Computers in Biology and Medicine 54 (2014) 14–23

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

New layers in understanding and predicting α-linolenic acid content in plants using amino acid characteristics of omega-3 fatty acid desaturase ☆

Zahra Zinati [a], Fatemeh Zamansani [a], Amir Hossein KayvanJoo [b], Mahdi Ebrahimi [c], Mansour Ebrahimi [b], Esmaeil Ebrahimie [d,*], Manijeh Mohammadi Dehcheshmeh [e,**]

[a] Department of Crop Production and Plant Breeding, Faculty of Agriculture, Shiraz University, Shiraz, Iran
[b] Bioinformatics Research Group and Department of Biology, School of Basic Sciences, University of Qom, Qom, Iran
[c] Saarland University, Department of Informatics, Saarbrucken, Germany
[d] The University of Adelaide, School of Molecular and Biomedical Science, Adelaide, SA, Australia
[e] The University of Adelaide, School of Agriculture, Food, and Wine, Adelaide, SA, Australia

ARTICLE INFO

Article history:
Received 17 April 2014
Accepted 17 August 2014

Keywords:
α-Linolenic acid
Amino acids
Bioinformatics
Discriminant function
Feature selection
Machine learning: modelling
Omega-3
Prediction
Random Forest model

ABSTRACT

α-linolenic acid (ALA) is the most frequent omega-3 in plants. The content of ALA is highly variable, ranging from 0 to 1% in rice and corn to >50% in perilla and flax. ALA production is strongly correlated with the enzymatic activity of omega-3 fatty acid desaturase. To unravel the underlying mechanisms of omega-3 diversity, 895 protein features of omega-3 fatty acid desaturase were compared between plants with high and low omega-3. Attribute weighting showed that this enzyme in plants with high omega-3 content has higher amounts of Lys, Lys-Phe, and Pro-Asn but lower Aliphatic index, Gly-His, and Pro-Leu. The *Random Forest* model with *Accuracy* criterion when run on the dataset pre-filtered with *Info Gain* algorithm was the best model in distinguishing high omega-3 content based on the frequency of Lys–Lys in the structure of fatty acid desaturase. Interestingly, the discriminant function algorithm could predict the level of omega-3 only based on the six important selected attributes (out of 895 protein attributes) of fatty acid desaturase with 75% accuracy. We developed "Plant omega3 predictor" to predict the content of α-linolenic acid based on structural features of omega-3 fatty acid desaturase. The software calculates the 6 key structural protein features from imported Fasta sequence of omega-3 fatty acid desaturase or utilizes the imported features and predicts the ALA content using discriminant function formula. This work unravels an underpinning mechanism of omega-3 diversity via discovery of the key protein attributes in the structure of omega-3 desaturase offering a new approach to obtain higher omega-3 content.

1. Introduction

Omega-3s are polyunsaturated fatty acids that the body needs but cannot make them [1]. α-linolenic acid (ALA), a type of omega-3, has important roles in human health [1]. The use of ALA can reduce the risk of cardiac arrhythmias and sudden cardiac death [2]. ALA is synthesized via introducing a desaturation bond into linoleic acid (LA) by omega-3 fatty acid desaturase [3]. In plant species, ALA level is strongly correlated with the enzymatic activity of omega-3 fatty acid desaturase. This enzymatic reaction can take place in the membranes of both the plastids and microsomes [3,4].

The low level of omega-3 fatty acid desaturase activity in seeds of cereal has resulted in the low ALA content in these plants. It has been reported that ALA synthesis in seeds is mostly catalyzed by microsomal fatty acid desaturase [5]. Interestingly, introducing heterologous microsomal fatty acid desaturase genes under the control of the CaMV 35S and Ubi-1 promoters resulted in a high-

Abbreviations: ALA, α-linolenic acid; ANOVA, analysis of variance; DT, Decision Tree; FCdb, final cleaned dataset; LA, linoleic acid; PCA, principle component analysis; SVM, support vector machine.

☆Availability of software: The developed software in this study, "Plant omega3 predictor" is shared at the following link: https://drive.google.com/folderview?id=0B2Npj-saFbgeNXRiT3kxUXlOQXM&usp=sharing.

* Correspondence to: School of Molecular and Biomedical Science, The University of Adelaide, Adelaide 5005, North Tce, SA, Australia. Mobile: +449121357, Tel.: +61883132522.

** Correspondence to: School of Agriculture, Food, and Wine, Plant Research Centre, Waite Campus, The University of Adelaide, Adelaide 5064, SA, Australia. Mobile: +452241356, Tel.: +61883137224.

E-mail addresses: zahra.zinati@gmail.com (Z. Zinati), fatemeh.zamansani@gmail.com (F. Zamansani), a.h.kaivanjo@live.com (A. Hossein KayvanJoo), ebrahimi@mpi-inf.mpg.de (M. Ebrahimi), mansour@future.org (M. Ebrahimi), esmaeil.ebrahimie@adelaide.edu.au (E. Ebrahimie), manijeh.mohammadidehcheshmeh@adelaide.edu.au (M. Mohammadi Dehcheshmeh).

http://dx.doi.org/10.1016/j.compbiomed.2014.08.019
0010-4825/

level accumulation of ALA in transgenic rice seeds (up to 2.5 and 13-fold, respectively) [6,7]. Hua et al., showed that rice seeds overexpressing endoplasmic reticulum localized soybean-originated microsomal fatty acid desaturase and rice-originated microsomal fatty acid desaturase gained 23.8- and 27.9-fold higher ALA than that of non-transformants, respectively [8]. Human beings cannot convert LA into ALA because of the lack of omega-3 fatty acid desaturase [9]. Therefore, ALA must be obtained through the diet. The major dietary sources of ALA are restricted to some deep sea fishes and specific oilseed plants such as flax (*Linum usitatissimum*), soybean (*Glycine soja)*, rape (*Brassicca napus*), and perilla (*Perilla frutescens*). More recently, decreasing global fish supply and heavy metal pollution in sea water have limited the fish consumption [8]. There are few numbers of oilseed plants with high ALA content. However, high ALA content reduces the oxidative stability of oil and leads to rancidity in food products [10]. Therefore, in recent years, developing new varieties with decreased ALA content has been one of the goals in oilseed breeding [11,12]. Altogether, there is a considerable demand to develop non-oilseed plants with high ALA contents [8]. Indeed, oils with high polyunsaturated fatty acid have industrial applications including paints, ink carriers, and fuel [13,14]. Understanding the underlying molecular bases of omega-3 fatty acid desaturase activity in high ALA containing plants and subsequent engineering of this enzyme in non-oilseed plants are main pathways to achieve the high level of ALA.

Many attributes are required to illustrate different characteristics of a protein molecule. there is a high probability that some aspects of protein structure do not capture when a limited number of protein features is considered [15–17]. As a result, feature extraction to construct a proper and comprehensive feature vector of proteins is vital in success of any classifier. We recently demonstrated that increasing feature number and adding features such as dipeptides significantly contribute in achieving high prediction accuracy [18–20].

Structural and physicochemical features of proteins, such as amino acid compositions, dipeptide compositions, pseudo amino acid compositions, normalized Moreau-Broto autocorrelation, Moran autocorrelation, Geary autocorrelation are important class of employed features in this context [21,22]. In particular, 400 dipeptide features are good candidates which monitor protein structural alteration in amino acid level [15,19,20]. In our previous work on salinity, increasing the number of features to more than 800 resulted in higher accuracy of predictive models in discriminating halotolerant from halo-sensitive proteins [20]. Also, more robust training trough increasing the range and number of amino acid features provided the feasibility of predicting thermostability, irrespective of sequence similarity from any input protein sequence [19]. In line with this discussion, Das Roy and Dash showed that large and relevant feature set gives an opportunity to select the most relevant ones and develop more robust classifiers [23]. It has been discussed that for a successful prediction, a proper coding of proteins is necessary where using a full and proper protein feature set gives the best result [24].

Sting or repeats of amino acid/nucleotide is another type of features which has been used in “composition vectors based methods” for alignment-free classification/clustering strategy [25–27]. This approach counts the number of these repeats within different genomes or genes ranging from 1 to 7 (in maximum). Due to their limited numbers, this class of features is not able to compete with physic-chemical protein features in providing a comprehensive view on protein structure.

To prevent losing the information of sequence-order, pseudo amino acid composition (PseAAC) has been introduced [28–30] and extensively used in the study of protein structure [31–34]. The exponential increase in the number of protein attributes highlights the need for more advanced analytical tools compared to the commonly used multivariate based methods.

Fuzzy logic, neural networks, decision trees, and support vector machine (SVM) are the most effective modeling techniques within the available mathematical models [16,19,35,36]. Numerous studies have considered modeling and prediction of food characteristics [35]. As example, Yalçin and Tasdemir developed a fuzzy expert system for prediction of ALA content of eggs from hens fed with flaxseed [37]. They achieved a high correlation between experimental values and expert system based prediction [37]. Rheological features were also used to model fatty acid composition of 7 vegetable oils (hazelnut, soybean, sunflower, olive, canola, corn, and cotton seed) using neuro fuzzy inference and neural network [38]. The effects of natural antioxidant compounds on hazelnut oil oxidation were also modeled by neuro-fuzzy inference system and neural network [35]. Fuzzy logic has been also used to model the effects of environmental treatments on food [39]. Decision trees are interesting models as a Decision Tree looks for regularities in data, determines the features to add at the next level of the tree to minimize the entropy impurity [40,41]. Decision Tree is of great attention in the recent prediction investigations since it presents the hierarchical ranking of important features and provides a clear image of protein structure variation [36,42]. Recently, Decision Tree models have been used to identify the key physiological and agronomic traits as well as maker discovery in plants [36,43,44]. It should be noted that the combination of attribute weighting algorithms with the above mentioned predictive models has the potential to increase the accuracy of prediction and decrease the complexity of analysis.

Possible link between ALA content and amino acid characteristics of omega-3 fatty acid desaturase has not been investigated yet. In this study, using amino acid attributes, we compared the molecular structure of omega-3 fatty acid desaturase between plants with low and high ALA. To this end, we made a big dataset of protein attributes of omega-3 fatty acid desaturase proteins in plants with high and low ALA content by the help of computational biology. Then, attribute weighting (feature selection) algorithms, predictive Decision Tree models as well as statistical univariate and multivariate were used to find the key protein features of omega-3 fatty acid desaturase (contributing to its function). This provided statistical models for prediction of enzyme performance based on the structural amino acid attributes. We developed a software, Plant omega3 (α-linolenic acid), to predict high and low content of α-linolenic acid based on the sequence-based structural protein features of omega-3 fatty acid desaturase. These results would be advantageous for engineering novel omega-3 fatty acid desaturase in future investigations.

2. Material and methods

The following steps are required for developing a reliable predictor for a protein system as well as finding the key distinguishing the features which significantly contribute in function/phenotype: (1) making a big dataset of protein attributes for training and testing, (2) application of different attribute weighting (features election) models to unravel the most important attributes which have correlation with target variable (label), (3) fitting a reliable mathematical formula (algorithm/model) for prediction, (4) measuring the prediction accuracy to evaluate the strength of the developed predictive system, (5) providing a UIRL or a software based on the important discovered features to provide a predictive application [45].

2.1. Data collection

From ExPASY (http://www.expasy.org/) and NCBI (http://www.ncbi.nlm.nih.gov/) databases, amino acid and coding sequences of