Mol Divers
DOI 10.1007/s11030-017-9749-4

ORIGINAL ARTICLE

A study of the Immune Epitope Database for some fungi species using network topological indices

Severo Vázquez-Prieto[1,2] · Esperanza Paniagua[3] · Hugo Solana[1] · Florencio M. Ubeira[3] · Humberto González-Díaz[4,5]

Received: 19 December 2016 / Accepted: 9 May 2017

Abstract In the last years, the encryption of system structure information with different network topological indices has been a very active field of research. In the present study, we assembled for the first time a complex network using data obtained from the Immune Epitope Database for fungi species, and we then considered the general topology, the node degree distribution, and the local structure of this network. We also calculated eight node centrality measures for the observed network and compared it with three theoretical models. In view of the results obtained, we may expect that the present approach can become a valuable tool to explore the complexity of this database, as well as for the storage, manipulation, comparison, and retrieval of information contained therein.

Keywords Network theory · Topological indices · B-cells epitopes

Introduction

Fungi are ubiquitous in the environment. There are approximate 1.5 million different species, but only about 300 are known to cause disease in humans. Infections caused by these organisms have increased dramatically during the past decades, mainly concomitant with other diseases (e.g., AIDS, diabetes) or caused by treatment with chemotherapeutics, corticosteroids, or tumor necrosis factor inhibitors [1].

On the other hand, as expressed by González-Díaz et al. [2], the number of systems that can be represented and studied with network theory in nature is so vast that many authors regard this theory as a science [3]. A network is a real system in which the vertices correspond to parts of the real entity that is intended to represent and the edges to the relationships of different nature that are established between them [4]. One can numerically describe a network by what are known as topological indices (TIs) [5–8]. These parameters have the advantage of having a straightforward theoretical base, which can be understood by scientists non-expert on computational techniques, and being not time-consuming in terms of computational resources [9]. Consequently, over time, the use of this type of indices has been extended to the encoding of information contained in complex networks of very diverse fields [10–15].

The body of the Immune Epitope Database (IEDB; www.iedb.org) has considerably increased in the last few years, providing a wealth of data potentially useful for basic and clinical applications [16]. Given the complexity of this database, a complex network approach may be applied to analyze the huge amount of information contained

Electronic supplementary material The online version of this article (doi:10.1007/s11030-017-9749-4) contains supplementary material, which is available to authorized users.

✉ Severo Vázquez-Prieto
severovazquezprieto@gmail.com

[1] Laboratorio de Biología Celular y Molecular, Centro de Investigación Veterinaria de Tandil (CIVETAN), CONICET, Facultad de Ciencias Veterinarias, Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), Campus Universitario, 7000 Tandil, Argentina

[2] Instituto de Ciencias Biomédicas, Facultad de Ciencias de la Salud, Universidad Autónoma de Chile, Temuco, Chile

[3] Laboratorio de Parasitología, Departamento de Microbiología y Parasitología, Facultad de Farmacia, Universidad de Santiago de Compostela, Campus Vida, 15782 Santiago de Compostela, Spain

[4] Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940 Leioa, Spain

[5] IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain

Published online: 31 May 2017

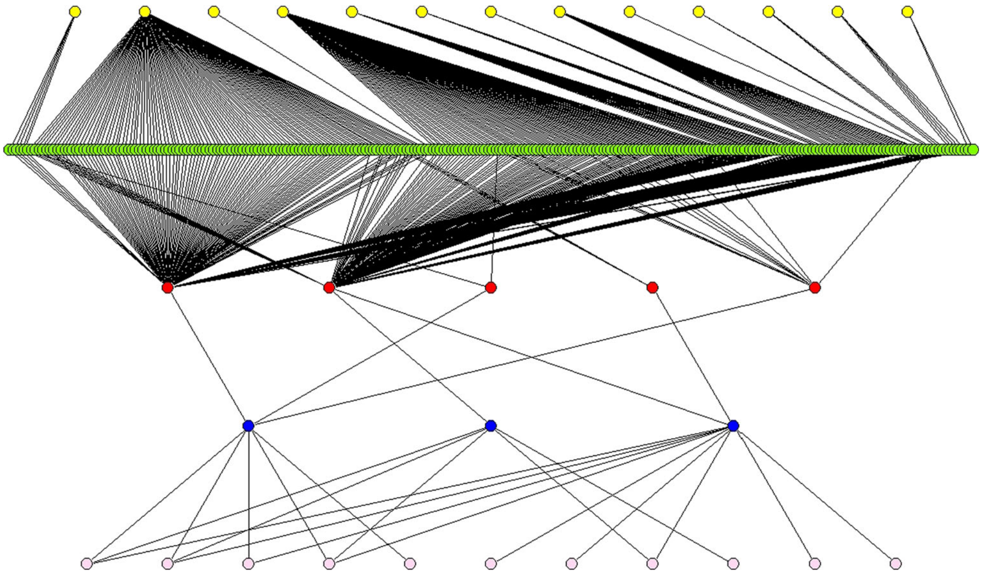


Fig. 1 Observed network for data obtained from the Immune Epitope Database for fungi species. Source organisms (*yellow*), peptide sequences (*green*), immunological processes (*red*), host organisms (*blue*), and experimental techniques (*pink*).The network is 5th-partite in such a way that nodes of a given class are connected with nodes of other classes, but nodes of the same class are never connected to each other. Thus, the information encoded by a link depends on the classes of the two nodes interconnected. For example, if one node belongs to the class host organism and the other to the class experimental technique, this link indicates the technique by which *i*th molecule was determined to be an immune epitope for the corresponding host organism. (Color figure online)

therein. Moreover, understanding the topology of a network may give direct insight into various network characteristics [17–20].

In the present study, we proposed the application of a topological network approach for the analysis of information obtained from the IEDB on fungi species. For it, we defined and built for the first time a complex network based on data obtained from the IEDB for some of such organisms, and we then considered the general topology, the node degree distribution and the local structure of this network. We also calculated eight node centrality measurements for the observed network and compared it with three theoretical models.

Results and discussion

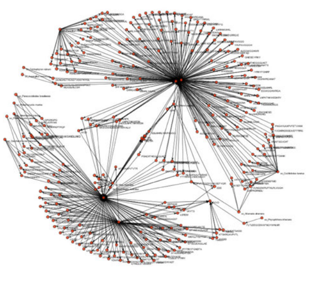
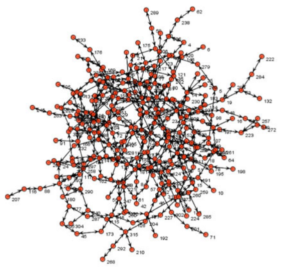
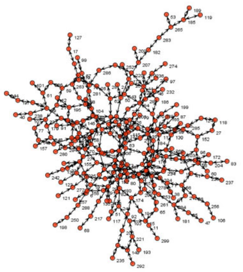
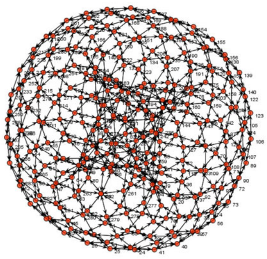
In the present study, we assembled for the first time a complex network using data obtained from the IEDB on fungi species. The observed network contained a total of 292 nodes (13 source organisms, 260 peptide sequences, 5 immunological processes, 3 host organisms, and 11 experimental techniques) interconnected by 559 directed edges (Fig. 1).

The names and codes for all nodes are listed in Table S1 of the Supplementary material. We provide the top 10 values for each of the node centrality measures calculated in this work in Table S2 of the Supplementary material. The node degree distributions, based on the Kolmogorov–Smirnov and the Chi-Square tests, did not fit any previously studied distribution (Table S3, Supplementary material). The three types of generated random networks showed some remarkable properties. For example, the Erdös–Rényi network (ERN) and Eppstein Power Law network (EPLN) presented a similar average degree (*Ad*), although relatively lower than that showed by 2D-lattice network (2D-L). The graphical representation and numerical comparison of the observed and random networks are shown in Table 1.

An interesting review on the applications of networks' TIs for the study of small molecules, macromolecules, and other networks may be read in González-Díaz et al. [7]. The general topologies of the corresponding ideal theoretical networks with respect to that of the real network in terms of the relative difference percentage (RD%) are given in Table 2.

The lower differences for different features were 2D-L and ERN for number of nodes (n) (RD% 1.0 and 1.4%, respectively), and EPLN for number of edges (m) (RD% = −4.5%). The absolute values means of RD%s for EPLN, ERN, and 2D-L were 56.4, 60.2, and 81.4%, respectively. Therefore, the observed network does not match with any of ideal behaviors studied here. We show the triadic census analysis results (local structure) for the complex network constructed in this work in Table 3. The most triads were null triads (type 1-003), which coincide with the behavior of complex social networks where this type of triads accounts for more than 50% of the total [21]. The 5-021U and 4-021D triads had values higher than the expected ones; however,

Table 1 Comparison of observed versus random networks

Observed network	Value	TIs[a]	Value	Erdös–Rényi network
	292	*n*	288	
	559	*m*	744	
	0.007	*d*	0.009	
	3.829	*Ad*	5.167	
	55568	*M1*	9680	
	115228	*M2*	30912	
	55.171	*Xr*	134.036	
	54450	*F*	8192	
Eppstein Power Law network	**Value**	**TIs**	**Value**	**2D-Lattice network**
	247	*n*	289	
	584	*m*	1156	
	0.01	*d*	0.014	
	4.729	*Ad*	8	
	7104	*M1*	18496	
	21072	*M2*	73984	
	114.22	*Xr*	144.5	
	5936	*F*	16184	

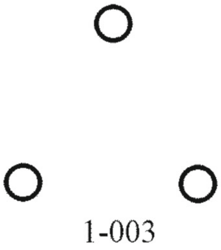
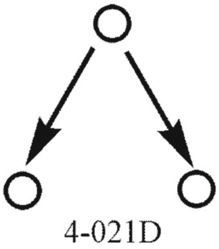
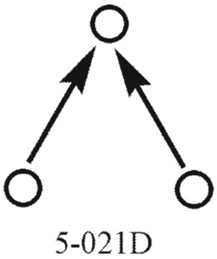
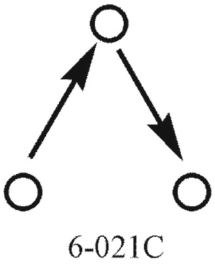
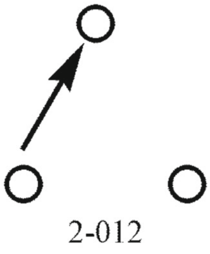
[a] The TIs used are number of nodes (*n*), the total adjacency index or the number of edges (*m*), the density (*d*), the average degree (*Ad*), the Zagreb group index 1 (*M1*), the Zagreb group index 2 (*M2*), the Randic connectivity index (*Xr*), and the Platt index (*F*)

Table 2 Summary of the comparative study of the observed versus random networks

Parameters[a]	ERN	EPLN	2D-L
n	1.4	15.4	1.0
m	−33.1	−4.5	−106.8
d	−28.6	−42.9	−100.0
Ad	−34.9	−23.5	−108.9
$M1$	82.6	87.2	66.7
$M2$	73.2	81.7	35.8
Xr	−142.9	−107.0	−161.9
F	85.0	89.1	70.3

[a] The TIs used are: number of nodes (*n*), the total adjacency index or the number of edges (*m*), the density (*d*), the average degree (*Ad*), the Zagreb group index 1 (*M1*), the Zagreb group index 2 (*M2*), the Randic connectivity index (*Xr*), and the Platt index (*F*). The results are expressed as relative difference percentage, defined as RD% = (TIreal − TIideal) * 100/TIreal. ERN, EPLN, and 2D-L means Erdös–Rényi network, Eppstein Power Law network, and 2D-lattice network, respectively

Table 3 Triadic census analysis of the real network

Triad representation	Triad type	Number of triads (ni)	Expected (ei)	(ni—ei)/ei	Triad details and examples [a]
	1 - 003	3972095	3947512.89	0.00	Null triad, e.g., isolated nodes of one source organism, one epitope, and one immunological process totally unrelated
	4 - 021D	9618	519.34	17.52	Divergent triad, e.g., one peptide that is epitope in two different immunological processes
	5 - 021U	16910	519.34	31.56	Convergent triad, e.g., two peptides that are epitopes in the same immunological process
	6 - 021C	697	1038.68	–0.33	Transitivity triad; e.g., one peptide isolate from one source organism that is epitope in one specific immunological process
	2 - 012	107660	156847.38	–0.31	Not transitive triad, e.g., the peptide isolated from the source organism is not epitope for this specific immunological process

[a] Please, take into consideration the existence of only the following hierarchy relationships (and not others) inside the network source organisms => peptide sequences => immunological processes => host organisms => experimental techniques (see Fig. 1) In consequence, all other triads do not appear due to the asymmetrical nature of the network

the 6-021C and 2-012 triads presented a number lower than the expected value (see Table 3 for triad details and examples).

Using graphical approaches, one can obtain easily an intuitive image that provides useful information about the complex biological system under study. Consequently, different approaches of this type have been fruitfully applied to a broad spectrum of biological topics, such as enzyme-catalyzed systems [22–25], protein folding kinetics [26], analysis of codon usage [27,28], HIV-1 reverse transcriptase inhibition mechanisms [29,30], base frequencies in the anti-sense strands [31], analysis of DNA sequence [32], and the parasite *Fasciola hepatica* [33]. In the field of mycology, a network approach can be used to identify drugs with similar mechanism of action [34] or predict the antifungal activity of drugs against different species [35,36]. On the other hand, the study of protein–protein interaction networks in this type of organisms may be a key tool in understanding the basic principles that govern their biological processes [37–40].

Conclusions

We have demonstrated that TIs are promising indices of general use at different structural organization levels. In particular, we have confirmed that it is possible to use these parameters in the study of data obtained from IEDB for fungi species. Consequently, we may expect that the present approach can become a valuable tool to explore the complexity of this database, as well as for the storage, manipulation, comparison, and retrieval of information contained therein. On the other hand, this relatively simple approach may be very useful for mycological research. For example, it could provide a fast way to carry out a preliminary evaluation that allow us to make a decision on the best experimental conditions to determine whether the molecules being studied are immune epitopes or not, by detecting groups of nodes highly connected among each other. Thus, this methodology could help us to somewhat rationalize this process and reducing costs in terms of material resources and time. In addition, the present study could be the seed for further development of new software programs, webs-serves, and/or theoretical methods for handling structure-function information and data mining in this field in the near future.

Experimental

Using Microsoft Excel, we constructed a network from data obtained from the database utilized by González-Díaz et al. [41] and Vázquez-Prieto et al. [42]. Once the network is constructed, the *.mat file was uploaded in CentiBin [43] and the program tools were used to prepare the network for the calculation of eight node centralities directed: in-Degree, out-Degree, Eigenvector, Hubbel Index, Bargaining, PageRank, HITS-Authority, and HITS-Hubs. We then generated random networks by using three different algorithms: the Erdös–Rényi network (ERN), the Eppstein Power Law network (EPLN), and the 2D-Lattice network (2D-L). The networks were generated with a number of nodes and edges as close as possible to the observed network. Pajek software [44] was used to calculate several global measures of network structure for both observed and random networks, including number of nodes (n) and edges (m), density (d), average degree (Ad), Zagreb group index 1 ($M1$), Zagreb group index 2 ($M2$), Randic connectivity index (Xr), and Platt index (F). We used these parameters to compare the topology of the observed network with random networks in terms of the relative difference percentage, defined as RD% = (TIreal-TIideal) * 100 / TIreal. We also performed a triadic census analysis using Pajek software. Finally, all node in- and out-degrees were used as input in STATISTICA 6.0 software package [45] in order to study the distribution of the observed network and compare it with other ideal network distributions, including Normal, Exponential, Poisson and Chi-Square.

Acknowledgements This work was supported by Grants GPC2014/058 from the Xunta de Galicia and AGL2011-30563-C03 from the Ministerio de Ciencia e Innovación, Spain. S.V.P. is supported by a postdoctoral fellowship from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

References

1. Ali T, Kaitha S, Mahmood S, Ftesi A, Stone J, Bronze MS (2013) Clinical use of anti-TNF therapy and increased risk of infections. Drug Healthc Patient Saf 5:79–99. doi:10.2147/DHPS.S28801
2. González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. Proteomics 8:750–778. doi:10.1002/pmic.200700638
3. Menke R (2004) Linked: the new science of networks (review). Perspect Biol Med 47:300–303. doi:10.1353/pbm.2004.0030
4. Newman M (2003) The structure and function of complex networks. SIAM Rev 45:167–256. doi:10.1137/S003614450342480
5. Balaban AT, Beteringhe A, Constantinescu T, Filip PA, Ivanciuc O (2007) Four new topological indices based on the molecular path code. J Chem Inf Model 47:716–731. doi:10.1021/ci6005068
6. Estrada E (2001) Generalization of topological índices. Chem Phys Lett 336:248–252. doi:10.1016/S0009-2614(01)00127-0
7. González-Díaz H, Vilar S, Santana L, Uriarte E (2007) Medicinal chemistry and bioinformatics - current trends in drugs discovery with networks topological indices. Curr Top Med Chem 7:1025–1039. doi:10.2174/156802607780906771
8. Ivanciuc O, Ivanciuc T, Klein DJ, Seitz WA, Balaban AT (2001) Wiener index extension by counting even/odd graph distances. J Chem Inf Comput Sci 41:536–549. doi:10.1021/ci000086f
9. González-Díaz H, Romaris F, Duardo-Sanchez A, Pérez-Montoto LG, Prado-Prado F, Patlewicz G, Ubeira FM (2010) Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applica-

tions, and legal issues. Curr Pharm Des 16:2737–2764. doi:10.2174/138161210792389234

10. Agüero-Chapin G, González-Díaz H, Molina R, Varona-Santos J, Uriarte E, González-Díaz Y (2006) Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. FEBS Lett 580:723–730. doi:10.1016/j.febslet.2005.12.072
11. Bielinska-Waz D, Nowak W, Waz P, Nandy A, Clark T (2007) Distribution moments of 2D-graphs as descriptors of DNA sequences. Chem Phys Lett 443:408–413. doi:10.1016/j.cplett.2007.06.088
12. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19. doi:10.1093/bioinformatics/bth466
13. González-Díaz H, Cruz-Monteagudo M, Molina R, Tenorio E, Uriarte E (2005) Predicting multiple drugs side effects with a general drug-target interaction thermodynamic Markov model. Bioorg Med Chem 13:1119–1129. doi:10.1016/j.bmc.2004.11.030
14. Liao B, Ding K (2005) Graphical approach to analyzing DNA sequences. J Comput Chem 26:1519–1523. doi:10.1002/jcc.20287
15. Mandado M, Gonzáles-Moa MJ, Mosquera RA (2007) QTAIM N-center delocalization indices as descriptors of aromaticity in mono and poly heterocycles. J Comput Chem 28:1625–1633. doi:10.1002/jcc.20647
16. Vaughan K, Peters B, Larche M, Pomes A, Broide D, Sette A (2013) Strategies to query and display allergy-derived epitope data from the immune epitope database. Int Arch Allergy Immunol 160:334–345. doi:10.1159/000343880
17. Flórez AF, Park D, Bhak J, Kim BC, Kuchinsky A, Morris JH, Espinosa J, Muskus C (2010) Protein network prediction and topological analysis in *Leishmania major* as a tool for drug target selection. BMC Bioinform 11:484. doi:10.1186/1471-2105-11-484
18. Han HW, Ohn JH, Moon J, Kim JH (2013) Yin and Yang of disease genes and death genes between reciprocally scale-free biological networks. Nucleic Acids Res 41:9209–9217. doi:10.1093/nar/gkt683
19. Kotlyar M, Fortney K, Jurisica I (2012) Network-based characterization of drug regulated genes, drug targets, and toxicity. Methods 57:499–507. doi:10.1016/j.ymeth.2012.06.003
20. Yang L, Wang J, Wang H, Lv Y, Zuo Y, Li X, Jiang W (2014) Analysis and identification of essential genes in humans using topological properties and biological information. Gene 551:138–151. doi:10.1016/j.gene.2014.08.046
21. Moody J (1998) Matrix methods for calculating the triad census. Soc Netw 20:291–299. doi:10.1016/S0378-8733(98)00006-9
22. Andraos J (2008) Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Can J Chem 86:342–357. doi:10.1139/v08-020
23. Chou KC (1989) Graphical rules in steady and non-steady enzyme kinetics. J Biol Chem 264:12074–12079 PMID 2745429
24. Kuzmic P, Ng KY, Heath TD (1992) Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. Anal Biochem 200:68–73. doi:10.1016/0003-2697(92)90278-F
25. Lin SX, Neet KE (1990) Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy. J Biol Chem 265:9670–9675
26. Chou KC (1990) Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys Chem 35:1–24. doi:10.1016/0301-4622(90)80056-D
27. Chou KC, Zhang CT (1992) Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res Hum Retrovir 8:1967–1976. doi:10.1089/aid.1992.8.1967
28. Zhang CT, Chou KC (1994) A graphic approach to analyzing codon usage in 1562 *Escherichia coli* protein coding sequences. J Mol Biol 238:1–8. doi:10.1006/jmbi.1994.1263
29. Althaus IW, Gonzales AJ, Chou JJ, Romero DL, Diebel MR, Chou KC, Kezdy FJ, Resnick L, Busso ME, So AG, Downey KM, Thomas RC, Aristoff PA, Tarpley WG, Reusser F (1993) The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem 268:14875–14880
30. Chou KC, Kezdy FJ, Reusser F (1994) Review: steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal Biochem 221:217–230. doi:10.1006/abio.1994.1405
31. Chou KC, Zhang CT, Elrod DW (1996) Do "antisense proteins" exist? J Protein Chem 15:59–61. doi:10.1007/BF01886811
32. Qi XQ, Wen J, Qi ZH (2007) New 3D graphical representation of DNA sequence based on dual nucleotides. J Theor Biol 249:681–690. doi:10.1016/j.jtbi.2007.08.025
33. Vázquez-Prieto S, González-Díaz H, Paniagua E, Vilas R, Ubeira FM (2014) A QSPR-like model for multilocus genotype networks of *Fasciola hepatica* in Northwest Spain. J Theor Biol 343:16–24. doi:10.1016/j.jtbi.2013.11.005
34. González-Díaz H, Prado-Prado F (2008) Unified QSAR and network-based computational chemistry approach to antimicrobials, Part 1: multispecies activity models for antifungals. J Comput Chem 29:656–657. doi:10.1002/jcc.20826
35. González-Díaz H, Prado-Prado FJ, Santana L, Uriarte E (2006) Unify QSAR approach to antimicrobials. Part 1: predicting antifungal activity against different species. Bioorg Med Chem 14:5973–5980. doi:10.1016/j.bmc.2006.05.018
36. Prado-Prado FJ, Borges F, Perez-Montoto LG, González-Díaz H (2009) Multi-target spectral moment: QSAR for antifungal drugs vs. different fungi species. Eur J Med Chem 44:4051–4056. doi:10.1016/j.ejmech.2009.04.040
37. Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, Larsen B, Lin ZY, Breitkreutz BJ, Stark C, Liu G, Ahn J, Dewar-Darch D, Reguly T, Tang X, Almeida R, Qin ZS, Pawson T, Gingras AC, Nesvizhskii AI, Tyers M (2010) A global protein kinase and phosphatase interaction network in yeast. Science 328:1043–1046. doi:10.1126/science.1176495
38. Lin YY, Qi Y, Lu JY, Pan X, Yuan DS, Zhao Y, Bader JS, Boeke JD (2008) A comprehensive synthetic genetic interaction network governing yeast histone acetylation and deacetylation. Genes Dev 22:2062–2074. doi:10.1101/gad.1679508
39. Nandy SK, Jouhten P, Nielsen J (2010) Reconstruction of the yeast protein-protein interaction network involved in nutrient sensing and global metabolic regulation. BMC Syst Biol 4:68. doi:10.1186/1752-0509-4-68
40. Shi MG, Huang DS, Li XL (2008) A protein interaction network analysis for yeast integral membrane protein. Protein Pept Lett 15:692–699. doi:10.2174/092986608785133627
41. González-Díaz H, Pérez-Montoto LG, Ubeira FM (2014) Model for vaccine design by prediction of B-epitopes of IEDB given perturbations in peptide sequence, in vivo process, experimental techniques, and source or host organisms. J Immunol Res. doi:10.1155/2014/768515
42. Vázquez-Prieto S, Paniagua E, Ubeira FM, González-Díaz H (2016) QSPR-perturbation models for the prediction of B-epitopes from immune epitope database: a potentially valuable route for predicting "in silico" new optimal peptide sequences and/or boundary conditions for vaccine development. Int J Pept Res Ther 22:445–450. doi:10.1007/s10989-016-9524-x
43. Junker BH, Koschützki D, Schreiber F (2006) Exploration of biological network centralities with CentiBiN. BMC Bioinform 7:219. doi:10.1186/1471-2105-7-219
44. Batagelj V, Mrvar A (1998) Pajek: a program for large network analysis. Connections 21:47–57. doi:10.1017/cbo9780511996368
45. StatSoft, Inc. (2002) STATISTICA (data analysis software system), version 6.0. www.statsoft.com