出國報告（出國類別：其他：國際會議）

出席印尼峇里島 2016 ICCBS 國際研討會報告書

服務機關：國立高雄應用科技大學

姓名職稱：吳國銓/博士生

派赴國家：印尼

出國期間：2016.06.23-2016.07.04

報告日期：2016.07.07

摘要

我們的著作之研究論文 Breast Cancer Risk Prediction in SNP-SNP Interaction using Ions Motion Optimization 被 2016 5th International Conference on Bioinformatics and Biomedical Science (ICBBS 2016)國際研討會接受為口論報告論文。因此在研討會議程時間(6 月 25 日至 27 日)赴印尼峇里島出席並口頭報告本論文。本論文被安排在下午 Bioinformatics & Medical 的場次進行，早上聽完四場演講場次後，亦多聽了 Biomedicine 場次的報告。完成會議後，協會詢問我是否有意願成為他們的審稿者，亦是另一種收獲。而每次參與國際研討會最大的收獲就是可以提升自己的國際觀之外，亦可累積自己上台演說的經驗。最後由衷的感謝學校能提供補助的機會，讓我們出國學習之餘，可以大大的減輕我們旅費的負擔。

關鍵詞：Breast Cancer、SNP-SNP Interaction、Ions Motion Optimization、ICBBS 2016 國際研討會、印尼峇里島。

目次

一、目的

我們的著作之研究論文 Breast Cancer Risk Prediction in SNP-SNP Interaction using Ions Motion Optimization 被 2016 5th International Conference on Bioinformatics and Biomedical Science (ICBBS 2016)國際研討會接受為口論報告論文，因此在研討會議程時間(6 月 25 日至 27 日)赴印尼峇里島出席研討會並口頭報告本論文。

二、過程

會議地點

Patra Jasa Bali Resort & Villas, Bali, Indonesia.

研討會簡述

ICBBS 2016 國際研討會由 association of the scientists and engineers in Chemical, Biological, & Environmental Engineering (CBEES) 協會主辦，本屆研討會主要研究領域包含三大主題：Bioinformatics and Computational Biology (包含 Protein structure, function and sequence analysis、Computational proteomics 及 Algorithms, models, software, and tools in Bioinformatics 等多項主題)、Biomedical Engineering (Biomedical imaging, image processing & visualization、Bioelectrical and neural engineering 及 Biomechanics and bio-transport)及 Other Related Topics (Biostatics、Biometric 及 Biomeasurement)。發表的論文將會分別收錄在 Journal of Life Sciences and Technologies (JOLST, ISSN: 2301-3672)與 International Journal of Pharma Medicine and Biological Sciences (IJPMBS, ISSN: 2278-5221)兩個期刊之中。

會議議程

Brief Schedule for Conferences

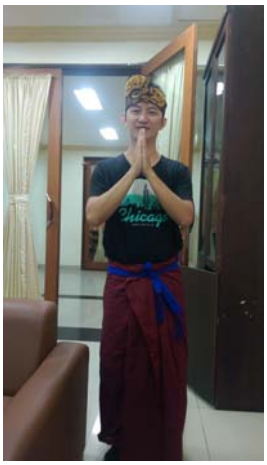
<table>
<tr><td>Day 1</td><td colspan="2">Afternoon, June 25, 2016 (Saturday) Venue: Lobby
Arrival Registration 13:30~17:00
(Committee Meeting 14:00~16:00)</td></tr>
<tr><td rowspan="9">Day 2</td><td colspan="2">June 26, 2016 (Sunday) 8:50~17:30
Venue: Gianyar Room & Klungkung Room
Arrival Registration, Keynote Speech, and Conference Presentation</td></tr>
<tr><td colspan="2">Morning Conference</td></tr>
<tr><td colspan="2">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 Chulaongkorn 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)</td></tr>
<tr><td colspan="2">Lunch 12:00~13:00 Venue: The Coffee Shop</td></tr>
<tr><td colspan="2">Afternoon Conferences</td></tr>
<tr><td>Session 1: 13:00~15:00
Venue: Gianyar Room
8 presentations-Topic: “Food Science & Biochemistry”</td><td>Session 2: 13:00~15:00
Venue: Klungkung Room
8 presentations-Topic: “Biomedicine”</td></tr>
<tr><td colspan="2">Coffee Break 15:00~15:30</td></tr>
<tr><td>Session 3: 15:30~17:30
Venue: Gianyar Room
8 presentations-Topic: “Environment”</td><td>Session 4: 15:30~17:30
Venue: Klungkung Room
8 presentations-Topic: “Bioinformatics & Medical”</td></tr>
<tr><td colspan="2">Dinner 17:40 Venue: The Coffee Shop</td></tr>
<tr><td>Day 3</td><td colspan="2">June 27, 2016 (Monday) 9:00~17:00
One Day Visit & Tour</td></tr>
</table>

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

會議過程

Day 1 (2016.06.25)

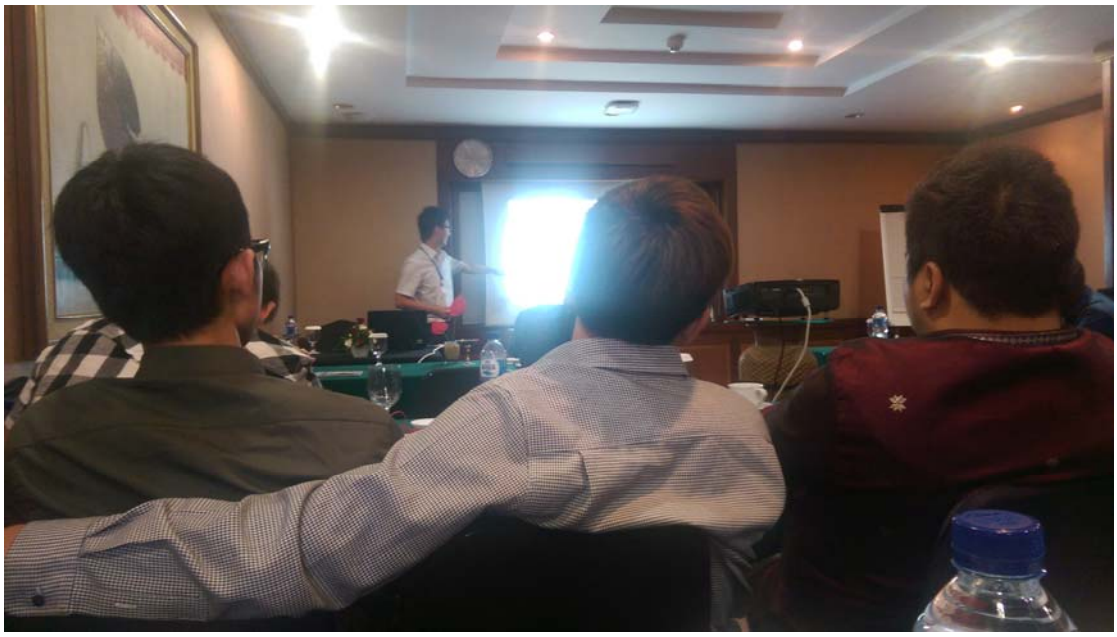
由於同行的研究室同學，有兩位印尼同學，因此早上印尼同學與峇里島當地大學接洽，帶我們前往 Bali State Polytechnic 參觀校區及剛好適逢當地的慶典，更深入感受不同的文化。在體驗在地文化過後，前往會議現場簽到。



Day 2 (2016.06.26)

我們一早依議程時間走，早上四場演講，第一位演講者來自泰國 The Environmental Research Institute of Chulaongkorn University 的 Orawan Siriratpiriya 教授，講述關於環境相關的議題，包含水源回收再利用等。第二位來自印尼 Engineering Faculty, Udayana University 的 Tjokorda Gde Tirta Nindhia 教授，探討印尼野生蠶繭用於生技材料上的應用。第三位來自美國 Rutgers, The State University of New Jersey 的 Helmut Zarbl 教授，講述細胞內 DNA 損害反應及修復的日夜生理時鐘的觀念。最後由來自美國 Environmental Engineering Civil and Environmental Engineering Department, Idaho State University 的 Solomon W. Leung 教授探討重金屬在農村的地下水之影響。早上議程結束拍下全體合影，作為四場精彩演講的完美收尾。

接著下午接續有四個會議，分成兩個房間進行，我們選擇參與 Biomedicine 與 Bioinformatics & Medical 兩個主題，而我們的論文歸為 Bioinformatics & Medical。先聽完 Biomedicine 場次後，發現隔行如隔山，很多東西都蠻陌生的，不過聽完報告後亦能略知一些，也發現很多很有趣的研究議題可以探討，如一些生醫的一些系統，如復健及肌肉感知等。另也有一些生醫的應用，如蠶繭的應用等。結束後，接著是 Bioinformatics & Medical 主題，也就是我們的研究主題會議。輪到我報告時，英文程度沒有相當流利的我，即便已經潤稿許多次，上台之後仍然緊張，腦袋一片空白，所準備的一切都彷彿都已消失。所幸對自己所研究的題目有一定的熟悉度，加上準備的夠充足，所以也算順利完成報告。台下有位先進詢問了我一個問題，也順利的回答了。會後，有位在哈薩克任教的大陸人，來鼓勵我說報告講的很棒，聊了一陣子後，隨後也交換了聯絡方式，並尋求未來合作的機會。



Day 3 (2016.06.27)

跟著研討會參訪行程走，其行程介紹如附錄所示。一路上我們參訪了峇里島的大學、醫院及著名景點，同行的有美國人、印尼人、泰國人、南非人還有中國大陸的人，旅途間不斷的溝通交流，不僅多的是英語能力的交流，更多的是各國文化之間的交流。下面照片剛好被主辦協會捕捉到與印度人交流的情況。

三、心得及建議事項

在預算有限的情況下，我們選擇了廉航。從高雄出發至吉隆坡轉機至峇里島，加上航班誤點全程耗時 10 個多小時。第一次坐這麼久的飛機，也因為廉航的關係，任何額外的服務都需要加錢，如食物、水。因此就這樣挨渴挨餓的情況下到達了峇里島，也算是另一種的體驗。

由於提前議程時間一天，所以我們自己安排了一天遊峇里島的行程，當地的文化和我之前去過的國家(如日本、香港)差異相當的大，當地也幾乎沒有中文字，所以對我來說算是非常新鮮的體驗，雖然食衣住行等溝通上沒什麼大問題，但還是深深覺得自己的英文仍需加強，不過也體認古人說，讀萬卷書不如行萬里路。

我們第一天報到時，研討會承辦人員劈哩啪啦講了一大堆英文後，得知我們會講中文，突然眉開眼笑了起來一直跟我們講中文，雖然是大陸四川人，但仍有一種異鄉遇到同鄉，他鄉遇故知的感覺，突然覺得是人生一大樂事！

第二天正式議程時，發現這研討會竟只有一位承辦人員，忙東忙西的，感覺有點誇張，覺得這部份研討會應當改善，但也由得佩服這位承辦人員。早上的議程主要是四場演講，裡面只有一個議題和我的研究領域較為相關，不僅報告的相當流暢專業，所做的簡報亦有相當水準，整體報告的方式很值得我學習，也激勵自己，未來也能用英文如此報的如此精彩。

回想起碩士班時，投了十幾篇的研討會，上台報告了數十次，因此上台時都能用很從容的心情面對，相對表現的都十分得宜。反之英文報告的次數就相對少了許多，不過也因為之前經歷過一場台中、兩場香港的訓練，讓這場上台時減少了很多不安定感。雖然報告的是相當的流暢，但也算是報告的蠻順利的。會後協會詢問是否成為他們的審稿會員，也算是另類的收獲。

最後由衷的感謝學校能提供補助的機會，讓我們出國學習之餘，可以大大的減輕我們旅費的負擔。

附錄

研討會行程

One Day Visit & Tour

June 27, 2016 (Monday) 9:00-17:00

(Tip: We will depart on time, please arrive at the Lobby before 9 a.m.)

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

The Turtle Conservation and Education Center (TCEC) opened by the governor of Bali, Mr Dewa Barata (20 January 2006) on Serangan island of Bali. TCEC is developed as part of the comprehensive strategy to eradicate illegal turtle trading on the island. Established on a land of 2.4 ha, the TCEC is trying to support the community of Serangan to find the alternatives beside illegal turtle business. The centre harnesses the potential of education, tourism, conservation and research, with a liberal sprinkling of business, to give endangered turtles one more chance on Serangan.

The four fundamental aspects to the centre include putting a definitive end to turtle trade, by encouraging the public not to consume turtle products (religious use or otherwise), and to generally support turtle conservation; providing turtles for rituals - without their killing - and monitoring turtle size and numbers, so that their use can be strictly controlled and regulated; offering employment opportunities for locals from Serangan; and finally, acting as a watchdog for turtle trade - in Serangan in particular and Bali in general.

2. Visit Udayana University (University hospital, Institute of peace and Democracy

(Photo session in front of Rectorat Building) 11:00-12:00

In the beginning of the 1960s, the people of Bali aspired to have a Tertiary Institution on the island. In order to realize this aspiration, on May 12th 1961, several figures from the educational sector, government, and community leaders conducted a conference led by Prof.Dr. Purbatjaraka,and assisted by Prof. Dr. Ida Bagus Mantra as secretary.

The conference discussed the steps required for the preparation of the establishment of a tertiary institution in Bali. An agreement was also reached for the formation of a committee led by dr. Anak Agung Made Djelantik, Head of the Board of Health in Bali, with a team of eight members.

Subsequently, the committee formed an institution named the Tertiary Education Institution of Bali, chaired by Ir. Ida Bagus Oka (Coordinator of Public Works Boards in the Southeast Islands Region); vice chaired by Dr. I Gusti Ngurah Gede Ngurah, assisted by two secretaries, Prof. Dr. Ida Bagus Mantra, and Drh. G.D. Teken Temadja. This institution succeeded in forming the Preparatory Committee for the establishment of Udayana University Bali on January 15th, 1962.

By a decision of the Directorate General of Higher Education, Ministry of Education and Culture of Indonesia, Udayana University (UNUD) was officially founded in August 17, 1962. Initially Unud consisted of four

faculties: Letters, Medicine, Veterinary Sciences and Animal Husbandry and Education and Teacher Training. The Faculty of Letters was actually established on 29th September 1958, however, the time it was a subsidiary of the Faculty of Letters of Airlangga University in Surabaya (East Java). This Faculty was thenintegrated into Udayana University in 1962. Although it was founded on August 17, the anniversary date of Udayana University is not August 17, but was choosen to be on September 29 to commemorate the date of establishment of the Faculty of Letters in 1958. Unud has develop rapidly, in 2015 the university has 13 faculties, 25 master programs and 10 doctoral programs.

Udayana University today's is listed as one of the 50"Promising Universities of Indonesia" published by the Ministry of Education of Republic Indonesia, out of nearly 2.500 higher education institutions around the country. The university has a strong position as one of the leading university particularly in the Eastern Indonesian Territory.

3. Lunch at Garuda Wisnu Kencana

Mandala Garuda Wisnu Kencana, or **Garuda Wisnu Kencana** (GWK), is a cultural park covering approximation 60 ha area located in Ungasan, Badung Regency, or about 10–15 minutes driving from Bali Ngurah Rai International Airport. It is devoted to the Hindu God Vishnu, and his mount, Garuda, the mythical bird who become his companion.

Currently, the statue of Vishnu is 23 metres (75.5 ft) high, although the original plan was for a 120-metre (390 ft) gold-plated Vishnu riding Garuda on top of an 11-storey entertainment complex. Garuda wing span will be 64 metres (210.0 ft) across. The idea was not without controversy, and religious authorities on the island complained that its massive size might disrupt the spiritual balance of the island, and that its commercial nature was inappropriate, but some groups agree with the project, because it will make new tourist attraction over barren land.

In 2013 Nyoman Nuarta and PT Alam Sutera Realty Tbk (IDX:ASRI) joined to build villas and apartments in the GWK area in exchange for Rp150 billion ($14.4 million). Nuarta plans to spend Rp20 billion to make another bust and to move the existing bust to another site 300 meters from the original site. It plans to spend additional Rp29 billion to make the new statue of stainless steel instead of galvanized steel as proposed previous design.

4. Tour to Uluwatu Temple

Uluwatu Temple (Indonesian: ***Pura (Luhur) Uluwatu***) is a Balinese sea temple (*pura segara*) in Uluwatu (Kuta South, Badung). The temple is regarded as one of the *sad kahyangan* and is dedicated to Sang Hyang Widhi Wasa in his manifestation as Rudra.

The temple (*pura* in Balinese) is built at the edge (*luhur*) of a 70 meter high cliff or rock (*watu*) projecting into the sea. In folklore, this rock is said to be part of Dewi Danu's petrified barque.

Though a small temple was claimed to have existed earlier, the structure was significantly expanded by a Javanese sage, Empu Kuturan in the 11th Century. Another sage from East Java, Dang Hyang Nirartha is credited for constructing the padmasana shrines and it is said that he attained moksha here, an event called *ngeluhur* ("to go up") locally. This has resulted in the temple's epithet *Luhur*.

5. Dinner (farewell party) at Muaya Beach Jimbaran

研討會論文簡報

2016 5th International Conference on Bioinformatics and Biomedical Science (ICBBS 2016)

Breast Cancer Risk Prediction using Ions Motion Optimization Algorithm

Cheng-Hong Yang[a], Kuo-Chuan Wu[b], Li-Yeh Chuang[c,*]

[a] Department of Electronic Engineering, National Kaohsiung University of Applied Sciences
E-mail address: chyang@kuas.edu.tw
[b] Department of Electronic Engineering, National Kaohsiung University of Applied Sciences
Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences
E-mail address: 1099405116@kuas.edu.tw
[c] Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan
E-mail address: chuang@isu.edu.tw

OUTLINE

CHAPTER 1

- Breast cancer
  - most common female malignancy in the world and has recently increased significantly in Taiwan
  - morbidity and mortality rate were estimated 1.7 million cases and 521,900 deaths among females worldwide in 2012
  - early detection of the disease and discovery of number of genetic variants are very important and valuable for prognosis and treatment
- In epidemiological studies
  - several high-penetrance breast cancer genes were discovered, such as BRCA1 and BRCA2 increase breast cancer risk up to 20-fold
  - more than 70 single nucleotide polymorphisms (SNPs) were identified that influence breast cancer risk in genome-wide association studies (GWAS)

CHAPTER 1

- Single nucleotide polymorphisms (SNPs)
  - known as the most common type of DNA sequence variation
  - play an essential role for high risk disease identification
- Genome-wide association studies (GWAS)
  - complex diseases with high correlation rates in case-control studies
  - SNP-SNP interaction (or epistasis) to determine genetic multifactorial associated with biological mechanisms and individual risk prediction
- Traditionally, the parametric linear statistical model is a poor epistasis detection approach
  - some limitations designed for single-locus
  - the advantage of computational methods in the machine learning were presented to solve the limitations of linear parametric statistical methods

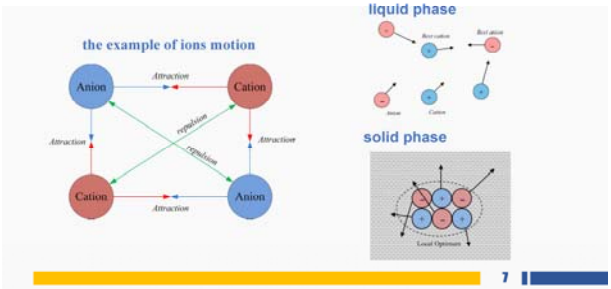
CHAPTER 1

- Data mining
  - a tool based on machine learning concepts for high dimensional data
  - several approaches have been proposed such as
    - multifactor-dimensionality reduction (MDR)
    - random forest (RF)
    - Bayesian
    - multi-objective ant colony optimization (MACOED)
- In this study
  - wrapper algorithm of attribute selection using ions motion optimization (IMO) for epistasis detection
  - evaluate whether the selected combination of locus affect the high risk prediction accuracy rate

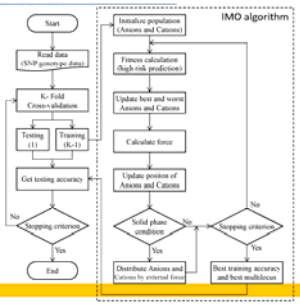
CHAPTER 2 Ions motion optimization (IMO) 1/2

- A physics based metaheuristic optimization algorithm has been proposed for global optimization problems by Javidy et al. in 2015
  - inspired by the nature of the ion motion
  - the ions were divided into anion (negative charge) and cation (positive charge) that represents candidate solutions
  - ions motion utilized attraction/repulsion forces between two ions to move the position around the feasible search space
  - forces as acceleration of ions motion
    - anions were according to the best fitness of cation and cations were according best fitness of anion
  - two strategies of the movement calculation
    - liquid phase and solid phase for diversification and intensification in search

CHAPTER 2 Ions motion optimization (IMO) 2/2



CHAPTER 2 IMO for SNP epistasis detection 1/6



Population initialization

A population consists of *N* Anion/Cation moving around in a *D*-dimensional search space

Anion: $A_i = (a_{i1}, a_{i2}, \ldots, a_{iD}) \rightarrow \boldsymbol{A} \in \{a_1, a_2, \ldots, a_N\}$

Cation: $C_i = (c_{i1}, c_{i2}, \ldots, c_{iD}) \rightarrow \boldsymbol{C} \in \{c_1, c_2, \ldots, c_N\}$

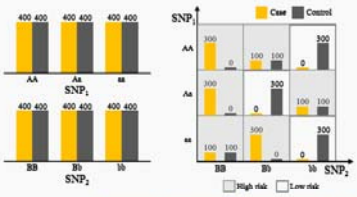
initialize population of Anions and Cations with random position

each position of an ion is a candidate solution for multi-locus genotypes

Fitness calculation

accuracy estimation from INME disease model

SAMPLE \ SNP	SNP_1	SNP_2	...	SNP_m
S_1	AA	BB	...	cc
S_2	AA	Bb	...	CC
S_3	Aa	bb	...	CC
⋮	⋮	⋮	⋮	⋮
S_n	aa	BB	...	Cc



diseases risk estimation from cases/controls ratio

Fitness calculation

A prediction contingency table was generated from the case-control, high-low risk and training-testing samples, then the accuracy was calculated

		Actual (or disease)	
		+	-
Predicted (or test)	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Force evaluation (liquid phase)

$$AF_{i,j} = \frac{1}{1 + e^{-0.1/AD_{i,j}}} \rightarrow A_{i,j} = A_{i,j} + AF_{i,j} \times (Cbest_j - A_{i,j})$$

$$CF_{i,j} = \frac{1}{1 + e^{-0.1/CD_{i,j}}} \rightarrow C_{i,j} = C_{i,j} + CF_{i,j} \times (Abest_j - C_{i,j})$$

$AD_{i,j} = |A_{i,j} - Cbest_j|$ is the distance between anion and best cation

$CD_{i,j} = |C_{i,j} - Abest_j|$ is the distance between cation and best anion

AF_{ij} and CF_{ij} represent resultant attraction force of anions and cations respectively

External force distribution (solid phase)

if (CbestFit ≥ CworstFit / 2 AND AbestFit ≥ AworstFit / 2)
  if $rand_1() > 0.5$
    $A_i = A_i + \Phi_1 \times (Cbest - 1)$
  else
    $A_i = A_i + \Phi_1 \times Cbest$
  end if
  if $rand_2() > 0.5$
    $C_i = C_i + \Phi_2 \times (Abest - 1)$
  else
    $C_i = C_i + \Phi_2 \times Abest$
  end if
  if $rand_3() < 0.05$
    Re – initialized A_i and C_i with random position
  end if
end if

- Φ_1 and Φ_2 are random numbers in range of -1 to 1
- $rand_1()$, $rand_2()$ and $rand_3()$ are random numbers in range of 0 to 1
- *AworstFit* and *CworstFit* are worst fitness solutions of anion and cation respectively

554 case-control samples include seven SNPs with breast cancer

SNP (Genes)	Chr.	Genotype	Control	Case	OR	*p*-value
rs12812942 (CD4)	12	1-AA	174	128	–	–
		2-AT	141	76	0.733	0.10
		3-TT	19	16	1.145	0.72
rs3136685 (CCR7)	17	1-GG	107	77	–	–
		2-AG	180	114	0.880	0.57
		3-AA	47	29	0.857	0.68
rs2228014 (CXCR4)	2	1-CC	254	151	–	–
		2-CT	73	63	1.452	0.07
		3-TT	7	6	1.442	0.57
rs1801157 (CXCL12)	10	1-GG	175	106	–	–
		2-AG	136	98	1.189	0.37
		3-AA	23	16	1.149	0.73
rs3025039 (VEGF)	6	1-CC	211	155	–	–
		2-CT	117	59	0.687	0.05
		3-TT	6	6	1.361	0.77
rs2287074 (MMP2)	16	1-GG	164	113	–	–
		2-AG	139	93	0.971	0.93
		3-AA	31	14	0.655	0.25
rs10506957 (KITLG)	12	1-TT	182	133	–	–
		2-CT	133	69	0.709	0.08
		3-CC	19	18	1.296	0.08

Chr: Chromosome; OR: odds ratio

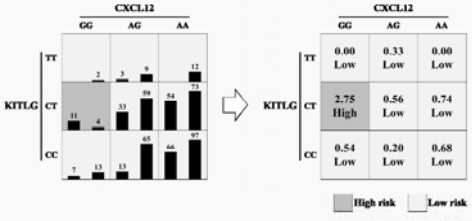
220 pathologically confirmed female breast cancer patients
334 female control participants

Breast cancer of results from IMO prediction

	Mean ± SAMs		
Number of loci	Training ACC (%)	Testing ACC (%)	CV Consistency
2	61.86 ± 0.04	**57.99 ± 0.94**	**7.34 ± 0.84**
3	63.82 ± 0.07	56.93 ± 0.94	5.53 ± 0.81
4	66.34 ± 0.08	54.02 ± 0.98	2.91 ± 0.58
5	**70.22 ± 0.07**	54.32 ± 1.10	4.41 ± 0.67

*SAMs: standard accuracy of the means, ACC: accuracy, CV: cross-validation

Summary of two-locus genotype combinations associated with high risk and with low risk for breast cancer



CHAPTER 4 Discussion

- MDR
  - non-parametric statistical method and no need to assume a specific genetic model for the detection of high-dimensional SNP interactions
  - has been used in epistasis classically and widely over the past decade
  - a robust statistical computing algorithm and can effectively identify high multi-factorial SNP interactions in the lack of marginal effects
- However, remains an exhaustive search that leads to computationally very intense and depends on the computing resources largely

CHAPTER 4 Discussion

- We have implemented IMO as an approach for SNP-SNP interaction of multi-locus information detection
  - to search the classification of best combinations in SNPs with marginal effects that associate with the risk for complex diseases
  - has the advantage of few number of tuning parameters, low computational complexity, fast convergence and high local optima avoidance
  - a potential epistasis detection algorithm for high-dimension data

CHAPTER 5

- Our proposed approach can obtain a better identification ability based on the cross-validation consistency and prediction accuracy
- For future work, the IMO algorithm will be implemented on the simulation data and other complex diseases data

Reporter: Kuo-Chuan Wu
Email: kuo.chuan.wu@gmail.com

論文接受原稿

Breast Cancer Risk Prediction using Ions Motion Optimization Algorithm

Cheng-Hong Yang
Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan 80708
Email: chyang@kuas.edu.tw

Kuo-Chuan Wu
Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan 80708
Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan 80708
Email: kuo.chuan.wu@gmail.com

Li-Yeh Chuang
Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan 80041
Email: chuang@isu.edu.tw

***Abstract*—Breast cancer is the most common female malignancy in the world and has recently increased markedly in Taiwan. In epidemiological studies, several high-penetrance breast cancer genes, such as BRCA1 and BRCA2, were discovered to increase breast cancer risk up to 20-fold. Recently, there are more than 70 single nucleotide polymorphisms (SNPs) identified that influence breast cancer risk in genome-wide association studies (GWAS). This study collects 554 samples of breast cancer and non-breast cancer data, and implements a wrapper algorithm of attribute selection using ions motion optimization (IMO) to identify the SNP-SNP interaction. The experimental results have shown our proposed method has reasonable power to identify SNPs interactions for multi-locus interaction associated with a common complex multifactorial disease.**

***Index Terms*—breast cancer, single nucleotide polymorphisms, SNP-SNP interaction, ions motion optimization**

I. Introduction

Breast cancer is the most common female malignancy in the world and has recently increased significantly in Taiwan [1]. The morbidity and mortality rate of breast cancer were estimated 1.7 million cases and 521,900 deaths among females worldwide in 2012 [2]. Although breast cancer is curable when diagnosed early in cancer treatment, if it is recurrence or distant organs metastasis, it will become incurable and mortal [3]. Thus, early detection of the disease and discovery of number of genetic variants are very important and valuable for prognosis and treatment. Breast cancer is known to be caused by multiple genetic and nongenetic (environmental) factors, perhaps the interaction between the two factors [4]. In epidemiological studies, several high-penetrance breast cancer genes were discovered, such as BRCA1 and BRCA2 increase breast cancer risk up to 20-fold. Recently, there are more than 70 single nucleotide polymorphisms (SNPs) were identified that influence breast cancer risk in genome-wide association studies (GWAS) [5]-[7].

SNPs play a very important role in high risk disease identification. It is known as the most common type of DNA sequence variation and can affect the gene expression. The SNPs are defined when a nucleotide (A, T, C and G) changes more than 1% or greater within the human population [8]. New opportunities and challenges had been presented to find association between genetic polymorphisms and phenotypes to GWAS with the completion of the human genome project (HGP) and the international haplotype map project (HapMap) in the last decade. GWAS are most of the complex diseases with high correlation rates in case-control studies from SNP that examine SNP-SNP interaction to determine genetic multifactorial associated with biological mechanisms and individual risk prediction [9]-[12].

Traditionally, the parametric linear statistical model is a poor epistasis detection approach which has some limitations designed for single-locus, such as it does work on statistical modeling of non-linear interactions and considering multiple SNPs simultaneously [13]. The advantages of computational methods in the machine learning are presented to solve the limitations of linear parametric statistical methods. Data mining is a tool based on machine learning concepts for high dimensional data which has been applied in GWAS. Several data mining approaches have been proposed such as multifactor-dimensionality reduction (MDR) [9], [12], random forest (RF) [14], [15], multi-objective ant colony optimization (MACOED) [16] and Bayesian epistasis association mapping [17].

Manuscript received April 25, 2016; revised June 25, 2016.

In this paper, we implement a wrapper algorithm of attribute selection using ions motion optimization (IMO) for SNP-SNP interaction identification. IMO was used to select the polymorphisms and evaluate whether the selected combination of locus affect the high risk. The accuracy rate of prediction was determined using SNP interactions displaying no marginal effects (INME) [10] with k-fold cross-validation. This study collected 554 samples of breast cancer and non-breast cancer data, and the genotyping was determined by PCR-restriction fragment length polymorphism (RELP). Those SNP data were analyzed by machine learning approach for SNP-SNP interaction with disease risk prediction.

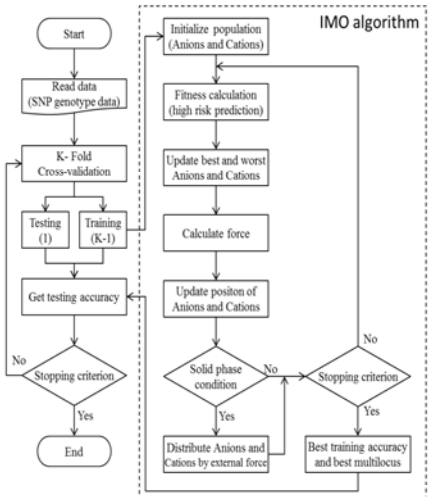


Figure 1. The flowchart of IMO for SNP-SNP interaction identification

II. METHODS

A. Ions Motion Optimization (IMO) Algorithm

A physics based metaheuristic optimization algorithm has been proposed for global optimization problems by Javidy et al. in 2015 [18]. It is inspired by the nature of the ion motion. The ions are divided into anion (negative charge) and cation (positive charge) that represents candidate solutions to the particular problem. The ions motion utilizes attraction/repulsion forces between two ions to move the position around the feasible search space. The forces are represented as acceleration of ions motion, anions are according to the best fitness of cation and cations are according to the best fitness of anion. Two strategies of the movement calculation are called liquid phase and solid phase for diversification and intensification in search.

B. IMO for SNP-SNP Interaction Identification

This paper presents an IMO algorithm for identification of SNP-SNP interaction revealing high-order interactions problem from case-control studies. We used the heuristic algorithm to find the best multi-locus combination to identify the high risk multi-locus genotypes. The flowchart is shown in **错误！未找到引用源。** and the detail process is described as follows.

Step 1) Initialize population

A population consists of N Anion/Cation moving around in a D-dimensional search space. The position of the ith Anion and ith Cation can be represented by $A_i = (a_{i1}, a_{i2}, \ldots, a_{iD})$, $C_i = (c_{i1}, c_{i2}, \ldots, c_{iD})$, respectively. Initialize population of Anions and Cations with uniform random position $\boldsymbol{A} \in \{a_1, a_2, \ldots, a_N\}$ and position $\boldsymbol{C} \in \{c_1, c_2, \ldots, c_N\}$, each position of an ion is a candidate solution for multi-locus genotypes.

Step 2) Fitness calculation

The fitness calculation is used as accuracy estimation from INME disease model (example shown in **错误！未找到引用源。**). SNP interaction refers to combinatorial effect of genetic variants that observes the marginal effect on the frequency in cases and controls for the combination between the SNPs phenotype distribution. This model can easily obtain disease risk estimation from the cases: controls ratio. If the ratio exceeds threshold (e.g., ≥ 1.0) which is labeled as high-risk, or as low-risk, if the ratio is not exceeded. A prediction contingency table (**错误！未找到引用源。**) is generated from the case-control, high-low risk and training-testing samples, then the accuracy is calculated.

TABLE I. A PREDICTION CONTINGENCY TABLE

		Actual (or disease)	
		+	-
Predicted (or test)	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)

Step 3) Update best and worst Anions and Cations

Determine the global best and individual current worst solution *Abest* and *Cbest* according to the fitness evaluation results (i.e., training accuracy).

Step 4) Force evaluation

The repulsion forces are ignored for search space [18]. The attraction force computation is evaluated from distance between ions; the measurement can be defined as follow:

$$AF_{i,j} = \frac{1}{1+e^{-0.1/AD_{i,j}}} \quad (1)$$

$$CF_{i,j} = \frac{1}{1+e^{-0.1/CD_{i,j}}} \quad (2)$$

where $AD_{i,j} = |A_{i,j} - Cbest_j|$ is the distance between anion and best cation, $CD_{i,j} = |C_{i,j} - Abest_j|$ is the

distance between cation and best anion. $AF_{i,j}$ and $CF_{i,j}$ represent resultant attraction force of anions and cations, respectively.

Step 5) Update position of Anions and Cations

In IMO, each ion is updated based on the attraction force as following equations:

$$A_{i,j} = A_{i,j} + AF_{i,j} \times (Cbest_j - A_{i,j}) \quad (3)$$

$$C_{i,j} = C_{i,j} + CF_{i,j} \times (Abest_j - C_{i,j}) \quad (4)$$

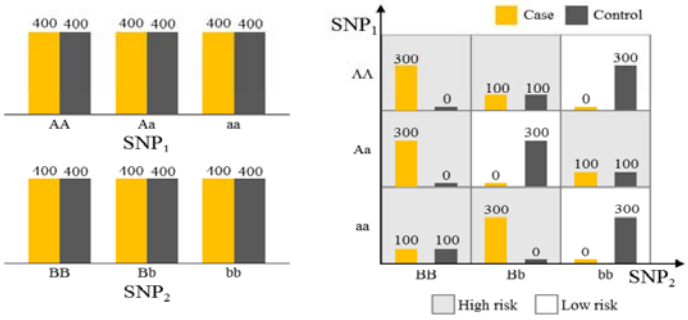


Figure 2. An illustration of the INME disease model. The left indicates the marginal effect of SNP1 and SNP2. The right indicates the frequency in cases and controls for the combinations between SNP1 and SNP2. In gray cells, the allele combinations represent high risk effect (from the threshold of cases: controls ratio e.g., ⩾1.0)

Step 6) Distribute Anions and Cations by external force

In order to escape entrapment in local optima, when the solid phase condition is satisfied, the external force is calculated, and the formula is shown as follows:

$$
\begin{aligned}
&\textbf{if } (\text{CbestFit} \geq \text{CworstFit} / 2 \\
&\qquad \text{AND AbestFit} \geq \text{AworstFit} / 2) \\
&\quad \textbf{if } rand_1() > 0.5 \\
&\quad\quad A_i = A_i + \Phi_1 \times (Cbest - 1) \\
&\quad \textbf{else} \\
&\quad\quad A_i = A_i + \Phi_1 \times Cbest \\
&\quad \textbf{end if} \\
&\quad \textbf{if } rand_2() > 0.5 \\
&\quad\quad C_i = C_i + \Phi_2 \times (Abest - 1) \\
&\quad \textbf{else} \\
&\quad\quad C_i = C_i + \Phi_2 \times Abest \\
&\quad \textbf{end if} \\
&\quad \textbf{if } rand_3() < 0.05 \\
&\quad\quad \text{Re} - \text{initialized } A_i \text{ and } C_i \text{ with random position} \\
&\quad \textbf{end if} \\
&\textbf{end if}
\end{aligned}
\quad (5)
$$

where Φ_1 and Φ_2 are random numbers in range of -1 to 1. $rand_1(\ )$, $rand_2(\ )$ and $rand_3(\ )$ are random numbers in range of 0 to 1. *AworstFit* and *CworstFit* are the worst fitness solutions of anion and cation, respectively.

Step 7) Stopping criterion

Repeat steps 2-6 until a certain number of iterations have been completed. Consequently, the best training accuracy and best multi-locus are obtained.

III. Results

A. Data Sets

The experimental data sets of this study were obtained from our previous breast cancer association study [19]. The data include 220 pathologically confirmed female breast cancer patients and 334 female control participants. The SNPs of seven genes were used in this data included CD4, CCR7, CXCR4, CXCL12, VEGF, MMP2 and KITLG. The cases and controls study was conducted from Kaohsiung Medical University in Taiwan. The baseline characteristics of breast cancer data for cases and controls is shown in **错误！未找到引用源。**.

B. Experimental Results

错误！未找到引用源。 summarizes the means and the standard accuracy of the means (SAMs), the cross-validation consistency, the training accuracy and the testing accuracy obtained from IMO analysis of the breast cancer case-control data, for each number of loci evaluated. The reported cross-validation consistency is the number of cross-validation intervals that a particular SNP combination was chosen by IMO from the average of 100 runs. One tow-locus model had a higher cross-validation consistency of 10, the accuracy are 61.86 ± 0.04 (training) and 57.99 ± 0.94 (testing), and the best

model chosen was CXCL12 (rs1801157) and KITLG (rs10506957). **错误！未找到引用源。** shows the identification of SNP-SNP interaction which is corresponding to the distribution of genotype frequencies of cases and controls in the best combination of two-locus by IMO. As shown in **错误！未找到引用源。**, each multi-locus genotype combinations are divided into the cells, the left bar in cells represents the corresponding distribution of cases and controls on the right bar.

TABLE II. DESCRIPTIVE LIST OF BREAST CANCER CASES AND CONTROLS

				Number of		Scoring function					
	SNP (Genes)	Chr.	Genotype	Control	Case	CC	SN	SP	AVG	OR	*p*-value
1.	rs12812942 (CD4)	12	1-AA	174	128	–	–	–	–	–	–
			2-AT	141	76	0.482	0.372	0.552	0.469	0.733	0.10
			3-TT	19	16	0.564	0.111	0.902	0.526	1.145	0.72
2.	rs3136685 (CCR7)	17	1-GG	107	77	–	–	–	–	–	–
			2-AG	180	114	0.462	0.587	0.373	0.474	0.880	0.57
			3-AA	47	29	0.523	0.274	0.695	0.357	0.857	0.68
3.	rs2228014 (CXCR4)	2	1-CC	254	151	–	–	–	–	–	–
			2-CT	73	63	0.586	0.294	0.777	0.552	1.452	0.07
			3-TT	7	6	0.622	0.382	0.973	0.659	1.442	0.57
4.	rs1801157 (CXCL12)	10	1-GG	175	106	–	–	–	–	–	–
			2-AG	136	98	0.320	0.480	0.562	0.524	1.189	0.37
			3-AA	23	16	0.597	0.131	0.884	0.537	1.149	0.73
5.	rs3025039 (VEGF)	6	1-CC	211	155	–	–	–	–	–	–
			2-CT	117	59	0.498	0.276	0.643	0.472	0.687	0.05
			3-TT	6	6	0.574	0.037	0.972	0.528	1.361	0.77
6.	rs2287074 (MMP2)	16	1-GG	164	113	–	–	–	–	–	–
			2-AG	139	93	0.505	0.451	0.541	0.499	0.971	0.93
			3-AA	31	14	0.553	0.110	0.841	0.510	0.655	0.25
7.	rs10506957 (KITLG)	12	1-TT	182	133	–	–	–	–	–	–
			2-CT	133	69	0.486	0.342	0.578	0.469	0.709	0.08
			3-CC	19	18	0.568	0.119	0.905	0.531	1.296	0.08

*Chr: chromosome, CC: correct, SN: sensitivity, SP: Specificity, AVG: average, OR: Odds Ratio

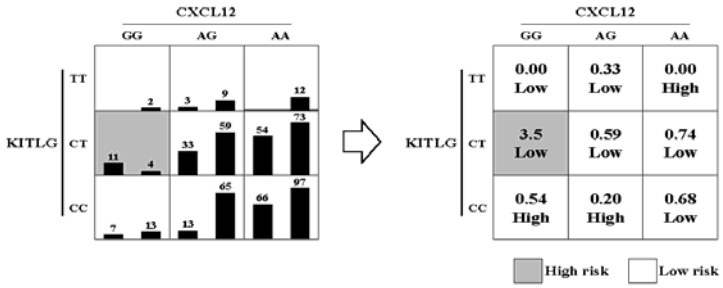


Figure 3. Summary of two-locus genotype combinations associated with high risk and with low risk for breast cancer

TABLE III. ANALYSIS RESULTS FROM IMO ALGORITHM.

	Mean ± SAMs		
Number of loci	Training ACC (%)	Testing ACC (%)	CV Consistency
2	61.86 ± 0.04	**57.99 ± 0.94**	**7.34 ± 0.84**
3	63.82 ± 0.07	56.93 ± 0.94	5.53 ± 0.81
4	66.34 ± 0.08	54.02 ± 0.98	2.91 ± 0.58
5	**70.22 ± 0.07**	54.32 ± 1.10	4.41 ± 0.67

*SAMs: standard accuracy of the means, ACC: accuracy, CV: cross-validation

C. *Analysis of the Differences between Cases and Controls in Two-Locus*

Table IV shows the difference between cases and controls in two-locus. The most difference between case and control was found in the SNP4-genotype GG and SNP7-genotype CT (i.e. 11 – 4 = 7), indicating that this combination is dominant in case. The SNP3-genotype TT and SNP4-genotype GG showed the most difference between control and case (i.e. 137 – 69 = 68), indicating that this combination is dominant in case. Consequently, the difference in control-dominant is larger than the difference in case-dominant. The difference distribution between case and control shows a lower risk rate in this data except SNP4 and SNP7. However, our proposed approach can find the best model for SNP4 (CXCL12) and SNP7 (KITLG).

IV. DISCUSSION

MDR is a non-parametric statistical method and no need to assume a specific genetic model for the detection of high-dimensional SNP interactions [9], [12], [13]. It has been used in epistasis detection classically and widely over the past decade. MDR has a robust statistical computing algorithm and can effectively identify high multi-factorial SNP interactions in the lack of marginal effects. However, MDR remains an exhaustive search that leads to computationally very intense and depends on the computing resources largely [10], [11]. Assuming that there are n dimension genotype data and m-locus are chosen, then we get the combination formula: $\mathbf{C(n,m)} = \frac{n!}{(n-m)!\,m!}$. Therefore, this study presents a heuristic search algorithm to replace brute-force search algorithm. The goal of our proposed approach is solving high-dimension data for SNP-SNP interaction identification of multi-locus.

In machine learning area, overfitting problem appears when computationally intensive search algorithms. If the set of training data are too close, it will lead to the poor quality prediction. In order to avoid overfitting, some additional techniques have been proposed, such as cross-validation, regularization, and early termination or resampling [20]. However the best way to avoid overfitting is to use an abundant amount of training data. In this paper, the k-fold cross-validation was used.

We have implemented IMO as an approach for SNP-SNP interaction of multi-locus information detection, to search the classification of best combinations in SNPs with marginal effects that associate with the risk for complex diseases. IMO algorithm has the advantages of few number of tuning parameters (only population size and number of iterations), low computational complexity, fast convergence and high local optima avoidance [18]. In this paper, the parameters of IMO were set with the values, the population size 40 and the number of iteration 100. run 100 times for detecting interactions independently. Although the dimension of breast cancer data just only seven SNPs, it has been verified to possess potential epistasis detection algorithm for high-dimension data. In fact, many heuristic algorithms had been proposed in this area and obtained high performance outcome, such as genetic algorithm [21] and ant colony optimization [16].

TABLE IV. THE DIFFERENCES BETWEEN CASES AND CONTROLS FOR THE WHOLE COMBINATIONS IN TWO-LOCUS.

SNPs	Genotype	case	control	Δ	SNPs	Genotype	case	control	Δ	SNPs	Genotype	case	control	Δ
4,7	GG-CT	11	4	7	2,7	GG-CT	9	11	-2	4,6	AG-AG	39	58	-19
4,6	AG-GG	11	7	4	3,4	CC-AG	4	6	-2	4,7	AA-CT	54	73	-19
1,5	AT-TT	9	6	3	3,6	TT-AA	1	3	-2	1,2	AT-AG	69	89	-20
3,5	CC-CT	48	45	3	4,7	GG-TT	0	2	-2	1,6	TT-GG	64	84	-20

3,7	CC-CT	5	2	3	5,6	TT-AA	2	4	-2	2,4	AG-AG	47	67	-20
5,7	CC-CT	6	3	3	2,5	AG-CC	8	11	-3	1,4	AT-AG	31	52	-21
1,2	TT-AA	9	7	2	3,7	CT-CT	21	24	-3	5,6	CT-AG	24	45	-21
1,5	TT-CT	3	1	2	3,7	TT-CC	3	6	-3	5,6	TT-GG	81	102	-21
1,7	AT-CC	8	6	2	4,6	GG-GG	0	3	-3	1,7	TT-TT	70	92	-22
1,3	TT-CC	5	4	1	6,7	GG-TT	0	3	-3	1,7	TT-CT	56	78	-22
1,4	AT-AA	7	6	1	1,3	AT-CT	23	27	-4	1,5	AT-CT	24	47	-23
1,6	AA-AA	1	0	1	1,4	TT-GG	8	12	-4	2,5	AG-TT	21	44	-23
2,3	AG-CC	9	8	1	5,7	CC-TT	12	16	-4	4,5	AG-CT	25	48	-23
2,3	AA-CT	4	3	1	1,3	AT-CC	35	40	-5	6,7	AG-CT	31	54	-23
2,4	GG-GG	3	2	1	1,4	AA-AG	5	10	-5	1,7	AT-TT	49	73	-24
2,5	AA-TT	2	1	1	1,7	TT-CC	7	12	-5	2,7	AG-CC	22	46	-24
2,7	GG-TT	2	1	1	3,4	CC-GG	12	17	-5	6,7	AA-CT	54	78	-24
3,5	TT-CC	6	5	1	3,7	CC-TT	12	17	-5	1,6	AT-AG	33	58	-25
3,6	CT-AA	4	3	1	1,2	AA-AG	11	17	-6	1,4	TT-GG	58	84	-26
3,6	TT-AG	37	36	1	1,5	AA-CC	99	105	-6	4,7	AG-CT	33	59	-26
3,7	CC-CC	1	0	1	1,5	AA-TT	7	13	-6	5,6	CT-GG	66	92	-26
3,7	CT-CC	2	1	1	2,4	GG-AG	8	14	-6	5,7	CT-CT	18	44	-26
4,7	AA-TT	13	12	1	2,6	AG-GG	8	14	-6	2,3	GG-TT	54	82	-28
5,6	CC-AA	1	0	1	2,7	AA-CC	48	54	-6	2,6	AA-AG	55	83	-28
5,6	CT-AA	3	2	1	4,5	GG-CC	11	17	-6	5,6	TT-AG	30	58	-28
5,7	TT-CC	4	3	1	4,6	GG-AA	8	14	-6	1,6	TT-AG	41	70	-29
6,7	GG-CT	8	7	1	4,7	GG-CC	7	13	-6	3,4	CT-GG	70	100	-30
6,7	GG-CC	10	9	1	4,7	AG-TT	3	9	-6	2,6	AG-AG	52	83	-31
1,4	AA-AA	1	1	0	1,6	AA-GG	11	18	-7	2,7	AG-CT	35	66	-31
1,7	AA-TT	9	9	0	2,6	GG-AG	7	14	-7	4,6	AG-AA	43	74	-31
1,7	AA-CC	1	1	0	3,4	CT-AG	26	33	-7	4,7	AA-CC	66	97	-31
2,3	AA-TT	2	2	0	2,5	GG-TT	54	62	-8	2,5	AG-CT	30	62	-32
2,5	AA-CT	4	4	0	4,6	GG-AG	6	14	-8	4,6	AA-AA	55	87	-32
2,7	GG-CC	7	7	0	4,6	AA-GG	5	13	-8	4,7	AG-CC	33	65	-32
3,4	TT-AA	4	4	0	1,2	AA-GG	18	27	-9	2,7	AA-CT	70	103	-33
3,5	CC-TT	4	4	0	2,3	AG-CT	33	42	-9	2,5	GG-CT	80	114	-34
3,6	CC-AG	5	5	0	2,6	GG-AA	5	14	-9	4,5	AG-TT	30	64	-34
3,6	CC-AA	1	1	0	2,6	AG-AA	33	42	-9	4,5	GG-TT	72	107	-35
4,5	AA-CC	1	1	0	2,7	AG-TT	12	21	-9	5,7	TT-CT	35	70	-35
4,5	AA-CT	1	1	0	5,6	CC-GG	8	17	-9	3,6	CT-GG	68	104	-36
4,5	AA-TT	4	4	0	5,6	CC-AG	5	14	-9	3,7	TT-TT	93	129	-36
6,7	AG-TT	8	8	0	2,7	AA-TT	15	25	-10	5,7	CT-TT	49	86	-37
1,3	AT-TT	5	6	-1	3,7	TT-CT	37	47	-10	1,5	AT-CC	26	64	-38
1,7	AA-CT	8	9	-1	1,6	AA-AG	2	13	-11	1,4	TT-AG	40	79	-39
2,5	AA-CC	0	1	-1	2,6	AA-GG	19	30	-11	2,4	AA-AG	59	99	-40
2,6	GG-GG	2	3	-1	3,6	CT-AG	21	32	-11	6,7	AG-CC	30	71	-41
3,4	CT-AA	2	3	-1	4,6	AA-AG	53	64	-11	1,3	AA-CC	88	130	-42
3,4	TT-AG	33	34	-1	6,7	AA-CC	73	84	-11	1,7	AT-CT	12	54	-42
3,5	CT-TT	2	3	-1	2,4	AA-AA	37	49	-12	1,2	AT-AG	38	82	-44
3,5	TT-CT	0	1	-1	2,6	AA-AA	39	51	-12	1,5	AA-CT	49	93	-44
4,5	AG-CC	4	5	-1	3,5	CT-CT	15	27	-12	3,5	CT-CC	42	87	-45
5,7	CT-CC	2	3	-1	2,5	GG-CC	21	35	-14	3,6	TT-GG	75	125	-50
1,2	AT-AA	7	9	-2	6,7	AA-TT	6	20	-14	2,3	GG-CT	77	135	-58
1,3	AA-TT	11	13	-2	1,2	TT-AG	27	42	-15	1,3	AA-CT	52	111	-59
1,3	TT-CT	1	3	-2	4,5	GG-CT	72	87	-15	3,5	CC-CC	103	162	-59
1,4	AA-GG	10	12	-2	5,7	TT-TT	94	109	-15	3,7	CT-TT	46	108	-62
1,5	TT-CC	3	5	-2	2,4	AG-AA	35	51	-16	3,4	TT-GG	69	137	-68
1,6	AT-AA	7	9	-2	1,2	TT-GG	41	58	-17	1,2	AA-AA	-	-	-
1,6	TT-AA	8	10	-2	2,3	GG-CC	20	37	-17	1,3	TT-TT	-	-	-
2,3	AG-TT	21	23	-2	2,4	AA-GG	10	27	-17	1,5	TT-TT	-	-	-
2,3	AA-CC	0	2	-2	3,6	CC-GG	8	25	-17	3,4	CC-AA	-	-	-
2,4	GG-AA	5	7	-2	1,4	AT-GG	60	78	-18	3,5	TT-TT	-	-	-
2,4	AG-GG	16	18	-2	1,6	AT-GG	53	72	-19	5,7	CC-CC	-	-	-

"Δ: difference between case and control

V. Conclusion

In this paper, we have proposed a detection method, IMO algorithm with the k-fold cross-validation, for SNP-SNP interaction: using the SNP data of breast cancer. The results of this study demonstrated that our proposed approach can obtain a better identification ability based on the cross-validation consistency and prediction accuracy. For future work, the IMO algorithm will be implemented on the simulation data and other complex diseases data.

Acknowledgment

This study was partly supported by the National Science Council of Taiwan for Grant NSC 103-2221-E-151-024-MY3.

References

[1] M. J. Chen, W. Y. Wu, A. M. Yen, J. C. Fann, S. L. Chen, and S. Y. Chiu, et al., " Body mass index and breast cancer: analysis of a nation-wide population-based prospective cohort study on 1 393

985 Taiwanese women," *Int J Obes (Lond)*, vol. 40, pp. 524-30, May, 2016.
[2] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA Cancer J Clin*, vol. 65, pp. 87-108, May 2015.
[3] H. Cao, Z. Zhang, S. Zhao, X. He, H. Yu, Q. Yin, et al., "Hydrophobic interaction mediating self-assembled nanoparticles of succinobucol suppress lung metastasis of breast cancer by inhibition of VCAM-1 expression," *J Control Release*, vol. 205, pp. 162-71, May 10, 2015.
[4] L. Fejerman, M. C. Stern, E. M. John, G. Torres-Mejia, L. M. Hines, and R. K.Wolff, et al., "Interaction between common breast cancer susceptibility variants, genetic ancestry, and nongenetic risk factors in Hispanic women," *Cancer Epidemiol Biomarkers Prev*, vol. 24, pp. 1731-8, Nov., 2015.
[5] M. R. Stratton and N. Rahman, "The emerging landscape of breast cancer susceptibility," *Nat Genet*, vol. 40, pp. 17-22, Jan., 2008.
[6] N. Mavaddat, A. C. Antoniou, D. F. Easton, and M. Garcia-Closas, "Genetic susceptibility to breast cancer," *Mol Oncol*, vol. 4, pp. 174-91, June, 2010.
[7] O. Fletcher and F. Dudbridge, "Candidate gene-environment interactions in breast cancer," *BMC Med*, vol. 12, pp. 195, 2014.
[8] B. S. Shastry, "SNP alleles in human disease and evolution," *J Hum Genet*, vol. 47, pp. 561-6, 2002.
[9] 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. 376-82, Feb., 12 2003.
[10] P. Li, M. Guo, C. Wang, X. Liu, and Q. Zou, "An overview of SNP interactions in genome-wide association studies," *Brief Funct Genomics*, vol. 14, pp. 143-55, May. 2015.
[11] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau, "A survey about methods dedicated to epistasis detection," Front Genet, vol. 6, pp. 285, 2015.
[12] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, et al., "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *Am J Hum Genet*, vol. 69, pp. 138-47, July. 2001.
[13] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, vol. 26, pp. 445-55, Feb., 15 2010.
[14] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, et al., "Identifying SNPs predictive of phenotype using random forests," *Genet Epidemiol*, vol. 28, pp. 171-82, Feb., 2005.
[15] D. F. Schwarz, I. R. Konig, and A. Ziegler, "On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data," *Bioinformatics*, vol. 26, pp. 1752-8, July. 15 2010.
[16] P. J. Jing and H. B. Shen, "MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies," *Bioinformatics*, vol. 31, pp. 634-41, May 1 2015.
[17] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nat Genet*, vol. 39, pp. 1167-73, Sep., 2007.
[18] B. Javidy, A. Hatamlou, and S. Mirjalili, "Ions motion algorithm for solving optimization problems," *Applied Soft Computing*, vol. 32, pp. 72-79, 2015.
[19] G. T. Lin, H. F. Tseng, C. K. Chang, L. Y. Chuang, C. S. Liu, C. H. Yang, et al., "SNP combinations in chromosome-wide genes are associated with bone mineral density in Taiwanese women," *Chin J Physiol*, vol. 51, pp. 32-41, Feb., 29, 2008.
[20] C. Schaffer, "Overfitting Avoidance as Bias," *Machine Learning*, vol. 10, pp. 153-178, Feb 1993.
[21] C. H. Yang, L. Y. Chuang, Y. H. Cheng, Y. D. Lin, C. L. Wang, C. H. Wen, et al., "Single nucleotide polymorphism barcoding to evaluate oral cancer risk using odds ratio-based genetic algorithms," *Kaohsiung J Med Sci*, vol. 28, pp. 362-8, July. 2012.

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 1988and 1992, respectively. His main areas of research are evolutionary computation, bioinformatics, and assistive tool implementation.

Kuo-Chuan Wu is a PhD. student of Department of Electronic Engineering at National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan. He received his Master degree from the Department of Computer Science and Information Engineering at National Kaohsiung University of Applied Sciences. His research interests include pattern classification, 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.