SciForum
Mol2Net

Complex Networks of anti-HIV Drugs Activity *vs.* Prevalence of AIDS in US Counties Using Symmetry Information Indices

Diana María Herrera-Ibatá [1,*] and Ricardo Alfredo Orbegozo-Medina [2]

[1] Department of Information and Communication Technologies, University of A Coruña UDC, 15071 Ferrol, A Coruña, Spain

[2] Department of Microbiology and Parasitology, University of Santiago de Compostela (USC), 15782 Santiago de Compostela, A Coruña, Spain

* Author to whom correspondence should be addressed; E-Mail: dianamariahi@gmail.com.

Received: 6 October 2015 / Accepted: 6 October 2015 / Published: 2 December 2015

Abstract: Different aspects about the epidemiology, drugs, targets, chem-bioinformatics, and systems biology methods, related to AIDS/HIV have been reviewed. Next, we developed a new model to predict complex networks of the AIDS prevalence in U.S. counties taking into consideration the Gini coefficient (income inequality) and activity/structure data of anti-HIV drugs in preclinical assays. First, we trained different Artificial Neural Networks (ANNs) using as input Markov and Symmetry information indices of social networks and of molecular graphs, respectively. We obtained the data about AIDS prevalence and Gini coefficient from the AIDSVu database of the Rollins School of Public Health at Emory University and the data about anti-HIV compounds from ChEMBL database. To train/validate the model and predict the complex network we needed to analyze 43,249 data points including values of AIDS prevalence in 2310 US counties *vs.* ChEMBL results for 21,582 unique drugs, 9 viral or human protein targets, 4856 protocols, and 10 possible experimental measures. The best model found was a Linear Neural Network (LNN) with Accuracy, Specificity, Sensitivity, and AUROC above 0.72-0.73 in training and external validation series. The new linear equation was shown to be useful to generate complex network maps of drug activity *vs*. AIDS/HIV epidemiology in U.S. at county level.

Keywords: anti-HIV drugs, Gini coefficient, neighborhood symmetry indices; complex networks.

Mol2Net YouTube channel: *http://bit.do/mol2net-tube*

1. Introduction

Human immunodeficiency virus (HIV) is a retrovirus belonging to the family of lentiviruses that causes AIDS. Retroviruses[1] can use their RNA and host DNA to make viral DNA, and are known for their long incubation periods. There are two types of HIV: HIV type 1 and HIV type 2. Despite progresses, HIV[2] remains a public health challenge. After thirty years in the AIDS epidemic, there are over 34 million people living with HIV[3], and still 2.5 million new infections and 1.7 million deaths each year.

A useful chemoinformatics-pharmacoepidemiology model must be multi-level to account molecular and population structure. We need to process diverse types of input data. Initially, we need the information about the anti-HIV drugs, such as chemical structure of the drug (level i) and preclinical information, like biological targets (level ii), organisms (level iii), or assay protocols (level iv). Afterwards, we need to incorporate population structure descriptors (level v) that quantify the epidemiological and socioeconomic factors affecting the population selected for the study.

2. Results and Discussion

After analysis of the previous results, we decided to test the predictive power of these indices in a simpler model using the STATISTICA 6.0[4] software. In so doing, we trained the LNN predictors using only each family of information indices of drugs ($^{q}IC_{5f}$) of 5- order, their MA operators ($\Delta^{q}IC_{5fj}$)) and the fifth MA operator of the U.S. counties ($\Delta I^{a}5s$). The LNN model based on qIC_{51} (LNN-IC_{51}) presented the higher values of Sn = 72.04/72.81 and Sp = 72.38/72.50 in training/ and external validation sets (see Table 1). LNN-IC_{51} presented also the higher values for the AUROC in train and validation series (0.73 and 0.74 respectively). Analyzing all the previous results for this dataset, we found that the IC_k index appears to be the most important to predict the drug structure-activity relationships. We can conclude it by comparison to the other indices, which have lower values of classification. The equation of LNN-IC_{51} this model is the following:

$$\begin{aligned} S_{aq}(\dot{c}_j) = & -25.48\cdot {}^{q}IC_{51} + 1081.64\cdot\Delta^{q}IC_{51}(c_1) + 29.36\cdot\Delta^{q}IC_{51}(c_2) \\ & - 1084.52\cdot\Delta^{q}IC_{51}(c_3) - 0.7727\cdot\Delta^{q}IC_{51}(c_4) \\ & - 0.0792\cdot\Delta I^{a}{}_{5}(s) - 0.5025 \end{aligned}$$

Last, we used this LNN-ALMA model to generate/predict a complex network of the prevalence of AIDS in the United States at county level with respect to the preclinical activity of anti-HIV drugs (Figure 1). The bipartite network has two types of nodes (counties vs. drug). Thus, this is a multiscale network similar to bipartite networks of drugs vs. target proteins reported by other groups[5-7]. However, the nodes in the present network contain information about the molecules, i.e., chemical structure as well as assay conditions (target protein, organism, experimental measure, etc.). Additionally, the other set of nodes contain information about socioeconomic factors, such as the income inequality in the county.

Multiscale networks of this type have been discussed by Barabasi et al.[8] as one of the more important tools to perform trans-disciplinary research. The links of this complex network are the outputs Laq(cj)pred = 1 of our model. In Figure 1, we illustrate the sub-network of AIDS prevalence vs. Anti-HIV drug preclinical activity for the state of Florida. For instance, the model predicts a high effectively for the drug Zidovudine to treat AIDS in Nassau County.

Table 1. LNN classifier for symmetry information indices of 5-order

Type of Index	Observed	$L_{pq} = 1$	$L_{pq} = 0$	$L_{pq} = 1$	$L_{pq} = 0$	AUROC
		Training		**Validation**		
$^{q}IC_{51}$	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
	Predicted	72.04	72.38	72.81	72.50	0.73 / 0.73
	$L_{pq} = 1$	8255	5746	2775	1908	
	$L_{pq} = 0$	3203	15060	1036	5031	

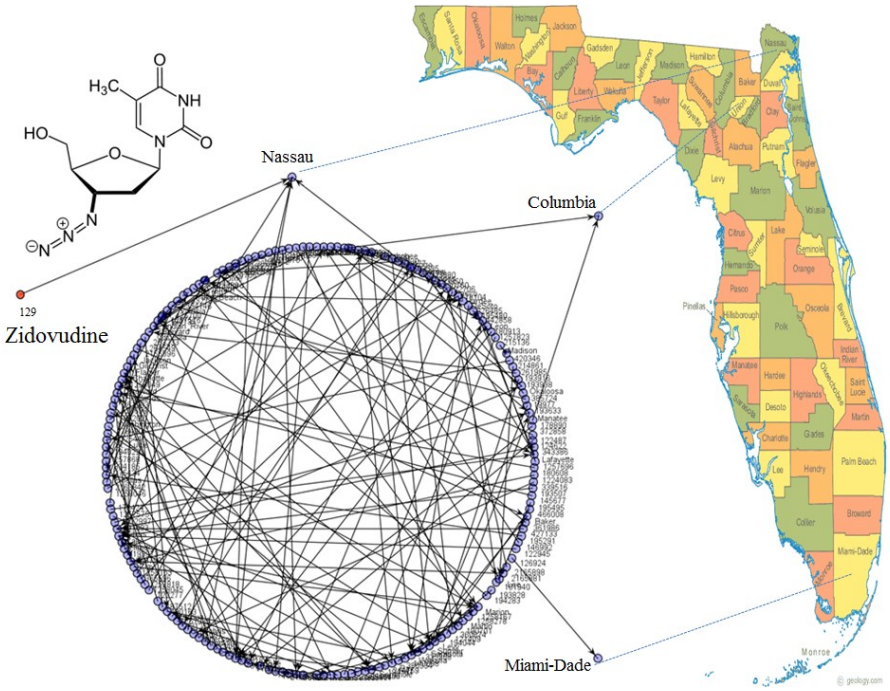

Figure 1. Sub-network of AIDS prevalence *vs.* Anti-HIV drug activity for U.S. state of Florida (FL)

3. Materials and Methods

In the present paper, we changed the Balaban information indices ($I^{q}k$) by Symmetry information content indices ($^{q}IC_{kf}$)[9]. These indices are calculated for H-included molecular graph and based on neighbor degrees and edge multiplicity.[10, 11] The symmetry information

indices are calculated by partitioning graph vertices into equivalence classes; the topological equivalence of two vertices is that the corresponding neighborhoods of the kth order are the same. However, we used the $I^a k(s)$ indices to characterize the different populations. We used the software DRAGON[12] to calculate the $^qIC_{kf}$ indices for the molecules of the ChEMBL dataset of anti-HIV drugs. In this case we calculated a total of $N_{indices}$ = Nk•Nf = 6*5 = 30 values of $^qIC_{kf}$ indices with Nk = 6 different orders (k) that belong to Nf = 5 different families of descriptors (f). We have used Markov chains to calculate Shannon information indices of different systems including simulations of disease spreading relevant to epidemiology. [13]

The codification of the chemical structure of the compounds is the first step here. We have data about a large number of assays developed in very different conditions (c_j) for equal or different targets (molecular or not). The non-structural information here refers to different assay conditions (c_j) like concentrations, temperature, targets, organisms, *etc.* A solution may rely upon the use of the idea of Moving Average (MA) operators used in time series analysis with a similar purpose. We have developed a similar approach called ALMA (Assessing Links with Moving Averages) using also MA operators. ALMA models remember those used in ARIMA models of time series analysis[14]. They are adaptable to all molecular descriptors and/or graphs invariants or descriptors for complex networks. In consonance with the previous section, we use a similar terminology. The inputs of one ALMA model are the descriptors D^q_k of type kth of the qth system (compound or drug d_q in this case) represented by a matrix M. On the other hand, the outputs of one ALMA model are the links (Laq = 1 or Laq = 0) of a complex network with Boolean matrix L and formed by different pairs of input systems. We developed different ANN models using all the set of parameters as well as simple models using different sub-sets of descriptors. The new ALMA model developed using these other set of indices has the following general form:

$$
\begin{aligned}
S_{aqj} = & \sum_{k=0}^{k=5}\sum_{f=1}^{f=5} e_{kf}\cdot {}^{q}IC_{kf} \\
& + \sum_{k=0}^{k=5}\sum_{f=1}^{f=5}\sum_{j=1}^{j=4} e_{kfj}\cdot \Delta {}^{q}IC_{kfj} \\
& + \sum_{k=1}^{k=5} e_{ak}\cdot \Delta I^{a}{}_{ks} + e_0 \\
= & \sum_{k=0}^{k=5}\sum_{f=1}^{f=5} e_{kf}\cdot {}^{q}IC_{kf} \\
& + \sum_{k=0}^{k=5}\sum_{f=1}^{f=5}\sum_{j=1}^{j=4} e_{kfj}\cdot \left({}^{q}IC_{kf} - \left\langle {}^{q}IC_{kf} \right\rangle_j \right) \\
& + \sum_{k=1}^{k=5} e_k \cdot \left(I^{a}{}_{k} - \left\langle I^{a}{}_{k} \right\rangle_s \right) + e_0
\end{aligned}
$$

4. Conclusions

This work presents a review of several aspects of the disease, including the epidemiology, pathophysiology, treatments, etc. We also developed a model called LNN-ALMA to generate complex networks of the prevalence of AIDS in the counties of the U.S. with respect to the preclinical activity of anti-HIV drugs. The best classifier found was the LNN-$IC_{51;}$ this classifier has only six inputs based on neighborhood information content indices, compared to the other models, the IC_k index seems to be the most important to predict the drug structure-activity relationships. The new model has similar performance but is notably simpler than a previous model based on Balaban's information indices with >20 inputs.

Acknowledgments

R.O.M acknowledges financial support of FPI fellowship associated to research project (AGL2011-30563-C03-01) funded by MECD (Spanish Ministry of Education, Culture and Sport).

Conflicts of Interest

"The authors declare no conflict of interest".

References and Notes

1. Lindemann, D.; Steffen, I.; Pohlmann, S., Cellular entry of retroviruses. *Advances in experimental medicine and biology* **2013**, 790, 128-49.
2. Moss, J. A., HIV/AIDS Review. *Radiologic technology* **2013**, 84, 247-67; quiz p.268-70.
3. Piot, P.; Quinn, T. C., Response to the AIDS pandemic--a global health model. *The New England journal of medicine* **2013**, 368, 2210-8.
4. *STATISTICA*, version 6.0; StatSoft Inc.: Tulsa, Oklahoma, 2001.
5. Prado-Prado, F.; Garcia-Mera, X.; Escobar, M.; Alonso, N.; Caamano, O.; Yanez, M.; Gonzalez-Diaz, H., 3D MI-DRAGON: new model for the reconstruction of US FDA drug- target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for AChE. *Current topics in medicinal chemistry* **2012**, 12, 1843-65.
6. Prado-Prado, F.; Garcia-Mera, X.; Abeijon, P.; Alonso, N.; Caamano, O.; Yanez, M.; Garate, T.; Mezo, M.; Gonzalez-Warleta, M.; Muino, L.; Ubeira, F. M.; Gonzalez-Diaz, H., Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica. *European journal of medicinal chemistry* **2011**, 46, 1074-94.
7. Vina, D.; Uriarte, E.; Orallo, F.; Gonzalez-Diaz, H., Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. *Molecular pharmaceutics* **2009**, 6, 825-35.
8. Barabasi, A. L.; Gulbahce, N.; Loscalzo, J., Network medicine: a network-based approach to human disease. *Nature reviews. Genetics* **2011**, 12, 56-68.
9. González-Díaz, H.; Herrera-Ibatá, D. M.; Duardo-Sanchez, A.; Munteanu, C. R.; Orbegozo-Medina, R. A.; Pazos, A., Model of the Multiscale Complex Network of AIDS prevalence in US at county level vs. Preclinical activity of anti-HIV drugs based on information indices of molecular graphs and social networks. *Journal of chemical information and modeling* **2014**, 54, 744-755.
10. Magnuson, V. R.; Harriss, D. K.; Basak, S. C. In *Studies in Physical and Theoretical Chemistry; King, R.B.*; Elsevier: Amsterdam (The Netherlands), 1983, pp 178-191.
11. Todeschini, R.; Consonni, V., *Handbook of Molecular Descriptors*. Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000.
12. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M., In; Talete srl: Milano, Italy, 2005.
13. 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. *Journal of theoretical biology* **2012**, 293, 174-88.
14. Langenfeld, M. C.; Cipani, E.; Borckardt, J. J., Hypnosis for the control of HIV/AIDS-related pain. *The International journal of clinical and experimental hypnosis* **2002**, 50, 170-88.