

出國報告（出國類別：國際會議）

參加 2016 ICBBS
國際研討會心得報告

服務機關：國立高雄應用科技大學

姓名職稱：吳承翰

派赴國家：印尼, 峇里島

出國期間：2016.6.26-2015.7.4

報告日期：2015.6.26

摘要

本次出席 2016 5th International Conference on Bioinformatics and Biomedical Science (ICBBS 2016)，雖然僅僅五屆，但觀看研討會的品質及規模，在 Bioinformatics 及 Biomedical Science 收錄的論文品質及可看出成熟度。本次參與研討會，目的在會議當中聆聽許多不同領域專家，閱覽對人類有用的研究主題中發表的研究成果，並透過發表本次報告題目，與眾學者互相討論交流，找尋改進方向。

關鍵詞： Bioinformatics, Biomedical Science。

目次

一、目的.....	1
二、過程.....	2
三、心得及建議事項.....	7
附錄.....	8

一、目的

Conference General Co-Chairs

Prof. Orawan Siriratpiriya, Environmental Research Institute of Chulaongkorn University, Thailand

Prof. Tjokorda Gde Tirta Nindhia, Engineering Faculty, Udayana University, Bali, Indonesia

Prof. Helmut Zarbl, Rutgers, The State University of New Jersey, USA

Program Co-Chairs

Prof. Kevin Fong-Ray Liu, Ming Chi University of Technology, Taiwan

Prof. Manoj R. Tarambale, Marathwada Mitra Mandal's College of Engineering, Pune, India

Prof. Zuraini Ahmad, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Malaysia

Contact Chair

Ms. Flora Feng, Asia-Pacific Chemical, Biological & Environmental Engineering Society

Technical Committee

		
Prof. Dr. Guyeux Christophe University of Franche- Comté, France	Prof. Mark Segal University of California at San Francisco, USA	Dr. Christian Theil Have Copenhagen University, Denmark
		
Dr. I-Fang Cheng National Applied Research Laboratories, Taiwan	Dr. Bassam B. J. ALKINDY University of Franche-Comte, France	Dr. Yingqiu Xie Nazarbayev University, Kazakhstan

		
Dr. Fahmida Gulshan Bangladesh University of Engineering and Technology, Bangladesh	Dr. Momen Ahmad Orabi Cairo University, Egypt	Dr. Bo Hai Texas A&M Health Science Center, China
		
Dr. Rajendran Ananthan National Institute of Nutrition, India	Dr. Arjon Tumip Indonesian Institute of Sciences, Indonesia	Dr. Agnes Sri Harti School of Health Science, Surakarta, Indonesia

以上為研討會 chairs 及委員會成員，感謝他們為研討會的幫助，研討會才得以順利進行。本次前往印尼峇里島進行口頭報告，也因為此行結識了許多在生物資訊領域一同研究的夥伴，使用著新穎的技術提升人類的醫療品質，著實值得得到學術領域的鼓勵，在此感謝他們對學術上的付出，也讓我這趟研討會論文發表之旅增加了不少價值。

二、過程

本次投稿 2016 ICBBS 研討會，投稿題目為 Identification of SNP-SNP interaction using Entropy-based multifactor dimensionality reduction in Case-Control studies，主要目的在使用 entropy-based information gain 作為前處理工具，篩檢去除 gain 值低落的欄位，以減少 multifactor dimensionality reduction 所需要探索的欄位數。

本人於研討會中午才與會，參加的議程為 session 2, session 4，分別以 Biomedicine 及 Bioinformatics & Medical 為議題的論文題目，大會議程時間表如下所示。

Brief Schedule for Conferences

Day 1	Afternoon, June 25, 2016 (Saturday) Venue: Lobby Arrival Registration 13:30~17:00 (Committee Meeting 14:00~16:00)	
	June 26, 2016 (Sunday) 8:50~17:30 Venue: Gianyar Room & Klungkung Room Arrival Registration, Keynote Speech, and Conference Presentation	
Day 2	Morning Conference	
	Venue: Gianyar Room Opening Remarks 8:50~8:55 (Prof. Tjokorda Gde Tirta Nindhia, Engineering Faculty, Udayana University, Bali, Indonesia)	
	Keynote Speech I 8:55~9:30 Topic: "Sustainable Use and Zero Waste for Water Resources" (Prof. Orawan Siriratpiriya, Environmental Research Institute of Chulalongkorn University, Thailand)	
	Keynote Speech II 9:30~10:05 Topic: "Indonesian Wild Silkworm Cocoon as Biomaterial" (Prof. Tjokorda Gde Tirta Nindhia, Engineering Faculty, Udayana University, Bali, Indonesia)	
	Coffee Break & Photo Taking 10:05~10:40 Keynote Speech III 10:40~11:15 Topic: "Dietary Methylselenocysteine Prevents Mammary Carcinogenesis by Recoupling the Expression DNA Damage and Response Genes to the Circadian Clock" (Prof. Helmut Zarbl, Rutgers, The State University of New Jersey, USA)	
	Keynote Speech IV 11:15~11:50 Topic: "In Situ Arsenic Removal in Groundwater for Rural Communities by Iron Sorption and Arsenic Immobilization" (Prof. Solomon W. Leung, Environmental Engineering Civil and Environmental Engineering Department, Idaho State University)	
	Lunch 12:00~13:00 Venue: The Coffee Shop	
	Afternoon Conferences	
	Session 1: 13:00~15:00 Venue: Gianyar Room 8 presentations-Topic: "Food Science & Biochemistry"	Session 2: 13:00~15:00 Venue: Klungkung Room 8 presentations-Topic: "Biomedicine"
	Coffee Break 15:00~15:30	
Session 3: 15:30~17:30 Venue: Gianyar Room 8 presentations-Topic: "Environment"	Session 4: 15:30~17:30 Venue: Klungkung Room 8 presentations-Topic: "Bioinformatics & Medical"	
Dinner 17:40 Venue: The Coffee Shop		
Day 3	June 27, 2016 (Monday) 9:00~17:00 One Day Visit & Tour	

Tips: Please arrive at the conference room 10 minutes before the session begins to upload PPT into the laptop.

在聆聽這些題目時，發現大部分的生物醫學論文都指向應用甲殼素這類生物素材，具有良好的生物相容性、無毒性、強度佳等優點，但在 **biomedicine** 場次中，較多為使用統計模型計算，而學生使用的較多將生物問題轉為數學問題進而使用電腦演算法解決問題，故在聆聽之時體會較沒有採取演算法做研究時來得深刻，但這也代表著以後的研究可以以演算法為主，統計方法為副，以進行論文研討。

本次與同行的同學及學長們一起參與此次印尼國際研討會，報告時一同討論著相關議題，此程研討會議與本人以往所參加的研討會收穫更多，不僅在報告時台下與會學者提出更多具建設性的提問，也與學者們一同討論議題未來走向。

2016 / 06 / 27 為研討會參訪活動日，活動流程為下

1. Visit Turtle conservation at Serangan Island 09:00 - 11:00

http://www.wwf.or.id/en/about_wwf/whatwedo/marine_species/how_we_work/en_dangered_marine_species/tcec.cfm

2. Visit Udayana University (University hospital, Institute of peace and Democracy 11:00-12:00

Photo session in front of Rectorat Building

3. Lunch at Garuda Wisnu Kencana

https://en.wikipedia.org/wiki/Garuda_Wisnu_Kencana

4. Tour to Uluwatu Temple

https://en.wikipedia.org/wiki/Uluwatu_Temple

5. Dinner (farewell party) at Muaya Beach Jimbaran

Muaya Beach Cafe Area Jimbaran

活動照為下





三、心得及建議事項

本次出國因護照未更新護照使用期限，差點參加不了研討會，下次若還有機會參加研討會議，必須得小心打理所有出國相關物件，本次前往的國家為印尼，印度尼西亞地處於熱帶，整趟行程若處於非雨天或者沒有冷氣的空間，實在是難以習慣的氣後，另外出國到其他國家，希望大家不要把垃圾亂丟在野外，破壞了美景。

附錄

1. 本文
2. 報告 ppt

Identification of SNP-SNP interaction using Entropy-based multifactor dimensionality reduction in Case-Control studies

Cheng-Hong Yang

Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
Email: chyang@cc.kuas.edu.tw

Cheng-Han Wu

Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
Email: jackmel030@gmail.com

Li-Yeh Chuang

Department of Chemical Engineering, I-Shou University, Kaohsiung, Taiwan
Email: chuang@isu.edu.tw

Abstract—Diseases susceptibility plays an important role in genome-wide association study (GWAS). There are complex relationships between genotypes and environment factors in diseases. Due to the nonlinear relationship, the identification methods are met a challenge to detect gene-gene interaction or gene-environment interactions. In this study, entropy-based multifactor dimensionality reduction (EMDR) was used for identification of the single nucleotide polymorphisms (SNPs) interaction effects. MDR method is able to identify the interaction by trying n -locus interaction brute force. The proposed method uses K -way entropy based information gain as the filter for preprocessing, and then picks the suggested percentage of n -locus SNP combinations. Entropy-based interaction was compared with the searching way of MDR based on the ranking of interaction gain value. The Gametes simulation datasets were used to test the top percentage chosen for MDR, and the real kidney data was used to proof the ability of EMDR.

Index Terms—Entropy-based interaction Gain, SNP-SNP interaction, multifactor dimensionality reduction, Gametes

I. INTRODUCTION

Due to the huge amount of data produced from the next generation sequencing technologies (NGSTs), the genome wide association study (GWAS) becomes an important role in bioinformatics. In fact, an individual factor may not be the cause of a disease, but may lead to a high risk of disease susceptibility. To determine an individual is in the risk group (case) or in the low risk group (control) depends on the interactions between genetic and environment factors. Single nucleotide polymorphisms (SNPs), the common genetic variants between different human beings, have become the main elements to determine the particular disease susceptibility. According to the features of SNPs, GWAS is widely applied on the identification of gene-gene interaction or gene-environment interactions to determine the disease susceptibility [1, 2].

An efficiency method to identify the specific interactions from a huge amount of SNP or environment

factors is become the hot topic in GWAS. Most of the traditional statistical methods can only detect a SNP factor in linear relationship between genetic marker with disease [3, 4]. However, the relationship between SNP and environment interactions is nonlinear in a complex disease. Even a SNP has been identified to have small effect on the heritability of complex disease; the combination of several SNPs could be highly associated with the disease susceptibility. The high-dimension of SNP data remains the other challenge in the analyzing methods due to lack of computation effectiveness in traditional linear parametric methods.

To overcome the challenges of identification in SNP and environment factors interactions, many algorithms were proposed to conquer the problems, e.g., Genetic Algorithm (GA) [5], logic regression [6-8], polymorphism interaction analysis (PIA) [9] and multifactor dimensionality reduction (MDR) [10]. Here, the main structure of the proposed method is based on MDR algorithm. MDR classifies the high-dimension data into high risk group or low risk group by reorganizing the amount of SNP-SNP characterization in a "Cell", the base unit of MDR. The cell can be declared as the high risk group through the threshold of the cell. In this study, the process of MDR method was maintained, but the selection step of the variables number was discarded. The entropy-based MDR uses the interaction gain [11] as a preprocess of MDR. EMDR leaves a number of SNP combinations that doesn't pass the threshold of gain limit. In this way, we can save the computational time through dislodging the SNP combinations with low gain value.

In this study, two 2-order GAMETES [12] model and two 3-order GAMETES models were selected for testing the cut line, and each Gametes model has 2 sets of SNP number (50SNP, 100SNP), 3 sets of case-control number (Case: Control = 200: 200, Case: Control = 500: 500, Case: Control = 1000: 1000), and each combine setting has 100 experiments. In total, there are $4 \times 2 \times 3 \times 100 = 1200$ simulation data for testing the gain list cut line.

II. METHODS

A. Multifactor dimensionality reduction

The results of this study were obtained by the MDR method with cross-validation. Before the MDR processing, the number of polymorphisms was decided, and then started the MDR processing. In the first step, the data was divided into 9/10 training dataset and 1/10 test dataset (for a 10 fold cross-validation); the data should be random shuffle the order of samples in the whole data by a random seed. The random shuffle process is able to avoid the case or control samples concentrated to a certain subset or area. Even though the data is concentrated to a certain subset after the random shuffle process, the data division can be reorganized through several times of cross-validation by using different seed to random shuffle the sample order in dataset. Hence, the final results turn to avoid the particular situation. In the second step, the number of a polymorphisms set was selected from the beginning. In the third step, the number of different class (case and control) in a polymorphisms set with different genotype was calculated. Then, the ratio of case number to control number in whole data was used as the threshold. If the ratio of the cell is higher than the threshold, then the cell is described as the high risk group. For example, assume the MDR cell is 2 polymorphisms and there are 3 available genotypes in each polymorphism. Then, the MDR cell has 9 sub cells in it as shown in Fig. 1. Each sub cell was calculated to obtain a ratio of case to control. Then, based on the ratio value, the sub cell could be described as high risk group if the ratio is met or exceed the threshold.

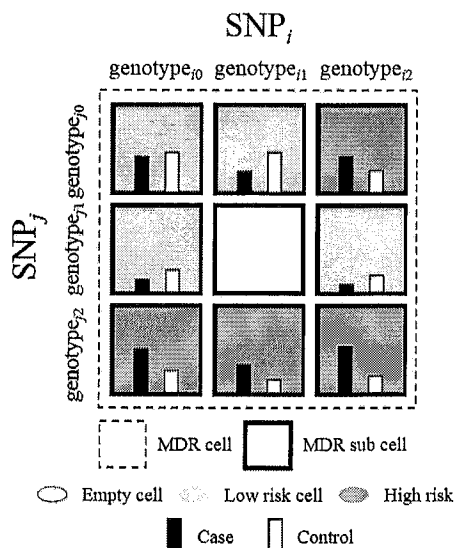


Figure 1. The description of a base MDR unit.

In MDR, the error rate is the evaluation value of the SNP-SNP interaction. The error rate is according to the contingency table that is divided into 4 groups. The sample in case group and with the high risk cell is labeled as True Positive group (TP). The sample in control group and with the high risk cell is labeled as False Positive group (FP). The sample in control group and with the low

risk cell is labeled as True Negative group (TN). The sample in case group and with the low risk cell is labeled as False Negative group (FN).

The error rate is calculated to determine the prediction ability of a filter. Therefore, the error rate would be a ratio of false prediction amount to all sample amount. The accuracy of training data sets and test data sets is calculated as the equation (1) by the contingency table of SNP problem in Table I.

Table I. THE CONTINGENCY TABLE IN SNP PROBLEM

	High risk cell	Low risk cell
Case ^t	TP	FN
Control	FP	TN

$$Error\ rate(X_i) = \frac{FP_i + FN_i}{TP_i + FP_i + FN_i + TN_i} \quad (1)$$

B. Entropy-based interaction gain

1) Definitions

The entropy is the value measuring the uncertainty associated with a random variable or a random system. The entropy $H(X)$ of a discrete random variable X is defined as equation (2):

$$H(X) = -E[\log P(X)] = -\sum_{x \in N} p(x) \log p(x) \quad (2)$$

The disease status of an individual is denoted by D . If $D = 0$, the individual is labeled as control. If $D = 1$, the individual is labeled as case. To identify the interactions between two di-allelic markers, A, B and C (SNP_A , SNP_B , and SNP_C), the genotypes of SNP_A , SNP_B and SNP_C are denoted as G_A , G_B and G_C , shown as the equation (3) below.

$$G_A = \begin{Bmatrix} 2 & AA \\ 1 & Aa \\ 0 & aa \end{Bmatrix}, G_B = \begin{Bmatrix} 2 & BB \\ 1 & Bb \\ 0 & bb \end{Bmatrix}, G_C = \begin{Bmatrix} 2 & CC \\ 1 & Cc \\ 0 & cc \end{Bmatrix} \quad (3)$$

2) Interaction gain

In the literature, genetic markers and environmental factors are treated as attributes. Using the entropy definition (1), we can define the entropy $H(A)$ of marker A in the general population and the conditional entropy $H(A|D)$ in the affected population as equation (4) and (5).

$$H(A) = -\sum_{i=0}^{genotype} P(G_A = i) \log P(G_A = i) \quad (4)$$

$$H(A|D) = -\sum_{i=0}^{genotype} P(G_A = i | D = case) \log P(G_A = i | D = case) \quad (5)$$

3) Two way interaction gain

The mutual information measures the interaction between two markers. In the general population, the mutual information of markers A and B , $I(A, B)$, is defined as

$$I(A, B) = H(A) + H(B) - H(A, B) \\ = - \sum_{i=0}^2 \sum_{j=0}^2 P(G_A = i, G_B = j) \log \frac{P(G_A = i, G_B = j)}{P(G_A = i)P(G_B = j)} \quad (6)$$

In the affected population, the mutual information of markers A and B is defined as

$$I(A, B | D) = H(A | D) + H(B | D) - H(A, B | D) \\ = - \sum_{i=0}^2 \sum_{j=0}^2 P(G_A = i, G_B = j | D = \text{cas}) \log \frac{P(G_A = i, G_B = j | D = \text{cas})}{P(G_A = i | D = \text{cas})P(G_B = j | D = \text{cas})} \quad (7)$$

The information gain of markers A and B in the presence of a disease can be defined as the difference between the mutual information in the affected population and that in the general population [13].

$$IG(AB | D) = I(A, B | D) - I(A, B) \quad (8)$$

4) Three way interaction gain

In the general population, we denote the joint genotype probabilities for markers A , B and C by $P_{ijc} = P(G_A = i, G_B = j, G_C = c)$. In the affected population, we denote the joint conditional genotype probabilities by $Q_{ijc} = P(G_A = i, G_B = j, G_C = c | D = 1)$.

Denote $P_{i\cdot} = \sum_{j=0}^2 \sum_{c=0}^2 P_{ijc}$, $P_{\cdot j} = \sum_{i=0}^2 \sum_{c=0}^2 P_{ijc}$, $P_{\cdot\cdot c} = \sum_{i=0}^2 \sum_{j=0}^2 P_{ijc}$. Similarly, $Q_{i\cdot}$, $Q_{\cdot j}$, and $Q_{\cdot\cdot c}$ can be defined in a similar manner. Denote $P_{ij\cdot} = \sum_{c=0}^2 P_{ijc}$, $P_{\cdot j\cdot} = \sum_{i=0}^2 P_{ijc}$, $P_{i\cdot c} = \sum_{j=0}^2 P_{ijc}$. Similarly, $Q_{i\cdot}$, $Q_{\cdot j}$, and $Q_{\cdot\cdot c}$ can be defined in a similar manner.

$$f_{ijc} = P_{ijc} \log \frac{P_{ijc}}{P_{i\cdot} P_{\cdot j} P_{\cdot\cdot c}}, g_{ijc} = Q_{ijc} \log \frac{Q_{ijc}}{Q_{i\cdot} Q_{\cdot j} Q_{\cdot\cdot c}} \quad (9)$$

Denote $P_{ij\cdot} = \sum_{c=0}^2 P_{ijc}$, $P_{\cdot j\cdot} = \sum_{i=0}^2 P_{ijc}$, $P_{i\cdot c} = \sum_{j=0}^2 P_{ijc}$. Similarly, $Q_{i\cdot}$, $Q_{\cdot j}$, and $Q_{\cdot\cdot c}$ can be defined in a similar manner.

$$h_{ijc} = P_{ijc} \log \frac{P_{ijc} P_{i\cdot} P_{\cdot j} P_{\cdot\cdot c}}{P_{ij\cdot} P_{\cdot j\cdot} P_{i\cdot c}}, l_{ijc} = Q_{ijc} \log \frac{Q_{ijc} Q_{i\cdot} Q_{\cdot j} Q_{\cdot\cdot c}}{Q_{ij\cdot} Q_{\cdot j\cdot} Q_{i\cdot c}} \quad (10)$$

The interaction information gain A , B and C would be denoted as

$$IIG(ABC | D) = I(A, B, C | D) - I(A, B, C) \\ = \sum_{i=0}^2 \sum_{j=0}^2 \sum_{c=0}^2 l_{ijc} - \sum_{i=0}^2 \sum_{j=0}^2 \sum_{c=0}^2 h_{ijc} \quad (11)$$

C. Entropy-based MDR

The entropy-based MDR uses gain value of interaction information as the suggestion. The suggestion is a

ranking list in order. The SNP sets with a higher gain value would be put at the front of the list. The ranking list is similar as a pool that is feeded with the polymorphisms sets into the MDR process in order. In the way, we can reduce the redundancy SNP sets to join the MDR process for saving the computational time meaningfully. The pseudocode is show as Table II.

Table II. EMDR PSEUDO CODE

1. **Interaction gain phase:**
2. **For** $S = 1$ to the last SNPs combination.
3. Calculating the interaction gain value by n way interaction gain.
4. **End** S
5. **End** interaction gain phase
6. **MDR phase:**
7. Divide data into 10 subsets randomly.
8. **For** $D = 1$ to 10 subsets
9. **Training data:**
10. **For** $S = 1$ to the outline of interaction gain
11. **For** $C = 1$ to all combination of genotypes
12. Determine the high/low risk groups in C MDR sub-cell.
13. **End** C
14. Compute the error rate of S SNP combination.
15. **End** S
16. Choose the best combination with the least error rate.
17. **End** training data
18. **Test data:**
19. Compute the best combination in the test data.
20. **End** test data
21. Collect the best combination into consistency set.
22. **End** D
23. Compute cross-validation consistency from consistency set.
24. Choose the best combination with the least error rate in test data.
25. **End** MDR phase

III. RESULTS AND DISCUSSION

A. Results

1) GAMETES datasets

GAMETES is a tool for generating 2-locus, 3-locus models with random architectures. GAMETES is focused on generating the lower heritability models that typically used in simulation studies. In the case, the extremely strict models can be tried to evaluate the identification algorithms for SNP interaction. There are two 2-locus GAMETES models and two 3-locus GAMETES models were used in this study.

Two of the 2-locus models are the models with marginal effects. Junghyun Namkung et al. developed Models 1 (Table III) and 2 (Table IV) [14] by varying the strength of genetic effects while fixing the interaction

structure, the minor allele frequency (MAF) and prevalence.

Table III. PREVALENCE VALUES OF 2-LOCUS MODEL 1.

	AA	Aa	aa
BB	0.060	0.010	0.010
Bb	0.010	0.208	0.208
bb	0.010	0.208	0.208

Table IV. PREVALENCE VALUES OF 2-LOCUS MODEL 2.

	AA	Aa	aa
BB	0.061	0.017	0.017
Bb	0.017	0.136	0.136
bb	0.017	0.136	0.136

Two of the 3-locus models are called XOR model [15] (Table V) and ZZ model [16, 17] (Table VI), respectively. XOR model is a nonlinear epistasis model, and the high risk of disease is dependent on inheriting a heterozygous genotype from one locus or a heterozygous genotype from another locus, but not all loci. In ZZ model, the high risk of disease is dependent upon inheriting exactly two high risk alleles from two loci. The simulation data was set at 50 or 100 SNP number and 400, 1000 or 2000 sample size for each model.

Table V. PENETRANCE VALUES FOR COMBINATIONS OF GENOTYPES FROM THREE SNPs EXHIBITING INTERACTIONS IN THE ABSENCE OF INDEPENDENT MAIN EFFECTS. (XOR MODEL)

	CC			Cc			cc		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
BB	0.4	0.9	0.7	0.2	0.2	0.6	1.0	0.4	0.5
Bb	0.9	0.0	0.9	0.6	0.9	0.0	0.3	0.1	0.6
bb	0.1	0.2	0.6	0.3	0.6	0.3	0.3	0.9	1.0

Table VI. PENETRANCE VALUES FOR COMBINATIONS OF GENOTYPES FROM TWO GENES EXHIBITING INTERACTIONS TO 3-LOCUS EXTENSION. (ZZ MODEL)

	CC			Cc			cc		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
BB	0	0	1	1	0	1	1	0	1
Bb	0	0	0	0	0.5	0	0	0	0
bb	1	0	1	1	0	1	1	0	0

2) Cut line

EMDR uses MDR as a based algorithm and entropy-based interaction gain as preprocess. The ranking list lists the gain values in order. Only a percentage of data from the ranking list is sent into the MDR for figuring out the best training model. Therefore, the definition of cutline is important. A good cutline of ranking list can abandon the SNP combinations with lower gain values into the SNP combination pool of MDR, but still maintains the high accuracy as the standard MDR.

Here, the cutline of ranking list is discussed by detecting the ranking of SNP combinations in the simulation data. As shown in Fig. 2, the ranking of target SNP combinations revealed lower than 50 percentage of

the SNP combinations in the selection pool except that of the 3-locus of ZZ model.

The obtained gain values are shown in Fig. 3. The gain values obtained from 2-locus models were close to 0. However, the gain values of 3-locus models were increased and higher than 2-locus models. In addition, no matter the SNP number is getting higher, the average gain value is kept at the same level in same model.

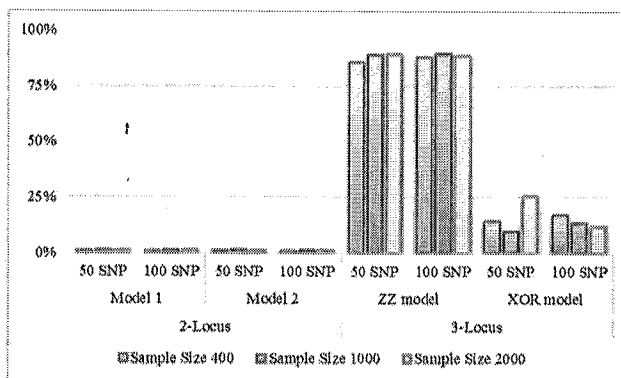


Figure 2. The ranking of two 2-locus marginal effect model and 3-locus ZZ, XOR model with different SNP and sample size.

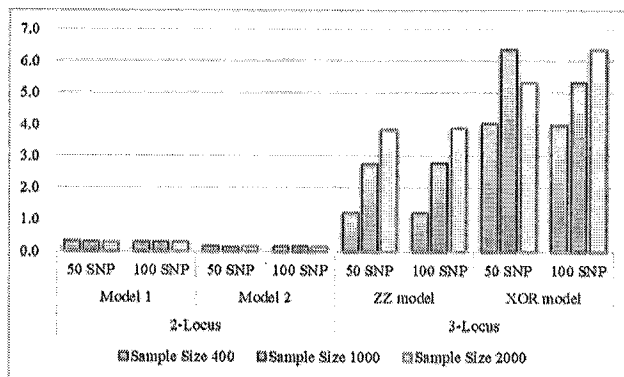


Figure 3. The gain value of two 2-locus marginal effect model and 3-locus ZZ, XOR model with different SNP and sample size.

3) Comparison of MDR and EMDR

Fig. 4 - 7 show the identification accuracy of MDR and EMDR in different GAMETES simulation data sets (symbol A to F). Symbol A is 50 SNPs with 400 sample size. Symbol B is 50 SNPs with 1000 sample size. Symbol C is 50 SNPs with 1000 sample size. Symbol D is 100 SNPs with 400 sample size. Symbol E is 100 SNPs with 1000 sample size. Symbol F is 100 SNPs with 2000 sample size. As shown in Fig. 4, the accuracy line of MDR and EMDR are same in the 2-locus model 1. In the 2-locus model 2, the accuracy line of MDR and EMDR become different (Fig. 5); EMDR method showed higher accuracy than MDR for the GAMETES data sets A and D. However, the MDR method revealed higher accuracy than EMDR in 3-locus ZZ model for all of the GAMETES data sets (Fig. 6). Same as the 2-locus model 1, the accuracy line of MDR and EMDR are same in the 3-locus XOR model (Fig. 7).

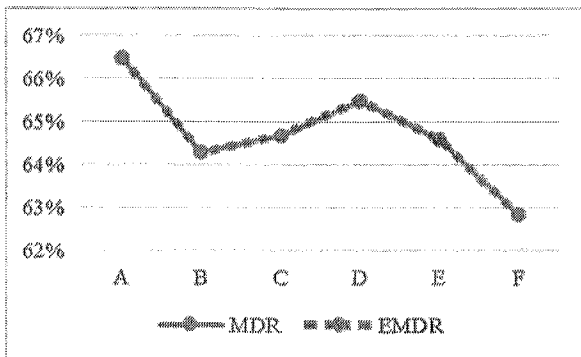


Figure 4. The identification accuracy of MDR and EMDR in different GAMETES simulation data setting. (2-locus model1)

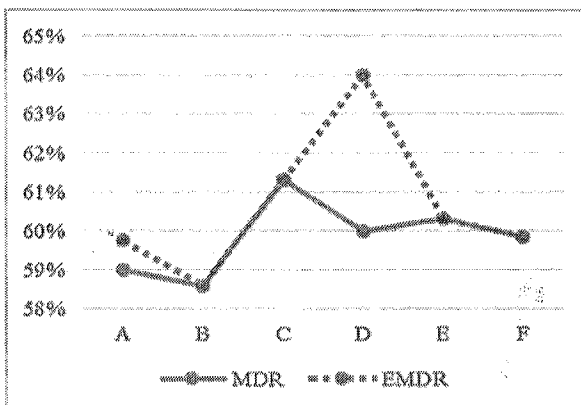


Figure 5. The identification accuracy of MDR and EMDR in different GAMETES simulation data setting. (2-locus model2)

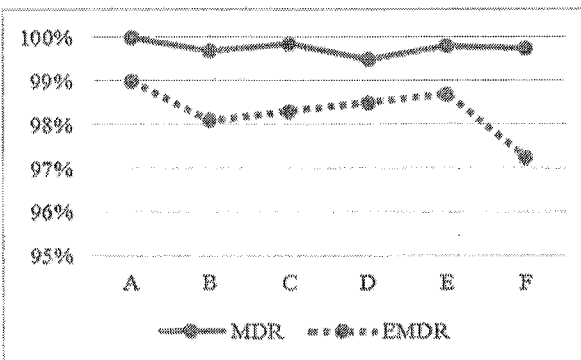


Figure 6. The identification accuracy of MDR and EMDR in different GAMETES simulation data setting. (3-locus ZZ model)

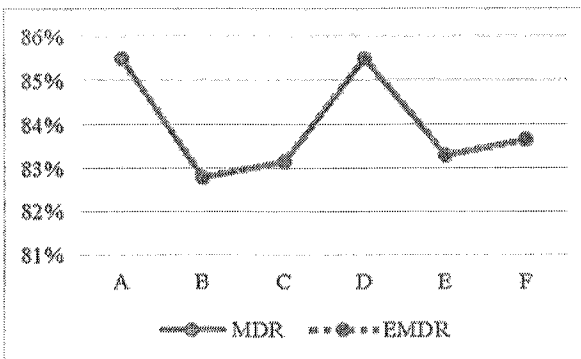


Figure 7. The identification accuracy of MDR and EMDR in different GAMETES simulation data setting. (3-locus XOR model)

B. Discussion

In this paper, the ability of EMDR and MDR is evaluated for the identification of SNPs interaction. In the experiment of interaction gain ranking list, the most interesting part is the ZZ model in the whole experiment. The ranking percentage of target SNP combinations is significant high, even higher than 75%. The reason might be due to inexistence of the main effect in the ZZ model, and the other 3 models in 2-locus or 3-locus might exist more or less of the main effects. Thus, we hypothesize the existence of main effect may influence the information gain value of the combination items.

In the accuracy test experiment, we compared the identification accuracy between MDR and EMDR methods. Before the experiment, we assumed that if the target SNP combination didn't get into the MDR selection pool, the EMDR won't select the correct SNP combination, and the accuracy of EMDR would be lower than MDR. In Fig. 4 and Fig. 7, the MDR performance is exactly same as EMDR. That is because of the target SNP combination is in the gain ranking list, and MDR is not lose control. The MDR method randomly shuffles the order of samples in the whole data to avoid the unbalance sub-data in any fold. If the unbalance sub-data exist, the MDR would made a mistake to choose the SNP combination with a lower interaction. As shown in Fig. 5, comparison of the identification accuracy of MDR and EMDR, the EMDR can keep the stability of accuracy for all of the test data sets. The superior performance of EMDR is due to EMDR removed the redundancy pair of SNP combination. Without the perturbation of redundancy pair of SNP combination, EMDR can easily keep the high accuracy even the fold is unbalance. In Fig. 6, EMDR showed worse performance than MDR. That might be caused by the interaction information gain ranking list. The list missed the useful SNP combinations due to divide the wrong group among the gain value of SNP combinations. EMDR was not able to search the correct SNP combination in the selection pool. Thus, EMDR obtained lower accuracy than MDR in the 3-locus of ZZ model.

IV. CONCLUSION

The method of multifactor dimensionality reduction (MDR), a nonparametric method, is a good tool that divides high dimensional data into one dimension. MDR plays an important role on the identification of gene interaction. However, there is a drawback for MDR due to its unbalance folding. With the unbalance folding of MDR, the lower accuracy or missing selection is come behind.

EMDR is proposed to solve the problem of unbalance folding. The results indicated that EMDR provided a better identification while the important SNP combinations were in the gain ranking list for interaction information. EMDR is able to use the interaction information gain list to pool the SNP combinations to target group and remove the redundancy group meaningfully. With abandoning the redundancy group of SNP combinations, EMDR would save the computational

time on identification of the interaction of SNP combinations.

However, based on the current study results, the entropy-based interaction gain can't correctly pool the SNP combinations to the target group and remove the redundancy group. That would make EMDR never select the target SNP combination, instead select the close solution. We expect raising the precision of interaction information gain can be proposed in the future work.

ACKNOWLEDGMENT

This study was partly supported by the National Science Council of Taiwan for Grant NSC 103-2221-E-151-029-MY3.

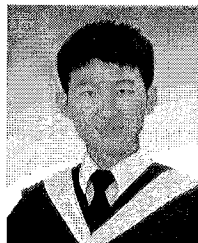
REFERENCES

- [1] P. Li, M. Guo, C. Wang, X. Liu, and Q. Zou, "An overview of SNP interactions in genome-wide association studies," *Briefings in Functional Genomics*, p. elu036, 2014.
- [2] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, vol. 26, pp. 445-455, 2010.
- [3] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, pp. 781-791, 2006.
- [4] X. Wu, H. Dong, L. Luo, Y. Zhu, G. Peng, J. D. Reville, et al., "A novel statistic for genome-wide interaction analysis," *PLoS Genet*, vol. 6, p. e1001131, 2010.
- [5] H.-W. Chang, L.-Y. Chuang, C.-H. Ho, P.-L. Chang, and C.-H. Yang, "Odds ratio-based genetic algorithms for generating SNP barcodes of genotypes to predict disease susceptibility," *OMICS A Journal of Integrative Biology*, vol. 12, pp. 71-81, 2008.
- [6] I. Dinu, S. Mahasirimongkol, Q. Liu, H. Yanai, N. S. Eldin, E. Kreiter, et al., "SNP-SNP interactions discovered by logic regression explain Crohn's disease genetics," *PLoS one*, vol. 7, p. e43035, 2012.
- [7] S. Prabhu and I. Pe'er, "Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease," *Genome research*, vol. 22, pp. 2230-2240, 2012.
- [8] M. Li, S. W. Erickson, C. A. Hobbs, J. Li, X. Tang, T. G. Nick, et al., "Detecting Maternal - Fetal Genotype Interactions Associated With Conotruncal Heart Defects: A Haplotype - Based Analysis With Penalized Logistic Regression," *Genetic epidemiology*, vol. 38, pp. 198-208, 2014.
- [9] L. E. Mechanic, B. T. Luke, J. E. Goodman, S. J. Chanock, and C. C. Harris, "Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions," *BMC bioinformatics*, vol. 9, p. 1, 2008.
- [10] L. W. Hahn, M. D. Ritchie, and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics*, vol. 19, pp. 376-382, 2003.
- [11] R. Fan, M. Zhong, S. Wang, Y. Zhang, A. Andrew, M. Karagas, et al., "Entropy - based information gain approaches to detect and to characterize gene - gene and gene - environment interactions/correlations of complex diseases," *Genetic epidemiology*, vol. 35, pp. 706-721, 2011.
- [12] R. J. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. M. Fisher, and J. H. Moore, "GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures," *BioData mining*, vol. 5, p. 1, 2012.
- [13] J. H. Moore, J. C. Gilbert, C.-T. Tsai, F.-T. Chiang, T. Holden, N. Barney, et al., "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *Journal of theoretical biology*, vol. 241, pp. 252-261, 2006.
- [14] R. J. Urbanowicz, A. L. Granizo-Mackenzie, J. Kiralis, and J. H. Moore, "A classification and characterization of two-locus, pure, strict, epistatic models for simulation and detection," *BioData mining*, vol. 7, p. 1, 2014.
- [15] J. H. Moore, L. W. Hahn, M. D. Ritchie, T. A. Thornton, and B. C. White, "Routine discovery of complex genetic models using genetic algorithms," *Applied soft computing*, vol. 4, pp. 79-86, 2004.
- [16] J. H. Moore, L. W. Hahn, M. D. Ritchie, T. A. Thornton, and B. C. White, "Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics," in *Proceedings of the Genetic and Evolutionary Computation Conference/GECCO. Genetic and Evolutionary Computation Conference, 2002*, p. 1150.
- [17] W. N. Frankel and N. J. Schork, "Who's afraid of epistasis?," *Nature genetics*, vol. 14, pp. 371-373, 1996.

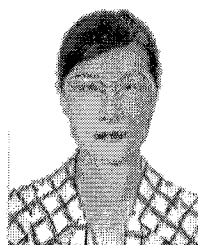
Cheng-Hong Yang is a professor of the Department of Electronic Engineering at National Kaohsiung University of Applied Sciences, Taiwan. He received his M.S. and Ph.D. degrees in computer engineering from North Dakota State University in 1988 and 1992, respectively. His main areas of research are evolutionary computation, bioinformatics, and assistive tool implementation.



Cheng-Han Wu is a master student of the Department of Electronic Engineering at National Kaohsiung University of Applied Sciences, Taiwan. He received his bachelor degree from the Department of Computer Science and Information Engineering at National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan. His research interests include evolutionary computation, data mining, machine learning, and bioinformatics.



Li-Yeh Chuang is a professor and director of the Department of Chemical Engineering & Institute of Biotechnology and Chemical Engineering at I-Shou University, Kaohsiung, Taiwan. She received her M.S. degree from the Department of Chemistry at the University of North Carolina in 1989 and her Ph.D. degree from the Department of Biochemistry at North Dakota State University in 1994. Her main areas of research are Bioinformatics, Biochemistry and Genetic Engineering.



**Identification of SNP-SNP
 interaction using Entropy-based
 multifactor dimensionality reduction
 in Case-Control studies**

Reporter: Cheng-Han Wu

Outline

- ▶ Introduction
- ▶ Method
- ▶ Results
- ▶ Discussion
- ▶ Conclusion

Introduction
*Single nucleotide polymorphisms (SNP), interaction factors, entropy based
 interaction gain,*

Single nucleotide polymorphisms (SNP)

- ▶ Single nucleotide polymorphisms (SNPs), the common genetic variants between different human beings, have become the main elements to determine the particular disease susceptibility.

Single nucleotide polymorphisms (SNP)

- ▶ According to the features of SNPs, GWAS is widely applied on the identification of gene-gene interaction or gene-environment interactions to determine the disease susceptibility [1, 2].

[1] P. Li, M. Guo, C. Wang, X. Liu, and Q. Zou, "An overview of SNP interactions in genome-wide association studies," *Briefings in Functional Genomics*, p. eht036, 2014.

[2] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, vol. 26, pp. 445-455, 2010.

Current challenges

- ▶ Most of the traditional statistical methods can only detect a SNP factor in linear relationship between genetic marker with disease [3, 4].
- ▶ The relationship between SNP and environment interactions is nonlinear in a complex disease.

[3] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, pp. 781-791, 2006.

[4] X. Wu, H. Dong, L. Luo, Y. Zhu, G. Peng, J. D. Revell, et al., "A novel statistic for genome-wide interactions analysis," *PLoS Genet*, vol. 6, p. e1001131, 2010.

Proposed methods

- ▶ To overcome the challenges of identification in SNP and environment factors interactions, many algorithms were proposed to conquer the problems, e.g., Genetic Algorithm (GA) [5], logic regression [6-8], polymorphism interaction analysis (PIA) [9] and multifactor dimensionality reduction (MDR) [10].

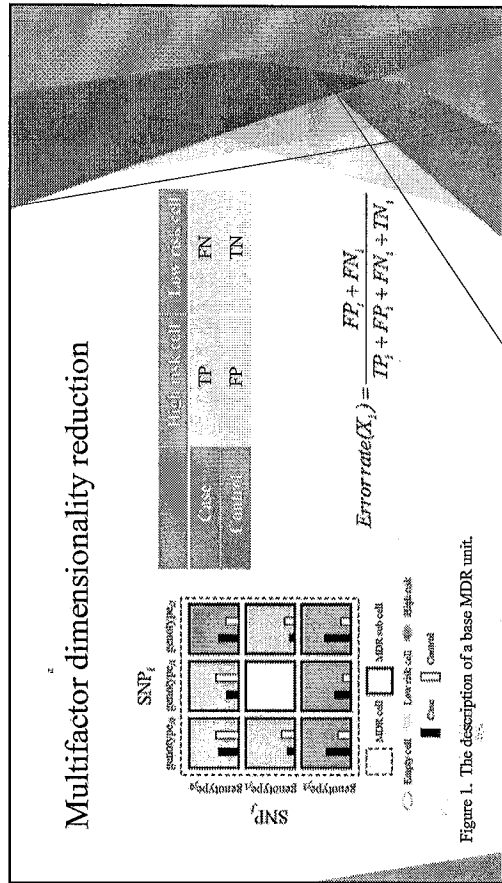
[10] L. W. Hahn, M. D. Ritchie, and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics*, vol. 19, pp. 376-382, 2003.

Entropy-based multifactor dimensionality reduction

- ▶ The entropy-based MDR uses the interaction gain [11] as a preprocess of MDR.
- ▶ EMDR leaves a number of SNP combinations that doesn't pass the threshold of gain limit.
- ▶ In this way, we can save the computational time through dislodging the SNP combinations with low gain value.

[11] R. Fan, M. Zheng, S. Wang, Y. Zhang, A. Andrew, M. Karagas, et al., "Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions for relations of complex diseases," *Genetic epidemiology*, vol. 35, pp. 706-721, 2011.

Methodology
Multifactor dimensionality reduction, entropy-based information gain, entropy-based multifactor dimensionality reduction



Entropy-based information gain

1. Definition

$$H(X) = -E[\log P(X)] = -\sum_{x \in X} p(x) \log p(x)$$

$$G_A = \begin{Bmatrix} 2 & AA \\ 1 & Aa \\ 0 & aa \end{Bmatrix}, G_B = \begin{Bmatrix} 2 & BB \\ 1 & Bb \\ 0 & bb \end{Bmatrix}, G_C = \begin{Bmatrix} 2 & CC \\ 1 & Cc \\ 0 & cc \end{Bmatrix}$$

Entropy-based information gain

2. Interaction gain

$$H(A) = -\sum_{i=0}^{SNIP} P(G_A = i) \log P(G_A = i)$$

$$H(A | D) = -\sum_{i=0}^{SNIP} P(G_A = i | D = case) \log P(G_A = i | D = case)$$

Entropy-based information gain

3. Two way interaction gain

An item (X) has its entropy $(H(X))$, but what about the covariance items? $(X, Y$ or $X, Y, Z)$

$$I(A, B) = H(A) + H(B) - H(A, B)$$

$$I(A, B | D) = H(A | D) + H(B | D) - H(A, B | D)$$

$$IG(AB | D) = I(A, B | D) - I(A, B)$$

entropy-based multifactor dimensionality reduction

So what's the main part that EMDR is different from MDR?

Ans. Uses the interaction gain as a preprocess of MDR.

Results and Discussion

Results and Discussion

Results and Discussion

Results and Discussion

Datasets

	AA	Aa	aa	CC
BB	0.060	0.010	0.010	0.061
Bb	0.010	0.208	0.208	0.017
bb	0.010	0.208	0.208	0.017

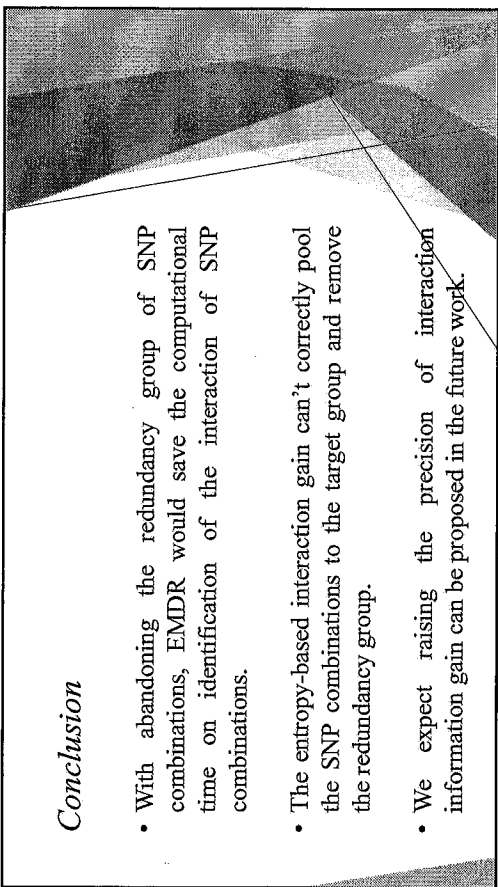
2-locus model 1

	AA	Aa	aa	AA	Aa	aa
BB	0.4	0.9	0.7	0.2	0.6	1.0
Bb	0.9	0.0	0.6	0.9	0.0	0.3
bb	0.1	0.2	0.6	0.3	0.6	0.3

3-locus XOR model

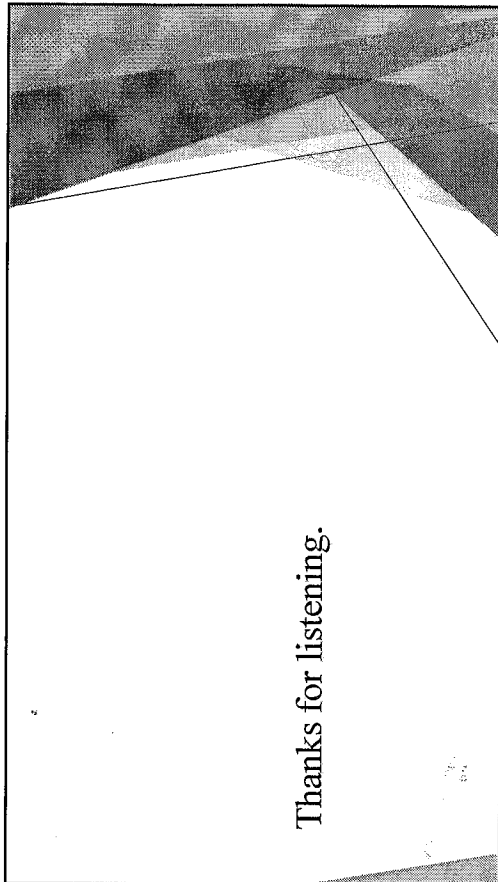
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
BB	0	1	1	1	0	1	1	0	1
Bb	0	0	0	0	0.5	0	0	0	0
bb	1	0	1	1	0	1	1	0	0

3-locus ZZ model

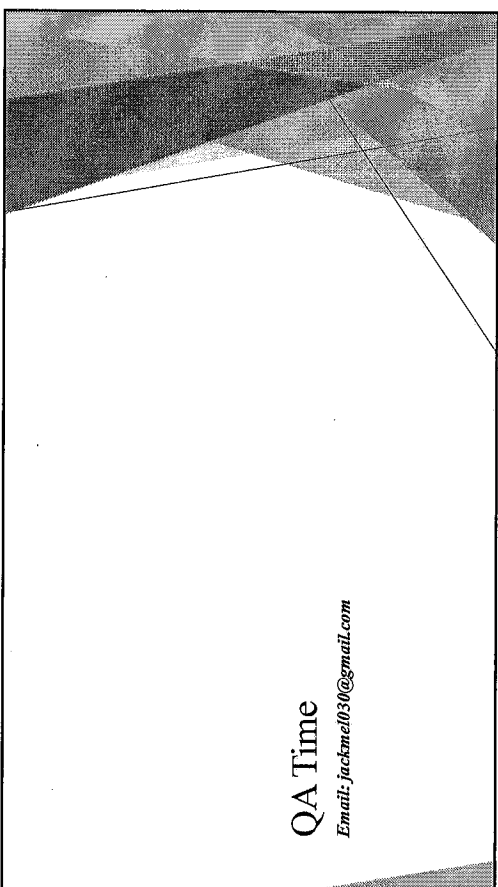


Conclusion

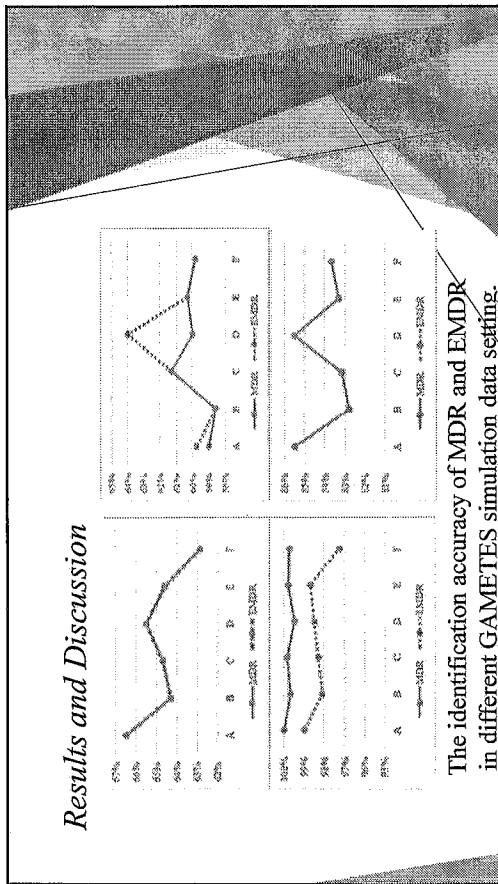
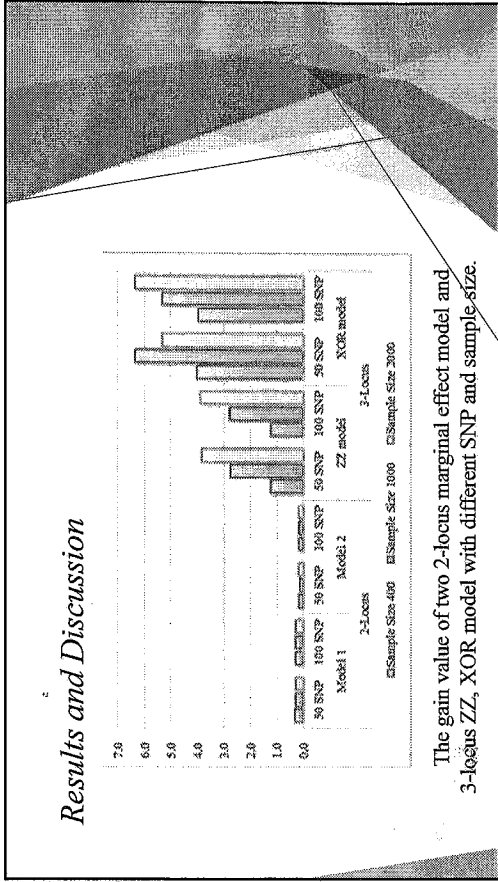
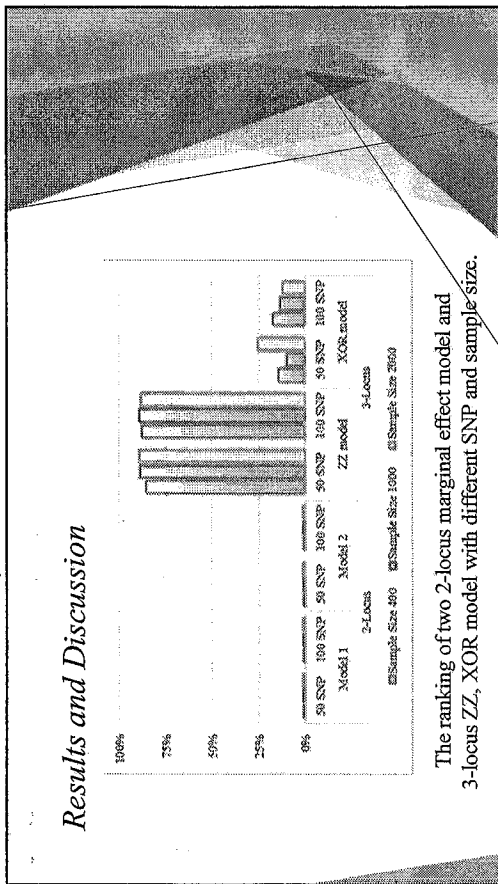
- With abandoning the redundancy group of SNP combinations, EMDR would save the computational time on identification of the interaction of SNP combinations.
- The entropy-based interaction gain can't correctly pool the SNP combinations to the target group and remove the redundancy group.
- We expect raising the precision of interaction information gain can be proposed in the future work.



Thanks for listening.



QA Time
Email: jackme030@gmail.com



Conclusion

Conclusion