Journal of Theoretical Biology 384 (2015) 50–58

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Classification of signaling proteins based on molecular star graph descriptors using Machine Learning models

Carlos Fernandez-Lozano [a,*], Rubén F. Cuiñas [a], José A. Seoane [b], Enrique Fernández-Blanco [a], Julian Dorado [a], Cristian R. Munteanu [a,c]

[a] Information and Communications Technologies Department, Faculty of Computer Science, University of A Coruna, Campus de Elviña s/n, 15071 A Coruña, Spain

[b] Bristol Genetic Epidemiology Laboratories, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS82BN, UK

[c] Department of Bioinformatics – BiGCaT, Maastricht University, P.O. Box 616, UNS50 Box 19, NL-6200 MD Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 23 March 2015
Received in revised form
20 July 2015
Accepted 27 July 2015
Available online 20 August 2015

Keywords:
Feature selection
SVM-RFE
Topological indices
Signal transduction pathway

ABSTRACT

Signaling proteins are an important topic in drug development due to the increased importance of finding fast, accurate and cheap methods to evaluate new molecular targets involved in specific diseases. The complexity of the protein structure hinders the direct association of the signaling activity with the molecular structure. Therefore, the proposed solution involves the use of protein star graphs for the peptide sequence information encoding into specific topological indices calculated with S2SNet tool. The Quantitative Structure–Activity Relationship classification model obtained with Machine Learning techniques is able to predict new signaling peptides. The best classification model is the first signaling prediction model, which is based on eleven descriptors and it was obtained using the Support Vector Machines-Recursive Feature Elimination (SVM-RFE) technique with the Laplacian kernel (RFE-LAP) and an AUROC of 0.961. Testing a set of 3114 proteins of unknown function from the PDB database assessed the prediction performance of the model. Important signaling pathways are presented for three UniprotIDs (34 PDBs) with a signaling prediction greater than 98.0%.

1. Introduction

Responding to immediate changes in the environment is essential for the survival of cells. A cell is able to receive many signals simultaneously and to integrate the information into a unified action plan. In addition, a cell can also send signals to the environment. These signals may be chemical, mechanical, thermal or optical (Rhee, 2006).

Examples of chemical signals are growth factors, hormones, neurotransmitters, and extracellular matrix components. Due to the fact that any molecule can travel at distance, these signals may have a local or long distance effect (Jordan et al., 2000). An example of short-range signaling molecules is that of the neurotransmitters. They can travel a tiny distance between adjacent neurons or between neurons and muscle cells. On the other hand, an example of long-range molecules is that of the follicle-stimulating hormone, which travels from the mammalian brain to the ovary, in order to trigger the egg release. Sensory cells in the skin, ear, and human vascular system could receive mechanical stimuli.

A cell can receive and modulate the signals through hundreds of protein receptor types (Wong et al., 2012). They cause a physiological response specific for different types of molecules. The receptors for dopamine are different than those for insulin, and others can even respond directly to light or pressure. Receptors are generally transmembrane proteins, with an outside part which binds to specific signaling molecules and with an internal part that can initiate specific signaling pathways in the cell.

The membrane receptors could be grouped into three major classes: G-protein-coupled receptors (Kobilka, 1992; Rohrer and Kobilka, 1998), ion channel receptors, and enzyme-linked receptors (Evans and Levitan, 1986). Therefore, these receptors transform external signals into internal ones via protein action, ion channel opening, or enzyme activation, respectively. Due to the internal and external part of the receptors, they allow cellular actions of an external molecule without entering the cell of the signal molecules. Other receptors are localized deep inside the cell, even in the nucleus and they may bind to molecules which are able to pass through the plasma membrane (ex: gases, and steroid hormones) or that are products of the internal metabolism (Dykstra et al., 2003).

* Corresponding author. Tel.: +34 981167000x1302; fax: +34 981167160.
E-mail addresses: carlos.fernandez@udc.es (C. Fernandez-Lozano), ruben.fcuinas@udc.es (R.F. Cuiñas), j.seoane@bristol.ac.uk (J.A. Seoane), efernandez@udc.es (E. Fernández-Blanco), julian@udc.es (J. Dorado), crm.publish@gmail.com (C.R. Munteanu).

http://dx.doi.org/10.1016/j.jtbi.2015.07.038
0022-5193/© 2015 Elsevier Ltd. All rights reserved.

The receptors can launch a series of biochemical reactions within the cell due to the conformational changes determined by the signal interaction. These signal transduction cascades can amplify the message, producing multiple intracellular signals. In the first step, the activation of the receptors can trigger the synthesis of small molecules called second messengers, such as cyclic AMP (cAMP), which initiate and coordinate intracellular signaling pathways (Sassone-Corsi, 2012). The activation of adenylyl cyclase produces the synthesis of hundreds/thousands of cAMPs that activate the enzyme protein kinase A (PKA), which then phosphorylates multiple protein substrates. The cAMP signaling stops when cAMP is degraded by the enzyme phosphodiesterase.

The signaling proteins are the best targets for drugs in order to modify any biological activity. Therefore, searching for new proteins with signaling function is essential. Due to the fact that the experimental methods are expensive and time-consuming, the theoretical methods could offer the practical solution for this screening. Therefore, classification models that link the protein structure to the signaling activity could be obtained using Machine Learning techniques. The molecular information can be encoded into invariant molecular descriptors based on molecular graph topology, 3D protein conformation, peptide sequence and physical-chemical properties of the amino acids.

The classification model represents a Quantitative-Structure-Activity-Relationship (QSAR) (Archer, 1978) between the protein structure and the biological function. The QSAR models (Puzyn et al., 2010) were extensively used for antifungal (Gonzalez-Diaz et al., 2006), antiviral (Prado-Prado et al., 2011b), and antimalarial (Katritzky et al., 2006) drugs and they were extended to macromolecules such as proteins (González-Díaz and Uriarte, 2005; González-Díaz et al., 2009; Ivanciuc, 2009; Randic et al., 2007, 2006)_ENREF_19_ENREF_20 or nucleic acids (Gonzalez-Diaz et al., 2005; Li and Wang, 2004; Randić et al., 2000)_ENREF_21. Previous studies on other protein functions were focused on anti-oxidant (Fernandez-Blanco et al., 2012), transporter (Fernandez-Lozano et al., 2013), enzyme regulator (Concu et al., 2009; Fernandez-Lozano et al., 2014b), cell death-related (Fernandez-Lozano et al., 2014a) or cancer-related (Aguiar-Pulido et al., 2012; Munteanu et al., 2009) proteins. There are also other works which involves QSAR models for the prediction of protein inhibitors, where the structures or sequence of the proteins have been considered (Prado-Prado et al., 2012, 2011a; Speck-Planche et al., 2012, 2013; Viña et al., 2009).

The scientific community pays special attention to signaling protein classification problems and further efforts are needed in order to complete the scientific knowledge in this field, as the classification of proteins is a key step in computational genomics. Simple Machine Learning approaches were used in the past to automate this process: for protein classification (Ahmad et al., 2014), classification of proteins known to be frequently mutated in human cancer (U et al., 2014), classification of proteins in the early clinical stage of rheumatoid arthritis (Pratt et al., 2012) or to explore the non-linear relationships between mitochondrial morphology and apoptotic signaling (Reis et al., 2012). Those Supervised Machine Learning techniques for signaling protein classification are mainly Artificial Neural Networks, Support Vector Machines, Naïve Bayes, Decision Trees or Random Forest. Even though the use of Machine Learning approaches for protein classification is a common topic, it is difficult to compare the performance of newer approaches. This is due mainly to the lack of information about the sequences, date of extraction, concrete database or information about the configuration of the techniques in publications. In order to get a reproducible experimentation all this information was in the current paper and the datasets were made available at http://dx.doi.org/10.6084/m9.figshare.1330132.

The aim of the current study is to obtain a classification model for signaling protein prediction based on star graph descriptors of protein sequences and Machine Learning techniques. The current paper is organized as follows: Section 2 describes the methodology and the particular methods used in this work, as well as the process to generate our dataset; Section 3 includes a comparison of the proposed algorithm with the above-mentioned state-of-the-art algorithms for protein classification and an experimental analysis of the performance of the proposed method is performed,

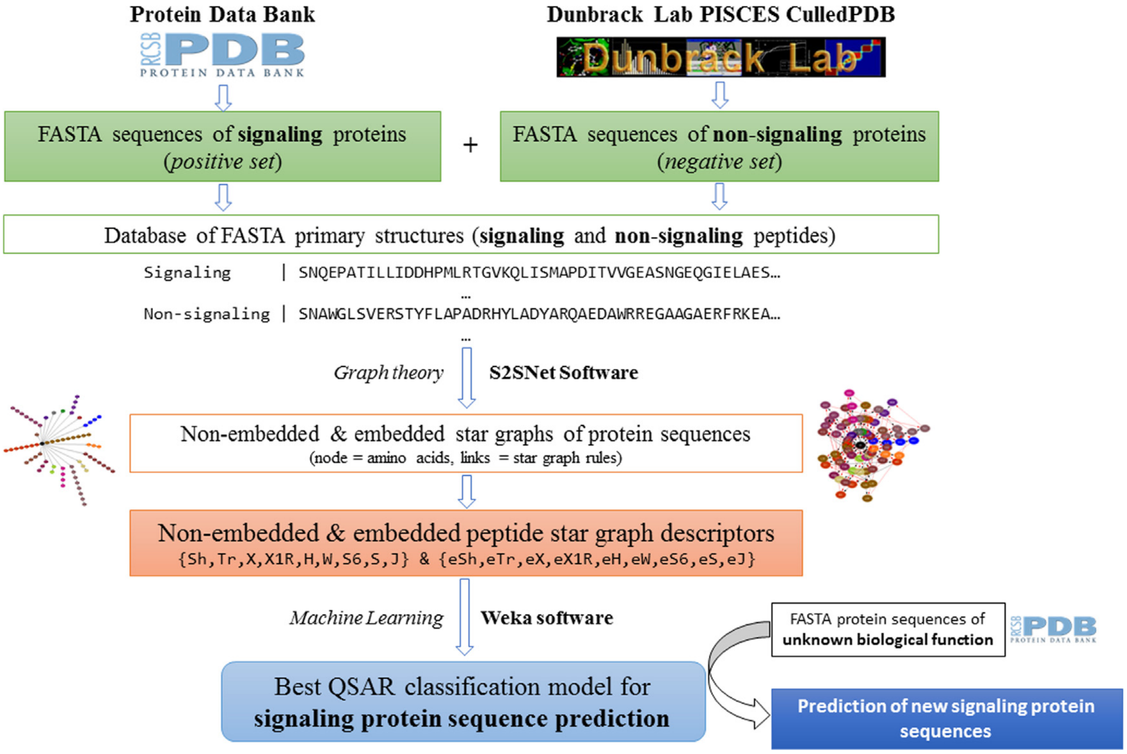
Protein Data Bank
PDB PROTEIN DATA BANK
Dunbrack Lab PISCES CulledPDB
Dunbrack Lab
FASTA sequences of signaling proteins (positive set)
+
FASTA sequences of non-signaling proteins (negative set)
Database of FASTA primary structures (signaling and non-signaling peptides)
Signaling | SNQEPATILLIDDHPMLRTGVKQLISMAPDITVVGEASNGEQGIELAES...
Non-signaling | SNAWGLSVERSTYFLAPADRHYLADYARQAEDAWRREGAAGAERFRKEA...
Graph theory
S2SNet Software
Non-embedded & embedded star graphs of protein sequences
(node = amino acids, links = star graph rules)
Non-embedded & embedded peptide star graph descriptors
{Sh,Tr,X,X1R,H,W,S6,S,J} & {eSh,eTr,eX,eX1R,eH,eW,eS6,eS,eJ}
Machine Learning
Weka software
FASTA protein sequences of unknown biological function
PDB
Best QSAR classification model for signaling protein sequence prediction
Prediction of new signaling protein sequences

Fig. 1. Methodology flowchart including molecular graphs and Machine Learning for signaling protein classification model.