A Fuzzy Query Method Based on Human-Readable Rules for Predicting Protein Stability Changes

Liang-Tsung Huang*,1, Lien-Fu Lai2 and Chao-Chin Wu²

1Department of Computer Science and Information Engineering, Mingdao University, Changhua 523, Taiwan

2Department of Computer Science and Information Engineering, National Changhua University of Education, Changhua 500, Taiwan

Abstract: Predicting protein mutant stability changes is important for protein design. Many methods have tried to improve prediction accuracy by various models. However, it will be difficult to employ them when the required input information is incomplete. This paper presents a fuzzy query method (named FQ-STAB), which cooperates with a human-readable rule base to predict stability changes upon single mutations. Firstly, we have effectively established a set of classification rules as a knowledge representation from the thermodynamic database. Next, we applied the proposed method on the rules to predict stability changes under the condition without sufficient information. Further, FQ-STAB has been tested on a data set of single point mutants. The results show that it can be applied to the prediction of stability changes using partial input information, and can also provide the explanation capabilities for the predicted outcome. We have developed a web server for predicting protein stability changes upon single mutations by using fuzzy query mechanism and it is available at http://bioinformatics.myweb.hinet.net/fqstab.htm.

Keywords: Protein stability, prediction, fuzzy query, rule base.

INTRODUCTION

Accurate prediction of protein stability changes upon single amino acid substitutions can provide valuable clues to a better understanding of the nature of protein structure and function, which is essential to the design of the novel protein. In earlier studies, different approaches have been adopted to predict stability changes upon single point mutations: (i) energy-oriented, (ii) method-oriented and (iii) feature-oriented. The first approach derives energy functions by using physical [1, 2], statistical [3, 4] or empirical [5-7] potentials. The second develops machine learning models such as artificial neural networks [8], support vector machines [9], decision tree [10], etc. Further, the third pays attention to discuss significant features to improve prediction accuracy [11, 12].

Those methods have found many ways to predict stability changes from different kinds of input information. It implies that giving sufficient information is necessary. However, in practice, the required input information may be imprecise, incomplete, and even unavailable. In this situation, it will be difficult to employ the methods for accurate prediction. Therefore, it would be constructive to develop an effective method to overcome the practical problem.

Fuzzy logic [13, 14] is a generalization of classic logic and it provides a formal way to deal with uncertainty and imprecision. Several recent studies have adopted the approach to solve problems with molecular biology and bioinformatics research [15], such as protein function prediction [16] and metabolic control analysis [17]. Based on the framework of fuzzy logic, we have proposed a fuzzy query method (named FQ-STAB) that cooperates with a human-readable rule base to predict stability changes upon single mutations. Firstly, we have effectively established a set of classification rules as a knowledge representation from the thermodynamic database. Next, we applied a fuzzy query mechanism on the rules to predict stability changes under the condition without sufficient information.

MATERIALS

Data Set

In the present work, we have compiled a data set of 1859 single mutants of 64 different proteins from ProTherm thermodynamic database [18, 19]. We have removed the duplicate mutants that have same mutated and mutant residues, mutated position, experimental conditions (pH and temperature) and ΔG values. Further, we kept only one mutant data for the mutants with different values but the same conditions by calculating their average. The analysis of data showed that 31.4% of mutants are stabilizing and 68.6% of them are destabilizing. Regarding solvent accessibility (SA), 42% are buried (SA < 20%), 26% are partially buried (20% < SA < 50%) and 32% are exposed (SA > 50%). Further, the occurrence of mutants in helical, strand, turn and coil regions are 47%, 20%, 11% and 22%, respectively.
Rule Base

In our earlier work, we have developed a human-readable rule generator for integrating amino acid sequence information and stability of mutant proteins [20]. Here, we utilized this tool to convert the data set into 104 classification rules along with related information. And then we calculated several measures of the rule, including (i) number of data, the number of data for completely matching the antecedent of the rule; (ii) support, the proportion of number of data to the whole data set; and (iii) accuracy: the accuracy of the rule for applying it to the data set. The statistics for all rules are also computed. 74 and 30 rules predict the stability of protein mutants as stabilizing and destabilizing, respectively. For stabilizing, the mean number of data, support and accuracy are 10.35, 0.005 and 0.688 respectively. For destabilizing, they are 33.53, 0.018 and 0.966, respectively. Full details of the rules have been available on the web.

A rule consists of antecedent and consequent. The consequent indicates the result (stabilizing and destabilizing) when all condition elements of the antecedent are satisfied. The antecedent includes the information about the original and substituted residue types, and three neighboring residue type for both sides.

METHODS

Fuzzy Query Method

In the framework of the fuzzy query mechanism, the prediction problem with incomplete input information is considered as a query problem from the rule base. In other words, the prediction is made by the rules with high confidence level.

For achieving this aim, we apply three key concepts: (i) fuzzy matching is applied to compute the degree of matching between the input partial information and the antecedent of rules. (ii) Fuzzy numbers [21, 22] are utilized to represent linguistic terms and (iii) fuzzy weighted average (FWA) [23-25] is applied to calculate the confidence level of a rule.

Fuzzy Matching

The incomplete input information can be categorized into the unknown and the known parts. For quantifying the match degree of the known part with each rule, the elements of the known part are compared to the condition elements of the antecedent of rules. For simplicity, let us assume that we have N rules with only two condition elements:

\[ R_1: \text{IF E}_1^1 \text{ AND E}_1^2 \text{ THEN C}_1, \]
\[ R_2: \text{IF E}_2^1 \text{ AND E}_2^2 \text{ THEN C}_2, \]
\[ \vdots \]
\[ R_n: \text{IF E}_n^1 \text{ AND E}_n^2 \text{ THEN C}_n, \]

where \( E_i^1 \) and \( E_i^2 \) denotes the condition elements of rule \( R_i \), and \( C_i \) is the consequent of rule \( R_i \). If the elements of the known part are given as set \( E \), then the score of the match degree for rule \( R_n \), is calculated by:

\[ \text{MD}_n = \Phi/N, \]  

where \( \Phi \) is the number of the intersection between sets \( E \) and \( \{E_i^1, E_i^2\} \).

Fuzzy Number

The confidence level of a rule mainly depends on its accuracy and support. FQ-STAB provides the adjustment in the importance of these two factors to improve the resulted predictions of fuzzy queries. To distinguish between various degrees of importance, “importance” is expressed in terms of “unimportant”, “rather unimportant”, “moderately important”, “rather important”, and “very important”. The five linguistic terms can be represented as fuzzy numbers (Fig. 1).

A fuzzy number \( \tilde{A} \) can be defined by a triplet \((a, b, c)\) and the membership function \( \mu_{\tilde{A}}(x) \) is defined as:

\[ \mu_{\tilde{A}}(x) = \begin{cases} 
0, & x < a \\
\frac{x - a}{b - a}, & a \leq x \leq b \\
\frac{c - x}{c - b}, & b \leq x \leq c \\
0, & x > c 
\end{cases} \]  

Fig. (1). The membership functions of “importance”. 

144  The Open Structural Biology Journal, 2009, Volume 3  Huang et al.
Fuzzy Weighted Average

According to the given importance, the confidence level of a rule can be computed by aggregating its accuracy and support. We apply the fuzzy weighted average (FWA) to calculate the confidence level of a rule using fuzzy numbers. FWA is a weighted average involving type-1 fuzzy sets. In FQ-STAB, the accuracy and the support are indicators \( y_i \) that rate the confidence level of a rule. The degrees of importance are weights \( w_j \) that act upon indicators. The fuzzy weighted average \( y \) is defined as

\[
y = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j} = f(x_1, \ldots, x_n, w_1, \ldots, w_n),
\]

where \( x_j \) and \( w_j \) are represented by fuzzy sets or fuzzy numbers; \( n \) is the number of the factors. Since \( w_j \) are no longer crisp numbers, \( \sum_{j=1}^{n} w_j = 1 \) is not imposed as a requirement. For example, consider a rule \( R_i \) having the accuracy rate = 0.78 and the relative quantity = 0.54. In the case that the accuracy is very important and the support is moderately important, we can calculate the confidence level of the rule as follows:

\[
CL_i = \frac{0.78 \odot (0.75, 1, 1) \odot 0.54 \odot (0.25, 0.5, 0.75)}{(0.75, 1, 1) \odot (0.25, 0.5, 0.75)} = \frac{0.997}{1.433} = 0.6957
\]

Applying mathematical operations to fuzzy numbers [26, 27], we have two fuzzy numbers \((0.72, 1.05, 1.185)\) and \((1, 1.5, 1.75)\). The centre of gravity is adopted to defuzzify a fuzzy number [28], which is achieved by mathematical integral. The resulted predictions to a fuzzy query are ranked by the reference level:

\[
RL_i = MD_i \times CL_i,
\]

Namely, the multiplication of \( MD_i \) (the match degree between the query and rule \( R_i \)) and \( CL_i \) (the confidence level of rule \( R_i \)).

FuzzyCLIPS [29] is a knowledge-base programming language designed especially for developing fuzzy expert systems. FQ-STAB utilizes FuzzyCLIPS to deal with imprecision and uncertainty in fuzzy query and to offer the capability of fuzzy matching and fuzzy reasoning.

RESULTS AND DISCUSSION

Prediction of Complete Input Information

The proposed method is based on a rule base, which is derived from a large data set of single mutants by a rule generator tool. The rule base reflects an integrative view of the data set and can be used to predict stability changes.

When complete input information is available, the prediction can be made by the rule-based model. We have tested the performance of the rule base and the results showed that it can improve the accuracy up to 79.1% and 80.2% for 10- and 20-fold cross-validation tests, respectively. The performance is comparable to other methods in the literature such as artificial neural networks (ANN) [8], support vector machines (SVM) [30] and the method based on torsion and atom potentials [31].

Characteristics of Rule Base

From a practical point of view, the rule base provides more useful information than a prediction model. The rule can be regarded as a representation of the knowledge about predicting stability changes. For example, one of the rules is that “IF mutant residue = P and N2 = K THEN increase”, which indicates that if the mutant residue is Pro, and its second neighbor towards the N-terminal is Lys, then the predicted stability change is increasing, namely stabilizing. The representation gives a reason why a prediction was made. And the IF-THEN structure also provides the foundation on which the rules can be inferred and queried.

Prediction of Incomplete Input Information

In this study, one of the main aims is to predict protein stability changes under insufficient conditions. For a further examination, we have given a query with partial input information to demonstrate the proposed method.

Let \( Q_1 \) is a query with the following conditions:

(i) Mutated residue is Ala.
(ii) Mutant residue is Pro.
(iii) The second neighbor towards the N-terminal is Lys.
(iv) The importance of the rule accuracy is very important.
(v) The importance of the rule support is moderately important.

As the original prediction model requires mutated and mutant residues, and three neighboring residue type for both sides, the query can not provide complete details of information required. In FQ-STAB, the unavailable elements are labeled as the unknown, and then the elements of the known part can be utilized to calculate the reference level to each rule.

The results from submitting query \( Q_1 \) to FQ-STAB are shown in Table 1. For the antecedent, \( Ni/Ci \) denotes the \( i \)-th residue of an extended sequence from the mutated residue towards the N-/C-terminus, where the residue type is encoded as one letter. Md and Mt are the mutated and mutant residues, respectively. Fuzzy matching firstly finds all the rules which include at least one of the known part elements of the query. For example, rule no. 5 satisfies the conditions (ii) and (iii) in \( Q_1 \). In other words, the antecedent of rule no. 5 and query \( Q_1 \) have two common elements: mutant residue Ala. Therefore, there are total eight responded rules, antecedents of which have different intersections with query \( Q_1 \). The rules found are then considered as the candidates for predicting stability changes.
Although all eight rules are referable to the query, it is uncertain which one is most suitable for the query by instinct. For example, compared with rule no. 5, rule no. 93 has significantly higher support but lower accuracy. Moreover, the two rules have led to contradictory conclusions. To deal with the problem of uncertainty, FQ-STAB calculates reference level to each rule. According to the importance of accuracy and support, the rule no. 5 has the highest RL value of 0.657, which indicates that this rule and the query are a good match and the query may have the same consequent, stabilizing, as the rule.

**Readability and Flexibility**

The proposed method is based on the rules with the IF-THEN structure, which is easy to understand for people. The representation has several advantages. Firstly, it provides the query with a clear explanation, which is easy to compare with other known reports. For example, rule no. 93 in Table 1 can be directly read as “IF the mutated residue is Ala and its first neighbor towards the N-terminal is Gln, then the predicted stability change is decreasing”.

Secondly, it can reflect the actual facts of the experimental results. For example, in Table 1, rule no. 93 shows that 43 instances of the experimental data set fit the antecedent of this rule, and 40 instances agree with the rule. Thirdly, the rule base is fairly flexible about additions and deletions. Through the fuzzy query mechanism, adding a new rule leads to the calculation of the reference level for this rule, which does not come into conflict with existing rules.

**Fuzzy Query on Web**

We have developed a web server for employing FQ-STAB and it is available at http://bioinformatics.myweb.hinet.net/fqstab.htm. On the web, only four steps are required to submit a query (see Fig. 2): (i) Select

Table 1. The Rules Responding to Query Q1 in Order of Reference Level

<table>
<thead>
<tr>
<th>No.</th>
<th>Antecedent</th>
<th>Consequent</th>
<th>Rule size</th>
<th>Number of data</th>
<th>Correct</th>
<th>Support</th>
<th>Accuracy</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Mt = P, N2 = K</td>
<td>I</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>0.004</td>
<td>1.00</td>
<td>0.657</td>
</tr>
<tr>
<td>48</td>
<td>Md = A, Mt = P, N1 = N</td>
<td>I</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>0.004</td>
<td>0.87</td>
<td>0.385</td>
</tr>
<tr>
<td>93</td>
<td>Md = A, N1 = Q</td>
<td>D</td>
<td>2</td>
<td>43</td>
<td>40</td>
<td>0.023</td>
<td>0.93</td>
<td>0.332</td>
</tr>
<tr>
<td>77</td>
<td>Md = A, N1 = L</td>
<td>D</td>
<td>2</td>
<td>9</td>
<td>9</td>
<td>0.005</td>
<td>1.00</td>
<td>0.330</td>
</tr>
<tr>
<td>76</td>
<td>Mt = P, N1 = A</td>
<td>D</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>0.003</td>
<td>1.00</td>
<td>0.328</td>
</tr>
<tr>
<td>2</td>
<td>Md = A, N1 = H</td>
<td>I</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0.002</td>
<td>1.00</td>
<td>0.326</td>
</tr>
<tr>
<td>3</td>
<td>Md = A, N1 = T</td>
<td>I</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.001</td>
<td>1.00</td>
<td>0.325</td>
</tr>
<tr>
<td>69</td>
<td>Md = A, N1 = G</td>
<td>I</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>0.002</td>
<td>0.66</td>
<td>0.220</td>
</tr>
</tbody>
</table>

Md: mutated residue; Mt: mutant residue; Ni and Ci denote the i-th residue of an extended sequence from the mutated residue towards the N-/C-terminus, I: increasing or stabilizing; D: decreasing or destabilizing; RL: reference level.

![Fig. (2). A snapshot of the prediction page with the input procedure for predicting stability changes.](image-url)
"Prediction" from the main page to open an input sub page. (ii) Set the residue type for the mutated residue, the mutant residue and the neighboring residues from the drop-down list. (iii) Give the importance for accuracy and support, respectively. (iv) Press "Submit" button to run.

Regarding the residue information, seven drop-down lists on the top represent N3, N2, N1, Md, C1, C2; and C3 and the bottom list represents Mt, where the residue type is encoded as three letters. The degree of the importance can distinguish accuracy and support, which are subjective judgments of rule importance.

Fig. (3) shows the details of the rules responding to the query. Total five main columns give further information to each rule, including: (i) Rule no., the rule number in the rule base; (ii) Antecedent, the hyperlink to the antecedent of the rule; (iii) Consequent, the prediction for the query; (iv) Reference level, the quantification of the rule importance; and (v) Percentage, a graphical representation of the reference level. The results show that the rule no. 5 is a good match for the query and give an "increase" (stabilizing) prediction.

CONCLUSIONS

We have developed a novel method for predicting protein stability changes especially when input information is insufficient. The proposed method applied fuzzy query mechanism to a set of readable rules generated from experimental data of single mutants, and made each rule the reference level to quantify the uncertainty of rule importance. Moreover, the predicted results exhibited a rule representation, which is human-readable and easily examined.

To sum up, the following are the strong points of the contribution in this work: (i) the prediction can be accurately made using partial input information; (ii) it provides the explanation capabilities for the predicted outcome; (iii) the rule base can be automatically generated from a large database and additions and deletions are flexible on the rule base; and (iv) rules can be easily examined by comparing them with other known reports.

We suggest that our method can be applied to the prediction of protein stability changes upon single mutations using sufficient or insufficient input information. And the predicted results can also promote more understanding of mutant protein stability.

REFERENCES


