineering and Process Development Division, National Chemical Laboratory, Pune, India - 411008
[2] Center for Development of Advanced Computing, Pune University Campus, Pune, India - 411007
vsundar@cdac.in, jayaramanv@cdac.in
[3] School of Informatics, Indiana University, Indianapolis, IN, USA, 46202

Abstract. Defensins represent a class of antimicrobial peptides synthesized in the body acting against various microbes. In this paper we study defensins using a non-linear signal analysis method Recurrence Quantication Analysis (RQA). We used the descriptors calculated employing RQA for the classification of defensins with Random Forest Classifier.The RQA descriptors were able to capture patterns peculiar to defensins leading to an accuracy rate of 78.12% using 10-fold cross validation.

1 Introduction

Defensins are a class of antimicrobial peptides usually rich in cystine residues. They can be defined as a family of potent antibiotics synthesized within the body by neutrophils (a type of white blood cell) and macrophages (cells that can engulf foreign particles). Defensins act against bacteria, viruses and fungi by binding to their membranes and kill cells by forming voltage regulated polymeric channels in the susceptible cell's membrane[1]. The defensins are classified into alpha-defensins, beta-defensins and theta-defensins on the basis of their sequence homology. Defensins, owing to their small and potent antimicrobial effects can be used effectively for development of new clinically applicable antibiotics. They are also known to harbor anti-tumor activity, mutagen activity and/or behave as signaling molecules [1]. A few of their characteristics that have made them preferred candidates as peptide drugs are their short length, fast and efficient action against microbes and low toxicity to mammals [1]. Protein sequences can be visualized as linear heteropolymers that are formed by non-periodic arrangement of 20 different amino acids. In spite of the non-linear arrangement of amino acids in a protein, biologically stable proteins and functional are formed only with certain arrangements of amino acids. Literature sources indicate that some of the amino acids are preserved during the

⋆ Corresponding authors.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 152–157, 2009.

course of evolution, the conservation of these sequences points that the signals for a particular functionality of proteins are preserved in the sequence. Advanced techniques that can be applicable for investigating protein structure functional relationship like Fourier analysis, Wavelet analysis, chaos based methods can explore the underlying correlations within the protein sequences. Proteins can be viewed as a time series, where the order of amino acids indicates the role of time. There are examples of protein sequence analysis using signal processing based techniques [2]. Techniques like recurrence plots are used to visualize time series in a meaningful manner and give idea about the nature of the time series. Recurrence quantification analysis (RQA) has earlier been applied to study structure activity relationships in various proteins[3,4].and for other protein science and engineering applications providing interesting insights[5].

In this work we utilize the recurrence property to study the protein sequences that are characterized as defensins which play important role in innate immunity. Further, we employ Random Forests as the machine learning algorithm to classify sequence as defensin based on the RQA descriptors. Literature sources[3] indicate that hydrophobic patterns govern the native structure of proteins, also influencing the activity of proteins so we decided to use the Kyte-Doolittle Hydrophobicity scale for the conversion of the protein sequences into numerical time series to calculate RQA descriptors.

In the following sections brief description of RQA and Random Forest will be given along with the methodology followed by the results and discussion.

2 Methodology

2.1 Dataset

Literature on defensins was searched using search engines like Pubmed, iHop, Google Scholar and HubMed. Uniprot was also queried in order to get defensin sequences. A dataset of 238 non-redundant sequences which were annotated as defensins was complied. The defensin data set constituted the positive training examples, for the negative examples Uniprot was randomly sampled for sequences which are not annotated as defensins having length less than 150 amino acids containing cystine residues.

2.2 Recurrence Quantification Analysis (RQA)

The concept of Recurrence Plots was introduced by Eckmann et al.[6] Recurrence Plots are used as a tool that can enable the visualization of n dimensional phase space trajectory as a 2-D or 3-D plot of the recurrenences. The notion of recurrence is based on two facts viz. similar situations evolve in a similar way and some situations occur quite often. Thus, the recurrent points represent a state at position i which recurs at position j. Both the axes of the recurrence plots represent the time series that is under consideration. Recurrence plots are based on the recurrence matrix which is calculated as follows:

$$R_{i,j} = \Theta(\epsilon - ||\boldsymbol{x_i} - \boldsymbol{x_j}||) \quad i, j = 1 \cdots N \tag{1}$$

Where N is the number of states considered for analysis, ϵ is a threshold distance, $||\ ||$ is the norm, and Θ is the Heavyside function. For given point to be recurrent, it must fall within the threshold distance, ϵ. The recurrence plot inspection gives us the idea about correlation in the states at higher dimensions visually. The quantification of the recurrence plots [7] provides a number of parameters that describe small scale structures in the plot such as dots, diagonal and horizontal lines. Following parameters obtained by RQA quantification reveal the small scale structures in the recurrence matrix:

- *REC*: Quantification parameter *REC* is a count of single isolated points in the plot which signify that these states are rare, and do not fluctuate or persist consistently. These points do not necessarily imply noise in the data.
- *DET*: Certain points in the time series are parallel to each other i.e. they have same values but are placed at a different interval. This is expressed mathematically as: $R_{i+k,j+k} = 1$ (for k,$\cdots$,l; where l is the length of the stretch). These stretches can be visualized as diagonal lines in the recurrence matrix. These diagonal lines are called *deterministic*, because they represent a deterministic pattern in the series. *DET* is the count of all diagonal lines in a recurrence matrix. *DET*, thus, quantifies the deterministic local evolution of states in the series.
- *LAM*: Quantification parameter *LAM* represents the states which change or changes very slowly and as vertical lines in Recurrence Plot. Mathematically, this is $R_{i,j+k} = 1$ (for k $1\cdots$,v, where v is the length of the vertical line). A vertical line represents a state in which the series is trapped; and *LAM* is a count of all vertical lines in a matrix.

In addition to the above mentioned descriptors additional descriptors were suggested by Webber et. al [7] in order to describe the recurrence plots. These include *ENT* and *MAXLINE*, which are the Shannon's Entropy of the distribution of the diagonal lines and the maximum length of a diagonal line, respectively, and *TRAPT*, which is the average length of vertical lines. Complete list of the descriptors is given in Figure 1:

3 Classification Algorithms

Random Forest is an algorithm for classification and regression developed by Leo Breiman[8] that uses an ensemble of trees for classification and regression. Each of the classification trees is built using a bootstrap sample of the data, and at each split the candidate set of variables is a random subset of the variables (*mtry*). Thus, Random Forest uses both bagging (bootstrap aggregation), a successful approach for combining unstable learners, and random variable selection for tree building. Each tree is unpruned (grown fully). The trees grown then vote on the class for a given input, the class that gets the maximum number of votes is then assigned to the input. Random forest exhibits excellent performance in classification tasks. It has been successfully employed to solve various problems from the life sciences domain[9,10,11]. It has several characteristics that make it

Name of Descriptor	Description	Equation
Recurrence, *REC*	Represents percentage of points in Recurrence Matrix.	$REC = \frac{1}{N^2}\sum_{i,j=1}^{N} R_{i,j}$
Determinism, *DET*	The percentage of recurrence points that form diagonal lines.	$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i,j=1}^{N} R_{i,j}}$
Laminarity, *LAM*	The percentage of recurrence points that form vertical lines.	$LAM = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)}$
Entropy, *ENT*	Shannon's entropy of the distribution of the diagonal lines.	$ENT = -\sum_{l=l_{min}}^{N} p(l)\ln p(l)$
Trend, *TRND*	Paling of the recurrence plot towards the edges.	$TRND = \frac{\sum_{i=1}^{N-2}[i-(N-2)(RR_i-\langle RR_i\rangle)]}{\sum_{i=1}^{N-2}[i-(N-2)/2]^2}$
Trapping time, *TRAPT*	Average length of vertical lines.	$TRAPT = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=v_{min}}^{N} P(v)}$
Longest diagonal line, *LMAX*	The length of the longest diagonal line.	$LMAX = \max(\{l_i; i = 1...N_l\})$

Fig. 1. RQA Descriptors

ideal for these data sets. We used the Random Forest implementation in R[12] by Andy Liaw and Matthew Wiener[13] based on original code in FORTRAN by Breiman. Following methodology was adopted for classification:

1. All the protein sequences were converted to their numerical equivalents using Kyte-Doolittle Hydrophobicity scale.
2. Protein sequences were partitioned into train (80%) and test (20%) splits. The test split was used for the evaluation purpose.
3. Individual sequence which is now represented as a time series was subjected to RQA using parameters embedding dimension as 3, delay as 1, and radius as 6 using programs from Webber et al§.
4. RQA descriptors were calculated for all the protein sequences in the train and test splits of the data.
5. A model for classification was built using RQA descriptors and Random Forest algorithm(using best *mtry* parameter)on the training data and the model was evaluated using 10-fold cross validation to check the robustness. The performance of the model on test data was also evaluated.

[1] §http://homepages.luc.edu/∼cwebber/RQA131.EXE

Table 1. Summary of Performance on Test Set

Algorithm	Sensitivity	Specificity	MCC	Accuracy
Random Forest (with *mtry =3*)	0.736	0.812	0.588	79.16%

4 Results and Discussion

Performance of the model was evaluated on the basis of the cross validation accuracy and the performance on the test set in terms of Mathew's Correlation Coefficient(MCC) and other standard metrics used for evaluation of classification performance. The 10-fold cross validation accuracy of the model is 78.2%. The results on the test set are given in Table 1. Random Forest algorithm returns the variables that are most important in classification. We tried eliminating the features that have a low importance but it did not improve the classification performance thus only a list of ranking is presented. The order of importance of the RQA descriptors used for classification of defensins is as follows:

1. Recurrence *REC*
2. Trend *TRND*
3. Determinism *DET*
4. Laminarity *LAM*
5. Entropy *ENT*
6. L_{max}
7. Trapping Time *TRAPT*
8. V_{max}

Recurrence being the most important feature suggests that single isolated points representing rare states which do not fluctuate or persist consistently are important signals in that discriminate defensins. RQA descriptors such as *TRND*, *DET*, *LAM*, and *ENT* are also amongst the important features in terms of classification of defensins. In current study, classification of proteins as defensins based on features extracted from RQA could be achieved with 78.2% accuracy (cross validation accuracy)this suggests that RQA based on the representation of proteins in terms of their numeric equivalent (Kyte-Doolittle Hydrophobicity Scale in this work) captures the essential signals characteristic of defensins and can hence be used as an effective tool for exploring sequence function relationships and classification.

Acknowledgements. VKJ gratefully acknowledes financial assistance from Department of Science and Technology New Delhi, INDIA.

References

1. Ganz, T.: Defensins: antimicrobial peptides of vertebrates. Comptes Rendus Biologies 327(6), 539–549 (2004)
2. Giuliani, A., Benigni, R., Sirabella, P., Zbilut, J.P., Colosimo, A.: Nonlinear methods in the analysis of protein sequences: A case study in rubredoxins. Biophysics Journal 78(1), 136–149 (2000)

3. Zbilut, J.P., Giuliani, A., Webber, C.L.J., Colosimo, A.: Recurrence quantification analysis in structure-function relationships of proteins: an overview of a general methodology applied to the case of tem-1 beta-lactamase. Protein Eng. 11(2), 87–93 (1998)
4. Angadi, S., Kulkarni, A.: Nonlinear signal analysis to understand the dynamics of the protein sequences. The European Physical Journal - Special Topics 164(1), 141–155 (2008)
5. Mitra, J., Mundra, P.K., Kulkarni, B.D., Jayaraman, V.K.: Using recurrence quantification analysis descriptors for protein sequence classification with support vector machines. Journal of Biomolecular Structure and Dynamics 25(3), 141 (2007)
6. Eckmann, J.P., Kamphorst, S.O., Ruelle, D.: Recurrence plots of dynamical systems. EPL (Europhysics Letters) (9), 973 (1987)
7. Webber Jr., C.L., Zbilut, J.P.: Dynamical assessment of physiological systems and states using recurrence plot strategies. J. Appl. Physiol. 76(2), 965–973 (1994)
8. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
9. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(1), 3 (2006)
10. Hamby, S., Hirst, J.: Prediction of glycosylation sites using random forests. BMC Bioinformatics 9, 500 (2008)
11. Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E., Zhao, H.: Pathway analysis using random forests classification and regression. Bioinformatics (2006)
12. R Development Core Team: R: A Language and Environment for Statistical Computing. In: R. Foundation for Statistical Computing, Vienna, Austria (2009) ISBN 3-900051-07-0
13. Liaw, A., Wiener, M.: Classification and regression by randomforest. R. News 2(3), 18–22 (2002)