A novel predicting algorithm for Thermostable Proteins based on Hurst exponent and Maximized L-measure

HSIANG-CHUAN LIU
Department of Bioinformatics
Asia University
No. 500, Lioufeng Rd., Wufeng, Taichung County, 41354
TAIWAN, ROC
lhc@asia.edu.tw / fpan0366@yahoo.com.tw

HSIEN-CHANG TSAI
Department of Biology,
National Changhua University of Education
No. 1, Jin-De Rd., Changhua, 50001
TAIWAN, ROC
bihft@cc.ncue.edu.tw

SHIN-WU LIU
Laboratory of Viral Diseases
National Institutes of Health
No. 9000 Rockville Pike Bethesda, Maryland 20892
USA
liushin@mail.nih.gov

Abstract: - Establishing a good algorithm for predicting temperature of thermostable proteins is an important issue. In this study, a new thermostable proteins prediction method by using Hurst exponent and Choquet integral regression model with respect to maximized L-measure is proposed. The main idea of this method is to integrate the physicochemical properties, long term memory property and Choquet integral regression model with respect to maximized L-measure for amino symbolic sequences of different lengths. For evaluating the performance of this new algorithm, a 5-fold Cross-Validation MSE is conducted. Experimental result shows that this new prediction algorithm is better than the Choquet integral regression model with respect to other well known fuzzy measure, Lambda-measure, P-measure, and L-measure, respectively and the traditional prediction models, ridge regression and multiple linear regression models, respectively.

Key-Words: - Hurst exponent, Lambda-measure, P-measure, L-measure, Maximized L-measure

1 Introduction

Many experiments and chemical reactions have to be performed in high temperature environment in many fields such as medical industry, foodstuff industry, and cosmetics industry. Furthermore, many materials employ the thermostable proteins as its basic component. These make the research about thermostable proteins a new and important field nowadays. One important issue about thermostable proteins is to predict the temperature of thermostable proteins. In this paper, a prediction algorithm of thermostable proteins by using Hurst exponent and Choquet integral regression model with maximized L-measure and γ-support is proposed. The contribution of this method is to integrate the physicochemical properties, long term memory property and Choquet integral regression model based on our proposed fuzzy measure, maximized L-measure, for amino symbolic sequences with different lengths.

Basically, the procedure of the proposed method is as follows. A thermostable proteins data set was downloaded from the Protein Data Bank (PDB), http://www.rcsb.org, [1]. By substituting four physicochemical quantities of each residue of amino acid in sequence of the thermostable proteins using the four feature
scaling estimators, we can obtain four non-symbolic sequences of the thermostable proteins. Then we compute the Hurst exponents of each non-symbolic sequence of the thermostable proteins, so that we can obtain four features of Hurst exponents in each sequences of the thermostable protein. With these extracted features, the Choquet integral regression model based on our proposed fuzzy measure, maximized L-measure, is used to predict the temperature of the 40 thermostable proteins.

2 Four scaling estimators of physicochemical properties for each amino acid

Four physicochemical scaling estimators, solvent-accessible surface area, solvent accessibility percentage, electrostatic interaction, and contact energy in the situation of single amino acid were described as follows.

2.1 Solvent accessible surface area (ASA)

The ASA of a protein was obtained using POPS [2-3] on the following web site (mathbio.nimr.mrc.ac.uk/~ff-ranca/POPS/), selecting output residue areas (POPS_R). Both the polar (hydrophilic) and apolar (hydrophobic) surface areas can be obtained from the output residue areas, which were then changed to the percentage of apolar area for each residue in a protein.

2.2 Solvent accessibility percentages

The solvent accessibility percentages of the residues were obtained using the ASAView [4] data base (www.netasa.org/asaview/). Residues were classified to be buried in a protein core as the values between 0-50%, and those were considered to be exposed to solvent when the percentage exceeded 50%.

2.3 Electrostatic interactions

The number of ion pairs (electrostatic interactions) was calculated according to the following criterion [4]: two oppositely charged residues were considered an ion pair if the distance between the oppositely charged atoms of these residues was less than 18Å. Asp, Glu, Arg, Lys and His residues were used to calculate the ion pairs.

2.4 Contact energies

A 20×20 matrix of effective contact energies, the interaction energies between all amino acids pairs, was developed by Miyazawa and Jernigan [5-6], which was also called MJ matrix. The MJ effective energy (eij), which is the element of MJ matrix, was derived from all the possible interaction energies, including hydrophobic and solvation energies. Furthermore, the hydrophobic interaction is the dominant contribution to the MJ effective energy. The eij can be presented as the following equation

\[
e_{ij} = e'_{ij} + e_{ii} + e_{jj} / 2
\]

The e'ij is the mixing term, which is the free energy change upon the mixing of residues of type i and residues of type j when the contacts in self-pairs i-i and j-j are separated to form i-j pairs. The eii or eij is the free energy change after the desolvation of residue i or of residue j to form self-pairs i-i or j-j. The values of eii or eij should have high correlation with the hydrophobicity of residue type i or residue type j [5-6].

The average contact energy of each type of amino acid, ei, was used in this work, and it is defined as: [5-6].

\[
e_i = \frac{\sum_{j=1}^{20} e_{ij} N_{ij}}{N_{ir}}
\]

where

\[
N_{ij} = \sum_{p} n_{ij,p}, \quad n_{ir} = \sum_{q} n_{iq}
\]
The Hurst exponent (H) is a statistical measure used to classify time series for long term memory and predictability [7-8].

In this study, R/S method is used for the estimation of the Hurst exponent: R/S method [10] is based on empirical observations by Hurst in 1965 and estimates H are based on the R/S statistic. It indicates (asymptotically) second-order self-similarity. H is roughly estimated through the slope of the linear line in a log-log plot, depicting the R/S statistics over the number of points of the aggregated series. That is, given a time sequence of observations, \( w_t \) define the Series

\[
W(t,\tau) = \sum_{i=1}^{t} (w_i - \bar{w}_\tau), 1 \leq t \leq \tau \tag{4}
\]

where

\[
\bar{w}_\tau = \frac{1}{T} \sum_{i=1}^{t} w_i \tag{5}
\]

Define

\[
R(\tau) = \max_{t=1}^{\tau} W(t,\tau) - \min_{t=1}^{\tau} W(t,\tau) \tag{6}
\]

and

\[
S(\tau) = \sqrt{\frac{1}{T} \sum_{i=1}^{t} (w_i - \bar{w}_\tau)^2} \tag{7}
\]

In plotting \( \log R(\tau) / S(\tau) \) against \( \log \tau \), we expect to get a line whose slope determines the Hurst exponent.

4 Fuzzy Measures

The two well known fuzzy measures, the \( \lambda \)-measure proposed by Sugeno in 1974, and P-measure proposed by Zadah in 1978, are concise introduced as follows.

3.1 Fuzzy Measures [9-10]

A fuzzy measure \( \mu \) on a finite set \( X \) is a set function \( \mu : 2^X \rightarrow [0, 1] \) satisfying the following axioms:

1) \( \mu(\emptyset) = 0, \mu(X) = 1 \) (boundary conditions) \tag{8}

2) \( A \subseteq B \Rightarrow \mu(A) \leq \mu(B) \) (monotonicity) \tag{9}

3.2 Fuzzy density function [9-10]

A fuzzy density function, or singleton measure, of a fuzzy measure \( \mu \) on a finite set \( X \) is a function \( s : X \rightarrow [0, 1] \) satisfying:

\[
s(x) = \mu(\{x\}), x \in X \tag{10}
\]

\( s(x) \) is called the fuzzy density of singleton \( x \).

If \( \sum_{x \in X} s(x) = 1 \), then the fuzzy density function is called normalized density function.

3.1 \( \lambda \)-measure [10]

For given singleton measures \( s(x) \) on a finite set \( X \), a \( \lambda \)-measure, \( g_{\lambda} \), is a fuzzy measure on \( X \), satisfying:

1) \( \mu(\emptyset) = 0, \mu(X) = 1 \) (boundary conditions)

2) \( A, B \in 2^X, A \cap B = \emptyset, A \cup B \neq X \)

\[
\Rightarrow g_{\lambda}(A \cup B) = g_{\lambda}(A) + g_{\lambda}(B) + \lambda g_{\lambda}(A) g_{\lambda}(B) \tag{11}
\]

3) \( \prod_{i=1}^{n} \left[1 + \lambda s(x_i)\right] = \lambda + 1 > 0, s(x_i) = s(\{x_i\}) \) \tag{12}

Note that \( \lambda \)-measure has a unique solution without closed form.

3.4 P-measure [11]

For given singleton measures \( s(x) \) on a finite set \( X \), a P-measure, \( g_P \), is a fuzzy measure on \( X \), satisfying:

\[
\forall A \subseteq 2^X \Rightarrow g_P(A) = \max_{x \in A} \left\{ s(x) \right\} = \max_{x \in A} \left\{ g_P(\{x\}) \right\} \tag{13}
\]

Note that for any subset of \( X \), A, P-measure considers only the maximum value and will lead to insensitivity.

4 multivalent fuzzy measures

4.1 Definition of multivalent fuzzy measure

If a fuzzy measure has only one fuzzy measure solution, then it is called a univalent fuzzy measure, otherwise it is called a multivalent fuzzy measure.

4.2 L-measure [14]
4.2.1 Definition of L-measure

For given singleton measure \( s(x) \) on a finite set \( X \), \( x \in X \), a L-measure, \( g_L \), is a fuzzy measure on \( X \), satisfying:

1) \( g_L(\emptyset) = 0, g_L(X) = 1 \)
2) \( L \in [0, \infty) \), \( \forall A \subset X, A \neq X \) \Rightarrow

\[
g_L(A) = \max_{x \in A} \{ s(x) \} + \frac{L(|A|-1) \sum_{x \in A} s(x) \left[ 1 - \max_{x \in A} \{ s(x) \} \right]}{|X| - |A| + L(|A|-1) \sum_{x \in X} s(x)}
\] (14)

4.2.2 Important properties of L-measure

Theorem 1
(i) L-measure is increasing and continuous function of \( L \) on \( [0, \infty) \).
(ii) L-measure has infinitely many fuzzy measure solutions for \( L \in [0, \infty) \).
(iii). L-measure is a multivalent fuzzy measure
(iv) If \( L = 0 \) then L-measure is just the P-measure
(v) L-measure is more sensitive then L-measure

Proof: (v) Since the denominator of \( g_L(A) \),
\[
|A| - |A| \sum_{x \in A} s(x) + |X| - |A| \sum_{x \in X} s(x)
\]

is more sensitive than the denominator of \( g_L(A) \),
\[
|X| - |A| + L(|A|-1) \sum_{x \in X} s(x)
\]

6 \( \gamma \)-support [13-15]

For given singleton measure \( s(x) \) of a fuzzy measure \( \mu \) on a finite set \( X \), if \( \sum_{x \in X} s(x) = 1 \), then \( s \) is called a fuzzy support measure of \( \mu \), or a fuzzy support of \( \mu \), or a support of \( \mu \). One kind of fuzzy supports is introduced as below.

Let \( \mu \) be a fuzzy measure on a finite set \( X = \{ x_1, x_2, \ldots, x_n \} \), \( y_i \) be global response of subject \( i \) and \( f_i(x_j) \) be the evaluation of subject \( i \) for singleton \( x_j \), satisfying:

\[
0 < f_i(x_j) < 1, i = 1, 2, \ldots, N, j = 1, 2, \ldots, n
\] (18)

\[
s(x_j) = \frac{1 + r(f(x_j))}{\sum_{x \in X} \left[ 1 + r(f(x)) \right]}, j = 1, 2, \ldots, n
\] (19)

where \( r(f(x_j)) = \frac{S_{j,x_j}}{S_jS_{x_j}} \) (20)

\[
S_j = \frac{1}{N} \sum_{i=1}^N \left( y_i - \frac{1}{N} \sum_{i=1}^N y_i \right)^2
\] (21)

\[
S_{ij} = \frac{1}{N} \sum_{i=1}^N \left( f_i(x_j) - \frac{1}{N} \sum_{i=1}^N f_i(x_j) \right)^2
\] (22)
7 Choquet Integral Regression Models

7.1 Choquet Integral [12]

Let $\mu$ be a fuzzy measure on a finite set $X$. The Choquet integral of $f_i : X \rightarrow \mathbb{R}$ with respect to $\mu$ for individual $i$ is denoted by

$$
\left[ c f_d \mu = \sum_{j=1}^{N} f_i(x_{(j)}) - f_i(x_{(i-1)}) \right] \mu(A^i_j), i = 1, 2, ..., N
$$

where $f_i(x_{(0)}) = 0$, $f_i(x_{(j)})$ indicates that the indices have been permuted so that

$$
0 \leq f_i(x_{(1)}) \leq f_i(x_{(2)}) \leq \ldots \leq f_i(x_{(N)})
$$

$A^i_j = \{x_{(j)}, x_{(j+1)}, \ldots, x_{(a)}\}$

7.2 Choquet Integral Regression Models [13-15]

Let $y_1, y_2, \ldots, y_N$ be global evaluations of $N$ objects and $f_1(x_j), f_2(x_j), \ldots, f_N(x_j), j = 1, 2, \ldots, N$, be their evaluations of $x_j$, where $f_i : X \rightarrow \mathbb{R}, i = 1, 2, \ldots, N$.

Let $\mu$ be a fuzzy measure, $\alpha, \beta \in \mathbb{R}$,

$$
y_i = \alpha + \beta \int c f_d \mu + e_i, \quad e_i \sim N(0, \sigma^2), \quad i = 1, 2, \ldots, N
$$

$$
(\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta} \left[ \sum_{i=1}^{N} \left( y_i - \alpha - \beta \int c f_d \mu \right)^2 \right]
$$

then $\hat{y}_i = \hat{\alpha} + \hat{\beta} \int c f_d \mu, \quad i = 1, 2, \ldots, N$ is called the Choquet integral regression equation of $\mu$, where

\[
\hat{\beta} = \frac{S_{\gamma}}{S_{\beta}}
\]

$$
\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} y_i - \hat{\beta} \frac{1}{N} \sum_{i=1}^{N} \int f_i d \mu
$$

\[
\sum_{i=1}^{N} \left( y_i - \frac{1}{N} \sum_{i=1}^{N} y_i \right) \left( \int f_i d \mu - \frac{1}{N} \sum_{i=1}^{N} f_i d \mu \right) = \sum_{i=1}^{N} \left( \int f_i d \mu - \frac{1}{N} \sum_{i=1}^{N} f_i d \mu \right)^2
\]

8 Experiment and Result

A thermostable proteins data set is provided by the Protein Data Bank (PDB). There are 40 instances with two classes. We can obtain four features represented as Hurst exponents respectively in each sequences of the thermostable protein. The transformed data can be found in our previous paper [8]. It is omitted here for lack of space.

The generated data with four features in terms of Hurst exponents is applied to evaluate the 5-fold Cross-Validation MSE of six prediction models: the Choquet integral regression model based on P-measure, $\lambda$-measure, $L$-measure and maximized $L$-measure, respectively, the multiple linear regression, and ridge regression model.

### TABLE 1 MSE OF REGRESSION MODELS

<table>
<thead>
<tr>
<th>Regression model</th>
<th>5-fold CV MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE-Chequet integral Regression model with fuzzy measure</td>
<td>L&lt;sub&gt;\text{P}&lt;/sub&gt; 21.4673</td>
</tr>
<tr>
<td></td>
<td>L 21.7794</td>
</tr>
<tr>
<td></td>
<td>$\lambda$ 22.0117</td>
</tr>
<tr>
<td></td>
<td>$\mu$ 22.5051</td>
</tr>
<tr>
<td></td>
<td>HE-Ridge regression 23.9718</td>
</tr>
<tr>
<td></td>
<td>HE-Multiple linear regression 25.3937</td>
</tr>
</tbody>
</table>

**ISSN**: 1790-2769

The experimental results are listed in Table 1. We can find that the Choquet integral regression model based on maximized L-measure has the best performance.

10 Conclusion

In this paper, an improved prediction algorithm of thermostable proteins combining four feature scaling estimators, Hurst exponent, and the Choquet integral regression model with maximized L-measure is proposed. For evaluating the performance of this new algorithm, a thermostable proteins data set by using 5-fold Cross-Validation MSE is conducted. Experimental result shows that this new prediction algorithm is useful and better than other prediction models.

In the future, we will consider looking for some improving prediction algorithm of thermostable proteins by using Hurst exponent and Choquet integral regression model based on other improved fuzzy measures.

11 Acknowledgment

This paper is partially supported by the National Science Council grant of Taiwan Government (NSC 98-2410-H-468-014)

References:

[1] Protein Data Bank (PDB), http://www.rcsb.org/