JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

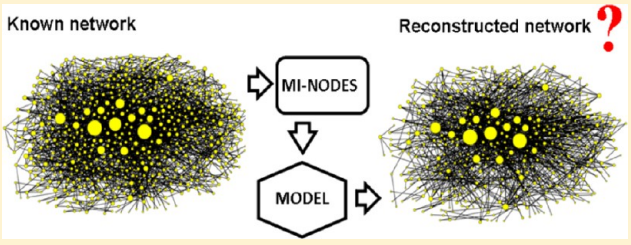
# New Markov-Autocorrelation Indices for Re-evaluation of Links in Chemical and Biological Complex Networks used in Metabolomics, Parasitology, Neurosciences, and Epidemiology

Humberto González-Díaz* and Pablo Riera-Fernández

[1]Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela (USC), 15782 Santiago de Compostela, Spain

Supporting Information

**ABSTRACT:** The development of new methods for the computational re-evaluation of links in chemical and biological complex networks is very important to save time and resources. The Moreau–Broto autocorrelation indices (MBis) are well-known topological indices (TIs) used in QSAR/QSPR studies to encode the structural information contained in molecular graphs. In addition, MBis and similar autocorrelation measures have been used to study other systems like, for example, proteins. In the present work, MBis are combined with Markov chains to develop a general class of stochastic MBis of order $k$ ($MB_k$) that is used to encode the structural information contained in different types of large complex networks. The $MB_k$ values obtained for the nodes (centralities) of these networks are used as input variables to seek QSPR-like equations (by means of linear discriminant analysis) in which the outputs are numerical scores $S(L_{ij})$ that allow us to discriminate between connected and nonconnected nodes and therefore re-evaluate the connectivity of the whole network. The models developed in this work produced the following results in terms of overall accuracy for network reconstruction: metabolic networks (72.10%), parasite–host networks (88.70%), CoCoMac brain cortex coactivation network (81.89%), and fasciolosis spreading network (86.39%).




## 1. INTRODUCTION

In the 1980s, Moreau and Broto applied an autocorrelation function to the molecular graph in order to measure the distribution of atomic properties on the molecular topology. This measure was called autocorrelation of topological structure (ATS) or more commonly Moreau–Broto autocorrelation.[1−3] After its introduction, this topological indice (TI) has been successfully applied in the search of quantitative structure–activity/property relationship (QSAR/QSPR) models.[4−9] These models predict the biological activity (QSAR) or physicochemical properties (QSPR) of molecules (output) using as input numerical parameters of the molecular structure (like MBis or others). In parallel, the idea of ATS has been reformulated in different ways in order to incorporate more information to study different molecular systems. For example, Moro et al.[10] used the autocorrelation of molecular electrostatic potential surface properties for the prediction of the activity of human A(3) adenosine receptor antagonists. More recently, the use of MBis and other ATS-like indices have been extended to carry out QSPR/QSAR-like studies in proteins. Almost all the models reported use as input some type of amino acid sequence autocorrelation (AASA) vectors, as in the excellent works of Caballero and Fernández.[11−14] Some of the models developed have been also implemented as online web servers, such as IUPforest-L[15] and PROFEAT.[16] Another extension of MBis in a different direction is their incorporation into a new computer program called S2SNet (Sequence to Star Networks).[17] This program was designed by our group in order to transform any character sequence into shining (sequence memory less) or sequence-embedded star networks. As seen in these works, it is possible to extend the use of MBis to study not only drugs or proteins but also mass spectra signals of proteins, 1D NMR signals, IR spectra, time series data, texts, and any other type of string data. This fact suggests that MBis should be good candidates to be used in the study of more complex systems.

Many complex network-like systems can be represented by very large graphs. This has led to the accumulation of a large number of network models that reflect drug–target interactions, protein structure, protein interactions, metabolic pathways, brain structure, disease spreading, social–legal relationships, landscape connectivity, internet structure, etc.[18−27] In all these systems there are different experimental and/or theoretical methods to assign node–node links depending on the type of network we want to construct. Unfortunately, many of these methods are expensive in terms of time or resources, and therefore it is difficult or impossible to repeat the measures to confirm the results obtained. In addition, a common way to construct complex networks is to use information obtained from various scientific works in which the sampling effort and/or the methods used to study the same type of interaction are different. These factors are responsible

**Received:** July 9, 2012
**Published:** November 3, 2012

   dx.doi.org/10.1021/ci300321f | J. Chem. Inf. Model. 2012, 52, 3331−3340

for the introduction of bias in the constructed networks. One possible solution to this problem is the use of QSAR/QSPR-like models based on TIs.[28−34] The idea is to seek a QSPR-like model able to discriminate reliable connections in a complex network using as input the structural information contained in its nodes and encoded by means of TIs.

Recently, our group has developed a computer program called MI-NODES (March-Inside NOde DEScriptors) that combines Markov chains with classic TIs. In this work, we introduce and use for the first time $MB_k(j)$ values calculated with MI-NODES to develop different QSPR-like models able to assess the quality of the connectivity of new complex networks assembled with information obtained from many sources not totally accurate. In these models (linear equations obtained after a linear discriminant analisys), the $MB_k(j)$ values calculated for a pair of nodes are used as input, and the output is used to decide if the link between the two components of the pair is possible. Although very different systems were studied, the same workflow was used in all the experiments (see Figure 1). In the first experiment we studied metabolic pathway networks of four different organisms (bacteria, yeast, nematode, plant). In the second experiment we studied four parasite−host interaction networks (PHIs). The third experiment consisted of carrying out a study regarding connectivity quality in the collations of connectivity data on the *Macaque* brain (CoCoMac) cerebral cortex coactivation network.[35] In the fourth experiment we studied a macroscopic landscape parasitism-spreading network for cattle fasciolosis in NW Spain.

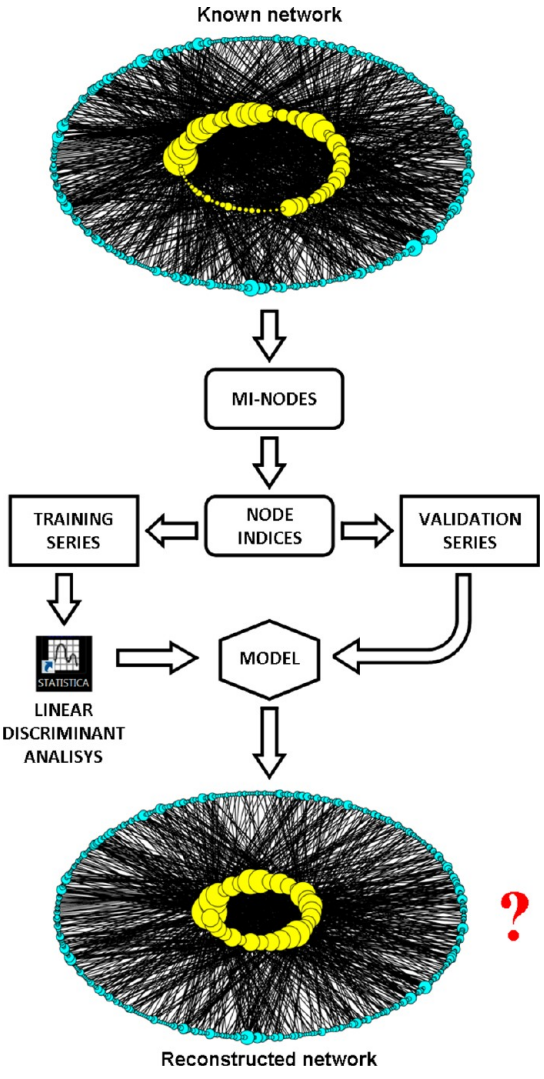



**Figure 1.** General workflow used in this work.

## 2. MATERIALS AND METHODS

**2.1. Data Sets.** *2.1.1. Metabolic Pathway Networks.* Metabolic network data were downloaded directly from Barabasi's group Web site (http://www.nd.edu/~networks/resources.htm) as a gzipped ASCII file. In this file each number represents a substrate in the metabolic network of corresponding organism. Data format is: From → To (directed link). The information studied was previously obtained by Jeong et al. from the 'intermediate metabolism and bioenergetics' portions of the WIT database and used in order to try to understand the large-scale organization of metabolic networks.[36]

*2.1.2. Parasite−Host Complex Networks.* The four parasite−host networks used here (parasite−fish, parasite−ungulates, parasite−carnivores, and parasite−primates) were constructed and published by our group in previous works,[37,38] and the original data were obtained from two databases: the interaction web database (IWDB) (http://www.nceas.ucsb.edu/interactionweb/index.html) and the global mammal parasite database (GMPD) (http://www.mammalparasites.org/).

*2.1.3. Cerebral Cortex Coactivation Network.* Modha and Singh, in their work 'network architecture of the long-distance pathways in the macaque brain',[35] studied the information contained in the CoCoMac (http://www.cocomac.org-) neuroinformatic database in order to construct the most comprehensive long-distance network of the *Macaque* brain. The final network obtained consists of 383 hierarchically organized regions spanning cortex, thalamus, and basal ganglia; models the presence of 6602 directed long-distance connections (three times larger than any previously derived brain network) and contains subnetworks corresponding to classic corticocortical, corticosubcortical, and subcortico-subcortical fiber systems.

*2.1.4. Complex Network for Fasciolosis Spreading in NW Spain.* The data set reported by Mezo et al.[39] in a previous work was used by our group to construct a network of farm-to-farm spreading of fasciolosis in cattle for Galicia (NW Spain) in another work.[40] In this case, each farm was considered as a node of the network associated to a Boolean or connectivity matrix **C** with elements $C_{ij}$ (links). The existence of a connection *ij* implies the existence of the inverse connection *ji*. Loops, connections from *j*th to the same *j*th farm (representing self-infection of animals inside the same farm) were allowed. We place an arc (directed edge) connecting the *i*th farm with the *j*th farm if they meet the condition given in the Microsoft excel command (see next equations) that is used to truncate the farm-to-farm distance function. The connectivity of the network **C** depends on the input parameters: spatial coordinates $(x_i, y_i)$ of the farm $(f_i)$, altitude of the place $(h_i)$, and strength of the drug treatment $(Tr_i = 0, 1, 2, 3)$ used for the pre-existent fasciolosis in this farm ($Tr_i = 0$ indicates that the disease was not detected). Consequently the matrix **C** quantifies the propensity $C_{ij} = 1$ of the disease to spread between farms immediately after treatment. On the other hand, the matrix **L** includes two criteria: the pre-existence of a high propensity for disease spreading $C_{ij} = 1$ and the experimental confirmation of a high risk ratio $(RR_{ij})$ of prevalence after


dx.doi.org/10.1021/ci300321f | J. Chem. Inf. Model. 2012, 52, 3331−3340

Table 1. Training and Validation Results for All Models Developed in This Work[a]

| training series | | | | model | validation series | | | |
|---|---|---|---|---|---|---|---|---|
| NL | L | group | % | parameters | % | group | NL | L |
| Metabolic Pathway Networks | | | | | | | | |
| 46594 | 17925 | NL | 72.22 | specificity | 72.28 | NL | 15545 | 5962 |
| 3013 | 7467 | L | 71.25 | sensitivity | 71.24 | L | 1,005 | 2489 |
| | | total | 72.08 | accuracy | 72.13 | total | | |
| Parasite−Host Networks | | | | | | | | |
| 39047 | 5581 | NL | 87.49 | specificity | 87.67 | NL | 12972 | 1824 |
| 0 | 4590 | L | 100 | sensitivity | 100 | L | 0 | 1521 |
| | | total | 88.66 | accuracy | 88.82 | total | | |
| Cerebral Cortex Co-activation Network | | | | | | | | |
| 16832 | 3172 | NL | 84.14 | specificity | 84.42 | NL | 5658 | 1044 |
| 1352 | 3600 | L | 72.70 | sensitivity | 71.88 | L | 464 | 1186 |
| | | total | 81.87 | accuracy | 81.94 | total | | |
| Fasciolosis Spreading Network (NW Spain) (L) | | | | | | | | |
| 19178 | 2830 | NL | 87.14 | specificity | 87.34 | NL | 6381 | 925 |
| 374 | 995 | L | 72.68 | sensitivity | 75.78 | L | 109 | 341 |
| | | total | 86.29 | accuracy | 86.67 | total | | |

[a]Rows: Observed classifications; columns: predicted classifications; NL: not linked; L: linked.

treatment ($PAT_j$) of the disease. See the definition of these networks in mathematical terms using excel functions:

$$L_{ij} = \text{if}\,(\text{AND}(C_{ij} = 1,\ R_{ij} > 1) \tag{1}$$

$$RR_{ij} = (PAT_i + 1)/(PAT_j + 1) \tag{2}$$

$$C_{ij} = \text{if}\left(\text{OR}\left(d_{ij} > d_{\text{cutoff}} \times \frac{1}{m}\text{Sum}(d_{ij}),\ d_{ij} = 0\right),\ 0,\ 1\right) \tag{3}$$

$$d_{ij} = 0.5 \times (h_i + h_j) \times \text{Tr}_i \times \text{Tr}_j \times \text{SQR}((x_i - x_j)^2 + (y_i - y_j)^2) \tag{4}$$

**2.2. Computational Methods.** *2.2.1. Stochastic MB Centralities for Nodes.* The classical Markov matrix ($^1\mathbf{\Pi}$) for each network is constructed as follows: First, we obtain from different sources (see the previous section) the connectivity matrix (**L**, a $n \times n$ matrix, where $n$ is the number of nodes) or the data about the links between nodes to assemble it. Next, based on **L**, the Markov matrix $^1\mathbf{\Pi}$ is built. It contains the probability of reaching a node from another in one step ($^1p_{ij}$). Essentially the same class of Markov or transition matrices have been used before to define different numerical parameters useful in the study of complex networks.[37,41−44] The probability matrix is raised to the power $k$, resulting $(^1\mathbf{\Pi})^k$. These matrices contain the probabilities of reaching a node from another in $k$ steps and are used for the calculation of the stochastic Moreau−Broto centrality values ($MB_k(j)$ of $j$th node):

$$^kMB_k(j) = \frac{1}{2}\sum_{i=1}^{n} {}^kp_{ij} \cdot {}^kp_{ji} \tag{5}$$

*2.2.2. MI-NODES Software for Calculation of Stochastic MBis.* MI-NODES is a GUI Python/wxPython application used for the calculation of a new class of centralities/topological indices of nodes, subnetworks, or whole networks. Actually, it should be considered as the generalization of the software March-Inside to manage any kind of complex networks (this program was originally designed to study drugs, proteins, and nucleic acid structures).

MI-NODES calculates new types of node centralities $^kC_c(j)$, based on Markov normalized node probabilities, without removing each node previously to perform calculations. It also calculates Markov generalizations of different topological indices $^kTI_c(G)$ of class $c$ and power $k$ for the graph $G$. The tool is both Pajek and CentiBin compatible (it reads networks in the following formats: .net, .dat, and .mat.).[43]

*2.2.3. LDA Models.* Linear Discriminant Analysis (LDA) is a multivariate statistical method used to find a linear combination of features that discriminates between two or more classes of objects or events. Possibly, this is the most common technique used in QSPR/QSAR studies with TIs of molecular graphs, protein, and RNA structure networks and biomolecular complex networks. In this work LDA is used to seek a linear function able to discriminate between two classes of pairs of nodes, linked and nonlinked (see Figure 1). The data necessary to train the model are obtained from the different systems studied and include two types of pairs of nodes (categorical dependent variable): linked ($L_{ij} = 1$) and nonlinked ($L_{ij} = 0$). Since the number of linked pairs is much lower than the not linked ones, we did not use all the possible negative cases in order to avoid statistical bias. In addition, previous to the training, 25% of the cases were chosen randomly and separated to carry out a cross-validation of the model. The continuous dependent variables used are: $MB_k(i)$, $MB_k(j)$ and $[MB_k(i) - MB_k(j)]$, where the terms $i$ and $j$ are referred to the nodes that compose the pair and $k$ is calculated for 0−5. Therefore we have a total of 18 variables that encode information of a node and its neigbors (placed at a topological distance $d = k$). For each group ($L_{ij} = 1$ and 0) in the training set, we can determine the location of the point (centroid) that represents the means for all variables in the multivariate space defined by the variables in the model. The idea is to find the linear function that maximizes the Mahalanobis distance between these centroids. This discriminant function has the following form:


dx.doi.org/10.1021/ci300321f | J. Chem. Inf. Model. 2012, 52, 3331−3340

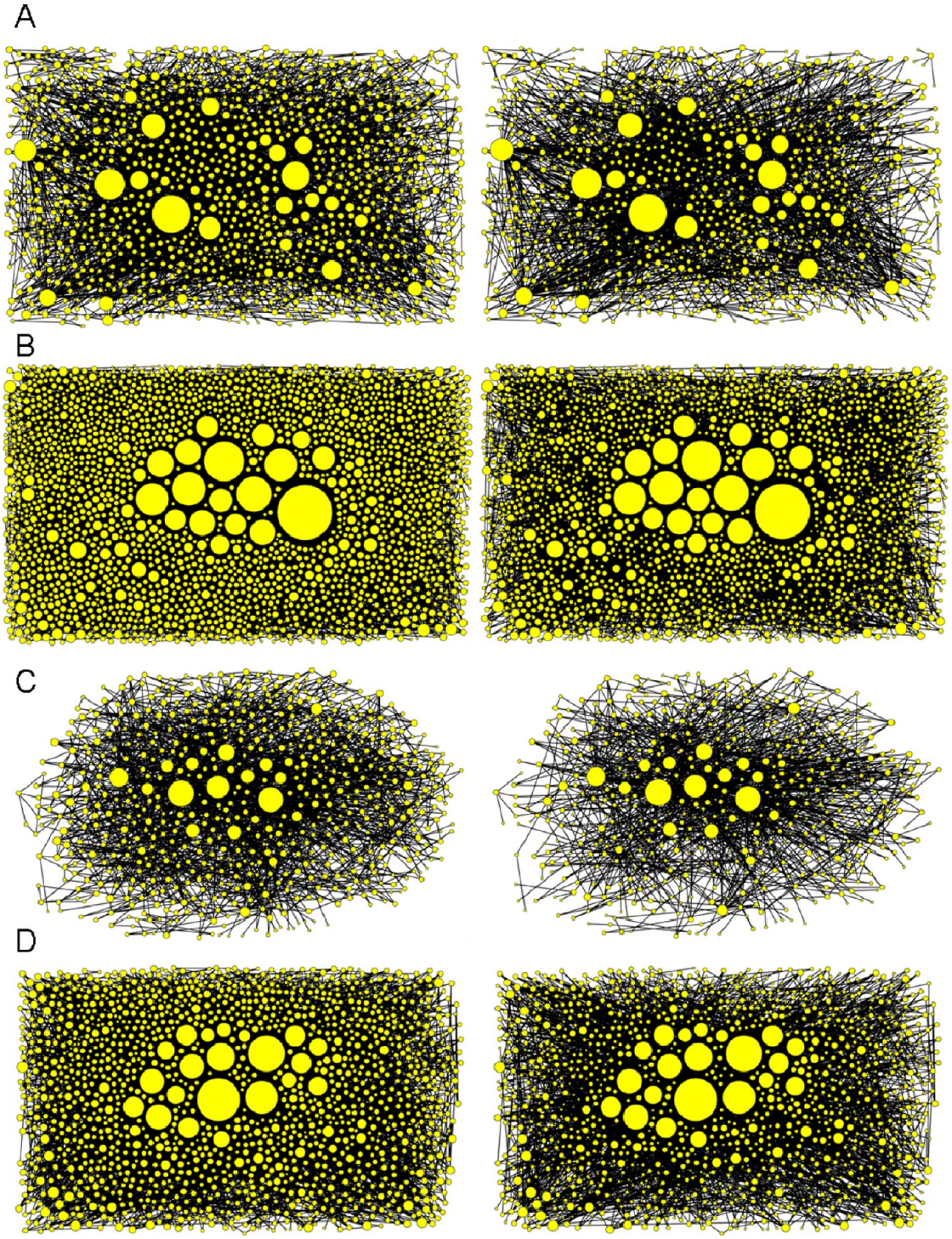



**Figure 2.** Observed (left) and reconstructed (right) metabolic networks. (A): CE, (B): EC, (C): OS, (D) SC. The size of each node represents its degree.

$$S(L_{ij}) = \sum_{k=0}^{5} a_{ik}\cdot \mathrm{MB}_k(i) + \sum_{k=0}^{5} a_{jk}\cdot \mathrm{MB}_k(j) + \sum_{k=0}^{5} a_{ijk}\cdot [\mathrm{MB}_k(i) - \mathrm{MB}_k(j)] + a_0 \quad (6)$$

where $a_{ik}$, $a_{jk}$, and $a_{ijk}$ are the coefficients of the variables and $a_0$ the independent term. $S(L_{ij})$ is the output variable (a real number). It is important to bear in mind that the objective is not only to find an equation with the maximum discrimination power but also with the least number of variables. This can be achieved by using a stepwise procedure. For example, in our case we have chosen a forward stepwise approach (starting with no variables and testing the inclusion of each variable). Once we have the discriminant function, we can use it to classify new cases. To do this, we need a decision boundary that can be calculated as the mean of the two $S(L_{ij})$ values obtained for the centroids. Thus, a case is classified into the first or the second group depending on whether its $S(L_{ij})$ value is below or above


dx.doi.org/10.1021/ci300321f | J. Chem. Inf. Model. 2012, 52, 3331−3340

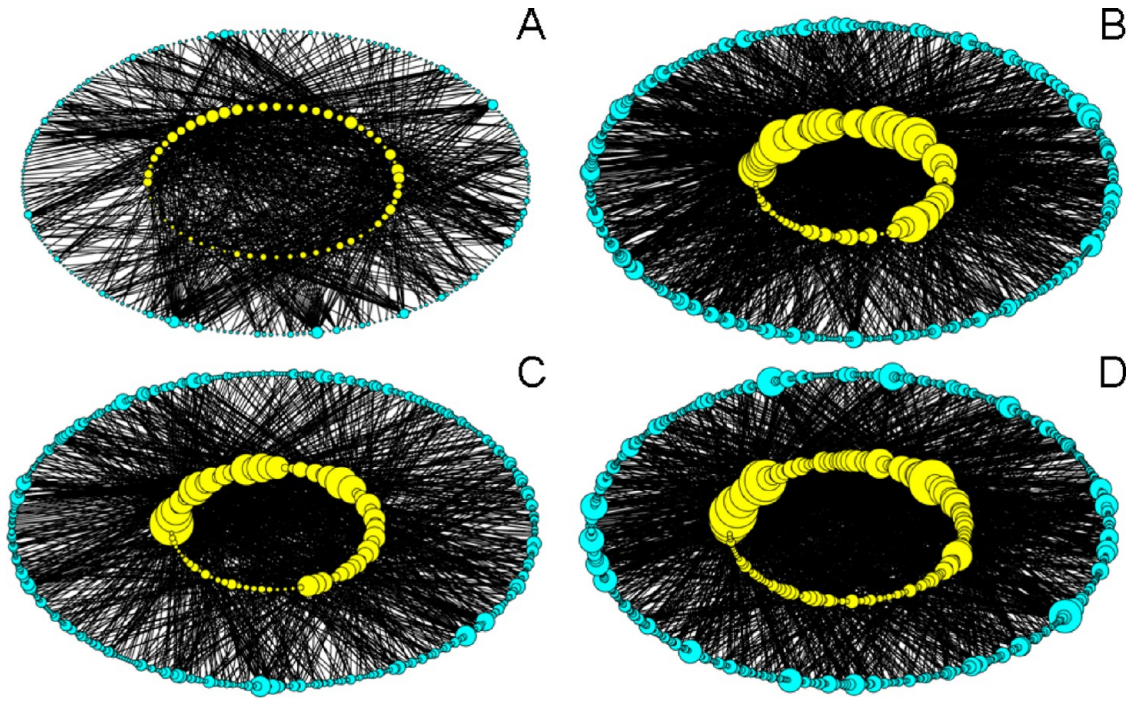



**Figure 3.** Parasite−host networks: (A) Parasite−fish, (B) parasite−ungulate, (C) parasite−carnivore, (D) parasite−primate. The size of each node represents its degree. Outer nodes = parasites and inner nodes = hosts. The observed and reconstructed networks are equal for each complex system, and therefore they are only represented once.

the decision limit. An additional factor that should be taken into consideration is that sometimes the probabilities of belonging to one or another group are different. If these a priori probabilities are known, it is possible to adjust the decision boundary to improve the classification. Different parameters can be used to evaluate the statistical significance and validate the goodness-of-fit of the equation obtained. In the results section, in addition to the discriminant function, we include the following information: $n$ = number of cases, $\chi^2$ = chi squared, and $p$ = p level as well as the classification matrix for the training and external validation series.[45] From these matrices we can obtain the sensitivity = true positives/(true positives + false negatives), specificity = true negatives/(true negatives + false positives), and accuracy = true negatives + true positives/(true negatives + true positive + false negatives + false positives).

Fortunately, there are many statistical packages able to carry out a LDA. In our case we have used the program STATISTICA.

## 3. RESULTS AND DISCUSSION

**3.1. Model 1: Metabolic Pathway Networks.** Study of metabolic networks is relevant to shed light over the mechanisms that control metabolic diseases as well as to control infections by making use of the metabolic differences between human beings and pathogens.[46] However, Jeong et al.[36] showed that, despite significant variation in their individual constituents and pathways, these metabolic networks have the same topological scaling properties and show striking similarities to the inherent organization of complex non-biological systems. Many pathways are not totally confirmed experimentally but have been computationally deduced using protein or gene alignment techniques. On the other hand, the experimental determination of the full metabolome is a hard experimental task. All this aspects determine the necessity of alignment-free techniques to assess network connectivity quality in existing models of metabolic pathway networks. For this analysis we have used metabolic networks of four model organisms belonging to different domains of the tree of life. These organisms are: *Escherichia coli* (EC), *Saccharomyces cerevisiae* (SC), *Caenorhabditis elegans* (CE), and *Oryza sativa* (OS). EC is a gram negative bacterium that is commonly found in the lower intestine of warm-blooded organisms. Most EC strains are harmless, but some serotypes can cause serious food poisoning in humans. In addition, many models of metabolic networks are based on this prokaryotic organism.[47−59] Yeast, SC is of the major relevance for biotechnology industry, being the first eukaryotic genome fully sequenced and annotated.[60] CE (the first multicellular organism to have its genome completely sequenced) is a free-living nematode very useful for genetic research, since it is easy to maintain and has a very fast life cycle.[61,62] That is why the comparison of CE vs other parasitic nematodes is an invaluable tool to unravel parasitism-related gene in parasitology.[63] CE is also a model for the study of antihelmintic drugs that are used against parasitic worm infections of humans and livestock.[64] Finally, rice, OS, is a plant of great economic importance for the human being and the most widely studied model for cereals.[65] The best QSPR model found for the metabolic networks of all these organisms was

$$S(L_{ij}) = 23.44\cdot \mathrm{MB}_5(e_i) - 5.59\cdot[\mathrm{MB}_3(e_i) - \mathrm{MB}_3(p_j)] - 0.73$$
$$n = 74999, \quad \chi^2 = 20143, \quad p < 0.001 \tag{7}$$

In this equation, $S(L_{ij})$ is a real-valued output variable that scores the propensity of the $i$th input or educt $(e_i)$ (reactant or substrate) to undergo a metabolic transformation into the product $(p_j)$. The model presents good values of accuracy,


dx.doi.org/10.1021/ci300321f | J. Chem. Inf. Model. 2012, 52, 3331−3340

Table 2. Some Examples of $\Delta MB_2$ Values Used As Input in the Model for PHIs

| parasite (node A) | host (node B) | $\Delta MB_2$ | parasite (node A) | host (node B) | $\Delta MB_2$ |
|---|---|---|---|---|---|
| *Anisakis pacificus* | *Phoca fasciata* | −0.465 | *Plasmodium knowlesi* | *Presbytis melalophos* | −0.461 |
| *Anisakis similis* | *Eumetopias jubatus* | −0.461 | *Plasmodium pitheci* | *Pongo pygmaeus* | −0.465 |
| *Anisakis simplex* | *Phoca hispida* | −0.462 | *Plasmodium schwetzi* | *Gorilla gorilla* | −0.467 |
| *Anisakis simplex* | *Lutra lutra* | −0.462 | *Plasmodium shortii* | *Macaca radiata* | −0.462 |
| *Anisakis simplex* | *Phoca fasciata* | −0.462 | *Plasmodium shortii* | *Macaca sinica* | −0.462 |
| *Entamoeba coli* | *Colobus angolensis* | −0.465 | *Plasmodium simiovale* | *Macaca sinica* | −0.463 |
| *Entamoeba coli* | *Colobus badius* | −0.465 | *Toxoplasma gondii* | *Enhydra lutris* | −0.465 |
| *Entamoeba coli* | *Erythrocebus patas* | −0.465 | *Toxoplasma gondii* | *Felis silvestris* | −0.465 |
| *Entamoeba coli* | *Gorilla gorilla* | −0.465 | *Toxoplasma gondii* | *Lontra canadensis* | −0.465 |
| *Entamoeba coli* | *Macaca fuscata* | −0.465 | *Toxoplasma gondii* | *Lynx canadensis* | −0.465 |
| *Entamoeba coli* | *Macaca sinica* | −0.465 | *Toxoplasma gondii* | *Lynx lynx* | −0.465 |
| *Entamoeba histolytica* | *Colobus badius* | −0.466 | *Toxoplasma gondii* | *Lynx rufus* | −0.465 |
| *Entamoeba histolytica* | *Macaca fuscata* | −0.466 | *Toxoplasma gondii* | *Ursus americanus* | −0.465 |
| *Entamoeba histolytica* | *Macaca sinica* | −0.466 | *Toxoplasma gondii* | *Vulpes vulpes* | −0.465 |
| *Entamoeba histolytica* | *Mandrillus sphinx* | −0.466 | *Treponema endemicum* | *Erythrocebus patas* | −0.466 |
| *Entamoeba histolytica* | *Gorilla gorilla* | −0.466 | *Treponema pallidum* | *Gorilla gorilla* | −0.466 |
| *Entamoeba histolytica* | *Papio cynocephalus* | −0.466 | *Treponema pertenue* | *Gorilla gorilla* | −0.467 |
| *Entamoeba histolytica* | *Papio hamadryas* | −0.466 | *Treponema pertenue* | *Papio anubis* | −0.467 |
| *Fasciola gigantica* | *Sus scrofa* | −0.466 | *Trichinella britovi* | *Ursus thibetanus* | −0.466 |
| *Fasciola gigantica* | *Kobus leche* | −0.466 | *Trichinella britovi* | *Vulpes vulpes* | −0.466 |
| *Fasciola gigantica* | *Tragelaphus strepsiceros* | −0.466 | *Trichinella britovi* | *Ursus arctos* | −0.466 |
| *Fasciola hepatica* | *Lutra lutra* | −0.463 | *Trichinella murrelli* | *Ursus americanus* | −0.467 |
| *Fasciola hepatica* | *Oryx gazella* | −0.467 | *Trichinella murrelli* | *Vulpes vulpes* | −0.467 |
| *Fasciola hepatica* | *Sus scrofa* | −0.467 | *Trichinella murrelli* | *Procyon lotor* | −0.467 |
| *Giardia duodenalis* | *Gorilla gorilla* | −0.465 | *Trichinella nativa* | *Gulo gulo* | −0.465 |
| *Giardia duodenalis* | *Papio anubis* | −0.465 | *Trichinella spiralis* | *Lutra lutra* | −0.466 |
| *Giardia duodenalis* | *Papio ursinus* | −0.465 | *Trichinella spiralis* | *Martes americana* | −0.466 |
| *Giardia duodenalis* | *Saimiri oerstedii* | −0.465 | *Trichinella spiralis* | *Mustela vison* | −0.466 |
| *Giardia duodenalis* | *Colobus badius* | −0.465 | *Trichinella spiralis* | *Ursus americanus* | −0.466 |
| *Giardia duodenalis* | *Papio hamadryas* | −0.465 | *Trichinella spiralis* | *Vulpes vulpes* | −0.466 |
| *Necator americanus* | *Papio anubis* | −0.466 | *Trichinella spiralis* | *Ursus arctos* | −0.466 |
| *Necator americanus* | *Papio cynocephalus* | −0.466 | *Trichuris suis* | *Sus scrofa* | −0.467 |
| *Necator congolensis* | *Pan troglodytes* | −0.467 | *Trichuris trichiura* | *Gorilla gorilla* | −0.466 |
| *Plasmodium brasilianum* | *Saimiri sp.* | −0.460 | *Trichuris trichiura* | *Macaca tonkeanna* | −0.466 |
| *Plasmodium brasilianum* | *Pithecia sp.* | −0.460 | *Trichuris trichiura* | *Papio anubis* | −0.466 |
| *Plasmodium coatneyi* | *Presbytis cristata* | −0.465 | *Trichuris trichiura* | *Papio papio* | −0.466 |
| *Plasmodium coatneyi* | *Macaca fascicularis* | −0.465 | *Trypanosoma brucei* | *Perodicticus potto* | −0.459 |
| *Plasmodium cynomolgi* | *Macaca cyclopis* | −0.463 | *Trypanosoma brucei* | *Miopithecus talapoin* | −0.459 |
| *Plasmodium cynomolgi* | *Macaca fascicularis* | −0.463 | *Trypanosoma brucei* | *Ourebia ourebi* | −0.459 |
| *Plasmodium eylesi* | *Hylobates lar* | −0.447 | *Trypanosoma cruzi* | *Saguinus geoffroyi* | −0.461 |
| *Plasmodium falciparum* | *Pithecia pithecia* | −0.465 | *Trypanosoma cruzi* | *Saguinus midas* | −0.461 |
| *Plasmodium falciparum* | *Saimiri sciureus* | −0.465 | *Trypanosoma cruzi* | *Saguinus nigricollis* | −0.461 |
| *Plasmodium knowlesi* | *Macaca fascicularis* | −0.461 | *Trypanosoma vivax* | *Miopithecus talapoin* | −0.454 |
| *Plasmodium knowlesi* | *Macaca nemestrina* | −0.461 | *Trypanosoma vivax* | *Perodicticus potto* | −0.454 |

sensitivity, and specificity for the recognition of links both in training and external validation series (see Table 1). In Figure 2 we depict the metabolic networks observed vs those obtained after re-evaluating link quality with our QSPR model.

**3.2. Model 2: Parasite−Host Networks.** Due to the importance for the human and animal health and therefore for the economy, much attention has been focused on PHIs. The study of these interactions can help us to understand the role of phylogenetics and ecological factors on the parasite−host specificity[66−68] and to know how parasites affect the ecosystem functioning.[69−71] In this sense, network theory is a useful tool for analyzing this type of interactions.[72] However, the high experimental difficulty inherent to the accurate in situ determination of PHIs makes the possibility of curate PHIs networks using a computational model very interesting. In this work, we used $MB_k$ to seek a QSPR-like model able to score the quality of PHIs in known networks. The best model found was:

$$S(L_{ij}) = 4.59\cdot[MB_2(p_i) - MB_2(h_j)] + 0.21$$
$$n = 49218, \quad \chi^2 = 15801, \quad p < 0.001 \qquad (8)$$

In this equation, $S(L_{ij})$ is a real-valued output variable that scores the propensity of the $i$th parasite specie $(p_i)$ to infect a given host specie $(h_j)$. The $\chi^2$ presents a low $p$-level < 0.001, which indicates a significant discrimination between well-established host−parasite relationships and not confirmed parasitism. The observed and predicted (network reconstructed using this model) have been graphically represented in Figure 3.

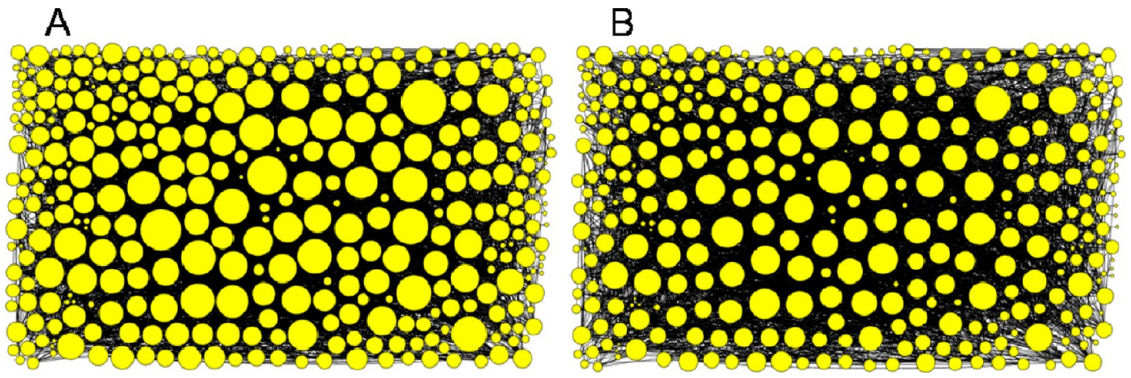



**Figure 4.** Graphical representation of the cortex coactivation network: (A) observed and (B) reconstructed networks. The size of each node represents its degree.

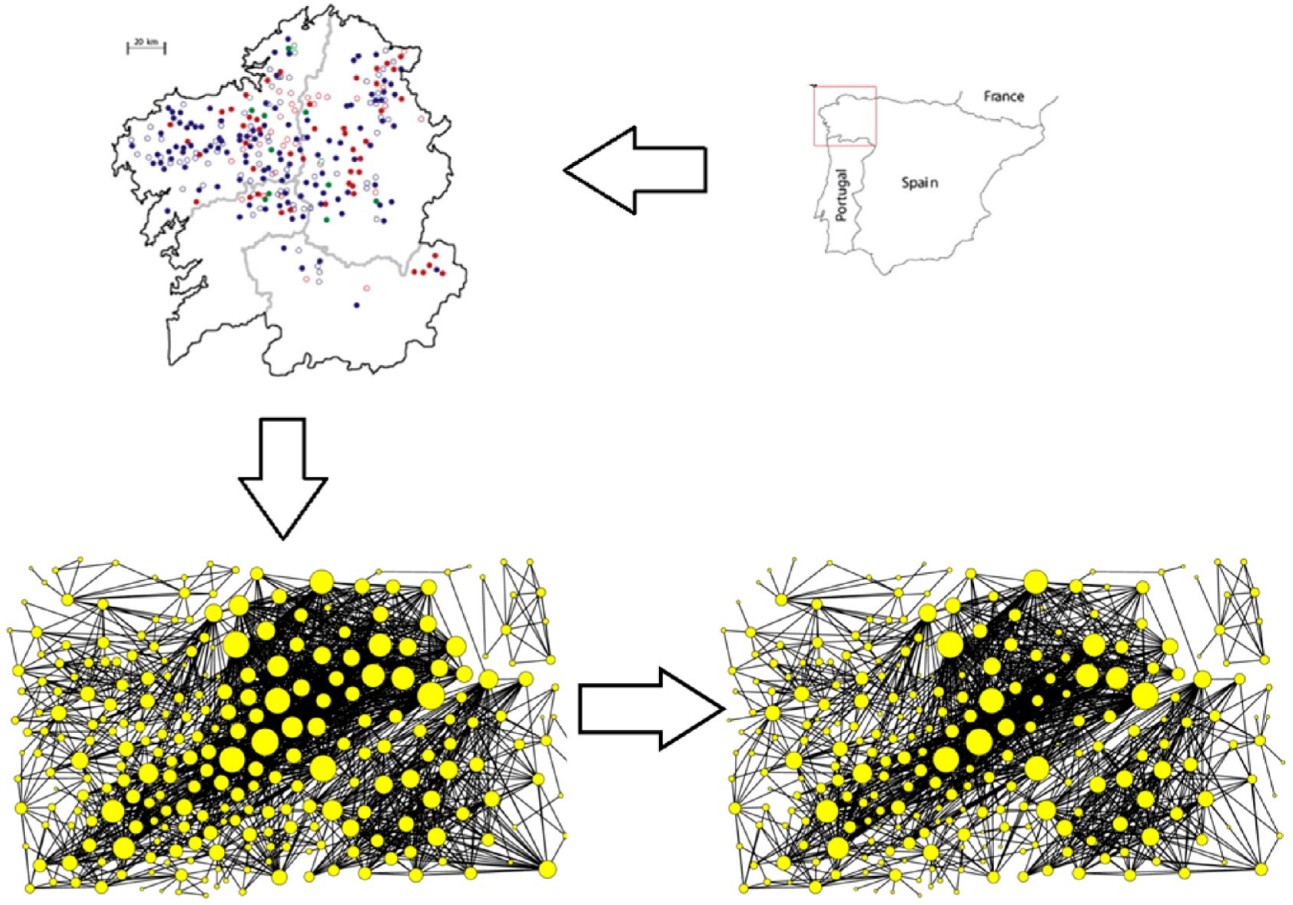



**Figure 5.** Top: Geographical map of Galicia (NW Spain) showing the location of the 275 sampled farms: the status of infection (empty circles: *F. hepatica* free and solid circles: *F. hepatica* infected) and the treatment administered on each farm are shown (blue, none; red, an anthelmintic effective against fluke mature stages; and green, a fasciolicide effective against immature and mature stages). Bottom: Observed fasciolosis landscape-spreading network (left) and reconstructed network (right). The size of each node represents its degree.

The model presents very good values of accuracy, sensitivity, and specificity for the recognition of parasite−host relationships (links) both in training and external validation series (see Table 1). Consequently, with this simple linear model we could re-evaluate the connectivity quality in the already-known PHIs networks in a fast and nonexpensive way (without having to experimentally resample all PHIs in the corresponding ecological niche). In the Supporting Information of this model, we describe the training/cross validation sets including: host and parasite names, host−parasite $\Delta MB_2$ value (= $[MB_2(p_i) - MB_2(h_j)]$) and posterior probability $p(L)$. In addition, in Table 2 we depict the values of $\Delta MB_2$ for some examples of parasite−host pairs used both in training and cross validation. All of them have been correctly predicted with $p(L)$ values > 0.9. In this example we can see that the QSPR model is able to predict the infection of other species by parasites that are also important human pathogens, such as *Plasmodium falciparum* (causal agent of Malaria), *Trypanosoma cruzi* (Chagas disease), or *Toxoplasma gondii* (opportunistic infections in HIV-infected patients) among others. This could be important for the study of the transmission of infectious agents of emerging zoonosis in multihost systems.[73,74]


dx.doi.org/10.1021/ci300321f | J. Chem. Inf. Model. 2012, 52, 3331−3340

**3.3. Model 3: Cerebral Cortex Coactivation Network.** Connectivity is the key to understanding distributed and cooperative brain functions. Detailed and comprehensive data on large-scale connectivity between primate brain areas have been collated systematically from published reports of experimental tracing studies.[75] Databasing the brain's anatomical connectivity as delivered by tracing studies is of particular importance as these data characterize fundamental structural constraints of the complex and poorly understood functional interactions between the components of real neural systems. The eventual impact and success of connectivity databases, however, will require the resolution of several methodological problems that currently limit their use. These problems comprise four main points: (i) objective representation of coordinate-free, parcellation-based data; (ii) assessment of the reliability and precision of individual data, especially in the presence of contradictory reports; (iii) data mining and integration of large sets of partially redundant and contradictory data; and (iv) automatic and reproducible transformation of data between incongruent brain maps.[76] In order to address points (ii) and (iv), we have developed a specific model for the CoCoMac database (http://www.cocomac.org). The best model found was

$$S(L_{ij}) = 12.74\cdot[\mathrm{MB}_1(i) - \mathrm{MB}_1(j)] - 0.80$$
$$n = 24956, \quad \chi^2 = 9422, \quad p < 0.001 \qquad (9)$$

In this equation, $S(L_{ij})$ is a real-valued output variable that scores the propensity of the $i$th cerebral cortex region to undergo coactivation with the $j$th region in the CoCoMac network. The parameter $\mathrm{MB}_k$ quantifies the information related to the position of the $i$th region and their direct neighbors ($j$th regions) in the network after $k$ steps. As in the previous equation, the $\chi^2$ statistics corresponds to a p-level < 0.001, which indicates a significant discrimination between coactivated regions and not coactivated ones. The model presents good values of accuracy, sensitivity, and specificity (see Table 1). The observed and reconstructed networks are represented graphically in Figure 4.

**3.4. Model 4: Fasciolosis Spreading Network (NW Spain).** Fasciolosis is a parasitic infection caused by *Fasciola hepatica* (liver fluke) that has become an important cause of lost productivity in livestock worldwide. Considered a secondary zoonotic disease until the mid-1990s, human fasciolosis is at present emerging or re-emerging in many countries, including increases of prevalence and intensity and geographical expansion. In fact, research in recent years has justified the inclusion of fasciolosis in the list of important human parasitic diseases. At present, fasciolosis is the vector-borne disease presenting the widest latitudinal, longitudinal, and altitudinal distribution known. In addition, it presents a range of epidemiological characteristics related to a wide diversity of environments.[77] In this sense, the study of geographical spreading of fasciolosis becomes a subject of great interest. In fact, in a recent work we have constructed a network to study the landscape spreading of fasciolosis in Galicia (NW Spain).[40] However, we do not have quantitative criteria on the quality of the network connectivity, and resampling of all data to re-evaluate this connectivity in a field study is a hard and expensive task in terms of time and resources. This situation has prompted us to seek models in order to assess the quality of the networks previously assembled. The best QSPR model found was

$$S(L_{ij}) = -11.50\cdot\mathrm{MB}_3(f_i) - 18.26\cdot[\mathrm{MB}_1(f_i) - \mathrm{MB}_1(f_j)] - 0.07$$
$$n = 23377, \quad \chi^2 = 3897, \quad p < 0.001 \qquad (10)$$

The values $\mathrm{MB}_k(f_i)$ and $\mathrm{MB}_k(f_j)$ used in this equation quantify information about the connectivity patterns between farms in the network. As can be seen in the equations described in Materials and Methods section, the connectivity of **C** depends on the spatial coordinates $(x_i, y_i)$ of the farm $(f_i)$, altitude of the place $(h_i)$, and antiparasitic drug treatment $(\mathrm{Tr}_j)$ used to prevent fasciolosis in this farm. Consequently the matrix **C** quantifies the a priori propensity $C_{ij} = 1$ of this disease to spread between farms immediately after treatment depending on geographical conditions. On the other hand, matrix **L** includes both criteria: (i) the preexistence of a high propensity for disease spreading $C_{ij} = 1$ and (ii) the experimental confirmation $L_{ij} = 1$ of a high risk ratio $(\mathrm{RR}_{ij})$ of prevalence after treatment $(\mathrm{PAT}_j)$ for this disease in farms. The QSPR model developed (based in **L**) presents good values of accuracy, sensitivity, and specificity (see Table 1). Both observed and reconstructed networks are represented graphically in Figure 5.

## 4. CONCLUSIONS

In this work we confirm that the MB autocorrelation descriptors can be combined with Markov chains in order to calculate higher order parameters $(\mathrm{MB}_k)$. We also show that these parameters can be used to quantify information about local and global node−node connections in different types of complex networks. To do this we have used MI-NODES, a new tool for the study of complex networks, which is an upgrade of the software March-Inside, classically used to study drugs and proteins. The parameters obtained can be used as inputs of LDA models to computationally assess the quality of connectivity patterns in known and new complex networks. These QSPR-like models are useful to reconstruct and/or collate in a simple and cheap way a network, as alternative to high-cost experimental re-evaluation of all links. This is a topic of the major importance because of the increasing use of existing complex networks in many areas of research.

## ■ ASSOCIATED CONTENT

### Supporting Information

Description of the training/cross validation sets. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: gonzalezdiazh@yahoo.es. Phone: +34 981 167 000.

### Notes

The authors declare no competing financial interest.

## ■ REFERENCES

(1) Moreau, G.; Broto, P. Autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360.

(2) Moreau, G.; Broto, P. Autocorrelation of molecular structures, application to SAR studies. *Nouv. J. Chim.* **1980**, *4*, 757−764.

(3) Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. *Eur. J. Med. Chem.* **1984**, *19* (1), 71−78.

(4) Saiz-Urra, L.; Gonzalez, M. P.; Teijeira, M. 2D-autocorrelation descriptors for predicting cytotoxicity of naphthoquinone ester derivatives against oral human epidermoid carcinoma. *Bioorg. Med. Chem.* **2007**, *15* (10), 3565−71.
(5) Nohair, M.; Zakarya, D. Prediction of solubility of aliphatic alcohols using the restricted components of autocorrelation method (RCAM). *J. Mol. Model. (online)* **2003**, *9* (6), 365−71.
(6) Gonzalez, M. P.; Caballero, J.; Helguera, A. M.; Garriga, M.; Gonzalez, G.; Fernandez, M. 2D autocorrelation modelling of the inhibitory activity of cytokinin-derived cyclin-dependent kinase inhibitors. *Bull. Math. Biol.* **2006**, *68* (4), 735−51.
(7) Gancia, E.; Bravi, G.; Mascagni, P.; Zaliani, A. Global 3D-QSAR methods: MS-WHIM and autocorrelation. *J. Comput.-Aided Mol. Des.* **2000**, *14* (3), 293−306.
(8) Caballero, J.; Fernandez, M.; Saavedra, M.; Gonzalez-Nilo, F. D. 2D Autocorrelation, CoMFA, and CoMSIA modeling of protein tyrosine kinases' inhibition by substituted pyrido[2,3-d]pyrimidine derivatives. *Bioorg. Med. Chem.* **2008**, *16* (2), 810−21.
(9) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: The Netherlands, 1999.
(10) Moro, S.; Bacilieri, M.; Cacciari, B.; Spalluto, G. Autocorrelation of molecular electrostatic potential surface properties combined with partial least squares analysis as new strategy for the prediction of the activity of human A(3) adenosine receptor antagonists. *J. Med. Chem.* **2005**, *48* (18), 5698−704.
(11) Fernández, L.; Caballero, J.; Abreu, J. I.; Fernández, M. Amino Acid Sequence Autocorrelation Vectors and Bayesian-Regularized Genetic Neural Networks for Modeling Protein Conformational Stability: Gene V Protein Mutants. *Proteins* **2007**, *67*, 834−852.
(12) Caballero, J.; Garriga, M.; Fernandez, M. 2D Autocorrelation modeling of the negative inotropic activity of calcium entry blockers using Bayesian-regularized genetic neural networks. *Bioorg. Med. Chem.* **2006**, *14* (10), 3330−40.
(13) Caballero, J.; Fernández, L.; Garriga, M.; Abreu, J. I.; Collina, S.; Fernández, M. Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. *J. Mol. Graphics Modell.* **2007**, *26* (1), 166−178.
(14) Caballero, J.; Fernandez, L.; Abreu, J. I.; Fernandez, M. Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants. *J. Chem. Inf. Model.* **2006**, *46* (3), 1255−68.
(15) Han, P.; Zhang, X.; Norton, R. S.; Feng, Z. P. Large-scale prediction of long disordered regions in proteins using random forests. *BMC Bioinf.* **2009**, *10*, 8.
(16) Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* **2006**, *34* (Web Server issue), W32−7.
(17) Munteanu, C. R.; Gonzalez-Diaz, H.; Borges, F.; de Magalhaes, A. L. Natural/random protein classification models based on star network topological indices. *J. Theor. Biol.* **2008**, *254* (4), 775−83.
(18) Bornholdt, S.; Schuster, H. G. *Handbook of Graphs and Complex Networks: From the Genome to the Internet*; WILEY-VCH GmbH & CO. KGa.: Weinheim, Germany, 2003.
(19) Bonchev, D. Complexity analysis of yeast proteome network. *Chem. Biodiversity* **2004**, *1* (2), 312−26.
(20) Bonchev, D.; Thomas, S.; Apte, A.; Kier, L. B. Cellular automata modelling of biomolecular networks dynamics. *SAR QSAR Environ. Res.* **2010**, *21* (1), 77−102.
(21) Managbanag, J. R.; Witten, T. M.; Bonchev, D.; Fox, L. A.; Tsuchiya, M.; Kennedy, B. K.; Kaeberlein, M. Shortest-path network analysis is a useful approach toward identifying genetic determinants of longevity. *PLoS ONE* **2008**, *3* (11), e3802.
(22) Mazurie, A.; Bonchev, D.; Schwikowski, B.; Buck, G. A. Phylogenetic distances are encoded in networks of interacting pathways. *Bioinformatics* **2008**, *24* (22), 2579−85.
(23) Thomas, S.; Bonchev, D. A survey of current software for network analysis in molecular biology. *Hum. Genomics* **2010**, *4* (5), 353−60.
(24) Estrada, E.; Kalala-Mutombo, F.; Valverde-Colmeiro, A. Epidemic spreading in networks with nonrandom long-range interactions. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2011**, *84* (3 Pt 2), 036110.
(25) Estrada, E. Generalized walks-based centrality measures for complex biological networks. *J. Theor. Biol.* **2010**, *263* (4), 556−65.
(26) Estrada, E. Universality in protein residue networks. *Biophys. J.* **2010**, *98* (5), 890−900.
(27) Estrada, E.; Bodin, O. Using network centrality measures to manage landscape connectivity. *Ecol. Appl.* **2008**, *18* (7), 1810−25.
(28) González-Díaz, H.; Prado-Prado, F.; García-Mera, X. *Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences*; Transworld Research Network: Kerala, India, 2011; pp 001−142.
(29) Duardo-Sanchez, A.; Patlewicz, G.; González-Díaz, H. A Review of Network Topological Indices from Chem-Bioinformatics to Legal Sciences and back. *Cur. Bioinf.* **2011**, *6* (11), 53−70.
(30) Gonzalez-Diaz, H.; Romaris, F.; Duardo-Sanchez, A.; Perez-Montoto, L. G.; Prado-Prado, F.; Patlewicz, G.; Ubeira, F. M. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues. *Curr. Pharm. Des.* **2010**, *16* (24), 2737−64.
(31) Gonzalez-Diaz, H.; Duardo-Sanchez, A.; Ubeira, F. M.; Prado-Prado, F.; Perez-Montoto, L. G.; Concu, R.; Podda, G.; Shen, B. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. *Curr. Drug Metab.* **2010**, *11* (4), 379−406.
(32) Gonzalez-Diaz, H. QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neurosciences. *Curr. Pharm. Des.* **2010**, *16* (24), 2598−600.
(33) Gonzalez-Diaz, H. Network topological indices, drug metabolism, and distribution. *Curr. Drug Metab.* **2010**, *11* (4), 283−4.
(34) González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics* **2008**, *8*, 750−778.
(35) Modha, D. S.; Singh, R. Network architecture of the long-distance pathways in the macaque brain. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (30), 13485−90.
(36) Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* **2000**, *407* (6804), 651−4.
(37) Riera-Fernández, P.; Munteanu, C. R.; Dorado, J.; Martín-Romalde, R.; Duardo-Sanchez, A.; González-Díaz, H., From Chemical Graphs in Computer Aided Drug Design to General Markov-Galvez Indices of Drug-Target, Proteome, Drug-Parasitic Disease, Technological, and Social-Legal Networks. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, (?), ??
(38) Riera-Fernández, P.; Munteanu, C. R.; Pedreira-Souto, N.; Martín-Romalde, R.; Duardo-Sanchez, A.; González-Díaz, H. Definition of Markov-Harary Invariants and Review of Classic Topological Indices and Databases in Biology, Parasitology, Technology, and Social-Legal Networks. *Curr. Bioinf.* **2011**, *6* (1), 94−121.
(39) Mezo, M.; Gonzalez-Warleta, M.; Castro-Hermida, J. A.; Ubeira, F. M. Evaluation of the flukicide treatment policy for dairy cattle in Galicia (NW Spain). *Vet. Parasitol.* **2008**, *157* (3−4), 235−43.
(40) González-Díaz, H.; Mezo, M.; González-Warleta, M.; Muíño-Pose, L.; Paniagua, E.; Ubeira, F. M., Network prediction of fasciolosis spreading in Galicia (NW Spain). In *Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks*, González-Díaz, H., Munteanu, C. R., Eds; Transworld Research Network: Kerala (India), 2010; pp 191−204.
(41) Estrada, E. Information mobility in complex networks. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2009**, *80* (2 Pt 2), 026104.
(42) Riera-Fernandez, I.; Martin-Romalde, R.; Prado-Prado, F. J.; Escobar, M.; Munteanu, C. R.; Concu, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. From QSAR models of Drugs to Complex

Networks: State-of-Art Review and Introduction of New Markov-Spectral Moments Indices. *Curr. Top. Med. Chem.* **2012**.
(43) Riera-Fernandez, P.; Munteanu, C. R.; Escobar, M.; Prado-Prado, F.; Martin-Romalde, R.; Pereira, D.; Villalba, K.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks. *J. Theor. Biol.* **2012**, *293*, 174−88.
(44) Riera-Fernandez, P.; Munteanu, C. R.; Pedreira-Souto, N.; Martin-Romalde, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. Definition of Markov-Harary Invariants and Review of Classic Topological Indices and Databases in Biology, Parasitology, Technology, and Social-Legal Networks. *Curr. Bioinf.* **2011**, *6* (1), 94−121.
(45) Hill, T.; Lewicki, P. *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*; StatSoft: Tulsa, OK, 2006; Vol. *1*, p 813.
(46) Rosa da Silva, M.; Sun, J.; Ma, H. W.; He, F.; Zeng, A. P., Metabolic networks. In *Analysis of biological networks*, Junker, B. H.; Schreiber, F., Eds. Wiley & Sons: Hoboken, NJ, 2008; pp 233−253.
(47) Baldazzi, V.; Ropers, D.; Markowicz, Y.; Kahn, D.; Geiselmann, J.; de Jong, H. The carbon assimilation network in Escherichia coli is densely connected and largely sign-determined by directions of metabolic fluxes. *PLoS Comput. Biol.* **2010**, *6* (6), e1000812.
(48) Costa, R. S.; Machado, D.; Rocha, I.; Ferreira, E. C. Hybrid dynamic modeling of Escherichia coli central metabolic network combining Michaelis-Menten and approximate kinetic equations. *Biosystems* **2010**, *100* (2), 150−7.
(49) Gerlee, P.; Lizana, L.; Sneppen, K. Pathway identification by network pruning in the metabolic network of Escherichia coli. *Bioinformatics* **2009**, *25* (24), 3282−8.
(50) Fowler, Z. L.; Gikandi, W. W.; Koffas, M. A. Increased malonyl coenzyme A biosynthesis by tuning the Escherichia coli metabolic network and its application to flavanone production. *Appl. Environ. Microbiol.* **2009**, *75* (18), 5831−9.
(51) Konig, R.; Schramm, G.; Oswald, M.; Seitz, H.; Sager, S.; Zapatka, M.; Reinelt, G.; Eils, R. Discovering functional gene expression patterns in the metabolic network of Escherichia coli with wavelets transforms. *BMC Bioinf.* **2006**, *7*, 119.
(52) Imielinski, M.; Belta, C.; Rubin, H.; Halasz, A. Systematic analysis of conservation relations in Escherichia coli genome-scale metabolic network reveals novel growth media. *Biophys. J.* **2006**, *90* (8), 2659−72.
(53) Shi, H.; Nikawa, J.; Shimizu, K. Effect of modifying metabolic network on poly-3-hydroxybutyrate biosynthesis in recombinant Escherichia coli. *J. Biosci. Bioeng.* **1999**, *87* (5), 666−77.
(54) Lin, H.; Bennett, G. N.; San, K. Y. Chemostat culture characterization of Escherichia coli mutant strains metabolically engineered for aerobic succinate production: a study of the modified metabolic network based on metabolite profile, enzyme activity, and gene expression profile. *Metab. Eng.* **2005**, *7* (5−6), 337−52.
(55) Ghim, C. M.; Goh, K. I.; Kahng, B. Lethality and synthetic lethality in the genome-wide metabolic network of Escherichia coli. *J. Theor. Biol.* **2005**, *237* (4), 401−11.
(56) Schmid, J. W.; Mauch, K.; Reuss, M.; Gilles, E. D.; Kremling, A. Metabolic design based on a coupled gene expression-metabolic network model of tryptophan production in Escherichia coli. *Metab. Eng.* **2004**, *6* (4), 364−77.
(57) Light, S.; Kraulis, P. Network analysis of metabolic enzyme evolution in Escherichia coli. *BMC Bioinf.* **2004**, *5*, 15.
(58) Burgard, A. P.; Maranas, C. D. Probing the performance limits of the Escherichia coli metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.* **2001**, *74* (5), 364−75.
(59) Edwards, J. S.; Palsson, B. O. Robustness analysis of the Escherichia coli metabolic network. *Biotechnol. Prog.* **2000**, *16* (6), 927−39.
(60) Goffeau, A. The yeast genome directory. *Nature* **1997**, *387* (6632 Suppl), 5.
(61) Burglin, T. R.; Lobos, E.; Blaxter, M. L. Caenorhabditis elegans as a model for parasitic nematodes. *Int. J. Parasitol.* **1998**, *28* (3), 395−411.
(62) Consortium, T. C. e. S., Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **1998**, *282*, (5396), 2012-8.
(63) Bird, D. M.; Opperman, C. H. Caenorhabditis elegans: A Genetic Guide to Parasitic Nematode Biology. *J. Nematol.* **1998**, *30* (3), 299−308.
(64) Holden-Dye, L.; Walker, R. J. Anthelmintic drugs. *WormBook*, 2007; pp 1−13.
(65) Muller, B.; Grossniklaus, U. Model organisms-A historical perspective. *J. Proteomics* **2010**, *73* (11), 2054−63.
(66) Desdevises, Y.; Morand, S.; Legendre, P. Evolution and determinants of host specificity in the genus Lamellodiscus (Monogenea). *Biol. J. Linn. Soc.* **2002**, *77*, 431−443.
(67) Detwiler, J.; Janovy, J., Jr. The role of phylogeny and ecology in experimental host specificity: Insights from a eugregarine-host system. *J. Parasitol.* **2008**, *94* (1), 7−12.
(68) Poulin, R.; Krasnov, B. R.; Mouillot, D. Host specificity in phylogenetic and geographic space. *Trends Parasitol.* **2011**, In press.
(69) Hatcher, J. M.; Dick, J. T. A.; Dunn, A. M. How parasites affect interactions between competitors and predators. *Ecol. Lett.* **2006**, *9* (11), 1253−1271.
(70) Price, P. W.; Westoby, M.; Rice, B.; Atsatt, P. R.; Fritz, R. S.; Thompson, J. N.; Mobley, K. Parasite Mediation in Ecological Interactions. *Annu. Rev. Ecol. Syst.* **1986**, *17*, 485−505.
(71) Anderson, R. M.; May, R. M. Population biology of infectious diseases: Part I. *Nature* **1979**, *280*, 361−367.
(72) Poulin, R. Network analysis shining light on parasite ecology and diversity. *Trends Parasitol.* **2010**, *26* (10), 492−8.
(73) Dobson, A. Population dynamics of pathogens with multiple host species. *Am. Nat.* **2004**, *164* (Suppl5), S64−78.
(74) Roche, B.; Guegan, J. F. Ecosystem dynamics, biological diversity and emerging infectious diseases. *C. R. Biol.* **2011**, *334* (5−6), 385−92.
(75) Kotter, R. Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database. *Neuroinformatics* **2004**, *2* (2), 127−44.
(76) Stephan, K. E.; Kamper, L.; Bozkurt, A.; Burns, G. A.; Young, M. P.; Kotter, R. Advanced database methodology for the Collation of Connectivity data on the Macaque brain (CoCoMac). *Philos. Trans. R. Soc., B* **2001**, *356* (1412), 1159−86.
(77) Mas-Coma, S. Epidemiology of fascioliasis in human endemic areas. *J. Helminthol.* **2005**, *79* (3), 207−16.