World Scientific Research Journal Volume 7 Issue 5, 2021
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0010

Design of Protein Crystal Detection System Based on Mask RCNN

Jiangping Qin[1], Yan Zhang[1, *]

School of Logistics Engineering, Shanghai Maritime University, Shanghai, 201306, China

Abstract

Protein crystallography has developed rapidly since the first protein crystal structure was resolved by British scientists more than 50 years ago. The analysis of protein structure in protein crystallography based on X-ray diffraction experiments mainly depends on the high-diffraction resolution protein crystals screened by large-scale crystallization conditions, and the problem of protein crystallization is still a bottleneck in protein crystallography research. Therefore, an automatic detection system which is based on the Mask R-CNN instance segmentation algorithm is designed in this paper. This system could recognize the protein crystal image and send information about whether there are crystals in the image to researchers, which could reduce the waste of time of researchers.

Keywords

Protein crystal; Image segmentation; Automatic detection system.

1. INTRODUCTION

Protein is not only the fundamental matter of life, but it is also the main undertaker of life activities. Therefore, there is no life without protein. Proteins are involved in each cell and all important parts of the body, so protein plays a very key role in the process of self-regulation of the living body, such as immunity of antibodies (immunoglobulin), regulation of hormones and carrier transportation. Therefore, understanding the functional characteristics of protein which belongs to organic macromolecule, has always been the hot-button of structural biology research.

The characteristics of functions of the molecular mainly depend on its unique three-dimensional structure [1], so the three-dimensional structural characterization of protein is very important in order to understand its mechanism of action. It also can understand the mechanism of synergy between protein and other biological molecules by means of resolving protein structure. Meanwhile, it has a profound impact on the development of the field of biomedical science.

Protein crystallography is an important subject to study the three-dimensional structure of proteins and also an important branch of structural biology. Protein crystallography has many applications, such as the study of protein function, structure-based drug design, and the discovery of budding compounds (especially in the fragment-based drug screening) [2].

Researchers have been studying proteins since 1838, when the thing "protein" was first given, but early research focused on how the protein can be purified more easily. In 1895, the German physicist Johann Rontgen accidentally discovered X-rays. The discovery of X-rays not only won the first Nobel Prize in 1901, but also laid the foundation for the birth of protein crystallography. In 1957, Perutz resolved the low resolution spatial structure of the first protein that was myoglobin by means of isomorphic substitution, which It was also the first time that the structure of proteins [3], a biological macromolecules, was resolved by human by means of X-ray diffraction. Since that, protein crystallography had been a rapid development, methodology

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0010

also had been mature at the same time. In 1989, there been about only 300 the resolved structures of proteins in the Protein Data Bank (PDB). However, there had been 45000 structure in the PDB in 2007 which was a large progress. In 2021, 175000 protein structures have been deposited in the PDB, and more than 88% of them are resolved by the X-ray crystal diffraction method. Hence protein crystallography is still the main method to get protein three-dimensional structure. But there are great differences for crystallization conditions of different protein. Therefore, the problem of protein crystallization is still one of the bottlenecks in protein crystallography. So, how to design an automated system for protein crystallography from protein purification to crystal growth, has been becoming an urgent requirement in the field of life sciences [4].

Some work has been done to classify the images of protein crystals [5, 6], but due to insufficient computing of hardware resources, it is impossible to get good results. Bruno et al. [7] proposed a classification algorithm based on the deep convolutional neural network to classify protein crystallization results in 2018, which can achieve about 94% of the classification effect. However, it is only able to show whether there are crystals in the droplet and it couldn't provide the location information of crystals. At the same time, a number of automated crystallization devices have been developed by commercial companies to satisfy the requirement from academic centers and pharmaceutical companies. The fully automatic crystallization imaging system, Rockimager1500, was designed by the Formulatrix company [8]. Compared with the above mentioned studies, none of them could well meet the requirements of researchers.

In this paper, a software platform for protein crystal automatic recognition was designed based on the previous work [9], to achieve one-click recognition of protein crystal images. And, this system could save the result of protein crystals recognition to local. It can show researchers about the condition of crystal growth in the droplet by the result, without using artificial method to observe the crystal.

2. ALGORITHM INTRODUCTION

Mask R-CNN was proposed by the Facebook Researcher KM He [10], which integrates target detection and instance segmentation. The output of Mask R-CNN is divided into three parts: the prediction box regression, the image classification, and the mask branch. The network structure of Mask R-CNN is shown in Figure 1.

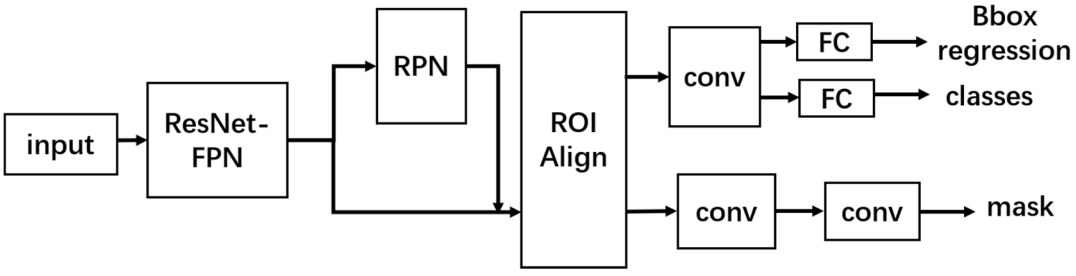

Figure 1. The network structure of Mask R-CNN

In Mask R-CNN structure, the protein crystal image is input into the network, and then different feature maps are output by means of a series of convolution and pooling in feature pyramid networks (FPN). After that, different feature maps are delivered into the region proposal networks (RPN) so as to extract the region of interest (ROI).

Then the ROI is input to the ROI Align to perform pixel correction on the feature map for subsequent target classification and bounding box regression. In the mask branch, the original

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0010

images are cropped using the corrected bounding box, and then the images in ROI are performed by mask prediction. Therefore, the object in the bounding box belongs to the two-class classification problem (0: background, 1: object). This can avoid inter-class competition and the final result belongs to instance segmentation. The total loss function of Mask R-CNN is defined as:

$$L_{total} = L_{cls} + L_{box} + L_{mask} \quad (1)$$

Where L_{cls} is classification loss; L_{box} is regression loss of bounding box; L_{mask} is semantic segmentation loss.

3. RESULTS AND ANALYSIS

3.1. Experiment Platform

The software environment for the experiment platform is based on Windows 10. The software platform is designed based on PyQT5 library of Python, and the Mask R-CNN algorithm is designed based on Keras 2.2.4 and TensorFlow 1.13. CPU is AMD R5 3600, the memory is 16G, graphics processing unit (GPU) is NVIDIA RTX2060. The interface of the protein crystal recognition assistant system is shown in Figure 2.

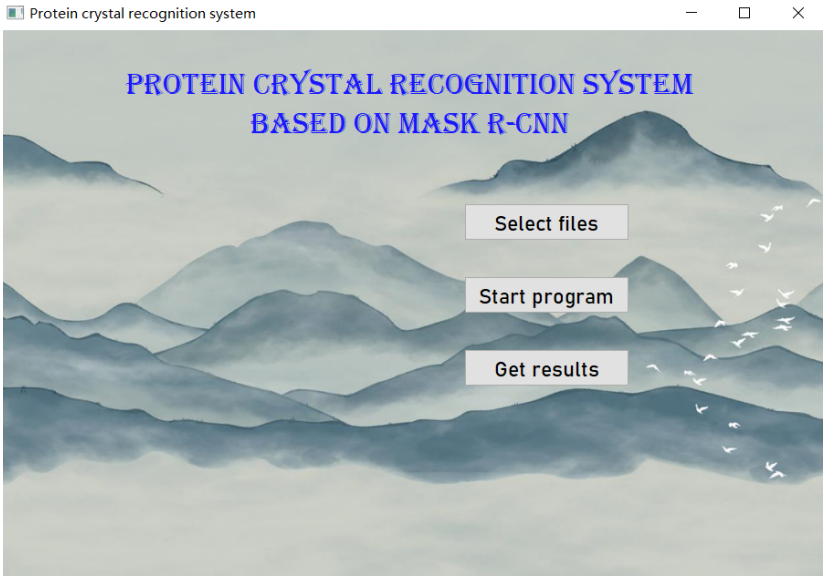

Figure 1. The interface of protein crystal recognition system

3.2. Experiment results

Firstly, select the folder where the protein crystal images is stored by clicking the "Select Files" button. Secondly, click "Start program" button to perform program for protein image segmentation. The program will read all images data in the folder, and then, these protein crystal images will be send to Mask R - CNN model one by one by program for image segmentation. If program identify those images which protein crystals are in the drop, the images id and identification results will be saved to the local database by program. Finally, clicking "Get result" button, the program result can be completely derived as TXT file (export). If the program is used in conjunction with a protein crystal viewer, the results would show the exact location of the holes where there are protein crystals in drop. The result example is shown in the Figure 2.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0010

(1,3) 4 crystals

Figure 2. The result of program. The coordinate value of the hole and the number of protein crystals are displayed respectively

According to the result list can know the condition of protein crystallization, Thus, the observation time is greatly saved.

4. CONCLUSION

The design of software platform for protein crystal automatic recognition can help researchers save a lot of time, and the result data can be used to analyze the crystallization conditions of the protein. According to the result, the researchers were also able to quickly retrieve suitable protein crystals from the plates.

REFERENCES

[1] F.X. B, X. Chen: Progresses on Protein Crystallization Process, Journal of Salt Science and Chemical Industry, Vol.46 (2017) No.7, p.1-3.

[2] M Spiliopoulou, A Valmas, D-P Triandafillidis, et al. Applications of X-ray Powder Diffraction in Protein Crystallography and Drug Screening, Crystals, Vol.10 (2020) No.2, p.1-35.

[3] L.F. Li, H. Nan, X.D. Su: Protein Crystallography Technology - Past, Present and Future, Acta Biophysica Sinica, Vol.23 (2007) No.4, p.247-255.

[4] P. Theveneau, P. Baker, R. Barrett, et al. The Upgrade Programme for the Structural Biology beamlines at the European Synchrotron Radiation Facility—High throughput sample evaluation and automation. In Proceedings of the 11th International Conference on Synchrotron Radiation Instrumentation (Lyon, France, July 9–13, 2012).

[5] G. Spraggon, S.A. Lesley, A. Kreusch and J.P. Priestle: Computational analysis of crystallization trials, Acta Crystallographica Section D, Vol.58 (2002) No.11, p.1915-1923.

[6] E.H. Snell, J.R. Luft, S.A. Potter, et al. Establishing a training set through the visual analysis of crystallization trials. Part I: approximately 150,000 images, Acta Crystallogr D Biol Crystallogr, Vol.64 (2008) No.11, p.1123-1130.

[7] A.E. Bruno, P. Charbonneau, J. Newman, et al. Classification of crystallization outcomes using deep convolutional neural networks, PLoS One, Vol.13 (2018) No.6, p.1-16.

[8] H.G. Jones, D. Wrapp, M.S. A Gilman, et al. Iterative screen optimization maximizes the efficiency of macromolecular crystallization, Acta Crystallogr F Struct Biol Commun, Vol.75 (2019) No.2, p.123-131.

[9] J.P. Qin, Y. Zhang, H. Zhou, et al. Protein Crystal Instance Segmentation Based on Mask R-CNN, Crystals, Vol.11 (2021) No.2, p.1-8.

[10] K.M. He, G. Gkioxari, P. Dollar and R. G: Mask R-CNN, IEEE Trans Pattern Anal Mach Intell, Vol.42 (2020) No.2, p.386-397.