Journal of Theoretical Biology 343 (2014) 16–24

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A QSPR-like model for multilocus genotype networks of *Fasciola hepatica* in Northwest Spain

Severo Vázquez-Prieto [a,*], Humberto González-Díaz [b,c], Esperanza Paniagua [a], Román Vilas [d], Florencio M. Ubeira [a]

[a] Laboratory of Parasitology, Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, Campus Vida, Santiago de Compostela 15782, Spain
[b] Department of Organic Chemistry II, University of the Basque Country UPV/EHU, Leioa 48940, Spain
[c] IKERBASQUE, Basque Foundation for Science, Bilbao 48011, Spain.
[d] Department of Genetics, Faculty of Biology, University of Santiago de Compostela, Campus Vida, Santiago de Compostela 15782, Spain

HIGHLIGHTS

- Two complex networks of *Fasciola hepatica* are built based on codominant genetic markers.
- A quantitative structure–property relationship (QSPR) like model is developed.
- This type of model could be used to manage and prevent the spread of fasciolosis.

ARTICLE INFO

Article history:
Received 28 May 2013
Received in revised form
8 November 2013
Accepted 11 November 2013
Available online 20 November 2013

Keywords:
Graph theory
Complex networks
Fasciolosis
Codominant genetic markers
Linear discriminant analysis

ABSTRACT

Fasciola hepatica is a parasitic trematode that infects wild and domesticated mammals, particularly cattle and sheep, and causes significant economic losses to global livestock production. In the present study, we used codominant genetic markers to define and build, for the first time, complex genotype networks for *F. hepatica* isolated from cattle and sheep in NW Spain. We generated three types of random networks with a number of nodes and edges as close as possible to the observed networks, and we then calculated 14 node centrality measures for both observed and random networks. Finally, using Linear Discriminant Analysis (LDA) and these measures as inputs, we constructed a quantitative structure–property relationship (QSPR)-like model able to predict the propensity of a specific genotype of *F. hepatica* to infect different infrapopulations, farms and/or host species. The accuracy, sensitivity and specificity of the model were $>90\%$ for both training and cross-validation series. We also assessed the applicability domain of the model. This type of QSPR model is a potentially powerful tool for epidemiological studies and could be used to manage and prevent the spread of fasciolosis.

1. Introduction

Network theory is useful for representing and analyzing complex systems of interacting agents (Porter et al., 2005). Consequently, the approach is increasingly being used in areas as diverse as linguistics, technology and sociology (Bornholdt and Schuster, 2003; Boccaletti et al., 2006; Dehmer and Emmert-Streib, 2009). Likewise, network analysis methods are widely used today to study various important biological and biomedical problems (see, e.g., (Aguero-Chapín et al., 2008; Chen et al., 2010a, 2010b; Cruz-Monteagudo et al., 2008; He et al., 2010; Hu et al., 2011, 2012, 2011; Huang et al., 2010, 2012a, 2012b; Jiang et al., 2013; Li et al., 2012; Prado-Prado et al., 2009)). In the context of this theory, a network can be defined as a set of items (vertices or nodes) with connections between them (links or edges) (Newman, 2003). Each of the nodes represents one part of a complex system and the edges symbolize structural or functional relationships between these parts. One important aspect of networks is that the connectivity patterns can be described by numerical parameters. In particular, we can calculate Topological Indices (TIs), Connectivity Indices (CIs) and node centralities $C_t(j)$ of type t from the graph/network representation of a system to describe full complex network topology, node connectivity and sub-graph branching, respectively (Sabidussi, 1966; Freeman, 1977; Todeschini and

* Corresponding author. Tel.: +34 881 815004; fax: +34 981 593316.
E-mail addresses: severo.vazquez@usc.es (S. Vázquez-Prieto), humberto.gonzalezdiaz@ehu.es (H. González-Díaz), mesperanza.paniagua@usc.es (E. Paniagua), roman.vilas@usc.es (R. Vilas), fm.ubeira@usc.es (F.M. Ubeira).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.11.005

Consonni, 2002). The latter descriptors, i.e. the node centrality measures, allow the importance of a node in a graph to be quantified on the basis of its position and connections. Although each node centrality has its own symbol, for the sake of simplicity, we use here the notation $C_t(j)$, where C represents centrality, jth represents a node, and t represents the type of centrality.

Transformation of graphs into numbers allows storage, manipulation, comparison and retrieval of information, but also facilitates the search for quantitative structure–property relationship (QSPR) models (Katritzky and Gordeeva, 1993; Estrada and Molina, 2001; Yang and Zhong 2003; Liu et al., 2007; Xu et al., 2007). This type of model enables prediction of the properties of complex systems on the basis of numerical parameters that describe the structure of the systems (Puzyn et al., 2010). Although QSPR models have traditionally been restricted to the study of small molecules, in recent years many QSPR-like models have been used to study macromolecules, such as DNA and proteins, and also brain cortex, population sociology and other complex systems (Altmann, 1993; Gozalbes et al., 2002; Hahn and Kern, 2005; Bonchev and Buck, 2007; Honey et al., 2007; McDonald, 2007; González-Díaz and Munteanu, 2010). However, current use of TIs, CIs and $C_t(j)$ measures in QSPR research by no means covers all the potential applications of the indices, and new openings for biological QSPR-like models are waiting to be explored (Roy et al., 2006; Roy and Ghosh, 2010; Marrero-Ponce et al., 2012; Speck-Planche et al., 2012a; Barigye et al., 2013).

Here, we propose the application of this type of approach to the study of *Fasciola hepatica* (the common liver fluke), a parasitic trematode that infects wild and domesticated mammals, particularly cattle and sheep. Fasciolosis is currently the most widespread vector-borne disease (in terms of latitude, longitude and altitude) known (Mas-Coma, 2005), and it causes significant economic losses to livestock production worldwide. Effective strategies for control of the disease are mainly based on the use of anthelmintic drugs; however, resistance to the drug of choice, triclabendazole, has been reported in several countries (Overend and Bowen, 1995; Mitchell et al., 1998; Moll et al., 2000; Mooney et al., 2009; Olaechea et al., 2011). Fasciolosis is endemic in Galicia (NW Spain), where the flukicide treatment policy probably contributes to the development of drug resistance because triclabendazole is often administered without prior diagnosis (Mezo et al., 2008). Vilas et al. (2012) have recently suggested that *F. hepatica* infecting cattle and sheep have different patterns of population genetic structure in the study area, and that these different patterns may increase the risk of development and rapid spread of drug resistance. In this sense, a complex network analysis of the information generated by the genetic characterization of *F. hepatica* populations from the study area is of great interest. In the present study, we used codominant genetic markers to build, for the first time, two complex networks for *F. hepatica* isolated from cattle and sheep in NW Spain. We also calculated several $C_t(j)$ measures for all the nodes of these networks. Finally, we developed a QSPR-like model that may predict the tendency or propensity of a specific genotype of *F. hepatica* to infect different infrapopulations, farms and/or host species. We also assessed the applicability domain of the model.

2. Materials and methods

2.1. Data set used

Vilas et al. (2012) used 12 polymorphic genetic markers to characterize 20 infrapopulations (i.e. all parasites in an individual host) of *F. hepatica* isolated from 10 cows and 10 sheep in NW Spain. The studied farms were located at Abegondo (ABG), Bóveda (BOV), Chantada (CHAN), Colunga (COL), Mesía (MES), Cervantes (CER), Melide (MEL), Cerceda (CCD), Guntín (GUN) and Mabegondo (MBO) (Fig. 1). All the sheep were from the same farm (MBO) and were designated MBO1-MBO10. Each of the individual cattle belonged to a farm from a different location. Only parasites from two cattle hosts were collected from two farms from the same locality (BOV1 and BOV2). Thus, samples from 11 farms were analyzed, 10 of which were cattle farms. The following enzymes were analyzed electrophoretically: aconitase (ACO, EC 4.2.1.3), adenylate kinase (AK, EC 2.7.4.3), glutamate oxalacetate transaminase (GOT, EC 2.6.1.1), hexokinase (HK, EC 2.7.1.1), isocitrate dehydrogenase (IDH, 1.1.1.42), phosphogluconate dehydrogenase (PGD, EC 1.1.1.44) and phosphoglucomutase (PGM, EC 2.7.5.1). These enzymes were presumably encoded by eight polymorphic loci in *F. hepatica* populations from the study area (Vázquez-Prieto et al., 2011). The four microsatellite markers examined, previously reported by Hurtrez-Boussès et al. (2004), were FH15, FH23, FH25 and FH222CBP. The authors used the variation in these 12 codominant genetic markers, eight allozymes and four microsatellite loci, to obtain multilocus genotypes of each worm.

2.2. Network construction and centrality calculation

In the present study, we used the above data set to construct two multilocus genotype networks of *F. hepatica*—one for the parasite isolated from cattle and the other for the parasite isolated from sheep. As the parasite is a diploid organism, each locus was

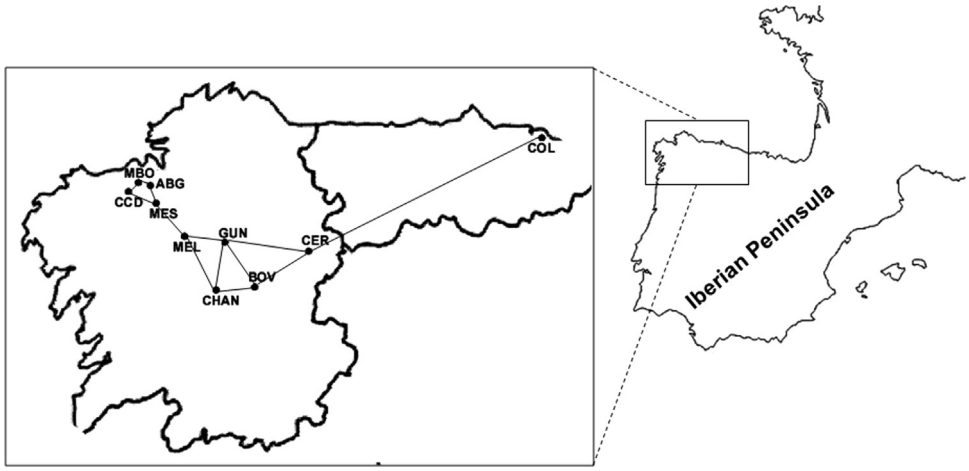

Fig. 1. Map showing the 10 Spanish locations where sampling was conducted: Abegondo (ABG), Bóveda (BOV), Chantada (CHAN), Mesía (MES), Cervantes (CER), Melide (MEL), Cerceda (CCD), Guntín (GUN) and Mabegondo (MBO), all of which are in Galicia, and Colunga (COL), which is in Asturias.