Lecture Notes in Networks and Systems 325

Miguel Rocha
Florentino Fdez-Riverola
Mohd Saberi Mohamad
Roberto Casado-Vara *Editors*

Practical Applications of Computational Biology & Bioinformatics, 15th International Conference (PACBB 2021)

Springer

Lecture Notes in Networks and Systems

Volume 325

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at http://www.springer.com/series/15179

Miguel Rocha · Florentino Fdez-Riverola ·
Mohd Saberi Mohamad ·
Roberto Casado-Vara
Editors

Practical Applications of Computational Biology & Bioinformatics, 15th International Conference (PACBB 2021)

Springer

Editors

Miguel Rocha
Department de Informática
Universidade do Minho
Braga, Portugal

Mohd Saberi Mohamad
Department of Genetics and Genomics
United Arab Emirates University
Abu Dhabi, United Arab Emirates

Florentino Fdez-Riverola
Superior de Ingeniería Informática
Universidade de Vigo, Escuela
Ourense, Spain

Roberto Casado-Vara
BISITE, Digital Innovation Hub
University of Salamanca
Salamanca, Salamanca, Spain

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-030-86257-2 ISBN 978-3-030-86258-9 (eBook)
https://doi.org/10.1007/978-3-030-86258-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The success of bioinformatics in recent years has been prompted by research in molecular biology and molecular medicine in several initiatives. These initiatives gave rise to an exponential increase in the volume and diversification of data, including nucleotide and protein sequences and annotations, high-throughput experimental data, biomedical literature, among many others. Systems biology is a related research area that has been replacing the reductionist view that dominated biology research in the last decades, requiring the coordinated efforts of biological researchers with those related to data analysis, mathematical modeling, computer simulation and optimization.

The accumulation and exploitation of large-scale databases prompt for new computational technology and for research into these issues. In this context, many widely successful computational models and tools used by biologists in these initiatives, such as clustering and classification methods for gene expression data, are based on computer science/artificial intelligence (CS/AI) techniques. In fact, these methods have been helping in tasks related to knowledge discovery, modeling and optimization tasks, aiming at the development of computational models so that the response of biological complex systems to any perturbation can be predicted. The 15th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB) aims to promote the interaction among the scientific community to discuss the applications of CS/AI with an interdisciplinary character, exploring the interactions between sub-areas of CS/AI, bioinformatics, chemoinformatic and systems biology. The PACBB'21 technical program includes 17 papers of authors from many different countries (Australia, Colombia, Egypt, Germany, India, Malaysia, Portugal, Saudi Arabia, Slovakia, South Korea, Spain, Switzerland, Turkey, United Arab Emirates, UK and USA) and different subfields in bioinformatics and computational biology. There will be special issues in JCR-ranked journals, such as Interdisciplinary Sciences: Mathematical Biosciences and Engineering, Integrative Bioinformatics, Information Fusion, Neurocomputing, Sensors, Processes and Electronics. Therefore, this event will strongly promote interaction among researchers from international research groups working in diverse fields. The scientific content will be innovative, and it will help improve the valuable work that is being carried out by the participants.

This symposium is organized by the University of Salamanca with the collaboration of the United Arab Emirates University, the University of Minho and the University of Vigo. We would like to thank all the contributing authors, the members of the program committee and the sponsors IBM, Indra, AEPIA, APPI, AIIS, EurAI and AIR Institute. We thank for funding support to the project: “Intelligent and sustainable mobility supported by multi-agent systems and edge computing” (Id. RTI2018-095390-B-C32), and finally, we thank the local organization members and the program committee members for their valuable work, which is essential for the success of PACBB’21.

Miguel Rocha
Florentino Fdez-Riverola
Mohd Saberi Mohamad
Roberto Casado-Vara

Organization

Program Committee Chairs

Mohd Saberi Mohamad	United Arab Emirates University, United Arab Emirates
Miguel Rocha	University of Minho, Portugal

General Co-chairs

Florentino Fdez-Riverola	University of Vigo, Spain
Roberto Casado Vara	University of Salamanca, Spain

Advisory Committee

Grabriella Panuccio	Istituto Italiano di Tecnologia, Italy

Organizing Committee

Juan M. Corchado Rodríguez	University of Salamanca, Spain AIR Institute, Spain
Roberto Casado Vara	University of Salamanca, Spain
Fernando De la Prieta	University of Salamanca, Spain
Sara Rodríguez González	University of Salamanca, Spain
Javier Prieto Tejedor	University of Salamanca, Spain AIR Institute, Spain
Pablo Chamoso Santos	University of Salamanca, Spain
Belén Pérez Lancho	University of Salamanca, Spain
Ana Belén Gil González	University of Salamanca, Spain
Ana De Luis Reboredo	University of Salamanca, Spain
Angélica González Arrieta	University of Salamanca, Spain

Emilio S. Corchado Rodríguez	University of Salamanca, Spain
Alfonso González Briones	University of Salamanca, Spain
Yeray Mezquita Martín	University of Salamanca, Spain
Javier J. Martín Limorti	University of Salamanca, Spain
Alberto Rivas Camacho	University of Salamanca, Spain
Elena Hernández Nieves	University of Salamanca, Spain
Beatriz Bellido	University of Salamanca, Spain
María Alonso	University of Salamanca, Spain
Diego Valdeolmillos	AIR Institute, Spain
Sergio Marquez	University of Salamanca, Spain
Marta Plaza Hernández	University of Salamanca, Spain
David García Retuerta	University of Salamanca, Spain
Guillermo Hernández González	AIR Institute, Spain
Ricardo S. Alonso Rincón	University of Salamanca, Spain
Javier Parra	University of Salamanca, Spain

Program Committee

Vera Afreixo	University of Aveiro, Portugal
Manuel Álvarez Díaz	University of A Coruña, Spain
Carlos Bastos	University of Aveiro, Portugal
Lourdes Borrajo	University of Vigo, Spain
Ana Cristina Braga	University of Minho, Portugal
Fernanda Brito Correia	DEIS/ISEC/Polytechnic Institute of Coimbra, Portugal
Rui Camacho	University of Porto, Portugal
Angel Canal	University of Salamanca, Spain
Roberto Casado Vara	University of Salamanca, Spain
Yingbo Cui	National University of Defense Technology, China
Fernando De La Prieta	University of Salamanca, Spain
Sergio Deusdado	IPB-Polytechnic Institute of Bragança, Portugal
Oscar Dias	University of Minho, Portugal
Florentino Fdez-Riverola	University of Vigo, Spain
João Diogo Ferreira	University of Lisbon, Faculty of Sciences, Portugal
Nuno Filipe	University of Porto, Portugal
Nuno A. Fonseca	University of Porto, Portugal
Dino Franklin	Federal University of Uberlandia, Spain
Narmer Galeano	Universidad Catolica de Manizales, Colombia
Rosalba Giugno	University of Verona, Italy
Gustavo Isaza	University of Caldas, Colombia

Paula Jorge	IBBCEB Center of Biological Engineering, Portugal
Rosalia Laza	Universidad de Vigo, Spain
Thierry Lecroq	University of Rouen, France
Giovani Librelotto	Universidade Federal de Santa Maria, Brazil
Hugo López-Fernández	Instituto de Investigação e Inovação em Saúde, i3S, Portugal
Eva Lorenzo Iglesias	University of Vigo, Spain
Marcos Martínez-Romero	Stanford University, USA
Mohd Saberi Mohamad	United Arab Emirates University, United Arab Emirates
Loris Nanni	University of Padua, Italy
José Luis Oliveira	University of Aveiro, Portugal
Joel P. Arrais	University of Coimbra, Portugal
Cindy Perscheid	Hasso Plattner Institute, Germany
Armando Pinho	University of Aveiro, Portugal
Ignacio Ponzoni	Planta Piloto de Ingeniería Química, PLAPIQUI-UNS-CONICET, Argentina
Miguel Reboiro-Jato	University of Vigo, Spain
Jose Ignacio Requeno	Western Norway University of Applied Sciences, HVL, Norway
Miguel Rocha	Center for Computer Science and Technologies, CCTC, University of Minho, Portugal
João Manuel Rodrigues	DETI/IEETA, University of Aveiro, Portugal
Gustavo Santos-Garcia	Universidad de Salamanca, Spain
Ana Margarida Sousa	University of Minho, Portugal
Niclas Ståhl	University of Skövde, Sweden
Carolyn Talcott	SRI International, USA
Rita Margarida Teixeira Ascenso	ESTG-IPL, Portugal
Antonio J. Tomeu-Hardasmal	University of Cadiz, Spain
Alicia Troncoso	Universidad Pablo de Olavide, Spain
Eduardo Valente	IPCB, Portugal
Alejandro F. Villaverde	Instituto de Investigaciones Marinas (C.S.I.C.), Spain
Pierpaolo Vittorini	University of L'Aquila, Department of Life, Health and Environmental Sciences, Portugal

PACBB 2021 Sponsors

Sponsors

Organizers

Support from National Associations

Contents

Computational Methods for the Identification of Genetic Variants in Complex Diseases

Débora Antunes[1(✉)], Daniel Martins[2,3], Fernanda Correia[4], Miguel Rocha[1], and Joel P. Arrais[2]

[1] Department of Informatics, University of Minho, Braga, Portugal
mrocha@di.uminho.pt
[2] CISUC, University of Coimbra, Coimbra, Portugal
jpa@dei.uc.pt
[3] CNC, University of Coimbra, Coimbra, Portugal
[4] ISEC, Polytechnic Institute of Coimbra, Coimbra, Portugal
fernanda@isec.pt

Abstract. Complex diseases, as Type 2 Diabetes, arise from dysfunctional complex biological mechanisms, caused by multiple variants on underlying groups of genes, combined with lifestyle and environmental factors. Thus far, the known risk factors are not sufficient to predict the manifestation of the disease. Genome-Wide Association Studies (GWAS) data were used to test for genotype-phenotype associations and were combined with a network-based analysis approach. Three datasets of genes associated with this disease were built and features were extracted for each of these genes. Machine learning models were employed to develop a predictor of the risk associated with Type 2 Diabetes to help the identification of new genetic markers associated with the disease. The obtained results highlight that the use of gene regions and protein-protein interaction networks can identify new genes and pathways of interest and improve the model performance, providing new possible interpretation for the biology of the disease.

1 Introduction

Complex diseases are conditions influenced by mutations in a group of genes that interact with each other and environmental factors. Contrary to the case of monogenic disorders, the genes associated to complex diseases do not have any effect individually, hindering their identification [7].

The case study for this work was the complex disease Type 2 Diabetes (T2D), a subtype of diabetes that accounts for 90% of diabetes worldwide. In this subtype of diabetes the cells of the organism cannot respond to an essential hormone called insulin, leading not only to high blood sugar levels (hyperglycemia) but also to an increase in insulin production. According to Stančáková et al. [11], until 2016, more than 80 variants were associated with this condition, mostly

M. Rocha et al. (Eds.): PACBB 2021, LNNS 325, pp. 1–10, 2022.
https://doi.org/10.1007/978-3-030-86258-9_1

through Genome-Wide Association Studies (GWAS) and considering independent effects. However, those variants only explained about 10% of the T2D variability within a population, producing little information that can be used in a medical context.

GWAS finds genetic variations associated with a particular disease by scanning sets of DNA or genomes of many individuals. Frequently, they are focused in Single-Nucleotide Polymorphisms (SNPs) and phenotypes, inside a population, sequencing the genotypes of tag-SNPs, representative of a haplotype, a group of genes that are inherited together. The identified genetic associations can be used to predict, detect and treat the disease risk and to produce knowledge about the underlying biological entities and processes. Such studies are particularly useful in finding genetic variations that contribute to complex diseases like T2D [12]. However, their relationships are not easy to understand because of their complex pathways and growing number of variables [8].

In single SNP association analysis, an association between each SNP and the phenotype is tested. In some cases, the SNPs are treated as independent and some of the methods used to study this association are generalised linear models, statistical tests, like χ^2 test, and Bayesian approaches [8]. Multiple SNP association analysis examines the relationship between the phenotype and the combined effect of multiple SNPs. Different types of analysis have been proposed to account for these associations: haplotype-based methods, SNP-SNP interaction models and models based in biological knowledge [8]. Machine Learning (ML) approaches were used to analyse the SNPs produced in GWAS and provide information about the relation between variants and phenotypes in complex diseases [3].

This paper describes a new proposed complex disease predictor of the risk associated with T2D complex disease that allows the identification of new genetic variants associated with the disease. The pipeline used GWAS data, grouped them into gene regions and applied a network-based analysis approach. Using these methods, new subsets of genes were defined, their most relevant features were selected and ML techniques were applied to predict the risk of T2D.

2 Methods

2.1 Data Preparation

The first step involved choosing and preparing the datasets. The Final dataset construction involved two initial datasets, Case and Control. Information about the sizes of the datasets at each stage of the dataset construction is summarised in Table 1.

The Case dataset originated from a privately owned gzip compressed Variant Call Format (VCF) file that contained information of exomes from 71 Portuguese patients diagnosed with T2D, and 57,142,453 loci. Since the focus of the study were the genetic factors, only the genomic data was used. An initial filtering restricted the type of variants to Insertion/Deletions (INDELs) or SNPs. The quality control of these data revealed that from the 71 individuals, two of them were related. For this reason, the one with more missing variants was excluded from the study. Also, the variants that did not follow the Hardy-Weinberg Equilibrium (HWE) theorem or that had less than 20 of quality score were removed.

The Control dataset resulted from a selection of VCF files collected from the Iberian Populations in Spain (IBS) in the Phase 3 release of the 1000 Genome Project [1]. This group was chosen to prevent bias that would tend to distinguish cases and controls, based on population genetic divergences.

To create a dataset with all the information, the two previous datasets were merged into the Merged dataset, by finding the common variants. A variant was considered equal if their location (chromosome and position) was the same and their REF and ALT allele were matching. To produce the Final dataset, the variants with a rate of missing genotypes of more than 10% were removed (Table 1).

Table 1. Summary of the number of samples and variants on the different datasets.

	Case dataset[a]	Control dataset	Merged dataset	Final dataset
Number of samples	70	107	177	177
Number of variants	228,301[b]	81,404,605	172,629	168,715

[a] After preprocessing; [b] Includes 225 070 SNPs and 3 231 INDELs

2.2 Gene Selection

The first analysis made was a single SNP association. For that, we used the χ^2 test, which measures the probability of association of each variant with the disease considering each one of them independent. The statistical association between the variants and the phenotype using the χ^2 test with a p-value of 0.05, showed that of 168,715 variants, 9427 (5.6%) presented a statistical association with the phenotype

Three sets of genes were selected and used to build different datasets for this study, the dataset of Known Risk Genes, the dataset of Significant Genes and the dataset of Central Genes.

To select these sets of genes, the position of each variant was associated with a gene ID using the python package *pyensembl* and a reference GTF file from the GRCh37 version of Ensembl [13]. The variants were grouped by these gene IDs and the associated gene p-value was calculated as the average of p-values of the set of variants. The final list contained 16,513 genes and the respective p-values. The set of Known Risk Genes were the genes on the final list that matched the list of 75 known risk genes from Type 2 Diabetes Knowledge Portal [10] (the T2D_related and CAUSAL genes). In total, 67 Known Risk Genes were selected (Table 2). The set of Significant Genes included 82 genes with a p-value of less than 0.05 (Table 2).

To select the Central Genes a Protein-Protein Interaction (PPI) network was built. This PPI network can be represented by a graph, whose nodes are genes (proteins) and edges are their interactions. The PPI network file was downloaded from BioGRID [9], and after a prepossessing that selected the interactions with experimental data associated and with both interactors being human proteins, the file had 366,327 interactions and included 17,940 proteins. The PPIs file and the final list of genes (proteins) were used as input for the R package *dmGWAS*. This tool implements a dense module searching method and outputs a list of modules associated with the disease, ranked by significance. In this study, the top 50 modules were extracted and combined into a subnetwork of significant PPIs containing 252 genes. Using the R package *igraph*, three network metrics (degree, betweenness and closeness), that measured the centrality of each gene of this subnetwork, were applied. Choosing the genes that were in the top 100 of each metric, 77 genes were selected for the set of Central Genes. From these 77 genes, three were in common with the Known Risk Genes, namely CAV1, PCBD1 and WFS1 (in bold in Table 2).

2.3 Feature Extraction and Reduction

At this point, there were three sets of genes selected, the 67 known risk genes, the 82 significant genes and the 77 central genes. First, feature extraction was applied to each gene, using the information present in the corresponding group of variants. Four features were extracted for each gene, the first component after applying the Principal Component Analysis (PCA), the first component after applying t-distributed Stochastic Neighbor Embedding (t-SNE) and two statistical measures, the mean and variance. The Known Risk Genes, Significant Genes and Central Genes datasets had 268, 328 and 308 features, respectively.

Feature reduction was performed in two of these datasets, the Significant Genes and the Central Genes datasets. For each dataset, 1000 Extremely Randomized Trees (Extra-Trees) models were trained. In every training cycle, the top 100 most important features were registered and, at the end, their frequencies were calculated. The 25 features with higher frequency were selected, for each dataset. The 25 features from the Top 25 Significant Genes dataset belonged to 25 different genes, while the ones from the Top 25 Central Genes dataset belonged to 12 different genes (in grey boxes in Table 2).

Three machine learning models, Support Vector Machines (SVM), Decision Tree and Logistic Regression, were trained for the following genes datasets: Significant Genes, Top 25 Significant Genes, Known Risk Genes, Central Genes and Top 25 Central Genes. The classifiers were run 1000 times using a 5-fold cross-validation. A grid search was performed for each dataset and the overall best parameters were selected and used for the study. The final parameters used are shown in Table 3. For the evaluation of the models, three metrics were used, Accuracy, F1-score and Area Under Curve (AUC).

Table 2. Known Risk Genes, Significant Genes and Central Genes selected for this study. In the grey boxes are the genes selected in the dimensionality reduction. In bold are the common genes between the Known Risk Genes the Central Genes lists.

Known risk genes		Significant genes		Central genes	
ABCC8	PCSK1	AAMDC	MAST1	APP	MYC
AKT2	PDX1	AKT1S1	MSTN	ATXN1	NCL
ANGPTL4	PLCB3	AMMECR1L	MYPOP	BAG3	NEK6
ANKH	PNPLA3	ANP32A	NBPF14	BAIAP2	NFKBIA
APOE	POC5	B3GALNT1	NBPF4	BTRC	NSMF
APPL1	POLD1	BCL2L10	NDUFB6	CALM1	OPTN
BLK	PPARG	C10orf95	NGLY1	CASP1	**PCBD1**
BSCL2	PPP1R15B	C1orf162	NKX2-1	CASP8	PCNA
CAV1	PTF1A	C20orf202	NUFIP2	**CAV1**	PICK1
CDKN1B	QSER1	CDKN2C	OLIG1	CDC37	PIK3R1
CEL	RFX6	CGB5	OR13J1	CDH1	PIN1
EIF2AK3	RREB1	CHKA	OR2T5	CDK2	PLK1
ERAP2	SIX2	CLEC18A	P4HTM	CDKN1A	PPP1CA
GATA4	SIX3	CNOT11	PAGR1	CEP70	PTPN6
GATA6	SLC16A11	CSRP2	PARP11	DEAF1	RAC1
GCG	SLC19A2	CXCL13	PNMT	DISC1	RPS6KB1
GCK	SLC30A8	CXCL5	PNRC2	ENO1	SFN
GCKR	SLC5A1	DCAF16	POTED	ERBB2	SKP1
GIPR	TBC1D4	DDTL	PPP1R7	ESR1	SMAD3
GLIS3	TM6SF2	DEXI	PRAMEF13	GFAP	SPRED1
GLP1R	TRMT10A	DLEU1	PRDX6	GRB2	STK11
GRB10	WARS	DLX6	PRRT2	HLA-B	STX1A
HNF1A	**WFS1**	DOK1	PTRF	HNRNPC	SYK
HNF1B	WSCD2	GLIPR1	RAX	HSP90AB1	TGFBR2
HNF4A	ZFP57	GPR25	RNASE10	HSPA8	TNF
IGF1	ZNF771	HEXIM2	RNF182	HSPD1	TRAF6
IRS2		HFE2	RYBP	HTT	TRIM54
KCNJ11		HMOX2	S100A16	INCA1	TSC22D1
KLF11		HNRNPAB	SCG5	IQUB	UBC
LPL		HOXB8	SKIL	JPH3	UBE2Z
MC4R		HSD3B1	SOX21	KANK2	USP2
MNX1		ID2	SYT4	KCTD13	VCP
MTNR1B		IFNA13	TADA2B	KDR	**WFS1**
NAT2		IL33	TAF11	KIFC3	YWHAE
NEUROD1		IL36RN	TEX22	KRT34	YWHAG
NEUROG3		JOSD1	TLX1NB	LMO4	YWHAZ
NKX2-2		KCNMB4	TMEM178A	LNX1	
PAM		KRTAP5-1	TMEM60	LNX2	
PAX4		LSMEM1	TPST1	MAP3K1	
PAX6		MAFA	TRAM1L1	MDFI	
PCBD1		MAFF	WDR45B	MEOX2	

3 Results and Discussion

The obtained results are illustrated graphically in Fig. 1.

The Decision Tree models produced higher values for all statistical measures ($\geq$0.87) compared to Logistic Regression models and SVM models, which shows that this model could better address the complexity of the data. Both Regression models and SVM models used linear functions for the classification, which indicates that these functions have a higher difficulty in explaining the underlying structure of the data. Also, the values of each metric were similar, which shows the robustness of the results.

The lower values obtained when using the Known Risk Genes as input was expected because it is known that these gene associations do not account for a high percentage of the heritability. The Significant Genes dataset had features that were extracted directly from the most significant genes of the original dataset and, as expected, produced good results. When just the top 25 features were used, which had less than 8% of the full dataset's size, the three metrics kept relatively good values. Lower values in the results from the central genes dataset were expected, given that the features were extracted from genes central to a network of significance and generally not significant themselves. Although the values from Accuracy, F1-score and AUC were lower than the values from the Significant Genes dataset, the results were still good. The best results were from the Top 25 Central Genes dataset, which had less than 9% of the full dataset's size. With only the 25 features of this dataset it was possible to predict the risk of disease with a good degree of success.

To add biological context to the genes, a functional annotation of the Known Risk Genes, the Significant Genes and the Central Genes was conducted, using the online platform Database for Annotation, Visualization and Integrated Discovery (DAVID) [5,6]. This platform finds the most relevant and over-represented biological terms related to the gene lists provided. The results from the biological processes annotation from Gene Ontology (GO) revealed that from 29,683 biological process terms, 160 were found to be terms in common between the genes of the three lists. Knowing that most of the genes are different across the three lists, it is observable that many of the identified genes (either Significant Genes or Central Genes) share the same terms as the Known Risk Genes (Fig. 2).

Table 3. Parameters and respective values chosen for SVM, Decision Tree and Logistic Regression models after grid search.

SVM		Decision tree		Logistic regression	
kernel	linear	n_estimators	50	penalty	l1
C	0.25	criterion	entropy	C	0.4
tol	1e−3	min_samples_leaf	3	tol	1e−3
gamma	25	min_samples_split	5	solver	liblinear
degree	1	max_leaf_nodes	50	–	–