

Project 1: Automated Particle Picking in Cryo-EM

(Due March 27, 2016)

Overview: In this project, you are asked to implement an algorithm to pick particles from micrograph in cryo-electron microscopy (cryo-EM). For final submission, you need to write a report in *English* to describe your algorithm, evaluate the performance of your algorithm through numerical results, and make a discussion.

Policy: You are allowed to use any reference from papers, books or materials from Internet. If so, please cite these references in your report. If you have discussed with other people (e.g., professors, friends or classmates), please mention these discussions in the acknowledgement section of your report. In a team-based report (with ≤ 3 persons), please describe clearly the contribution of each team member.

Questions: If you have any question about this homework, please contact TA (Wang Fangpin, wfp15@mails.tsinghua.edu.cn).

Academic Honor Code: This homework must be done independently. We have zero tolerance on cheating and plagiarism.

1 Problem Description

In this problem, you are given 60 micrographs with center coordinates of particles (40 for training and 20 for testing). The coordinates of center of particles are manually picked by human experts. Your goal is to develop an algorithm pick the particles from micrographs. You should train your algorithm on the training dataset, test the performance of your algorithm on the test dataset and then analyze the results.

2 Getting started

2.1 Data

The data can be downloaded from <http://pan.baidu.com/s/1gdPddYV> with password: 75sj.

The protein complex of the dataset is called TRPV1. The micrographs are stored in .mrc files (for more details, see [https://en.wikipedia.org/wiki/MRC_\(file_format\)](https://en.wikipedia.org/wiki/MRC_(file_format)) and http://bio3d.colorado.edu/imod/doc/mrc_format.txt). The coordinates of particles in micrographs are stored in corresponding .star files, where the first two columns are coordinates of center of particles. You only need to use these two columns.

2.2 Helpful materials

[1] is a good reference to know about cryo-EM. [2] shows some terms to judge the your algorithm of particle picking. [3][4] are some particle picking algorithms, we encourage you to design a new algorithm or improve the previous algorithms instead of using the existing ones. The dataset used in our homework comes from [5].

In addition, Coursera has a good free online course on cryo-EM. Also, the following textbook is a a good introduction on single-particle data analysis (Library should have the electronic version): “Three-Dimensional Electron Microscopy of Macromolecular Assemblies”.

2.3 Implementation

Implement your algorithm in any programming language that you are familiar with, such as Java, C/C++, Matlab, Python, etc. You are allowed to call any other available public package in your program. If so, please include the library in your final submission.

2.4 Measure your algorithm

When you test your algorithm on the test dataset, you are required to output the coordinates of particles you picked as the same format as we gave. Here we define a distance threshold $l = 0.2 * particleSize$, where the particle size (*particleSize*) is 180 pixels. If the distance between a predicted particle coordinates and the expert picking coordinates is lower than the distance threshold (e.g., $0.2 * 180$ pixels), then we regard the prediction as a correct one. Let TP denote the number of particles that are picked by both your approach and human experts, FP denote the number of particles that are picked by your approach but not in the reference set, and FN denotes the number of particles that are only picked by human experts (i.e., not in the set of particles identified by your method). Then recall and precision scores are defined as $recall = TP / (FN + TP)$, and $precision = TP / (FP + TP)$. Compute and report the average precision and recall scores of your algorithm on test dataset. Other measurements like curves of precision and recall, f-measure can also be used to demonstrate the performance. Also you need to analyze your result. For example, you can plot the wrong prediction results and try to see why your algorithm fails.

3 Requirement of Report

In your final report, you should address the following points:

- (1) Details of your algorithm, such as overview, pseudo-code (or flow chart), etc.
- (2) Performance evaluation of your algorithm.
- (3) Discussion about strength and limitation of your algorithm.
- (4) Give the final recall and precision on the test data as well as the coordinate files.

4 Final Submission

For final submission, you need to provide: (1) report; (2) source code and binary executable file of your program, and a short readme file that describes how to compile and run your program.

References

- [1] Yifan Cheng, Nikolaus Grigorieff, Pawel A. Penczek, and Thomas Walz. A primer to single-particle cryo-electron microscopy. *Cell*, 161(3):438 – 449, 2015.
- [2] Robert Langlois and Joachim Frank. A clarification of the terms used in comparing semi-automated particle selection algorithms in cryo-em. *Journal of Structural Biology*, 175(3):348 – 352, 2011.
- [3] Robert Langlois, Jesper Pallesen, Jordan T. Ash, Danny Nam Ho, John L. Rubinstein, and Joachim Frank. Automated particle picking for low-contrast macromolecules in cryo-electron microscopy. *Journal of Structural Biology*, 186(1):1 – 7, 2014.
- [4] C.O.S. Sorzano, E. Recarte, M. Alcorlo, J.R. Bilbao-Castro, C. San-Martín, R. Marabini, and J.M. Carazo. Automatic particle selection from electron micrographs using machine learning techniques. *Journal of Structural Biology*, 167(3):252 – 260, 2009.
- [5] Maofu Liao, Erhu Cao, David Julius, and Yifan Cheng. Structure of the trpv1 ion channel determined by electron cryo-microscopy. *Nature*, 504(7478):107112, December 2013.