Computational Prediction of ATC Codes of Drug-Like Compounds Using Tiered Learning

Thomas Olson[1*] and Rahul Singh[1*]
[1]Department of Computer Science, San Francisco State University

Abstract—The Anatomical Therapeutic Chemical (ATC) Code System is a World Health Organization (WHO) proposed classification that assigns codes to compounds based on their therapeutic, pharmacological and chemical characteristics as well as the in-vivo site of activity. The ability to predict the ATC code of an arbitrary compound with high accuracy can go a long way in selecting molecules for lead identification. We propose a computational approach to this problem that utilizes a natural pharmacological constraint, namely, that anatomical-therapeutic biological activity of certain types must preclude activities of many other types. The method proposed here utilizes machine learning in a tiered architecture; prediction of the ATC code at a certain level is constrained by the ATC code at the higher levels. Using this learning architecture, we have built classifiers that incorporate information from a compound's structure, as well as its chemical and protein interactions. The proposed approach has been validated using 2335 drugs from the ChEMBL database in both cross-validation and test setting. The prediction accuracy obtained with this approach is 78.72% and is comparable or better than the prediction accuracy of other methods at the state of the art.

I. Introduction

In this paper, we present a propose a tiered learning architecture for predicting ATC codes. Several methods have been published on ATC code prediction. Amongst these, Chen [1] *et al.* made predictions based on biochemical information and structural information. SuperPred [2] utilizes 3D structure similarity and fragment-based similarity and NetPredATC [3] used a Support Vector Machine (SVM) for predictions based on compound structural similarity and target similarity.

II. Method Summary

The dataset used for training and testing the proposed learning architecture consisted of 2335 molecules selected from the ChEMBL database [4], with interaction data associated with these compounds retrieved from the STITCH database [5].

In the proposed tiered learning architecture, in order to predict the ATC code at the k_{th} level, the ATC code of the previous (higher) k-1_{st} level has to be available. In other words, successive levels of ATC hierarchy map to increasingly conserved parts of the chemical space. Operationally, after the first letter of ATC code is predicted, the training data is filtered so that only compounds sharing the predicted ATC code of the prior level are used for predicting the subsequent level.

III. Summary of Expeirments

Six learning algorithms including, a Bayesian classifier, a Multilayer Perceptron, a Support Vector Machine (SVM), a Random Forest Classifier, a J48 Decision Tree, and a Reduced Error Pruning (REP) Tree were investigated. The Bayesian classifier performed the best among these and was chosen to compare and test our method against prior works. The results from these comparisons are presented in Table 1.

Table 1 Comparison of the methods for predicting ATC codes at varying depths. Chen *et al.* and NetPredATC do not report prediction accuracy at depth 3 and beyond and are indicated through a "--".

Method	***Method details***			
	Data size	***Features used***	***Accuracy at depth 1***	***Maximum Accuracy at any depth***
Proposed (Bayesian classifier)	**2335**	**Chemical and Protein Interactions, and Structural Similarity**	**78.7%**	**Depth 3 89.7%**
SuperPred [8]	1040	3D Structure and Fragment Similarity	80.90%	Depth 5 75.1%
Chen *et. al.* [5]	3947	Chemical interactions, structural similarity, and ontologies	75.9%	--
NetPredATC [10]	790	Compound and target structural similarities	74%	--

Acknowledgments

This research was funded in part by the NSF grant IIS 0644418 (CAREER) and a grant from the Center for Computing in Life Sciences at San Francisco State University.

References

[1] L. Chen, J. Lu, N. Zhang, T. Huang, and Y. Cai, "A hybrid method for prediction and repositioning of drug Anatomical Therapeutic Chemical classes." *Molecular BioSystems* , 2014, pp. 868-877.

[2] J. Nickel,B-O. Gohlke,J. Erehman *et al*., "SuperPred: update on drug classification and target prediction". *Nucleic Acids Research* 2014, Vol. 42 (Web Server issue), W26-W31

[3] Y. Wang, S.L Chen, N.Y. Deng and Y. Wang. "Network predicting drug's anatomical therapeutic chemical code" *Bioinformatics* 2013, Vol. 29 no. 10 2013, pp 1317–1324

[4] A. Gaulton, L. Bellis, "ChEMBL: a large-scale bioactivity database for drug discovery", *Nucleic Acids Research* 2012, Vol. 40 (D1): D1100-D1107

[5] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T. H. Blicher, C. Von Mering, L. J. Jensen, and P. Bork. "STITCH 4: integration of protein–chemical interactions with user data" *Nucleic Acids Research* 2014, Vol. 42 (D1): D401-D407

*Equal Contributors

978-1-4673-9963-9/15/$31.00 ©2015 IEEE