IBS 2.0: an upgraded illustrator for the visualization of biological sequences

Yubin Xie[1,#], Huiqin Li[1,#], Xiaotong Luo[2,#], Hongyu Li[1], QiuYuan Gao[1], Luowanyue Zhang[1], Yuyan Teng[1], Qi Zhao[2], Zhixiang Zuo[2] and Jian Ren[1,2,*]

[1] School of Life Sciences, Precision Medicine Institute, the First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510060, China

[2] State Key Laboratory of Oncology in South China, Cancer Center, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University, Guangzhou 510060, China

[*]Correspondence to: Jian Ren (renjian@sysucc.org.cn).

[#] The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

SUPPLEMENTARY METHODS

Construction of a deep learning-based model for recognizing biological sequence diagrams from published literatures

To identify biological sequence diagrams from a large number of published papers, a deep convolutional neural network was designed based on the basic structure of Resnet152 architecture. The network model consists of 5 stages of convolution processing. Among them, the first stage consists of 64 convolutional layers with a step size of 2 and a max-pooling layer for feature dimensionality reduction. The second to fifth stages consist of 3, 8, 36, and 3 stacked bottleneck structures, respectively. After the fifth stage of convolution processing is completed, the high-dimensional features will be dimensionally reduced by the average-pooling layer and then output by the fully connected layer with the activation function setting to softmax.

For model training, we first collected a total of 11,440 figures from published literatures and DocFigure(1). According to their contents, we manually divided the collected figures into 15 different types, including biological sequence diagrams, bar charts, box plots, circos plots, etc. Detailed statistics were listed in Table S1. Another 20,682 pictures were collected from literatures published in Science and Cell journals from 2016 to 2020 as a test dataset.

Generally, obtaining a deep convolutional neural network model with excellent prediction accuracy and generalization ability requires a huge amount of training data to optimize millions of network parameters. However, in actual application, collecting training data on such a scale is unrealistic. Therefore, in considering the relatively small size of training data we have collected so far, a network-based migration learning method was applied in our study to perform model training. As shown in Figure S1A, we first trained the backbone network structure of Resnet152 based on the ImageNet datasets. Next, we transferred the pre-trained model into our project, and then fine-tune the model on our collected training dataset.

Based on our constructed model, we further evaluated its prediction performance using 4-fold cross-validation and independent test. Figure S1B and Figure S1C present the results of the confusion matrix evaluation, which show that our constructed model has strong robustness and excellent prediction performance in recognizing biological sequence diagrams. For reproducibility purpose, the training dataset had been deployed to Zenodo with a DOI number of 10.5281/zenodo.6477060. In addition, the source code was also released to GitHub, and can be freely available at https://github.com/RenLabBioinformatics/IBS.

SUPPLEMENTARY REFERENCES

1. Jobin, K.V., Mondal, A. and Jawahar, C.V. (2019), *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pp. 74-79.

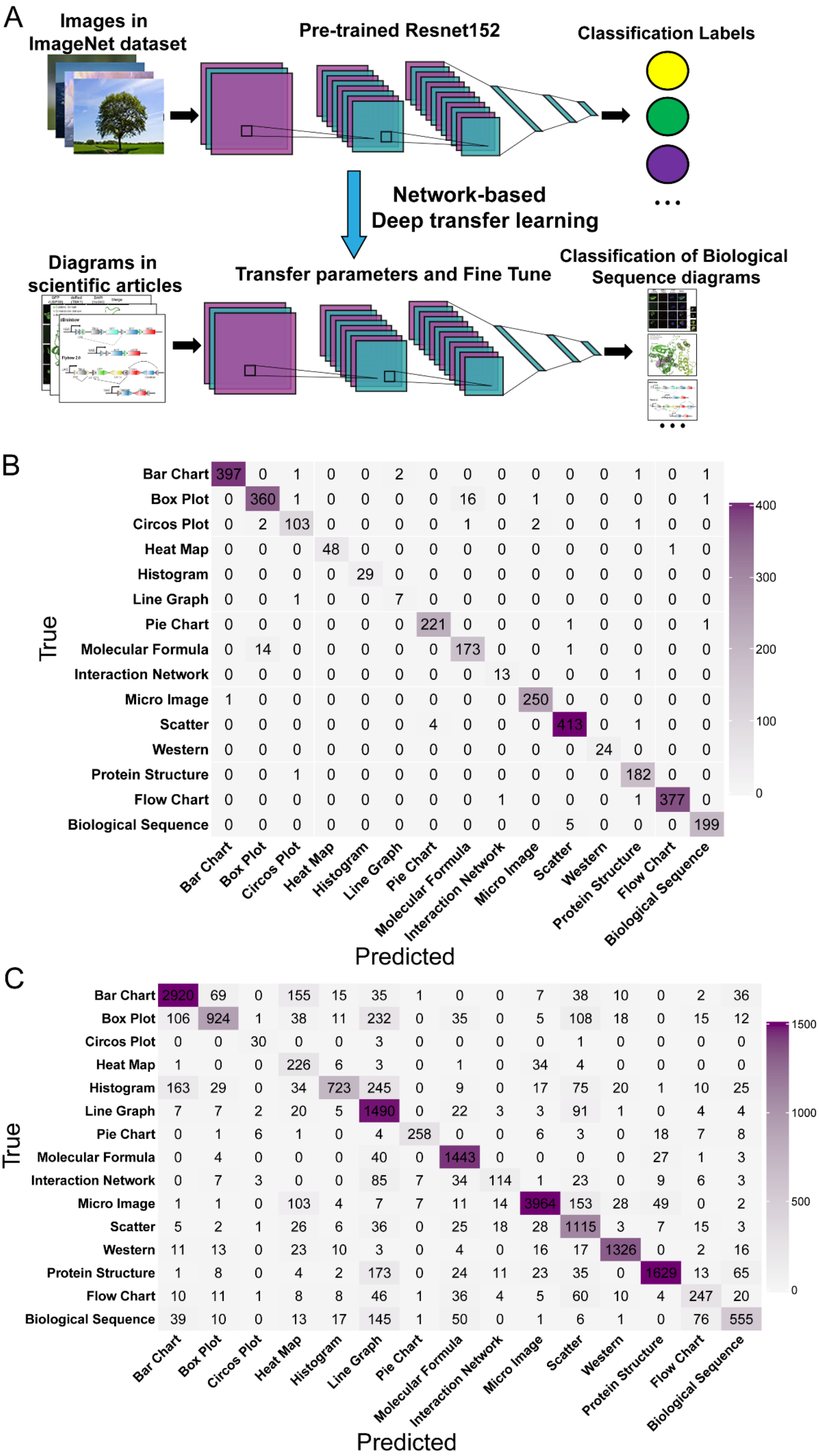
SUPPLEMENTARY TABLES

Table S1. The statistics of collected 15 types of diagrams for the training dataset.

Classification	Number
Bar Chart	1,516
Box Plot	436
Circos Plot	115
Heat Map	894
Histogram	752
Line Graph	1,006
Pie Chart	96
Molecular Formula	195
Interaction Network	53
Microscopic Image	1,672
Scatter	732
Western	815
Protein Structure	1,517
Flow Chart	32
Biological Sequence	1,609

SUPPLEMENTARY FIGURES AND LEGENDS

Figure S1. Construction of a deep learning-based model for recognizing biological sequence diagrams by network-based transfer learning. (**A**) Schematic diagram of the construction of the algorithm model. (**B**) The confusion matrix evaluation under 4-fold cross-validation on the training dataset. (**C**) The confusion matrix evaluation on the independent test dataset.



B

True \ Predicted	Bar Chart	Box Plot	Circos Plot	Heat Map	Histogram	Line Graph	Pie Chart	Molecular Formula	Interaction Network	Micro Image	Scatter	Western	Protein Structure	Flow Chart	Biological Sequence
Bar Chart	397	0	1	0	0	2	0	0	0	0	0	0	1	0	1
Box Plot	0	360	1	0	0	0	0	16	0	1	0	0	0	0	1
Circos Plot	0	2	103	0	0	0	0	1	0	2	0	0	1	0	0
Heat Map	0	0	0	48	0	0	0	0	0	0	0	0	0	1	0
Histogram	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0
Line Graph	0	0	1	0	0	7	0	0	0	0	0	0	0	0	0
Pie Chart	0	0	0	0	0	0	221	0	0	0	1	0	0	0	1
Molecular Formula	0	14	0	0	0	0	0	173	0	0	1	0	0	0	0
Interaction Network	0	0	0	0	0	0	0	0	13	0	0	0	1	0	0
Micro Image	1	0	0	0	0	0	0	0	0	250	0	0	0	0	0
Scatter	0	0	0	0	0	0	4	0	0	0	413	0	1	0	0
Western	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0
Protein Structure	0	0	1	0	0	0	0	0	0	0	0	0	182	0	0
Flow Chart	0	0	0	0	0	0	0	0	1	0	0	0	1	377	0
Biological Sequence	0	0	0	0	0	0	0	0	0	0	5	0	0	0	199

C

True \ Predicted	Bar Chart	Box Plot	Circos Plot	Heat Map	Histogram	Line Graph	Pie Chart	Molecular Formula	Interaction Network	Micro Image	Scatter	Western	Protein Structure	Flow Chart	Biological Sequence
Bar Chart	2920	69	0	155	15	35	1	0	0	7	38	10	0	2	36
Box Plot	106	924	1	38	11	232	0	35	0	5	108	18	0	15	12
Circos Plot	0	0	30	0	0	3	0	0	0	0	1	0	0	0	0
Heat Map	1	0	0	226	6	3	0	1	0	34	4	0	0	0	0
Histogram	163	29	0	34	723	245	0	9	0	17	75	20	1	10	25
Line Graph	7	7	2	20	5	1490	0	22	3	3	91	1	0	4	4
Pie Chart	0	1	6	1	0	4	258	0	0	6	3	0	18	7	8
Molecular Formula	0	4	0	0	0	40	0	1443	0	0	0	0	27	1	3
Interaction Network	0	7	3	0	0	85	7	34	114	1	23	0	9	6	3
Micro Image	1	1	0	103	4	7	7	11	14	3964	153	28	49	0	2
Scatter	5	2	1	26	6	36	0	25	18	28	1115	3	7	15	3
Western	11	13	0	23	10	3	0	4	0	16	17	1326	0	2	16
Protein Structure	1	8	0	4	2	173	0	24	11	23	35	0	1629	13	65
Flow Chart	10	11	1	8	8	46	1	36	4	5	60	10	4	247	20
Biological Sequence	39	10	0	13	17	145	1	50	0	1	6	1	0	76	555