ALGORITHMICS - Bachelor of Software Engineering 2022

LAB GUIDE. SESSION 6

GOALS:

- **Backtracking: high resolution image reconstruction.**

1. High resolution image reconstruction for cryo-electron microscopy

By preserving molecular structure at atomic level, cryo-electron microscopy (cryo-EM) has solved the atomic structure of many proteins during last years including COVID19 virus spike, which has considerably accelerated the development of vaccines. Cryo-EM relies greatly on computing, in fact Joachim Frank (Germany 1940) was Nobel prize awarded in 2017 (shared with other two researchers) for developing the original image processing algorithms for cryo-EM.

An electron microscope generates high-resolution images from a purified sample with many copies of the same protein. Nevertheless, these images are corrupted by high levels of noise and distortions because they are acquired in cryo ("frozen atoms") conditions, the image must be radiated with a very low dose to keep intact protein structure at atomic level (Fig1.A). Firstly, the particles (projected protein shadows) are detected, green boxes in Fig1.A. The key step comes afterwards, particles with the same shape are grouped together an averaged to obtain a clean averaged single image (Fig1.B). Finally, these different projections of the same protein are aligned to recover its 3D shape (Fig 1.C) at near atomic resolution (Fig 2.D).

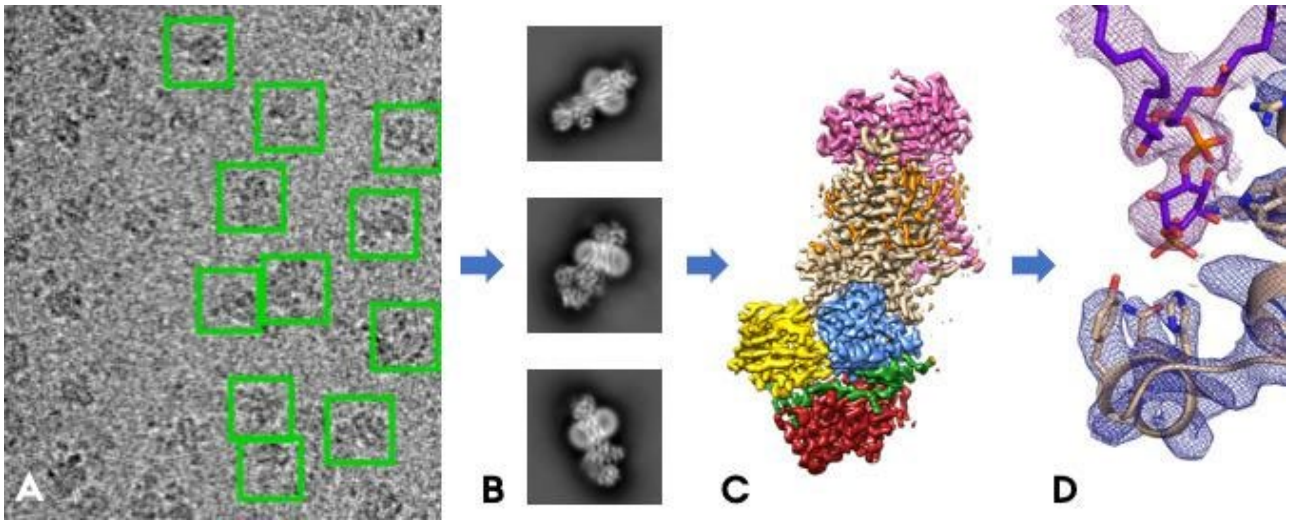

Figure 1. A) cryo-micrograph with some protein 2D projections selected. B) Clean images obtained after averaging three differently oriented projections. C) 3D reconstruction and D) zoom to its atomic model.

2. Image averaging

Here we deal with the image average problem necessary to get rid of noise, distortions, and badly selected particles to recover a clean version of the true embedded pattern.

This problem can be expressed in terms of combinatorial optimization. We must find the vector $X = \{x_1, x_2, \dots, x_n\}$ which generates the maximum *Zero Mean Cross-Correlation* (ZNCC). The values of an element x_k of X can only be: 0 image of an erroneous (bad) particle, 1 image for half subset 1, 2 image for half subset 2.

Similarity between two gray valued images, I_1 and I_2 with the same size, can be measured by computing Zero Mean Cross-Correlation (ZNCC):

$$ZNCC(I_1, I_2) = \frac{\sum_{i,j}(I_1(i,j) - \mu_1)(I_2(i,j) - \mu_2)}{N \cdot \sigma_1 \cdot \sigma_2}$$

Where N is the number pixels of an image, $I(i,j)$ is the gray value of the image pixel at coordinates i and j, μ is the image mean gray value and σ is the standard deviation.

Since we do not rely in any external prior reference image, we are going to divide the input dataset into two completely separated subsets, images labeled as 1 or 2 in X. The function to maximize is $ZNCC(\bar{I}_1, \bar{I}_2)$ where $\bar{I}_1$ and $\bar{I}_2$ are the averaged images of the subsets 1 and 2 respectively.

2.1. The input dataset

Two classes **Image.java** and **ImageAverager.java** are provided to generate synthetically the input image dataset. This dataset contains images containing the reference pattern (good) and its negative (bad), both corrupted by synthetic distortions:

- An image half region is deleted to simulate distortions, there are four possibilities: up, bottom, left and right.
- Additive Gaussian noise with certain standard deviation is added (try with value 5).

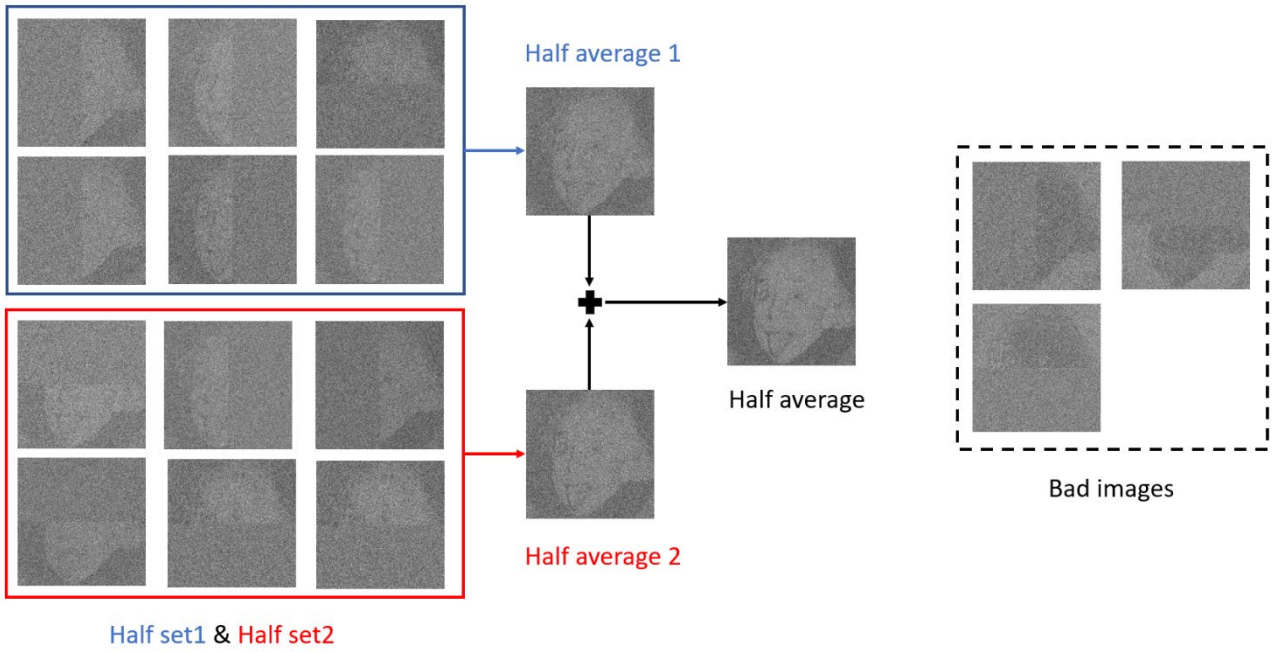

Figure 2. A solution of an image dataset of 15 images. Solution with 6 images in the subsets 1 and 2, 3 are identified as bad ones. It's noticeable that individual images contain partially the reference pattern, the ideal solution finds averages with the complete pattern and a reduced noise. In addition, the higher similarity between subsets 1 and 2 the higher ZNCC value.

2.2. Greedy approximation

Current implementations for solving this problem use greedy approximations. In essence random subsets of images are generated and their averaged images computed, after several tries the best result is kept.

2.3. Backtracking

Because greedy approximations do not guarantee to obtain the optimal solution. Here we propose to solve this task by using a backtracking approach.

Balancing condition: here we introduce this condition to reduce space search, the difference between half subsets in number of images must be lower or equal to 1. Intuitively, we can assume that a balanced partition will be close to the optimal because having a similar number of images ensures similar noise levels.

3. Work to be done

3.1. Coding

Implement the next code:

- The method for a greedy approach in **ImageAverager.java** class.
- The method for a backtracking approach in **ImageAverager.java** class.
- The method for a backtracking approach including balancing condition in **ImageAverager.java** class.

3.2. Measurements

Complete the next table considering **n** the number of dataset images (for greedy uses **n** as the number of tries too). Complete the main method in class **ImageAveragerBench.java**:

n	Time_BT	Time_BT_balancing	ZNCC_greedy	ZNNC_BT	ZNNC_BT_balancing
2					
3					
4					
5					
6					
…					
Until not tractable					

Since input dataset is generated by a random process, values for ZNCC may change from different executions.

3.3. Questions

Write a justified answer to the next questions:

a) State the algorithm that provides better results and explain why.
b) Which algorithm will you use for processing a realistic dataset a million of images? Explain why.
c) Determine the theoretical time complexity for backtracking (without balancing condition) and validate this analysis from the experimental results.

d) Determine the advantage of including the balancing condition in terms of time for backtracking, does it affect the quality of the results?

You should include in your Java project a new `algstudents.s6` package with the following content inside it:

- All the requested source files for that package.
- The requested PDF document called `session6.pdf` with the corresponding activities.

This practice takes two sessions (or weeks), the requested files must be delivered one day before the next topic practice.