Chemometrics and Intelligent Laboratory Systems 145 (2015) 7–16

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A quantitative sequence–aggregation relationship predictor applied as identification of self-assembled hexapeptides

Chen Chen [a], Yonglan Liu [a], Jin Zhang [a], Mingzhen Zhang [b], Jie Zheng [b], Yong Teng [c,*], Guizhao Liang [a,*]

[a] Key Laboratory of Biorheological Science and Technology, Ministry of Education, School of Bioengineering, Chongqing University, Chongqing 400044, PR China
[b] Department of Chemical and Biomolecular Engineering, The University of Akron, Akron, OH 44325, USA
[c] School of Life Sciences, Chongqing University, Chongqing 400044, PR China

ARTICLE INFO

Article history:
Received 30 December 2014
Received in revised form 3 April 2015
Accepted 12 April 2015
Available online 24 April 2015

Keywords:
Quantitative sequence–aggregation relationship (QSAR)
Hexapeptide
Factor analysis scale of generalized amino acid information (FASGAI)
Supporting vector machine (SVM)
Aggregation

ABSTRACT

It is essential to predict aggregation-forming sequences for elucidation of protein misfolding mechanisms and the design of effective antiamyloid inhibitors. In this work, we predict and characterize self-assembled hexapeptides by a quantitative sequence–aggregation relationship (QSAR) model, which involves characterization of factor analysis scale of generalized amino acid information (FASGAI) and modeling of supporting vector machine (SVM) with radial basis function kernel. The QSAR model achieves maximum accuracy of 78.33% and area under the receiver operating characteristic curve of 0.83 with leave one out cross-validation on 180 training hexapeptides. We determine "hotspots" and key factors that largely contribute to the self-assembly of these hexapeptides by analyzing their sequence–aggregation relationships. We also explore the applications of the present model, e.g., the first is to identify the aggregation-forming sequences within both β-amyloid peptide ($A\beta_{42}$) and human islet amyloid polypeptide (hIAPP) using a 6-residue slide window, and acquire good agreement with previous experimental observations, the second is to perform in silico design of potential aggregation-forming hexapeptides which are validated by all-atom molecular dynamics simulation and density functional theory calculations, and the third is to predict the potential self-assembled tri-, tetra- and pentapeptides, in which hydrophobic amino acids such as isoleucine, leucine, valine, phenylalanine, and methionine occur at higher frequencies. The present QSAR model is helpful for (i) predicting self-assembled behaviors of peptides, (ii) scanning and identifying aggregation-forming sequences within proteins, (iii) understanding action mechanisms of peptide/protein aggregation, and (iv) designing potential self-assembled sequences applied as drug discovery and nano-materials.

1. Introduction

In organisms, proteins will implement fold processing and maturation after being coded by ribosomes, which is a process to search for conformational stability and execute different biological functions [1]. Unfortunately, protein misfolding may not only lead to the existence of abnormal conformation [2], but also bring out the occurrence of the unfolding regions of proteins, which induces further protein aggregation [1]. So far, there are more than 20 kinds of illnesses related to protein misfolding such as Alzheimer's disease, Parkinson's disease and Type II diabetes [1,3,4]. Their common pathogenic characteristic is that the protein secondary structures changed from random coil and α-helix to β-folded or self-assembled amyloid deposition [5,6].

There has been a great deal of controversy about the pathogenic mechanisms and theories of abnormal accumulation of proteins [7]. One of the most persuasive ideas is that the polypeptide fragments in sensitive areas have caused protein aggregation [8]. Investigations on aggregation characteristics of two representative proteins in Alzheimer's disease, β-amyloid peptide (Aβ) and Tau protein, provide good interpretations of the mechanism above [9]. It has been reported the 15–21 aggregation-forming segment (QKLVFFA) in Aβ with 42 amino acids ($A\beta_{42}$) could induce $A\beta_{42}$ form amyloid fibrils [10–12], and the VQIVYK sequence in Tau protein could bring out abnormal protein aggregation [13]. Therefore, identification of aggregation-forming fragments is helpful for elucidating mechanisms of protein aggregation and designing antiaggregation inhibitors [14].

As we know, X-ray scattering, nuclear magnetic resonance or electron microscopy [15] are common experimental approaches used to observe the structural features of proteins. However, separation and purification of proteins commonly requires complicated parameters, and is a very time, resource, and found consuming process [16]. Experimental testing of all possible amino acid combinations is currently not feasible, which makes computational methods for predicting structures and functions of proteins very attractive [11]. Currently, there is an increasing number of noteworthy methods for predicting protein functions from sequence and structural data [11]. These methods can be generally classified

* Corresponding authors. Tel./fax: +86 23 65112677.
E-mail addresses: usyteng@163.com (Y. Teng), gzliang@cqu.edu.cn (G. Liang).

http://dx.doi.org/10.1016/j.chemolab.2015.04.009
0169-7439/

into sequence-based and structure-based predictions [11,17]. Several computational methods have been proposed to predict propensity for aggregation-forming fragments. Tompson et al. proposed a 3D profile method to identify fibril-forming segments in proteins [18]. Fernandez-Escamilla et al. used a statistical mechanics algorithm, TANGO, based on the physicochemical principles of β-sheet formation, extended by the assumption that the core regions of an aggregate are fully buried, to predict protein aggregation [19]. Garbuzynskiy et al. introduced two characteristics involving the expected probability of hydrogen bond formation and expected packing density of residues to detect amyloidogenic regions in protein sequences [20]. It should be mentioned that molecular dynamics (MD) simulations [21,22] and density functional theory (DFT) [23,24] calculations can yield valuable information about the structural changes that arise at the atomic level upon the formation of aggregation-forming fragments, while such information is difficult to be experimentally obtained. Prediction of protein functions from their structures is usually undertaken when sequence-based methods have failed [17]. The aggregating behaviors of proteins/peptides are strongly determined by the intrinsic properties of their amino acid sequences [25]. Therefore, it is possible to make an accurate prediction about whether proteins/peptides will aggregate from the knowledge of their sequences [26].

Our aim was to establish a quantitative sequence–aggregation relationship (QSAR) model to predict the self-assembled characteristics of hexapeptides. We explored the aggregation-driving forces as well as hotspots and key factors with major contribution to peptide aggregation. The predictive model was then applied as (i) identification of aggregation-forming fragments within $A\beta_{42}$ and human islet amyloid polypeptide (hIAPP) evaluated by comparison with experimental observations, (ii) design of aggregation-forming fragments validated by MD simulations and DFT calculations, and (iii) prediction of aggregation-forming tri-, tetra-, and pentapeptides. This work is beneficial to identifying aggregation-forming sequences, understanding mechanisms of protein/peptide misfolding, and designing potentially effective antiamyloid inhibitors or building blocks applied as nanomaterials.

2. Principles and methods

2.1. Data set

The training set was constituted by two datasets. The first dataset was derived from the AmylHex dataset containing 158 hexapeptides [18] (67 aggregation-forming and 91 non-aggregation-forming samples). The second dataset was derived from our own 22 hexapeptides [14]. We built and applied a QSAR model based on Index of Natural and Non-natural Amino Acids (NNAAIndex) to design about 8000 hexapeptides, then screened 22 hexapeptides to examine their self-assembling properties and structural features using atomistic molecular dynamics simulations and experiments such as atomic force microscopy and circular dichroism, and finally obtained 18 aggregation-forming and 4 non-aggregation-forming samples [14]. Thus, a total of 180 training samples including 85 aggregation-forming hexapeptides as positive samples and 95 non-aggregation-forming hexapeptides as negative samples (Table S1) was used to train the QSAR model.

The predictive capability of the model was validated by one test set including 109 experimentally identified hexapeptides (Table S2). The test set including 48 positive samples and 61 negative samples was collected by removing the duplicate or contradictory samples from 120 hexapeptides reported by Maurer-Stroh et al. [27].

2.2. Structure characterization

Factor analysis scales of generalized amino acid information (FASGAI) proposed by our group [28] was used to represent the structural features of 180 training hexapeptides. Briefly, the FASGAI characterized a total of 335 physicochemical and other properties for each of the 20 natural amino acids, followed by clustering these 335 properties into 6 fingerprint factors named as hydrophobicity, alpha and turn propensity, bulky property, local flexibility, compositional characteristics, and electronic property (Table S3) [29]. Based on fingerprint factor scores for each amino acid, the FASGAI method can generally represent sequence and structural features of any peptide by simply constructing $6 \times n$ matrices, where n is the number of residues. Thus, the structural features of any hexapeptide can be readily characterized by a 6×6 FASGAI matrix.

2.3. Supporting vector machine modeling

Supporting vector machine (SVM), as a machine learning algorithm, finds an optimal hyperplane to maximize the classification of the two types of sample intervals [30]. In linearly separable cases, SVM constructs a hyperplane which separates two different classes of vectors with a maximum margin. In nonlinearly separable cases, SVM maps the input variables into a high-dimensional feature space with the kernel function which can effectively deal with dimensional puzzledom, calculation complexity, etc. In this work, the radial basis function kernel was used for our SVM modeling. The two parameters, the regularization parameter C and the kernel width parameter γ, were adjusted based on the accuracy criteria of leave one out (LOO) cross validation using a grid search approach as follows: First, a possible interval of C (or γ) with the grid space was provided. Then, all grid points of (C, γ) were tried to see which one gave the highest cross validation accuracy. Finally, the best parameters were used to train the whole training dataset and generate the final model. The software LIBSVM 3.2 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used to construct the SVM model.

2.4. Model validation

The validity of the established model was assessed by LOO cross validation and external validation. In the LOO test, to avoid any biased propensity by using single randomly selected sample in the dataset, 180 samples were iteratively removed one at a time, and the predictive performance was recalculated each time and then averaged by 180 times [31]. In external validation, the test set was employed to validate the predictive capability of the predictor.

2.5. Model evaluation

The predictive performance was evaluated using the statistical parameters as follows [32]: Accuracy (*Acc*), Sensitivity (S_n), Specificity

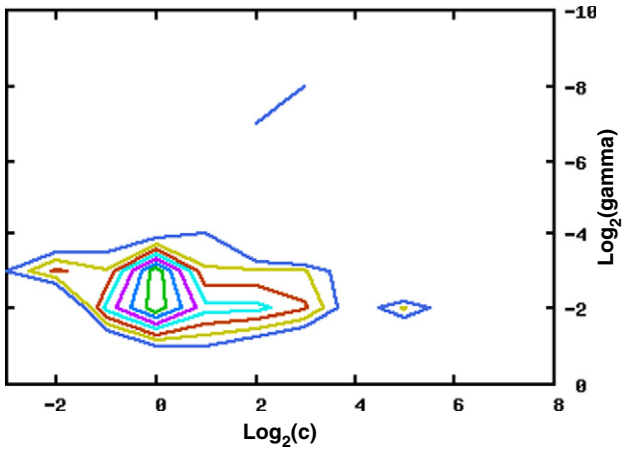

Fig. 1. Parameters by the grid search with LOO cross validation on 180 training hexapeptides. The SVM model peaks at $Acc = 78.33\%$ with $C = 1$ and $\gamma = 0.25$.