Applied Mathematics & Information Sciences
An International Journal

Data Mining Application in Biomedical Informatics for Probing into Protein Stability upon Double Mutation

Liang-Tsung Huang[1,], Chao-Chin Wu[2], Lien-Fu Lai[2], M. Michael Gromiha[3], Chang-Sheng Wang[4] and Yet-Ran Chen[5]*

[1] Department of Biotechnology, Mingdao University, Changhua, Taiwan
[2] Department of Computer Science and Information Engineering, National Changhua University of Education, Changhua, Taiwan
[3] Department of Biotechnology, Indian Institute of Technology (IIT) Madras, Tamilnadu, India
[4] Department of Agronomy, National Chung Hsing University, Taichung, Taiwan
[5] Agricultural Biotechnology Research Center (ABRC), Academia Sinica, Taipei, Taiwan

Abstract: To explore the mechanism of protein stability change is one of the important topics in protein design. The accurate prediction of protein stability change upon mutation is very useful for enhancing the experimental efficiency in many biological and medical studies. In this work, we aimed at effectively introducing data mining technologies for investigating the understanding of protein stability change upon double mutation. We constructed a non-redundant dataset of protein mutants with various attributes and applied systematically analyses on the dataset. Therefore, we developed general knowledge from the dataset by several data mining techniques, including decision tree, decision table and association rule algorithms. Furthermore, we interpreted, evaluated, and compared those knowledge outcomes obtained from different techniques. The observations on the experimental results demonstrated that the present method may serve as an effective tool in biomedical informatics to understand the prediction of protein stability change upon double mutation.

Keywords: data mining, protein stability, double mutation, biomedical informatics

1 Introduction

The prediction of stability change for protein mutants is one of the important tasks in protein engineering. It is necessary for understanding the mechanisms responsible for protein stability as well as for designing temperature sensitive protein mutants [1].

The protein mutants have been frequently used to study the importance of the specific amino acid in the protein function and stability for biological and medical research [2]. For example, wild-type green fluorescent protein (wtGFP) has been modified to a variety of mutants with a single amino acid substitution. These GFP mutants may elicit altered spectral properties and display different colors of fluorescence [3][4][5].

For the past several years, many different methods have been developed to address this issue for predicting stability change upon single mutation [6][7][8][9]. Moreover, earlier studies revealed that the methods developed for predicting protein stability change upon point mutation are not suitable for predicting the stability of the mutants with double mutation [10][11]. Therefore, the prediction upon double or multiple mutations has attracted more and more attention [12][13]. However, it is necessary to provide complete information as input when employing these prediction methods. Whereas the input information may be incomplete and unavailable in practice, we have developed an effective approach to make reliable predictions from partial input information [14].

In this work, we aimed at effectively introducing data mining technologies for investigating the protein stability changes upon double mutation. We have constructed a non-redundant dataset of protein mutants with various attributes and applied systematically analyses on the dataset. Therefore, we have developed general knowledge from the dataset by several data mining techniques, including decision tree, decision table and association rule algorithms. Furthermore, we have interpreted, evaluated, and compared those knowledge outcomes

* Corresponding author e-mail: larry@mdu.edu.tw

obtained from different techniques. Knowledge outcomes obtained from various techniques were compared, interpreted and discussed.

2 Materials and Protein Mutants

2.1 Summary statistics of mutant dataset

We collected a set of double mutants from ProTherm database [15][16] in which the free energy changes of mutants were obtained with thermal denaturation. The final non-redundant dataset (named S180) consists of 180 mutants acquired from 27 different proteins. The $\Delta\Delta G$ ranges between -8.9kcal/mol and 6.2kcal/mol and the number of stabilizing and destabilizing mutants are 93 and 87, respectively. More details about the dataset are available at http://bioinformatics.myweb.hinet.net/ wetstab.htm [12].

2.2 Rule-based knowledge representation

The form of knowledge representation (KR) is critical to the utility of the information. There are different forms for specific applications. In this study, we employed a rule-based knowledge representation, the advantages of which include [17] : (i) Rules are easy to construct. (ii) Rules are quick to test. (iii) Rules provide intuitive interpretation for understanding.

Accordingly, we utilized two sets of attributes to represent the information of a double mutant. Each set refers to a mutation site including wild-type residue, mutant residue and three neighboring residues of the mutation site along both directions containing 8 attributes for each mutation site.

3 Methods and Performance Measurement

This study presented a knowledge discovery approach to effectively developing the knowledge in predicting protein stability change upon double mutation. Firstly, the dataset were systematically analyzed. Secondly, different data mining techniques were applied to the dataset. Finally, we interpreted, evaluated, and compared the mined knowledge.

For data mining, we implemented three kinds of rule induction algorithms, i.e. decision tree [18], decision table [19] and association rule [20] algorithms. The last one generated association rules and the others generated classification rules.

3.1 Decision tree algorithm

The PART algorithm [21] is based on partial decision trees [22] to generate accurate rules. It adopts the separate-and-conquer strategy which contained three main steps: (i) building a rule; (ii) removing the instances that it covers; and (iii) creating rules recursively for the remaining instances until none is left.

3.2 Decision table algorithm

In this study, the inducer of decision table majority algorithm [19] was implemented to build decision tables. A decision table consists of two components: a schema and a body. The schema is a set of attributes that are included in the table and the body is a list of instances with values defined by the attributes in the schema. The set of instances with the same values for attributes in the schema is named a cell and the cell can be presented by if-then rules.

Once a decision table is available, the outcome of a query instance can be predicted by the following criteria. Let L be the set of instances in the cell that matches the query instance over the attributes in the schema:

(i) If $L \neq \emptyset$, return the majority class of instances in L for the nominal class (or the average class for numeric classes.

(ii) Otherwise ($L = \emptyset$), return the majority (or average) class of instances in the decision table.

3.3 Association rule algorithm

The Apriori [20] is a classic algorithm for learning association rules. The algorithm generates the candidate itemsets by using only the itemsets found large in the previous pass. Therefore, the candidate itemsets having k items can be generated by joining large itemsets having k-1 items, and deleting those that contain any subset that is not large. This procedure results in generation of a much smaller number of candidate itemsets.

In this study, we implemented an altered version of association rule algorithm, PredictiveApriori algorithm [23], which proposes a trade-off between confidence and support and maximizes the chance of correct predictions on unseen data.

3.4 Performance measurement

In this study, various measure scores were used to evaluate the performance of the mined rules. Given a rule R: IF A THEN B, where A is the antecedent and B is the consequent, support is calculated by $SU_R = N_{A \cup B} / N *100\%$, where N is the number of total instances and $N_{A \cup B}$ is the number of instances with both antecedent A and consequent B to rule R. Similarly, confidence (CO_R) is calculated by $CO_R = N_{A \cup B} / N_A *100\%$, where N_A is the number of instances with only antecedent A of rule R. Support (SU_R) may quantify the coverage of the rule on a
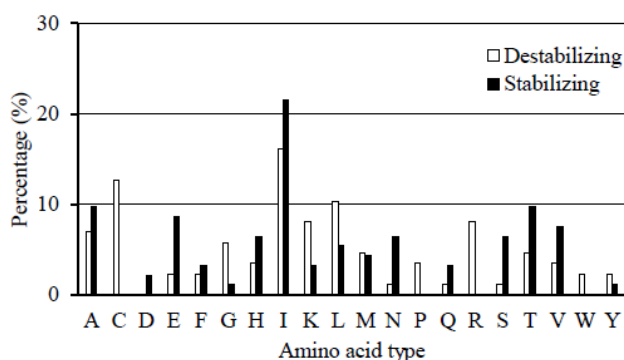
Fig. 1: Distribution of amino acid types for the wild residues at the mutation site near N-terminus.
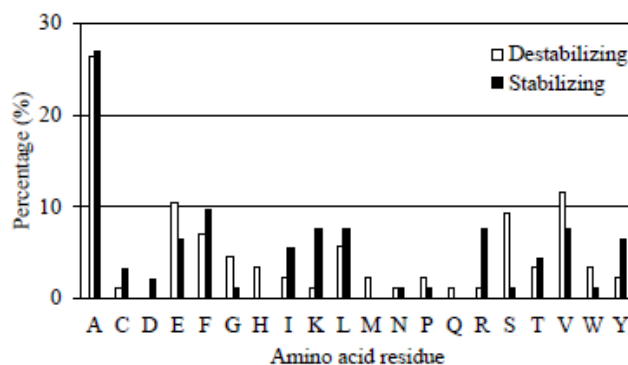

Fig. 2: Distribution of amino acid types for the mutant residues at the mutation site near N-terminus.

dataset and confidence (CO_R) may indicate the correctness of the rule to a dataset.

Furthermore, for evaluating the performance of the rule induction algorithm in predicting the stabilizing and destabilizing mutants, accuracy is calculated by the following expression: $AC=(TP+TN)/(TP+TN+FP+FN)*100\%$, where TP, TN, FP and FN refer to the number of true positives, true negatives, false positives and false negatives, respectively. Additionally, sensitivity (SE) indicating the correct prediction of stabilizing mutants and specificity (SP) about the destabilizing ones are calculated by $SE=TP/(TP+FN)*100\%$ and $SP=TN/(TN+FP)*100\%$, respectively.

4 Results & Discussions

4.1 Direct effects of wild and mutant residues to protein stability change

There are total 16 attributes with various properties on the dataset F180. Firstly, we mined F180 with two basic attributes, the wild and mutant residues. As shown in **Figure 1**, we analyzed the distribution of wild residues based on different amino acid types for the mutation site near N-terminus. The significant differences between stabilizing and destabilizing mutants appear at types C (Cysteine) with 12.6% and R (Arginine) with 8 %. The results showed the wild residue with type C or R seems to prefer destabilizing.

Further, we also analyzed the distribution of mutant residues. **Figure 2** exhibits the percentage difference between stabilizing and destabilizing mutants for different types of mutant residues. The results suggested that the mutant residue with type S (Serine) prefer destabilizing on the dataset F180.

4.2 Environmental effects of neighboring residues on improving knowledge discovery

From the above results, only two attributes were observed, which may be insufficient to precisely understand the stability change in double mutants. Therefore, we further mined the dataset by appending the attributes of neighboring residues to integrate the environmental effects from neighboring residues.

We employed decision table algorithm on S180 dataset and then mined total 60 rules. **Table 1** lists 4 reliable rules (2 for stabilizing and 2 for destabilizing) along with performance measures. An IF-THEN rule assumes the form: IF A THEN B, where A is called the antecedent or premise and B is the consequent or conclusion. N_j^i / C_j^i denotes the i-th residue along N- / C-terminus at mutation site j; W_j is the wild residue at mutation site j; and M_j is the mutant residue at mutation site j (i=1...2 and j=1...3 here). For the first rule, "W_2 is Ile" in the antecedent means that the wild residue at mutation site near the C-terminus is Isoleucine. "N_2^1 is Gly" means that the first residue along N-terminus at this mutation site is Glycine. Hence, "C_2^3 is Asp" means that the corresponding residue is Aspartate. The consequent indicated that the antecedent would lead to the conclusion of destabilizing.

Further, **Table 2** lists 4 rules with higher performance. Interestingly, the third rule, "IF N_2^3 is Lys and C_2^1 is Val THEN stabilizing" produced a higher SU (14.44%) and CO (96%).

The results showed that the rule induction algorithm mined some rules with high confidence from the dataset, and that the attributes of neighboring residues may increase the effectiveness of the knowledge discovery.

Table 1: Antecedent and consequent of the rules obtained from decision table algorithm.

Antecedent	C*	SU %	CO %
W_2 is Ile, N_2^1 is Gly, and C_2^3 is Asp	-	4.44	100
W_2 is Cys, N_2^1 is Asp, and C_2^3 is Thr	-	3.33	100
W_2 is Glu, N_2^1 is Lys, and C_2^3 is none	+	1.67	100
W_2 is Glu, N_2^1 is Gln, and C_2^3 is Ser	+	1.11	100

C*: Consequent; +: Stabilizing; -: Destabilizing; SU: Support; CO: Confidence.

Table 2: Antecedent and consequent of the rules derived from decision tree algorithm.

Antecedent	C*	SU %	CO %
N_2^2 is Glu	-	8.89	100
C_2^1 is Val	-	5.00	82
N_2^3 is Lys and C_2^1 is Val	+	14.44	96
N_2^2 is Lys	+	7.22	93

C*: Consequent; +: Stabilizing; -: Destabilizing; SU: Support; CO: Confidence.

4.3 Reliability validity of rule induction algorithms

Based on 10-fold cross-validation, the rules derived by decision table algorithm discriminate the stabilizing and destabilizing mutants with AC, SE, and SP of 82.2%, 76.3% and 88.5%, respectively [12]. For 5-fold, AC, SE and SP are 78.3%, 66.7% and 90.8%, respectively. In addition, we employed a back-propagation neural network with three layers, showing lower accuracy of 68.3%. The results showed that the induction algorithm may mine reliable rules from the protein mutant dataset.

On the other hand, we also tested the performance of decision tree algorithm. The results showed that the obtained AC, SE and SP were 70.0%, 62.4% and 78.2%, respectively.

4.4 Correctness and coverage analysis for mined rules

We further analyzed the rules mined by decision table algorithm. **Figure 3** exhibits the distribution of rules based on confidence. It showed that a large proportion of rules have high confidence and there are 83% of rules with confidence above 70%. Interestingly, 46 rules out of 60 rules achieved the confidence of 100%.

In addition, we observed the coverage of rules for the dataset. The support of the rule lies between 0.56% and 13.3% and the cumulative support of total rules reach 85%. For example, the first rule in **Table 1** shows support and confidence of 4.44% and 100%, respectively.
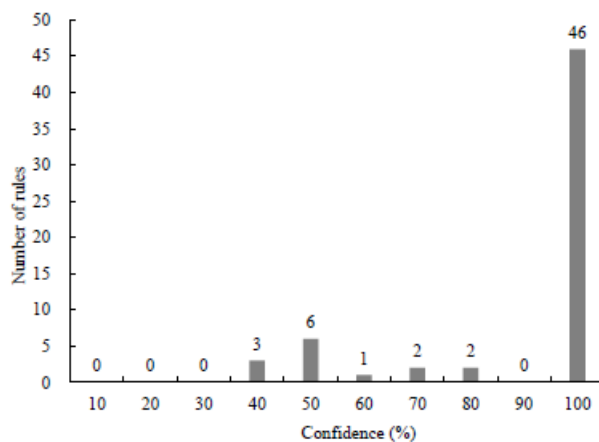
Fig. 3: Distribution of rules based on confidence.

Table 3: Summary of mined rules by different induction algorithms.

Rule set*	Number of attributes	Number of rules	Average support (%)	Average confidence (%)
RS_A	8	6	8.80	86
RS_B	8	60	1.42	89
RS_C	8	1	7.78	100

*Rule sets RS_A, RS_B and RS_C are obtained from decision tree, decision table, and association rule algorithms, respectively.

4.5 Characteristic comparison of rule induction algorithms

We further compared the characteristic of rule induction algorithms by analyzing three sets of rules obtained from different algorithms and calculated the overall scores for each rule set.

Firstly, our results showed that three rule sets were of slightly different rule length when using the same 8 attributes. The average lengths for rule sets RS_A, RS_B and RS_C were 2, 4 and 3, where rule sets RS_A, RS_B and RS_C were obtained from decision tree, decision table, and association rule algorithms, respectively.

Secondly, the numbers of three rule sets varied greatly from 1 to 60 as shown in **Table 3**. Interestingly, RS_A and RS_C showed less number of rules but higher SU (8.8 and 7.78, respectively) and CO (86% and 89%, respectively). The results indicated that the association and decision tree algorithms produced more general rules.

Finally, we analyzed the common rules among the three rule sets. The results showed that there was no identical rule found among the sets. Namely, the three kinds of rule induction algorithms may be complementary to this dataset.

Fig. 4: Snapshot of output page with predicted results.

Fig. 5: Snapshot of input page with related options.

4.6 The application of knowledge in discriminating stability change from partial sequence information

For demonstrating the application of mined rules, we developed a method which is based on the rule set to discriminate the stability change upon double mutation from partial sequence information [14].

The method evaluates both (i) the similarity between the rule antecedent and the input information, and (ii) the rule importance to rank the rules in order of the reference level (RL). The consequent of the rule with the highest reference level may provide the most possible solution.

The similarity is quantified by the degree of fit (DF). Assuming that we have N rules (e.g. $R_i, i=1 \ldots N$) with only two elements of the antecedent (e.g. $E_i^1, E_i^2, i=1 \ldots N$):

R_1: IF E_1^1 AND E_1^2 THEN C_1,

R_2: IF E_2^1 AND E_2^2 THEN $C_2, \ldots,$

R_N: IF E_N^1 AND E_N^2 THEN C_N,

where C_i is the consequent of rule R_i. If the elements of the known sequence information are given as set E, then the score of the degree of fit (DF) for rule R_i, is calculated by: $DF_i = \Phi / 2$, where Φ is the number of the intersection between sets E and $\{ E_i^1, E_i^2 \}$.

The importance of each rule was evaluated by the important level (IL) which is calculated from different measures of a rule. In this case, we have two measures (i.e. SU and CO) and users can weight the measures in terms of fuzzy numbers [24],[25] "unimportant", "rather unimportant", "moderately important", "rather important", and "very important". Then the importance of the rule is calculated by fuzzy weighted average [26][27][28] as the following expression:

$y = \sum_{j=1}^{n} w_j\, x_j / \sum_{j=1}^{n} w_j = \mathrm{f}(x_1,\ldots,x_n,w_1,\ldots,w_n)$,

where w_j (represented by fuzzy sets or fuzzy numbers) is the weight for the measure value x_j; n is the number of the measures.

Accordingly, the reference level of a rule R_i is calculated by: $\mathrm{RL}_i = \mathrm{DF}_i \times \mathrm{IL}_i$, namely, the multiplication of DF_i (the degree of fit between the input and rule R_i) and IL_i (the importance level of rule R_i).

We have also implemented a web service for discriminating stability change upon double mutation from partial sequence information using the proposed method (named TANDEM, predicting stability change upon double mutation). The web service is free to access at http://bioinformatics.myweb.hinet.net/tandem.htm. The input page for the service is shown in **Figure4**, which takes the information of wild-type, mutant and three neighboring residues for two mutation sites as input. Users may assign the known or unknown residue information on two mutation sites for a prediction. Further, users may also assign different intuitive terms (e.g. unimportant or very important) to weight each measure. The output page after users submit is shown in **Figure5**, which displays the information about the rule with the highest reference level along with the antecedent, consequent and reference level.

5 Conclusions and Future Works

In conclusion, we presented a knowledge discovery approach on predicting stability change for protein mutants. This approach integrates different data mining techniques to develop reliable rules, then evaluates and exams the knowledge.

Furthermore, we also demonstrated that the application of knowledge of mined rules by developing a fuzzy query method, which may discriminate protein stability change upon double mutation from partial sequence information.

The observations on the experimental results demonstrated that the present approach may serve as an effective tool in biomedical informatics for helping our understanding in predicting protein stability change upon double mutation. It is suggested that the present approach can be adopted to develop the potential knowledge and increase the understanding of predicting protein stability change upon double mutation.

Although this study has already outlined the framework of knowledge discovery in predicting protein stability change, it also brings several new future works. Firstly, more attributes including secondary structure, conservation score, and contacts may be included into the present approach for more analyses. Another promising future direction is to extend the knowledge discovery approach to other important issues in biology and medicine and more researches will benefit by this approach.

Acknowledgement

References

[1] M. M. Gromiha, Prediction of protein stability upon point mutations, Biochem Soc Trans, **35**, 1569-1573 (2007).

[2] R. F. Boyer, Concepts in biochemistry, 3rd ed., Wiley, Hoboken, NJ, (2006).

[3] J. D. Ropp, C. J. Donahue, D. Wolfgang-Kimball, J. J. Hooley, J. Y. Chin, R. A. Cuthbertson, K. D. Bauer, Aequorea green fluorescent protein: simultaneous analysis of wild-type and blue-fluorescing mutant by flow cytometry, Cytometry, **24**, 284-288 (1996).

[4] B. D. Moyer, J. Matteson, W. E. Balch, Expression of wild-type and mutant green fluorescent protein-Rab1 for fluorescence microscopy analysis, Methods Enzymol, **329**, 6-14 (2001).

[5] Y. M. Zheng, Z. X. An, X. E. Zhao, F. S. Quan, H. Y. Zhao, Y. R. Zhang, J. Liu, X. Y. He, X. N. He, Comparation of enhanced green fluorescent protein gene transfected and wild-type porcine neural stem cells, Res Vet Sci, **88**, 88-93 (2010).

[6] K. Saraboji, M. M. Gromiha, M. N. Ponnuswamy, Average assignment method for predicting the stability of protein mutants, Biopolymers, **82**, 80-92 (2006).

[7] L. T. Huang, M. M. Gromiha, S. Y. Ho, iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations, Bioinformatics, **23**, 1292-1293 (2007).

[8] A. Grosfils, Y. Dehouck, D. Gilis, M. Rooman, P. Bogaerts, Neural networks to predict protein stability changes upon mutation, in: Proceedings of the 17th International Federation of Automatic Control World Congress, International Federation of Automatic Control, Seoul, Korea, 12619-12624 (2008).

[9] M. Masso, I. I. Vaisman, Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis, Bioinformatics, **24**, 2002-2009 (2008).

[10] J. A. Wells, Additivity of mutational effects in proteins, Biochemistry, **29**, 8509-8517 (1990).

[11] A. Y. Istomin, M. M. Gromiha, O. K. Vorov, D. J. Jacobs, D. R. Livesay, New insight into long-range nonadditivity within protein double-mutant cycles, Proteins, **70**, 915-924 (2008).

[12] L. T. Huang, M. M. Gromiha, Reliable prediction of protein thermostability change upon double mutation from amino acid sequence, Bioinformatics, **25**, 2181-2187 (2009).

[13] L. Montanucci, P. Fariselli, P. L. Martelli, R. Casadio, Predicting protein thermostability changes from sequence upon multiple mutations, Bioinformatics, **24**, i190-195 (2008).

[14] L.-F. Lai, C.-C. Wu, L.-T. Huang, Predicting Protein Stability Change upon Double Mutation from Partial

Appl. Math. Inf. Sci. **8**, No. 1L, 125-132 (2014) / www.naturalspublishing.com/Journals.asp

131

Sequence Information Using Data Mining Approach, Advanced Intelligent Computing Theories and Applications, in: D.-S. Huang, Z. Zhao, V. Bevilacqua, J. Figueroa (Eds.), Springer Berlin / Heidelberg, 664-671 (2010).

[15] M. M. Gromiha, J. An, H. Kono, M. Oobatake, H. Uedaira, A. Sarai, ProTherm: Thermodynamic Database for Proteins and Mutants, Nucleic Acids Research, **27**, 286-288 (1999).

[16] M. D. Kumar, K. A. Bava, M. M. Gromiha, P. Prabakaran, K. Kitajima, H. Uedaira, A. Sarai, ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions, Nucleic Acids Research, **34**, D204-206 (2006).

[17] K. Morik, Knowledge representation and organization in machine learning, Springer-Verlag, Berlin ; New York, (1989).

[18] J. R. Quinlan, Induction of decision trees, Machine Learning, **1**, 81-106 (1986).

[19] R. Kohavi, The Power of Decision Tables, in: Proceedings of the 8th European Conference on Machine Learning, Springer-Verlag, 174-189 (1995).

[20] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, in: VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., 487-499 (1994).

[21] E. Frank, I. Witten, Generating accurate rule sets without global optimization, in: Proc. 15th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, 144-151 (1998).

[22] J. R. Quinlan, C4.5 : programs for machine learning, Morgan Kaufmann Publishers, San Mateo, Calif., (1993).

[23] T. Scheffer, Finding Association Rules That Trade Support Optimally against Confidence, Principles of Data Mining and Knowledge Discovery, in: L. De Raedt, A. Siebes (Eds.), Springer Berlin / Heidelberg, 424-435 (2001).

[24] A. Kaufmann, M. M. Gupta, Introduction to fuzzy arithmetic : theory and applications, Van Nostrand Reinhold Co., New York, N.Y., (1985).

[25] E. W. T. Ngai, F. K. T. Wat, Fuzzy decision support system for risk analysis in e-commerce development, Elsevier Science, (2005).

[26] C. Kao, S.-T. Liu, Competitiveness of manufacturing firms: an application of fuzzy weighted average, Systems, Man and Cybernetics, Part A, IEEE Transactions on, **29**, 661-667 (1999).

[27] S.-M. Guu, Fuzzy weighted averages revisited, Fuzzy Sets Syst., **126**, 411-414 (2002).

[28] P. T. Chang, K. C. Hung, K. P. Lin, C. H. Chang, A comparison of discrete algorithms for fuzzy weighted average, Fuzzy Systems, IEEE Transactions on, **14**, 663-675 (2006).

Liang-Tsung Huang is currently an Associate Professor in the Department of Biotechnology, Mingdao University, Taiwan. His main research interests include bioinformatics, medical informatics, data mining, and artificial intelligence. He has been selected for Who's Who in Science and Engineering, 11th Edition, Marquis, USA (2011-2012); and Man of the Year, International Biographical Centre, Cambridge, England (2012). More details are available at: http://www.mdu.edu.tw/ larry.

Chao-Chin Wu is an Associate Professor in the Department of Computer Science and Information Engineering, National Changhua University of Education, Taiwan. He received the M.S. and Ph.D. degrees in Computer Science and Information Engineering from National Chiao Tung University, Taiwan, in 1992 and 1998, respectively. His research interests are in parallel and distributed computing, cloud computing, grid computing, and intelligent systems.
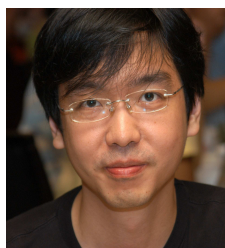
Lien-Fu Lai received his PhD in 1999 and MS in 1995, both in computer science from National Central University in Taiwan. He is currently an associate professor in the Department of Computer Science and Information Engineering at National Changhua University of Education in Taiwan. His research interests include fuzzy logic, knowledge engineering, web services, and search engines.

M. Michael Gromiha received his Ph.D in Physics from Bharathidasan University, Tiruchirappalli, India in 1994. He pursued his post doctoral research at the International Center for Genetic Engineering and Biotechnology (ICGEB), Trieste, Italy and The Institute of Physical and Chemical Research (RIKEN), Tsukuba, Japan. He worked as a Research Scientist/Senior Research Scientist at the Computational Biology Research Center (CBRC) of the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan from 2002-2010. Currently, he is serving as an Associate Professor at the Department of Biotechnology, Indian Institute of Technology Madras, India. His main research interests are structural analysis, prediction, folding and stability of globular and membrane proteins, protein interactions and development of databases and online tools in Bioinformatics. He has published over 130 research articles, 30 reviews and written a book on Protein Bioinformatics: From Sequence to Function. He is an Associate Editor of BMC Bioinformatics, Editor-in-Chief of Open Structural Biology Journal and a member of Nature Reader Panel. He is also a program committee member of ISMB, ECCB and InCoB.

Chang-Sheng Wang is an associate professor at the Department of Agronomy, National Chung-Hsing University, Taichung, Taiwan. He is working on rice mutagenesis and molecular breeding with the integration of technologies and information from genetics and genomics. He developed several rice mutant pools using sodium azide and studied its mechanism from phenotype screening to genotype identification to discover functional genes and markers for crop improvement.

Yet-Ran Chen received his PhD from Department of Chemistry, National Taiwan University in 2002. He is currently an assistant research fellow of Agricultural Biotechnology Research Center at Academia Sinica and assistant professor of Institute of Biotechnology at National Taiwan University. His research is focus on the development and application of mass spectrometry based approach for the elucidation of important plant and mammalian cellular signaling events.