J. Indian Soc. Coastal agric. Res. 33(2): 69-72 (2015)

An Exploratory Statistical Analysis of Phytoplasma Protein Sequence Associated with Coconut (*Cocos nucifera*) Root Disease Isolated from Malabar Coastal Region of India

SANDIP SHIL* and KISHORE K. DAS[1]

ICAR-Central Plantation Crops Research Institute, Research Centre, Guwahati - 781 017, Assam
[1]Department of Statistics, Gauhati University, Guwahati- 781014, India

Received:27.02.2015 Accepted : 20.04.2015

Phytoplasma, the causal agent of coconut root disease (CRD) has been considered as a serious threat to coconut production in the Malabar coastal region of India. The disease is non lethal, but causes a considerable economic loss in India every year. It has already been reported that species of the phytoplasma belonging to ribosomal group 16SrXI, are primarily associated this disease. In present study, our objective is to explore more information that is already hidden within this phytoplasma protein sequence, using important statistical techniques. Therefore, we explored all important statistics with regard to the CRD phytoplasma protein sequence and further, predicted its secondary structure using Chou-Fasman algorithm. We obtained the value 8.3 of iso-electronic point of this protein, and also found that its predicted secondary structure consists of 5 helical regions, 6 beta sheet regions and 2 beta turn regions. A comparative analysis of the sequence with the sequences of phytoplasma related to other economically important crops revealed that there was no evidence of significant difference within the secondary structures of the selected sequences at 5% level of significance.

(***Keywords:*** *Chou-Fasman algorithm, Secondary structure, Coconut root disease, Protein sequence analysis)*

Phytoplasma, the causal agent of coconut root disease (CRD), has been considered as a serious threat to coconut (*Cocos nucifera L.*) production in the Malabar coastal region of India, especially southern state like Kerala (Ramjegathesh *et al.*, 2012). The disease is non lethal, but causes a considerable economic loss of approximately 968 million nuts, annually. Plants infected with these tend to exhibit various symptoms, which include phyllody (development of green leaf like structures, instead of flowers), flaccidity (bending of leaf-lets), virescence (development of green flowers), foliar yellowing, marginal necrosis of the older leaves, sterility of flowers, proliferation of axillary buds/shoots resulting bushy type or "witches' broom" appearance, abnormal elongation of internodes and generalized stunting. The phytoplasma associated with CRD is a comparatively new group of plant pathogens related to bacteria, and associated with numerous yellow-type diseases in hundreds of economically important plant species. Several studies already confirmed that phytoplasma is obligate parasites having diameters (less than 1 micrometer), non-spiroplasma, non-helical, filamentous shaped, pleomorphic in nature and cell-wall less prokaryotes, expected to be found in phloem of yellow-type diseased plants (Bertaccini and Duduk, 2010). They have small genomes ranging from 530 to 1350 kb (among the smallest known for any self-replicating organisms) and a low G+C content in their DNA sequence (23.0–29.5) %. Various molecular studies also disclosed that a specific species of phytoplasma, belonging to ribosomal group 16SrXI, are primarily associated with CRD (Sharmila *et al.*, 2004; Edwin and Mohankumar, 2007). However, other species of different ribosomal groups of phytoplasma, which has been reported to be associated with similar kinds of coconut and other related palm diseases across the globe, may also be associated. Some of them are lethal yellowing disease of palms in American countries, Caribbean region, New Guinea and Republic of Cuba Cape St Paul wilt of coconut palm in Ghana coconut yellow decline in Malaysia, Weligama coconut leaf wilt disease in Sri Lanka, date palm disease in North Africa and so on (Nipah *et al.*, 2007; Perera *et al.*, 2012; Myrie *et al.*, 2014). Although coconut yellowing disease symptoms in different names across the globe indicate that causal phytoplasma species are somehow related, but phylogenetic based studies (Harrison *et al.*, 2008) delineated that different 16Sr group phytoplasma may be associated. In a subsequent study in Sri Lanka disclosed that the association of phytoplasma, belonging to the

**Corresponding author*: E-mail: sandip.iasri@gmail.com

16SrXI '*Candidatus* Phytoplasma *oryzae'* group, is mainly associated with Weligama disease (Perera *et al.*, 2012), and the sequence (GenBank: EU635503) is almost identical with recent obtained phytoplasma sequence of CRD (GenBank: GQ850122). Recent molecular characterization and phylogenetic analysis of the sequence (GenBank: JX394030) based on less well-conserved *secA* gene, further validated that there is an association of 16SrXI group phytoplasma associated with CRD, also revealing position in plant kingdom and this phytoplasma has been identified as '*Candidatus* Phytoplasma *oryzae'* closely related strain, belonging to 16SrXI-B group (Manimekalai *et al.*, 2014). Therefore, the sequence (GenBank: JX394030) is latest representative nucleotide sequence of phytoplasma associated with CRD isolated from Malabar Coast of India (Manimekalai *et al.*, 2014). In present study, we select a less well-conserved *secA* gene based protein sequence of phytoplasma (GenBank: AFS50101), which is analogous to that nucleotide sequence (GenBank: JX394030). Our objective is to explore more information that is already hidden within this phytoplasma protein sequence, using important statistical techniques. We believe that our results may further serve as tools to predict protein structures, functions and also help to understand underlying biology about CRD.

MATERIALS AND METHODS

Data preparation

A homologous protein sequence set, of size 100, from the GenBank of National Centre for Biotechnology Information (NCBI) with the CRD phytoplasma protein sequence (GenBank: AFS50101) as a query sequence was collected using the standard nucleotide basic local alignment search tool (BLAST) (Altschul *et al.*, 1990). Out of which, we chose only 12 sequences of interests, which mostly include the protein sequences of phytoplasma related to economically various important crops such as coconut, napier grass, arecanut, bermuda grass, apricot, soybean, pepper and so on.

Computation of protein sequence statistics

We computed the physical properties of the protein sequence including percentages of amino acid residues within a sequence, percentage of residues of various physico-chemical classes (such as tiny, small, aliphatic, aromatic, non-polar, polar, charged, positive and negative), protein molecular weight (measured in Daltons) and, also the iso-electric point (Bastolla *et al.*, 2007).

Prediction of secondary structure using Chou-Fasman algorithm

Secondary structure refers the pattern of H-bonds of a protein (such as alpha helices, beta sheets and beta turns) that are observed in a primary sequence structure. However, this structure does not describe any specific identity of amino acids in primary sequence. But, this kind of structure often serves as an intermediate tool to predict protein tertiary structures or, functions. Here, we used an empirical secondary structure prediction method, which is popularly known as Chou–Fasman algorithm (Chou and Fasman, 1978). The method is very simple to understand, and predicts the secondary structure on the basis of a set of probability parameters. These parameters are usually derived from the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on known protein primary structures solved with X-ray crystallography and are used to predict the probability that a given sequence of amino acids would form a helix, a beta strand, or a beta turn in a protein.

All computational steps, including implementation of Chou-Fasman algorithm, were performed in statistical software R using "seqinr" package (Charif and Lobry, 2007; Team, 2014). The written programs would be available on request to corresponding author.

RESULTS AND DISCUSSIONS

In this present study, we computed important sequence statistics about that protein, as given in table 1. These statistics explored the physical properties of the specified protein sequence, which may provide extra useful information to do further biological experiments with this protein. For example, the value of iso-electronic point measure of this protein is 8.93. As we know, the iso-electronic point or iso-ionic point is the pH at which the amino acid does not migrate in an electric field or, amino acid is neutral. Therefore, this information may help to calculate the charge of the protein at a given pH.

Next, we predicted the secondary structure of the sequence using Chou-Fasman algorithm, as given in figure 1. The probability information of the parameters for each amino acid residue, derived in (Chou and Fasman, 1978), was incorporated in the algorithm to alpha helices, beta sheets, and turns. We found that the predicted secondary structure consists of 5 helical regions, 6 beta sheet regions and 2 beta turn regions. This information can further be used to predict the tertiary structure of this protein and define its functionality. We

Table 1. *Protein sequence statistics of CRD phytoplasma sequence*

Sequence Statistics			
Percentage of Alanine	6.92	Percentage of tiny residues	23.27
Percentage of Arginine	1.26	Percentage of small residues	38.99
Percentage of Aspartic acid	4.4	Percentage of aliphatic residues	21.38
Percentage of Asparagine	8.18	Percentage of aromatic residues	13.84
Percentage of Cysteine	5.66	Percentage of non-polar residues	49.69
Percentage of Glumatic acid	5.66	Percentage of polar residues	50.31
Percentage of Glutamine	3.14	Percentage of charged residues	30.82
Percentage of Glycine	11.32	Percentage of basic residues	18.24
Percentage of Histidine	11.95	Percentage of acidic residues	12.58
Percentage of Isoleucine	8.18	Iso-electronic point	8.93
Percentage of Leucine	1.89	Protein molecular weight (in Daltons)	18339.92
Percentage of Lysine	7.55		
Percentage of Methionine	1.89		
Percentage of Phenylalanine	2.52		
Percentage of Proline	3.14		
Percentage of Serine	3.77		
Percentage of Threonine	5.66		
Percentage of Tryptophan	1.89		
Percentage of Tyrosine	0.00		
Percentage of Valine	5.03		

Table 2. *A comparative analysis of CRD phytoplasma sequence with the sequences of phytoplasma related to other economically important crops*

Sl. No.	GenBank accession number	Sequence Description	Amino acid length	% identity with query	Number of predicted helical region	Number of predicted beta sheet region	Number of predicted beta turn region
1	AFS50101.1	Coconut root wilt phytoplasma	159	100	5	6	2
2	ABY48841.1	Napier grass stunt phytoplasma	161	89	3	7	3
3	ACY39971.1	Malaysia Bermuda grass white leaf phytoplasma	161	87	4	7	1
4	AFS50100.1	Arecanut yellow leaf disease phytoplasma	141	99	3	5	2
5	ACD10534.1	Coconut lethal yellowing phytoplasma	161	75	5	5	2
6	ABY48831.1	Cape St. Paul wilt disease phytoplasma	161	74	4	6	1
7	ABY48828.1	Coconut lethal yellowing phytoplasma	161	71	5	8	2
8	ABY48829.1	Coconut lethal yellowing phytoplasma	161	70	5	8	2
9	ABY48830.1	Tanzanian lethal decline phytoplasma	161	69	5	8	2
10	ABY48818.1	Soybean phyllody phytoplasma	161	70	3	5	3
11	ABY48843.1	Pepper Stolbur phytoplasma	161	69	3	5	3
12	ABY48815.1	Apricot chlorotic leaf roll phytoplasma	163	69	5	7	4

also performed a comparative analysis of CRD phytoplasma sequence with the sequences of phytoplasma related to other economically important crops, as given in table 2. All the selected sequences were highly homologous, which were determined based on E-values (range between 3E-91 and 4E-63). The amino acid residue lengths and percentages of identities with the query sequence also varied from (141 to 161) and (69% to 89%), respectively. We applied permutation t-test to test the respective mean difference within the regions of the selected sequences, and found that there were no significant difference with the respective mean value of

Primary structure of CRD phytoplasma sequence

"DEARTPLIISQIVKETKNLYKEAQRFVRTLKNSHYLIELETKTIELTEEGITKAE
NFFQIDNLYNIEHASLLHHIKNALKAAFTMHKDKDYLVDYKDGQVLIIDQFT
GRALPGRQFSDGLHQALEAKEGVLIKEETSIGATITYQNFFRLYHKLSGMTG
TAKT"

Predicted secondary structure of CRD phytoplasma sequence using Chou-Fasman algorithm

"---EEEEE--HHHHHHHHHHHHHHHHH--
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEEEE---
HHEEEEEHHHHHHHHHHHHHHHHHHHHHHEEEEEEEEEECCCCCCCC------
HHHHHHHHHHHHHHHHHHHHHHEEEEECCCCCEEEE----------"

where, H stands for alpha helix, E stands for beta sheet and C stands for beta turn

Fig. 1. Predicted secondary structure of CRD phytoplasma sequence using Chou-Fasman algorithm

the helical regions, beta sheet regions and beta turn regions at 5% as well as 1% level of significance. The calculated p-values were 0.981, 0.993 and 1, respectively. This means that there was no evidence of significant difference within the secondary structures of the selected sequences. However, there may be some difference among their physical properties.

From this study, we obtained important useful sequence statistics about phytoplasma associated with CRD protein sequence and also its secondary structure using Chou-Fasman algorithm. We also studied the difference between CRD phytoplasma sequence and the sequences of phytoplasma related to other economically important crops. We do believe that our analysis results may further serve as tools to predict protein structures, functions and/or, understand underlying more biology about CRD to biologists so that this disease can be controlled.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology* **215**(3): 403-410.

Bastolla, U., Porto, M., Roman, E., and Vendruscolo, M. (2007). *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*, Springer Science & Business Media.

Bertaccini, A. and Duduk, B. (2010). Phytoplasma and phytoplasma diseases: a review of recent research. *Phytopathologia mediterranea* **48**(3): 355-378.

Charif, D. and Lobry, J. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: *Structural approaches to sequence evolution: Molecules, networks, populations* , U. Bastolla, M. Porto, H. Roman and M. Vendruscolo, M. (eds.), Biological and Medical Physics, Biomedical Engineering, Springer Verlag, New York. pp 207-232.

Chou, P. Y. and Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annual review of biochemistry* **47**(1): 251-276.

Edwin, B. T. and Mohankumar, C. (2007). Kerala wilt disease phytoplasma: Phylogenetic analysis and identification of a vector,*Proutista moesta*. *Physiological and Molecular Plant Pathology* **71**(1): 41-47.

Harrison, N. A., Helmick, E. E., and Elliott, M. L. (2008). Lethal yellowing-type diseases of palms associated with phytoplasmas newly identified in Florida, USA. *Annals of Applied Biology* **153**(1): 85-94.

Manimekalai, R., Nair, S., and Thomas, G. V. (2014). Molecular characterization identifies 16srxi-b group phytoplasma (' *Candidatus* phytoplasma oryzae'-related strain) associated with root wilt disease of coconut in india. *Scientia Horticulturae* **165**: 288-294.

Myrie, W., Harrison, N., Douglas, L., Helmick, E., Gore-Francis, J., and Oropeza, C. and. McLaughlin, W. (2014). First report of lethal yellowing disease associated with subgroup 16SrIV-A phytoplasmas in Antigua, West Indies. *New Disease Reports* **29**(1): 12.

Nipah, J. O., Jones, P., and Dickinson, M. J. (2007). Detection of lethal yellowing phytoplasma in embryos from coconut palms infected with cape st paul wilt disease in ghana. *Plant Pathology* **56**: 777-784.

Perera, L., Meegahakumbura, M. K., Wijesekara, H. R. T., Fernando, W. B. S., and Dickinson, M. J. (2012). A phytoplasma is associated with the weligama coconut leaf wilt disease in Sri Lanka. *Journal of Plant Pathology* **94**(1): 205-209.

Ramjegathesh, R., Karthikeyan, G., Rajendran, L., Johnson, I., Raguchander, T., and Samiyappan, R. (2012). Root (wilt) disease of coconut palms in south asia–an overview. *Archives of Phytopathology and Plant Protection* **45**(20): 2485-2493.

Sharmila, L. B., Bhasker, S., Thelly, M. T., Edwin, B. T., and Mohankumar, C. (2004). Cloning and Sequencing of Phytoplasma Ribosomal DNA (rDNA) Associated with Kerala Wilt Disease of Coconut Palms. *Journal of Plant Biochemistry and Biotechnology* **13**: 1-5.

Team, R. C. (2014). R*: A language and environment for statistical computing*, R Foundation for Statistical Computing.