Scientific Research Publishing

Advances in Bioscience and Biotechnology, 2020, 11, 153-159
https://www.scirp.org/journal/abb
ISSN Online: 2156-8502
ISSN Print: 2156-8456

iATC_Deep-mISF: A Multi-Label Classifier for Predicting the Classes of Anatomical Therapeutic Chemicals by Deep Learning

Zhe Lu[1], **Kuo-Chen Chou**[1,2]

[1]Computer Science, Jingdezhen Ceramic Institute, Jingdezhen, China
[2]Gordon Life Science Institute, Boston, MA 02478, USA
Email: 454170054@qq.com, kcchou@gordonlifescience.org, kcchou38@gmail.com

How to cite this paper: Lu, Z. and Chou, K.-C. (2020) iATC_Deep-mISF: A Multi-Label Classifier for Predicting the Classes of Anatomical Therapeutic Chemicals by Deep Learning. *Advances in Bioscience and Biotechnology*, **11**, 153-159.
https://doi.org/10.4236/abb.2020.115012

Received: April 8, 2020
Accepted: May 8, 2020
Published: May 11, 2020

Abstract

The recent worldwide spreading of pneumonia-causing virus, such as Coronavirus, COVID-19, and H1N1, has been endangering the life of human beings all around the world. To provide useful clues for developing antiviral drugs, information of anatomical therapeutic chemicals is vitally important. In view of this, a CNN based predictor called "iATC_Deep-mISF" has been developed. The predictor is particularly useful in dealing with the multi-label systems in which some chemicals may occur in two or more different classes. To maximize the convenience for most experimental scientists, a user-friendly web-server for the new predictor has been established at http://www.jci-bioinfo.cn/iATC_Deep-mISF/, which will become a very powerful tool for developing effective drugs to fight pandemic coronavirus and save the mankind of this planet.

Keywords

Pandemic Coronavirus, Multi-Label System, Anatomical Therapeutic Chemicals, Learning at Deeper Level, Five-Steps Rule

1. Introduction

According to the ATC (Anatomical Therapeutic Chemical) system (http://www.whocc.no/atc/structure_and_principles) as recommended by WHO (World Health Organization), the drug compounds are categorized into the following 14 main groups: 1) alimentary tract and metabolism; 2) blood and blood forming organs; 3) cardiovascular system; 4) dermatologicals; 5) genitourinary system and sex hormones; 6) systemic hormonal preparations, excluding sex hormones and insulins; 7) anti-infectives for systemic use; 8) antineoplastic and immunomodulating agents; 9) musculoskeletal system; 10) nervous system; 11) antiparasitic products, insecticides and repellents; 12) respiratory system; 13) sensory organs; 14) various. Given an uncharacterized compound, can we identify which ATC-class it belongs to? It is no doubt a significant problem for both basic research and drug development.

In 2017, a powerful predictor called "iATC-mISF", was developed, which is overwhelmingly superior to its counterparts. But the method has not been further treated with the Deep Learning yet, a very powerful technique [1] [2]. The present study was devoted to doing so.

According to the 5-step guidelines [3] and demonstrated in a series of recent publications (see, e.g., [4] [5]), to develop a statistical predictor that not only can be easily used by experimental scientists but also can stimulate theoretical scientists to develop more relevant ones, we should make the following five steps crystal clear: 1) benchmark dataset, 2) sample formulation, 3) operation algorithm, 4) anticipated accuracy, and 5) web-server. Below, we are to elaborate how to deal with these procedures one-by-one.

2. Materials and Methods

2.1. Benchmark Dataset

The benchmark dataset used in this study is exactly the same as that in iATC-mSMF [6]; *i.e.*,

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \cdots \cup \mathbb{S}_m \cup \cdots \cup \mathbb{S}_{13} \cup \mathbb{S}_{14} \tag{1}$$

where the subset $\mathbb{S}_m$ only contains the samples from the m-th ATC class $(m = 1, 2, 3, \cdots, 14)$, and $\cup$ denotes the symbol for "union" in the set theory. See Online Supporting Information S2 for a breakdown of the benchmark dataset according to the 14 subsets in Equation (1).

2.2. Installing Deep-Learning for Three Deeper Levels

In this study, we use multilayer perceptron neural network model, which consists of 3 fully connected layers and was used to predict classes of multi-label ATC classes, as illustrated in **Figure 1**. We set input layer with 14 neural unGranits which correspond to 14 features. Too many hidden layers would make network complexity bigger and suffer from the vanishing gradient problem while a model is constructed. Here, only two hidden layers are included. The hidden layer 1 is set as 200 neural units. The activation function is set as "relu". The second hidden layer has 100 neural units. The activation function is set the same as the hidden layer 1. We end the model with 14 neural units and sigmoid activation. To go with it, we use the binary_crossentropy loss and the adam (adaptive moment estimation) optimizer to train the model. The metrics is set as "accuracy". The batch size is set as 28, and the epochs is 100. The predicted results

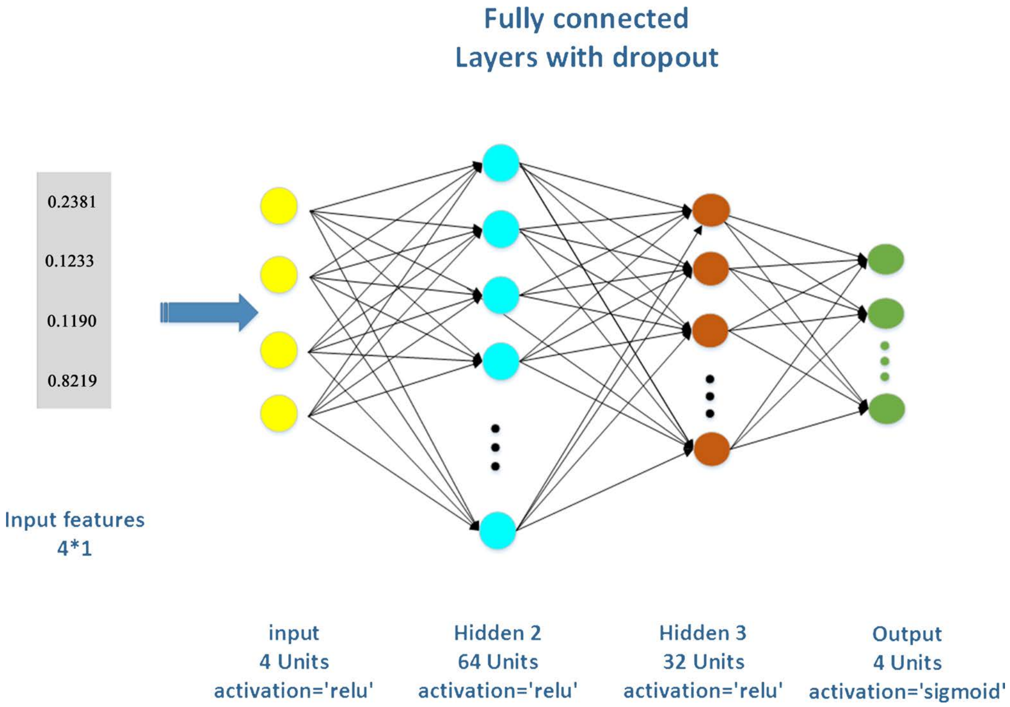

Figure 1. An illustration to show a dense neural network with 3 fully connected layers. Adapted from [1] with permission.

were decided by the output of the threshold θ. If the output is greater than 0.5, the outcome was true; otherwise, false. For more information about this, see [1], where the details have been clearly elaborated and hence there is no need to repeat here.

The new predictor developed via the above procedures is called "iATC_Deep-mISF", where "iATC_Deep" stands for "predict anatomical therapeutic chemicals", and "mISF" for "multi-label classes".

3. Results and Discussion

According to the 5-step rules [3], one of the important procedures in developing a new predictor is how to properly evaluate its anticipated accuracy. To deal with that, two issues need to be considered. 1) What metrics should be used to quantitatively reflect the predictor's quality? 2) What test method should be applied to score the metrics?

3.1. A Set of Five Metrics for Multi-Label Systems

Different from the metrics used to measure the prediction quality of single-label systems, the metrics for the multi-label systems are much more complicated. To make them more intuitive and easier to understand for most experimental scientists, here we use the following intuitive Chou's five metrics [7] or the "global metrics" that have recently been widely used for studying various multi-label systems (see, e.g., [8] [9]). For the current study, the set of global metrics can be formulated as:

$$\begin{cases} \text{Aiming} \uparrow = \dfrac{1}{N^{\mathrm{q}}} \sum_{k=1}^{N^{\mathrm{q}}} \left(\dfrac{\left\| \mathbb{L}_k \cap \mathbb{L}_k^* \right\|}{\left\| \mathbb{L}_k^* \right\|} \right), & [0,1] \\ \text{Coverage} \uparrow = \dfrac{1}{N^{\mathrm{q}}} \sum_{k=1}^{N^{\mathrm{q}}} \left(\dfrac{\left\| \mathbb{L}_k \cap \mathbb{L}_k^* \right\|}{\left\| \mathbb{L}_k \right\|} \right), & [0,1] \\ \text{Accuracy} \uparrow = \dfrac{1}{N^{\mathrm{q}}} \sum_{k=1}^{N^{\mathrm{q}}} \left(\dfrac{\left\| \mathbb{L}_k \cap \mathbb{L}_k^* \right\|}{\left\| \mathbb{L}_k \cup \mathbb{L}_k^* \right\|} \right), & [0,1] \\ \text{Absolute true} \uparrow = \dfrac{1}{N^{\mathrm{q}}} \sum_{k=1}^{N^{\mathrm{q}}} \Delta\left(\mathbb{L}_k, \mathbb{L}_k^* \right), & [0,1] \\ \text{Absolute false} \downarrow = \dfrac{1}{N^{\mathrm{q}}} \sum_{k=1}^{N^{\mathrm{q}}} \left(\dfrac{\left\| \mathbb{L}_k \cup \mathbb{L}_k^* \right\| - \left\| \mathbb{L}_k \cap \mathbb{L}_k^* \right\|}{M} \right), & [1,0] \end{cases} \tag{2}$$

where N^{q} is the total number of query proteins or tested proteins, M is the total number of different labels for the investigated system (for the current study it is $L_{\mathrm{cell}} = 4$), $\| \ \|$ means the operator acting on the set therein to count the number of its elements, $\cup$ means the symbol for the "union" in the set theory, $\cap$ denotes the symbol for the "intersection", $\mathbb{L}_k$ denotes the subset that contains all the labels observed by experiments for the *k*-th tested sample, $\mathbb{L}_k^*$ represents the subset that contains all the labels predicted for the *k*-th sample, and

$$\Delta\left(\mathbb{L}_k, \mathbb{L}_k^* \right) = \begin{cases} 1, & \text{if all the labels in } \mathbb{L}_k^* \text{ are identical to those in } \mathbb{L}_k \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In Equation (4), the first four metrics with an upper arrow ↑ are called positive metrics, meaning that the larger the rate is the better the prediction quality will be; the 5th metrics with a down arrow ↓ is called positive metrics, implying just the opposite meaning.

From Equation (2) we can see the following: 1) the "Aiming" defined by the 1st sub-equation is for checking the rate or percentage of the correctly predicted labels over the practically predicted labels; 2) the "Coverage" defined in the 2nd sub-equation is for checking the rate of the correctly predicted labels over the actual labels in the system concerned; 3) the "Accuracy" in the 3rd sub-equation is for checking the average ratio of correctly predicted labels over the total labels including correctly and incorrectly predicted labels as well as those real labels but are missed in the prediction; 4) the "Absolute true" in the 4th sub-equation is for checking the ratio of the perfectly or completely correct prediction events over the total prediction events; 5) the "Absolute false" in the 5th sub-equation is for checking the ratio of the completely wrong prediction over the total prediction events.

3.2. Comparison with the State-of-the-Art Predictor

Listed in Table 1 are the rates achieved by the current iATC_Deep-mISF predictor via the cross validations on the same experiment-confirmed dataset as

Table 1. Comparison with the state-of-the-art method in predicting iATC-mISF[a].

Predictor	Aiming (↑)[a]	Coverage (↑)[a]	Accuracy (↑)[a]	Absolute true (↑)[a]	Absolute false (↓)[a]
iATC-mISF	67.83%	67.10%	66.41%	60.98%	5.85%
iATC_Deep-mISF[c]	74.7%	73.91%	71.57%	67.01%	0%

[a]See Equation (2) for the definition of the metrics. [b]See [6], where the reported metrics rates were obtained by the jackknife test on the benchmark dataset of Supporting Information S1 that contains experiment-confirmed proteins only. [c]The proposed predictor; to assure that the test was performed on exactly the same experimental data as reported in [6] for iATC-mISF.

used in [6]. For facilitating comparison, listed there are also the corresponding results obtained by the iATC-mISF predictor [6], the existing most powerful method for predicting the classes of anatomical therapeutic chemicals. As shown in Table 1, the newly proposed predictor iATC_Deep-mISF is remarkably superior to the existing state-of-the-art predictor iATC-mISF in all the five metrics. Particularly, it can be seen from the table that the absolute true rate achieved by the new predictor is over 67%, which is about 7% higher than iATC-mISF [6]. This is because it is extremely difficult to enhance the absolute true rate of a prediction method for a multi-label system as clearly elucidated in [6]. Actually, to avoid embarrassment, many investigators even chose not to mention the metrics of absolute true rate in dealing with multi-label systems (see, e.g., [10] [11]).

Meanwhile, as a byproduct, the present paper has also stimulated some very interesting or provoked papers (see, e.g., [12]-[17]).

3.3. Web Server and User Guide

As pointed out in [18], user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors. Actually, user-friendly web-servers will significantly enhance the impacts of theoretical work because they can attract the broad experimental scientists [19]. In view of this, the web-server of the current iATC_Deep-mISF predictor has also been established at http://www.jci-bioinfo.cn/iATC_Deep-mISF/, by which users can easily get their desired data without the need to go thru the mathematical details.

4. Conclusion

It is anticipated that the iATC_Deep-mISF predictor holds very high potential to become a useful high throughput tool in identifying the classes of anatomical therapeutic chemicals. Most important is that the predictor will become a very useful tool for fighting against the coronavirus to save mankind on this planet.

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (No. 31560316, 61261027, 61262038, 61202313 and 31260273), the Province National Natural Science Foundation of JiangXi (No.

20132BAB201053), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No.20120BDH80023), the Department of Education of JiangXi Province (GJJ160866).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Maxwell, A., Li, R., Yang, B., Weng, H., Ou, A., Hong, H., Zhou, Z., Gong, P. and Zhang, C. (2017) Deep Learning Architectures for Multi-Label Classification of Intelligent Health Risk Prediction. *BMC Bioinformatics*, **18**, 523. https://doi.org/10.1186/s12859-017-1898-z

[2] Khan, Z.U., Ali, F., Khan, I.A., Hussain, Y. and Pi, D. (2019) iRSpot-SPI: Deep Learning-Based Recombination Spots Prediction by Incorporating Secondary Sequence Information Coupled with Physio-Chemical Properties via Chou's 5-Step Rule and Pseudo Components. *Chemometrics and Intelligent Laboratory Systems* (*CHEMOLAB*), **189**, 169-180. https://doi.org/10.1016/j.chemolab.2019.05.003

[3] Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition (50th Anniversary Year Review, 5-Steps Rule). *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024

[4] Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) iCar-PseCp: Identify Carbonylation Sites in Proteins by Monto Carlo Sampling and Incorporating Sequence Coupled Effects into General PseAAC. *Oncotarget*, **7**, 34558-34570. https://doi.org/10.18632/oncotarget.9148

[5] Liu, B., Long, R. and Chou, K.C. (2016) iDHS-EL: Identifying DNase I Hypersensitive Sites by Fusing Three Different Modes of Pseudo Nucleotide Composition into an Ensemble Learning Framework. *Bioinformatics*, **32**, 2411-2418. https://doi.org/10.1093/bioinformatics/btw186

[6] Cheng, X., Zhao, S.G., Xiao, X. and Chou, K.C. (2017) iATC-mISF: A Multi-Label Classifier for Predicting the Classes of Anatomical Therapeutic Chemicals. *Bioinformatics*, **33**, 341-346. (Corrigendum, ibid., 2017, Vol. 33, 2610) https://doi.org/10.1093/bioinformatics/btx387

[7] Chou, K.C. (2013) Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Molecular Biosystems*, **9**, 1092-1100. https://doi.org/10.1039/c3mb25555g

[8] Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N.D., Webb, G.I. and Chou, K.C. (2018) iProt-Sub: A Comprehensive Package for Accurately Mapping and Predicting Protease-Specific Substrates and Cleavage Sites. *Brief in Bioinform*, **20**, 638-658. https://doi.org/10.1093/bib/bby028

[9] Zhang, M., Li, F., Marquez-Lago, T.T., Leier, A., Fan, C., Kwoh, C.K., Chou, K.C., Song, J. and Jia, C. (2019) MULTiPly: A Novel Multi-Layer Predictor for Discovering General and Specific Types of Promoters. *Bioinformatics*, **35**, 2957-2965. https://doi.org/10.1093/bioinformatics/btz016

[10] Huang, C. and Yuan, J. (2013) Using Radial Basis Function on the General Form of Chou's Pseudo Amino Acid Composition and PSSM to Predict Subcellular Locations of Proteins with Both Single and Multiple Sites. *Biosystems*, **113**, 50-57. https://doi.org/10.1016/j.biosystems.2013.04.005

[11] Pacharawongsakda, E. and Theeramunkong, T. (2013) Predict Subcellular Locations of Singleplex and Multiplex Proteins by Semi-Supervised Learning and Dimension-Reducing General Mode of Chou's PseAAC. *IEEE Transactions on Nanobioscience*, **12**, 311-320. https://doi.org/10.1109/TNB.2013.2272014

[12] Chou, K.C. (2020) The Development of Gordon Life Science Institute: Its Driving Force and Accomplishments. *Natural Science*, **12**, 202-217. https://doi.org/10.4236/ns.2020.124018

[13] Chou, K.C. (2020) The Most Important Ethical Concerns in Science. *Natural Science*, **12**, 35-36. https://doi.org/10.4236/ns.2020.122005

[14] Chou, K.C. (2020) Other Mountain Stones Can Attack Jade: The 5-Steps Rule. *Natural Science*, **12**, 59-64. https://doi.org/10.4236/ns.2020.123011

[15] Chou, K.C. (2020) The Problem of Elsevier Series Journals Online Submission by Using Artificial Intelligence. *Natural Science*, **12**, 37-38. https://doi.org/10.4236/ns.2020.122006

[16] Chou, K.C. (2020) Proposing 5-Steps Rule Is a Notable Milestone for Studying Molecular Biology. *Natural Science*, **12**, 74-79. https://doi.org/10.4236/ns.2020.123011

[17] Chou, K.C. (2020) Using Similarity Software to Evaluate Scientific Paper Quality Is a Big Mistake. *Natural Science*, **12**, 42-58. https://doi.org/10.4236/ns.2020.123008

[18] Chou, K.C. and Shen, H.B. (2009) Recent Advances in Developing Web-Servers for Predicting Protein Attributes. *Natural Science*, **1**, 63-92. https://doi.org/10.4236/ns.2009.12011

[19] Chou, K.C. (2017) An Unprecedented Revolution in Medicinal Chemistry Driven by the Progress of Biological Science. *Current Topics in Medicinal Chemistry*, **17**, 2337-2358. https://doi.org/10.2174/1568026617666170414145508