Advances in Intelligent Systems and Computing 232

Luis F. Castillo • Marco Cristancho
Gustavo Isaza • Andrés Pinzón
Juan Manuel Corchado Rodríguez
Editors

Advances in Computational Biology

Proceedings of the 2nd Colombian Congress on Computational Biology and Bioinformatics (CCBCOL)

Springer

Advances in Intelligent Systems and Computing

Volume 232

Series Editor

Janusz Kacprzyk, Warsaw, Poland

For further volumes:
http://www.springer.com/series/11156

Luis F. Castillo · Marco Cristancho
Gustavo Isaza · Andrés Pinzón
Juan Manuel Corchado Rodríguez
Editors

Advances in Computational Biology

Proceedings of the 2nd Colombian Congress on Computational Biology and Bioinformatics (CCBCOL)

Springer

Editors
Luis F. Castillo
University of Caldas
Manizales
Colombia

Marco Cristancho
Cenicafé - Centro Nacional de
Investigaciones del Café en
Colombia

Gustavo Isaza
University of Caldas
Manizales
Colombia

Andrés Pinzón
BIOS - Centro Bioinformática y Biologia
Computacional de Colombia
Manizales
Colombia

Juan Manuel Corchado Rodríguez
Department of Computer Science
School of Science
University of Salamanca
Salamanca
Spain

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-01567-5 ISBN 978-3-319-01568-2 (eBook)
DOI 10.1007/978-3-319-01568-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013944537

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Bioinformatics and Computational Biology are areas of knowledge that have emerged due to advances that have taken place in the Biological Sciences and its integration with Information Sciences. The expansion of projects involving the study of genomes has led the way in the production of vast amounts of sequence data which needs to be organized, analyzed and stored to understand phenomena associated with living organisms related to their evolution, behavior in different ecosystems, and the development of applications that can be derived from this analysis.

In Colombia the main reference to consider for the advancement of Science is the National Development Plan 2010-2014, which is based on what the government calls, the locomotives of growth. The areas or economic activities to be included that have been selected with priority in the four years' duration are (i) new innovation-based sectors, (ii) agriculture and rural development, (iii) Housing and friendly cities, (iv) energy mining development and expansion, (v) transport infrastructure.

The first locomotive is focused on the need to promote the development of emerging sectors based on innovation, which features information technology and telecommunications, and biotechnology, among others. Several strategies have been proposed to generate knowledge that can be applied to production processes and the solution of problems affecting the community. This requires the formation of human resources, project financing and organization of institutions to promote research and innovation.

Bioinformatics becomes a crucial area of development in the National policy "COMMERCIAL DEVELOPMENT OF BIOTECHNOLOGY FROM THE SUSTAINABLE USE OF BIODIVERSITY", because it can support the process of search and discovery of molecules, genes or active ingredients that are present in our biodiversity, so through biotechnology they can be industrially produced in a sustainable scheme.

We can envision Colombia's effort to strengthen this field of research, on data from the Biodiversity Information System of Colombia - SiB, analyzed in the Report on the State of Renewable Natural Resources and the Environment

2009, which describes the departments with the largest number of known species and include Quindío, Risaralda, Caldas, Cundinamarca, Valle, Antioquia and Boyacá, located in the central Andean region.

The CCBCOL'13 Congress has the following objectives: Submit progress in research in computational biology and related areas and their relations and international scope, identify strengths and weaknesses in relation to infrastructure, research training and development strategies of computational biology in Colombia, advance the establishment of agreements that allow the integration of infrastructure, cooperation and development of research projects relevant and competitive, nationally and internationally, advance the establishment of agreements that allow the integration of infrastructure, cooperation and development of research projects relevant and competitive, nationally and internationally, encourage contact between scientists from multiple disciplines (computer science, biology, mathematics, statistics, chemistry, etc.) that conduct research in computational biology and related areas in the country, and launch the Colombian Society for Computational Biology.

Another important achievement and articulated with this event, is the founding of the Colombia's Computational Biology and Bioinformatics Center (BIOS), which is headquartered in the city of Manizales (the central Colombian coffee production zone), a leading national supercomputing facility devoted to providing services to government, academia and businesses interested in Biotechnology research, Development and Bioprospecting.

This volume compiles accepted contributions for the 2nd Edition of the Colombian Computational Biology and Bioinformatics Congress CCBCOL, after a rigorous review process in which 54 papers were accepted for publication from 119 submitted contributions.

Luis F. Castillo
Marco Cristancho
Gustavo Isaza
Andrés Pinzón
CCBCOL'13 Programme Co-chairs

Organization

General Co-chairs

Luis Fernando Castillo	Universidad de Caldas, Colombia
Andrés Pinzón	BIOS Centro de Bioinformática y Biología Computacional, Colombia
Marco Cristancho	Cenicafé, Colombia

Organizing Committee

Luis Fernando Castillo	Universidad de Caldas, Colombia
Gustavo A. Isaza E.	Universidad de Caldas, Colombia
Marco Cristancho	Cenicafé, Colombia
Andrés Pinzón	BIOS Centro de Bioinformática y Biología Computacional, Colombia
Dago Bedoya	BIOS Centro Bioinformática y Biología Computacional, Colombia
Emiliano Barreto	Universidad Nacional de Colombia, Colombia
Diego Mauricio Riaño	Universidad de los Andes, Colombia

Scientific Committee

Luis Fernando Castillo	Universidad de Caldas, Colombia
Gustavo Isaza E.	Universidad de Caldas, Colombia
Carlos Alberto Ruiz Villa	Universidad de Caldas, Colombia
Oscar Julián Sanchez	Universidad de Caldas, Colombia
Lucimar Gomes Dias	Universidad de Caldas, Colombia
German LopezGartner	Universidad de Caldas, Colombia
Andrés Paolo Castaño	Universidad de Caldas, Colombia
Maria Helena Mejía	Universidad de Caldas, Colombia
Maria Mercedes Zambrano	Corpogen, Colombia
Juan Manuel Anzola	Corpogen, Colombia

Jorge Duitama	CIAT, Colombia
Adriana Muñoz	University of Maryland, USA
Patricia Vélez	Universidad del Cauca, Colombia
Mauricio Rodriguez	Centro Bioinformática y Biología Computacional, Colombia
Andrés Pinzón	Centro Bioinformática y Biología Computacional, Colombia
Alvaro Gaitán	Cenicafé, Colombia
Marco Cristancho	Cenicafé, Colombia
Carlos Ernesto Maldonado	Cenicafé, Colombia
Emiliano Barreto	Universidad Nacional de Colombia, Colombia
NestorDario Duque	Universidad Nacional de Colombia, Colombia
Luis Fernando Cadavid	Universidad Nacional de Colombia, Colombia
Luz Mary Salazar Pulido	Universidad Nacional de Colombia, Colombia
Edgar Antonio Reyes	Universidad Nacional de Colombia, Colombia
Luis Fernando Niño	Universidad Nacional de Colombia, Colombia
Silvia Restrepo	Universidad de los Andes, Colombia
Diego Mauricio Riaño	Universidad de los Andes, Colombia
Adriana Bernal	Universidad de los Andes, Colombia
Felipe García Vallejo	Universidad del Valle, Colombia
Pedro Antonio Moreno	Universidad del Valle, Colombia
Mauricio Corredor Rodríguez	Universidad de Antioquia, Colombia
Omar Triana	Universidad de Antioquia, Colombia
Carlos Muskus	Universidad de Antioquia, Colombia
Juan Manuel Corchado	Universidad de Salamanca, Spain
Juan F. de Paz	Universidad de Salamanca, Spain
Emilio Corchado	Universidad de Salamanca, Spain
Sara Rodriguez	Universidad de Salamanca, Spain
Florentino Fdez-Riverola	Universidad de Vigo, Spain
Daniel Glez-Peña	Universidad de Vigo, Spain
Miguel Reboiro Jato	Universidad de Vigo, Spain
Hugo López F.	Universidad de Vigo, Spain
Analia Lourenco	Universidad de Vigo, Spain
David Torrens	Centro Supercomputación Barcelona (BSC), Spain
Jorge Enrique Gómez	Universidad del Quindio, Colombia
Jannet Gonzalez	Universidad Javeriana, Colombia
Nelson Fernández	Universidad de Pamplona, Colombia
Manuel Alfonso Patarroyo	Fundación Instituto de Inmunología de Colombia (FIDIC), Colombia
Juan F. Alzate Restrepo	Centro Nacional de Secuenciación Genómica-CNSG, Colombia
Sarah Ayling	The Genome Analysis Centre (TGAC), United Kingdom
Leonardo Mariño	NCBI, USA

Acknowledgement

This book has been sponsored by Colciencias through the National Call 612/2013 to form a bank of eligible scientific events National and International component of social appropriation of knowledge that take place in Colombia between the second half of 2013 and the first half of 2014.

Contents

Predictive Modeling of Signaling Transduction Mediated by Tyrosine-Kinase Receptors

Ivan Mura

EAN University
Carrera 11 No. 78 - 47 Bogotá - Colombia
imura@ean.edu.co

Abstract. HER members of the tyrosine-kinase family of transmembrane receptors are initiators of signaling cascades driving crucial cellular process, such as gene transcription, cell cycle progression, apoptosis. Given their capacity of oncogenic transformation these receptors are the target of selective anticancer drugs, which in-vivo are however not as effective as anticipated by in-vitro experiments. Translating HER inhibitors into effective therapies to block the oncogenic signaling cascades will be facilitated by models that can provide reliable predictions for the evolution of the intricate HER mediated signaling networks. This work presents a process-algebra based approach to compactly specify and simulate HER signaling models. The proposed HER activation model can be easily reused as a building block in larger models of signaling.

Keywords: Tyrosine kinase receptors, Signaling pathways, Cancer therapies, Computational modeling, Stochastic simulation.

1 Introduction

The tyrosine-kinase family of transmembrane receptors includes at least 17 different classes of receptors, among which the human epidermal growth factor receptors (hereafter, HER). There are four structurally related HER receptors: HER1, HER2, HER3, and HER4. HERs play a crucial role in cell signaling, mediating cell proliferation, migration, differentiation, apoptosis, and cell motility, owing to their ability to activate important cytoplasmic signaling relaying molecules such as PI3K, Ras, Stat3, Grb2 among others [1].

HER family receptors are often over-expressed, amplified or mutated in many forms of cancer. HER1 is found to be over-expressed in more than 80% of head and neck cancers, 50% of gliomas, 10 to 15% of non-small cell lung cancers in the west [7]. Amplification and overexpression of HER2 is seen in about 25 to 30% of breast cancers [13]. Given their frequent altered expression or dysregulation in human tumors, HERs are target of selective anticancer drugs.

In spite of the effective inhibition shown in vitro, only a small percent of the patients that receive HER antagonist therapies respond to it (for instance, 20-30% in the case the HER2 inhibitor *trastuzumab* [6]). This indicates the existence of a complex set of intertwined relationships that the HER family members

L.F. Castillo et al. (eds.), *Advances in Computational Biology*,
Advances in Intelligent Systems and Computing 232,
DOI: 10.1007/978-3-319-01568-2_1,

exhibit. A drug is attempting to switch off the signaling cascade by specifically targeting one receptor type, while at the same time other HER members are compensating for the effects of the drug with sustaining signaling activation [2]. This unexpectedly complex behavior may well explain the disappointing results of various trials of candidate HER antagonists drugs and offers an excellent challenge for the deployment of Systems Biology modeling approaches [12].

In this paper we consider the modeling of the initial events of the HER signaling pathways, which are triggered by the binding of the HER receptors with their ligands, the dimerization of receptors and the phosphorylation of their intracellular domains. More specifically, inspired by the recent study reported in [2], we consider a scenario where the HER1 and HER2 receptors are sinergistically working to sustain signaling. This scenario only considers a few molecules, yet the temporal evolution of the system results in an combinatorial mesh of interacting partners. This complexity offers at least two types of challenges to the modeling tools: From a model specification point of view, such complexity makes model definition long and cumbersome (and hence error prone), whereas from the solution point of view it makes model simulation a computationally intensive task, due to the large number of possible reactions.

To tackle these issues, we shall be using an approach based on the modeling language BlenX [3], which adopts a process-algebra model specification to nicely manage the complexity inherent to the combinatorial explosion of the number of species configurations. We offer two contributions in this paper. The first one is represented by the example of application of the process-algebra modeling approach, which finds in the complexity of the HER activation models an ideal application area. The second contribution consists in the model itself, which can be used as a building block in larger models of signaling or easily adapted to define activation models for other tyrosine-kinase receptors.

This paper is organized as follows. We provide in Section 2 an introduction to the structure and function of the HER family receptors, and then we focus in Section 3 on the definition of a HER activation and phosphorylation model in BlenX. Section 4 shows the results of model validation and finally Sections 5 and 6 provide a short discussion and conclusions for the paper, respectively.

2 HER Receptors

We provide in this section a short description of the molecular details of the HER activation process. Our discussion is limited to HER1 and HER2, but it readily applies to HER3 and HER4 as well.

HER receptors are mostly located on the cell membrane, and are made up of an extracellular region or ectodomain, a single transmembrane-spanning region, and a cytoplasmic tyrosine kinase domain. HER proteins are capable of forming homodimers, heterodimers, and possibly higher-order oligomers upon activation by a subset of potential growth factor ligands. There are many growth factors that activate HER1 receptors, among which EGF, TGF-α and neuregulins. Although unliganded homodimers and heterodimers can also form, they

are unlikely to be active [10]. Multiple phosphorylation sites exist in the intracellular domains of HERs. For instance, at least 5 sites appear to be relevant for downstream signal rely in HER1 [5] and at least 4 in HER2 [9]. The dimer formation is reversible; HER dimers can dissociate and reassociate regardless of their phosphorylation status.

This means that each single HER molecule can exist in many possible configurations. If each HER molecule carried 4 phosphorylation sites, the total number of possible configurations would be of the order of 2^9. Such a combinatorial number of configurations makes most modeling approaches cumbersome if not impracticable. Any modeling tool requiring the explicit encoding of all possible species configurations and of all the reactions they participate in would be impossible to use. We shall see in the next section how the system can be easily encoded with the BlenX approach, which does not require fully unfolding the set of possible configurations of the species.

3 Modeling Methods

BlenX [3] is a modeling language based on the process calculi and rule-based paradigms. It is specifically designed to account for the complexity of biological networks. The advantages of a rule-based approach become evident when the biological system exhibits a combinatorial number of possible configurations as in the case of the HER early signaling network. Given the space limitation, we just provide in this section a few clues about the BlenX modeling approach. A complete explanation with examples of application, can be found in [4].

BlenX uses a general abstraction of a biological network that separately considers biological entities and their interaction capabilities. Biological entities are encoded in BlenX as objects called *boxes*. Boxes expose interfaces called *binders*, which mimic domains of interaction, for instance for complexation and phosphorylation. The interaction capabilities of binders are determined by its type attribute, which is controlled by each box internal *process*, which updates them according to the box state.

BlenX resembles a normal programming language. A box for the EGF ligand of HER1 would be defined as follows:

```
let egf : bproc = #(egfrec,egfrec)[nil];
```

This text is declaring the box *egf* as having one binder named *egfrec*, having a type *egfrec*. This binder models the site across which EGF binds to the receptor. The text within square brackets is the box process. In the case of EGF, the process is `nil`, i.e. the null process, which does nothing and hence the type of binder *egfrec* will ever be changed. This models the fact that the complexation with HER1 is always possible for EGF molecules. A box for a HER1 molecule needs more binders: one called *h1lig* for the interaction with the ligand, one called *h1dim* for the dimerization with another HER molecule, plus at least one binder *h1ph* to model a single phosphorylation site. Moreover, it would need a non-null internal process to model the phosphorylation process so that it can

happen only after ligand binding and dimerization. We declare HER1 and HER2 boxes (for the sake of conciseness, each one with just one phosphorylation site), as follows:

```
let her1 :: bproc = #(h1lig,h1lig),#(h1dim,h1dim),
    #(h1ph,free)[h1_proc];
let her2 :: bproc = #(h2dim,h2dim),#(h2ph,free),
    [h2_proc];
```

Binding interaction capabilities are called *affinities* in BlenX, and are defined for pairs of binder types. For instance, to model the reversible complexation of EGF and HER1, we declare in BlenX a tuple as follows: `(egfrec,h1lig,`$\mathtt{k}_{on}$`,`$\mathtt{k}_{off}$`)`, where $\mathtt{k}_{on}$ and $\mathtt{k}_{off}$ are the rate of complex formation and dissociation, respectively. The rate information is used at simulation time. BlenX uses a discrete-space discrete-time interpretation of model evolution, according to Gillespie stochastic molecular dynamics [8]. Intuitively, the rates are proportional to the speed with which the biochemical transformations occur.

The internal processes keep track of the history of the boxes and appropriately change the state of the binders. For instance, process `h1_proc` of box `her1` will change the type of binder `h1ph` from `free` to `phospho` when an `egf` box binds `h1lig` and an HER molecule binds `h1dim`.

To model our scenario of HER1 and HER2 interaction, we just need the 3 box declarations given above, and the specification of the two internal processes `h1_proc` and `h2_proc`, a BlenX program that is as compact as 15 lines of code. Additionally, we need 4 affinities declaration, one for the HER1-EGF binding, and 3 for the hetero and homodimerizations. Overall, a very compact BlenX model accounts for all the possible configurations of the species and their reactions.

4 Results

We present in this section the results of simulation experiments aiming at validating the HER1-HER2 interaction model. We start our experiments by reproducing an experimental setting that was used in the paper [11], where the activation of HER1 in hepatocytes is considered. In that study, an in-vitro culture of HER1 cells is stimulated by a single EGF pulse, and the phosphorylation level of the receptors is measured over time by immunoprecipitation. By using the kinetics of EGF binding, dimerization and phosphorylation given in [11] to instanciate the BlenX model, we could reproduce the HER1 time course of phosphorylation over the time window [0-120] seconds, as shown in Fig.1.

Then, we used the same rates in the complete model, when also HER2 is considered. Without introducing any new model parameter, we reproduced the relative proportions of HER1-HER1 homodimers and HER1-HER2 heterodimers at equilibrium (10 minutes after EGF stimulation) in four breast cancer cell lines. Table 1 shows the very good match of simulation results with respect to

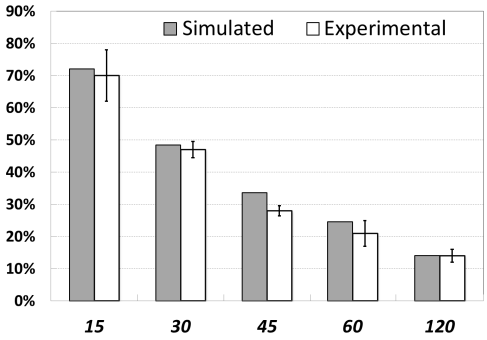

Fig. 1. Comparison of experimental and simulated HER1 phosphorylation over time. Experimental results are provided with error bars. Simulation results were estimated at 95% confidence level and are within 10% of the estimated value.

Table 1. Results of HER1-HER1 homodimer versus HER1-HER2 heterodimer ratio simulations for various breast cancer cell lines. Confidence intervals of simulation are within 10% of the estimated value.

Cell Line	Num HER1	Num HER2	Ratio Homodimers/Heterodimers	
			Experimental results	Simulation results
AU565	204560	1447688	0.071	0.072±0.0012
SKBR3	143559	1402832	0.042	0.050±0.0014
SKOV3	387771	657088	1.627	1.393±0.0035
H1650	158872	53810	15.625	14.934±0.0776

the experimental data obtained in [2]. These results validate the BlenX model and demonstrate its predictive capabilities.

5 Discussion

The BlenX model presented in this paper appears to be able to reproduce the experimentally observed behavior of HER receptors under different conditions. Main advantages of the proposed modeling approach are its compactness and easy extensibility. For instance, to pass from a pure HER1 model, i.e. one considered in the experimental setting of Kholodenko, to the HER1-HER2 interaction model, we just *added* the HER2 specifications to the existing code, without having to *change* it. This is due to the basic modeling choice of not specifying explicitly the reactions each species can participate in, but just to define its possible interactions. Specifically, the definition of interactions is made at the level of the binders, which represent the domains of molecules. Such a subtle difference obviate the necessity of enumerating the set of reactions, their reactants and products.

6 Conclusions

We showed in this paper the preliminary results of a HER model based on the process-algebra language BlenX. The foundational aspects of the modeling approach are presented, and a model of HER1-HER2 signaling sketched to provide some clues about the expressiveness of the language.

The proposed model can be easily extended to encode, in a very compact way, systems that include thousands of species and reactions, which would be otherwise impossible to specify. We validated the model with respect to experimental data coming from the literature, taking into consideration two different studies. The model could be further extended to consider additional HER members. In particular, we plan to consider HER3, as HER2-HER3 heterodimers are among the most active complexes in relaying growth factor signals.

References

1. Citri, A., Yarden, Y.: EGF-ERBB signalling: towards the systems level. Nat. Rev. Mol. Cell Biol. 7, 505–516 (2006)
2. DeFazio-Eli, L., Strommen, K., Dao-Pick, T., Parry, G., et al.: Quantitative assays for the measurement of HER1-HER2 heterodimerization and phosphorylation in cell lines and breast tumors. Breast Cancer Res. 13 (2011)
3. Dematté, L., Priami, C., Romanel, A.: Modelling and simulation of biological processes in BlenX. Perform Eval. Rev. 35, 32–39 (2008)
4. Dematté, L., Priami, C., Romanel, A.: The blenX language: A tutorial. In: Bernardo, M., Degano, P., Zavattaro, G. (eds.) SFM 2008. LNCS, vol. 5016, pp. 313–365. Springer, Heidelberg (2008)
5. Downward, J., Parker, P., Waterfield, M.D.: Autophosphorylation sites on the epidermal growth factor receptor. Nature 311, 483–485 (1984)
6. Esteva, F.J., Valero, V., Booser, D., Guerra, L.T., et al.: Unraveling resistance to trastuzumab (Herceptin): insulin-like growth factor-I receptor, a new suspect. J. Clin. Oncol. 20, 1800–1808 (2002)
7. Frederick, L., Wang, X.Y., Eley, G., James, C.D.: Diversity and frequency of epidermal growth factor receptor mutations in human gliobastomas. Cancer Res. 60, 1383–1387 (2000)
8. Gillespie, D.T.: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comp. Physics 22, 403–434 (1976)
9. Hazan, R., Margolis, B., Dombalagian, M., Ullrich, A., et al.: Identification of autophosphorylation sites of HER2/neu. Cell Growth Differ. 1, 3–7 (1990)
10. Jura, N., Endres, N.F., Engel, K., Deindl, S., et al.: Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment. Cell 137, 1293–1307 (2009)
11. Kholodenko, B.N., Demin, O.V., Moehren, G., Hoek, J.B.: Quantification of short term signaling by the epidermal growth factor receptor. Biol. Chem. 274, 30169–30181 (1999)
12. Kitano, H.: Systems Biology: a brief overview. Science 295, 1662–1664 (2002)
13. Slamon, D.J., Godolphin, W., Jones, L.A., Holt, J.A., et al.: Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. Science 244, 707–712 (1989)

Bioinformatic Analysis of Two Proteins with Suspected Linkage to Pulmonary Atresia with Intact Ventricular Septum

Oscar Andrés Alzate Mejía[1] and Antonio Jesús Pérez Pulido[2]

[1] Docente Universidad Autónoma de Manizales, Colombia
[2] Docente Universidad Pablo de Olavide, Sevilla - España
oalzate@autonoma.edu.co, ajperez@upo.es

Abstract. Pulmonary atresia with intact ventricular septum (PA-IVS) is a congenital heart disease characterized by occlusion of the pulmonary valve causing complete obstruction of the outflow tract from the right ventricle to the lungs. Some authors attribute the origin of disease to genetic causes and the mutation of genes WFDC8 and WFDC9 have been proposed as related to with its pathogenesis. Based on this suspicion, a bioinformatic analysis to their gene products was made to find the relationship between the mutation and disease.

Were reviewed the annotations, domains and structures of these proteins to study their biological characteristics; to find equivalent sequences in other species the orthologous of proteins were searched, so it was made a phylogenetic analysis and were searched conserved domains using alignments. Similarly, were searched the tissues in which genes are expressed to find its relation to heart and was studied the intergenic sequence to uncover regulatory sequences associated with cardiac development.

The results suggest that the human proteins WFDC8 and WFDC9, currently related to the immune system, they are also related to extracellular matrix proteins, and they could be expressed in heart tissue and embryonic, in addition, was found that the corresponding intergenic sequence has different binding sites for factors transcription related to the development of heart and heart valves.

Keywords: Pulmonary atresia, WFDC8, WFDC9, heart tissue.

1 Introduction

Congenital heart diseases are disorders of the heart and great vessels that exist before birth. They describe structural or functional injuries of one or more of the four cardiac chambers or septum which separate them, or their respective valves.

The Pulmonary atresia with intact ventricular septum (PA-IVS) is a congenital heart defect characterized by a pulmonary valve atresia, a complete obstruction of the outflow tract from the right ventricle to the lungs. But unlike other diseases of heart septum that separates both ventricles, it is intact. The etiology of the disease begins to be known. Some familial cases suggest genetic basis (1). A report of PA-IVS was

L.F. Castillo et al. (eds.), *Advances in Computational Biology*,
Advances in Intelligent Systems and Computing 232,
DOI: 10.1007/978-3-319-01568-2_2,

published in monozygotic twins (2); in them and through an analysis of comparative genomic hybridization (aCGH) revealed a deletion of 55 kb at the level of the chromosome 20q13.12 involving WFDC8 and WFDC9 genes in both patients.

In humans, WFDC genes are encoding small proteins that can contain one or more domains WFDC. It is knows that the domain WFDC performs part of innate immunity, which is present in some inhibitors serine protease that stop endogenous peptidases secreted by cells proinflammatory, thus avoiding damage and inflammation of the tissues. Similarly, it has been shown to inhibit proteases secreted by exogenous microorganisms showing potent antibacterial, antifungal, and antiviral properties (3). Currently does not exist in the literature references linking the proteins WFDC8 and WFDC9 with the heart or great vessels. In this work, we analyze bioinformatics in order to find the relation of the mutation and pulmonary atresia with intact ventricular septum.

To this purpose, functional annotations, domain arquitecture and 3D structures of proteins were reviewed in order to analyze their biological characteristics; to find sequences equivalents in other species (orthologs), a phylogenetic analysis was made and we searched for domains maintained by alignments. Similarly, we looked for tissues where genes are expressed to find his relationship with the heart and, finally, we studied the intergenic sequence to discover regulatory sequences related to cardiac development.

The results reveal the relationship of the WAP for WFDC8 and WFDC9 domains with extracellular matrix proteins, which constitute the architecture of the heart valves. So the same, the expression of the proteins was found in cardiac and embryonic tissue. Finally, it shows that the sequence between both genes presents different binding sites for factors of transcription which are related to the development of the heart and heart valves.

2 Materials and Methods

2.1 Review of Annotations from WFDC8 and WFDC9

To assign the biological characteristics of proteins WFDC8 and WFDC9 the database UniProt was used. The available information about the origin, attributes, annotations, ontology and sequences were obtained from this important knowledge base. Review of annotations was extended thanks to cross-references offering UniProt. The program Rasmol V2.7.4.2 was used to study the regions of interest and view the structure obtained.

2.2 Search for Orthologs and Similarity

To search of homologous sequenceswe used three methods. First we looked for significant orthologs in the results provided by the UniProt Blast. Second, an algorithm written in Perl programming language based on the location of signals of short length sequence was used (4). Finally, we searched for orthologs with NCBI Blast tool tblastn program against database EST (sequences coming from mRNA sequencing).

2.3 Comparison of Sequences

We compared the foundorthologs using the Dot Plot program UGENE V1.10.4. We obtained arrays of points by comparing the human sequence with each of their possible orthologs to observe the best positions to be evolutionarily conserved.

2.4 Multiple Alignment and Phylogeny

We perform the multiple alignment between human proteins and their orthologues using ClustalX V2.0.10. To edit and analyze the alignments also applied the program Bioedit V7.0.4.1. Another purpose with alignments was to carry out phylogenetic analysis, for this purpose, the multiple alignment data were treated with TreeView V1.6.6 to visualize the phylogenetic trees

2.5 Gene Expression Data

Gene expression data was originated from results from different experiments stored in the ArrayExpress database of EBI and the database of proteins in human, nextprot BETA.

2.6 Analysis Bioinformatics to the Intergenic Sequence

The coordinates of deletion of the work of Stefano were presented (2) in the graphical interface UCSC Genome browser (2006), to obtain the graphic that showed the deletion reported in this paper. On the other hand, were sought in the JASPAR database factors of transcription of the human species related heart embryogenesis. In addition, another application of an algorithm in Perl programming language search with weight matrices common transcription factors in obtained alignments. Finally a file in format gff added as tracks in the Genomes of UCSC, allowed to analyze the relationship between the conserved intergenic sequence and transcription factors.

2.7 Experimental Analysis to the Intergenic Sequence

Different experiments are underway at the moment to check the regulatory elements of conserved intergenic sequence. To do so, we conduct important molecular biology techniques such as PCR, electrophoresis, techniques of purification obtaining of colonies and digestion with EcoR1.

3 Results

Only the work of Stefano in 2008 (2) has related both WFDC8 and WFDC9 proteins to PA-IVS. To perform a search in the UniProt database and your links we find that domains WFDC8 and WFDC9 are associated with the extracellular matrix. To apply the tblastn to proteins WFDC we find that the orthologs of WFDC8 and WFDC9 are expressed in embryonic and cardiac tissue. On the other hand the databases of gene expression, Array Express, for WFDC8 shows results of differential expression in organs different than the heart. However, it highlights expression of genes in different

cell types, including embryonic stem cells. The results stored in other databases of proteins reported the expression of WFDC8 in ESTs of fetuses. Using the coordinates of deletion of the work of De Stefano (2) was obtained the region deleted of the cases studied (Figure 1A), this region includes the WFDC9 gene, intergenic sequence with the promoter region of WFDC8 and the initial region of this gene. To observe the intergenic sequence was detailed the relationship of this sequence with JunD genes and c-Jun (Figure 1), which has been governing the heart expression (5), (6).

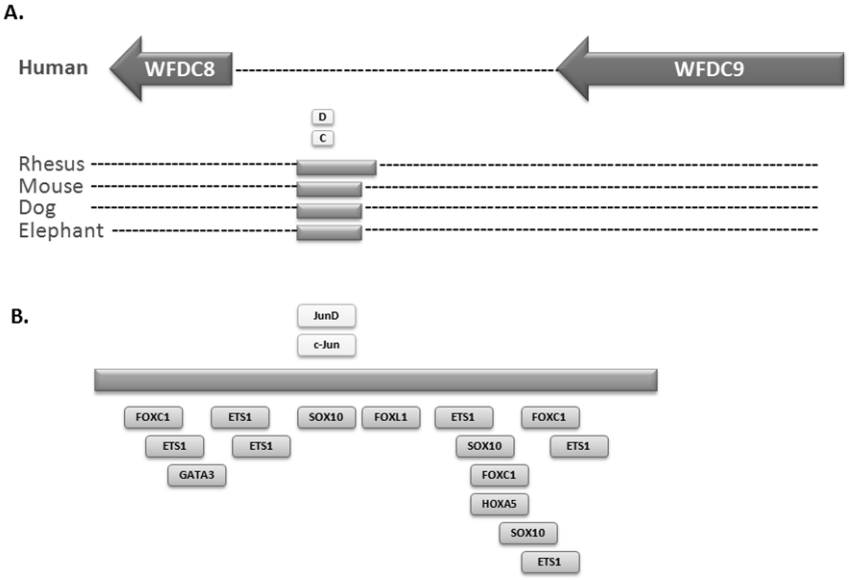

Fig. 1. Extent of mutation causative of PA-IVS (**A**). Show the WFDC9 gene, intergenic sequence and the region 5' of the WFDC8 gene. Details, in the midle of the intergenic sequence, a region preserved in 4 different species and the emergence in this region of binding sites for the transcription factors JunD and c-Jun. (**B**) Enlargement of the region preserved in mammals where we found binding sites for 15 more cardiac embryogenesis related transcription factors. Highlights the increased presence of ETS1 distributed into several blocks and is detailed in the middle of the sequence, the binding site for JunD and c-Jun shared with Sox10.

On the other hand, we found 16 factors of transcription in the human species related with the development of the heart, angiogenesis and the proteins of the extracellular matrix. So the factors of transcription of the human species were sought from Jaspar and analyzed its ontology using Ensembl database. Then and after applying an algorithm in programming language Perl was found in the conserved sequence 15 binding sites for transcription factors related to cardiac embryogenesis, with an identity of 80% compared to weight matrices. In addition, other important factors of transcription met identities a little lower, as it is the case of Sox 9 with 70% of identity against the arrays.

Finally, we analyze the important provision of these transcription factors on the intergenic sequence conserved. The factors complement the listed JunD and c-Jun (Figure 1B)

4 Discussion

Current research relates proteins WFDC8 and WFDC9 with the immune system. The bioinformatics analysis applied to these proteins allowed to clarify its relationship with the development of the heart, heart valves and PA-IVS.

According to the results of this work, the proteins WFDC8 and WFDC9 have WAP domains that bind the entire molecule to proteins of the extracellular matrix by means of its amino acids exposed. The mutation of the genes and the consequent lack of proteins can lead to the disruption of the extracellular matrix and therefore the desfuncionalización of the valves during the development of the heart.

The expression of the proteins WFDC8 and WFDC9 was confirmed with databases of gene expression in embryonic stem cells and human fetuses ESTs. Still, there are more preponderant evidence confirming its expression in heart and embryonic tissue. Their homologous sequences in rodents are quite significant for this fact.

In the present work, wehave found evidence suggesting that the mechanisms required for the structuring of the heart valves have relationship with different transcription factors. ETS1 is a transcription factor which has a wide range of biological functions including the regulation and cell growth, hematopoiesis, lymphocyte development, development and remodeling of vessels and mainly participation in the development of different organs (7). Some studies involve him significantly in the development of the heart of mammals and in the pathogenesis of some congenital heart (8).

On the other hand, the Sox proteins constitute a long family of factors involved in different processes of embryonic development. Sox 8, Sox 9 and Sox 10 exhibit a dynamic expression during embryogenesis of the heart correlated with the septation and differentiation of the connective tissue of the valve (9). Particularly, Sox 9 is required for proliferation and differentiation of cardiac valves progenitor cells, is required for the expression of collagen in the modeling of valve and its absence is related to diseases of the heart valve, including increases in calcium in the area (10).

The relationship of c-jun transcription factors, JunD, and conserved intergenic sequences also have significance to the development of the heart (5), (6). During embryogenesis, c-jun is required for the development of the cardiovascular system, but its regulation diminishes in the adult heart. For his part, JunD is dispensable for the cardiovascular development of the embryo, however is expressed in all cardiac development and is kept in the cells of the adult heart.

Finally the mutation of conserved intergenic sequence of WFDC8 and WFDC9 includes the deletion for different binding sites of ETS1, Sox 9, Sox10, c-jun, JunD, and other factors of transcription. This probably shows that the absence of sequence prevented the function of these important factors and proper development of the heart valve. Equally, the area of mutation as well as conserved sequence included the gen WFDC9, the promoter region and the initial part of WFDC8 (Figure 1), which possibly confirms the lack of expression of these two proteins.

5 Conclusions

An *in silico* analysis link to (o relates to) the domains WAP of WFDC8 and WFDC9 with extracellular matrix proteins, which could constitute the architecture of heart

valves, where the mutation of genes can lead to your disfunctionalization. So the same, we showed that proteins WFDC8 and WFDC9, are expressed in human embryonic tissue and in the heart, in any of its orthologs. Finally study the intergenic sequence finding a site conserved in different species, which were important binding sites for transcription factors related to the embryonic development of heart and heart valves, suggesting that the deletion could prevent valvular embryogenesis is promoted. Currently wish to strengthen these results, and now are looking for experimentally in the intergenic sequence enhancers that promote cardiac expression. Conclusions are expected to confirm the causes of PA-IVS.

References

1. Grossfeld, P.D., Lucas, V.W., Sklansky, M.S., Kashani, I.A., Rothman, A.: Familial occurrence of pulmonary atresia with intact ventricular septum. American Journal of Medical Genetics 72 (1997)
2. De Stefano, D., Li, P., Xiang, B., Hui, P., Zambrano, E.: Pulmonary Atresia With Intact Ventricular Septum (PA-IVS) in Monozygotic Twins. American Journal of Medical Genetics 146A, 525–528 (2008)
3. Clauss, A., Lilja, H., LundWall, A.: The evolution of a genetic locus encoding small serine proteinase inhibitors. Biochemical and Biophysical Research Communications 333, 383–389 (2005)
4. Mier, P., Perez Pulido, A.: Fungal Smn and Spf30 homologues are mainly present in filamentous fungi and genomes with many introns: Implications for spinal muscular atrophy. Gene. 491, 135–141 (2012)
5. Hilfiker-Kleiner, D., Hilfiker, A., Kaminsky, K., Schaefer, A., Park, J.-K., Michel, K., Quint, A., Yaniv, M., Weitzman, J., Drexler, H.: Lack of JunD Promotes Pressure Overload–Induced Apoptosis, Hypertrophic Growth, and Angiogenesis in the Heart. Circulation 112, 1470–1477 (2005)
6. Eferl, R., Sibilia, M., Hilberg, F., Fuschbichler, A., Kufferath, I., Guertl, B., Zenz, R., Wagner, E., Zatloukal, K.: Functions of c-Jun in Liver and Heart Development. The Journal of Cell Biology 145, 1049–1061 (1999)
7. Oikawa, T., Yamada, T.: Molecular Biology of the Ets family of transcription factors. Gene. 303, 11–34 (2003)
8. Ye, M., Coldren, C., Liang, X., Mattina, T., Goldmuntz, E., Benson, W., Ivy, D., Perryman, M.B., Garret-Shinha, L.A., Grossfeld, P.: Deletion of ETS-1, a gene in the Jacobsen syndrome critical region, causes ventricular septal defects and abnormal ventricular morphology in mice. Human Molecular Genetics 19, 648–656 (2010)
9. Montero, J.A., Giron, B., Arrechedera, H., Cheng, Y.-C., Scotting, P., Chimal-Monroy, J., Garcia-Porrero, J.A., Hurle, J.M.: Expression of Sox8, Sox9 and Sox10 in the developing valves and autonomic nerves of the embryonic heart. Mechanisms of Development 118, 199–202 (2002)
10. Lincoln, J., Kist, R., Sherer, G., Yutzey, K.: Sox9 is required for precursor cell expansion and extracellular matrix organization during mouse heart valve development. Developmental Biology 305, 120–132 (2007)

Construction and Comparison of Gene Co-expression Networks Based on Immunity Microarray Data from *Arabidopsis*, Rice, Soybean, Tomato and Cassava

Luis Guillermo Leal, Camilo López, and Liliana López-Kleine

Department of Statistics, Department of Biology,
Universidad Nacional de Colombia, Bogotá, Colombia
{lgleala,celopezc,llopezk}@unal.edu.co

Abstract. A big challenge in gene expression data analyses is to reveal the coordinated expression of different genes. Gene co-expression networks (GCNs) are graphic representations where nodes symbolize genes while edges reconstruct the coordinated transcription of genes to certain external stimuli. In this paper, an enhanced novel methodology for construction and comparison of GCNs is proposed. Microarray datasets from pathogen infected plants (*Arabidopsis*, rice, soybean, tomato and cassava) were used. Initially, similarity metrics that find linear and non-linear correlations between gene expression profiles were evaluated. A similarity threshold was chosen and GCNs were constructed. Afterwards, GCNs were characterized by graph variables and a principal component analysis on these variables was applied to differentiate them. The results allowed the discovery of topologically and non-topologically similar networks among species. Potentially conserved biological processes, like those related to immunity in plants could be studied from this work.

Keywords: Gene co-expression networks, similarity metrics, principal component analysis, plants immunity.

1 Introduction

The integration of expression data offers significant insights to understand the transcriptional mechanisms of living organisms. Gene co-expression networks (GCNs) are graphic representations for this purpose [1]. They depict the coordinated transcription of genes by means of linked nodes in a graph. Accordingly, GCNs show the functional associations between co-expressed genes when external treatments are induced [2].

Commonly, the procedure to construct a GCN involves a similarity metric assessed between expression profiles of genes [2].The Absolute value of Pearson Correlation Coefficient (APCC) has been the most used similarity metric [3]. However, as gene expression profiles are not always correlated linearly, many non-linearly correlated genes are not retained for inclusion in the final GCN [4]. Subsequently, a similarity threshold is selected to find relevant pairs of genes connected

L.F. Castillo et al. (eds.), *Advances in Computational Biology*,
Advances in Intelligent Systems and Computing 232,
DOI: 10.1007/978-3-319-01568-2_3,

in the network [5]. Once GCNs are constructed, they can be analyzed in order to provide functional annotations for genes which's function is unknown [6]. Besides, GCNs have been useful in studies of translational functional genomics such as networks alignment [6] and comparisons based on graph variables [7]. During the last case, networks have mainly been described by topological variables and a principal component analysis (PCA) on these variables has been applied to characterize them. Projections resulting from PCA reflect structural similarity through closeness on the PCA space. Nevertheless, networks comparison based on topological variables has for now only allowed the discovery of similar graph motifs while valuable biological conclusions remain unrevealed [5].

Because of the importance of these applications, both the construction and comparison of GCNs demands specific and well developed strategies. In this work, we concentrate on shortcomings of current approaches and propose an enhanced novel methodology to overcome them. We compared the performance of APCC with Mutual Information Coefficient (MIC) [4] and the Normalized Mean Residue Similarity (NMRS) [8], and chose the metric that better detects linear and non-linear correlations. To characterize the GCNs, we added new non-topological variables, such as tolerance to pathogen attacks and assortativity coefficients related to functional annotations. These variables describe better the networks from a biological perspective and are less dependent on network size. To evaluate our methodology, pathogen resistance microarray datasets from *Arabidopsis thaliana*, rice [*Oryza sativa*], soybean [*Glycine max*], tomato [*Solanum lycopersicum*] and cassava [*Manihot esculenta*] were queried from public repositories and used to construct and compare a total of 59 GCNs on different datasets.

Summarizing, we have considered not only the use of appropriated similarity metrics but also the comparison of networks using multivariate methods. This strategy allowed us to find functionally similar GCNs from immunity processes in plants. Moreover, the current biological knowledge on model organisms and less studied species is widened with our results and can be a starting point for emitting biological hypothesis related to plant immunity processes.

2 Materials and Methods

Raw microarray datasets from pathogen infected plants experiments were obtained from GEO DataSets repositories and previous studies [9]. Datasets were independently pre-processed through noise reduction, quantile normalization and log2 transformation. Expression matrices were obtained independently from each dataset after removing non-differentially expressed genes [10].

A square similarity matrix was calculated for every single expression matrix. The similarities ($s_{i,j}$) between pairs of genes i and j were calculated using a metric. As mentioned previously, we compared the similarities obtained with APCC ($s_{i,j}^{APCC}$), MIC ($s_{i,j}^{MIC}$) and NMRS ($s_{i,j}^{NMRS}$) [3][4][8]. These measures take values in the same interval [0,1], where 0 indicates non-dependence between expression profiles, and 1 indicates total dependence or maximum similarity. Subsequently, a similarity threshold (τ) was selected for each similarity matrix. The τ allowed