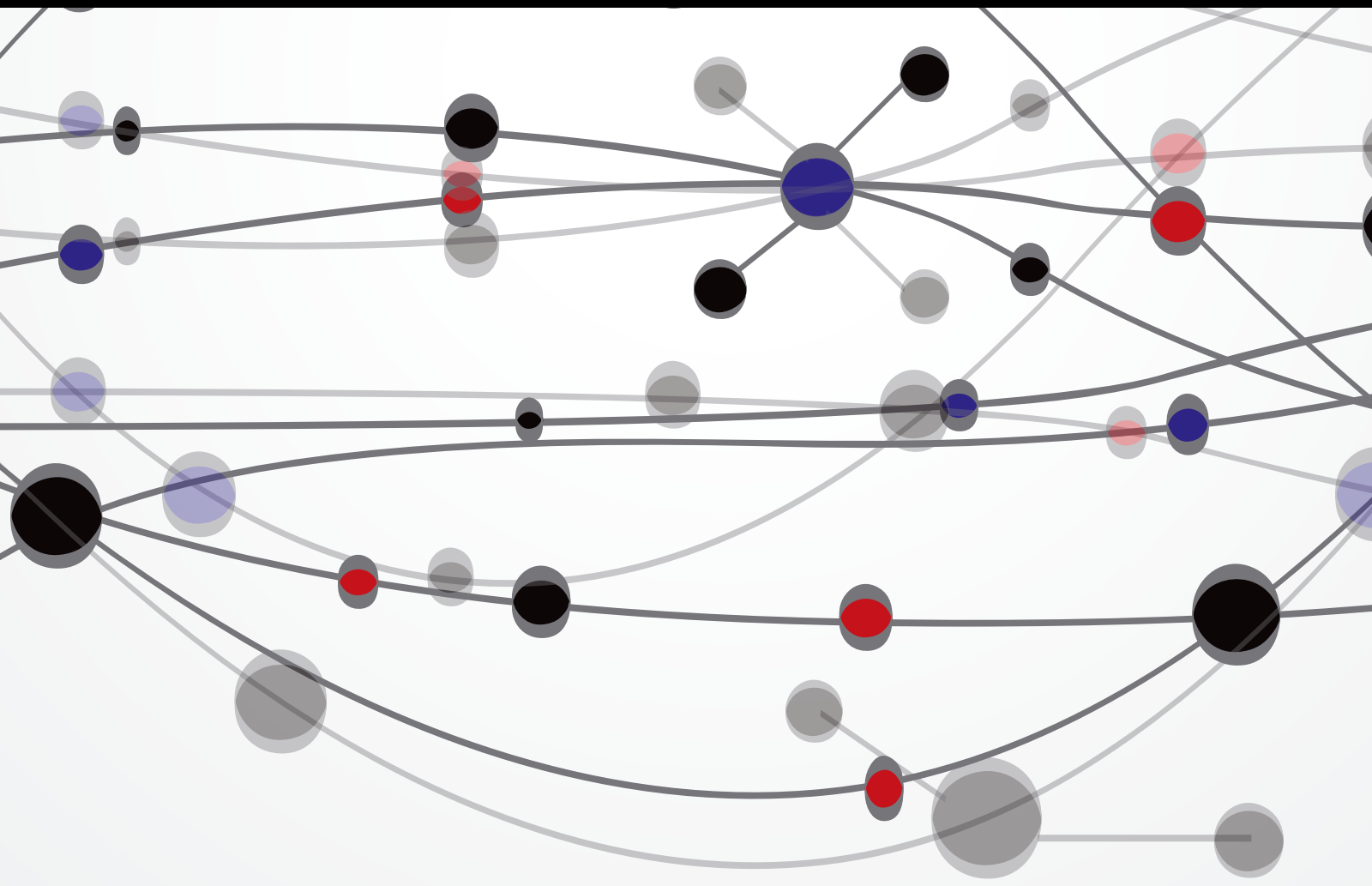
The Scientific World Journal

Biomedical Informatics and Computational Biology for High-Throughput Data Analysis

Guest Editors: Bairong Shen, Jianbo Lei, Jian Ma, Jiajun Wang, and Junbai Wang



Biomedical Informatics and Computational Biology for High-Throughput Data Analysis

The Scientific World Journal

Biomedical Informatics and Computational Biology for High-Throughput Data Analysis

Guest Editors: Bairong Shen, Jian Ma, Jiajun Wang, and Junbai Wang

Copyright © 2014 Hindawi Publishing Corporation. All rights reserved.

This is a special issue published in "The Scientific World Journal." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Contents

Biomedical Informatics and Computational Biology for High-Throughput Data Analysis, Bairong Shen, Jian Ma, Jiajun Wang, and Junbai Wang
Volume 2014, Article ID 398181, 2 pages

Tumor Necrosis Factor-α as a Diagnostic Marker for Neonatal Sepsis: A Meta-Analysis, Bokun Lv, Jie Huang, Haining Yuan, Wenying Yan, Guang Hu, and Jian Wang
Volume 2014, Article ID 471463, 14 pages

Genome-Wide Characterisation of Gene Expression in Rice Leaf Blades at 25°C and 30°C, Zhi-guo E, Lei Wang, Ryan Qin, Haihong Shen, and Jianhua Zhou
Volume 2014, Article ID 917292, 7 pages

Prediction and Analysis of Surface Hydrophobic Residues in Tertiary Structure of Proteins, Shambhu Malleshappa Gowder, Jhinuk Chatterjee, Tanusree Chaudhuri, and Kusum Paul
Volume 2014, Article ID 971258, 7 pages

Nonlinear-Model-Based Analysis Methods for Time-Course Gene Expression Data, Li-Ping Tian, Li-Zhi Liu, and Fang-Xiang Wu
Volume 2014, Article ID 313747, 9 pages

Adaptive $L_{1/2}$ Shooting Regularization Method for Survival Analysis Using Gene Expression Data, Xiao-Ying Liu, Yong Liang, Zong-Ben Xu, Hai Zhang, and Kwong-Sak Leung
Volume 2013, Article ID 475702, 5 pages

Molecular Dynamic Simulation to Explore the Molecular Basis of Btk-PH Domain Interaction with Ins(1,3,4,5)P4, Dan Lu, Junfeng Jiang, Zhongjie Liang, Maomin Sun, Cheng Luo, Bairong Shen, and Guang Hu
Volume 2013, Article ID 580456, 10 pages

Application of Bioinformatics in Chronobiology Research, Robson da Silva Lopes, Nathalia Maria Resende, Adenilda Cristina Honorio-França, and Eduardo Luzía França
Volume 2013, Article ID 153839, 8 pages

NCK2 Is Significantly Associated with Opiates Addiction in African-Origin Men, Zhifa Liu, Xiaobo Guo, Yuan Jiang, and Heping Zhang
Volume 2013, Article ID 748979, 8 pages

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 398181, 2 pages
http://dx.doi.org/10.1155/2014/398181

Editorial

Biomedical Informatics and Computational Biology for High-Throughput Data Analysis

Bairong Shen,[1] Jian Ma,[2] Jiajun Wang,[3] and Junbai Wang[4]

[1] *Center for Systems Biology, Soochow University, P.O. Box 206, Suzhou 215006, China*
[2] *Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[3] *School of Electronics and Information Engineering, Soochow University, Suzhou 215006, China*
[4] *Department of Pathology, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, 0310 Oslo, Norway*

Correspondence should be addressed to Bairong Shen; bairong.shen@suda.edu.cn

Received 8 December 2013; Accepted 8 December 2013; Published 12 February 2014

Copyright © 2014 Bairong Shen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The last few years have witnessed the emergence of several high-throughput (HTP) platforms that are based on the evolution of omic science through transcriptomics, proteomics, and metabolomics. With the huge data accumulation, the informatics and computational methods have become essential to understand the complexity of biomedical data. We lunched the special issue to address the demand for statistical models, biological omic data analyses, and meta-analysis of biomarker for complex diseases.

Statistical models are always important for the understanding of complex data; in this issue, Y. Liang et al. proposed a new adaptive L1/2 shooting regularization method for variable selection based on the Cox's proportional hazards mode. Simulation and the real gene expression dataset analysis showed that the method is more accurate for variable selection than Lasso and adaptive Lasso methods. L. Tian et al. employed a nonlinear model to analyze time course gene expression data. They firstly developed a method for estimating the parameters in the nonlinear model and then utilized the model to perform the significance analysis of individually differentially expressed genes and clustering analysis of a set of gene expression profiles. The simulation and real-life biological data analyses showed that their methods outperform some existing methods.

For the application of bioinformatics tools, Z. E et al. analyzed the gene expression in rice leaf blades at different temperatures. They analyzed the next-generation sequencing (NGS) datasets and characterized the transcription profiles of the samples in rice seedling leaf blades at 25°C and 30°C. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases were applied to study the effect of the temperature. Finally, they observed that temperature markedly regulates several superfamilies of transcription factors, including bZIP, MYB, and WRKY. Chronobiology is a field of biology that studies biological rhythms. R. Lopes et al. reviewed the bioinformatics or computational systems biology methods, tools, and databases which are very helpful to the understanding of the patterns and biological rhythms found in living organisms and the review will benefit the community for the application of bioinformatics in their future researches.

For the understanding of protein structure and structure-function relationship, two papers are selected in this issue. M. Shambhu et al. studied the surface accessibility of hydrophobic residues by considering their conservation score and knowledge of flanking regions. The accuracy of prediction is therefore improved. In the other work, D. Lu et al. applied molecular dynamics to investigate the effects of X-linked agammaglobulinaemia associated with amino acid mutations in pleckstrin homology (PH) domain on the binding of the domain with inositol 1,3,4,5-tetrakisphosphate (Ins(1,3,4,5)P4). The mutations were then classified as "functional mutations", and "folding mutations." The result provided new insights into the biological function of the Btk-PH domain and related mutation-causing diseases.

Finally, two works in this special issue are medical studies. The first is the work by B. Lv et al. They analyzed the 347 studies for the diagnostic value of tumor necrosis factor-α

(TNF-α) and found that the northern hemisphere group in the TNF-α test has higher sensitivity and specificity than those of southern hemisphere group. The second medical study is an association study. Z. Liu et al. reported the significantly association of NCK2 with opiates addiction in African-origin men by a comprehensive analysis of a dataset from the Study of Addiction: Genetics and Environment (SAGE). They employed both SNP and gene based methods of analysis and identified a strong and significant association between a SNP in the NCK2 gene on chromosome 2 with opiates addiction in African-origin men.

Bairong Shen
Jian Ma
Jiajun Wang
Junbai Wang

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 471463, 14 pages
http://dx.doi.org/10.1155/2014/471463

Research Article

Tumor Necrosis Factor-α as a Diagnostic Marker for Neonatal Sepsis: A Meta-Analysis

Bokun Lv,[1,2] Jie Huang,[1] Haining Yuan,[1,2] Wenying Yan,[3] Guang Hu,[2] and Jian Wang[1]

[1] *Systems Sepsis Team, Soochow University Affiliated Children's Hospital, Suzhou 215003, China*
[2] *Center for Systems Biology, Soochow University, Suzhou 215006, China*
[3] *Suzhou Zhengxing Translational Biomedical Informatics Ltd., Taicang 215400, China*

Correspondence should be addressed to Jian Wang; wangjian_sdfey@sina.com

Received 15 August 2013; Accepted 3 October 2013; Published 11 February 2014

Academic Editors: B. Shen, J. Wang, and J. Wang

Copyright © 2014 Bokun Lv et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Neonatal sepsis (NS) is an important cause of mortality in newborns and life-threatening disorder in infants. The meta-analysis was performed to investigate the diagnosis value of tumor necrosis factor-α (TNF-α) test in NS. Our collectible studies were searched from PUBMED, EMBASE, and the Cochrane Library between March 1994 and August 2013. Accordingly, 347 studies were collected totally, in which 15 articles and 23 trials were selected to study the NS in our meta-analysis. The TNF-α test showed moderate accuracy of the diagnosis of NS both in early-onset neonatal sepsis (sensitivity = 0.66, specificity = 0.76, $Q^* = 0.74$) and in late-onset neonatal sepsis (sensitivity = 0.68, specificity = 0.89, $Q^* = 0.87$). We also found the northern hemisphere group in the test has higher sensitivity (0.84) and specificity (0.83). A diagnostic OR analysis found that the study population may be the major reason for the heterogeneity. Accordingly, we suggest that TNF-α is also a valuable marker in the diagnosis of NS.

1. Introduction

Neonatal sepsis (NS), as one of the major causes, leads to neonatal mortality and morbidity, especially in neonates born preterm [1]. Early diagnosis and management of the newborn infant with NS play key roles in preventing severe and life-threatening complications [2]. During the first hours of life, reliable infection markers are absent in NS. Therefore, neonatologists often begin early antibiotic treatment in newborn infants with risk factors for infection, exposing many neonates to unnecessary treatments. Due to the limitation of the treatment strategy in early diagnosis of sepsis, the isolation of causative organisms from microbiological cultures takes up to 72 h, which cannot be used to identify most infected infants [3].

Procalcitonin (PCT), a 116-amino acid peptide considered as a precursor in calcium homeostasis, has been proved to be a valuable marker for distinguishing sepsis from other noninfectious disease [4, 5]. C-reactive protein (CRP), another biomarker, is a protein discovered in the blood and whose levels will rise after inflammation [6–8]. To date, CRP and PCT have been proposed for inclusion as two mostly used diagnostic markers in the international definition of sepsis [9, 10]. Recently, many studies have found that some new markers also play important roles in the diagnosis of neonatal sepsis. However, the systematic research and comparison of these biomarkers for diagnosing NS are limited. For example, we have investigated the diagnosis value of serum amyloid A (SAA) in NS [11]. Here, we continue to evaluate the value of the tumor necrosis factor-α (TNF-α), by considering it as a useful marker. TNF-α is a cytokine involved in systemic inflammation, which belongs to a member of a group of cytokines that stimulate the acute phase reaction [12].

Thus, the objective of this meta-analysis was to investigate the value of TNF-α for detecting NS. Although a lot of works indicate that both PCT and CRP are two superior markers for diagnosis of sepsis and infection, we suggest here the TNF-α is also a promising marker in NS. A deeper meta-analysis of these studies is thus currently needed.

2. Methods

2.1. Studies Retrieval and Selection. In order to perform a systematic analysis of the available evidence on the efficacy of

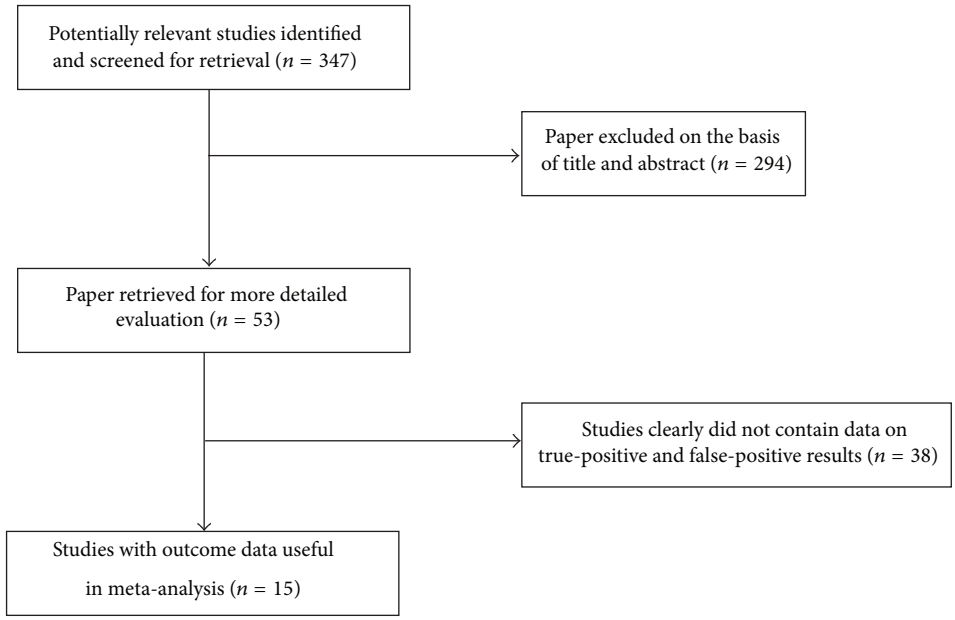


Figure 1: Summary of study assessment and inclusion in the meta-analysis of studies involving diagnosis of neonatal sepsis using a TNF-α test.

TNF-α in NS [13], the common approach of literature search was performed in PUBMED, EMBASE, and the Cochrane Library for relevant citations from March 1994 to August 2013. The search terms used were "TNF-α," "neonatal," "neonate," "sepsis," "infant," "newborn," and "tumor necrosis factors-α". The reference lists of all known primary and review articles were also searched. No language restriction was used, so that we have examined the references of known articles to fully retrieve the data.

If an article does not include enough data for calculating sensitivity and specificity (2 × 2 table), we asked the corresponding author to provide us with necessary data. If there was no response from the corresponding author, a reminder was sent after one week. If we still cannot achieve the data after this process, the study was excluded from meta-analysis. The selection of articles was performed by two investigators independently to ensure the high accuracy.

2.2. Data Extraction. Data collected from the studies included the first author, publication year, diagnostic cut-off point and time, test methods, and sensitivity and specificity data. So the numbers of true-positive, false-positive, false-negative, and true-negative results were extracted for each study. Accuracy data was extracted to construct 2 × 2 table at a specific time. We have requested the information from the authors, if no enough data on the criteria was found in the studies.

2.3. Statistical Analysis. We used Meta-Disc 1.4 software and Review Manager 5.0 to perform the statistical analysis [14]. Diagnoses were grouped into two groups according to the time of TNF-α test for diagnosis of NS. One group is the time points of TNF-α measurement for the diagnosis of early-onset neonatal sepsis (EONS), which were 0–72 h of age; the other group is the time points of TNF-α measurement for the diagnosis of late-onset neonatal sepsis (LONS), in which the age of neonates is older than 72 h. We calculated the sensitivity, specificity, diagnostic odds ratio (OR), and corresponding 95% confidence intervals (CI) from each study. We also gained the pooled sensitivity, specificity, and diagnosis OR from each group.

The diagnostic OR expresses how much greater the odds of having sepsis are for newborns who have a positive test result, relatively to newborns who have a negative result [15]. For the estimates of diagnostic OR, heterogeneity was assessed by using the Cochrane Q statistic. Normally, I^2 lies between 0% and 100%. If $I^2 < 50\%$, then there is a lot of homogeneity among studies in meta-analysis; whereas $I^2 > 50\%$ shows there is more heterogeneity among studies. A value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity [16]. We explored the reasons for heterogeneity by carrying out the subgroup analysis and examined characteristics of included studies.

In order to summarize these results, we constructed summary receiver operator characteristic (SROC) curves, which showed the relationship between sensitivity and the false positives (1-specificity). Q^* values was received from the SROC curves. Meanwhile, the area under the (SROC) curves was also calculated from the SROC curves, which have been proposed as a way to assess diagnostic data in the context of a meta-analysis [17].

3. Results

3.1. Study Selection. The literature search was completed in August 2013. We found 347 potentially relevant restudies, but only 15 articles met our inclusion criteria. Figure 1 shows

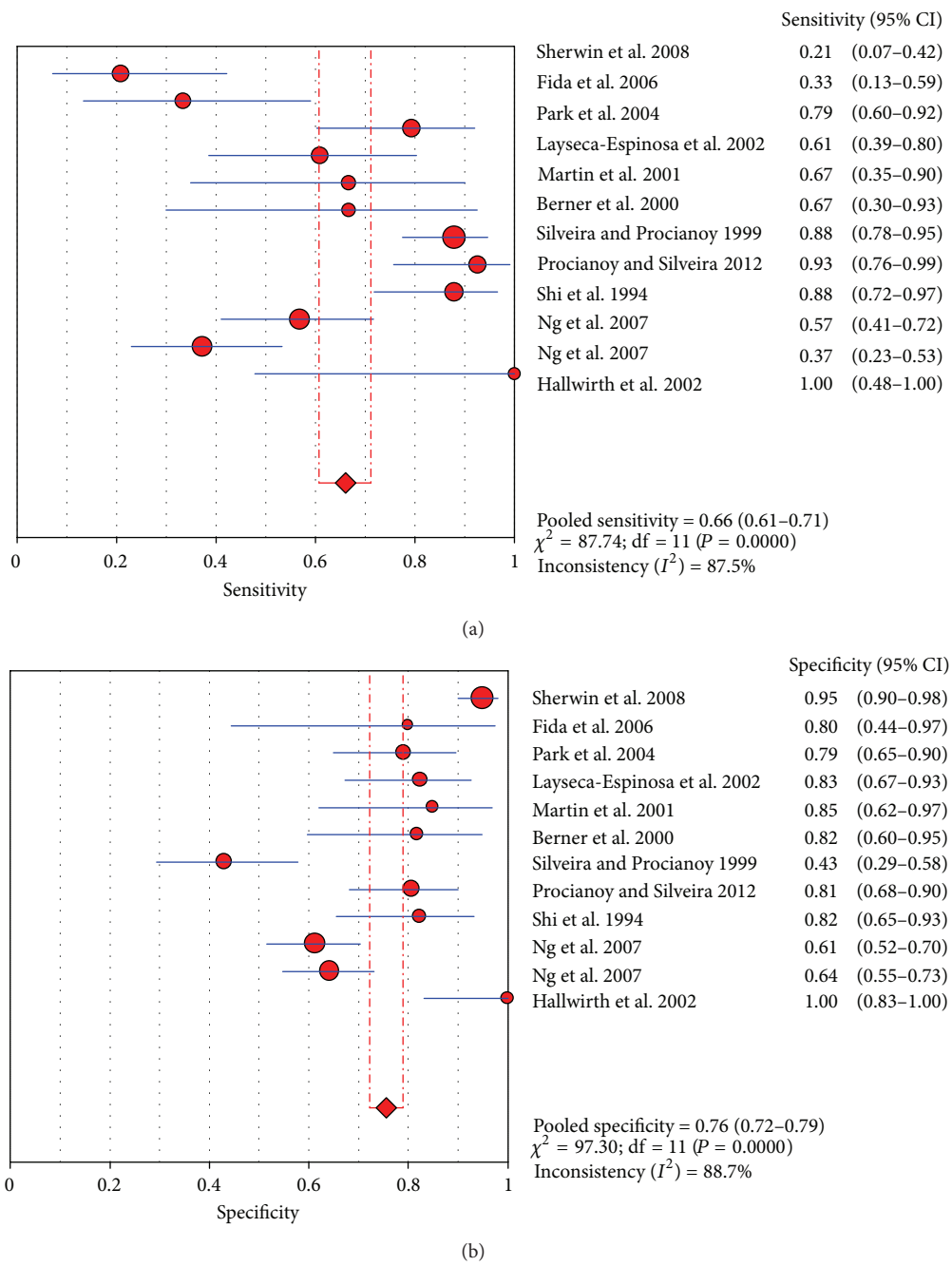


Figure 2: Forest plot for sensitivity and specificity of the TNF-α test to diagnose neonatal sepsis at the EONS.

the chart of literature search. Detailed information for each included study is presented in Table 1.

3.2. Accuracy of the TNF-α Test in the Diagnosis of Proven Early-Onset Neonatal Sepsis. Eleven articles and twelve trials were included to estimate the use of the TNF-α test in the diagnosis of proven early-onset neonatal sepsis (EONS). EONS was defined as the clinical sepsis in the 0–72 h after delivery, met the inclusion criteria in [18–27].

In these trials, we can get the TP, TN, FP, FN, sensitivity, specificity, DOR, PPV, and NPV from the articles. The sensitivity ranged from 20.8% to 100% and pooled sensitivity is 66.1% (95% CI 60.7%–70.1%), specificity ranged from 43.1% to 100% and pooled specificity is 75.6% (95% CI 72.2%–78.9%), and the detailed forest map is shown in Figure 2. We calculated the significant heterogeneity among studies (sensitivity, I^2 = 87.5%; specificity, I^2 = 88.7%); it indicated that patient selection or other relevant factors might be responsible for heterogeneity.

The value of DOR of TNF-α was 7.43 (95%CI 3.47–15.90), as shown in Figure 3. In these articles, we calculated the significant heterogeneity (I^2 = 77.9%). The SROC curve for

Table 1: Characteristics of studies included in the meta-analysis of the diagnosis of neonatal sepsis using a TNF-α test.

Study	Study population	Patients (n)	Sepsis diagnosis	Cut-off (pg/mL)	Time	Country
Hotoura et al. 2012 [1]	Cases: newborns with suspected sepsis Control: infection-free infants	82	Culture; clinical	30	LONS	Greece
Hotoura et al. 2012 [1]	Cases: newborns with suspected sepsis Control: infection-free infants	82	Culture; clinical	15	LONS	Greece
Sherwin et al. 2008 [28]	Cases: NICU newborns with suspected sepsis Control: neonates without sepsis	164	Culture; clinical	180	EONS	New Zealand
Sherwin et al. 2008 [28]	Cases: NICU newborns with suspected sepsis Control: neonates without sepsis	164	Culture; clinical	70	LONS	New Zealand
Sherwin et al. 2008 [28]	Cases: NICU newborns with suspected sepsis Control: neonates without sepsis	164	Culture; clinical	180	LONS	New Zealand
Kocabaş et al. 2007 [14]	Cases: neonates with a suspected clinical sepsis Control: healthy neonates without infectious	55	Culture; clinical	7.5	LONS	Turkey
Fida et al. 2006 [18]	Cases: neonates with clinical or proven or possible infected sepsis Control: disease without infection	28	Culture; clinical	29.86	EONS	Saudi Arabia
Park et al. 2004 [19]	Cases: newborns with suspected sepsis Control: neonates without sepsis	77	Culture; clinical	41	EONS	Korea
Layseca-Espinosa et al. 2002 [20]	Cases: neonates with clinical or proven sepsis Control: disease without infection	63	Culture; clinical	0.18	EONS	Spain
Martin et al. 2001 [21]	Cases: newborns with suspected sepsis Control: neonates without sepsis	32	Culture; clinical	20	EONS	Sweden
Berner et al. 2000 [22]	Cases: newborns with suspected sepsis Control: neonates without sepsis	31	Culture; clinical	48	EONS	Germany
Silveira and Procianoy 1999 [23]	Cases: newborn infants with clinical sepsis or probably infected with clinical sepsis Control: neonates without sepsis	117	Culture; clinical	12	EONS	Brazil
Ng et al. 1997 [29]	Cases: VLBW infants with suspected clinical sepsis Control: noninfected newborns	101	Culture; clinical	17	LONS	Hong Kong

Table 1: Continued.

Study	Study population	Patients (n)	Sepsis diagnosis	Cut-off (pg/mL)	Time	Country
Ng et al. 1997 [29]	Cases:VLBW infants with suspected clinical sepsis Control: noninfected newborns	101	Culture; clinical	17	LONS	Hong Kong
Ng et al. 1997 [29]	Cases: VLBW infants with suspected clinical sepsis Control: noninfected newborns	101	Culture; clinical	17	LONS	Hong Kong
Hotoura et al. 2011 [30]	Cases: full-term neonates with suspected or documented infection Control: infection-free infants	95	Culture; clinical	30	LONS	Greece
Hotoura et al. 2011 [30]	Cases: full-term neonates with suspected or documented infection Control: infection-free infants	95	Culture; clinical	15	LONS	Greece
Hallwirth et al. 2002 [24]	Cases: neonates with sepsis Control: neonates without sepsis	25	Culture; clinical	20000	LONS	Austria
Procianoy and Silveira 2012 [25]	Cases: very low birth weight infants with clinical sepsis Control: neonates without sepsis	84	Culture; clinical	30	EONS	Brazil
Shi et al. 1994 [26]	Cases: neonates with sepsis Control: neonates without sepsis	67	Culture; clinical	267.2	EONS	CHINA
Ng et al. 2007 [27]	Cases: very low birth weight infants with suspected sepsis Control: neonates without sepsis	155	Culture; clinical	0.6	EONS	Hong Kong
Ng et al. 2007 [27]	Cases: very low birth weight infants with suspected sepsis Control: neonates without sepsis	155	Culture; clinical	0.6	EONS	Hong Kong
Hallwirth et al. 2002 [24]	Cases: neonates with sepsis Control: neonates without sepsis	25	Culture; clinical	20000	EONS	Austria

TNF-α markers was plotted in Figure 4; the AUC was 0.81 with the standard error being 0.04. The pooled diagnostic accuracy (Q^*) of 0.7430 with the standard error was 0.04.

3.3. Accuracy of the TNF-α Test in the Diagnosis of Proven Late-Onset Neonatal Sepsis. Six articles and eleven trials were included to estimate the use of the TNF-α test in the diagnosis of proven late-onset neonatal sepsis (LONS). LONS was defined as the clinical sepsis 72 h after birth, met the inclusion criteria in [1, 14, 24, 28–30].

In these trials, we can also get much information from the articles. The sensitivity ranged from 23.1% to 100% and pooled sensitivity is 68.0% (95% CI 62.8%–72.8%), specificity ranged from 73.2% to 100% and pooled specificity is 88.5% (95% CI 85.9%–90.7%), and the detailed forest map is shown in Figure 5. We calculated the significant heterogeneity among studies (sensitivity, $I^2 = 91.9\%$; specificity, $I^2 = 87.5\%$), which indicated that patient selection or other relevant factors might be responsible for heterogeneity.

The value of DOR of TNF-α was 37.44 (95% CI 19.07–73.48), as shown in Figure 6. In these articles, we calculated the significant heterogeneity ($I^2 = 41.6\%$). The SROC curves for TNF-α markers were plotted in Figure 7; the AUC was 0.93 with the standard error being 0.017. The pooled diagnostic accuracy (Q^*) of 0.8696 with the standard error was 0.02.

3.4. Intensive Study of the TNF-α Test in the Diagnosis of Proven Late-Onset Neonatal Sepsis. In the LONS study, study

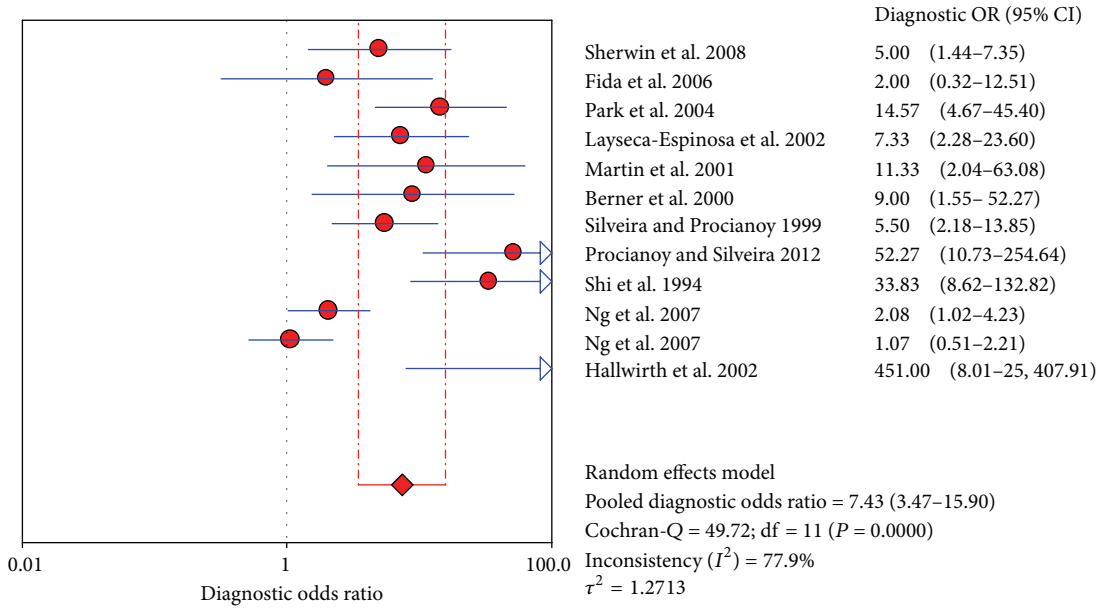


Figure 3: Forest plot for diagnostic OR of the TNF-α test to diagnose neonatal sepsis at the EONS.

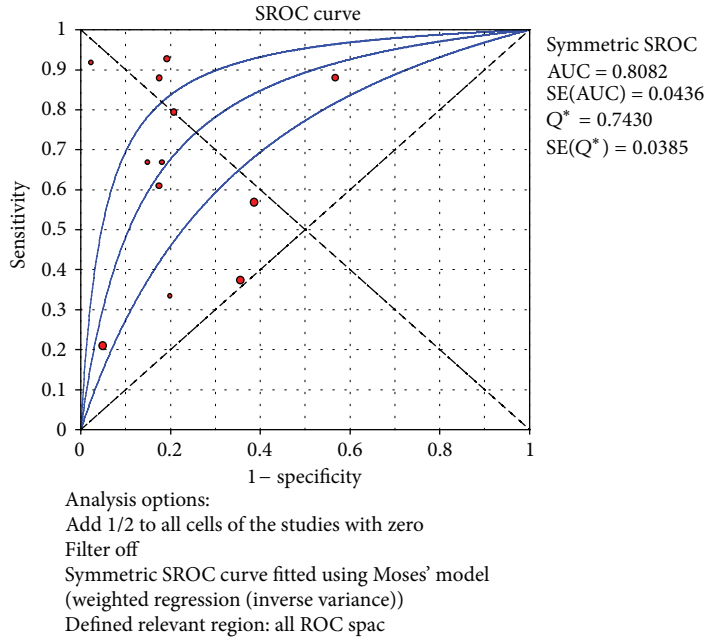


Figure 4: Summary receiver operating characteristic (SROC) curve of the TNF-α test for the diagnosis of early-onset neonatal sepsis. Each point represents one study in the SROC curve.

populations come from different countries, but in general they can be further divided into two regions: the northern hemisphere and the southern hemisphere.

Eight trials were included to estimate the use of the TNF-α test in the northern hemisphere at the diagnosis of proven late-onset neonatal sepsis [1, 14, 29, 30]. In these trials, sensitivity ranged from 61.4% to 100% and pooled sensitivity is 84.0% (95% CI 78.8%–88.4%), specificity ranged from 68.8% to 96.6% and pooled specificity is 83.3% (95% CI 79.6%–86.6%), and the detailed forest maps are shown in Figure 8. We calculated the significant heterogeneity among studies (sensitivity, I^2 = 77.4%; specificity, I^2 = 76.3%), which indicated that patient selection or other relevant factors might be responsible for heterogeneity.

The value of DOR of TNF-α test in the northern hemisphere at the diagnosis of proven late-onset neonatal sepsis was 44.94 (95% CI 20.71–97.50), as shown in Figure 9. In these articles, we calculated the significant heterogeneity (I^2 = 47.1%). The SROC curves for TNF-α markers were plotted in Figure 10; the AUC was 0.93 with the standard error

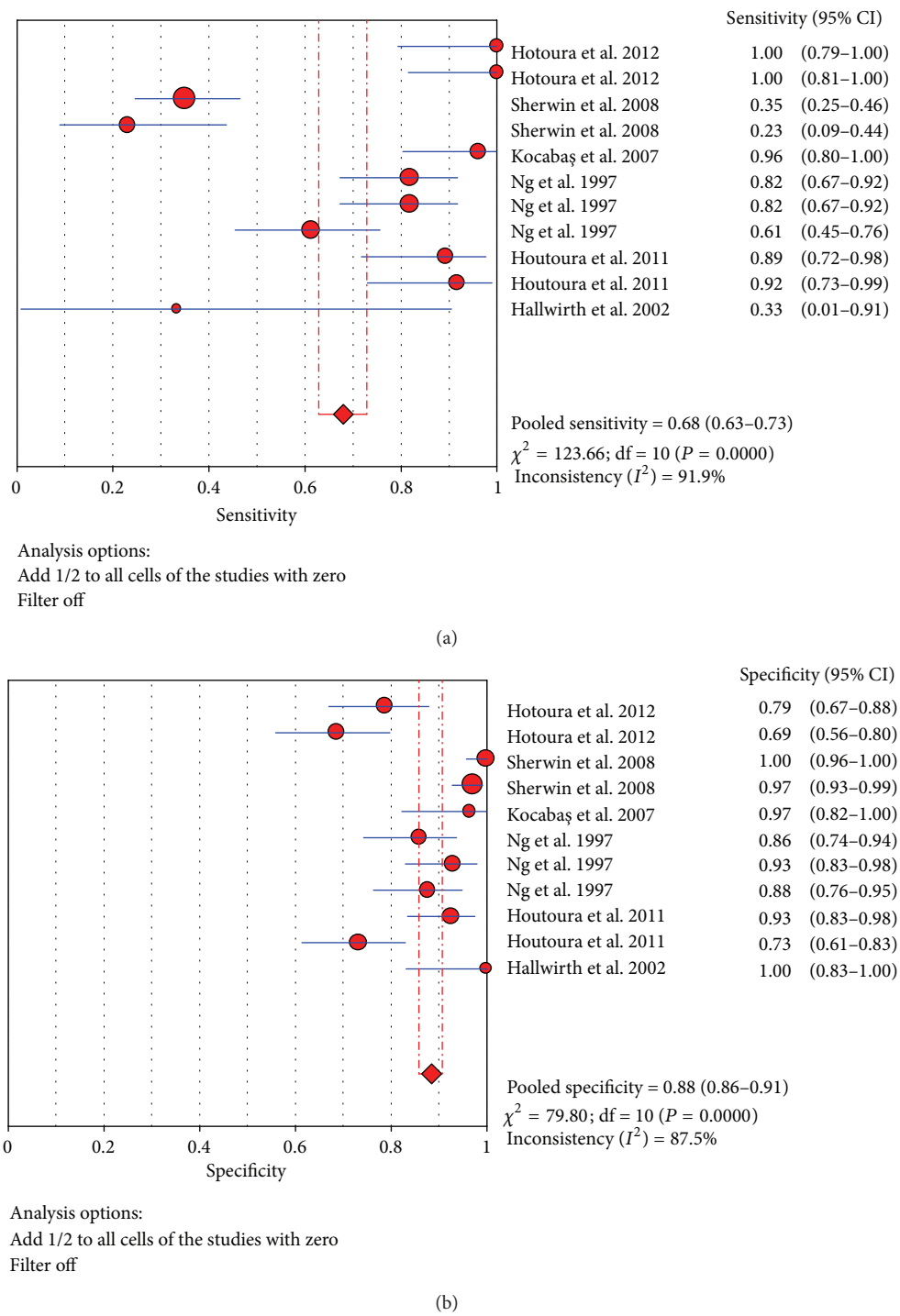


Figure 5: Forest plot for sensitivity and specificity of the TNF-α test to diagnose neonatal sepsis at the LONS.

being 0.017. The pooled diagnostic accuracy (Q^*) of 0.8710 with the standard error was 0.02.

Three trials were included to estimate the use of the TNF-α test in the southern hemisphere at the diagnosis of proven late-onset neonatal sepsis [24, 28]. In these trials, sensitivity ranged from 23.1% to 35% and pooled sensitivity is 32.1% (95% CI 23.5%–41.7%), specificity ranged from 97.1% to 100% and pooled specificity is 98.3% (95% CI 95.8%–99.5%), and the detailed forest map is shown in Figure 11. We calculated the significant heterogeneity among studies (sensitivity, I^2 = 0%; specificity, I^2 = 56.0%), which indicated that patient selection or other relevant factors might be responsible for heterogeneity.

The value of DOR of TNF-α test in the northern hemisphere at the diagnosis of proven late-onset neonatal sepsis was 20.88 (95% CI 3.84–113.49), as shown in Figure 12. In

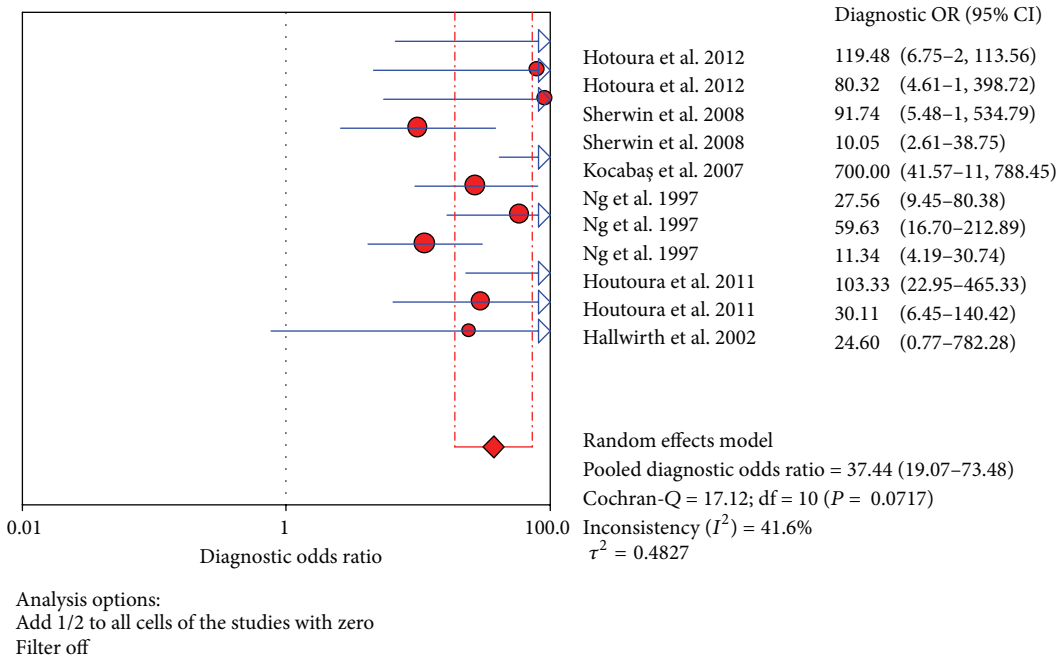


Figure 6: Forest plot for diagnostic OR of the TNF-α test to diagnose neonatal sepsis at the LONS.

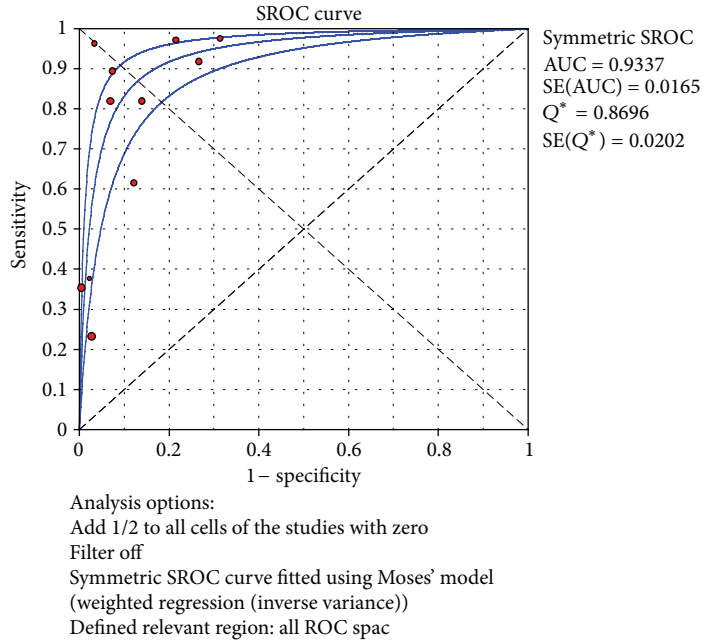
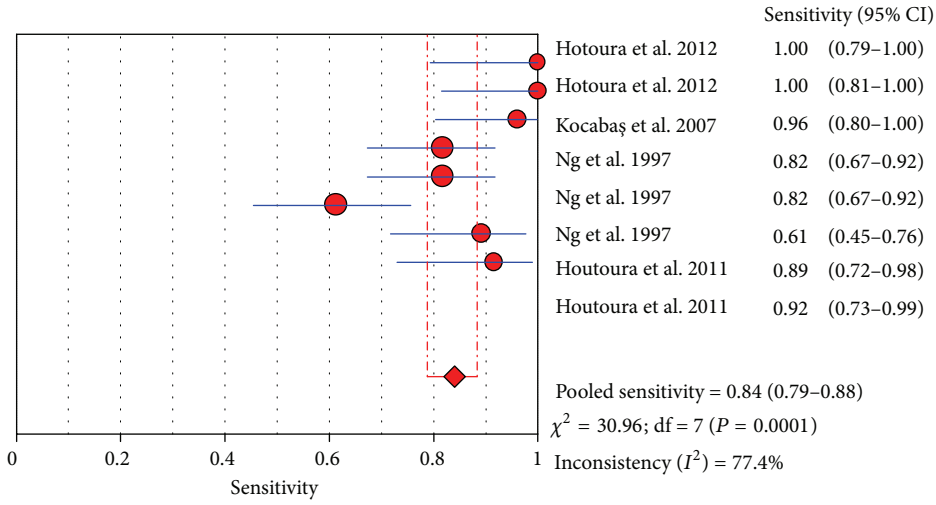


Figure 7: Summary receiver operating characteristic (SROC) curve of the TNF-α test for the diagnosis of late-onset neonatal sepsis. Each point represents one study in the SROC curve.

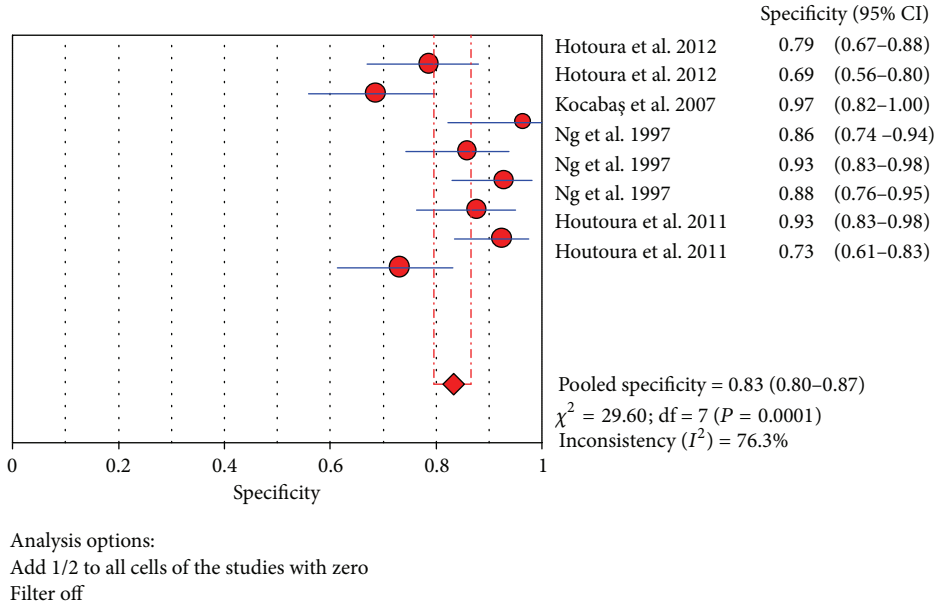
these articles, we calculated the significant heterogeneity (I^2 = 35.5%). The SROC curves for TNF-α markers were plotted in Figure 13; the AUC was 0.0468 with the standard error being 0.224. The pooled diagnostic accuracy (Q^*) of 0.1052 with the standard error was 0.309.

3.5. Comparison of the Diagnostic Accuracy of Markers for Neonatal Sepsis. In LONS, PCT and CRP have been proven to be useful markers for the diagnosis of NS [3, 9, 31]. In order to show the value of diagnosis of the TNF-α test for NS in LONS, we compared TNF-α with PCT and CRP in LONS. Six articles and eleven trials were used to evaluate the diagnosis of TNF-α. Compared with 55% (95% CI 45%–65%) for the CRP test and 72% (95% CI 63%–81%) for the PCT test [14], the pooled sensitivity for the TNF-α test was 68% (95% CI 63%–73%).

The pooled specificity for the TNF-α was slightly higher than for the CRP and PCT test (88.5% (95% CI 86%–91%) versus 85% (95% CI 81%–88%) versus 77% (95% CI 72%–81%)).



(a)



(b)

Figure 8: Forest plot for sensitivity and specificity of the TNF-α test to diagnose neonatal sepsis at the LONS in the northern hemisphere.

Furthermore, the pooled diagnostic OR for the TNF-α was higher than CRP and PCT (37.4 (95% CI 19.1–73.5) versus 8.6 (95% CI 3.5–21.0) versus 11.6 (95% CI 5.2–26.0)). The Q^* value was slightly higher for the TNF-α than CRP and PCT (0.87 versus 0.75 versus 0.77). In SROC curve the TNF-α's AUC is almost equal to CRP (0.93 versus 0.96). As many articles reported that PCT and CRP are good markers, the TNF-α is also a good marker for the diagnosis of NS in LONS.

3.6. Analysis of Heterogeneity. Heterogeneity is very critical in a meta-analysis, so we should try to explore the reason for the heterogeneity. Generally speaking, variations include several influence factors, for instance the cut-off value, study population, inject antibodies, and so forth.

Firstly, we consider the cut-off value. In our study, many data are so large, so we suppose that this is the possible reason for the heterogeneity. We excluded two studies of [24, 28] whose cut-off values were relatively large. But the I^2 (heterogeneity test) almost does not change (41.7% versus 41.6%). So the cut-off value is not the reason for the heterogeneity. Secondly, we consider the study population. The I^2 is reduced from 41.6% to 0% when two studies of

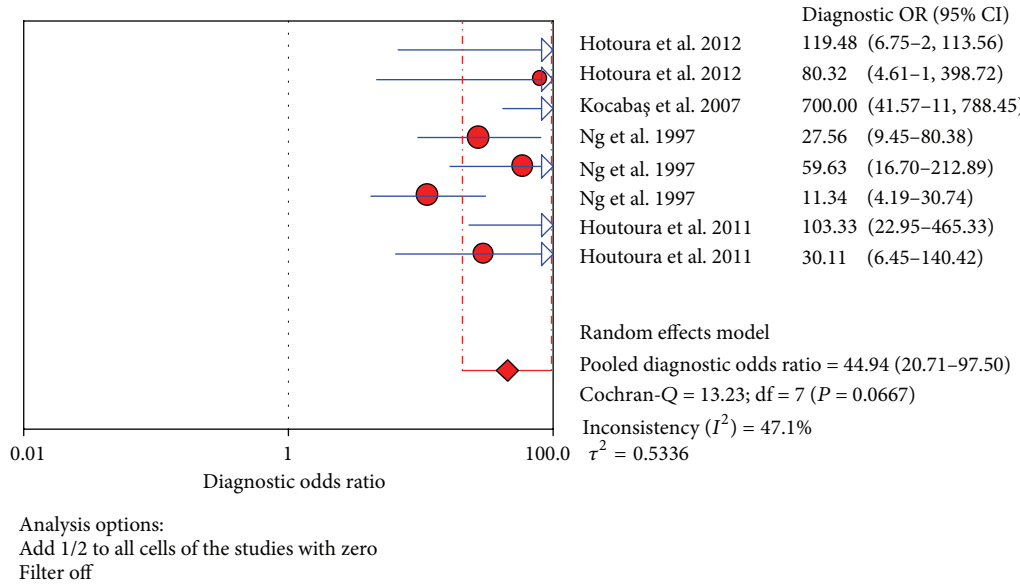


Figure 9: Forest plot for diagnostic OR of the TNF-α test to diagnose neonatal sepsis at the LONS in the northern hemisphere.

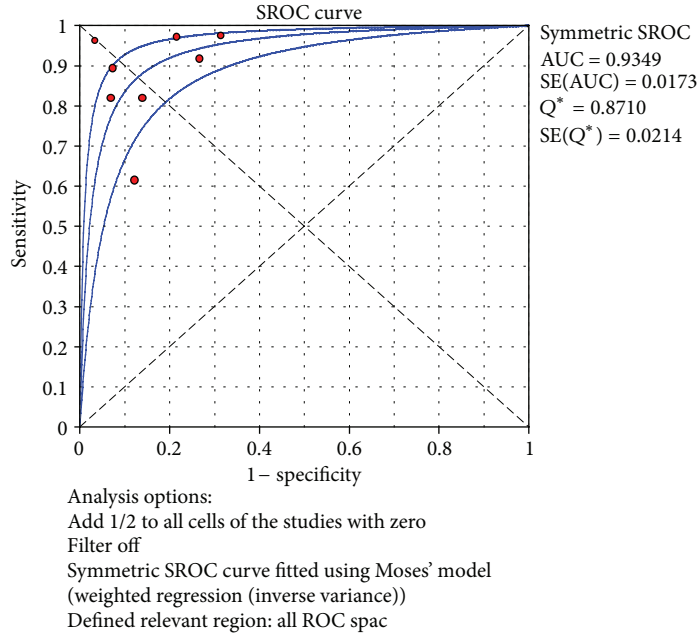


Figure 10: Summary receiver operating characteristic (SROC) curve of the TNF-α test for the diagnosis of late-onset neonatal sepsis in the northern hemisphere. Each point represents one study in the SROC curve.

[28, 29] were excluded from the meta-analysis. Their study population is different from others. So the study population may be the major reason for the heterogeneity.

3.7. Publication Bias. The publication bias is difficult to avoid in meta-analysis [32]. In each study, we often choose favorable results and give up negative results; many "blindness" of test results was never reported. The limit of current available data may bias our conclusion.

4. Discussion

NS is one of the most common diseases and life-threating disorder in neonate, and thus it can bring the high mortality and morbidity in infants. So the identification of biomarkers is very important to improve the diagnosis of NS. The clinical signs are nonspecific and laboratory indicators including blood culture are not reliable [33]. The sensitivities of markers are not always so high [34]. So it is necessary to find a good marker for NS.

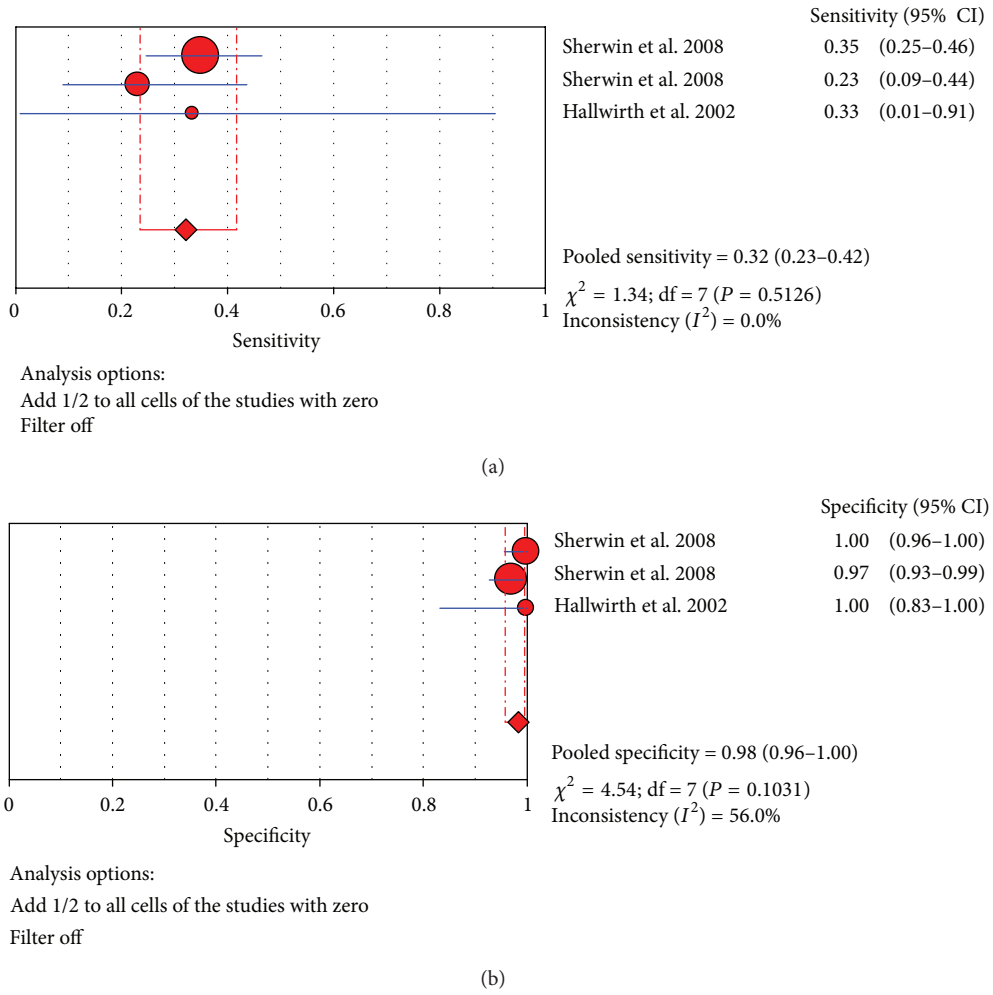


Figure 11: Forest plot for sensitivity and specificity of the TNF-α test to diagnose neonatal sepsis at the LONS in the southern hemisphere.

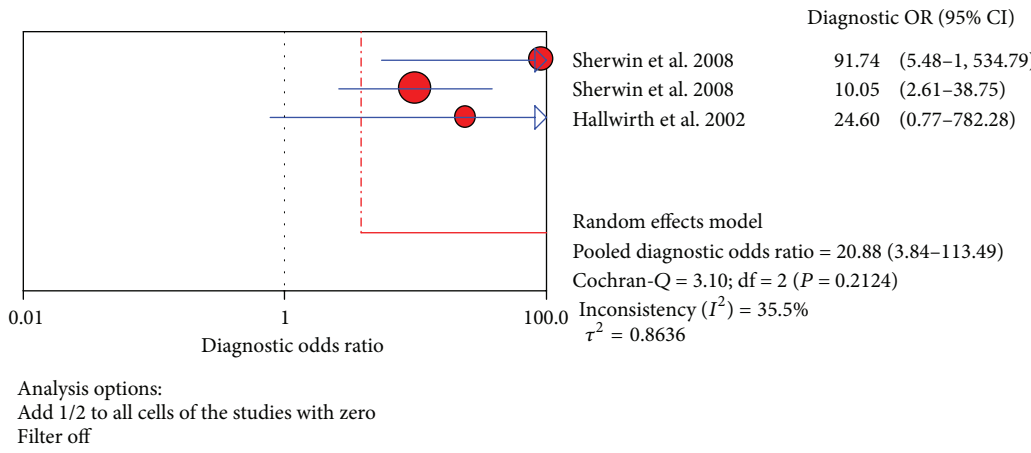


Figure 12: Forest plot for diagnostic OR of the TNF-α test to diagnose neonatal sepsis at the LONS in the southern hemisphere.

It is well known that an excellent marker should have high sensitivity and specificity. In our meta-analysis, the TNF-α tests' sensitivity is 0.66 for the diagnosis of early-onset neonatal sepsis, and the specificity is 0.46 and the Q^* is 0.74. At the late-onset neonatal sepsis, the TNF-α test's sensitivity is 0.68, whereas the specificity is 0.89 and the Q^* is 0.87. In particular, TNF-α shows a higher accuracy for the diagnosis of NS in LONS. Therefore, we have further analyzed the regional issues in LONS. To this end, we have classified studies into two groups, that is, northern hemisphere

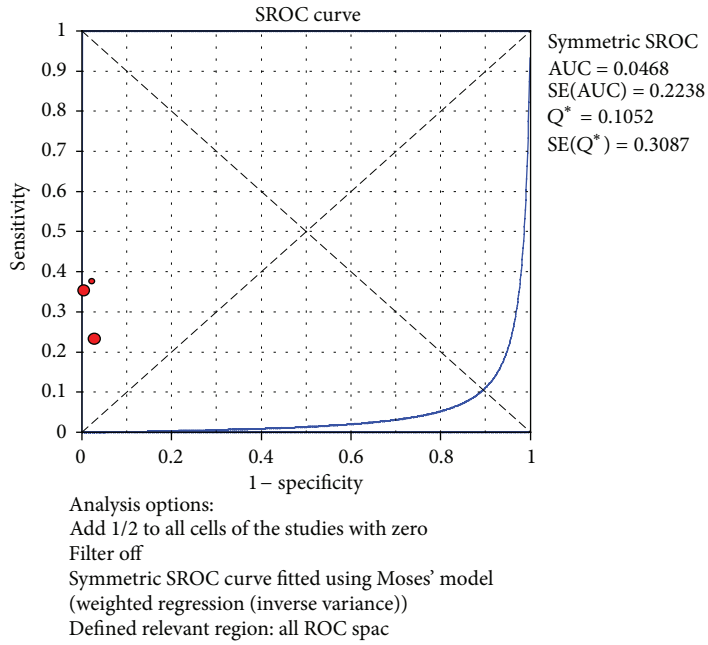


Figure 13: Summary receiver operating characteristic (SROC) curve of the TNF-α test for the diagnosis of late-onset neonatal sepsis in the southern hemisphere. Each point represents one study in the SROC curve.

group [1, 14, 29, 30] and southern hemisphere group [24, 28]. Our analysis found that the northern hemisphere group has higher sensitivity and specificity (sensitivity = 84%, specificity = 83%). The results show that the TNF-α has appropriate accuracy for the diagnosis of NS and thus is a good biomarker for the diagnosis of NS.

CRP is an excellent marker and has been applied in clinic [10]. The sensitivity of CRP is 30%–97%, and the specificity ranged from 75% to 100% [35]. In our meta-analysis, TNF-α' sensitivity ranged from 23.1% to 100%, and specificity ranged from 73.2% to 100%. The study from TNF-α is similar to CRP, so the TNF-α is a useful marker in the diagnosis of NS. In addition, PCT is more excellent marker which has better accuracy than CRP for the diagnosis of NS [3]. In our meta-analysis, the pooled sensitivity of TNF-α is slightly lower than that of the PCT test in EONS (66.1% versus 74%), the pooled specificity also in such (76% versus 86%). But in LONS, the pooled specificity of TNF-α is higher than that of the PCT test (89% versus 77%), although the pooled sensitivity is slightly lower than PCT (68% versus 72%). Generally speaking, the data of TNF-α and PCT is greater than CRP. This result again shows that TNF-α is good marker in the diagnosis of NS.

In conclusion, TNF-α shows the moderate accuracy in the diagnosis of NS, both in EONS and LONS. If we test the accuracy of TNF-α by further dividing data into two regions, the study in northern hemisphere shows a better result. Because of the relatively few testing data, the experiments results need to be further studied and the clinical validation is also needed.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Bokun Lv and Jie Huang equally contributed to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China Grants (81272143 and 21203131), the Natural Science Foundation of Jiangsu Province (K200509), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (12KJB180014), Jiangsu Innovation Team Grant LJ201141, and Program of Innovative and Entrepreneurial Talent.

References

[1] E. Hotoura, V. Giapros, A. Kostoula, P. Spyrou, and S. Andronikou, "Pre-inflammatory mediators and lymphocyte subpopulations in preterm neonates with sepsis," *Inflammation*, vol. 35, no. 3, pp. 1094–1101, 2012.

[2] U. K. Mishra, S. E. Jacobs, L. W. Doyle, and S. M. Garland, "Newer approaches to the diagnosis of early onset neonatal sepsis," *Archives of Disease in Childhood*, vol. 91, no. 3, pp. F208–F212, 2006.

[3] Z. Yu, J. Liu, Q. Sun, Y. Qiu, S. Han, and X. Guo, "The accuracy of the procalcitonin test for the diagnosis of neonatal sepsis: a meta-analysis," *Scandinavian Journal of Infectious Diseases*, vol. 42, no. 10, pp. 723–733, 2010.

[4] B. Uzzan, R. Cohen, P. Nicolas, M. Cucherat, and G.-Y. Perret, "Procalcitonin as a diagnostic test for sepsis in critically ill adults and after surgery or trauma: a systematic review and meta-analysis," *Critical Care Medicine*, vol. 34, no. 7, pp. 1996–2003, 2006.

[5] B. Müller, M. Christ-Crain, E. S. Nylen, R. Snider, and K. L. Becker, "Limits to the use of the procalcitonin level as a diagnostic marker," *Clinical Infectious Diseases*, vol. 39, no. 12, pp. 1867–1868, 2004.

[6] E. Bilavsky, H. Yarden-Bilavsky, S. Ashkenazi, and J. Amir, "C-reactive protein as a marker of serious bacterial infections in hospitalized febrile infants," *Acta Paediatrica*, vol. 98, no. 11, pp. 1776–1780, 2009.

[7] A. Kordek, M. Hałasa, and W. Podraza, "Early detection of an early onset infection in the neonate based on measurements of procalcitonin and C-reactive protein concentrations in cord blood," *Clinical Chemistry and Laboratory Medicine*, vol. 46, no. 8, pp. 1143–1148, 2008.

[8] L. Simon, F. Gauvin, D. K. Amre, P. Saint-Louis, and J. Lacroix, "Serum procalcitonin and C-reactive protein levels as markers of bacterial infection: a systematic review and meta-analysis," *Clinical Infectious Diseases*, vol. 39, no. 2, pp. 206–217, 2004.

[9] J. P. S. Caldas, S. T. M. Marba, M. H. S. L. Blotta, R. Calil, S. S. Morais, and R. T. D. Oliveira, "Accuracy of white blood cell count, C-reactive protein, interleukin-6 and tumor necrosis factor alpha for diagnosing late neonatal sepsis," *Jornal de Pediatria*, vol. 84, no. 6, pp. 536–542, 2008.

[10] M. M. Levy, M. P. Fink, J. C. Marshall et al., "2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference," *Intensive Care Medicine*, vol. 29, no. 4, pp. 530–538, 2003.

[11] H. Yuan, J. Huang, B. Lv et al., "Diagnosis value of the serum amyloid a test in neonatal sepsis: a meta-analysis," *BioMed Research International*, vol. 2013, Article ID 520294, 9 pages, 2013.

[12] L. P. Francesconi, K. M. Ceresér, R. Mascarenhas, L. Stertz, C. S. Gama, and P. Belmonte-de-Abreu, "Increased annexin-V and decreased TNF-alpha serum levels in chronic-medicated patients with schizophrenia," *Neuroscience Letters*, vol. 502, no. 3, pp. 143–146, 2011.

[13] W. M. Bernardo, F. T. Aires, R. M. Carneiro, F. P. Sá, V. E. Rullo, and D. A. Burns, "Effectiveness of probiotics in the prophylaxis of necrotizing enterocolitis in preterm neonates: a systematic review and meta-analysis," *Jornal de Pediatria*, vol. 89, no. 1, pp. 18–24, 2013.

[14] E. Kocabaş, A. Sarikçioğlu, N. Aksaray, G. Seydaoğlu, Y. Seyhun, and A. Yaman, "Role of procalcitonin, C-reactive protein, interleukin-6, interleukin-8 and tumor necrosis factor-α in the diagnosis of neonatal sepsis," *Turkish Journal of Pediatrics*, vol. 49, no. 1, pp. 7–20, 2007.

[15] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," *Journal of Clinical Epidemiology*, vol. 56, no. 11, pp. 1129–1135, 2003.

[16] J. P. T. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," *British Medical Journal*, vol. 327, no. 7414, pp. 557–560, 2003.

[17] S. D. Walter, "Properties of the Summary Receiver Operating Characteristic (SROC) curve for diagnostic test data," *Statistics in Medicine*, vol. 21, no. 9, pp. 1237–1256, 2002.

[18] N. M. Fida, J. A. Al-Mughales, and M. F. Fadellah, "Serum concentrations of interleukin-1 alpha, interleukin-6 and tumor necrosis factor-alpha in neonatal sepsis and meningitis," *Saudi Medical Journal*, vol. 27, no. 10, pp. 1508–1514, 2006.

[19] K. H. Park, B. H. Yoon, S.-S. Shim, J. K. Jun, and H. C. Syn, "Amniotic fluid tumor necrosis factor-alpha is a marker for the prediction of early-onset neonatal sepsis in preterm labor," *Gynecologic and Obstetric Investigation*, vol. 58, no. 2, pp. 84–90, 2004.

[20] E. Layseca-Espinosa, L. F. Pérez-González, A. Torres-Montes et al., "Expression of CD64 as a potential marker of neonatal sepsis," *Pediatric Allergy and Immunology*, vol. 13, no. 5, pp. 319–327, 2002.

[21] H. Martin, B. Olander, and M. Norman, "Reactive hyperemia and interleukin 6, interleukin 8, and tumor necrosis factor-alpha in the diagnosis of early-onset neonatal sepsis," *Pediatrics*, vol. 108, no. 4, p. E61, 2001.

[22] R. Berner, B. Tüxen, A. Clad, J. Forster, and M. Brandis, "Elevated gene expression of interleukin-8 in cord blood is a sensitive marker for neonatal infection," *European Journal of Pediatrics*, vol. 159, no. 3, pp. 205–210, 2000.

[23] R. C. Silveira and R. S. Procianoy, "Evaluation of interleukin-6, tumour necrosis factor-α and interleukin- 1β for early diagnosis of neonatal sepsis," *Acta Paediatrica*, vol. 88, no. 6, pp. 647–650, 1999.

[24] U. Hallwirth, G. Pomberger, D. Zaknun et al., "Monocyte phagocytosis as a reliable parameter for predicting early-onset sepsis in very low birthweight infants," *Early Human Development*, vol. 67, no. 1-2, pp. 1–9, 2002.

[25] R. S. Procianoy and R. C. Silveira, "Association between high cytokine levels with white matter injury in preterm infants with sepsis," *Pediatric Critical Care Medicine*, vol. 13, no. 2, pp. 183–187, 2012.

[26] Y. Shi, C. Shen, J. Wang, H. Li, S. Qin, and R. Liu, "Role of tumor necrosis factor in neonatal sepsis," *Chinese Medical Sciences Journal*, vol. 9, no. 1, pp. 45–48, 1994.

[27] P. C. Ng, K. Li, K. M. Chui et al., "IP-10 is an early diagnostic marker for identification of late-onset bacterial infection in preterm infants," *Pediatric Research*, vol. 61, no. 1, pp. 93–98, 2007.

[28] C. Sherwin, R. Broadbent, S. Young et al., "Utility of interleukin-12 and interleukin-10 in comparison with other cytokines and acute-phase reactants in the diagnosis of neonatal sepsis," *American Journal of Perinatology*, vol. 25, no. 10, pp. 629–636, 2008.

[29] P. C. Ng, S. H. Cheng, K. M. Chui et al., "Diagnosis of late onset neonatal sepsis with cytokines, adhesion molecule, and C-reactive protein in preterm very low birthweight infants," *Archives of Disease in Childhood*, vol. 77, no. 3, pp. F221–F227, 1997.

[30] E. Hotoura, V. Giapros, A. Kostoula, P. Spirou, and S. Andronikou, "Tracking changes of lymphocyte subsets and pre-inflammatory mediators in full-term neonates with suspected or documented infection," *Scandinavian Journal of Immunology*, vol. 73, no. 3, pp. 250–255, 2011.

[31] B. S. Naher, M. A. Mannan, K. Noor, and M. Shahiddullah, "Role of serum procalcitonin and C-reactive protein in the diagnosis of neonatal sepsis," *Bangladesh Medical Research Council Bulletin*, vol. 37, no. 2, pp. 40–46, 2011.

[32] T. B. Knudsen, T. B. Kristiansen, L. Simon et al., "Issues pertaining to data extraction and classification and publication bias in meta-analysis of the diagnostic accuracy of markers for bacterial infection," *Clinical Infectious Diseases*, vol. 40, no. 9, pp. 1372–1374, 2005.

[33] J. S. Gerdes, "Diagnosis and management of bacterial infections in the neonate," *Pediatric Clinics of North America*, vol. 51, no. 4, pp. 939–959, 2004.

[34] C. Chiesa, A. Panero, J. F. Osborn, A. F. Simonetti, and L. Pacifico, "Diagnosis of neonatal sepsis: a clinical and laboratory challenge," *Clinical Chemistry*, vol. 50, no. 2, pp. 279–287, 2004.

[35] T. Chan and F. Gu, "Early diagnosis of sepsis using serum biomarkers," *Expert Review of Molecular Diagnostics*, vol. 11, no. 5, pp. 487–496, 2011.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 917292, 7 pages
http://dx.doi.org/10.1155/2014/917292

Research Article

Genome-Wide Characterisation of Gene Expression in Rice Leaf Blades at 25°C and 30°C

Zhi-guo E,[1] Lei Wang,[1] Ryan Qin,[2] Haihong Shen,[3] and Jianhua Zhou[4]

[1] *China National Rice Research Institute, No. 359, Tiyuchang Road, Hangzhou 310006, China*
[2] *iBioinfo Group, Lexington, MA 02421, USA*
[3] *School of Life Science, Gwangju Institute of Science and Technology, Gwangju 500-712, Republic of Korea*
[4] *Nantong University, Nantong 226001, China*

Correspondence should be addressed to Jianhua Zhou; jianhua55@msn.com

Received 5 August 2013; Accepted 30 September 2013; Published 29 January 2014

Academic Editors: B. Shen, J. Wang, and J. Wang

Copyright © 2014 Zhi-guo E et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Rice growth is greatly affected by temperature. To examine how temperature influences gene expression in rice on a genome-wide basis, we utilised recently compiled next-generation sequencing datasets and characterised a number of RNA-sequence transcriptome samples in rice seedling leaf blades at 25°C and 30°C. Our analysis indicated that 50.4% of all genes in the rice genome (28,296/56,143) were expressed in rice samples grown at 25°C, whereas slightly fewer genes (50.2%; 28,189/56,143) were expressed in rice leaf blades grown at 30°C. Among the genes that were expressed, approximately 3% were highly expressed, whereas approximately 65% had low levels of expression. Further examination demonstrated that 821 genes had a twofold or higher increase in expression and that 553 genes had a twofold or greater decrease in expression at 25°C. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses suggested that the ribosome pathway and multiple metabolic pathways were upregulated at 25°C. Based on these results, we deduced that gene expression at both transcriptional and translational levels was stimulated at 25°C, perhaps in response to a suboptimal temperature condition. Finally, we observed that temperature markedly regulates several super-families of transcription factors, including bZIP, MYB, and WRKY.

1. Introduction

Rice (*Oryza sativa* L.) includes two major subspecies, *indica* and *japonica*. Owing to its importance in food security, extensive studies utilising genetic manipulation, improved cultivation and crossing of subspecies have been conducted in rice during recent decades to improve quality and yield. Like all other plants, rice constantly experiences environmental changes and hostile abiotic stress conditions, such as drought, cold, pollution due to heavy metals and salinity, in addition to biotic stresses, such as viral infection. To minimise abiotic and/or biotic stress-induced damage, plants have developed adaptations and stress tolerance during evolution through the regulation of gene expression and changes in cellular processes. Two major pathways, abscisic acid (ABA)-dependent and ABA-independent pathways, have been extensively studied in plants in response to biotic and/or abiotic stresses [1–3]. ABA is a hormone produced during metabolic reactions. In response to abiotic stress, the ABA-dependent pathway induces the expression of many stress-related genes by regulating the activities of transcription factors. Among many transcription factors, several super-families, including basic leucine zipper (bZIP) [4–6], MYB [7, 8], and WRKY [9–11], have been shown to play critical roles in the regulation of stress response genes in rice.

bZIP proteins include a family of transcriptional regulators that are exclusively present in eukaryotes. Furthermore, they characteristically harbour a bZIP domain composed of two structural features: a DNA-binding basic region and the leucine-zipper dimerisation region. They have been shown to regulate diverse plant-specific phenomena, including seed maturation and germination, flower induction and development, photomorphogenesis and stress and hormone signalling. There are approximately 90 bZIP transcription factor-encoding genes in the rice genome [5].

The *MYB* gene family includes at least 155 members that have been identified by a genome-wide analysis and represents one of the richest groups of transcription factors in rice. MYB proteins are characterised by a highly conserved MYB DNA-binding domain and can be classified into four major groups, 1R-MYB, 2R-MYB, 3R-MYB, and 4R-MYB, on the basis of the number and position of MYB repeats. MYB transcription factors are involved in plant development, secondary metabolism, hormone signal transduction, disease resistance, and abiotic stress tolerance [12].

WRKY genes encode transcription factors with a WRKY domain that belongs to zinc-finger proteins. WRKY proteins contain one or two conserved WRKY domains, which are encoded by approximately 60 N-terminal amino acid residues with a WRKYGQ(K/E)K sequence, followed by a C2H2 or C2HC zinc-finger motif. An exhaustive search for *WRKY* genes using HMMER and a hidden Markov model resulted in the identification of 98 and 102 *WRKY* genes in *O. japonica* and *O. indica* rice, respectively. WRKY genes play important roles in disease resistance, responses to salicylic and jasmonic acid, seed development and germination, senescence, abiotic stress responses and ABA responses in rice [13].

Despite all this knowledge, the mechanisms that regulate gene expression in rice are not completely understood. To investigate how external factors, such as temperature, affect rice development and growth through the regulation of gene expression, we searched the available transcriptome databases. We identified two transcriptome RNA-sequence (RNA-Seq) datasets of high quality from rice seedling leaf blades leaf blades grown at 25°C or 30°C. We found that the expression of more than 1300 genes in rice showed a twofold or higher difference between leaf blades that were grown at 25°C compared with those grown at 30°C. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses showed that transcription of many abiotic stress genes and genes involved in ribosome biogenesis were induced at 25°C, indicating that rice grown at 25°C has more active transcription and translation than rice grown at 30°C. Furthermore, we found that among the transcription factor super-families, bZIP, MYB, NAC, and WRKY were significantly regulated in rice at 25°C. Our studies provide useful information on the rice transcriptome in response to suboptimal temperatures.

2. Materials and Methods

2.1. Transcriptome Sequencing Datasets of Rice Seedling Leaf Blades. Two publicly available RNA-Seq datasets using deep-sequencing of rice seedling leaf blades were downloaded from Gene Expression Omnibus (GEO) under the accession number GSE42096 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42096) and used for primary analyses. The leaf blades analysed were obtained from wild-type seedlings grown at 30°C or 25°C. For each dataset, RNA-Seq was conducted by paired-end approaches using an Illumina HiSeq 2000 instrument. The read length was 90 bp.

2.2. Sequencing Analysis. Sequence alignment between the transcriptome reads was conducted and reads were checked for quality and mapped to the reference genome sequences by Bowtie 2 using the parameters "end-to-end" and "very-sensitive". The reference genome, transcript annotation, and GO datasets were downloaded from MSU Rice Genome Annotation Project, release 7. The number of reads for a gene was designated as reads per kb per million total reads (RPKM) after normalisation to the number of mapped genome locations. KEGG gene classifications were downloaded from its database.

2.3. Statistical Analyses. To determine whether expression was differentially regulated under different temperatures (25°C versus 30°C), we conducted statistical analyses based on the fold-changes in gene expression by adding median counts as a pseudocount. Pathway analyses were based on the binomial probability of observing a number of gene changes in a given pathway. Differences were considered statistically significant when the P value was <0.05.

3. Results

3.1. RNA-Seq Datasets of the Transcriptome from Rice. To accurately determine rice gene expression profiles, we took advantage of recent advances in deep-sequencing technologies. Many RNA-Seq datasets of the transcriptome of rice and other plants are publicly available. Using these datasets, we identified transcriptome sequencing libraries of two rice samples in a single GEO dataset, GSE42096, generated by the Chinese Academy of Sciences and the National Centre for Plant Gene Research (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42096). Profiling of these transcriptome RNAs by high-throughput sequencing was conducted by both single-end and paired-end approaches using an Illumina HiSeq 2000 platform. Paired-end sequencing provided 90 basepairs (bp) per read from each end. RNA-Seq of the transcriptome for each sample generated approximately 25 million reads. Sequence analysis indicated that the datasets were of exceptionally high quality with very low background noise (Supplementary Figure 1 available online on http://dx.doi.org/10.1155/2014/917292).

3.2. Gene Expression in Rice at 25°C and 30°C. To characterize expressed transcripts, a reference genome dataset was required. We searched existing databases and found that the MSU Rice Genome Annotation Project on *O. sativa japonica*, released on 31st Oct 2011, is the most complete rice genome database available with more than 56,143 annotated genes, slightly more than the 55986 reported by Kawahara et al. [14]. Therefore, we used this database as our reference for analysing gene expression in the two leaf blade transcriptome datasets. We used Bowtie 2 to align and map the transcriptome reads to rice genome. On the basis of our analysis, we calculated the number of reads as reads per billion (RPB) for each mRNA in two samples. We normalised RPB by read per kbp and RPKM. We found that at 25°C there were 19,766 annotated genes with >1 RPKM read, whereas at 30°C

Table 1: Top ranked pathways that are regulated (1374 genes) at 25°C by GO analysis.

Rank	Pathway	Pathway annotation	Pathway size	Observed	Ratio	P value
1	Process: metabolic process	rice:GO:0008152	7390	391	0.05	$4.3E-52$
2	Process: response to stress	rice:GO:0006950	3620	241	0.07	$1.9E-46$
3	Process: response to abiotic stimulus	rice:GO:0009628	2195	177	0.08	$1.5E-44$
4	Process: cellular process	rice:GO:0009987	7325	365	0.05	$3.3E-42$
5	Function: catalytic activity	rice:GO:0003824	3688	227	0.06	$2.7E-38$
6	Component: membrane	rice:GO:0016020	3728	220	0.06	$2.1E-34$
7	Process: biosynthetic process	rice:GO:0009058	4673	250	0.05	$1.6E-32$
8	Process: response to endogenous stimulus	rice:GO:0009719	1490	118	0.08	$4.9E-29$
9	Component: cytosol	rice:GO:0005829	2289	151	0.07	$2.1E-28$
10	Process: response to biotic stimulus	rice:GO:0009607	1081	96	0.09	$1.9E-27$
11	Component: ribosome	rice:GO:0005840	481	61	0.13	$9.5E-26$
12	Function: structural molecule activity	rice:GO:0005198	518	63	0.12	$1.5E-25$

the number of genes that had at least 1 RPKM was 19,350. The distribution of expressed genes with different RPKM levels was similar between 25°C and 30°C (Supplementary Figure 2). Approximately 30 genes had extremely high expression levels (>1000 RPKM), 3% of annotated genes (550–600) were highly expressed (>100 RPKM) and more than 33% of expressed genes (6900) were expressed at only a modest level (10–100 RPKM). Of the genes that were expressed, 64% (12,000) had fewer than 9 RPKM, suggesting that most genes were expressed at a low level. Expression of the other 36,000 genes was not detectable.

3.3. Differential Gene Expression in Rice between 30°C and 25°C. We then compared the genes with varied expression in rice growing at 30°C and 25°C. We calculated ratios between the number of reads at 30°C and the number of reads at 25°C. As shown in Supplementary Figure 3, the left part of the histogram shows the number of genes with an increased expression at 25°C. We found that the expression of 257 genes was upregulated more than threefold in rice grown at 25°C, whereas expression of 173 genes was downregulated more than threefold at 25°C. Moreover, there were more genes that were upregulated (564) than downregulated (380) at 25°C, with expression level changes between twofold and threefold. Among approximately 2712 genes with a 1.5–2.0-fold change in expression, 1617 genes were upregulated and 1095 genes (1.5–2.0-fold) were downregulated. Our results indicate that more genes were upregulated at 25°C than at 30°C, suggesting that at 25°C, rice plants need to respond to a suboptimal lower temperature by altering gene transcription.

3.4. GO Analysis of Genes That Are Upregulated and Downregulated. To examine the mechanisms of the molecular and cellular responses to a suboptimal lower temperature of 25°C, we performed GO and KEGG analysis in genes of the rice transcriptome that had an twofold or higher change in expression between 25°C and 30°C. In total, we found that the expression of 821 genes was upregulated by twofold or higher at 25°C, whereas 553 genes showed downregulated expression at 25°C. Among the 1374 genes that were either upregulated or downregulated, GO analysis indicated that 4 of the 10 top-ranked GO categories were stress related, with 17.54% of genes (241/1374) related to "response to stress" (ranked at no. 2 with $P = 1.88E-46$), 12.88% of genes (177/1374) related to "response to abiotic stimulus" (ranked at no. 3 with $P = 1.55E-44$), 8.59% of genes (118/1374) related to "response to endogenous stimulus" (ranked at no. 8 with $P = 4.90E-29$) and 6.99% of genes (96/1374) related to "response to biotic stimulus" (ranked at no. 10 with $P = 1.91E-27$), suggesting that 25°C could be considered as a cold-stress condition (Table 1). Other high-ranked GO categories included membrane processes (ranked no. 1) and metabolic processes and ribosome (ranked no. 11), indicating that metabolism and protein translation are perhaps also upregulated at 25°C.

3.5. KEGG Pathway Analysis of Upregulated and Downregulated Genes. To further characterise the pathways that are involved in temperature-induced stress responses, we performed KEGG pathway analysis. We observed that the expression of genes that are involved in ribosome biogenesis was significantly upregulated at 25°C, with an adjusted P value of $5.2E-29$ (Table 2). In addition, 57 of the potential 362 transcripts that had been annotated in the ribosome pathway had a twofold or higher increase in expression at 25°C compared with 30°C. These upregulated transcripts represent 15.7% of the genes in the ribosome pathway. In contrast, none of the 362 transcripts showed decreased expression. Taken together, these results strongly suggested that both transcription and translation were more active at 25°C than 30°C. Other major pathways showing significant changes in expression on KEGG analysis included metabolic and biosynthesis pathways. Of note, P values for pathways that were upregulated were markedly more significant than pathways that were downregulated.

3.6. Expression Analysis of bZIP, WRKY and MYB Transcription Factors. The relative growth rate (RGR) in rice is influenced by temperature, with an optimal growth rate at 30°C [15]. Rice also has a stress response mechanism that

Table 2: Top ranked pathways that are regulated (1374 genes) at 25°C by KEGG analysis.

Rank	Pathway	Pathway annotation	Pathway size	Observed	Ratio	*P* value
1	Ribosome	rice:osa03010	362	57	0.16	$5.2E-29$
2	Metabolic pathways	rice:osa01100	1565	119	0.08	$1.0E-27$
3	Biosynthesis of secondary metabolites	rice:osa01110	745	65	0.09	$1.3E-18$
4	Starch and sucrose metabolism	rice:osa00500	130	20	0.15	$5.9E-11$
5	Alpha-Linolenic acid metabolism	rice:osa00592	34	9	0.26	$8.7E-08$
6	Glyoxylate and dicarboxylate metabolism	rice:osa00630	62	11	0.18	$2.6E-7$
7	Carbon fixation in photosynthetic organisms	rice:osa00710	85	12	0.14	$9.6E-07$
8	Diterpenoid biosynthesis	rice:osa00904	24	7	0.29	$1.2E-06$
9	Photosynthesis	rice:osa00195	148	15	0.10	$3.2E-06$
10	Plant hormone signal transduction	rice:osa04075	150	14	0.09	$1.6E-05$
11	Biosynthesis of unsaturated fatty acids	rice:osa01040	44	7	0.16	$8.0E-05$
12	Phenylpropanoid biosynthesis	rice:osa00940	94	9	0.10	$4.0E-04$

Table 3: Top ranked transcription factors that are regulated in rice by temperature.

Rank	Pathway	Pathway annotation	Pathway size	Observed	Ratio	*P* value
1	Rice transcription factor: WRKY	rice:TF:WRKY	107	15	0.14	$4.9E-08$
2	Rice transcription factor: NAC	rice:TF:NAC	124	12	0.096	$4.5E-05$
3	Rice transcription factor: AP2-EREBP	rice:TF:AP2-EREBP	169	14	0.082	$5.9E-05$
4	Rice transcription factor: orphans	rice:TF:orphans	85	9	0.105	0.00019
5	Rice transcription factor: bZIP	rice:TF:bZIP	95	9	0.095	0.00043
6	Rice transcription factor: MYB	rice:TF:MYB	128	10	0.078	0.00092
7	Rice transcription factor: tify	rice:TF:tify	21	4	0.190	0.00140
8	Rice transcription factor: MYB-related	rice:TF:MYB-related	100	7	0.07	0.00835
9	Rice transcription factor: C2H2	rice:TF:C2H2	104	7	0.067	0.01004
10	Rice transcription factor: pseudo ARR-B	rice:TF:pseudo_ARR-B	9	2	0.222	0.01811
11	Rice transcription factor: G2-like	rice:TF:G2-like	48	4	0.083	0.02342
12	Rice transcription factor: WRKY	rice:TF:WRKY	107	15	0.140	4.95634

is triggered in response to lower temperatures. As described above, transcription is more active at 25°C than at 30°C, indicating that gene expression is stimulated at 25°C. To characterise transcription factors that may be involved in the regulation of gene expression in rice growth at different temperatures and to understand how rice responds to the suboptimal temperature of 25°C, we analysed transcription factor families in rice (Supplementary Table S1), including the expression distribution patterns of the bZIP, MYB, WRKY, and HLH transcription factor super-families (Table 3, Supplementary Table S1 and Figure 1). Approximately 9.5% of bZIP (9/95), 14.1% of WRKY (15/107), 7.8% of MYB (12/128) and 3.5% of HLH (5/145) transcription factors had a twofold or higher change in expression. In contrast, a random calculation suggested that <2.57% (1374/56,143) of the genes should be upregulated or downregulated by twofold or more. Therefore, we conclude that expression of bZIP, WRKY, and MYB super-families was significantly regulated by temperature, with *P* values of $4.96E{-}08$ (WRKY), $4.3E{-}04$ (bZIP) and 0.008 (MYB) between 25°C and 30°C. These data are consistent with previous reports [16, 17] showing that these transcription factor super-families are upregulated or downregulated under colder or warmer temperatures. Other highly regulated transcription factor super-families include NAC and AP2-EREBP. In contrast, other transcription factors such as bHLH and HB did not exhibit significant changes in expression.

4. Discussion

Like all plants, rice has to endure constant environmental changes. Among many factors, temperature has been shown to greatly influence rice growth. Rice can grow at a range of temperature, from as low as 12°C to as high as 40°C, but its optimal growth temperature is 30°C or warmer [15]. Unlike mammals, which have a constant body temperature, rice grows at temperatures that fluctuate daily between night and day. One of the mechanisms by which rice can adjust to temperature changes is through regulation of gene expression. Extensive studies have been conducted in rice to analyse the molecular basis of adaptation to both warmer temperatures and cold-stress conditions [3, 18–24]. However, global surveys of temperature-dependent changes in rice gene expression, particularly studies using next generation sequencing technology, are not extensive.

Because the rapid development and reduced cost of both next-generation sequencing and microarray technology, researchers have regularly deposited RNA-Seq datasets for

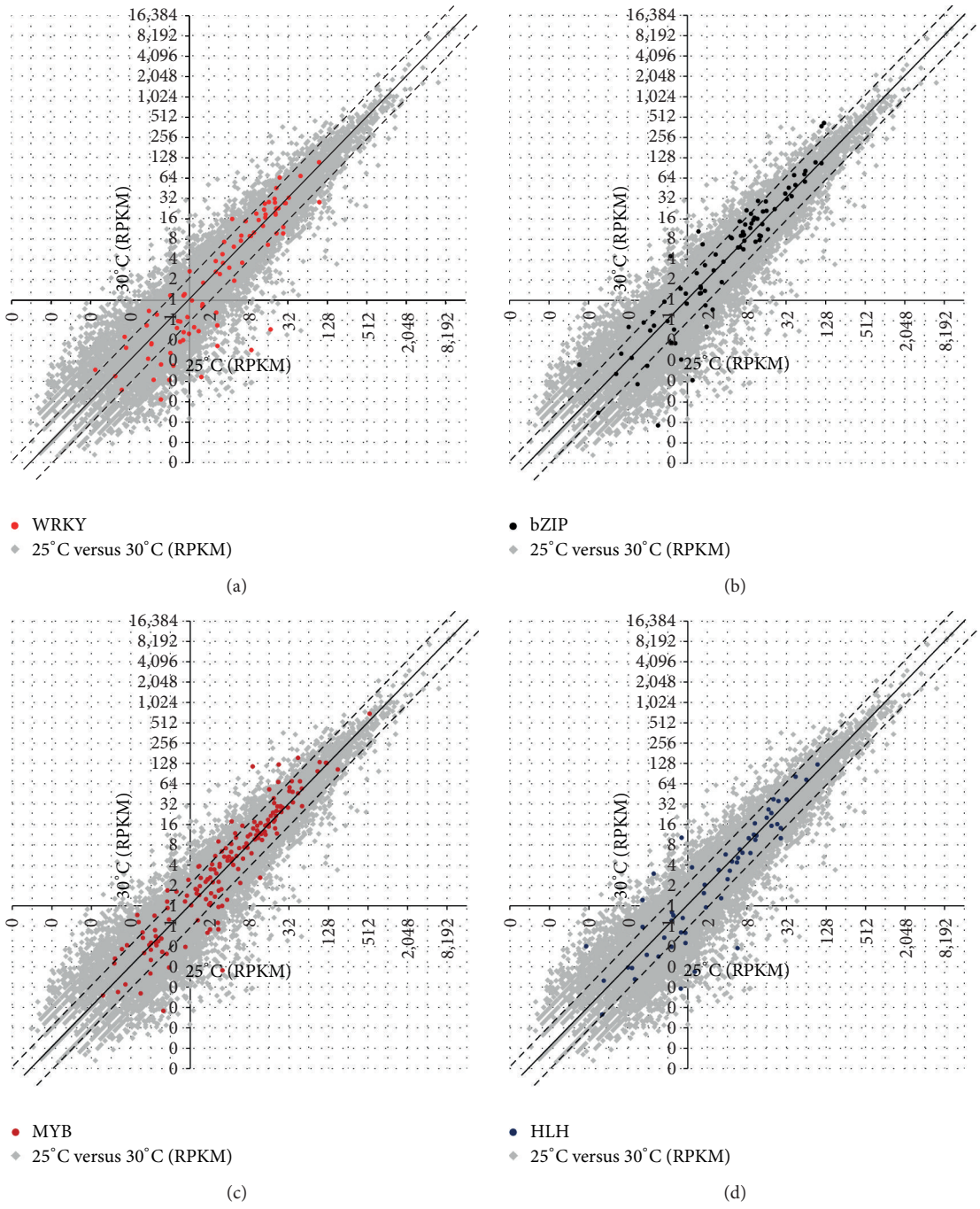


Figure 1: A scatter plot of gene expression values (per kb per million total reads, RPKM) in log2 scale for the 25°C versus 30°C datasets. Grey dots indicate each gene expressed in the two samples. Coloured dots indicate a specific group of genes. Two diagonal dashed lines indicate twofold changes (increases and decreases).

many organisms, including rice, in public domains. For example, several recent papers reported gene expression and splicing in rice using NGS and bioinformatics analysis [25–28]. To investigate how temperature may influence rice growth by affecting gene expression, we searched GEO databases (http://www.ncbi.nlm.nih.gov/geo/). We identified four sets of RNA-Seq transcriptome data from rice (GSE42096, GSE39307, GSE30490, and GSE27240) that were of a high quality. However, among these datasets, only four samples in the GSE42096 dataset, which includes wild-type and TOG1 mutant rice leaf blades, were related to different temperatures (25°C and 30°C). Because mutant TOG1 has not yet been fully described in a public domain, we analysed only two wild-type rice leaf blade samples that grew at both 25°C and 30°C. We realised that the number of samples (two) was limited, but because the quality of these RNA-Seq datasets was exceptionally high, with more than 25 M reads each sample, we believe that our analysis will provide useful

information on gene expression in rice to complement similar studies.

Gene expression is regulated at multiple levels. The most fundamental regulatory mechanisms that control the amount of proteins produced are transcription and translation. The *cis*-sequences in a gene, in particular the promoter, and transcription factors dictate how much RNA is transcribed from a gene, whereas ribosomes are directly related to the activities of protein translation. In our investigation, we demonstrated that the number of the genes that are expressed at 30°C (17,356) is similar to that at 25°C (17,966), indicating that it is necessary for only 1/3 of rice genes to be expressed to maintain growth under a given condition. Similarly, we observed that about 3% of genes were highly expressed at both 30°C and 25°C. The rest of the expressed genes had either a modest or a low expression level or no detectable expression.

However, the difference lies in the expression levels of specific genes between samples at 25°C and 30°C. We showed that 3986 genes had either increased or decreased expression levels of 1.50-fold or higher. Among these 3986 genes, 1374 had a twofold or higher increase (821) or decrease (553) when the temperature dropped from 30°C to 25°C, indicating perhaps that rice at 25°C has more active transcription. Consistent with this notion, by GO and KEGG analyses, we found that a significant number of genes in the ribosome pathways were upregulated at 25°C, suggesting that translation may be also more robust at 25°C. Considering previous reports that rice has a better growth rate at 30°C, or at least at temperatures warmer than 25°C [15], the more active transcription and translation at 25°C can be explained by the response of rice to a colder temperature. In fact, we observed in our GO analysis that 4 of the 12 top-ranked pathway categories were related to stress response. We deduced from our results that although 30°C is an optimal temperature for rice growth, transcription and translation for many genes are triggered at 25°C and this temperature (25°C) may be minimally sufficient to trigger the cold stress-response.

To further examine the molecular basis of the differential expression of rice genes between 25°C and 30°C, we examined rice transcription factor families (Figure 1, Table 3, Supplementary Table S1). In contrast to the WRKY super-family, bZIP and MYB transcription factor super-families had more downregulated than upregulated genes (1.5-fold cutoff) at 25°C. Although we were unable to draw definitive conclusions from this analysis, our results suggest that the WRKY super-family plays a positive role in the response to lower temperatures, whereas both bZIP and MYB super-families may have a negative impact on gene expression at 25°C, in agreement with a previous report that bZIP transcription factors, such as ZIP52, are negative regulators of cold stress [5]. In addition, we also showed that many transcription factors, including bHLH and HB, are not significantly regulated between 30°C and 25°C.

5. Conclusion

We concluded that only a small percentage of genes (3%) have a high expression level whereas more than 60% of genes have a very low expression level in rice. Both transcription and translation are more active at 25°C than at 30°C. Expression of bZIP, WRKY, and MYB is significantly regulated at 25°C.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this article.

Acknowledgments

This research was supported by the special fund for Agro-Scientific Research in the Public Interest (201203029, 201203031), the Special Foundation for Basic Research and the Development of Central Level Scientific Research Institutes in China (2012RG006) (ZGE, LW), and the Priority Academic Program Development of Jiangsu Higher Education Institution (PAPD) (JZ).

References

[1] Y. Fujita, M. Fujita, K. Shinozaki, and K. Yamaguchi-Shinozaki, "ABA-mediated transcriptional regulation in response to osmotic stress in plants," *Journal of Plant Research*, vol. 124, no. 4, pp. 509–525, 2011.

[2] S. C. Lee and S. Luan, "ABA signal transduction at the crossroad of biotic and abiotic stress responses," *Plant, Cell and Environment*, vol. 35, no. 1, pp. 53–60, 2012.

[3] G. T. Huang, S. L. Ma, L. P. Bai et al., "Signal transduction during cold, salt, and drought stresses in plants," *Molecular Biology Reports*, vol. 39, no. 2, pp. 969–987, 2012.

[4] H. Takahashi, T. Kawakatsu, Y. Wakasa, S. Hayashi, and F. Takaiwa, "A rice transmembrane bZIP transcription factor, OsbZIP39, regulates the endoplasmic reticulum stress response," *Plant and Cell Physiology*, vol. 53, no. 1, pp. 144–153, 2012.

[5] C. Liu, Y. Wu, and X. Wang, "bZIP transcription factor OsbZIP52/RISBZ5: a potential negative regulator of cold and drought stress response in rice," *Planta*, vol. 235, no. 6, pp. 1157–1169, 2012.

[6] M. A. Hossain, Y. Lee, J. I. Cho et al., "The bZIP transcription factor OsABF1 is an ABA responsive element binding factor that enhances abiotic stress signaling in rice," *Plant Molecular Biology*, vol. 72, no. 4, pp. 557–566, 2010.

[7] X. Dai, Y. Xu, Q. Ma et al., "Overexpression of an R1R2R3 MYB gene, OsMYB3R-2, increases tolerance to freezing, drought, and salt stress in transgenic Arabidopsis," *Plant Physiology*, vol. 143, no. 4, pp. 1739–1751, 2007.

[8] L. Zhang, G. Zhao, J. Jia, X. Liu, and X. Kong, "Molecular characterization of 60 isolated wheat MYB genes and analysis of their expression during abiotic stress," *Journal of Experimental Botany*, vol. 63, no. 1, pp. 203–214, 2012.

[9] P. Agarwal, M. P. Reddy, and J. Chikara, "WRKY: its structure, evolutionary relationship, DNA-binding selectivity, role in stress tolerance and development of plants," *Molecular Biology Reports*, vol. 38, no. 6, pp. 3883–3896, 2011.

[10] H. Chen, Z. Lai, J. Shi, Y. Xiao, Z. Chen, and X. Xu, "Roles of arabidopsis WRKY18, WRKY40 and WRKY60 transcription factors in plant responses to abscisic acid and abiotic stress," *BMC Plant Biology*, vol. 10, article 281, 2010.

[11] R. Sharma, D. De Vleesschauwer, M. K. Sharma, and P. C. Ronald, "Recent advances in dissecting stress-regulatory crosstalk in rice," *Molecular Plant*, vol. 6, pp. 250–260, 2013.

[12] A. Nijhawan, M. Jain, A. K. Tyagi, and J. P. Khurana, "Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice," *Plant Physiology*, vol. 146, no. 2, pp. 333–350, 2008.

[13] C. A. Ross, Y. Liu, and Q. J. Shen, "The WRKY gene family in rice (*Oryza sativa*)," *Journal of Integrative Plant Biology*, vol. 49, no. 6, pp. 827–842, 2007.

[14] Y. Kawahara, M. de la Bastide, J. P. Hamilton et al., "Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, article 4, 2013.

[15] T. Nagai and A. Makino, "Differences between rice and wheat in temperature responses of photosynthesis and plant growth," *Plant and Cell Physiology*, vol. 50, no. 4, pp. 744–755, 2009.

[16] Y. Xiang, N. Tang, H. Du, H. Ye, and L. Xiong, "Characterization of OsbZIP23 as a key player of the basic leucine zipper transcription factor family for conferring abscisic acid sensitivity and salinity and drought tolerance in rice," *Plant Physiology*, vol. 148, no. 4, pp. 1938–1952, 2008.

[17] N. Tang, H. Zhang, X. Li, J. Xiao, and L. Xiong, "Constitutive activation of transcription factor OsbZIP46 improves drought tolerance in rice," *Plant Physiology*, vol. 158, no. 4, pp. 1755–1768, 2012.

[18] Y. Cao, Q. Zhang, Y. Chen et al., "Identification of differential expression genes in leaves of rice (*Oryza sativa* L.) in response to heat stress by cDNA-AFLP analysis," *BioMed Research International*, vol. 2013, Article ID 576189, 11 pages, 2013.

[19] H. Chauhan, N. Khurana, P. Agarwal, and P. Khurana, "Heat shock factors in rice (*Oryza sativa* L.): genome-wide expression analysis during reproductive development and abiotic stress," *Molecular Genetics and Genomics*, vol. 286, no. 2, pp. 171–187, 2011.

[20] J. Zou, A. Liu, X. Chen et al., "Expression analysis of nine rice heat shock protein genes under abiotic stresses and ABA treatment," *Journal of Plant Physiology*, vol. 166, no. 8, pp. 851–861, 2009.

[21] D. Mittal, D. A. Madhyastha, and A. Grover, "Gene expression analysis in response to low and high temperature and oxidative stresses in rice: combination of stresses evokes different transcriptional changes as against stresses applied individually," *Plant Science*, vol. 197, pp. 102–113, 2012.

[22] D. Mittal, D. A. Madhyastha, and A. Grover, "Genome-wide transcriptional profiles during temperature and oxidative stress reveal coordinated expression patterns and overlapping regulons in rice," *PLoS ONE*, vol. 7, no. 7, Article ID e40899, 2012.

[23] R. B. Saad, D. Fabre, D. Mieulet et al., "Expression of the *Aeluropus littoralis AlSAP* gene in rice confers broad tolerance to abiotic stresses through maintenance of photosynthesis," *Plant, Cell and Environment*, vol. 35, no. 3, pp. 626–643, 2012.

[24] M. R. Park, K. Y. Yun, B. Mohanty et al., "Supra-optimal expression of the cold-regulated OsMyb4 transcription factor in transgenic rice changes the complexity of transcriptional network with major effects on stress tolerance and panicle development," *Plant, Cell and Environment*, vol. 33, no. 12, pp. 2209–2230, 2010.

[25] G. Zhang, G. Guo, X. Hu et al., "Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome," *Genome Research*, vol. 20, no. 5, pp. 646–654, 2010.

[26] T. Lu, G. Lu, D. Fan et al., "Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq," *Genome Research*, vol. 20, no. 9, pp. 1238–1249, 2010.

[27] L. Gu and R. Guo, "Genome-wide detection and analysis of alternative splicing for nucleotide binding site-leucine-rich repeats sequences in rice," *Journal of Genetics and Genomics*, vol. 34, no. 3, pp. 247–257, 2007.

[28] B. B. Wang and V. Brendel, "Genomewide comparative analysis of alternative splicing in plants," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 18, pp. 7175–7180, 2006.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 971258, 7 pages
http://dx.doi.org/10.1155/2014/971258

Research Article

Prediction and Analysis of Surface Hydrophobic Residues in Tertiary Structure of Proteins

Shambhu Malleshappa Gowder,[1] Jhinuk Chatterjee,[2] Tanusree Chaudhuri,[1] and Kusum Paul[1]

[1] *Department of Biotechnology, The Oxford College of Engineering, Bangalore 560068, India*
[2] *Department of Biotechnology, PES Institute of Technology, Bangalore 560085, India*

Correspondence should be addressed to Shambhu Malleshappa Gowder; shambumg13@gmail.com

Received 13 August 2013; Accepted 17 October 2013; Published 9 January 2014

Academic Editors: J. Wang and J. Wang

Copyright © 2014 Shambhu Malleshappa Gowder et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The analysis of protein structures provides plenty of information about the factors governing the folding and stability of proteins, the preferred amino acids in the protein environment, the location of the residues in the interior/surface of a protein and so forth. In general, hydrophobic residues such as Val, Leu, Ile, Phe, and Met tend to be buried in the interior and polar side chains exposed to solvent. The present work depends on sequence as well as structural information of the protein and aims to understand nature of hydrophobic residues on the protein surfaces. It is based on the nonredundant data set of 218 monomeric proteins. Solvent accessibility of each protein was determined using NACCESS software and then obtained the homologous sequences to understand how well solvent exposed and buried hydrophobic residues are evolutionarily conserved and assigned the confidence scores to hydrophobic residues to be buried or solvent exposed based on the information obtained from conservation score and knowledge of flanking regions of hydrophobic residues. In the absence of a three-dimensional structure, the ability to predict surface accessibility of hydrophobic residues directly from the sequence is of great help in choosing the sites of chemical modification or specific mutations and in the studies of protein stability and molecular interactions.

1. Introduction

Knowledge of protein stability is crucial for understanding of the basic thermodynamics of the process of folding. The hydrophobic effect is considered to be the major driving force for the folding of globular proteins [1]. The hydrophobic effect is driven by the entropy increase of the solvent water molecules; hydrophobic side chains are located predominantly in the interior of a protein. This arrangement stabilizes the folded polypeptide backbone, since unfolding it or extending it would expose the hydrophobic side chains to the solvent. The hydrophobicity analysis has remained at the central focus for understanding protein folding and stability. It has been hypothesized that hydrophobic interactions play a major role in organizing and stabilizing the architecture of proteins [2]. As their name implies, hydrophobic amino acids have essentially nonpolar side chains, for example, valine, leucine, isoleucine, phenylalanine, and methionine fit into this group. In proteins, hydrophobic residues tend to be buried in the interior of the protein away from the solvent and polar side chains are exposed to the solvent. The folding process of polypeptide chain depends on the hydrophobicity of the side chains. It is now widely accepted that hydrophobicity is a dominant force of protein folding [3, 4]. There is a linear relationship between the surface areas of amino acid residues (in a standard state) and the free energy changes associated with the transfer of the amino acids from water to organic solvent [5–7].

One strategy to increase the stability of proteins is to reduce the area of water-accessible hydrophobic surface [8].

Solvent accessibility plays an important role in the structure and functions of biological macromolecules. Generally amino acid residues located on the surface of a protein serve as active sites and/or interact with other molecules and ligands [9]. The concept of solvent accessibility is widely used to understand the location of amino acid residues in

protein structures and their contribution to the stability of the protein.

The folding process of soluble proteins decreases the surface in contact with the solvent. This is related to the secondary structures of proteins. Accurate knowledge of residue accessibility would thus aid the prediction of secondary structures. Different methods of prediction are based on the use of protein structure databases and on multiple sequence alignments. They have various efficiencies, notably, depending on the number of relative accessibility states that is, exposed, 2 buried, and in-between; [10–14].

The accessible surface area of the protein is calculable from a set of coordinates which measures the thermodynamic interaction between protein and water. Surface area accessibility calculations identify which residues are solvent exposed and which residues are buried, contributing to the hydrophobic stabilization of protein structure. In the case of the solvent accessibility prediction, using evolutionary information such as multiple sequence alignment and position-specific scoring matrix has generally given good prediction results [15]. From MSA (multiple sequence alignment), we analyzed how well solvent exposed and buried hydrophobic residues are evolutionarily conserved on the nonredundant data set of 218 monomeric proteins.

2. Materials and Methods

2.1. Data Set. In the present study, total of 4154 monomeric proteins were obtained from PIQSI (quaternary structure database) [16]. We have filtered out those proteins to get nonredundant monomeric proteins dataset from PDB (protein data bank) [17] which has the following features: (i) X-ray resolution less than 2 Å for better resolution, (ii) percentage of similarity cut-off less than 30%, (iii) having a biological assembly unit, and (iv) chain length not less than 50 residues, and finally nonredundant datasets of 218 proteins were obtained.

2.2. Computation of Solvent Access Surface Area. The ASA (accessible surface area) is defined as the locus of the center of the solvent molecule as it rolls over van der Waals surface of the protein [7]. The software NACCESS [18] was used to calculate ASA for all atoms in PDB file. The ASA is calculated using Lee-Richards (1971) formula [19], whereby a probe of a given radius is rolled around the surface of the molecule, and the path traced out by its center is the accessible surface:

$$\mathrm{ASA} = \sum \left[\frac{R}{\left(R^2 - Z_i^2\right)^{1/2}} \right] L_i \cdot D; \qquad D = \frac{\Delta Z}{2} + \Delta' Z, \tag{1}$$

where L_i is the length of the arc computed on a given section I, Z_i is the perpendicular distance from the center of the sphere to the section I, ΔZ is the spacing between the sections, and $\Delta' Z$ is $\Delta Z/2$ or $R - Z_i$, whichever is smaller.

2.3. Relative Solvent Accessibility. RSA (relative accessible surface area) is defined as the per residue ratio between ASA and references value for particular residue. RSA file containing summed atomic accessible surface areas over each protein or nucleic acid residue, as well as the relative accessibility of each residue calculated as the % accessibility compared to the accessibility of that residue type in an extended ALA-x-ALA tripeptide for amino acids [20].

The pictorial representation of such RSA values provides an easy understanding of the location of each residue in the structure of protein. It will also reveal the population of each residue on the surface and interior core of a protein.

Threshold to distinguish 2 states is also specified. We have classified residues based on threshold values of RSA cut-off used by Zhu and Blundell [21] If the RSA percentage is greater than 7, it will be considered as solvent exposed residue and RSA percentage is less than 7, it will be considered as buried residue.

2.4. Residue Propensity. During the process of protein folding, the amino acid residues along with the polypeptide chain interact with each other in a cooperative manner to form stable native structure and also form clusters. Zehfus reported that averages of 65% of hydrophobic residues are involved in residue clusters and each hydrophobic cluster contains at least five residues. Probably, hydrophobic residues (FMILYVW) occur frequently within buried area and flanking the gapped region [22, 23].

In order to analyze the hydrophobic cluster in proteins and to understand the influence of interresidue interactions to the formation of residue clusters, which are important for the folding and stability of protein structures, we have calculated propensity of each residue type on the surface and buried area in order to know each residue's natural tendency towards buried area and exposed area.

2.5. Propensity Calculation. The Following equations refer to propensity calculation towards surface and buried area:

$$\mathrm{SURFACEPROPENSITY} = \left(\frac{\text{Total no of solvent exposed specific type residues}}{\text{Total no of solvent exposed residues}} \right) \times \left(\frac{\text{Total no of specific type residues}}{\text{Total no of residues}} \right)^{-1}, \tag{2}$$

$$\mathrm{BURIEDPROPENSITY} = \left(\frac{\text{Total no of buried specific type residues}}{\text{Total no of buried residues}} \right) \times \left(\frac{\text{Total no of specific type residues}}{\text{Total no of residues}} \right)^{-1}. \tag{3}$$

Similarly as mentioned in (2) and (3), we have also calculated propensity of hydrophobic residues for flanking regions both for buried and exposed hydrophobic residues. (i) +1 and −1 (ii) +2 and −2 regions are considered for flanking residues.

2.6. Searching for Homologous Sequences for Each of 218 Monomer Proteins. By the nature of proteins, we know that solvent exposed hydrophobic residues are poorly conserved, but buried hydrophobic residues are highly conserved [24, 25]. In order to check the evolutionarily conserved hydrophobic residues on solvent exposed area and buried area, we used stand-alone BLASTP [26] for each individual protein against nonredundant dataset. Consider homologous sequences which have sequence identity greater than 30%.

2.7. Calculating Conservation Score Based on Hydrophobic Nature. Conservation score for all the residues in the protein can be obtained by comparing the sequence of a PDB chain with its respective homologous sequences using multiple sequence alignment. In our analysis, conservation score has been calculated based on hydrophobic nature evolutionarily in the alignment by applying following conditions:

(1) in the alignment when any of these hydrophobic residues occur (Val, Ile, Leu, Met, and Phe) are scored 1;

(2) similarly Ala and aromatic residues like Tyr and Trp that occur in the alignment are scored 0.5 because these three residues are partially hydrophobic and they tend to be buried and exposed equally;

(3) if any polar residues occur in the alignment, then they are scored as −2 because they are hydrophilic in nature;

(4) finally gap has been considered as −2 extra penalty is given for gap.

3. Results and Discussion

3.1. Interior and Surface Amino Acid Composition. To know the hydrophobic residues distribution in protein three dimensional structures we have performed structural analysis of 218 proteins using NACCESS server with respect to its RSA values, (details provided in supplementary file available online at http://dx.doi.org/10.1155/2014/971258) the following results were observed. 34.84% of hydrophobic residues occurred in total data set of proteins in which 77.1% of hydrophobic residues preferred in buried area and 22.9% of hydrophobic residues preferred in accessible surface area. Propensity of hydrophobic residues preference on protein surface and interior was calculated in order to analyze the hydrophobicity cluster (Figure 1).

The surface propensity and buried propensity for each residue described in the Figure 1 were calculated using (2) and (3), respectively. It has been observed from Figure 1 that large hydrophobic residues such as Val, Ile, Leu, Met, Phe, and including partial hydrophobic residues like Tyr, Try, and Ala have high propensity towards buried region compared to surface regions. Among hydrophobic residues, Ile has the highest propensity towards the buried region having a value of 1.96 and correspondingly Met has the highest tendency towards solvent exposed region (Figure 1).

Hydrophilic residues have high propensity towards surface region. Among all hydrophobic residues, His has a high tendency towards buried region and Lys has high tendency towards exposed region.

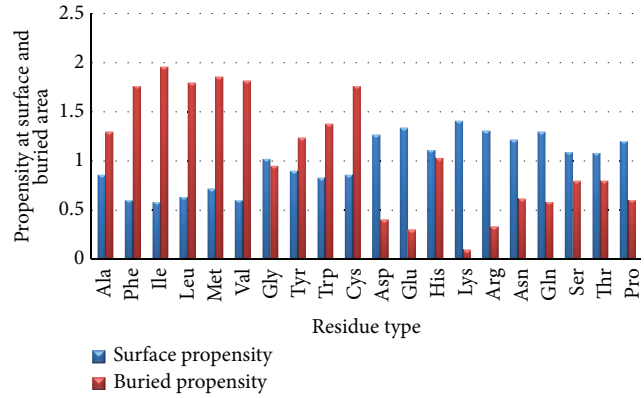


Figure 1: It refers to propensity of individual residues on surface and buried area.

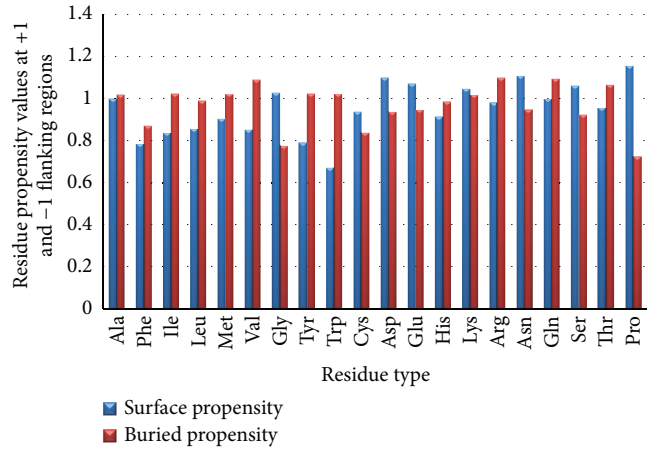


Figure 2: It refers to propensity of hydrophobic residues at +1 and −1 flanking position.

In order to analyze hydrophobic clusters appearing in surface or buried areas, flanking regions were considered. Figures 2 and 3 give a clear observation that hydrophobic cluster is more likely to come towards buried region than exposed region which is the range +1, −1 and +2, −2 present in the flanking region.

Significant changes for hydrophobic residues were not observed when flanking regions +1, −1 and +2, −2 were compared. The conservation score and knowledge of flanking regions of hydrophobic residues propensity towards buried and exposed area have been applied to the prediction of surface hydrophobic residues.

We needed to know how well surface and buried hydrophobic residues are conserved evolutionarily. Conservation score has been calculated for each residue of the query protein present in the complete data set using the knowledge of hydrophobic nature in the homologous sequences (Table 1).

Figure 4 refers to relative frequency of solvent exposed and buried hydrophobic residues in respective conservative score bin.

As observed from Figure 4, solvent exposed hydrophobic residues are dominant than buried hydrophobic residues at

conservation score range from 0 to 50. It has been observed that over 70% of exposed hydrophobic residues are falling in conservation score range from 0 to 50. Further, the observation leads to only 30% buried hydrophobic residues falling into the 0 to 50 conservation score ranges and remaining 70% falling into the 50 to 100 range. It shows that buried hydrophobic residues are highly conserved than exposed.

It has also been observed that at the conservation score range 60 to 70 there is an overlap, where in the buried hydrophobic residues start to take over exposed hydrophobic residues, they dominate in the conservation score range from 70 to 100.

3.2. Confidence Score Calculation. (a) Consider exposed hydrophobic residues

$$\text{CONFIDENCE SCORE} = \frac{\text{Solvent exposed hydrophobic residues normalized score at the range [0 to 10]}}{\text{Buried hydrophobic residues normalized score at the range [0 to 10]}}. \quad (4)$$

(b) Consider buried hydrophobic residues

$$\text{CONFIDENCE SCORE} = \frac{\text{Buried hydrophobic residues normalized score at the range [0 to 10]}}{\text{Solvent exposed hydrophobic residues normalized score at the range [0 to 10]}}. \quad (5)$$

Hydrophobic cluster analysis is based on a two-dimensional representation of the protein sequence, in which hydrophobic amino acids congregate into clusters [27, 28]. There is a need to assign the confidence score based on conservation score and knowledge of flanking region of hydrophobic residues. A confidence score has been assigned for each residue in the test protein. If the value of confidence score is more than or equal to 1, then residue is highly conserved and if the confidence score is less than 1, then the residue is variable (not well conserved evolutionarily) (Table 2).

Buried hydrophobic residues started to dominate while their confidence score was 2.07 at the range from 60 to 70. Hence, it can be concluded that the residue of the query protein is solvent exposed when it obtains a confidence score above 2.07 and the residues are buried hydrophobic residues if the value is below. (Table 2) (Figure 5).

4. Case Study

4.1. Results. For case study analysis, 10 proteins have been taken randomly from PDB which have chain length of around 300 residues. We assigned the confidence score based on query's homologous sequence to be buried and solvent exposed. After assigning the confidence score, we checked out accuracy of results based on its observed result from NACCESS server which is based on PDB structural results. Over 76% of expected results were accurate, after comparing result from case study proteins with its respective RSA value from NACCESS server.

4.2. Case Study Examples Representation Using Pymol Tool. Out of these 10 case studies, one protein has been chosen randomly to represent using Pymol tool [29]. Initially, the surface hydrophobic residues were taken into consideration from the randomly selected proteins (Figures 6(a) and 6(b)).

4.3. Accuracy Calculation. From the results obtained through the above case study, there was need for analyzing accuracy results by comparing with observed and predicted results.

Consider

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}. \quad (6)$$

In the above formula,

True positive = exposed hydrophobic residues as exposed

True negative = buried hydrophobic residue as buried

False positive = buried hydrophobic residue as exposed

False negative = exposed hydrophobic residue as buried.

After comparing result from case study examples with its respective RSA value from NACCESS server (Figure 7), we have observed that over 76% expected results were accurate. This accuracy has been improved to 78% by implementing knowledge of flanking residues hydrophobic nature.

Table 1: It represents percentage relative frequency of solvent exposed and buried hydrophobic residues in different conservation score range.

Conservation score range	% of relative frequency for solvent exposed hydrophobic residues	% of relative frequency for buried hydrophobic residues
<0	10.68	3.02
0 to 10	4.31	0.82
10 to 20	4.36	1.18
20 to 30	4.87	1.38
30 to 40	5.10	1.06
40 to 50	8.88	2.07
50 to 60	6.70	7.21
60 to 70	6.24	7.56
70 to 80	6.92	7.73
80 to 90	10.73	11.82
90 to 100	31.16	56.07

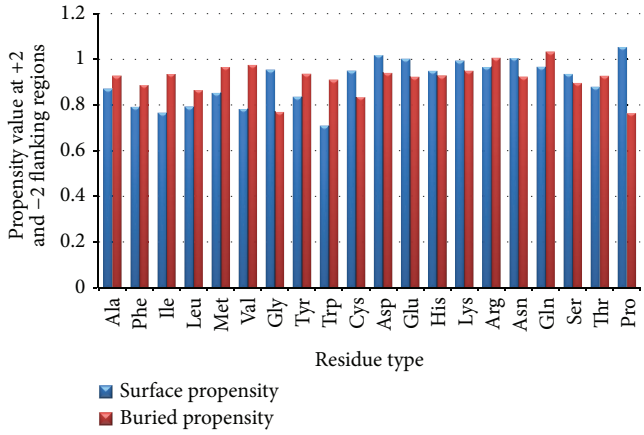


Figure 3: It refers to propensity of hydrophobic residues at +2 and −2 flanking position.

From Figures 2 and 3, we have analyzed exposed and buried residue propensity in flanking regions (+1, −1) and (+2, −2), respectively and implementation of the following points was obtained to improve the accuracy.

(i) When hydrophobic residues such as Phe, Ile, leu, Met, Val, and Cys occur in flanking regions, hydrophobic propensity values for these residues are considered to be 1. (ii) Hydrophobic propensity value is considered to be 0.75 for the partial hydrophobic residues such as Ala, Tyr, and Trp occurring in flanking regions. (iii) When Ser and Thr occur in flanking regions, the propensity value is considered as 0.35. (iv) When a hydrophilic residue occurs in flanking regions, the propensity value is considered as 0.15.

5. Conclusion

Present work is based on nonredundant dataset of monomeric proteins and we have observed that significant

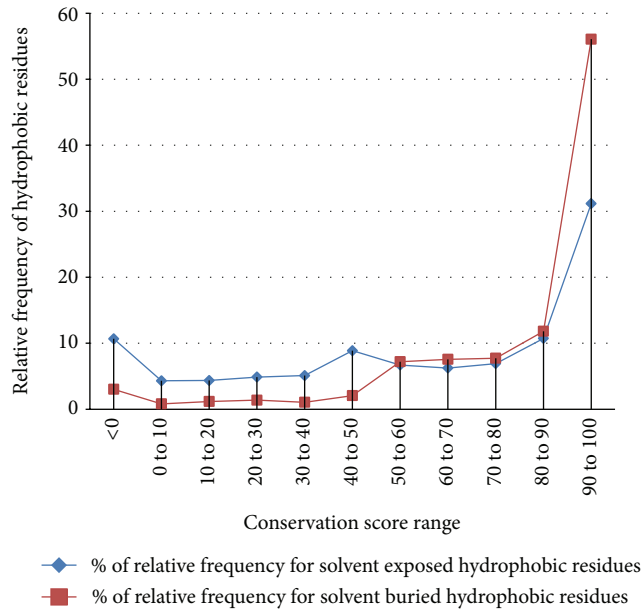


Figure 4: Relative frequency of hydrophobic residues versus conservation score range.

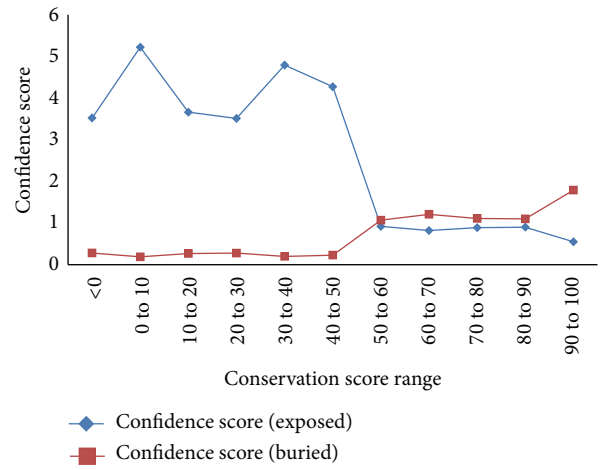


Figure 5: It represents confidence score versus conservation score range.

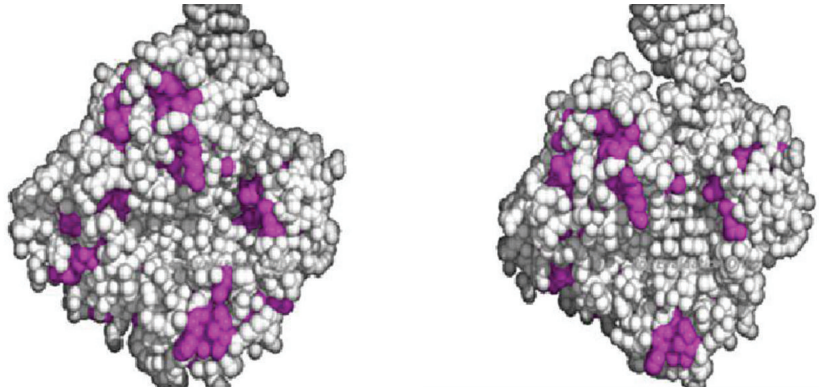
21.4% of hydrophobic residues are solvent exposed which is obtained from RSA analysis information.

After running multiple sequence alignment from the homologous sequences with respect to individual data set proteins, we came to know that exposed hydrophobic residues are poorly conserved and buried hydrophobic residues are highly conserved.

Based on the conservation score of hydrophobic residues obtained from MSA, we assigned confidence score to residues which are likely to be buried and exposed; after comparing the results from 10 proteins and doing a case study with its respective relative surface accessibility value from NACCESS server, we have observed that over 76% expected results were accurate but it has been improved to 78% by considering

Table 2: Confidence score table.

Conservation score range	% of relative frequency for solvent exposed hydrophobic residues	% of relative frequency for buried hydrophobic residues	Confidence score (exposed)	Confidence score (buried)
<0	10.68	3.02	3.52	0.282
0 to 10	4.31	0.82	5.22	0.19
10 to 20	4.36	1.18	3.66	0.27
20 to 30	4.87	1.38	3.51	0.28
30 to 40	5.10	1.06	4.79	0.20
40 to 50	8.88	2.07	4.27	0.23
50 to 60	6.70	7.21	0.92	1.07
60 to 70	6.24	7.56	0.82	1.21
70 to 80	6.92	7.73	0.89	1.11
80 to 90	10.73	11.82	0.90	1.10
90 to 100	31.16	56.07	0.55	1.79



(a) (b)

Figure 6: (a) 2UVW-*Sulfolobus solfataricus* P2 DNA polymerase IV (DPO4) (observed). (b) 2UVW-*Sulfolobus solfataricus* P2 DNA polymerase IV (DPO4) (predicted).

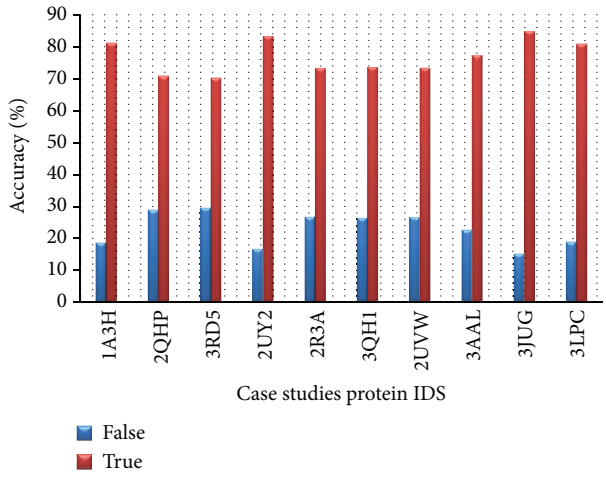


Figure 7: Case studies percentage of accuracy.

hydrophobic cluster, that is, flanking residues between +2 and −2 positions.

Knowledge on the solvation state of a residue would be used to identify the solvent exposed hydrophobic residues which can be targeted to increase stability. Hence in the work described here, the approach is adopted in developing a prediction methodology to identify the solvation state of a residue using only the information on sequence. Armed with the knowledge of only monomeric proteins, further research can be carried out to understand behavior of oligomers.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank N S Lab, IISc, Bangalore for the support and infrastructure provided and also TOCE & PES Institute of Technology, Bangalore, for their undiminished encouragement and valuable inputs in presenting the work.

References

[1] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, no. 31, pp. 7133–7155, 1990.

[2] W. Kauzmann, "Some factors in the interpretation of protein denaturation," *Advances in Protein Chemistry*, vol. 14, pp. 1–63, 1959.

[3] M. F. Perutz, J. C. Kendrew, and H. C. Watson, "Structure and function of haemoglobin: II. some relations between polypeptide chain configuration and amino acid sequence," *Journal of Molecular Biology*, vol. 13, no. 3, pp. 669–678, 1965.

[4] M. M. Gromiha, *Protein Bioinformatics: From Sequence to Function*, Academic Press, New York, NY, USA, 1st edition, 2010.

[5] C. Chothia, "Structural invariants in protein folding," *Nature*, vol. 254, no. 5498, pp. 304–308, 1975.

[6] G. D. Rose, A. R. Geselowitz, and G. J. Lesser, "Hydrophobicity of amino acid residues in globular proteins," *Science*, vol. 229, no. 4716, pp. 834–838, 1985.

[7] K. A. Sharp, A. Nicholls, R. Friedman, and B. Honig, "Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models," *Biochemistry*, vol. 30, no. 40, pp. 9686–9697, 1991.

[8] C. Strub, C. Alies, A. Lougarre, C. Ladurantie, J. Czaplicki, and D. Fournier, "Mutation of exposed hydrophobic amino acids to arginine to increase protein stability," *BMC Biochemistry*, vol. 5, article 9, pp. 1–6, 2004.

[9] M. M. Gromiha and S. Ahmad, "Role of solvent accessibility in structure based drug design," *Current Computer—Aided Drug Design*, vol. 1, pp. 223–235, 2005.

[10] G. Gianese, F. Bossa, and S. Pascarella, "Improvement in prediction of solvent accessibility by probability profiles," *Protein Engineering*, vol. 16, no. 12, pp. 987–992, 2003.

[11] B. Rost and C. Sander, "Conservation and prediction of solvent accessibility in protein families," *Proteins*, vol. 20, no. 3, pp. 216–226, 1994.

[12] B. Rost, "PHD: Predicting one-dimensional protein structure by profile-based neural networks," *Methods in Enzymology*, vol. 266, pp. 525–539, 1996.

[13] X. Li and X. M. Pan, "New method for accurate prediction of solvent accessibility from protein sequence," *Proteins*, vol. 42, no. 1, pp. 1–5, 2001.

[14] H. Naderi-Manesh, M. Sadeghi, S. Arab, and A. A. M. Movahedi, "Prediction of protein surface accessibility with information theory," *Proteins*, vol. 42, pp. 452–459, 2001.

[15] Z. Yuan, K. Burrage, and J. S. Mattick, "Prediction of protein solvent accessibility using support vector machines," *Proteins*, vol. 48, no. 3, pp. 566–570, 2002.

[16] E. D. Levy, "PiQSi: protein quaternary structure investigation," *Structure*, vol. 15, no. 11, pp. 1364–1367, 2007.

[17] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[18] "Naccess V2.1.1—solvent accessible area calculations," http://www.bioinf.manchester.ac.uk/naccess/nac_intro.html.

[19] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," *Journal of Molecular Biology*, vol. 55, no. 3, pp. 379–400, 1971.

[20] S. J. Hubbard, S. F. Campbell, and J. M. Thornton, "Molecular recognition. conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors," *Journal of Molecular Biology*, vol. 220, no. 2, pp. 507–530, 1991.

[21] Z.-Y. Zhu and T. L. Blundell, "The use of amino acid patterns of classified helices and strands in secondary structure prediction," *Journal of Molecular Biology*, vol. 260, no. 2, pp. 261–276, 1996.

[22] M. H. Zehfus, "Automatic recognition of hydrophobic clusters and their correlation with protein folding units," *Protein Science*, vol. 4, no. 6, pp. 1188–1202, 1995.

[23] M. S. S. Chang and S. A. Benner, "Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments," *Journal of Molecular Biology*, vol. 341, no. 2, pp. 617–631, 2004.

[24] J. M. Koshi and R. A. Goldstein, "Mutation matrices and physical-chemicalproperties: correlations and implications," *Proteins*, vol. 27, no. 3, pp. 336–344, 1997.

[25] M. D. Finucane and D. N. Woolfson, "Core-directed protein design. II. rescue of a multiply mutated and destabilized variant of ubiquitin," *Biochemistry*, vol. 38, no. 36, pp. 11613–11623, 1999.

[26] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[27] I. Callebaut, G. Labesse, P. Durand et al., "Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives," *Cellular and Molecular Life Sciences*, vol. 53, no. 8, pp. 621–645, 1997.

[28] C. Gaboriaud, V. Bissery, T. Benchetrit, and J. P. Mornon, "Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences," *FEBS Letters*, vol. 224, no. 1, pp. 149–155, 1987.

[29] "The PyMOL Molecular Graphics System," http://www.pymol.org/.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 313747, 9 pages
http://dx.doi.org/10.1155/2014/313747

Research Article

Nonlinear-Model-Based Analysis Methods for Time-Course Gene Expression Data

Li-Ping Tian,[1] Li-Zhi Liu,[2] and Fang-Xiang Wu[2,3]

[1] *School of Information, Beijing Wuzi University, No. 1 Fuhe Street, Tongzhou District, Beijing 101149, China*
[2] *Department of Mechanical Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK, Canada S7N 5A9*
[3] *Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK, Canada S7N 5A9*

Correspondence should be addressed to Fang-Xiang Wu; faw341@mail.usask.ca

Received 27 August 2013; Accepted 16 October 2013; Published 2 January 2014

Academic Editors: B. Shen, J. Wang, and J. Wang

Copyright © 2014 Li-Ping Tian et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Microarray technology has produced a huge body of time-course gene expression data and will continue to produce more. Such gene expression data has been proved useful in genomic disease diagnosis and drug design. The challenge is how to uncover useful information from such data by proper analysis methods such as significance analysis and clustering analysis. Many statistic-based significance analysis methods and distance/correlation-based clustering analysis methods have been applied to time-course expression data. However, these techniques are unable to account for the dynamics of such data. It is the dynamics that characterizes such data and that should be considered in analysis of such data. In this paper, we employ a nonlinear model to analyse time-course gene expression data. We firstly develop an efficient method for estimating the parameters in the nonlinear model. Then we utilize this model to perform the significance analysis of individually differentially expressed genes and clustering analysis of a set of gene expression profiles. The verification with two synthetic datasets shows that our developed significance analysis method and cluster analysis method outperform some existing methods. The application to one real-life biological dataset illustrates that the analysis results of our developed methods are in agreement with the existing results.

1. Background

To understand the mechanisms of dynamic biological processes, DNA microarray experiments have been employed to produce gene expression profiles at a series of time points, for example, the cell division cycle processes of yeast *Saccharomyces cerevisiae* [1, 2], bacterium *Caulobacter crescentus* [3], and human being [4]. Such time-course gene expression data provides a dynamic snapshot of most (if not all) of the genes related to the biological development process and thus can be useful in genomic disease diagnosis and genomic drug design. The challenge is how to uncover useful information from such data by proper analysis methods [5].

Although the behaviours of genome-wide genes can be monitored simultaneously with the current DNA microarray technology, not are all of monitored genes closely related to the biological process being studied or of interest. In addition, gene expression data are often contaminated by various noises or noisy genes. It is impossible to uncover some useful information without any preprocessing. Either excluding genes of interest or including noisy genes could degrade the significance of any analysis results. Therefore, it is critical to select the genes which are closely relevant to a biological process from gene expression profiles measured during the biological process. The selection of genes can be performed by the so-called significance analysis of gene expression profiles. Much attention has been paid to the significant analysis of static gene expression data over the past years. For gene expression data obtained from a pair of conditions (e.g., normal versus abnormal, or control versus treatment) with multiple replicates, one of the widely used approaches in early years is called the R-fold change method [6, 7]. The "R-fold change" method determines a gene to be significantly expressed if the ratio of expression values under two different conditions is greater than R or less than $1/R$, where R is a user-preset positive number. This approach has been improved by a resampling (bootstrapping) method called SAM [8, 9]. Another approach to the significance analysis is the use of

t-test, for example, on logarithm of the expression levels. In a t-test, the means and variances of gene expressions from a pair of conditions are used to compute a normalized distance so-called t-value. When the t-value exceeds a certain threshold depending on the confidence level selected, gene expression data from a pair of conditions are considered to be significantly different. Although R-fold and t-test approaches can be extended to apply for the analysis of gene expression data with multiple conditions, for example, SAM [8, 9] and RIT [10], these approaches need the assumption that multiconditional values are statistically independent. Therefore, it is not applicable to time-course gene expression profiles as they are not statistically independent but dynamically dependent. In recent year, we have developed several methods for significance analysis of time-course gene expression data. In [11, 12], we employ linear regression models to detect the significantly differentially expressed genes. In [13, 14], we employ nonlinear models to detect the periodically expressed genes.

Besides the significance analysis, the cluster analysis is another class of analysis methods to uncover the useful information from gene expression data [5]. A number of clustering methods have been proposed for cluster analysis of gene expression data. These include distance/correlation-based clustering methods (e.g., hierarchical clustering [15], k-means clustering [16], and self-organizing maps [17]) and static-model-based clustering methods [18, 19]. In these methods, gene expression profiles are viewed as multidimensional vectors. Distance/correlation-based clustering methods cluster genes based on the distance/correlation among their expression profiles. Static-model-based clustering methods assign genes to one of the clusters if their expression profiles are generated by a multivariate normal distribution. These methods do not take the dynamic of time-course gene expression data into account and thus are not efficient for periodically expressed gene data. Some dynamic-model-based clustering methods have been proposed to analyze time-course gene expression data [20, 21]. These methods employ linear autoregressive models to describe the dynamics of time-course gene expression data. Recently we propose the nonlinear-model-based method for clustering periodically expressed genes [22, 23].

As measured from typical nonlinear biological systems, time-course gene expression profiles should display the nonlinear properties. In this paper, we propose nonlinear-model-based methods for significance analysis and cluster analysis of time-course gene expression data. The proposed nonlinear model can be viewed as a generalization of many existing models [13, 14, 20–23]. A two-step method is proposed to estimate the model parameter. An F-test is employed to determine if a gene expression profile is significantly different from noisy data. A relocation-iteration algorithm is employed to assign each gene to an appropriate cluster. A bootstrapping method and an average adjusted Rand index (AARI) are employed to measure the quality of clustering. We employ two synthetic datasets to evaluate the performance of the proposed methods and apply them to one real-life biological dataset.

2. Methods

2.1. Nonlinear Model for Time-Course Gene Profiles. Let $x(t)$ $(t = 1, 2, \ldots, m)$ be a time-course gene expression profile generated from a dynamic biological process, where m is the number of time points at which gene expression is measured. Many nonlinear gene expression profiles contain a periodic component and a long-term decrease or increase component. In this study, we employ the following nonlinear model to describe time-course gene expression data:

$$x(t) = e^{\alpha t}\left[a\cos(\omega t) + b\sin(\omega t)\right] + ct + d + \varepsilon(t), \quad (1)$$

where parameter α represents the degradation rate of periodicity; parameters a and b are the coefficients of sine and cosine functions, respectively; parameter ω is the frequency of periodic expression data; parameters c and d are the coefficients of linear function; and $\varepsilon(t)$ represents random errors. This study assumes that the errors have a normal distribution independent of time with the mean of 0 and the variance of σ^2. This model generalizes several existing models; for example, setting $\alpha = c = d = 0$, model (1) is reduced to sinusoidal function model [24–30]:

$$x(t) = A\sin(\omega t + \Phi) + \varepsilon(t), \quad (2)$$

which is widely used to generate the synthetic periodic gene expression profiles [24] and to detect the periodically expressed genes [27–29]. In model (2), $A = \sqrt{a^2 + b^2}$ is called the magnitude and $\Phi = \arctan(a/b)$ is called the phase. Setting $\alpha = 0$, model (1) is reduced to a model used in [13], while, setting $c = d = 0$, model (1) is reduced to a model used in [14, 22]. As model (1) is the generalization of several existing models, it is expected that the analysis results based on this model are better than those reduced models.

To construct model (1) six parameters need to be estimated from a time-course gene expression profile $x(t)$ $(t = 1, 2, \ldots, m)$. Obviously estimating those parameters in model (1) is a nonlinear estimation problem as α and ω are nonlinear in the model. In general, all nonlinear optimization programs can be used to estimate parameters in model (1), for example, Gauss-Newton iteration method and its variants such as Box-Kanemasu interpolation method, Levenberg damped least squares methods, and Marquardt's method [31–33]. However, these iteration methods are sensitive to initial values. Another main shortcoming is that these methods may converge to the local minimum of the least squares cost function and thus cannot find the true values of the parameters.

Our observation is that noise free model (1)

$$x(t) = e^{\alpha t}\left[a\cos(\omega t) + b\sin(\omega t)\right] + ct + d \quad (3)$$

can be viewed as the general solution of the following second order ordinary differential equation:

$$\ddot{x}(t) + A\dot{x}(t) + Bx(t) = Ct + D, \quad (4)$$

which is independent of a and b and

$$\alpha = -\frac{A}{2}, \qquad \omega = \frac{\sqrt{4B - A^2}}{2}, \qquad c = \frac{C}{B}, \qquad d = \frac{DB - AC}{B^2}. \quad (5)$$

Now we can see that constant parameters A, B, C, and D are linear in (4). As long as we get the first and second derivatives, we can easily estimate the parameters A, B, C, and D by the linear least squares method. Then we can get the estimation of α, ω, c, and d from equations in (5). Finally we can use (3) to estimate the rest of parameters a and b. Therefore, we propose the following two-step parameter estimation methods to estimate all six parameters in model (1).

Step 1. Numerically calculate the first and second derivatives of $x(t)$. As time-course gene expression data are discrete, the first and second derivatives of $x(t)$ can be estimated by the central (second order) finite difference formulas as follows:

$$\dot{x}(t) = \frac{x(t+1) - x(t-1)}{2\Delta} \quad \text{for } t = 2, \ldots, m-1, \tag{6}$$

$$\ddot{x}(t) = \frac{x(t+1) + x(t-1) - 2x(t)}{\Delta^2} \quad \text{for } t = 2, \ldots, m-1, \tag{7}$$

respectively, where Δ is the time difference between two consecutive gene expression data points. If the number of data points in a gene expression profile is enough, one can choose a high order finite difference formula to get more accurate estimation of these derivatives.

Then, based on model (4), we use the linear least squares method to estimate parameter ω^2. In detail, let

$$Y = \begin{bmatrix} \ddot{x}(1) \\ \ddot{x}(2) \\ \vdots \\ \ddot{x}(l) \end{bmatrix}, \qquad X = \begin{bmatrix} -\dot{x}(1) & -x(1) & t_2 & 1 \\ -\dot{x}(2) & -x(2) & t_3 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ -\dot{x}(l) & -x(l) & t_{m-1} & 1 \end{bmatrix}. \tag{8}$$

From (6) and (7), we have $l = m-2$. Then by the least squares method, the paramters A, B, C, and D in model (4) can be estimated as

$$\begin{bmatrix} \widehat{A} \\ \widehat{B} \\ \widehat{C} \\ \widehat{D} \end{bmatrix} = \left(X^T X\right)^{-1} X^T Y. \tag{9}$$

Note that if the value of $4\widehat{B} - \widehat{A}^2$ calculated by (5) for a gene is negative, the expression of this gene will be judged not to be described by model (1).

Step 2. Substitute the estimated values of α, ω, c, and d into (3). Then we apply the least squares method to model (1) to estimate parameters a and b. In detail, let

$$Z = [z(1), \ldots, z(m)], \qquad E = \begin{bmatrix} \cos(\Delta\widehat{\omega}), \ldots, \cos(m\Delta\widehat{\omega}) \\ \sin(\Delta\widehat{\omega}), \ldots, \sin(m\Delta\widehat{\omega}) \end{bmatrix}; \tag{10}$$

by the least squares method, a and b can be estimated as

$$\begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \left(EE^T\right)^{-1} \left(EZ^T\right), \tag{11}$$

where

$$z(t) = e^{-\widehat{\alpha}t} \left[x(t) - \widehat{c}t - \widehat{d}\right] \quad \text{for } t = 1, 2, \ldots, m. \tag{12}$$

2.2. Nonlinear-Model-Based Significance Analysis. Significance analysis of gene expression data is to determine if a gene expression profile is significantly different from noisy data. This issue is not easy to answer through statistical inference [29, 30] yet, especially for time-course gene expression profiles as their data points are not statistically independent. However, a practical way in the literature [27–30] is to perform a statistical hypothesis test whether the gene expression profile is pure normal white noise or it fits a certain model as specified by (1). Along with this way, this study tests the null hypothesis of

(H_0)

$$x(t) = d + \varepsilon(t) \tag{13}$$

versus the alternative hypothesis of

(H_1) (see (1)).

Let

$$S_0^2 = \sum_{i=1}^{m} \left(x(t_i) - \widehat{d}\right)^2, \qquad \widehat{d} = \frac{1}{m} \sum_{i=1}^{m} x(t_i), \tag{14}$$

where S_0^2 is the residual of model (13) with estimated parameters, and

$$S_1^2 = \sum_{i=1}^{m} \left\{x(t_i) - e^{\widehat{\alpha}t} \left[a\cos\left(\widehat{\omega}t_i\right) + \widehat{b}\sin\left(\widehat{\omega}t_i\right)\right] - \widehat{c}t_i - \widehat{d}\right\}^2, \tag{15}$$

where S_1^2 is the residual of model (1) with estimated parameters. As the noise model (13) can be viewed as a special case of model (1), the statistic

$$F = \frac{\left(S_0^2 - S_1^2\right)/5}{S_1^2/(m-6)} = \frac{m-6}{5}\left(\frac{S_0^2}{S_1^2} - 1\right) \tag{16}$$

follows the F-distribution with the degrees of freedom $(5, m-6)$, according to statistics theory [21, 23].

When the value of F-statistic is large enough (greater than a specified threshold), model (13) is rejected; that is, the gene expression profile is not pure normal white noise, and otherwise the gene expression profile appears as white noises. According to degrees of freedom (which are related to the length of time-course data m and the number of parameters in the models) and a significance level (typically, 0.01, 0.05, 0.1, 0.2, or the like) specified by a user, the threshold

value can be determined from F-distribution table or by using a f-distribution table or a standard MATLAB function $icdf(\text{“}f\text{”}, 1-\gamma, 5, m-6)$, where γ is the significance level. A significance level is the probability that the null hypothesis is true. Therefore, the rejection of the null hypothesis at a smaller significance level indicates the more favourable to alternative hypothesis. That is, the smaller the significance level is, the more confidence one accepts that genes are not noises if its corresponding value of F-statistic is greater than the threshold.

2.3. Nonlinear-Model-Based Cluster Analysis

2.3.1. The Mixture Model. In this study, it is assumed that a time-course gene expression dataset is a collection of time series which belongs to several clusters and time series in each cluster can be described by model (1) with different parameters. Let $\theta_k = [\alpha_k, \omega_k, a_k, b_k, c_k, d_k]$ be parameters of model (1) for the kth cluster. Then the task of nonlinear-model-based clustering is as follows: for a given number of cluster K, divide a time-course gene expression dataset into a partition $C = \{C_1, \ldots, C_k, \ldots, C_K\}$ using model (1) with parameters $\theta_k = [\alpha_k, \omega_k, a_k, b_k, c_k, d_k] (k = 1, \ldots, K)$ which minimize

$$f(C \mid \Theta) = \sum_{k=1}^{K} \sum_{x \in C_k} \sum_{i=1}^{m} \left\{ x(i) - e^{\alpha_k \Delta i} \left[a_k \cos\left(i \Delta \omega_k\right) + b_k \sin\left(i \Delta \omega_k\right) \right] - c_k \Delta i - d_k \right\}^2, \tag{17}$$

where the parameters Θ consist of $\{\theta_k,\ k = 1, \ldots, K\}$.

2.3.2. Estimation of Cluster Parameters. According to the parameter estimation method proposed in previous section for single time-course expression profiles, for the kth cluster, parameters $\theta_k = [\alpha_k, \omega_k, a_k, b_k, c_k, d_k]$ can be estimated as

$$\widehat{\alpha}_k = -\frac{\widehat{A}_k}{2}, \qquad \widehat{\omega}_k = \frac{\sqrt{4\widehat{B}_k - \widehat{A}_k^2}}{2}, \qquad \widehat{c}_k = \frac{\widehat{C}}{\widehat{B}}, \qquad \widehat{d}_k = \frac{\widehat{D}_k \widehat{B}_k - \widehat{A}_k \widehat{C}_k}{\widehat{B}_k^2}, \tag{18}$$

where

$$\begin{bmatrix} \widehat{A}_k \\ \widehat{B}_k \\ \widehat{C}_k \\ \widehat{D}_k \end{bmatrix} = \left(\sum_{x \in C_k} X^T X \right)^{-1} \sum_{x \in C_k} X^T Y, \qquad \begin{bmatrix} \widehat{a}_k \\ \widehat{b}_k \end{bmatrix} = \left(\sum_{x \in C_k} E E^T \right)^{-1} \sum_{x \in C_k} E Z^T, \tag{19}$$

where $|C_k|$ represents the number of time series in cluster C_k, $\sum_{k=1}^{K} |C_k| = N$.

2.3.3. Algorithm for Clustering. This study employs the following relocation-iteration algorithm to estimate the parameters such that the cost function (17) is minimized:

(1) select an initial partition for the given number of clusters, K;

(2) iterate ($s = 1, 2, \ldots$):

(a) estimate the parameter Θ^s based on the current partition by using (18)-(19);

(b) generate a new partition by assigning each sequence x to cluster k where

$$k = \arg\min_{1 \le j \le K} \sum_{i=1}^{m} \left\{ x(i) - e^{\alpha_j^s \Delta i} \left[a_j^s \cos\left(i \Delta \omega_j^s\right) + b_j^s \sin\left(i \Delta \omega_j^s\right) \right] - c_j^s \Delta i - d_j^s \right\}^2; \tag{20}$$

(3) stop if the improvement of the cost function (17) is below a given threshold, or the cluster memberships of time series do not change significantly.

In 2(a), $\Theta^s = \{\theta_k^s, 1 \le k \le K\}$ represents the estimated parameters in cost function (17) at iteration s while in 2(b), parameters α_j^s, ω_j^s, a_j^s, b_j^s, c_j^s, and d_j^s represent the parameters of model j at iteration s.

3. Evaluation

In this section, we use two synthetic datasets to evaluate our proposed significance analysis method and cluster analysis method, respectively. To evaluate the significance analysis method, we generate one synthetic dataset that consists of 2000 noisy gene expression profiles based on model (13) and 2000 time-course gene expression profiles based on model (1). All 4000 expression profiles are depicted in Figure 1, from which one can hardly differentiate time-course gene expression profiles from noisy ones. To measure the performance of significance analysis, we employ two widely used indices: sensitivity and specificity, which can be defined as follows [34]:

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}},$$

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}, \tag{21}$$

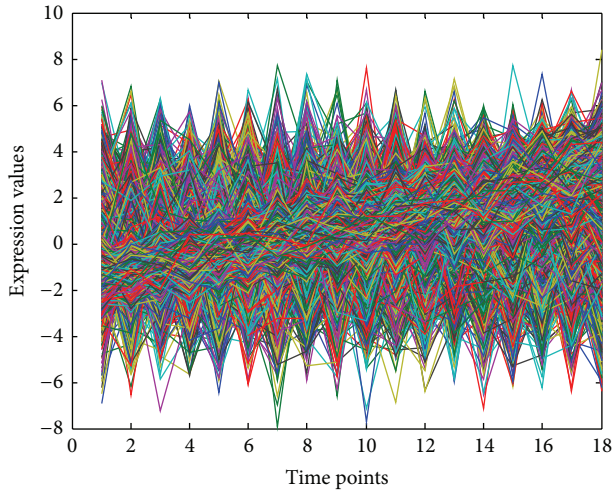


Figure 1: Plot of 4000 expression profiles for evaluating significance analysis method.

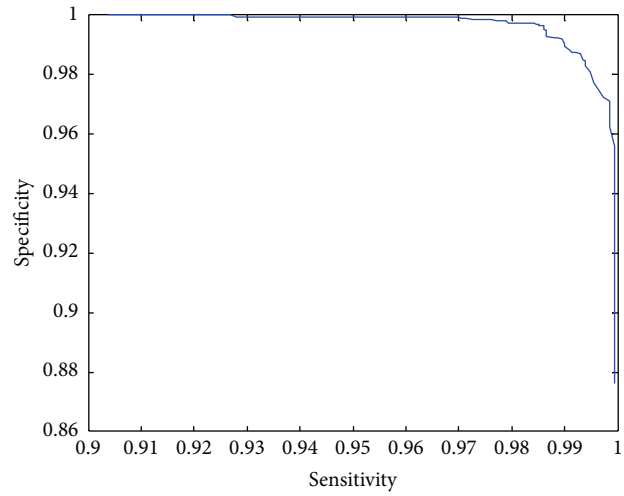


Figure 2: Plot of sensitivity versus specificity.

where

true positive is a time-course gene expression profile identified as it is;

false positive is a time-course gene expression profile identified as it is noisy;

true negative is a noisy gene expression profile identified as it is;

false negative is a noisy gene expression profile identified as it is time-course.

The sensitivity and the specificity depend on thresholds which determine if an expression profile is time-course or noisy. In general, the sensitivity is increasing, while the specificity is decreasing and vice versa. However, a good method is expected to have both high sensitivity and specificity. Figure 2 depicts the curves of sensitivity versus specificity over various thresholds. From this figure, we can see that both sensitivity and specificity can be as high as of 99% for a specific threshold, which indicates that our proposed significance analysis methods are excellent.

To evaluate our proposed cluster analysis method, another synthetic dataset consisting of six clusters is generated from model (1), where different clusters have different randomly selected parameters with some large variances. In each cluster, all profiles are generated with model parameters for this cluster with some random perturbations. All generated profiles are plotted in Figure 3, from which one can see that all time-course gene expression profiles are mixed up. To measure the quality of clustering results, we use the adjusted Rand index (ARI) [35], which originally is to measure the degree of agreement between two partitions of the same set of objects. Given two partitions of N objects, the r-cluster partition $U = \{u_1, \ldots, u_r\}$ and the s-cluster partition $V = \{v_1, \ldots, v_s\}$, the ARI is defined as follows [35]:

$$\text{ARI} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{s}\binom{n_{ij}}{2} - 1/T \sum_{i=1}^{r}\binom{n_{i.}}{2}\sum_{i=1}^{s}\binom{n_{.j}}{2}}{1/2\left[\sum_{i=1}^{r}\binom{n_{i.}}{2} + \sum_{i=1}^{s}\binom{n_{.j}}{2}\right] - (1/T)\sum_{i=1}^{r}\binom{n_{i.}}{2}\sum_{i=1}^{s}\binom{n_{.j}}{2}}, \quad (22)$$

where T is the number of pairs of N objects, n_{ij} is the number of objects that are both in clusters u_i and v_j, $i = 1, \ldots, r$, $j = 1, \ldots, s$, and $n_{i.}$ is the number of objects in cluster u_i, while $n_{.j}$ is the number of objects in cluster v_j. From these definitions, we have

$$T = \frac{N(N-1)}{2}, \qquad n_{i.} = \sum_{j=1}^{s} n_{ij}, \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}. \quad (23)$$

The expected value of ARI is 1 when two partitions agree perfectly and 0 when they are selected at random.

As the results of clustering are sensitive to the initial partition, we run our proposed clustering algorithm and competing clustering algorithms 30 times on the synthetic dataset and calculate the average ARI (AARI) for each algorithm. Figure 4 depicts the AARI of three algorithms named "algorithm with random initial," "algorithm with k-means initial," and "k-means" over several different numbers of clusters, where "algorithm with random initial" means our proposed clustering algorithm with randomly chosen initial partition, "algorithm with k-means initial" means our proposed clustering algorithm with k-means result as initial partition, and "k-means" is an algorithm coded in the MATLAB software for k-means clustering method. Those values of AARI are also listed in Table 1.

From Figure 4 and Table 1, one can see that our algorithm with random chosen initial partitions outperforms the other two algorithms. Particularly, at the correct number of clusters,

Table 1: The values of AARI for different clustering methods on synthetic data.

No. of clusters	2	3	4	5	6	7	8	9	10
Random initial	0.2915	0.5741	0.6636	0.7549	0.9787	0.9516	0.8862	0.826	0.7944
k-means initial	0.2915	0.4875	0.6741	0.7168	0.7732	0.7668	0.7666	0.7739	0.753
k-means	0.2915	0.5099	0.6352	0.7047	0.8001	0.7635	0.8189	0.7849	0.7827

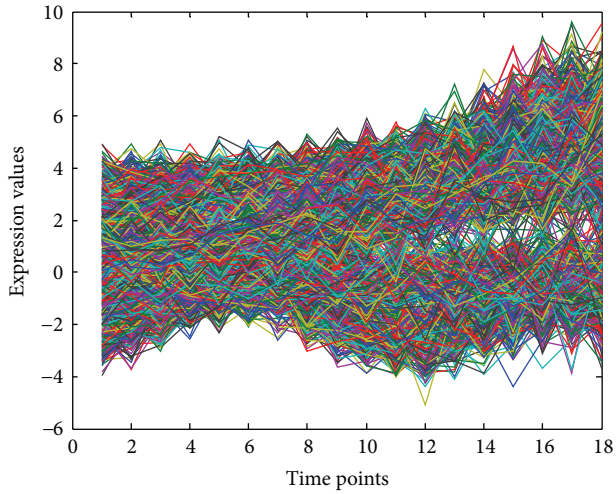


Figure 3: Plot of expression profiles for evaluating cluster analysis method.

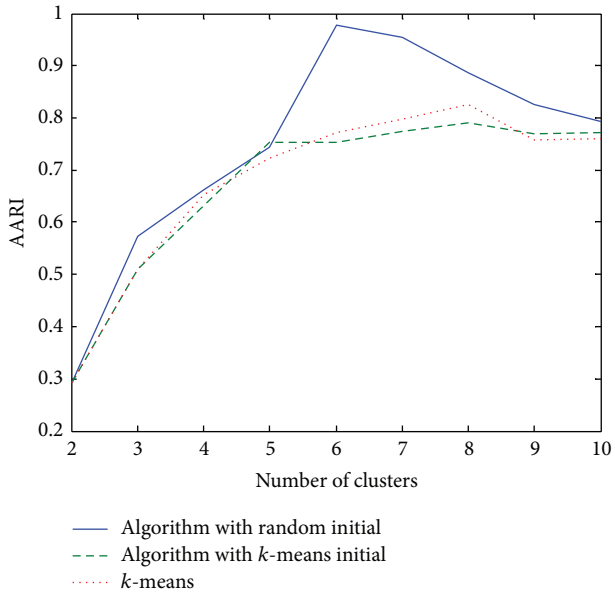


Figure 4: Plot of AARI with different numbers of clusters.

the ARRI from our algorithm with random chosen initial partitions reaches its maximum. The quality of our algorithm with k-means result as initial partitions is comparable with that of k-means, which indicates that after k-means falls in a local optimum, our algorithm cannot escape from that local optimum and thus inherits the drawbacks of k-means. This also suggests that our developed algorithm should combine with random chosen initial partitions.

4. Applications to a Real-Life Gene Expression Data

In this section, we apply our proposed significance analysis and cluster analysis method to a real-life gene expression dataset which is collected from the alpha-synchronized experiment [2]. To study the mitotic cell division cycle of yeast, Spellman et al. [2] have monitored more than 6000 genes of yeast (*Saccharomyces cerevisiae*) at 18 equally spacing time points in the alpha-synchronized experiment. The original dataset is publicly available at http://genome-www.stanford.edu. Genes with missing data are excluded in this study. The resultant dataset contains the expression profiles of 4489 genes.

We first apply our proposed significance analysis method to this dataset and set the significance level $\gamma = 0.1$. As a result, 846 genes are determined to be involved in the alpha-synchronized cell division cycle process, while the other 3643 genes are determined to be noises with respect to this process. Figure 5(f) depicts these 3643 expression profiles. From Figure 5(f), most of the expression profiles look like noises and are not related to the alpha-synchronized cell division cycle process according to the results in [2]. Then we apply our proposed clustering algorithm to the subset consisting of 846 genes involved in the alpha-synchronized cell division cycle process. According to the biological meaning of this process [2], the reasonable number of clusters is 5. The model parameters identified for each of the five clusters are listed in Table 2. From Table 2, for all clusters the values of parameter α_k are negative numbers, which are reasonable. As the cell division cycle is a stable biological system, after a perturbation such as the alpha synchronization, the system tends to its stable attractor. Therefore the degradation rate represented by α_k should be negative.

Furthermore, the values of model parameters a_k and b_k determine the importance of periodic components. From Table 2, the module of parameters a_k and b_k is the largest, while the absolute value of parameter α_k is small for Cluster 3. This indicates that 17 genes in Cluster 3 are periodically expressed in this process, which can be verified from Figure 5(c). Actually all 17 genes in this cluster have also been identified as periodically expressed genes in [2]. The module of parameters a_k and b_k is the second largest for Cluster 5, while the absolute value of parameter α_k is very large for Cluster 5. As a result, gene expression profiles in Cluster 5 are quickly degrading while hardly displaying periodicity as

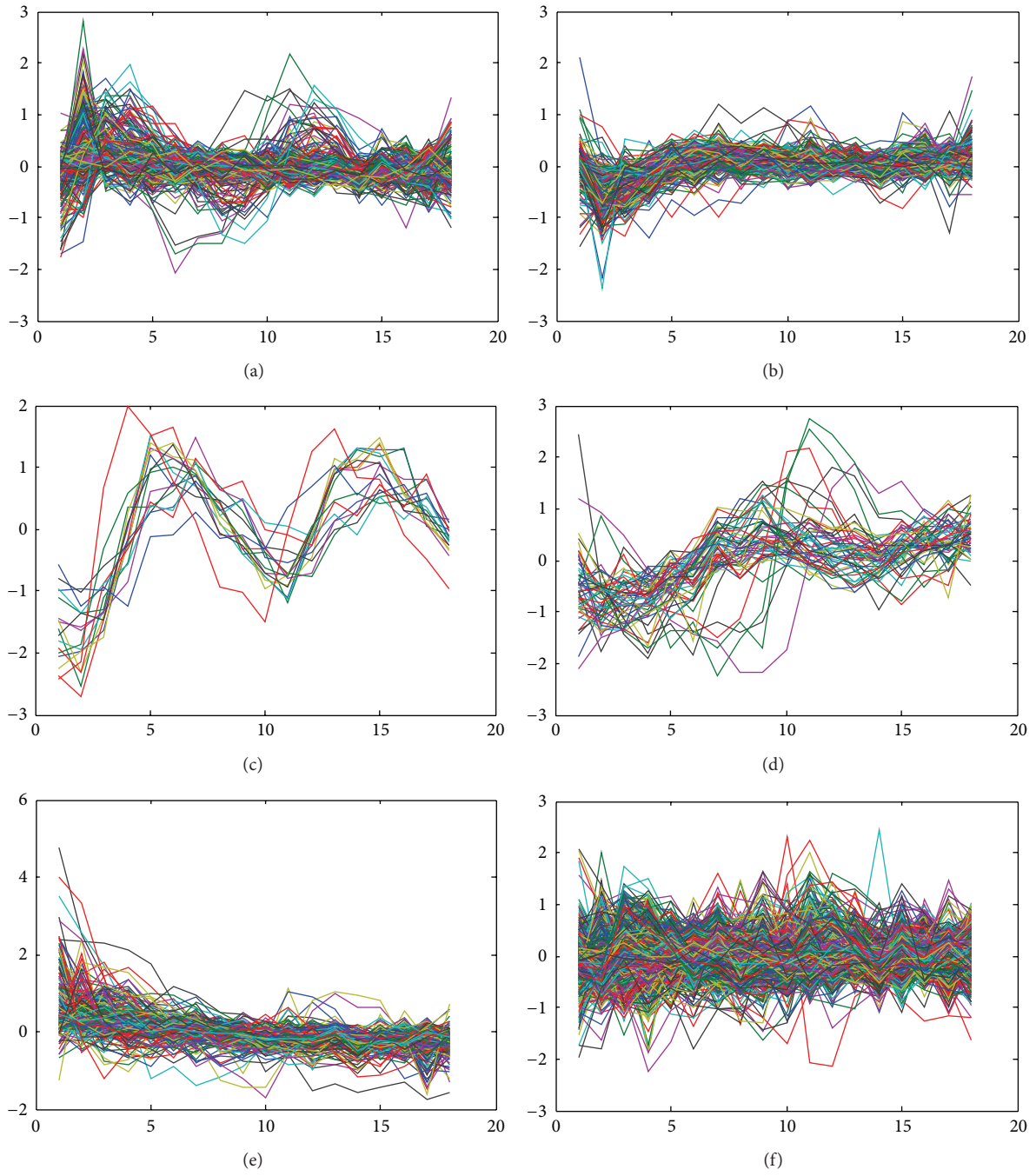


Figure 5: Plot of gene expression profiles. (a)–(e) show gene expression profiles for one of five clusters. (f) shows gene expression profiles which are determined as noises.

Table 2: The model parameters for each cluster.

Parameters	Cluster 1 (315)	Cluster 2 (233)	Cluster 3 (17)	Cluster 4 (53)	Cluster 5 (228)
α_k	−1.1543	−1.7033	−0.6612	−0.5111	−1.8483
ω_k	9.8108	9.8673	7.1631	7.0517	8.736
a_k	0.0234	0.2675	1.0948	0.0024	0.4427
b_k	0.1389	0.033	−1.2261	0.1248	−0.6807
c_k	−0.1287	0.1422	0.3353	0.5748	−0.3738
d_k	0.1383	−0.1372	−0.2723	−0.6011	0.3946

shown in Figure 5(e). According to the estimated values of model parameters, expression profiles in other clusters can similarly be explained.

5. Conclusions

In this paper, we have presented a significance analysis method and a cluster analysis method for time-course gene expression profiles. In these methods, time-course gene expression profiles are modeled by a nonlinear model, which is a generalization of several existing models. To estimate the parameters, which is key to the developed significance analysis method and a cluster analysis method, we propose a two-step linear least squares method. One synthetic dataset has been employed to verify our developed significance analysis method in terms of sensitivity and specificity, while another synthetic dataset has been employed to verify our developed cluster analysis method in terms of AARI. The results have shown that both of our developed methods outperform some existing methods. The application to one real-life biological dataset illustrates that the analysis results of our developed methods are in agreement with the existing results. The reconstruction of gene regulatory network from time-course gene expression data is a very important issue in systems biology [36]. Obviously, noisy genes should be excluded from gene expression data for reconstructing gene regulatory networks. In the future, we may combine our method with other methods as in [36] to reconstruct gene regulatory networks.

Conflict of Interests

The authors declare that there are no competing interests.

Acknowledgments

This research is supported by the Special Fund of Ministry of Education of Beijing for Distinguishing Professors and the Science and Technology Funds of Beijing Ministry of Education (SQKM201210037001) through the first author and Natural Sciences and Engineering Research Council of Canada (NSERC) through the other authors.

References

[1] R. J. Cho, M. J. Campbell, E. A. Winzeler et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[2] P. T. Spellman, G. Sherlock, M. Q. Zhang et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[3] M. T. Laub, H. H. McAdams, T. Feldblyum, C. M. Fraser, and L. Shapiro, "Global analysis of the genetic network controlling a bacterial cell cycle," *Science*, vol. 290, no. 5499, pp. 2144–2148, 2000.

[4] M. L. Whitfield, G. Sherlock, A. J. Saldanha et al., "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1977–2000, 2002.

[5] P. Baldi and G. W. Hatfield, *DNA Microrarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press, New York, NY, USA, 2002.

[6] A. A. Alizadeh, M. B. Elsen, R. E. Davis et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[7] J. M. Claverie, "Computational methods for the identification of differential and coordinated gene expression," *Human Molecular Genetics*, vol. 8, no. 10, pp. 1821–1832, 1999.

[8] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[9] G. K. Smyth, J. Michaud, and H. S. Scott, "Use of within-array replicate spots for assessing differential expression in microarray experiments," *Bioinformatics*, vol. 21, no. 9, pp. 2067–2075, 2005.

[10] R. Nilsson, J. M. Pena, J. Bjorkegren, and J. Tegner, "Detecting multivariate differentially expressed genes," *BMC Bioinformatics*, vol. 8, article 150, 2007.

[11] F. X. Wu and W. J. Zhang, "Dynamic-model-based method for selecting significantly expressed genes from time-course expression profiles," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 16–22, 2010.

[12] F. X. Wu, Z. H. Xia, and L. Mu, "Finding significantly expressed genes from time-course expression profiles," *International Journal of Bioinformatics Research and Applications*, vol. 5, no. 1, pp. 50–63, 2009.

[13] L. P. Tian, L. Z. Liu, and F. X. Wu, "Identification of pseudo-periodic gene expression profiles," in *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP '11)*, pp. 42–46, Las Vegas, Nev, USA, July 2011.

[14] L. P. Tian, L. Z. Liu, and F. X. Wu, "Detecting nearly periodically expressed genes from their microarray time-course profiles," in *Proceedings of the IASTED International Conference on Computational Bioscience (CompBio '10)*, pp. 612–619, November 2010.

[15] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[16] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.

[17] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, "Analysis of gene expression data using self-organizing maps," *FEBS Letter*, vol. 451, no. 2, pp. 142–146, 1999.

[18] D. Ghosh and A. M. Chinnaiyan, "Mixture modelling of gene expression data from microarray experiments," *Bioinformatics*, vol. 18, no. 2, pp. 275–286, 2002.

[19] G. J. McLachlan, R. W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, vol. 18, no. 3, pp. 413–422, 2002.

[20] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 14, pp. 9121–9126, 2002.

[21] F. X. Wu, W. J. Zhang, and A. J. Kusalik, "Dynamic model-based clustering for time-course gene expression data," *Journal*

of Bioinformatics and Computational Biology, vol. 3, no. 4, pp. 821–836, 2005.

[22] L. P. Tian, H. Y. Lu, and F. X. Wu, "Nonlinear model-based clustering for periodically expressed gene profiles," in *Proceedings of the 7th International Symposium on Bioinformatics Research and Applications (ISBRA '11)*, pp. 62–66, Hunan, China, May 2011.

[23] L. P. Tian, L. Z. Liu, Q. W. Zhang, and F. X. Wu, "Nonlinear model-based method for clustering periodically expressed genes," *TheScientificWorldJOURNAL*, vol. 11, pp. 2051–2061, 2011.

[24] S. L. Harmer, J. B. Hogenesch, M. Straume et al., "Orchestrated transcription of key pathways in Arabidopsis by the circadian clock," *Science*, vol. 290, no. 5499, pp. 2110–2113, 2000.

[25] S. Wichert, K. Fokianos, and K. Strimmer, "Identifying periodically expressed transcripts in microarray time series data," *Bioinformatics*, vol. 20, no. 1, pp. 5–20, 2004.

[26] R. A. Fisher, "Test of significance in harmonic analysis," *Proceedings of the Royal Society A*, vol. 125, no. 796, pp. 54–59, 1929.

[27] J. Chen, "Identification of significant periodic genes in microarray gene expression data," *BMC Bioinformatics*, vol. 6, article 286, 2005.

[28] E. F. Glynn, J. Chen, and A. R. Mushegian, "Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms," *Bioinformatics*, vol. 22, no. 3, pp. 310–316, 2006.

[29] J. Chen and K. C. Chang, "Discovering statistically significant periodic gene expression," *International Statistical Review*, vol. 76, no. 2, pp. 228–246, 2008.

[30] A. W. C. Liew, N. F. Law, X. Q. Cao, and H. Yan, "Statistical power of Fisher test for the detection of short periodic gene expression profiles," *Pattern Recognition*, vol. 42, no. 4, pp. 549–556, 2009.

[31] J. V. Beck and K. J. Arnold, *Parameter Estimation in Engineering and Science*, John Wiley & Sons, New York, NY, USA, 1977.

[32] A. van den Bos, *Parameter Estimation for Scientists and Engineers*, John Wiley & Sons, New York, NY, USA, 2007.

[33] I. Chou and E. O. Voit, "Recent developments in parameter estimation and structure identification of biochemical and genomic systems," *Mathematical Biosciences*, vol. 219, no. 2, pp. 57–83, 2009.

[34] P. Baldi and S. Brunak, *Bioinformatic: The Machine Learning Approach*, The MIT Press, Cambridge, Mass, USA, 2nd edition, 2001.

[35] A. M. Krieger and P. E. Green, "A generalized rand-index method for consensus clustering of separate partitions of the same data base," *Journal of Classification*, vol. 16, no. 1, pp. 63–89, 1999.

[36] W. Yan, H. Zhu, Y. Yang, J. Chen, Y. Zhang, and B. Shen, "Effects of time point measurement on the reconstruction of gene regulatory networks," *Molecules*, vol. 15, no. 8, pp. 5354–5368, 2010.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 475702, 5 pages
http://dx.doi.org/10.1155/2013/475702

Research Article

Adaptive $L_{1/2}$ Shooting Regularization Method for Survival Analysis Using Gene Expression Data

Xiao-Ying Liu,[1] Yong Liang,[1] Zong-Ben Xu,[2] Hai Zhang,[2] and Kwong-Sak Leung[3]

[1] *Faculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Macau 999078, China*
[2] *Faculty of Science, Xi'an Jiaotong University, Xi'an 710000, China*
[3] *Department of Computer Science and Technology, The Chinese University of Hong Kong, Hong Kong 999077, China*

Correspondence should be addressed to Yong Liang; yliang@must.edu.mo

Received 3 September 2013; Accepted 30 October 2013

Academic Editors: J. Ma, B. Shen, J. Wang, and J. Wang

Copyright © 2013 Xiao-Ying Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A new adaptive $L_{1/2}$ shooting regularization method for variable selection based on the Cox's proportional hazards mode being proposed. This adaptive $L_{1/2}$ shooting algorithm can be easily obtained by the optimization of a reweighed iterative series of L_1 penalties and a shooting strategy of $L_{1/2}$ penalty. Simulation results based on high dimensional artificial data show that the adaptive $L_{1/2}$ shooting regularization method can be more accurate for variable selection than Lasso and adaptive Lasso methods. The results from real gene expression dataset (DLBCL) also indicate that the $L_{1/2}$ regularization method performs competitively.

1. Introduction

In the study of the dependence of survival time T on covariances X, the Cox's proportional hazards model [1, 2] is the most widely used model in survival analysis. Suppose the dataset has a sample size of n to study survival time T on covariate X, we use the data form of $(t_1, \delta_1, X_1), \ldots, (t_n, \delta_n, X_n)$ to represent the individual's sample, where δ is the censoring indicator, the t_i denotes the survival time if $\delta_i = 1$ or otherwise censoring time.

By the Cox's proportional hazards model, the hazard function can be defined as

$$h(t \mid \beta) = h_0(t) \exp\left(\beta^T X\right), \tag{1}$$

where baseline hazard function $h_0(t)$ is unspecified or unknown and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the regression coefficient vector of p variables.

The Cox's partial log-likelihood is expressed as

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left\{ x_i^T \beta - \log\left(\sum_{j \in R_i} \exp\left(x_j^T \beta\right) \right) \right\}, \tag{2}$$

where $R_i = \{j \in 1, \ldots, n, t > t_i\}$ denotes ordered risk set at time t_i; t_i represents failure time.

In practice, not all the n covariates may contribute to the prediction of survival outcomes: some components of β may be zero in the true model. To select important variables under the proportional hazards model (2), Tibshirani [3], Fan and Li [4], and Zhang and Lu [5] proposed to minimize the penalized log partial likelihood function as

$$-\frac{1}{n} l(\beta) + \lambda \sum_{j=1}^{p} P\left(\beta_j\right). \tag{3}$$

The standard regularization algorithm cannot directly be applied for nonlinear Cox model to obtain parameter estimates. Therefore, Tibshirani [3] and Zhang and Lu [5] proposed iterative procedure to transform the Cox's partial log-likelihood function (2) to linear regression problem through an iteratively Newton-Raphon update. Here we follow the approach of Zhang and Lu [5]: define the gradient vector $\nabla l(\beta) = -\partial l(\beta)/\partial\beta$ and the Hessian matrix $\nabla^2 l(\beta) = -\partial l^2(\beta)/\partial\beta\partial\beta^T$, then apply the Choleski decomposition to obtain $X^T = \{\nabla^2 l(\beta)\}^{1/2}$, and generate the pseudoresponse vector $Y = (X^T)^{-1}\{\nabla^2 l(\beta)\beta - \nabla l(\beta)\}$. Then Zhang and Lu [5] suggested an optimization problem with the penalty function:

$$\widehat{\beta} = \arg\min \left\{ (Y - X\beta)^T (Y - X\beta) + \lambda \sum_{j=1}^{p} P\left(\beta_j\right) \right\}. \tag{4}$$

The Lasso penalty is $P(\beta_j) = |\beta_j|$, which shrinks small coefficients to zero and hence results in a sparse representation of the solution. However, estimation of large β's may suffer from substantial bias in λ if chosen too big and may not be sufficiently spare if λ is selected too small. Hence, Fan and Li [4] proposed the smoothly clipped absolute deviation (SCAD) penalty, which avoids excessive penalties on large coefficients and enjoys the oracle properties. The adaptive penalty is $P(\beta_j) = |\beta_j|/|\beta'_j|$, where the weights $1/|\beta'_j|$ are chosen adaptively by data. The values chosen for $1/|\beta'_j|$ are crucial for guaranteeing the optimality of the solution.

The above-mentioned series of Lasso methods were based on the L_1 penalty. Xu et al. [6, 7] and Liang et al. [8] have proposed $L_{1/2}$ regularization method which has the $L_{1/2}$ penalty $P(\beta_j) = |\beta_j|^{1/2}$. The theoretical analyses and experiments show that the $L_{1/2}$ regularization is more effective than Lasso both in theory and practice. In this paper, we investigate the adaptive $L_{1/2}$ shooting regularization to solve the Cox model.

The rest of the paper is organized as follows. Section 2 describes an adaptive $L_{1/2}$ shooting regularization algorithm to obtain estimates from the Cox model. Section 3 evaluates our method by simulation studies and application to real gene expression dataset (DLBCL). Finally we give a brief discussion.

2. Adaptive $L_{1/2}$ Shooting Regularization Method for the Cox Model

The log partial likelihood function of the Cox model with the $L_{1/2}$ penalty is

$$\beta_{1/2} = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2 + \lambda \sum_{i=1}^{p} |\beta_i|^{1/2} \right\}, \quad (5)$$

where λ is the tuning parameter.

In this section, we proposed the adaptive $L_{1/2}$ shooting algorithm to optimize the Cox model in an approximate linear form. The following is the complete algorithm procedure.

Step 1. Initial coefficients value $\beta^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_p^0) = (1, 1, \ldots, 1)$ and $t = 0$.

Step 2. Compute ∇l, $\nabla^2 l$, X, Y, and $\omega_j = 1/\sqrt{\beta_j^t}$ based on β_j^t $(1 \le j \le p)$, define RSS $= (Y - X\beta)^T (Y - X\beta)$, $S_j = \partial \text{RSS}/\partial \beta_j^t$ $(1 \le j \le p)$, and write β^t as $(\beta_j^t, (\beta_{-j}^t)^T)^T$, where β_{-j}^t is the $(p-1)$-dimensional vector consisting of all β^t's other than β_j^t, let $S_0 = S_j(0, \beta_{-j}^t)$ for each $j = 1, \ldots, p$.

Step 3. Solve $\beta^{t+1} = \arg\min\{(Y - X\beta)^T (Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^{1/2}\}$ $(1 \le j \le p)$, using the $L_{1/2}$ shooting regularization approach:

$$\beta_j^{t*} = \begin{cases} \dfrac{\lambda \cdot \omega_j - 2S_0}{4x_j^T x_j}, & \text{if } S_0 > \dfrac{1}{2}\lambda \cdot \omega_j, \\ \dfrac{-\lambda \cdot \omega_j - 2S_0}{4x_j^T x_j}, & \text{if } S_0 < \dfrac{1}{2}\lambda \cdot \omega_j, \\ 0, & \text{if } |S_0| \le \dfrac{1}{2}\lambda \cdot \omega_j. \end{cases} \quad (6)$$

Step 4. Solve $\beta^{t+1} = \arg\min\{(Y - X\beta)^T (Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| / \sqrt{\beta_j^t}\}$ $(1 \le j \le p)$, using the modified reweighed iterative approach of the L_1 shooting approach.

Step 4.1. Start with $\beta^{t,m} = (\beta_1^{t,m}, \beta_2^{t,m}, \ldots, \beta_p^{t,m}) = \beta^t$, set inner iteration count $m = 0$.

Step 4.2. At each iterative step m, for each $j = 1, \ldots, p$, update:

$$\beta_j^{t,m+1} = \begin{cases} \dfrac{\lambda \cdot \omega_j - S_0}{2x_j^T x_j}, & \text{if } S_0 > \lambda \cdot \omega_j, \\ \dfrac{-\lambda \cdot \omega_j - S_0}{2x_j^T x_j}, & \text{if } S_0 < \lambda \cdot \omega_j, \\ 0, & \text{if } |S_0| \le \lambda \cdot \omega_j, \end{cases} \quad (7)$$

where x_j is the jth column of X. A new estimator $\beta_j^{t,m}$ is formed after updating all β_j's and let $m = m + 1$.

Step 4.3. Update ω_j and S_0 and repeat Step 4.2 until $\beta^{t,m}$ converge.

Step 5. Let $t = t + 1$ and update $\beta_j^{t+1} = \min(\beta_j^{t,m}, \beta_j^{t*})$ and $j = 1, \ldots, p$ and repeat Steps 2, 3, and 4 until β^{t+1} does not change.

In Steps 2 and 4.3, we modify shooting algorithm with weight $1/\sqrt{|\beta_j^t|}$ based on last estimate β^t at each iteratively step. It is possible that some β^t become zero during the iterative procedure. So to guarantee the feasibly, we replace $1/\sqrt{|\beta_j^t|}$ with $1/\sqrt{|\beta_j^t + \varepsilon|}$ when implementing, where ε is any fixed positive real number. Steps 3 and 4 implement the shooting strategy of $L_{1/2}$ penalty and the reweighed iterative strategy of L_1 penalties, respectively. Step 5 selects the minimum of β^t, which is obtained by Steps 3 and 4, to improve the converge speed of the algorithm.

This algorithm gives exact zeros for some coefficients and it converges quickly based on our empirical experience. Similarly to Theorem 3 in Fu [9], we can show that the adaptive $L_{1/2}$ shooting regularization algorithm is guaranteed to converge to the global minimum of the log partial likelihood function of the Cox model (5).

3. Numerical Studies

3.1. Simulation Study for the High Dimensional Artificial Dataset. In this section, we compare the performance of the Lasso, the adaptive Lasso, and the adaptive $L_{1/2}$ shooting regularization method, under Cox's proportional hazards model. The cross-validated partial likelihood (CVPL) method is used to estimate the tuning parameter λ in these three algorithms. In our simulation studies, we use the Gempertz model suggested by Qian et al. [10] to generate the Cox model datasets in the setting:

$$\beta = \left(\frac{\overbrace{-0.7, -0.5, -0.3, -0.1, 0, 0, 0, 0, 0, 0, 0.4, 0, 0, 0.7}^{14}, \overbrace{0, \ldots, 0}^{986}}{1000} \right). \quad (8)$$

Table 1: The simulation results based on the high dimensional simulated dataset by the three methods over 100 replications. The columns include the average number of the selected variable (Var), the average number of the correct zeros (Corr), the average number of the incorrect zeros (Incorr), and the integrated Brier score (IBS). (Lasso: the Lasso method, A-L: the adaptive Lasso method, and $L_{1/2}$: the adaptive $L_{1/2}$ shooting regularization method).

n		25% censoring				40% censoring			
	Method	Var	Corr (994)	Incorr (0)	IBS	Var	Corr (994)	Incorr (0)	IBS
200	Lasso	81.29	917.29	0.26	0.1502	96.38	906.83	0.31	0.1516
	A-L	41.06	962.47	0.35	0.1474	59.05	948.89	0.43	0.1503
	$L_{1/2}$	17.79	984.28	0.42	0.1440	20.42	974.15	0.53	0.1498
250	Lasso	98.46	903.07	0.11	0.1462	148.87	883.85	0.15	0.1493
	A-L	64.10	949.46	0.17	0.1446	74.42	933.74	0.26	0.1478
	$L_{1/2}$	27.38	972.95	0.25	0.1421	31.91	968.03	0.34	0.1458
300	Lasso	167.82	883.18	0.01	0.1448	177.50	869.83	0.03	0.1479
	A-L	72.95	932.49	0.02	0.1436	80.97	927.42	0.06	0.1459
	$L_{1/2}$	33.45	967.12	0.03	0.1418	38.64	958.38	0.06	0.1427
350	Lasso	196.24	847.84	0.00	0.1441	204.22	834.53	0.00	0.1463
	A-L	82.80	928.07	0.00	0.1428	89.18	921.54	0.00	0.1441
	$L_{1/2}$	37.58	959.78	0.00	0.1405	40.15	948.63	0.00	0.1412

We considered the cases with 25% and 40% of censoring and used four samples, $n = 200$, 250, 300, and 350. The simulation results obtained by the three methods reported in Table 1. Since this simulation dataset has 6 relevant features (6 nonzero coefficients) in the 1000 ones, the idealized average numbers of variables selected (the Var column) and correct zeros (the Corr column) by each method are 6 and 994, respectively. From the Var and Corr columns of Table 1, the results obtained by the $L_{1/2}$ regularization method are obviously better than those of other methods for different sample sizes and censoring settings. For example, when $n = 200$ and the censoring is 25%, the average numbers (Var) from the Lasso, the adaptive Lasso, and the $L_{1/2}$ regularization methods are 81.29, 41.06, and 17.79 (best). The correct zeros' numbers (Corr) of the three methods are 917.29, 962.47, and 984.28 (best), respectively. The results obtained by the $L_{1/2}$ method are obviously close to the idealized values in the Var and Corr columns. Moreover, in the IBS (the integrated Brier score) column, the IBS's value of the Lasso, the adaptive Lasso, and the $L_{1/2}$ shooting regularization method are 0.1502, 0.1474, and 0.1440. This means that the $L_{1/2}$ shooting regularization method performs slight better than the other two methods for the prediction accuracy. Similar results are observed for the 40% censoring case.

As shown in the Incorr columns of Table 1, the idealized average number is 0 if the method can correctly identify all relevant variables at each run, whereas its maximal value is 6 if the method incorrectly identifies all the nonzero coefficients to zero in all runs. When the sample size is relative small ($n = 200$ and censoring rate = 25%), the average number of the incorrect zeros from the Lasso is 0.26, from the adaptive Lasso is 0.35 and from the $L_{1/2}$ regularization shooting method is 0.42. The adaptive $L_{1/2}$ shooting regularization method performs worse than the other two methods. When n increases to 350, all the three algorithms never evaluated the nonzero coefficients to zero. This means that the adaptive $L_{1/2}$ shooting regularization method shrinks the small effect covariates to zero more easily than the Lasso and the adaptive Lasso when the sample size is relative small. Similar results are observed for the 40% censoring case.

3.2. Experiments on the Real Gene Expression (DLBCL) Dataset. To further demonstrate the utility of the $L_{1/2}$ regularization shooting procedure in relating microarray gene expression data to censored survival phenotypes, we re-analyzed a published dataset of DLBCL by Rosenwald et al. [11]. This dataset contains a total of 240 patients with DLBCL, including 138 patient deaths during the followups with a median death time of 2.8 years. Rosenwald et al. [11] divided the 240 patients into a training set of 160 patients and a test set of 80 patients and built a multivariate Cox model. The variables in the Cox model included the average gene expression levels of smaller sets of genes in four different gene expression signatures together with the gene expression level of BMP6. It should be noted that in order to select the gene expression signatures, they performed a hierarchical clustering analysis for genes across all the samples (including both training and test samples). In order to compare our results with those in Rosenwald et al. [11], we used the same setting of training and test datasets in our analysis.

We applied the adaptive $L_{1/2}$ shooting regularization method to first build a predictive model using the training data of 160 patients and all the 7399 genes as features (predictors). Table 2 shows the GeneBank ID and a brief description of top ten genes selected by our proposed $L_{1/2}$ regularization method. It is interesting to note that eight of these genes belong to the gene expression signature groups defined in Rosenwald et al. [11]. These three signature groups include Germinal-center B-cell signature, MHC, and lymph-node signature. On the other hand, two genes selected by the $L_{1/2}$ method are not in the proliferation signature group defined by Rosenwald et al. [11].

Based on the estimated model with these genes, we estimated the risk scores using the method proposed by

Table 2: GeneBank ID and descriptions of the top 10 genes selected by the adaptive $L_{1/2}$ shooting regularization method based on the 160 patients in the training dataset. Indicated are the gene expression signature groups that these genes belong to; Germ: Germinal-center B-cell signature, MHC: MHC class II signature, and Lymph: lymph-node signature. Genes NM_005191 and X82240 do not belong to these signature groups.

GeneBank ID	Signature	Description
NM_005191		Homosapiens CD80 molecule (CD80), mRNA
AA714513	MHC	major histocompatibility complex, class II, DR beta 5
AA598653	Lymph	osteoblast specific factor 2 (fasciclin I-like)
AA767112	MHC	major histocompatibility complex, class II, DP beta 1
LC_24433	Lymph	
AA840067	Germ	TCL1A T-cell leukemia/lymphoma 1A
X82240		Homosapiens mRNA for T-cell leukemia
AA700997	Germ	cell associated 1
AA505045	Germ	Homosapiens, clone MGC:3963 IMAGE:3621362, mRNA, complete CDs
AA805575	Germ	Thyroxine-binding globulin precursor

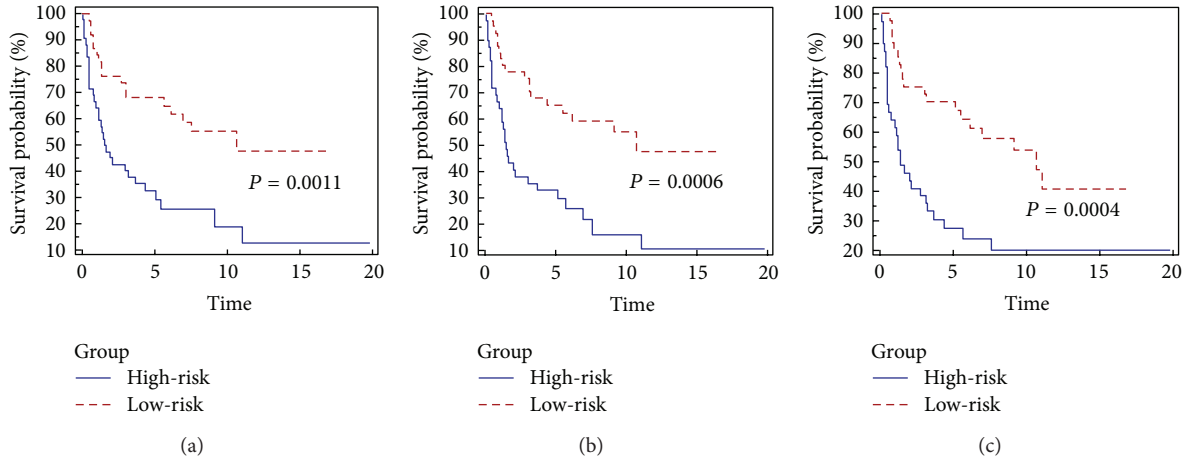


Figure 1: The Kaplan-Meier curves for the high- and low-risk groups defined by the estimated scores for the 80 patients in the test dataset. The scores are estimated based on the models estimated by the Lasso method (plot (a)), the adaptive Lasso method (plot (b)), and the $L_{1/2}$ regularization shooting method (plot (c)). The maximal follow-up time is 20 years.

Gui and Li [12]. To further examine whether clinically relevant groups can be identified by the model, we used zero as a cutoff point of the risk scores and divided the test patients into two groups based on whether they have positive or negative risk scores ($f(x) = \beta^T x$).

As a comparison, the Lasso, the adaptive Lasso, and the $L_{1/2}$ regularization methods are validated on the test dataset of 80 patients defined in Rosenwald et al. [11], and their corresponding Kaplan-Meier curves are shown in Figure 1. In Figure 1, the horizontal coordinate is the predictive survival time (years) and the vertical coordinate is the predictive survival probabilities. The P value (lower the better to indicate statistical significance) of the Lasso for the test dataset is 0.0011, which is significantly larger than 0.0006 and 0.0004 of the adaptive Lasso and the $L_{1/2}$ regularization methods. This means that lasso method performs the worst for the survival prediction compared with other two methods.

On the other hand, in order to assess how well the model predicts the outcome, we also use the idea of the integrated Brier score (IBS) for the test dataset including censored observations as our criteria. In Table 3, the IBS's value of the Lasso, the adaptive Lasso, and the adaptive $L_{1/2}$ shooting regularization method are 0.2306, 0.2026, and 0.2017. We can see that the adaptive Lasso and the adaptive $L_{1/2}$ shooting regularization methods perform slight better than Lasso for the prediction accuracy.

Table 3: The integrated Brier score (IBS) obtained by the Lasso, the adaptive Lasso and the adaptive $L_{1/2}$ shooting regularization method for DLBCL dataset. (Lasso: the Lasso method; A-L: the adaptive Lasso method; $L_{1/2}$: the adaptive $L_{1/2}$ shooting regularization method).

	Lasso	A-L	$L_{1/2}$
IBS	0.2306	0.2026	0.2017

4. Discussion and Conclusion

In this paper, we have presented the novel adaptive $L_{1/2}$ shooting regularization method, which is used for variable

selection in the Cox's proportional hazards model. Its performance is validated by both simulation and real case studies. In the experiments, we use the high-dimensional and low-sample size dataset, with applications to microarray gene expression data (DLBCL). Results indicate that our proposed adaptive $L_{1/2}$ shooting regularization algorithm is very competitive in analyzing high dimensional survival data in terms of sparsity of the final prediction model and predictability. The proposed $L_{1/2}$ regularization procedure is very promising and useful in building a parsimonious predictive model used for classifying future patients into clinically relevant high-risk and low-risk groups based on the gene expression profile and survival times of previous patients. The procedure can also be applied to select important genes which are related to patient's survival outcome.

Acknowledgments

This work was supported by the Macau Science and Technology Develop Funds (Grant no. 017/2010/A2) of Macau SAR of China and the National Natural Science Foundation of China (Grant nos. 11131006, 11171272).

References

[1] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society B*, vol. 34, no. 2, pp. 187–220, 1972.

[2] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[3] R. Tibshirani, "The lasso method for variable selection in the Cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[4] J. Fan and R. Li, "Variable selection for Cox's proportional hazards model and frailty model," *Annals of Statistics*, vol. 30, no. 1, pp. 74–99, 2002.

[5] H. H. Zhang and W. B. Lu, "Adaptive Lasso for Cox's proportional hazards model," *Biometrika*, vol. 94, no. 3, pp. 691–703, 2007.

[6] Z. B. Xu, H. Zhang, Y. Wang, X. Y. Chang, and Y. Liang, "$L_{1/2}$ regularization," *Science in China F*, vol. 53, no. 6, pp. 1159–1169, 2010.

[7] Z. B. Xu, X. Y. Chang, F. M. Xu, and H. Zhang, "$L_{1/2}$ regularization: a thresholding representation theory and a fast solver," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 7, pp. 1013–1027, 2012.

[8] Y. Liang, C. Liu, X. Z. Luan et al., "Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification2013," *BMC Bioinformatics*, vol. 14, p. 198.

[9] W. J. Fu, "Penalized regressions: the bridge versus the lasso," *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416, 1998.

[10] J. Qian, B. Li, and P. Chen, "Generating survival data in the simulation studies of Cox model," in *Proceedings of the 3rd International Conference on Information and Computing (ICIC '10)*, pp. 93–96, June 2010.

[11] A. Rosenwald, G. Wright, W. Chan et al., "The use of molecular profiling to predict survival afterchemotherapy for diffuse large B-cell lymphoma," *The New England Journal of Medicine*, vol. 346, pp. 1937–1946, 2002.

[12] J. Gui and H. Li, "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data," *Bioinformatics*, vol. 21, no. 13, pp. 3001–3008, 2005.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 580456, 10 pages
http://dx.doi.org/10.1155/2013/580456

Research Article

Molecular Dynamic Simulation to Explore the Molecular Basis of Btk-PH Domain Interaction with Ins(1,3,4,5)P4

Dan Lu,[1] Junfeng Jiang,[1] Zhongjie Liang,[1,2] Maomin Sun,[1] Cheng Luo,[1,2] Bairong Shen,[1,3] and Guang Hu[1]

[1] *Center for Systems Biology, Soochow University, Suzhou 215006, China*
[2] *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*
[3] *Department of Bioinformatics, Medical College, Soochow University, Suzhou 215123, China*

Correspondence should be addressed to Guang Hu; huguang@suda.edu.cn

Received 28 July 2013; Accepted 7 September 2013

Academic Editors: P. Minkiewicz, Y. Muto, and W. A. Thompson

Copyright © 2013 Dan Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Bruton's tyrosine kinase contains a pleckstrin homology domain, and it specifically binds inositol 1,3,4,5-tetrakisphosphate (Ins(1,3,4,5)P4), which is involved in the maturation of B cells. In this paper, we studied 12 systems including the wild type and 11 mutants, K12R, S14F, K19E, R28C/H, E41K, L11P, F25S, Y40N, and K12R-R28C/H, to investigate any change in the ligand binding site of each mutant. Molecular dynamics simulations combined with the method of molecular mechanics/Poisson-Boltzmann solvent-accessible surface area have been applied to the twelve systems, and reasonable mutant structures and their binding free energies have been obtained as criteria in the final classification. As a result, five structures, K12R, K19E, R28C/H, and E41K mutants, were classified as "functional mutations," whereas L11P, S14F, F25S, and Y40N were grouped into "folding mutations." This rigorous study of the binding affinity of each of the mutants and their classification provides some new insights into the biological function of the Btk-PH domain and related mutation-causing diseases.

1. Introduction

Bruton's tyrosine kinase (Btk) is a member of the Tec family of kinases and the only known one associated with human disease [1, 2]. Previous studies have indicated the significance of Btk in B-cell development, differentiation, and signaling [3, 4]. Once the Btk-dependent signal transduction pathway is inactivated, B cells remain at the pre-B-cells stage leading to X-linked agammaglobulinemia (XLA) in humans, which is one of the most frequently inherited immunodeficient disorders in human, and X-linked immunodeficiency (Xid) in mice [5–8].

The Btk protein contains Src-homology 2 and 3 domains (SH2 and SH3), a catalytic SH1 domain, a Tec-homology (TH) domain, and an N-terminal pleckstrin homology (PH) domain [9–11]. Studies have shown that XLA mutations in Btk can be mapped to all five domains of the kinase, which are critical for signal transmission. Several missense mutations in the PH domain have been widely studied and so far considered to be the only known cause of the disease [12]. The PH domain is responsible for binding with phosphatidylinositols, showing the importance for the regulation of membrane targeting. Thus, a mutation in the PH domain can influence the binding affinity with a ligand, membrane targeting, and the activation of Btk [13, 14]. Of the various phosphatidylinositols, the Btk-PH domain has higher specificity and binding affinity with inositol 1,3,4,5-tetrakisphosphate (Ins(1,3,4,5)P4) [15–17]. The crystal structure of a complex of the Btk-PH domain with Ins(1,3,4,5)P4 (PDB ID code: 2Z0P [18]) shows that the Btk-PH domain recognizes Ins(1,3,4,5)P4 in a canonical manner [19, 20].

The PH domain is a structural protein domain containing approximately 120 amino acid residues that retains a highly conserved three-dimensional organization of different

proteins, despite their poorly conserved primary sequences [21–23]. The core structure is a β-sandwich of two almost orthogonal β-sheets consisting of three and four strands, respectively. The opposite edge of the structure is capped by an amphipathic C-terminal α-helix. Six loops connect the β-strands, while the amino acids in the β1-β2 and β3-β4 loops form a hydrogen bond network that interacts with Ins(1,3,4,5)P4 [19].

According to previous studies, XLA-causing mutations can be mainly classified into two groups [24, 25]. The first group of mutations appears to prevent the formation of stable native-like structures, which are identified as "folding mutations." In the second group named as "functional mutations," the mutations do not affect the overall fold, though they may disrupt the ligand binding affinity leading to functional defects. Specially, K12 and R28 are significant in most functional mutations involved in the interaction with a ligand because a functionally based covariant pair is created, which contacts with a bound negatively charged ligand [26]. Generally, experimental studies have indicated the biological influence of amino acid mutation [24, 25, 27], but the proposed explanations do not exactly reveal much detailed molecular basis of the effects of amino acid mutations, such as a change in hydrogen bonds, binding sites, and binding free energy.

In this paper, we have studied Btk-PH domain binding with the Ins(1,3,4,5)P4 structure in wild type and 11 well-known mutants [24, 25, 27], L11P, K12R, S14F, K19E, F25S, R28C/H, Y40N, E41K, and K12R-R28C/H. The method of molecular mechanics/Poisson-Boltzmann solvent-accessible surface area (MM/PBSA) was used to compare the binding affinity [28–32]. Three main aims were mainly raised and discussed in this study: (1) the classification of the eleven mutants, (2) the effects of amino acid mutations on the biological function of the Btk-PH domain, and (3) the possible coordinate interactions of amino acid residues. Our simulations indicated that mutations K12R, K19E, R28C/H, E41K, and K12R-R28C/H were "functional mutations," because negative binding free energies were identified in the MM/PBSA calculation for these structures, whereas the others were considered to be "folding mutations" whose binding free energies are positive. The ligand clearly moves away from the binding pocket in R28C/H mutation structures. In K12R mutation structures, the ligand sites in the top of the binding Pocket, which results in the evident decrease of binding free energy. Thus, the biological function of Btk is weakened in all three mutant structures. The K19E mutation leads to a major change in the β1-β2 loop, leading to weak membrane targeting and the suppression of Btk activation. The E41K mutation increases the positive charge of the β3-strand, which increases the positively charged surface. This may provide a cavity for binding with another ligand, which could enhance the membrane targeting and activation of Btk. Remarkably, the two double amino acids mutant structures K12R-R28C/H retain a similar ligand binding affinity to the wild type structure, because new coordinate interactions by the amino acid pairs R12 and K53 are found to play important roles in the binding with Ins(1,3,4,5)P4. We hope that our findings help to explain the relationship between residue mutations and biological function, as well as the molecular basis of related diseases.

2. Materials and Methods

2.1. Btk PH-Ins(1,3,4,5)P4 and Mutation Structures. The X-ray structure of the PH domain of Btk (PDB ID code: 2Z0P (A chain); resolution: 2.58 Å) [18], which is bound to Ins(1,3,4,5)P4, was used as the initial structure for all simulations. The PH domain has the common PH domain fold, which has a seven-stranded antiparallel β-sheet grouped into a β-sandwich and a α1 helix. For far distance away from the binding pocket, the Zn^{2+} ion in the structure was deleted for simplification and the forces between the Zn^{2+} ion and original four amino acids, His143 and Cys154/155/165, were manually added to reduce the flexibility of protein structure. Mutant structures (L11P, K12R, S14F, K19E, F25S, R28C/H, Y40N, E41K, and K12R-R28C/H) were obtained and visualized using the Sybyl software package (Tripos, St. Louis, MO). Ligand of Ins(1,3,4,5)P4 was extracted from the PH domain complex as an independent small molecule for the molecular dynamics (MD) and MM/PBSA simulation. All initial mutation structures were minimized by Sybyl using a distance-dependent dielectric function, with a nonbonded cut-off of 8 Å. Amber charges were assigned to the protein while Gasteiger-Hückel charges were given to Ins(1,3,4,5)P4. The whole system was minimized until no more atom collisions were found.

2.2. MD Simulation. Before the MD simulation, the charge information of Ins(1,3,4,5)P4 was calculated using the RESP method [33] encoded in the AMBER suite program (version 9) [34], followed by Gaussian 03 for calculations at the Hartree-Fock (HF)/6-31G* level. After the application of the AMBER03 force field to proteins [35] and the general AMBER force field (gaff) to ligands [36], each complex was loaded into the AMBER9 program, adding Na^+ ions with a 1 Å grid to neutralize the system. Finally, an 8 Å water TIP3PBOX was loaded into the system to form a complex environment. Subsequently, energy minimization was performed to remove any inappropriate contacts.

MD simulations were conducted with a nonbonded cut-off of 8 Å, the dielectric constant of 1.0, integration step of a 2 fs time interval, and Particle Mesh Ewald (PEM) for long-range electrostatic interaction calculation [37], using the following protocol: (1) 500 ps of heating equilibration with weak restraints on each complex in Cartesian space using a harmonic potential, where the temperature was gradually increased to 300 K, with constant volume, and Langevin dynamics for temperature control; (2) 500 ps of density equilibration, 300 K thermal bath at 1.0 atm of pressure (atm = 101.3 kPa) periodic boundary conditions with isotropic position scaling, and Langevin dynamics for temperature control; (3) 1 ns of constant pressure equilibration with weak restraint on the ligand and 20 ns of constant pressure equilibration at 300 K.

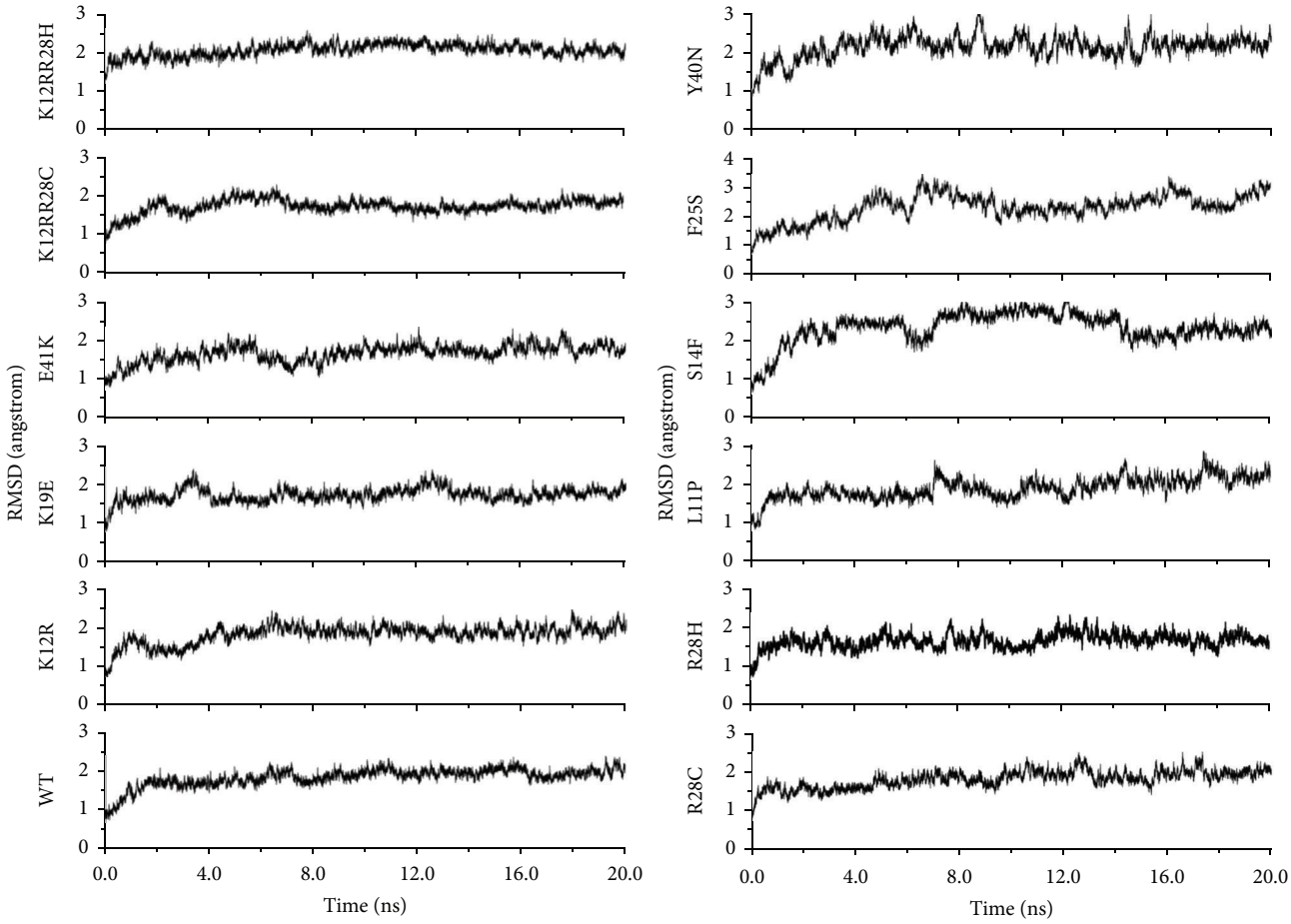


FIGURE 1: Time dependence of the root-mean-square deviations (RMSDs) for the C^{α} atoms from their initial structure of 20 ns MD simulations of all complex structures.

2.3. MM-PBSA Calculations for Binding Free Energy. Binding free energy (ΔG_{bind}) of ligand on Btk-PH domain was calculated by single trajectory method of MM/PBSA, according to the following equation:

$$\Delta G_{\text{bind}} = G^{\text{complex}} - G^{\text{PH}} - G^{\text{li}}, \quad (1)$$

where the free energies of complex, PH domain, and ligand are denoted as G^{complex}, G^{PH}, and G^{li}, respectively. Free energy (G) was calculated according to following equations:

$$G = E_{\text{gas}} + G_{\text{sol}} - TS_{\text{config}},$$
$$E_{\text{gas}} = E_{\text{vdw}} + E_{\text{ele}} + E_{\text{int}}, \quad (2)$$
$$G_{\text{sol}} = G_{\text{elec}} + G_{\text{SA}},$$

where E_{gas} defines the molecular mechanical energy, G_{sol} defines a solvation free energy, which is further decomposed to polar (G_{elec}) and nonpolar (G_{SA}) terms, and S_{config} defined as configurational entropy, which is normally derived from normal model analysis. E_{gas} was the sum of contributions from internal energies including bond, angle, and torsion angle energies (E_{int}), electrostatic energy (E_{ele}), and van der Waals energy (E_{vdw}), which were calculated using the same force field as that of MD simulations with no cutoff. Notably, the E_{int} was assigned to zero in (1), since the torsion angle energies of the complex and the separated part were calculated from the same trajectory. In the MM/PBSA calculation, 2000 snapshots of each model were extracted from the last 10 ns of the trajectories at time intervals of 5 ps.

3. Results and Discussions

MD simulations were used to investigate the complexes of 11 mutants and wild type PH domain and the interaction behavior between the PH domain and the ligand Ins(1,3,4,5)P4. Details of the simulation setup are provided in Materials and Methods. The root-mean-square deviation (RMSD) values of the C^{α} atoms relative to the initial coordinates were calculated using the 20 ns trajectory data. As shown in Figure 1, the RMSD tended to be flat after 5 ns' simulation in WT, K12R, K19E, E41K, K12R-R28C, K12R-R28H, R28C, and R28H at about 2 Å, indicating stable conformations. As for the RMSD values for other four mutants, relative large fluctuations were observed during the 20 ns' simulation, suggesting unstable structures and maybe poor binding affinities between

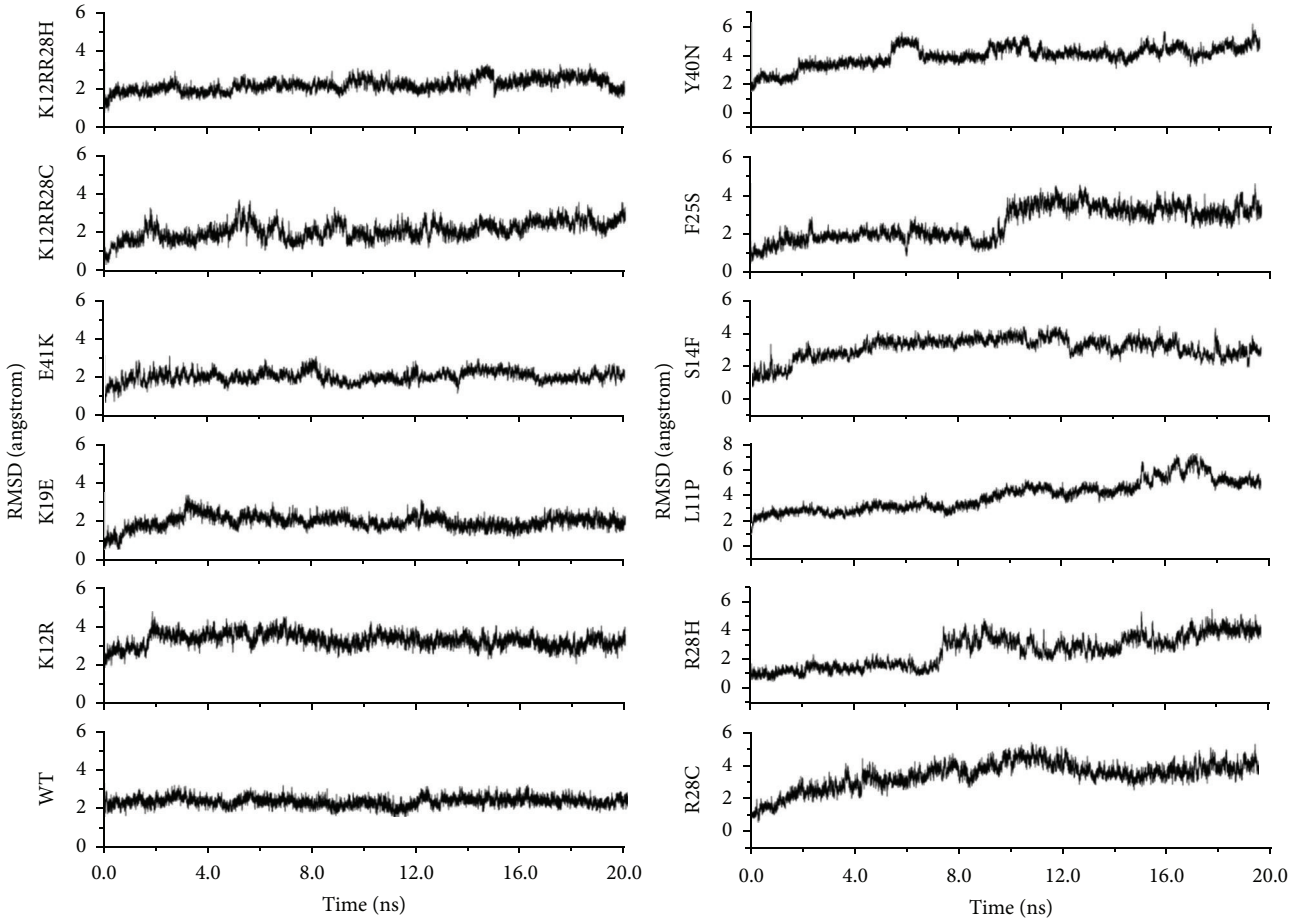


Figure 2: Time dependence of the root-mean-square deviations (RMSDs) for the ligand from its initial structure of 20 ns MD simulations of all complex structures.

the protein and the ligand. Meanwhile, the RMSD value of the ligand Ins(1,3,4,5)P4 was also calculated to further indicate the stability of the structures. In Figure 2, it was obvious that the ligand of WT, K12R, K19E, E41K, K12R-R28C, and K12R-R28H was more stable than that of R28C, R28H, L11P, S14F, F25S, and Y40N. Therefore, it was believed that R28C, R28H, L11P, S14F, F25S, and Y40N mutants might share some similar instabilities, which was in good accordance with the predefined classifications based on experimental results [27].

In order to offer a better criterion for classification of the mutants, the MM/PBSA method was used to estimate the binding affinities of the PH domain and the ligand in the 12 structures. Energy results for all the MM/PBSA calculations shown in Table 1 indicated that these mutations could be clearly divided into two groups. It was convinced that the "folding mutations" could not fold into a stable native-like structure to perform the function of the PH domain [27]. In our results, it is obvious that the four mutations, L11P, S14F, F25S, and Y40N, own a positive binding free energy, though the enthalpy change is favorable, implying the formation of complex is not allowed by thermodynamics. As aforementioned, these four mutations also had unstable RMSD value. Therefore, it is most likely that the four "folding mutations" will result in the loss of function of the PH domain completely, although S14F was considered to be a "functional mutation" in an earlier study [24, 25]. On the contrary, the other 7 mutations have negative binding free energies. With the stable RMSD values, it is proposed that the mutations of K12R, K19E, R28C/H, E41K, and K12R-R28C/H may be considered as "functional mutations," which do not affect the overall fold but can be expected to disrupt the binding affinities between the protein and the ligand [24, 27]. In this paper, the binding free energies of seven mutation structures were carefully compared to the wild type, using ΔG_{bind} as the criterion. Different mutational simulations provided important information about the affinity drift of the ligand-binding protein complex, which were further discussed below.

3.1. Structural and Binding Affinity Analysis of "Functional Mutations". According to the MM/PBSA simulations, the binding affinity of the 7 "functional mutations" was ranked to offer a better understanding of risks in disease, which was displayed in Table 1.

As aforementioned, the RMSD values of the ligand in R28C and R28H mutants were greater than 4, which was two times larger than that of wild type. Therefore, it is most

Table 1: Ranked calculated binding free energies (kcal/mol) using MM/PBSA for the twelve protein-ligand systems.

	ELE	VDW	PB	SA	PBTOT	$T\Delta S$	ΔG_{bind}
WT	−3031.77 (9.01[a])	−0.16 (0.07)	2961.60 (7.63)	−4.28 (0.47)	−74.61 (5.10)	−45.29 (1.43)	−29.32 (1.49)
K12R-R28C	−3119.66 (10.33)	−0.12 (0.05)	3044.04 (8.61)	−4.25 (0.18)	−79.99 (3.37)	−52.50 (1.45)	−27.49 (1.06)
K19E	−3143.08 (8.92)	−0.17 (0.03)	3069.08 (7.38)	−4.21 (0.16)	−78.38 (5.33)	−51.09 (1.60)	−27.29 (0.73)
K12R-R28H	−3029.46 (8.74)	−0.05 (0.02)	2962.73 (6.43)	−4.12 (0.21)	−73.30 (2.02)	−51.66 (1.37)	−21.64 (0.66)
E41K	−2855.47 (7.73)	−0.28 (0.16)	2789.39 (5.95)	−4.16 (0.23)	−70.52 (3.12)	−50.08 (1.52)	−20.44 (0.38)
K12R	−2922.34 (9.17)	−0.08 (0.03)	2859.80 (8.38)	−4.30 (0.21)	−66.92 (3.62)	−52.84 (1.57)	−14.08 (0.94)
R28C	−2612.17 (5.95)	−1.90 (0.23)	2552.82 (5.03)	−3.69 (0.16)	−64.94 (1.87)	−57.28 (2.47)	−7.66 (0.03)
R28H	−2591.08 (6.16)	−1.61 (0.17)	2533.38 (4.65)	−3.70 (0.17)	−63.01 (1.16)	−56.80 (2.54)	−6.21 (0.02)
S14F	−2573.79 (7.06)	−3.37 (0.74)	2545.38 (6.99)	−3.42 (0.25)	−35.20 (2.47)	−57.49 (3.02)	22.29 (0.98)
L11P	−2599.24 (7.61)	−1.75 (0.41)	2578.10 (6.70)	−3.94 (0.33)	−26.83 (1.42)	−58.75 (4.47)	31.92 (2.02)
Y40N	−2550.35 (7.0)	−2.57 (0.59)	2531.48 (6.23)	−3.92 (0.27)	−25.36 (1.84)	−58.34 (3.89)	32.98 (1.68)
F25S	−2501.74 (7.17)	−2.59 (0.69)	2487.39 (6.64)	−3.77 (0.19)	−20.71 (1.54)	−58.32 (4.15)	37.61 (1.82)

[a]The statistical error was estimated on the basis of the deviation between block averages.
ELE: electrostatic energy; VDW: van der Waals energy; PB: the PB solvation energy; SA: the surface area energy (nonpolar solvation energy); PBTOT: sum of ELE, VDW, PB, and SA.
The absolute temperature (T) was set to 300 K in the MM-PBSA calculations. ΔG_{bind}: subtracting $T\Delta S$ from PBTOT.

likely that the ligand completely moves away from the original binding pocket of the PH domain and this was consistent with the result of the binding free energy calculation for R28C/H mutants (Table 1). It is clear that the neutral charged cysteine has a sulfur atom in its side chain, which imposes a larger steric restriction compared with arginine. Although it has the same positive charge as arginine, histidine has a five-member ring that produces more steric hindrance than the side chain of arginine. Structural analysis in Figure 3 verified that the ligand was already out of the binding pocket. The mutations removed the hydrogen bond interaction between R28 and the ligand while forming new hydrogen bond interactions with residues on the β4 strand (K53) and β1-β2 loop (K17), respectively, which would not provide suitable side chains to bind with the ligand. Accordingly, it was considered that the R28C/H mutations had greatly reduced the binding affinity with the ligand. And the biological function of the two mutants was severely weakened, which may be considered as a primary factor that leads to associated diseases such as XLA and Xid [5].

As shown in Table 1, according to the MM/PBSA simulations, the free binding free energies of the mutants K19E, E41K, and K12R-R28C/H were similar to the wild type, whereas the K12R mutant was relatively smaller (−14.08 kcal/mol). Furthermore, compared with the wild type, the ligand was still located in the binding pocket of these mutants, as shown in Figures 3(d)–3(h), although there were different ligand-conformational changes due to hydrogen bond rupture. The time evolution of the root-mean-square fluctuation (RMSF) from the initial structure was also analyzed to further evaluate the ligand binding behavior of these mutations.

Structural analysis of the wild type showed that K12, which was located at the bottom of the binding pocket, formed hydrogen bonds with the 3- and 4-phosphates of the ligand, which is critical for the twist of β1-β2 loops [24]. In the K12R mutant, there was obvious decrease in binding free energy in the MM/PBSA simulation (Table 1), which could be explained by the difference in the conformation and the change in the hydrogen bond interactions of the ligand. Sharing the same positive change as lysine, the three atoms in arginine tail formed a conjugated system, which decreased the flexibility and increased the steric hindrance of this residue, leading to the conformational change required for ligand binding. The hydrogen bond with the 3-phosphates of the ligand was also fully ruptured in K12R mutant, resulting in a loss of specificity in the Btk-PH domain for the Ins(1,3,4,5)P4 ligand. As shown in Table 2, the K12R mutant abolished the hydrogen bond network formed by K12, R28, and the ligand; thereby the contribution of the conserved covariation pair on the biological function of the Btk-PH domain was removed.

In the wild type, as shown in Figure 3(a), K19 contacts D148 via a hydrogen bond, which moves the β1-β2 loop closer to the tail of the structure and moves it away from the

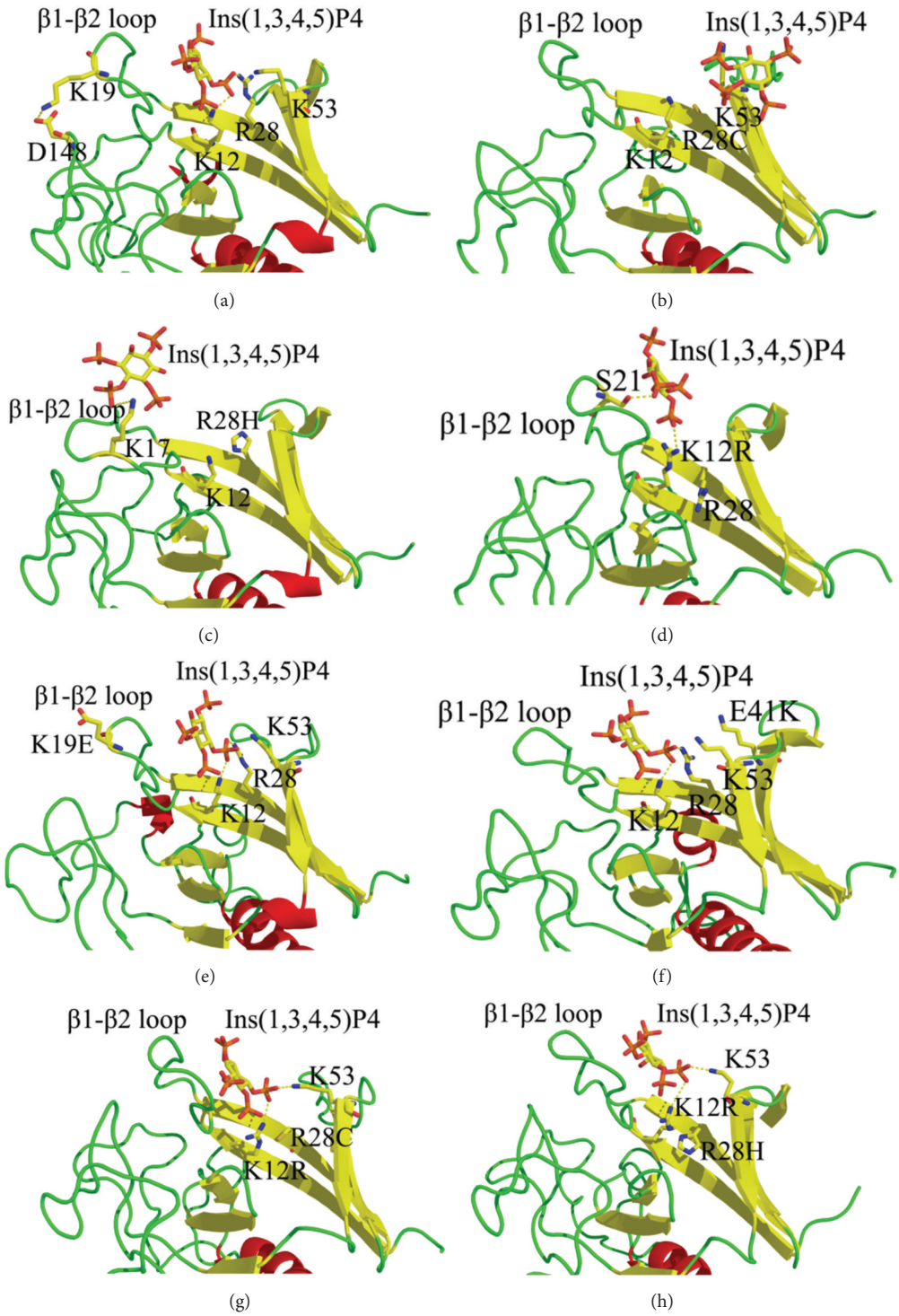


Figure 3: The snapshot structures taken at time 20 ns of wild type and all "functional mutations" complexes from MD simulation. Ligands and surrounding important residues are labeled and shown in stick, while hydrogen bonds are displayed in yellow dashed lines. (a) Wild type complex; (b) R28C mutation complex; (c) R28H mutation complex; (d) K12R mutation complex; (e) K19E mutation complex; (f) E41K mutation complex; (g) K12R-R28C mutation complex; (h) K12R-R28H mutation complex.

Table 2: Comparison of hydrogen bond interactions of five "functional mutations" and wild-type (WT) structures. Only hydrogen bonds with occupation ratio of the trajectory greater than 30%, atomic distance of two heavy atoms smaller than 3 Å, and atomic angle smaller than 60° are presented. The rank of amino acid according to its occupation ratio of the trajectory is from large to small.

WT	E41k	K19E	K12R	K12R-R28H	K12R-R28C
Arg28 3HB with 3-PG[a]	Arg28 3HB with 3-PG	Arg28 3HB with 3-PG	Arg12 2HB with 4-PG	Arg12 3HB with 3-PG	Arg12 3HB with 3-PG
Lys12 3HB with 3-PG	Lys12 3HB with 3-PG	Lys12 3HB with 3-PG	Ser 21 1HB with 5-PG	Ser 14 1HB with 4-PG	Lys53 2HB with 3-GP
Ser14 1HB with 4-PG	Ser 14 2HB with 4-PG	Ser 14 1HB with 4-PG	Lys17 2HB with 5-PG	Lys53 2HB with 3-PG	Ser 14 1HB with 4-PG
Lys26 4HB with 1-PG	Ser 21 2HB with 5-PG	Thr 20 1HB with 5-PG	Ser 14 1HB with 4-PG	Gln15 3HB with 4-PG	Gln15 3HB with 4-GP
Ser 21 2HB with 5-PG	Gln15 1HB with 4-PG	Lys17 2HB with 4-PG	Gln15 1HB with 4-PG	Lys17 2HB with 4-PG	Lys17 2HB with 4-PG
1HB with 6-OH$^{\beta}$	Lys26 2HB with 1-PG	Gln15 1HB with 4-PG	Lys26 1HB with 1-PG	Ser 21 1HB with 5-PG	Ser 21 2HB with 5-PG
Gln15 1HB with 4-PG	Lys53 2HB with 3-GP	Ser 21 2HB with 5-PG		1HB with 6-OH	1HB with 6-OH
Lys17 2HB with 4-PG		1HB with 6-OH		Lys26 1HB with 1-PG	Lys26 1HB with 1-PG
Lys53 2HB with 3-GP		Lys26 2HB with 1-PG			
		Lys53 2HB with 3-PG			

[a]HB and PG represent hydrogen bond and phosphate group, respectively.
$^{\beta}$OH represents OH group in the inositol ring.

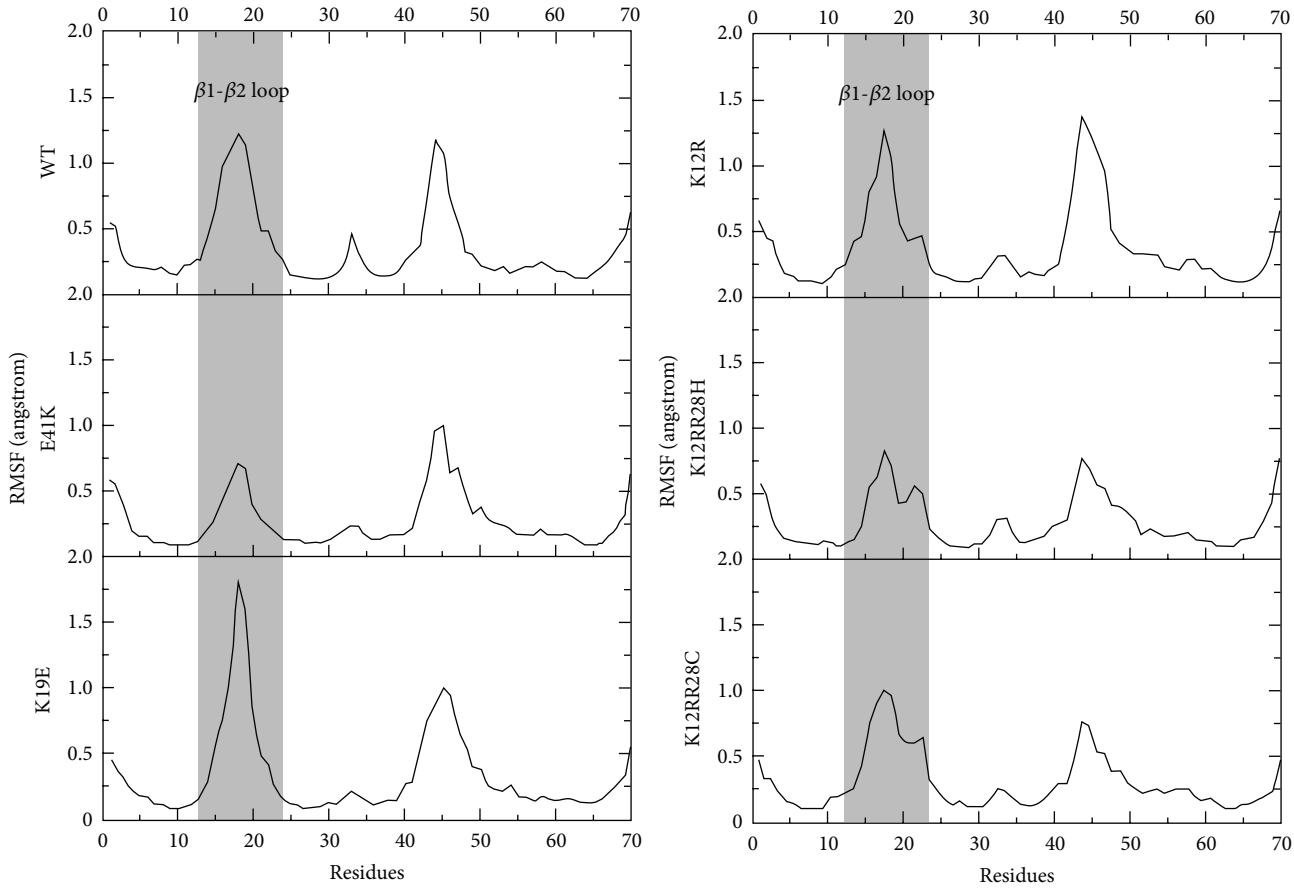


Figure 4: Residue fluctuations obtained by average residual fluctuations between residues 1 and 70 for six simulations, wild type (WT), K19E, E41K, K12R, and K12R-R28C/H, over the 20 ns simulations are illustrated in solid lines. The residues of β1-β2 loop are highlighted.

ligand. In the mutant K19E, a stable hydrogen bond network is formed among S14, T20, and the ligand after the rupture of the hydrogen bond between E19 and D148 (Table 2), which leads to a higher binding free energy with the ligand than the wild type while leading to an almost 90° change in the direction of the β1-β2 loop, as shown in Figure 3(e). The RMSF analysis in Figure 4 also confirmed that the β1-β2 loop had much greater fluctuations than the wild type. Therefore, it is possible the significant change in the β1-β2 loop of this mutation could eliminate the positively charged surface that interacts with the negatively charged membrane surface, resulting in poor membrane targeting and incomplete activation of Btk. This may mean that the biological function of Btk is not fully available. Therefore, K19E was also a loss-of-function

mutation, although it had a similar binding affinity with the ligand than the wild type, which was in good agreement with the experimental results [24, 27].

Compared with K19E, the mutant of E41K has the opposite effect on the biological function of Btk. As shown in Figure 3(a), E41 was located in the β3 strand with some distance from the ligand, but it is known to have great importance in the contact with the membrane. Despite the higher steric hindrance after the substitution of glutamate with lysine, the ligand slightly shifts to the β1-β2 loop in Figure 3(f), which results in a reduced flexibility of the β1-β2 loop (RMSF in Figure 4). The binding free energy of the E41K mutant was also smaller than the wild type, as shown in Table 1. Thus, E41K may retain most of the biological functionality of the wild type. However, the mutation in E41K increased the positively charged complex surface of the region located above the β3 and β4 strand, which was favorable for interaction with the negatively charged membrane surface, while it also offered a second ligand binding site. Therefore, it is likely that the E41K mutant may increase the activation of Btk, which is known as a gain of function [24, 27].

3.2. Covariant Pair Identification in K12R-R28C/H Mutations. Of the disease-causing mutations, the mutations in K12 and R28 always coexist with other mutations, although with lower frequencies, which may be due to the diversity of the PH domain. Single mutations of K12 and R28 have been known to weaken the biological function of the Btk-PH domain to different degrees, while it is also reported that K12 and R28 can form a conserved covariant pair that maintains the binding affinity of Ins(1,3,4,5)P4, which plays an important role in the biological function of the Btk-PH domain [26]. Therefore, the double residue mutations of K12R-R28C/H were closely analyzed to underly the molecular basis. Compared with the wild type structure, the ligand was still located in the bottom of the binding pocket of the two mutations (Figures 3(g) and 3(h)). The MM/PBSA simulations indicated a similar free binding energy for K12R-R28H and K12R-R28C (Table 1). This could be explained by the formation of a new covariant pair between R12 and K53 as K12-R28. Unlike the K12R mutation, the hydrogen bond network of 3-phosphates was kept through R12 and K53 in the mutation of K12R-R28C/H, which ensured the preservation of ligand specificity, whereas all hydrogen bond interactions between H28/C28 and ligand were ruptured (Table 2). Interestingly, a new conserved hydrogen bond was discovered between K53 and the ligand during the MD simulations. Instead of the covariant pair K12-R28, we believe that R12-K53 formed a similar odd positively charged pair that made contact with a bound negatively charged inositol phosphate and thereby can maintain the biological function of the K12R-R28C/H mutants.

4. Conclusions

Twelve complex systems, including eleven mutants and one wild type PH domain, have been comprehensively studied by integrating MD and MM/PBSA simulations, to reveal the detailed molecular basis. As a result, L11P, S14F, F25S, and Y40N were considered to be "folding mutations," whereas mutants K12R, K19E, R28C/H, E41K, and K12R-R28C/H were classified into "functional mutations." The effects of "functional mutations" on the biological function of the Btk-PH domain have also been predicted. In the R28C/H mutation, the ligands completely left the binding pocket and formed new interactions with the β1-β2 loop and the β4 strand, respectively. For the K12R mutation, the biological function of the Btk-PH domain may be reduced by the lack of hydrogen bond interaction with 3-phosphates when the ligand is formed, which determines the specificity of the Btk-PH domain. As for the K19E mutation, the change in the residue charge and the trend of the β1-β2 loop eliminated the positively charged surface that interacted with the negatively charged membrane surface, resulting in poor membrane targeting and incomplete activation of Btk. In contrast, E41K increased the positive surface of the Btk-PH domain, thereby increasing the possibility of the binding with a second ligand, which may result in a gain of function. Of note, the K12R-R28C/H mutant retained a high binding affinity without the displacement of the ligand from the binding pocket. A new odd positively charged pair was formed between R12 and K53 in the K12R-R28C/H mutation, which was similar to the K12-R28 in the wild type. This formation may interact with a bound negatively charged inositol phosphate, which may preserve the biological function of the Btk-PH domain and this should be further investigated experimentally.

In summary, our MD and MM/PBSA simulations of the twelve structures may provide some new insights into PH domain and their associated diseases. Based on the binding free energy, binding site of the ligand, and hydrogen bond interactions, we have predicted the effects of mutations on the biological function of the Btk-PH domain.

Authors' Contribution

Dan Lu and Junfeng Jiang contributed equally to this work.

Acknowledgments

The authors gratefully acknowledge financial support from the National Natural Science Foundation of China Grants (21203131, 31170795, 20872107, 20972174, and 91029704), the 863 Program (SS2012AA021103), and the Chinese Academy of Sciences (XDA01040305).

References

[1] J. Gomez-Rodriguez, J. A. Readinger, I. C. Viorritto, K. L. Mueller, R. A. Houghtling, and P. L. Schwartzberg, "Tec kinases, actin, and cell adhesion," *Immunological Reviews*, vol. 218, no. 1, pp. 45–64, 2007.

[2] C. I. Smith, T. C. Islam, P. T. Mattsson, A. J. Mohamed, B. F. Nore, and M. Vihinen, "The Tec family of cytoplasmic tyrosine kinases: mammalian Btk, Bmx, Itk, Tec, Txk and homologs in other species," *Bioessays*, vol. 23, no. 5, pp. 436–446, 2001.

[3] A. Maas and R. W. Hendriks, "Role of Bruton's tyrosine kinase in B cell development," *Developmental Immunology*, vol. 8, no. 3-4, pp. 171–181, 2001.

[4] G. Matthew and G. K. William, "Role of inositol phospholipid signaling in natural killer cell biology," *Front Immunology*, vol. 4, article 47, 2013.

[5] A. J. Mohamed, L. Yu, C. M. Bäckesjö et al., "Bruton's tyrosine kinase (Btk): function, regulation, and transformation with special emphasis on the PH domain," *Immunological Reviews*, vol. 228, no. 1, pp. 58–73, 2009.

[6] T. S. Rans and R. England, "The evolution of gene therapy in X-linked severe combined immunodeficiency," *Annals of Allergy, Asthma and Immunology*, vol. 102, no. 5, pp. 357–363, 2009.

[7] J. D. Thomas, P. Sideras, C. I. Smith, I. Vorechovsky, V. Chapman, and W. E. Paul, "Colocalization of X-linked agammaglobulinemia and X-linked immunodeficiency genes," *Science*, vol. 261, no. 5119, pp. 355–358, 1993.

[8] M. Vihinen, B. H. Belohradsky, R. N. Haire et al., "BTKbase, mutation database for X-linked agammaglobulinemia (XLA)," *Nucleic Acids Research*, vol. 25, no. 9, pp. 166–171, 1997.

[9] T. Kaneko, R. Joshi, S. M. Feller, and S. S. Li, "Phosphotyrosine recognition domains: the typical, the atypical and the versatile," *Cell Communication and Signaling*, vol. 10, no. 1, p. 32, 2012.

[10] G. Koytiger, A. Kaushansky, A. Gordus, J. Rush, P. K. Sorger, and G. Macbeath, "Phosphotyrosine signaling proteins that drive oncogenesis tend to be highly interconnected," *Molecular & Cellular Proteomics*, vol. 12, no. 5, pp. 1204–1213, 2013.

[11] K. Chudasama, J. Winnay, S. Johansson et al., "Short syndrome with partial lipodystrophy due to impaired phosphatidylinositol 3 kinase signaling," *The American Journal of Human Genetics*, vol. 93, no. 1, pp. 150–157, 2013.

[12] M. Vihinen, M. D. Cooper, G. de Saint Basile et al., "BTKbase: a database of XLA-causing mutations," *Immunology Today*, vol. 16, no. 10, pp. 460–465, 1995.

[13] T. G. Kutateladze, "Molecular analysis of protein—phosphoinositide interactions," *Current Topics in Microbiology and Immunology*, vol. 362, pp. 111–126, 2012.

[14] P. Mayinger, "Phosphoinositides and vesicular membrane traffic," *Biochimica et Biophysica Acta*, vol. 1821, no. 8, pp. 1104–1113, 2012.

[15] L. E. Rameh, A. Arvidsson, K. L. Carraway et al., "A comparative analysis of the phosphoinositide binding specificity of pleckstrin homology domains," *Journal of Biological Chemistry*, vol. 272, no. 35, pp. 22059–22066, 1997.

[16] A. M. Scharenberg, O. El-Hillal, D. A. Fruman et al., "Phosphatidylinositol-3,4,5-trisphosphate (Ptdlns-3,4,5-P3)/Tec kinase-dependent calcium signaling pathway: a target for SHIP-mediated inhibitory signals," *The EMBO Journal*, vol. 17, no. 7, pp. 1961–1972, 1998.

[17] K. Saito, A. M. Scharenberg, and J.-P. Kinet, "Interaction between the Btk PH Domain and Phosphatidylinositol-3,4,5-trisphosphate Directly Regulates Btk," *Journal of Biological Chemistry*, vol. 276, no. 19, pp. 16201–16206, 2001.

[18] K. Murayama, M. Kato-Murayama, C. Mishima et al., "Crystal structure of the Bruton's tyrosine kinase PH domain with phosphatidylinositol," *Biochemical and Biophysical Research Communications*, vol. 377, no. 1, pp. 23–28, 2008.

[19] T. J. Gibson, M. Hyvönen, A. Musacchio, M. Saraste, and E. Birney, "PH domain: the first anniversary," *Trends in Biochemical Sciences*, vol. 19, no. 9, pp. 349–353, 1994.

[20] C. C. Milburn, M. Deak, S. M. Kelly, N. C. Price, D. R. Alessi, and D. M. F. van Aalten, "Binding of phosphatidylinositol 3,4,5-trisphosphate to the pleckstrin homology domain of protein kinase B induces a conformational change," *Biochemical Journal*, vol. 375, no. 3, pp. 531–538, 2003.

[21] T. F. Jenny and S. A. Benner, "A prediction of the secondary structure of the pleckstrin homology domain," *Proteins*, vol. 20, no. 1, pp. 1–3, 1994.

[22] B. J. Mayer, R. Ren, K. L. Clark, and D. Baltimore, "A putative modular domain present in diverse signaling proteins," *Cell*, vol. 73, no. 4, pp. 629–630, 1993.

[23] H. S. Yoon, P. J. Hajduk, A. M. Petros, E. T. Olejniczak, R. P. Meadows, and S. W. Fesik, "Solution structure of a pleckstrin-homology domain," *Nature*, vol. 369, no. 6482, pp. 672–675, 1994.

[24] E. Baraldi, K. D. Carugo, M. Hyvönen et al., "Structure of the PH domain from Bruton's tyrosine kinase in complex with inositol 1,3,4,5-tetrakisphosphate," *Structure*, vol. 7, no. 4, pp. 449–460, 1999.

[25] T. Kojima, M. Fukuda, Y. Watanabe, F. Hamazato, and K. Mikoshiba, "Characterization of the pleckstrin homology domain of Btk as an inositol polyphosphate and phosphoinositide binding domain," *Biochemical and Biophysical Research Communications*, vol. 236, no. 2, pp. 333–339, 1997.

[26] B. Shen and M. Vihinen, "Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain," *Protein Engineering, Design and Selection*, vol. 17, no. 3, pp. 267–276, 2004.

[27] M. Hyvönen and M. Saraste, "Stucture of the PH domain and Btk motif from Bruton's tyrosine kinase: molecular explanations for X-linked agammaglobulinaemia," *The EMBO Journal*, vol. 16, no. 12, pp. 3396–3404, 1997.

[28] S. Fujiwara and T. Amisaki, "Identification of high affinity fatty acid binding sites on human serum albumin by MM-PBSA method," *Biophysical Journal*, vol. 94, no. 1, pp. 95–103, 2008.

[29] T. Hou and R. Yu, "Molecular dynamics and free energy studies on the wild-type and double mutant HIV-1 protease complexed with amprenavir and two amprenavir-related inhibitors: mechanism for binding and drug resistance," *Journal of Medicinal Chemistry*, vol. 50, no. 6, pp. 1177–1188, 2007.

[30] P. A. Kollman, I. Massova, C. Reyes et al., "Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models," *Accounts of Chemical Research*, vol. 33, no. 12, pp. 889–897, 2000.

[31] J. Wang, P. Morin, W. Wang, and P. A. Kollman, "Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA," *Journal of the American Chemical Society*, vol. 123, no. 22, pp. 5221–5230, 2001.

[32] T. Hou, J. Wang, Y. Li, and W. Wang, "Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations," *Journal of Chemical Information and Modeling*, vol. 51, no. 1, pp. 69–82, 2011.

[33] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model," *Journal of Physical Chemistry*, vol. 97, no. 40, pp. 10269–10280, 1993.

[34] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Automatic atom type and bond type perception in molecular mechanical calculations," *Journal of Molecular Graphics and Modelling*, vol. 25, no. 2, pp. 247–260, 2006.

[35] Y. Duan, C. Wu, S. Chowdhury et al., "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations," *Journal of Computational Chemistry*, vol. 24, no. 16, pp. 1999–2012, 2003.

[36] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general Amber force field," *Journal of Computational Chemistry*, vol. 25, no. 9, pp. 1157–1174, 2004.

[37] T. Darden, L. Perera, L. Li, and L. Pedersen, "New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations," *Structure*, vol. 7, no. 3, pp. R55–R60, 1999.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 153839, 8 pages
http://dx.doi.org/10.1155/2013/153839

Review Article

Application of Bioinformatics in Chronobiology Research

Robson da Silva Lopes, Nathalia Maria Resende, Adenilda Cristina Honorio-França, and Eduardo Luzía França

Institute of Biological and Health Science, Federal University of Mato Grosso, Rodovia 070, Km 5 s/n 78600-000, Barra do Garças, MT, Brazil

Correspondence should be addressed to Adenilda Cristina Honorio-França; denifran@terra.com.br

Received 7 June 2013; Accepted 18 August 2013

Academic Editors: B. Shen and J. Wang

Copyright © 2013 Robson da Silva Lopes et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Bioinformatics and other well-established sciences, such as molecular biology, genetics, and biochemistry, provide a scientific approach for the analysis of data generated through "omics" projects that may be used in studies of chronobiology. The results of studies that apply these techniques demonstrate how they significantly aided the understanding of chronobiology. However, bioinformatics tools alone cannot eliminate the need for an understanding of the field of research or the data to be considered, nor can such tools replace analysts and researchers. It is often necessary to conduct an evaluation of the results of a data mining effort to determine the degree of reliability. To this end, familiarity with the field of investigation is necessary. It is evident that the knowledge that has been accumulated through chronobiology and the use of tools derived from bioinformatics has contributed to the recognition and understanding of the patterns and biological rhythms found in living organisms. The current work aims to develop new and important applications in the near future through chronobiology research.

1. Introduction

Living organisms are constantly exposed to cyclical variations that occur throughout the day, night, week, month, and year. These cycles have presented challenges for the survival of all species and have provided mechanisms for development throughout its evolution to anticipate and adjust to the changes that are present in these cycles [1].

As a result of evolution, every living organism carries within itself a clock that regulates its biological functions and determines how it will react to changes and rhythmic cycles, adequately supporting the various functions that develop.

Chronobiology is the science that studies biological rhythms characterized by the recurrent, regular intervals and the cyclic physical, biochemical, and behavioral phenomena that occur in all living organisms [2, 3].

There are many types of biological rhythms governing the bodies of living beings, including the frequency domain, which can be classified into ultradian, circadian, and infradian rhythms [4].

Circadian rhythms (Latin circa diem, meaning "about a day") have a duration of approximately 24 hours. Rhythms that exceed 24 hours are considered infradian rhythms and include the menstrual cycle and the rate of production of blood platelets. The infradian period can range from approximately seven days to 100 years in the extreme case of the reproductive cycle of Chinese bamboo. Cycles that have duration of less than 24 hours are called ultradian rhythms. These include high-frequency oscillations with periods less than 20 hours, including periods in the order of milliseconds, such as the firing rate of neurons, or in the order of minutes, such as the rhythm of heartbeats [3].

Within the field of chronobiology, there is great interest in understanding how behavioral, psychological, biochemical, and cellular rhythmicity oscillations present within a 24-hour period and how these oscillations are regulated by an internal clock (endogenous).

The circadian rhythms are mainly influenced by endogenous signals, but they adapt to external stimuli (exogenous) from the environment. Temporal integration of the internal and external rhythm is coordinated by the suprachiasmatic nucleus (SCN) of the hypothalamus through the monitoring of temporal signals, called Zeitgebers (a German neologism

meaning time marker), which may be temperature, food intake, and the sleep/wake cycle [3, 5].

The circadian rhythm is the most well-studied biological rhythm because it controls the structures responsible for the generation and synchronization of biological rhythms that correspond to 24-hour environmental cycles, that is, day and night. It consists of an oscillatory machine, including a central pacemaker, the SCN in the hypothalamus, and peripheral oscillators located in most tissues and cells, such as cardiomyocytes, fibroblasts, smooth muscle cells, and vascular endothelial cells [6], abdominal adipose tissue [7], and skeletal muscle tissue [8].

Briefly, the SCN receives stimuli from the external environment, such as light, which thereafter serves as a synchronizing clock that coordinates internal and external stimuli, such as the light-dark cycle, over 24 hours. The harmony and timing of this system allows the organism to have an anticipatory capacity, enabling it to prepare for events and activities that are basic to sustaining life, such as feeding and resting.

Human circadian rhythmicity is necessary for the functioning of the biological clock, which has central components defined by specific genes, called the "*clock*" genes. The protein products of these genes are essential for the generation and regulation of circadian rhythms within individual cells [9]. The genetic approach has revealed a remarkable conservation of molecular and biochemical features of the biological clock, providing a major breakthrough in identifying the role of each gene.

In mammals, the functioning of the circadian rhythm involves positive and negative feedback mechanisms. The CLOCK and BMAL1 genes (*brain and muscle Arnt-like protein 1*) form a heterodimer that functions as a transcription factor for expression of the gene PER (*period*) and its counterparts 1–3 (Per1, Per2, and Per3), CRY (*cryptochrome*), and the receptor REV-ERB. The PER and CRY genes are mainly transcribed in the morning and are transported from the cytoplasm to the nucleus, where they block the action of CLOCK/BMAL1. When the receptor REV-ERB is absent, the CLOCK and BMAL1 genes are released from this inhibition and are able to begin a new circadian rhythm [10].

Another regulatory mechanism of circadian rhythmicity is the transcription of nuclear receptors REV-ERB (α and β) and ROR (α, β, and γ), which is activated by the CLOCK/BMAL1 genes. REV-ERB (α and β) and ROR (α, β, and γ) compete to bind ROREs (*retinoic-acid-related orphan receptor response elements*) present at the promoter of BMAL1. The ROR activates the transcription of the BMAL1 gene, while REV-ERB inhibits the process. Thus, the circadian oscillation of BMAL1 is both positively and negatively regulated [9, 10].

In addition to the environmental Zeitgebers, humans are influenced by social markers. Social commitments, such as work and/or school and social demands, can act as synchronizing agents that function in the establishment of a social rhythm, and these commitments conform to a circadian pattern. Therefore, it can be said that the human biological clock is influenced by three aspects: environmental, social, and biological.

Chronobiology is not specific to humans or animals; it occurs in all living organisms. Studies in plants, such as *Arabidopsis thaliana*, *Oryza sativa* L. ssp. *japonica*, and *Populus trichocarpa*, suggest that a large proportion of the transcriptome is subject to circadian regulation [11]. The idea of the existence for such regulation, which gave rise to the field of chronobiology, had its beginning in approximately 1729, when the astronomer Jean-Jacques D'Ortous of Mairan (1675–1774) showed that the daily rhythm of the opening and closing of leaves in plants was maintained even in constant darkness. After this experiment, others showed that biological rhythms are reflections of environmental fluctuations.

Chronobiology is currently composed of a branch of biomedical sciences that seek to understand the relationships between the rhythms of biological functions, health, and disease. There are other specialties of growing interest within the field of chronobiology, such as chronopharmacology, chronopharmacokinetic, chronoanesthesia, chronoenergy, chronotoxicology, and chronotherapy [1].

It is evident that the circadian rhythm plays an important role in maintaining and coordinating the biological processes necessary for the efficient functioning of the living organism. Modern life often disrupts the workflow of the circadian rhythm in humans. For example, when we travel through different time zones, we suffer from jetlag, also called desynchronization. The symptoms of desynchronization also occur when we have to work on rotating shifts for 24 hours a day. The consequences of frequent jetlag have serious effects on health, including sleep disorders, chronic memory deficits, obesity, diabetes, and other metabolic diseases [7, 12], as well as the development of cancer [13].

Various chronomic studies [6–8, 13–15] seek to understand the endogenous and exogenous mechanisms of synchronization and desynchronization.

However, sleep disturbances and depression symptoms correlated with the time of appearance have been observed since the middle XVIII century. These studies showed that the lack of light affects both mood and physiological functions [3, 16]. These studies were not conclusive because the internal temporal order supplied the elements for the formulation of pathophysiological hypotheses. It was shown that depression and some types of insomnia exhibit changes in human circadian rhythmicity [3, 17].

The posttranslational mechanisms of these genes affect the stability, degradation, and localization of proteins. This can affect metabolic functions related to power and energy metabolism, which may lead to metabolic diseases, such as obesity and diabetes [10].

Several approaches to biological and genetic models can provide more information about the maintenance of the rhythmicity system and the occurrence of metabolic disorders caused by circadian disorders. The desynchronization of the biological clock resulting from the ingestion of food at an inappropriate time or deprivation of sleep increases the likeliness of metabolic risk and may cause disease [18].

All these discoveries have contributed to the development of more complex experimental methods, such as the use of "omics" science and the emergence of chronomics.

2. Chronomics

The "omics" sciences integrate genomics, transcriptomics, proteomics, and metabolomics to better understand and describe the cellular and molecular mechanisms of living organisms [19]. The mapping of the temporal structure of each variable of biological diversity has led to the development of another "omics" science, chronomics. This designation was first mentioned by Halberg et al. [20], who recommended the use of both experimental procedures and laboratory bioinformatics to map the anticipation of experiences. Chronomics enables quantify the phenomenon as a cycle and its occurrence as a rhythm, making it possible to replicate its events [21].

Since then, many studies of chronomics in medicine have been conducted. Blood pressure, for example, has been prominently featured [21]. Changes in the blood pressure over a complete circadian rhythm can be monitored, and, thus, the chronomics can be used to support the correct diagnosis. This diagnosis can lead to effective treatment and prevent stroke or other diseases [22].

As chronomics has been shown to be effective in diagnosing disease, its use to detect changes that are still reversible may allow for prevention of disease [22, 23].

Bioinformatics and chronomics may provide a tool for mapping the anticipation of the components involved in any biological system or the interaction of these systems. It will be used to improve estimates over a range and to correlate a variable with the multifrequency and the modulations of biological rhythms. Thus, the diagnosis/analysis may be correlated with biological systems with different paces [24].

In chronomics, the main tools used to store, manage, and analyze the events are derived from bioinformatics. Using these tools, chronomes are mapped, grouped, and compared, resulting in the correlation of the multifrequencies and modulations of biological rhythms [25].

Chronomics requires improvements in bioinformatics techniques, including the development of additional tools with features that are adapted to this new area of research.

On the other hand, the variables of biological systems are referred to as the chronome when they are analyzed using the techniques of chronomics. Physiological chronomes and their unique features are calculated within a specific range that is much shorter than the range of the observations of the system as a whole. However, this integrated analysis system is an increment of diagnosis [20, 22–24].

There is evidence that the circadian rhythm involves mechanisms that are independent of the cell nucleus. However, it is also essential to understand the molecular basis for the timing of changes in gene expression through transcriptomics, based on the observation of enzymatic rhythms in the nuclei of red blood cells [26]. More recently, Johnston [27] analyzed the transcriptome of mice for 24 hours and detected the presence of circadian rhythms in adipocytes.

Chronomics, the mapping of chronomes or time structures, aims to provide maps. The research of Löckinger et al. [28] aimed at assessing the circadian time structure of circulating peptides that are secreted by peripheral neurons and have strong effects on blood vessels in 20 healthy, young adults. They concluded that plasma concentrations of vasoactive intestinal peptide and cortisol present circadian rhythms, while substance P and neuropeptide Y undergo variations on the scale of hours.

Another area of "omics" science in which chronomics is particularly important is sportomics [29]. In this area, the focus is on understanding how the rhythms of body temperature are influenced by exercise. There is a proportional correlation between body temperature and the release of the hormone cortisol, a marker of stress [30].

Other studies in this area include investigations of the effects of exercise on sleep patterns, which have been studied since the 1980s. In one of the first epidemiological surveys on sleep and exercise, 1600 respondents aged between 31 and 50 years reported that the quality of sleep and physical performance are impacted by the social and psychological conditions associated with the location used for sleep, the sleep pattern, the lifestyle, and the living conditions of the individual [31].

In a study by Passos et al. [30], the total duration of each stage of sleep and the sleep efficiency were different after acute or chronic physical exercise. It was noted that during the process of adaptation to exercise, the body is restructured and returned to cycling stages of sleep, similar to its state before physical exercise.

In studies on the circadian rhythm response curve comparing exercise and bright light exposure, it was found that physical exercise had a significant influence on the circadian rhythm system, similar to the effects of light [32, 33].

All these studies reveal the complexity of the biological functions influenced by endogenous rhythmicity, including the physiological and biochemical mechanisms, as well as exogenous rhythmicity, such as the environment and its changes. Chronomics becomes an area for facilitating the integration of all these variables, and thus organizes the temporal chronomes for better interpretation and understanding.

However, the interrelatedness of the data from both areas is still a challenge for bioinformaticians and chronobiologists. A major barrier to progress is the lack of a proper infrastructure for developing data integration software and research groups.

New tools in bioinformatics have been developed, such as EUCLIS (EUCLOCK Information System), so that new circadian models can be exploited. This system allows the field of chronobiology to exploit the advantages of systems biology research in genetics to investigate the timing of the circadian rhythm at the level of the genome, the transcriptome, the proteome, and the metabolome [15].

EUCLIS takes advantage of the advanced architecture of the database that is used in pioneering liver cell research, the HepatoSys. According to Roennberg [13], this framework divides the database into several distinct modules: the experiments base contains the experimental procedures and provides a visualization tool for time series data; the knowledge base is a digital library that contains modules for common components used by the experiments base; the models base categorizes experimental models in humans, rodents, and birds; the genes base provides a catalog of genes associated with circadian rhythms in experimental

models, including homologous sequences, the functions of genes, genes associated with metabolic pathways, and their phenotypes, chromosomal locations, and the correlated literature; the references base is a repository of references including lessons and materials for chronobiological study; the tools base contains software used for simulation and analysis of experimental data to produce graphs and images; and the image base provides a repertoire of images and their metadata.

EUCLIS also provides a place to document events within the community through the museum base, as well as a place to document scientific genealogy through the pedigree base [13, 15].

EUCLIS was developed by chronobiologists in the European community, known as EUCLOCK [14], and was created to facilitate interaction among researchers. This community originated in January 2006 through the integration of the efforts of 34 chronobiology laboratories at 29 institutions in 11 European countries over 5 years. This network integration helped provide an understanding of how circadian clocks are synchronized with the rhythm of a particular environment, whether endogenous or exogenous. For this purpose, the researchers used the most advanced methods of phenotypic and functional genomics and developed innovative techniques to measure the periodicity of the circadian CLOCK genes in human skin fibroblasts and the phase of the circadian CLOCK genes in the oral mucosa and blood leukocytes, thus aiming at the development of diagnostic tools for diseases related to circadian rhythms [13].

The contributions of information technologies are indisputable, especially biotechnology and bioinformatics, for the treatment, storage, management, analysis, and visualization of large amounts of data compiled from the databases within the "omics" sciences.

3. Artificial Intelligence Techniques

Bioinformatics is an interdisciplinary field of study that has recently emerged and that involves physical and chemical biology, cellular and molecular biology, mathematics and computer science. The main goal of bioinformatics is to resolve problems arising from bioscience using computer science techniques, such as artificial intelligence, data mining, neural networks and Bayesians, and evolutionary algorithms.

Studies on physical performance also employ bioinformatics. With this science, it is possible to understand how each pathway of a particular process is working under conditions at rest or in different situations, such as during physical stress or even a pathological state. Thus, each compound and its related interactions can be represented on a metabolic map (integrated representation of human metabolism), providing a complete map of the metabolome, which we call a sportomic. The metabolic responses to exercise are dynamic and influenced by time. Due to the large amount of data derived from this study, the tools of bioinformatics, MarkerLynx (Waters Corporation, USA) and ChromaLynx (Waters Corporation, USA), were used for identification and quantification of the relevant compounds for the metabolomic analysis in accordance with a database from the National Institute of Standards of Technology (NIST) library [29, 34].

The modern technologies used for high-throughput measurements in molecular biology, such as microarray and RT-PCR, produce a large amount of data. With these large datasets, the emphasis has been to move from traditional statistical tests to new methods of data mining [35], which derive their roots from statistics, artificial intelligence, and machine learning [36].

4. Data Mining

Data mining focuses on automating the discovery of knowledge that is contained within a database but that is not readily perceptible. The use of data mining in a database is considered the central step in a larger process called knowledge discovery in databases (KDD), which includes several other processes that can be divided into preprocessing and postprocessing steps [37].

KDD preprocessing includes several steps, such as integration, cleansing, data discretization, and the selection of relevant attributes for the task of data mining. Postprocessing through KDD improves the understanding of the results obtained from data mining.

There are different data mining tasks, each of which aims to obtain a certain type of knowledge that is associated with a specific problem. The main tasks performed by DM algorithms are the development of clustering, association, and dependency models. These tasks can discover hidden associations or sequences within datasets, as well as clustering and visualizing the relationships among these data, enabling the prediction of hidden patterns.

Data mining can also be used in studies of gene expression rhythms [35]. Chronobiological studies used data mining or, more specifically, the task of clustering, to extract patterns from data to related behaviors and gene expression patterns in humans, rodents, and plants [7, 8, 11, 12].

Clustering involves dividing a dataset into groups based on the measurement of one or more characteristics of the data. Thus, it combines the elements with the same characteristics into different groups [38].

Filichkin et al. [11] showed that investigation of the cyclical transcriptome of plants could help elucidate the similarities and differences, as well as the common points of regulation in monocotyledonous and dicotyledonous plants. For this purpose, the authors used a combination of oligonucleotide microarrays and data mining to examine the daily rhythms in the gene expression of rice (*Oryza sativa* ssp. *japonica*) and poplar (*Populus trichocarpa*). The transcriptomic data related to rhythm were observed in different periods and under varying temperature conditions.

In the same study, the results identified groups of coexpressed genes at specific times of the day and showed that the rhythmic expression of the transcriptome of rice and poplar spanned all hours of the day, with peak expression levels at dawn and dusk [11].

Other studies have described the regulation of circadian gene expression in skeletal muscle [8], white adipose tissue, [7] and lungs tissue [12] of the rat.

In a study by Almon et al. [8], mRNA expression was observed in the gastrocnemius muscles in a series of 54 animals that were euthanized at different times within the 18 cycle of 24 hours. Mining of such data identified 109 genes that were expressed with rhythmicity and grouped them into eight distinct categories corresponding to 11 functions in the context of temporal expression.

In the study by Sukumaran et al. [7], circadian oscillations in gene expression in the white adipose tissue of mice were identified, and the regulatory role of circadian timing in the function of this tissue was examined. In total, 190 sets of probes that showed different circadian oscillations were identified through mining of the microarray data. These circadian probes sets were divided into seven different temporal groups in which >70% of the genes showed maximum expression during wakefulness (dark). These genes were grouped into eight functional categories that were examined in the context of their temporal expression. The circadian oscillations were also observed in the measurement of plasma leptin, glucose, insulin, corticosterone, triglycerides, free fatty acids, and LDL cholesterol. This oscillation of physiological circadian rhythms together with the functional classification of these genes suggests an important role for circadian rhythms in controlling various functions in white adipose tissue, including adipogenesis, energy metabolism, and the regulation of the immune system.

Similar work was conducted by Sukumaran et al. [12] to determine the patterns of gene expression in rat lungs during the light-dark cycle. In this work, 646 genes that showed variations in expression were identified through data mining, and eight different temporal groups were analyzed. More than two-thirds of the probes showed peak expression during the cyclic period of light or dark. Many of the genes whose expression peaked during the light were related to the extracellular matrix, the cytoskeleton, and protein processing, which seem to be mainly involved in tissue repair and remodeling.

Several studies have been conducted using data mining tools in the area of health for understanding pathological mechanisms. A study by Wood et al. [38] used data from Oncomine, an online microarray database that contains a set of gene expression data from various human cancers along with normal tissue controls. These data were analyzed to compare the levels of PERIOD gene expression (Per1, Per2, and Per3) in colon cancer and rectal and intestinal adenomas. The results showed that the role of these genes in circadian desynchronization can be used in cancer therapy based on nocturnal exposure to light.

5. Artificial Neural Networks and Rule-Based Systems

Beyond data mining, other artificial intelligence techniques, such as artificial neural networks [39, 40] and rule-based systems [41, 42], have contributed to the study of chronobiology.

An artificial neural network is a parallel processing system that has nonalgorithmic computation as its main feature, which is reminiscent of the structure of the workings of the human brain neurons [43].

In troubleshooting this technique, the model initially passes through a learning phase in which a reduced set of examples is presented to the network. Then, the network automatically learns the characteristics required to represent the information provided and therefore recognizes any existing standard that can be used to later identify consistent data in an unknown data set.

In recent years, Artificial Neural Networks have received considerable attention as a computing tool in several areas, including the learning decision process, prediction problems, pattern similarity discovery, data filtering, automatic acquisition of knowledge, monitoring and rapid diagnosis, and incomplete information processing, among others.

In chronobiological studies, artificial neural networks were used to identify patterns in the circadian rhythms of humans. The work of Kolodyazhniy et al. [40] presented an approach to a regression nonlinear model based on artificial neural networks with the aim of recognizing patterns in the circadian rhythms of 25 healthy young humans. The data were derived using an ambulatory multichannel monitoring system during daily routines for one week. The devices collected large amounts of data related to physiological, behavioral, and environmental factors, including body temperature, cardiovascular and respiratory functions, motion, posture, local temperature, light intensity perceived at eye level, and sleep [44].

The results showed that this type of neural network significantly enhances the variance of error prediction compared with traditional approaches for determining the circadian phase based on individual indicators (body temperature, acceleration of movement, and sleep recording). Two sets of noninvasive measures were identified that, combined with the predictive model, can provide researchers and clinicians with an accurate measure of endogenous time. This method was validated in healthy young men and requires testing in a clinical population or people who suffer from sleep/wake disturbances.

In studies conducted by Jara et al. [42], an intelligent system to detect and predict myocardial diseases by analyzing the electrocardiogram curve was developed.

Intelligent systems seek to simulate the knowledge of an expert in the subject through a rule base that is used to review the information presented to the system. The system developed by Jara et al. [42] first detects the symptoms and then makes the prediction of the disease. The system presents an algorithm that is based on chronobiological studies demonstrating that a myocardial infarction could be predicted up to eight days before it occurs.

This study is based on the assumption that the rate of the heartbeat of a patient is highly variable, and it was necessary to perform monitoring for 24 hours. For each hour, the maximum and minimal heart rates were recorded. With these data, the system's intelligence-based Jess rules determined the difference between the minimum and maximum heart rate, which indicated a risk for myocardial disease. The authors emphasize the importance of further tests before substantiating a positive diagnosis.

Based on these studies, it is clear that the various techniques originating from computer science and integrated

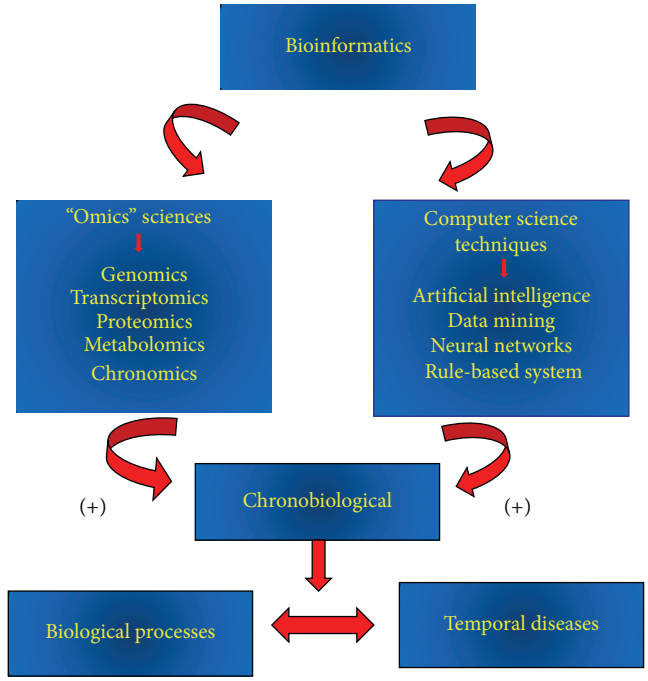


Figure 1: Relationship between bioinformatics, chronobiological, and diseases.

with the knowledge of cell and molecular biology, physiology, chemistry, medicine, and bioinformatics have assisted in the study and comprehension of how biological rhythms of living organisms influence the neuro-immuno-endocrine response. The results have become more consistent and indispensable as the amount of data generated in studies of the "omics" sciences and computer science techniques has increased. The variables involved are complex and numerous and it is possible that the bioinformatics and chronobiological can be applicable in understanding the biological rhythms and various diseases (Figure 1).

The observation and recording of endogenous and exogenous biological rhythmicity is the focus of chronomics in collaboration with bioinformatics. This approach will contribute to improvements in the diagnosis of disease and understanding the pathophysiology and treatment of diseases, as well as the temporal variables of disease, based on quantitative and qualitative interpretations of various disorders. The fields of chronomics with bioinformatics are applicable not only in diagnosis but also in understanding the chronomes according to an individual's rhythm.

6. Conclusion

In recent decades, chronobiology has revealed the existence of a temporal regulation system that synchronizes all body systems to environmental cycles, such as the day-night cycle. Therefore, it is now known that living organisms respond proactively to environmental rhythmicity and therefore actively prepare.

Although considerable progress has been achieved in the understanding of the biological rhythms in the SCN and its synchronization with the peripheral tissues and exogenous signals, many questions remain about the functioning and dynamics of biological rhythms and how they influence and are influenced by genetics and phenotype.

The studies of the "omics" sciences produce a large amount of information on the molecular biology of various living organisms, which is used for the production of new scientific knowledge through different experimental paradigms. The in silico approach is represented by methodologies based on the creation of algorithms and bioinformatics tools.

Maps of chronomes in circadian and noncircadian patterns may be useful for screening, diagnosis, and a better understanding of the cellular and molecular mechanisms of life. The complex maps of chronomes should be made available to the scientific community online so that investigators may study the effects of gender, age, ethnicity, patterns of behavior, and geographic location as a function of time.

Bioinformatics, as well as other well-established sciences, provides an approach based on scientific methods for the analysis of data generated by "omics" projects that can be used in studies of chronobiology.

Some of the studies presented in this paper made use of various bioinformatic techniques, such as data mining, artificial neural networks, and rule-based systems, to extract patterns in the rhythms of CLOCK gene expression and understand how these rhythms can influence rhythmic oscillations in the peripheral tissues. The results of studies that apply such techniques show how they have significantly aided the understanding of chronobiology.

However, no bioinformatics tool is sufficient to eliminate the need for an understanding of the field of research or the data to be manipulated, nor can such tools replace analysts and researchers. For the application of bioinformatics tools,

such as the KDD, in which preprocessing is necessary to select the relevant data from a repository and perform cleanup and discretization, the intrinsic knowledge of researchers and scientists in the field is extremely important. Similarly, in postprocessing KDD, it is often necessary to conduct an evaluation of the results of mining to determine the degree of reliability.

It is evident that our knowledge has already increased in the field of chronobiology, and the use of tools derived from bioinformatics can contribute to the recognition and understanding of the patterns in the biological rhythms of living organisms. New and important applications for chronobiology will be developed in the near future.

Conflict of Interests

The authors declare no conflict of interests and nonfinancial competing interests.

Acknowledgments

The authors are very grateful to the Postgraduate Program: DINTER UFMT/UFMG. This research was supported by CAPES (Coordination for Improvement of Higher Education Personnel, No. 23038008852).

References

[1] A. R. Acúrcio and L. M. Rodrigues, "The rhythms of life—an uptaded vision of applied chronobiology," *Revista Lusófona de Ciências e Tecnologia da Saúde*, vol. 6, pp. 216–234, 2009.

[2] J. Aschoff, "Circadian rhythms in man," *Science*, vol. 148, no. 3676, pp. 1427–1432, 1965.

[3] B. Lemmer, "Discoveries of rhythms in human biological functions: a historical review," *Chronobiology International*, vol. 26, no. 6, pp. 1019–1068, 2009.

[4] C. E. D. S. Ferreira and A. Andriolo, "Reference ranges in clinical laboratory," *Jornal Brasileiro de Patologia e Medicina Laboratorial*, vol. 44, no. 1, pp. 11–16, 2008.

[5] D. S. Pereira, S. Tufik, and M. Pedrazzoli, "Moléculas que marcamotempo: ImplicacõEs para os fenótipos circadianos," *Revista Brasileira de Psiquiatria*, vol. 31, no. 1, pp. 63–71, 2009.

[6] F. Portaluppi, R. Tiseo, M. H. Smolensky, R. C. Hermida, D. E. Ayala, and F. Fabbian, "Circadian rhythms and cardiovascular health," *Sleep Medicine Reviews*, vol. 16, no. 2, pp. 151–166, 2012.

[7] S. Sukumaran, B. Xue, W. J. Jusko, D. C. DuBois, and R. R. Almon, "Circadian variations in gene expression in rat abdominal adipose tissue and relationship to physiology," *Physiological Genomics*, vol. 42, no. 2, pp. 141–152, 2010.

[8] R. R. Almon, E. Yang, W. Lai et al., "Relationships between circadian rhythms and modulation of gene expression by glucocorticoids in skeletal muscle," *American Journal of Physiology*, vol. 295, no. 4, pp. R1031–R1047, 2008.

[9] I. Kwon, H. K. Choe, G. H. Son, and K. Kim, "Mammalian molecular clocks," *Experimental Neurobiology*, vol. 20, pp. 18–28, 2011.

[10] J. Delezie and E. Challet, "Interactions between metabolism and circadian clocks: reciprocal disturbances," *Annals of the New York Academy of Sciences*, vol. 1243, no. 1, pp. 30–46, 2011.

[11] S. A. Filichkin, G. Breton, H. D. Priest et al., "Global profiling of rice and poplar transcriptomes highlights key conserved Circadian-controlled pathways and cis-regulatory modules," *PLoS ONE*, vol. 6, no. 6, Article ID e16907, 2011.

[12] S. Sukumaran, W. J. Jusko, D. C. DuBois, and R. R. Almon, "Light-dark oscillations in the lung transcriptome: implications for lung homeostasis, repair, metabolism, disease, and drug action," *Journal of Applied Physiology*, vol. 110, no. 6, pp. 1732–1747, 2011.

[13] T. Roenneberg, "EUCLOCK Entrainment of the Circadian Clock," Integrated Project Funded by the 6th Framework Programme of the European Commission, 2011.

[14] R. T. B. Batista, D. B. Ramirez, R. D. Santos, M. C. I. del Rosario, and E. R. Mendoza, "EUCLIS—an information system for circadian systems biology," *IET Systems Biology*, vol. 1, no. 5, pp. 266–273, 2007.

[15] D. Hadrich, 2006, Project Info: EUCLOCK-Entrainment of the Circadian Clock. EUCLOCK-Entrainment of the Circadian Clock, http://www.euclock.org/index.php.

[16] I. N. Karatsoreos, "Effects of circadian disruption on mental and physical health," *Current Neurology Neuroscience Repiratory*, vol. 12, pp. 218–225, 2012.

[17] M. A. Quera Salva, S. Hartley, F. Barbot, J. C. Alvarez, F. Lofaso, and C. Guilleminault, "Circadian rhythms, melatonin and depression," *Current Pharmaceutical Design*, vol. 17, no. 15, pp. 1459–1470, 2011.

[18] M. Abe, E. D. Herzog, S. Yamazaki et al., "Circadian rhythms in isolated brain regions," *The Journal of Neuroscience*, vol. 22, no. 1, pp. 350–356, 2002.

[19] J. Boccard, J. Veuthey, and S. Rudaz, "Knowledge discovery in metabolomics: an overview of MS data handling," *Journal of Separation Science*, vol. 33, no. 3, pp. 290–304, 2010.

[20] F. E. Halberg, G. Cornélissen, K. Otsuka, O. Schwartzkopff, J. Halberg, and E. E. Bakken, "Chronomics," *Biomedicine and Pharmacotherapy*, vol. 55, no. 1, pp. 153–190, 2000.

[21] F. Halberg, G. Cornélissen, G. Katinas et al., "Transdisciplinary unifying implications of circadian findings in the 1950s," *Journal of Circadian Rhythms*, vol. 29, article 2, 2003.

[22] G. Cornélissen, F. Halberg, E. E. Bakken et al., "Chronobiology of high blood pressure," *Scripta Medica*, vol. 80, pp. 157–166, 2007.

[23] F. Halberg, G. Cornelissen, G. Katinas et al., "Chronobiology's progress. Part I, season's appreciations 2004-2005: time-, frequency-, phase-, variable-, individual-, age- and site-specific chronomics," *Journal of Applied Biomedicine*, vol. 4, no. 1, pp. 1–38, 2006.

[24] F. Halberg, G. Cornélissen, G. Katinas et al., "Chronobiology's progress. Part II, chronomics for an immediately applicable biomedicine," *Journal of Applied Biomedicine*, vol. 4, no. 2, pp. 73–86, 2006.

[25] D. Gonze, J. Halloy, J. Leloup, and A. Goldbeter, "Stochastic models for circadian rhythms: effect of molecular noise on periodic and chaotic behaviour," *Comptes Rendus*, vol. 326, no. 2, pp. 189–203, 2003.

[26] S. F. Mabood, P. F. J. Newman, and I. A. Nimmo, "Circadian rhythms in the activity of acetylcholinesterase of hyman erythrocytes incubated in vitro," *Biochemical Society Transactions*, vol. 6, no. 1, pp. 305–308, 1978.

[27] J. D. Johnston, "Adipose circadian rhythms: translating cellular and animal studies to human physiology," *Molecular and Cellular Endocrinology*, vol. 349, no. 1, pp. 45–50, 2012.

[28] A. Löckinger, D. Köberle, P. St. König et al., "Neuropeptide chronomics in clinically healthy young adults: circaoctohoran and circadian patterns," *Peptides*, vol. 25, no. 4, pp. 533–542, 2004.

[29] N. M. Resende, A. M. Magalhães-Neto, F. Bachini, L. E. V. Castro, A. Bassini, and L. C. Cameron, "Metabolic changes during a field experiment in a world-class windsurfing athlete: a trial with multivariate analyses," *OMICS*, vol. 15, pp. 695–704, 2011.

[30] G. S. Passos, D. L. Poyares, M. G. Santana, S. Tufik, and M. T. Mello, "Is exercise an alternative treatment for chronic insomnia?" *Clinics*, vol. 67, pp. 653–660, 2012.

[31] I. Vuori, H. Urponen, J. Hasan, and M. Partinen, "Epidemiology of exercise effects on sleep," *Acta Physiology Scandinavian*, vol. 574, pp. 3–7, 1988.

[32] V. A. Lemos, H. K. Antunes, R. V. dos Santos, F. S. Lira, S. Tufik, and M. T. de Mello, "High altitude exposure impairs sleep patterns, mood, and cognitive functions," *Psychophysiology*, vol. 49, pp. 1298–1306, 2012.

[33] G. Glickman, I. C. Webb, J. A. Elliott et al., "Gorman Photic sensitivity for circadian response to light varies with photoperiod," *Journal of Biological Rhythms*, vol. 27, pp. 308–318, 2012.

[34] F. Halberg, G. Cornélissen, G. Katinas et al., "Chronomics and genetics," *Scripta Medica*, vol. 80, pp. 133–150, 2007.

[35] A. Kallio, N. Vuokko, M. Ojala, N. Haiminen, and H. Mannila, "Randomization techniques for assessing the significance of gene periodicity results," *BMC Bioinformatics*, vol. 12, article 330, 2011.

[36] C. W. Hanson and B. E. Marshall, "Artificial intelligence applications in the intensive care unit," *Critical Care Medicine*, vol. 29, no. 2, pp. 427–435, 2001.

[37] A. A. Freitas, "A survey of evolutionary algorithms for data mining and knowledge discovery," in *Advances in Evolutionary Computation*, A. Ghosh and S. Tsutsui, Eds., Springer, New York, NY, USA, 2002.

[38] P. A. Wood, X. Yang, and W. J. M. Hrushesky, "The role of circadian rhythm in the pathogenesis of colorectal cancer," *Current Colorectal Cancer Reports*, vol. 6, no. 2, pp. 74–82, 2010.

[39] K. E. Hubbard, F. C. Robertson, N. Dalchau, and A. A. R. Webb, "Systems analyses of circadian networks," *Molecular BioSystems*, vol. 15, pp. 1502–1511, 2009.

[40] V. Kolodyazhniy, J. Späti, S. Frey et al., "An improved method for estimating human circadian phase derived from multichannel ambulatory monitoring and artificial neural networks," *Chronobiology International*, vol. 29, pp. 1078–1097, 2012.

[41] A. J. Jara, F. J. Blaya, M. A. Zamora, and A. F. G. Skarmeta, "An ontology and rule based intelligent information system to detect and predict myocardial diseases," in *Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine (ITAB '09)*, pp. 1–6, Larnaca, Cyprus, November 2009.

[42] A. J. Jara, M. A. Z. Izquierdo, and A. F. G. Skarmeta, "A wearable system for Tele-monitoring and Tele-assistance of patients with integration of solutions from chronobiology for prediction of illness," in *Ambient Intelligence and Smart Environments*, vol. 1, pp. 221–228, IOS Press Ebooks, 2009.

[43] A. D. P. Braga, T. B. Ludemir, and A. C. P. F. Carvalho, *Artificial Neural Networks: Theory and Applications*, LTC, Rio de Janeiro, Brazil, 2000.

[44] M. Garaulet and J. A. Madrid, "Chronobiological aspects of nutrition, metabolic syndrome and obesity," *Advanced Drug Delivery Reviews*, vol. 62, no. 9-10, pp. 967–978, 2010.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 748979, 8 pages
http://dx.doi.org/10.1155/2013/748979

Research Article

NCK2 Is Significantly Associated with Opiates Addiction in African-Origin Men

Zhifa Liu,[1] Xiaobo Guo,[1,2] Yuan Jiang,[3] and Heping Zhang[1]

[1] *Department of Biostatistics, Yale University School of Public Health, New Haven, CT 06520, USA*
[2] *Department of Statistical Science, School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou 510275, China*
[3] *Department of Statistics, Oregon State University, Corvallis, OR 97331, USA*

Correspondence should be addressed to Heping Zhang; heping.zhang@yale.edu

Received 31 December 2012; Accepted 18 January 2013

Academic Editors: J. Ma and B. Shen

Copyright © 2013 Zhifa Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Substance dependence is a complex environmental and genetic disorder with significant social and medical concerns. Understanding the etiology of substance dependence is imperative to the development of effective treatment and prevention strategies. To this end, substantial effort has been made to identify genes underlying substance dependence, and in recent years, genome-wide association studies (GWASs) have led to discoveries of numerous genetic variants for complex diseases including substance dependence. Most of the GWAS discoveries were only based on single nucleotide polymorphisms (SNPs) and a single dichotomized outcome. By employing both SNP- and gene-based methods of analysis, we identified a strong (odds ratio = 13.87) and significant (P value = $1.33E-11$) association of an SNP in the *NCK2* gene on chromosome 2 with opiates addiction in African-origin men. Codependence analysis also identified a genome-wide significant association between *NCK2* and comorbidity of substance dependence (P value = $3.65E-08$) in African-origin men. Furthermore, we observed that the association between the *NCK2* gene (P value = $3.12E-10$) and opiates addiction reached the gene-based genome-wide significant level. In summary, our findings provided the first evidence for the involvement of *NCK2* in the susceptibility to opiates addiction and further revealed the racial and gender specificities of its impact.

1. Introduction

Substance dependence is believed to result from a combination of genetic and environmental factors. Since substance dependence is a chronic brain disease, with high relapse rates, it causes serious social, economic, and medical consequences [1–3]. The World Health Organization (WHO) and the United Nations Office on Drugs and Crime (UNODC) reported that opiates dependence is associated with a high risk of HIV infection when opiates are injected using contaminated injection equipment [4]. Paulozzi et al. in 2006 reported that the number of deaths which involved prescription opioid analgesics increased from 2,900 in 1999 to at least 7,500 in 2004, an increase of 160% in just 5 years [5]. All available evidence indicated that the increasing numbers of deaths are significantly correlated to the increasing use of prescription drugs, especially opioid painkillers, among people during the working years of life. While exposure to drugs is the prerequisite for addiction, the most important question is as follows: who will be addicted after the exposure? Genes are believed to be a major factor, although it is most likely that there are multiple genes as well as gene-environment interactions. For this reason, understanding the genetic mechanisms behind vulnerability to drug addiction is critical to improve the quality of overall health and life.

Linkage and genome-wide association studies (GWASs) have implicated many regions and genes for dependence on alcohol, tobacco, and opiates. GABRA2, CHRM2, ADH4, PKNOX2, GABRG3, TAS2R16, SNCA, OPRK1, and PDYN have all been associated with alcohol dependence with various degrees of replication [6–21]. Associations of other candidate alcohol dependence genes, such as KIAA0040, ALDH1A1, and MANBA [18, 20, 22–25], remain to be confirmed. Several groups reported CHRNA5, CHRNA3,

CHRNB4, and CSMD1 to be associated with nicotine dependence [26–34]. Meanwhile, recent studies also reported that a group of genes, such as OPRM1 [35–37], OPRD1, OPRK1 [21, 38, 39], HTR1B [40], SLC6A4 [41], GABRG2 [42], and BDNF [43], to be associated or in linkage with opiates addiction.

Complex diseases may involve heterogeneous genetic effects in different ethnic and gender groups [7, 44–47]. Luo et al. [44] reported that African-origin smokers become dependent at a lower threshold (number of cigarettes per day) than European-origin smokers. Hartel et al. [46] found that men are more vulnerable to addiction when compared to women. In addition, Chen et al. [7] revealed that *PKNOX2* is associated with drug addiction in European-origin women. These examples underscore the necessity to consider demographic or even other covariates in genetic association studies.

Many of the reported genetic variants have been identified through single SNP association tests. Despite many of the successes, a single SNP tends to have a small effect, and the single SNP-based association tests require a very stringent significance level, which is likely a key factor to the so-called "missing heritability" problem [48, 49]. To overcome some of these limitations, gene-based analysis [50–52] has emerged to jointly analyze the SNPs within genes. Gene-based methods are less affected by the heterogeneity of a single locus; hence the results may be more robust across populations [53], which increases the likelihood of replication. Hence, we performed both single SNP-based and gene-based association analyses for the data from the Study of Addiction: Genetics and Environment (SAGE) [6] which includes well-characterized phenotypic data on substance dependence including addiction to nicotine, alcohol, marijuana, cocaine, opiates, and other drugs. In our analysis, we find a genome-wide significant association of *NCK2* gene on chromosome 2 with opiates dependence in African-origin men at both the SNP and gene levels. *NCK2* is a member of NCK family of adaptor proteins, which is associated with tyrosine-phosphorylated growth factor receptors of their cellular substrates [54]. However, to the best of our knowledge, *NCK2* has not been reported to be associated with any drug addiction outcomes in humans.

2. Materials and Methods

Phenotypes for multisubstance dependency and genome-wide SNP data from SAGE [6] were downloaded from dbGaP (http://www.ncbi.nlm.nih.gov/gap). SAGE is a large case-control association study which investigates the genetic variants for drug addiction. The samples were collected from three large-scale genome-wide association studies: Collaborative Study on the Genetic of Alcoholism (COGA), the Family Study of Cocaine Dependence (FSCD), and the Collaborative Genetic Study of Nicotine Dependence (COGEND) [16, 44, 55, 56]. The original data set contains 4,121 subjects with six categories of substance dependence data: addiction to alcohol, cocaine, marijuana, nicotine, opiates, and other drugs. Lifetime dependence on these six substances is diagnosed by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). The genotyping was performed by the illumina Human 1 M platform. In this study, we followed a quality control/quality assurance process similar to previous analyses [7, 57]. Individuals with call rates <90% and SNPs with minor allele frequency MAF <1% were excluded from the analysis. The P value for the Hardy-Weinberg equilibrium was set up by >0.0001. These steps reduced the level of noise in genotypes and increased the efficiency of analysis. There are 60 duplicate genotype samples and 9 individuals with ethnic backgrounds other than African origin or European origin. All of those individuals were removed from the subject list. Finally, there were a total of 3,627 unrelated samples with 859,185 autosomal SNPs for our final analysis. To alleviate the confounding by population substructure, we stratified the sample by race and sex. Finally, there are four sub-samples: 1,393 European-origin women, 1,131 European-origin men, 568 African-origin women and 535 African-origin men. The distribution of subjects diagnosed with lifetime dependence on substances in each of the six categories: nicotine, alcohol, marijuana, cocaine, opiates, or other drugs are presented in Table 1.

Table 1: Descriptive statistics of the key variables in the SAGE dataset stratified by sex and race.

	Black men	White men	Black women	White women	Overall
n	535	1131	568	1393	3627
Age (SD) yr.	40.9 (8.2)	38.7 (10.3)	39.7 (6.7)	38.2 (9.1)	39 (9.1)
Alcohol (%)	62.1	62.3	39.4	31.1	46.7
Cocaine (%)	46.4	27.3	36.3	12.5	25.8
Marijuana (%)	25.4	25.2	13.7	8.7	17.1
Nicotine (%)	47.5	46.7	47.7	41.1	44.8
Opiates (%)	8.2	9.9	6.2	4.8	7.1
Other drugs (%)	11.4	18	6.5	9.4	11.9
No drug (%)	27.1	31	38.9	50.1	39

3. Methods

Figure 1 displays the flow chart of our analytic strategy, and the details of the association analysis methods are described later.

3.1. Statistical Analysis for Single Trait. The SNP-based association is performed by the standard allelic test and logistic regression to obtain the P values for individual SNPs, and PLINK software (version 1.07) was used for analysis [58]. Meanwhile, a list of SNP pairs in linkage disequilibrium (LD) ($r^2 > 0.2$) is calculated for the gene-based association test.

For the gene-based analysis, we used the open-source tool: Knowledge-Based Mining System for Genome-Wide Genetic Studies (KGG, version 2.0)—based on the SNP association test results and LD files produced by PLINK. The procedure was performed as the following. We first calculate the effective number m_e of independent P value among m SNPs within a gene. Then, we sort the SNPs and calculate the effective number $m_{e(j)}$ of independent P-values among the

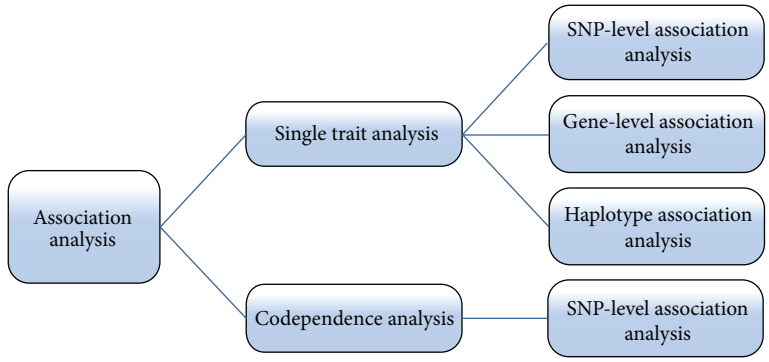


Figure 1: The pipeline of the association analysis.

top j significant SNPs. Finally, the modified Simes test [51] was employed to obtain a gene-based P value as follows,

$$P_G = \text{Min}\left(\frac{m_e P_{(j)}}{m_{e(j)}}\right), \quad (1)$$

where $P_{(j)}$ is the jth most significant among the m SNPs within a gene. We refer the interested readers to [51] for details.

In the gene-based method, SNPs within 20 kilo bases (kb) $5'$ upstream and 10 kilo bases (kb) $3'$ downstream of a gene's coding regions [59] were assigned to the gene. In addition, we included other SNPs if they are in strong LD ($r^2 > 0.9$) with the initially mapped SNPs within the gene [60].

Since there are about 20,000 protein coding genes in human genome, we used $0.05/20{,}000 = 2.5 \times 10^{-6}$ as the genome-wide significance threshold for the gene-based association test. In contrast, we used 5.0×10^{-8} as the genome-wide significance threshold for the SNP-based association test [61].

3.2. Codependence Association Analysis. Although logistic regression is commonly used to study a binary outcome, it is not suitable to evaluate comorbidity involving multiple outcomes. We use a nonparametric association test based on Kendall's tau [62] to study the comorbidity. The Kendall's tau-based association test proceeds as follows.

Suppose that we observe a p-dimensional vector of traits $Y_i = (Y_i^{(1)}, \ldots, Y_i^{(p)})^T$, genotype G_i, and a q-dimensional vector of covariates $Z_i = (Z_i^{(1)}, \ldots, Z_i^{(q)})^T$ for the ith subject in a population-based study with n subjects, and $\{(Y_i, G_i, Z_i) : i = 1, \ldots, n\}$ are independent samples. For subjects i and j, let Y_i and Y_j be their vectors of traits, respectively, and analogously, G_i and G_j and Z_i and Z_j are their genotypes and covariates. Generalized from Kendall's tau, a U statistic is defined to measure the association between Y and G as follows:

$$U = \binom{n}{2}^{-1} \sum_{i<j} \left(Y_i - Y_j\right)\left(G_i - G_j\right). \quad (2)$$

Without considering the covariates and conditioning on all phenotypes, U follows an asymptotically normal distribution in the absence of association [63]. To accommodate covariates, a weighted U statistic has been developed [64, 65]. We refer to Jiang and Zhang [64] for a detailed description of the method. For the purpose of comparison, we present the results with and without considering age as the covariate. Recall that our analysis is stratified by ethnicity and gender.

4. Results

4.1. Association Analysis at SNP Level. Table 2 summarizes the top four significant SNPs (with $P < 1.0 \times 10^{-4}$) in gene *NCK2* on chromosome 2 (2q12) for opiates dependence in African-origin men. We identified a genome-wide significant SNP (rs2377339 with $P = 1.33 \times 10^{-11}$) for the opiates dependence in African-origin men by the allelic test. Logistic regression also yielded strong evidence for the association between the SNP rs2377339 ($P = 1.01 \times 10^{-7}$) and opiates dependence although the P-value did not reach the genome-wide significance threshold. In addition, Table 2 presents the association results for the other five addictions with the four candidate SNPs. None of the four SNPs appeared significantly associated with the other five substance addictions.

4.2. Association Analysis at Gene Level. The gene-based association results are displayed in the last two rows of Table 2. Specifically, we included 39 SNPs in NCK2. The P values from the gene-*NCK2*-based tests that were obtained through the standard allelic test and logistic regression are 3.12×10^{-10} and 2.70×10^{-6}, respectively. The gene-based P value from the standard allelic test reached the genome-wide significance at gene level. The gene-based P value through logistic regression is very close to the gene-based genome-wide significance level. Therefore, both methods provided significant evidence that supports the association between the *NCK2* gene and opiates dependence in African-origin men. For the addiction of the other five substances in African-origin men, nicotine dependence had the most significant association with the *NCK2* gene ($P = 9.56 \times 10^{-3}$).

4.3. Haplotypes Analysis. We also examined association of haplotypes with opiate addiction in *NCK2* region. Figure 2 displays the linkage disequilibrium (LD) heat map of 14 SNPs

Table 2: Association of the most significant SNPs in the NCK2 gene with the six substances dependence in African-origin sen.

Gene	SNP	Method	Alcohol	Cocaine	Marijuana	Nicotine	Opiates	Others
NCK2	rs2377339	Logistic regression	$2.46E-2$	$5.09E-2$	$4.48E-2$	$7.01E-2$	**$1.10E-7$**	$3.84E-3$
		Standard allelic test	$6.03E-3$	$4.34E-2$	$3.89E-2$	$6.25E-2$	**$1.33E-11$**	$1.78E-3$
	rs7589342	Logistic regression	$2.84E-1$	$8.79E-1$	$9.35E-1$	$8.60E-1$	$1.45E-4$	$4.26E-1$
		Standard allelic test	$2.81E-1$	$8.78E-1$	$9.34E-1$	$8.59E-1$	$5.39E-5$	$4.22E-1$
	rs12995333	Logistic regression	$1.71E-1$	$7.51E-1$	$7.16E-1$	$9.86E-1$	$1.89E-4$	$4.69E-1$
		Standard allelic test	$1.68E-1$	$7.51E-1$	$7.15E-1$	$9.86E-1$	$7.82E-5$	$4.67E-1$
	rs12053259	Logistic regression	$1.39E-1$	$9.43E-1$	$7.42E-1$	$9.01E-1$	$2.31E-4$	$4.86E-1$
		Standard allelic test	$1.33E-1$	$9.42E-1$	$7.39E-1$	$9E-1$	$8.67E-5$	$4.80E-1$
NCK2	—	KGG-logistic	$1.83E-1$	$8.15E-1$	$5.92E-1$	$9.56E-3$	**$2.70E-6$**	$9.45E-2$
NCK2	—	KGG- Standard allelic test	$1.41E-1$	$8.16E-1$	$4.83E-1$	$8.71E-3$	**$3.12E-10$**	$4.17E-2$

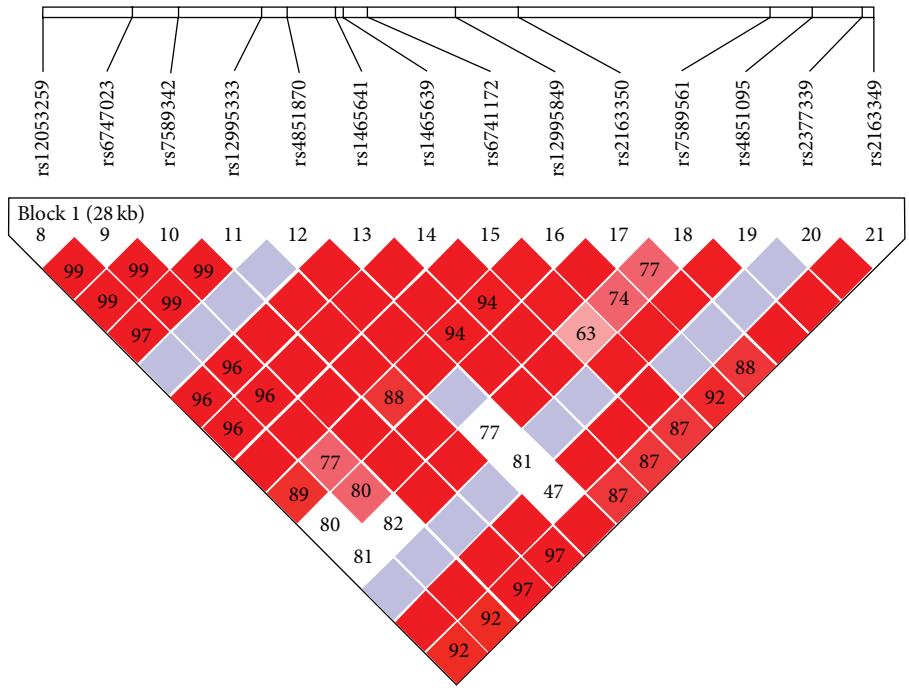


Figure 2: Linkage disequilibrium heat map near SNP rs2377339 on chromosome 2.

Table 3: Allele counts of rs2377339 in cases (opiates dependence) and controls (nonopiates dependence) in African-origin men.

	Genotype AA	Genotype AG	Genotype GG	Total
Case	35	9	0	44
Control	483	8	0	491

in 28 kb region [66]. Haplotype "AGTTCAGATCTCGT" with probability 0.016 yielded a P value of 1.66×10^{-11}. The genome-wide significant association between this haplotype and opiate addiction reduces the chance of a false discovery at the peak of a single SNP.

4.4. Contingency Table Analysis. We further examined the relationship between SNP rs2377339 and the opiates dependence in African-origin men. Table 3 depicts the allele frequencies of SNP rs2377339. The proportion of individuals having minor allele G is 21.43% in the case group and 1.63% in the control group. The odds ratio of SNP rs2377339 is 13.87, indicating that those who have the risk allele (G) for rs2377339 are at a significantly increased risk of being diagnosed with opiates dependence.

4.5. Stratification Analysis. Furthermore, in Table 4, we investigated the racial specificity and sex difference in the association between SNP rs2377339 and opiates dependence. This scrutiny required us to include all racial and gender groups. We observed that the MAF and P values vary between different races and genders. The association between rs2377339 and opiates dependence becomes less significant in the overall cohort, after we adjusted race and gender in logistic regression.

4.6. Codependence Association Analysis. In Table 5, we also presented the association results for *NCK2* and comorbidity

Table 4: Association between SNP rs2377339 and opiates dependence by race and sex.

	MAF	*P* value	OR
African-origin men	1.59%	$1.33E-11$	13.87
African-origin women	1.14%	$1.64E-1$	2.82
European-origin men	6.77%	$8.13E-2$	0.55
European-origin women	6.64%	$6.95E-1$	1.14
Combined*	5.04%	$4.37E-1$	1.17

* Logistic regression is used to adjust for sex and race.

Table 5: Association of most significant SNPs in NCK2 with codependence of six individual substances dependence outcomes (*P* value).

Gene	SNP	position	*P* value for unadjusted	*P* value for adjusted
NCK2	rs2377339	105823723	**3.65E − 8**	**2.03E − 8**
	rs6747023	105798288	$5.43E-3$	$2.58E-2$
	rs7589342	105799910	$3.13E-3$	$1.68E-2$
	rs12995333	105802798	$3.81E-3$	$1.83E-2$

of substance dependence. The most significant signal in *NCK2* was observed for SNP rs2377339 in men of African-origin with $P = 3.65 \times 10^{-8}$ in adjusted association test and $P = 2.03 \times 10^{-8}$ in unadjusted association test. *P* values of SNPs in *NCK2* for other ethnicity by gender groups were far from the genome-wide significance level and, hence, are omitted here.

5. Discussion

We found a genome-wide significant association between SNP rs2377339 and opiates dependence in African-origin men. The *NCK2* gene that contains SNP rs2377339 also achieved the genome-wide significance for opiates dependence at the gene level. For the addiction of the other five substances, nicotine dependence had the most significant association but not significant at the genome-wide level.

NCK2, a member of NCK family of adaptor proteins, is reported to be associated with tyrosine-phosphorylated growth factor receptors of their cellular substrate [54]. The association between *NCK2* and nicotine dependence has been suggested in humans [67, 68]. Our finding coupled with those human studies enhances the plausibility of a causality relationship between *NCK2* and drug addiction.

Importantly, about one-fifth of opiates addiction subjects in the African-origin men carried minor allele G of SNP rs2377339, which is more than 10-fold of the frequency in the nonopiates dependence group. This suggested that the minor allele G in SNP rs2377339 potentially elevates the risk for opiates dependence in African-origin men. We acknowledge that our analysis included only 44 African-origin men with opiates dependence. Therefore, it is important and necessary to validate our finding through independent and larger cohort studies. Specifically, there are two possible strategies to validate our finding. The direct approach is to replicate the association between SNP rs2377339 and opiates dependence in a larger cohort. An indirect approach is to evaluate whether SNP rs2377339 is associated with any substance dependence (opiates, alcohol, marijuana, etc.) as presented in Table 2.

A distinction of our analysis is to consider simultaneously multiple substance addictions rather than a single substance. This approach, which is a realistic depiction of substance dependence, confirmed that a novel susceptibility gene, *NCK2* is significantly associated with substance dependence in African-origin men.

This study has several limitations. First, we stratified by ethnicity and sex, which reduced sample sizes and affected the power of our analysis. Nonetheless, the significant associations revealed in African-origin men are consistent with the notion that men may be socially more prone to environmental influences that promote substance use and thus more vulnerable to addiction [46]. Second, for SNP rs2377339, we observed heterogeneous genetic effects, suggesting interactions between race, sex, and the gene, because the association is much weakened after adjusting for race and gender. Such interactions have been suggested in other addiction research [44, 45, 47]. Again, our result further supports the importance to examine interactions among genes, race, and sex in addiction.

Conflict of Interests

The authors declare that they have no conflict of interests.

Authors' Contribution

Zhifa Liu and Xiaobo Guo contributed equally.

Acknowledgments

This work was supported by Grant R01 DA016750-09 from the National Institute on Drug Abuse. Funding support for the Study of Addiction: Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative (GEI) (U01 HG004422). SAGE is one of the genome-wide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392), and the Family Study of Cocaine Dependence (FSCD; R01 DA013423). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, and the NIH Contract "High Throughput

Genotyping for Studying the Genetic Contributions to Human Disease" (HHSN268200782096C). The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1 through dbGaP Accession no. phs000092.v1.p.

References

[1] K. R. Merikangas, M. Stolar, D. E. Stevens et al., "Familial transmission of substance use disorders," *Archives of General Psychiatry*, vol. 55, no. 11, pp. 973–979, 1998.

[2] W. R. True, A. C. Heath, J. F. Scherrer et al., "Interrelationship of genetic and environmental influences on conduct disorder and alcohol and marijuana dependence symptoms," *American Journal of Medical Genetics*, vol. 88, no. 4, pp. 391–397, 1999.

[3] G. E. Uhl and I. Gregory, "Genetic influences in drug abuse," in *Psychopharmacology: The Fourth Generation of Progress*, pp. 1793–1806, 1995.

[4] The World Health Organization, "Substitution maintenance therapy in the management of opioid dependence and HIV/AIDS prevention," Annual Report World Health Organization, 2004.

[5] L. J. Paulozzi, D. S. Budnitz, and Y. Xi, "Increasing deaths from opioid analgesics in the United States," *Pharmacoepidemiology and Drug Safety*, vol. 15, no. 9, pp. 618–627, 2006.

[6] L. J. Bierut, A. Agrawal, K. K. Bucholz et al., "A genome-wide association study of alcohol dependence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 11, pp. 5082–5087, 2010.

[7] X. Chen, K. Cho, B. H. Singer, and H. Zhang, "The nuclear transcription factor PKNOX2 is a candidate gene for substance dependence in European-Origin Women," *PLoS ONE*, vol. 6, no. 1, Article ID e16002, 2011.

[8] J. Clarimon, R. R. Gray, L. N. Williams et al., "Linkage disequilibrium and association analysis of α-synuclein and alcohol and drug dependence in two American Indian populations," *Alcoholism: Clinical and Experimental Research*, vol. 31, no. 4, pp. 546–554, 2007.

[9] D. M. Dick, H. J. Edenberg, X. Xuei et al., "Association of GABRG3 with alcohol dependence," *Alcoholism: Clinical and Experimental Research*, vol. 28, no. 1, pp. 4–9, 2004.

[10] H. J. Edenberg, D. M. Dick, X. Xuei et al., "Variations in GABRA2, encoding the α2 subunit of the GABA a receptor, are associated with alcohol dependence and with brain oscillations," *American Journal of Human Genetics*, vol. 74, no. 4, pp. 705–714, 2004.

[11] H. J. Edenberg and T. Foroud, "The genetics of alcoholism: identifying specific genes through family studies," *Addiction Biology*, vol. 11, no. 3-4, pp. 386–396, 2006.

[12] H. J. Edenberg, J. Wang, H. Tian et al., "A regulatory variation in OPRK1, the gene encoding the κ-opioid receptor, is associated with alcohol dependence," *Human Molecular Genetics*, vol. 17, no. 12, pp. 1783–1789, 2008.

[13] T. Foroud, L. F. Wetherill, T. Liang et al., "Association of alcohol craving with α-synuclein (SNCA)," *Alcoholism: Clinical and Experimental Research*, vol. 31, no. 4, pp. 537–545, 2007.

[14] J. Gelernter, R. Gueorguieva, H. R. Kranzler et al., "Opioid receptor gene (OPRM1, OPRK1, and OPRD1) variants and response to naltrexone treatment for alcohol dependence: results from the VA Cooperative Study," *Alcoholism: Clinical and Experimental Research*, vol. 31, no. 4, pp. 555–563, 2007.

[15] T. Reich, "A genomic survey of alcohol dependence and related phenotypes: results from the Collaborative Study on the Genetics of Alcoholism (COGA)," *Alcoholism: Clinical and Experimental Research*, vol. 20, supplement 8s, pp. 133–137, 1996.

[16] T. Reich, H. J. Edenberg, A. Goate et al. et al., "Genome-wide search for genes affecting the risk for alcohol dependence," *American Journal of Medical Genetics*, vol. 81, no. 3, pp. 207–215, 1998.

[17] J. Song, D. L. Koller, T. Foroud et al., "Association of GABAA receptors and alcohol dependence and the effects of genetic imprinting," *American Journal of Medical Genetics*, vol. 117, no. 1, pp. 39–45, 2003.

[18] B. Tabakoff, L. Saba, M. Printz et al. et al., "Genetical genomic determinants of alcohol consumption in rats and humans," *BMC Biology*, vol. 7, article 70, 2009.

[19] J. C. Wang, A. L. Hinrichs, S. Bertelsen et al., "Functional variants in TAS2R38 and TAS2R16 influence alcohol consumption in high-risk families of African-American origin," *Alcoholism: Clinical and Experimental Research*, vol. 31, no. 2, pp. 209–215, 2007.

[20] K. S. Wang, X. F. Liu, Q. Y. Zhang, Y. Pan, N. Aragam, and M. Zeng, "A meta-analysis of two genome-wide association studies identifies 3 new loci for alcohol dependence," *Journal of Psychiatric Research*, vol. 45, no. 11, pp. 1419–1425, 2011.

[21] H. Zhang, H. R. Kranzler, B. Z. Yang, X. Luo, and J. Gelernter, "The OPRD1 and OPRK1 loci in alcohol or drug dependence: OPRD1 variation modulates substance dependence risk," *Molecular Psychiatry*, vol. 13, no. 5, pp. 531–543, 2008.

[22] G. Kalsi, P. H. Kuo, F. Aliev et al., "A systematic gene-based screen of chr4q22–q32 identifies association of a novel susceptibility gene, DKK2, with the quantitative trait of alcohol dependence symptom counts," *Human Molecular Genetics*, vol. 19, no. 12, pp. 2497–2506, 2010.

[23] G. Kalsi, P. H. Kuo, F. Aliev et al., "A systematic gene-based screen of chr4q22–q32 identifies association of a novel susceptibility gene, DKK2, with the quantitative trait of alcohol dependence symptom counts," *Human Molecular Genetics*, vol. 19, no. 20, pp. 4121–2506, 2010.

[24] P. H. Kuo, G. Kalsi, C. A. Prescott et al., "Association of ADH and ALDH genes with alcohol dependence in the Irish affected sib pair study of alcohol dependence (IASPSAD) sample," *Alcoholism: Clinical and Experimental Research*, vol. 32, no. 5, pp. 785–795, 2008.

[25] L. J. Zuo, J. Gelernter, C. K. Zhang et al. et al., "Genome-wide association study of alcohol dependence implicates IAA0040 on chromosome 1q," *Neuropsychopharmacology*, vol. 37, no. 2, pp. 557–566, 2012.

[26] L. J. Bierut, "Genetic variation that contributes to nicotine dependence," *Pharmacogenomics*, vol. 8, no. 8, pp. 881–883, 2007.

[27] L. J. Bierut, P. A. F. Madden, N. Breslau et al., "Novel genes identified in a high-density genome wide association study for nicotine dependence," *Human Molecular Genetics*, vol. 16, no. 1, pp. 24–35, 2007.

[28] N. Caporaso, F. Gu, N. Chatterjee et al., "Genome-wide and candidate gene association study of cigarette smoking behaviors," *PLoS ONE*, vol. 4, no. 2, Article ID e4653, 2009.

[29] L. S. Chen, E. O. Johnson, N. Breslau et al., "Interplay of genetic risk factors and parent monitoring in risk for nicotine dependence," *Addiction*, vol. 104, no. 10, pp. 1731–1740, 2009.

[30] N. L. Saccone, S. F. Saccone, A. L. Hinrichs et al., "Multiple distinct risk loci for nicotine dependence identified by dense coverage of the complete family of nicotinic receptor subunit (CHRN) genes," *American Journal of Medical Genetics B*, vol. 150, no. 4, pp. 453–466, 2009.

[31] S. F. Saccone, A. L. Hinrichs, N. L. Saccone et al., "Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs," *Human Molecular Genetics*, vol. 16, no. 1, pp. 36–49, 2007.

[32] G. R. Uhl, T. Drgon, C. Johnson et al., "Genome-wide association for smoking cessation success: participants in the patch in practice trial of nicotine replacement," *Pharmacogenomics*, vol. 11, no. 3, pp. 357–367, 2010.

[33] G. R. Uhl, Q. R. Liu, T. Drgon et al., "Molecular genetics of successful smoking cessation: convergent genome-wide association study results," *Archives of General Psychiatry*, vol. 65, no. 6, pp. 683–693, 2008.

[34] R. B. Weiss, T. B. Baker, D. S. Cannon et al., "A candidate gene approach identifies the CHRNA5-A3-B4 region as a risk factor for age-dependent nicotine addiction," *PLoS Genetics*, vol. 4, no. 7, Article ID e1000125, 2008.

[35] I. Deb, J. Chakraborty, P. K. Gangopadhyay, S. R. Choudhury, and S. Das, "Single-nucleotide polymorphism (A118G) in exon 1 of OPRM1 gene causes alteration in downstream signaling by mu-opioid receptor and may contribute to the genetic risk for addiction," *Journal of Neurochemistry*, vol. 112, no. 2, pp. 486–496, 2010.

[36] S. Kapur, S. Sharad, R. A. Singh, and A. K. Gupta, "A118G polymorphism in mu opioid receptor gene (OPRM1): association with opiate addiction in subjects of Indian origin," *Journal of Integrative Neuroscience*, vol. 6, no. 4, pp. 511–522, 2007.

[37] M. J. Kreek, G. Bart, C. Lilly, K. S. Laforge, and D. A. Nielsen, "Pharmacogenetics and human molecular genetics of opiate and cocaine addictions and their treatments," *Pharmacological Reviews*, vol. 57, no. 1, pp. 1–26, 2005.

[38] G. Gerra, C. Leonardi, E. Cortese et al., "Human kappa opioid receptor gene (OPRK1) polymorphism is associated with opiate addiction," *American Journal of Medical Genetics B*, vol. 144, no. 6, pp. 771–775, 2007.

[39] V. Yuferov, D. Fussell, K. S. LaForge et al., "Redefinition of the human kappa opioid receptor gene (OPRK1) structure and association of haplotypes with opiate addiction," *Pharmacogenetics*, vol. 14, no. 12, pp. 793–804, 2004.

[40] D. Proudnikov, K. S. LaForge, H. Hofflich et al., "Association analysis of polymorphisms in serotonin 1B receptor (HTR1B) gene with heroin addiction: a comparison of molecular and statistically estimated haplotypes," *Pharmacogenetics and Genomics*, vol. 16, no. 1, pp. 25–36, 2006.

[41] G. Gerra, L. Garofano, G. Santoro et al., "Association between low-activity serotonin transporter genotype and heroin dependence: behavioral and personality correlates," *American Journal of Medical Genetics*, vol. 126, no. 1, pp. 37–42, 2004.

[42] E. W. Loh, N. L. S. Tang, D. T. S. Lee, S. I. Liu, and A. Stadlin, "Association analysis of GABA receptor subunit genes on 5q33 with heroin dependence in a Chinese male population," *American Journal of Medical Genetics B*, vol. 144, no. 4, pp. 439–443, 2007.

[43] C. Y. Cheng, C. J. Hong, Y. W. Y. Yu, T. J. Chen, H. C. Wu, and S. J. Tsai, "Brain-derived neurotrophic factor (Val66Met) genetic polymorphism is associated with substance abuse in males," *Molecular Brain Research*, vol. 140, no. 1-2, pp. 86–90, 2005.

[44] Z. Luo, G. F. Alvarado, D. K. Hatsukami, E. O. Johnson, L. J. Bierut, and N. Breslau, "Race differences in nicotine dependence in the Collaborative Genetic study of Nicotine Dependence (COGEND)," *Nicotine and Tobacco Research*, vol. 10, no. 7, pp. 1223–1230, 2008.

[45] N. Breslau, E. O. Johnson, E. Hiripi, and R. Kessler, "Nicotine dependence in the United States: prevalence, trends, and smoking persistence," *Archives of General Psychiatry*, vol. 58, no. 9, pp. 810–816, 2001.

[46] D. M. Hartel, E. E. Schoenbaum, Y. Lo, and R. S. Klein, "Gender differences in illicit substance use among middle-aged drug users with or at risk for HIV infection," *Clinical Infectious Diseases*, vol. 43, no. 4, pp. 525–531, 2006.

[47] D. L. Karch, L. Barker, and T. W. Strine, "Race/ethnicity, substance abuse, and mental illness among suicide victims in 13 US states: 2004 data from the national violent death reporting system," *Injury Prevention*, vol. 12, no. 2, pp. 22–27, 2006.

[48] E. E. Eichler, J. Flint, G. Gibson et al., "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 446–450, 2010.

[49] T. A. Manolio, F. S. Collins, N. J. Cox et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[50] X. Guo, Z. Liu, X. Wang, and H. Zhang, "Genetic association test for multiple traits at gene level," *Genetic Epidemiology*, vol. 37, no. 1, pp. 122–129, 2013.

[51] M. X. Li, H. S. Gui, J. S. H. Kwan, and P. C. Sham, "GATES: a rapid and powerful gene-based association test using extended Simes procedure," *American Journal of Human Genetics*, vol. 88, no. 3, pp. 283–293, 2011.

[52] J. Z. Liu, A. F. McRae, D. R. Nyholt et al., "A versatile gene-based test for genome-wide association studies," *American Journal of Human Genetics*, vol. 87, no. 1, pp. 139–145, 2010.

[53] B. M. Neale and P. C. Sham, "The future of association studies: gene-based analysis and replication," *American Journal of Human Genetics*, vol. 75, no. 3, pp. 353–362, 2004.

[54] G. M. Rivera, S. Antoku, S. Gelkop et al., "Requirement of Nck adaptors for actin dynamics and cell migration stimulated by platelet-derived growth factor B," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 25, pp. 9536–9541, 2006.

[55] H. Begleiter, T. Reich, V. Hesselbrock et al., "The collaborative study on the genetics of alcoholism," *Alcohol Health & Research World*, vol. 19, no. 3, pp. 228–236, 1995.

[56] L. J. Bierut, J. R. Strickland, J. R. Thompson, S. E. Afful, and L. B. Cottler, "Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings," *Drug and Alcohol Dependence*, vol. 95, no. 1-2, pp. 14–22, 2008.

[57] X. Guo, Z. Liu, X. Wang, and H. Zhang, "Large scale association analysis for drug addiction: results from SNP to gene," *The Scientific World Journal*, vol. 2012, Article ID 939584, 6 pages, 2012.

[58] S. Purcell, B. Neale, K. Todd-Brown et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[59] I. Menashe, J. D. Figueroa, M. Garcia-Closas et al. et al., "Large-scale pathway-based analysis of bladder cancer genome-wide association data from five studies of European background," *Plos One*, vol. 7, no. 1, Article ID e29396, 2012.

[60] A. Christoforou, M. Dondrup, M. Mattingsdal et al. et al., "Linkage-disequilibrium-based binning affects the interpretation of GWASs," *American Journal of Human Genetics*, vol. 90, no. 4, pp. 727–733, 2012.

[61] F. Dudbridge and A. Gusnanto, "Estimation of significance thresholds for genomewide association scans," *Genetic Epidemiology*, vol. 32, no. 3, pp. 227–234, 2008.

[62] H. Zhang, C. T. Liu, and X. Wang, "An association test for multiple traits based on the generalized Kendall's tau," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 473–481, 2010.

[63] D. Rabinowitz and N. Laird, "A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information," *Human Heredity*, vol. 50, no. 4, pp. 211–223, 2000.

[64] Y. Jiang and H. Zhang, "Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder," *Genetic Epidemiology*, vol. 35, no. 2, pp. 125–132, 2011.

[65] H. Zhao, T. R. Rebbeck, and N. Mitra, "A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors," *Genetic Epidemiology*, vol. 33, no. 8, pp. 679–690, 2009.

[66] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics*, vol. 21, no. 2, pp. 263–265, 2005.

[67] M. Akiyama, K. Yatsu, M. Ota et al., "Microsatellite analysis of the GLC1B locus on chromosome 2 points to NCK2 as a new candidate gene for normal tension glaucoma," *British Journal of Ophthalmology*, vol. 92, no. 9, pp. 1293–1296, 2008.

[68] S. Bonovas, K. Filioussi, A. Tsantes, and V. Peponis, "Epidemiological association between cigarette smoking and primary open-angle glaucoma: a meta-analysis," *Public Health*, vol. 118, no. 4, pp. 256–261, 2004.