Open Access

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

Liang-Tsung Huang*[,1], Lien-Fu Lai[2] and Chao-Chin Wu[2]

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

[2]*Department 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 $\Delta\Delta G$ 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.

*Address correspondence to this author at the Department of Computer Science and Information Engineering, Mingdao University, Changhua 523, Taiwan; Tel: +886 4 8876660, Ext. 8119; Fax: +886 4 8782134; E-mail: larry@mdu.edu.tw

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: IF E_1^1 AND E_1^2 THEN C_1,

R_2: IF E_2^1 AND E_2^2 THEN C_2,

…

R_i: IF E_i^1 AND E_i^2 THEN C_i,

…

R_N: IF E_N^1 AND E_N^2 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_i, is calculated by:

$$MD_i = \Phi/N, \tag{1}$$

where Φ 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:

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 & ,x < a \\ \dfrac{x-a}{b-a} & ,a \le x \le b \\ \dfrac{c-x}{c-b} & ,b \le x \le c \\ 0 & ,x > c \end{cases} \tag{2}$$

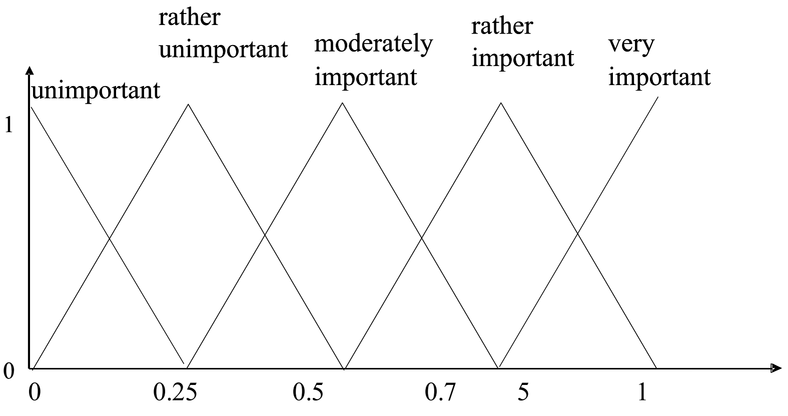


Fig. (1). The membership functions of "importance".

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 (x_j) 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

$$y = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j} = f(x_1,...,x_n,w_1,...,w_n), \tag{3}$$

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:

$$\mathrm{CL}_i = \frac{0.78 \otimes (0.75, 1, 1) \oplus 0.54 \otimes (0.25, 0.5, 0.75)}{(0.75, 1, 1) \oplus (0.25, 0.5, 0.75)}$$

$$= \frac{(0.72, 1.05, 1.185)}{(1, 1.5, 1.75)} \approx \frac{0.997}{1.433} = 0.6957$$

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:

$$\mathrm{RL}_i = \mathrm{MD}_i \times \mathrm{CL}_i, \tag{4}$$

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, N*i*/C*i* 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 Pro and the second neighbor towards the N-terminal Lys. Rule no. 93 satisfies the condition (i) in Q_1. Namely, the antecedent of rule no. 93 and the same query have another common one: mutated 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.

Table 1. The Rules Responding to Query Q_1 in Order of Reference Level

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

Md: mutated residue; Mt: mutant residue; N*i* and C*i* 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

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.

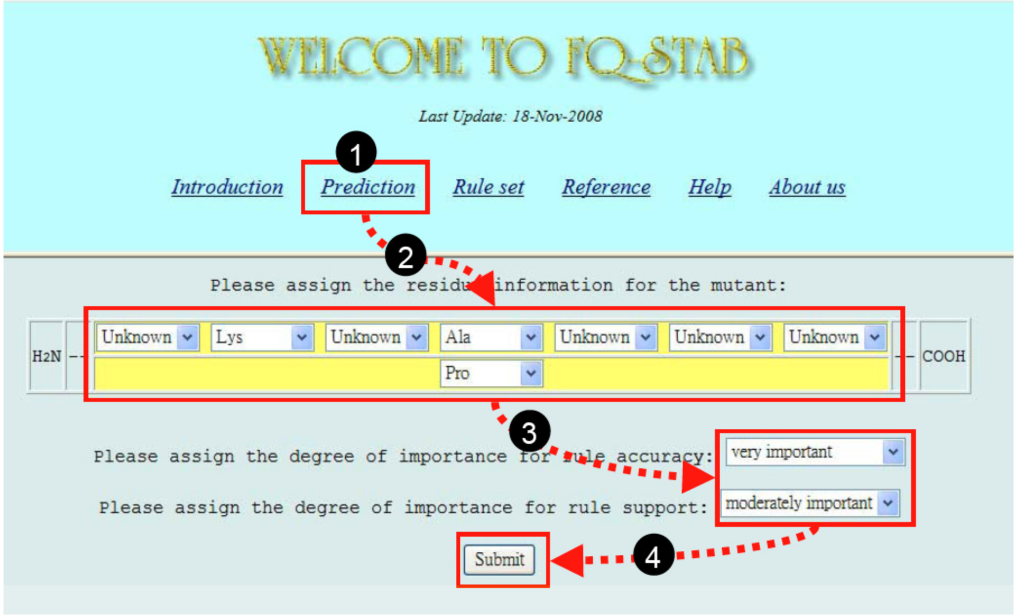
Readability and Flexibility

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



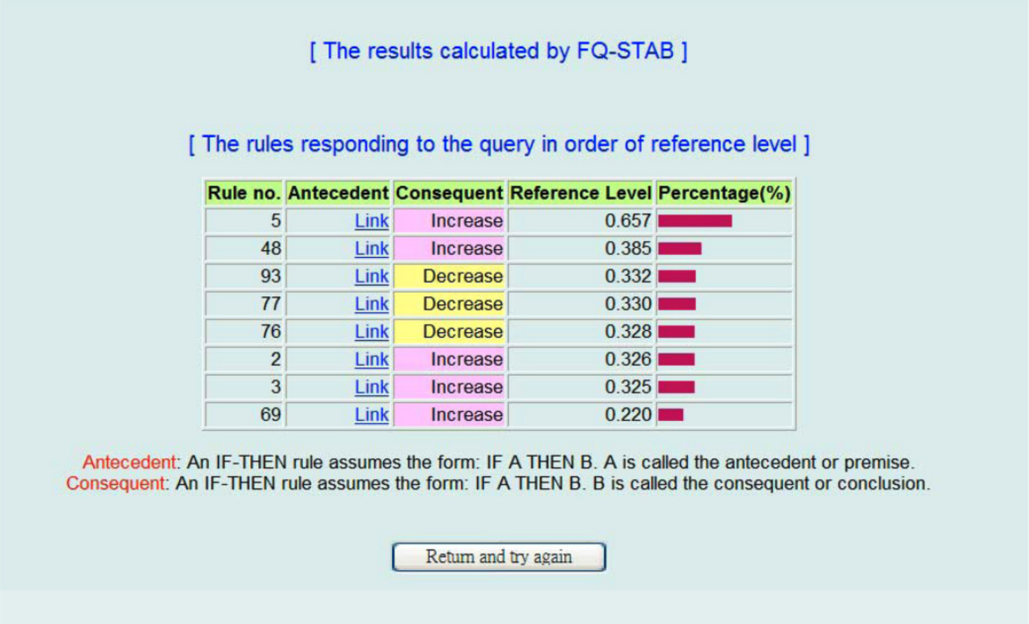
Please assign the residue information for the mutant:

Unknown | Lys | Unknown | Ala | Unknown | Unknown | Unknown
Pro

Please assign the degree of importance for rule accuracy: very important

Please assign the degree of importance for rule support: moderately important

Fig. (2). A snapshot of the prediction page with the input procedure for predicting stability changes.



[The results calculated by FQ-STAB]

[The rules responding to the query in order of reference level]

Rule no.	Antecedent	Consequent	Reference Level	Percentage(%)
5	Link	Increase	0.657	▬
48	Link	Increase	0.385	▬
93	Link	Decrease	0.332	▬
77	Link	Decrease	0.330	▬
76	Link	Decrease	0.328	▬
2	Link	Increase	0.326	▬
3	Link	Increase	0.325	▬
69	Link	Increase	0.220	▬

Antecedent: An IF-THEN rule assumes the form: IF A THEN B. A is called the antecedent or premise.
Consequent: An IF-THEN rule assumes the form: IF A THEN B. B is called the consequent or conclusion.

Fig. (3). The primary section of the FQ-STAB output along with a variety of rule summaries.

"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

[1] Prevost M, Wodak SJ, Tidor B, Karplus M. Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96----Ala mutation in barnase. Proc Natl Acad Sci USA 1991; 88(23): 10880-4.

[2] Lee C, Levitt M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature 1991; 352(6334): 448-51.

[3] Gilis D, Rooman M. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. J Mol Biol 1997; 272(2): 276-90.

[4] Zhou H, Zhou Y. Quantifying the effect of burial of amino acid residues on protein stability. Proteins 2004; 54(2): 315-22.

[5] Bordner AJ, Abagyan RA. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins 2004; 57(2): 400-13.

[6] Tan YH, Luo R. Protein stability prediction: a Poisson-Boltzmann approach. J Phys Chem B 2008; 112(6): 1875-83.

[7] Masso M, Vaisman, II. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics 2008; 24(18): 2002-9.

[8] Capriotti E, Fariselli P, Casadio R. A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 2004; 20 (Suppl 1): i63-8.

[9] Capriotti E, Fariselli P, Rossi I, Casadio R. A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics 2008; 9 (Suppl 2): S6.

[10] Huang L-T, Gromiha MM, Ho S-Y. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics 2007; 23(10): 1292-3.

[11] Huang L-T, Saraboji K, Ho S-Y, Hwang S-F, Ponnuswamy MN, Gromiha MM. Prediction of protein mutant stability using

classification and regression tool. Biophys Chem 2007; 125(2-3): 462-70.

[12] Parthiban V, Gromiha MM, Hoppe C, Schomburg D. Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. Proteins 2007; 66(1): 41-52.

[13] Zadeh LA. Soft computing and fuzzy logic. Software 1994; 11(6): 48-56.

[14] Zadeh LA. Fuzzy sets. Inf Control 1965; 8: 338-53.

[15] Torres A, Nieto JJ. Fuzzy logic in medicine and bioinformatics. J Biomed Biotechnol 2006; 2006(2): 91908.

[16] Gomez A, Cedano J, Espadaler J, Hermoso A, Pinol J, Querol E. Prediction of protein function improving sequence remote alignment search by a fuzzy logic algorithm. Protein J 2008; 27(2): 130-9.

[17] Franco-Lara E, Weuster-Botz D. Application of fuzzy-logic models for metabolic control analysis. J Theor Biol 2007; 245(3): 391-9.

[18] Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Sarai A. ProTherm: thermodynamic database for proteins and mutants. Nucleic Acids Res 1999; 27(1): 286-8.

[19] Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Res 2004; 32: D120-1.

[20] Huang L-T, Lai L-F, Gromiha MM. Human-readable rule generator for integrating amino acid sequence information and stability of mutant proteins. IEEE/ACM Trans Comput Biol Bioinform 2008 (Accept).

[21] Kaufmann A, Gupta MM. Introduction to Fuzzy Arithmetic: Theory and Applications. New York, NY: Van Nostrand Reinhold Co. 1985.

[22] Ngai EWT, Wat FKT. Fuzzy Decision Support System For Risk Analysis in e-Commerce Development: Elsevier Science 2005.

[23] Kao C, Liu S-T. Competitiveness of manufacturing firms: an application of fuzzy weighted average. Sys Man Cybernetics, Part A, IEEE Trans 1999; 29(6): 661-7.

[24] Guu S-M. Fuzzy weighted averages revisited. Fuzzy Sets Syst 2002; 126(3): 411-4.

[25] Chang PT, Hung KC, Lin KP, Chang CH. A comparison of discrete algorithms for fuzzy weighted average. Fuzzy Syst IEEE Trans 2006; 14(5): 663-75.

[26] Lai Y-J, Hwang CL. Fuzzy Mathematical Programming: Methods and Applications. Berlin; New York: Springer-Verlag 1992.

[27] Zimmermann HJ. Fuzzy Set Theory--and its Applications, 2nd ed. Boston: Kluwer Academic Publishers 1991.

[28] Tseng TY, Cerry MK. A new algorithm for fuzzy multicriteria decision making. Int J Approx Reasoning 1992; 6(1): 45-66.

[29] http://www.iit.nrc.ca/IR_public/fuzzy/fuzzyClips/ fuzzyCLIPSIndex.html

[30] Capriotti E, Fariselli P, Calabrese R, Casadio R. Predicting protein stability changes from sequences using support vector machines. Bioinformatics 2005; 21 (Suppl 2): ii54-8.

[31] Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 2006; 34(Web Server issue): W239-42.

Received: January 12, 2009 Revised: January 25, 2009 Accepted: February 2, 2009