2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635352

FAST GLOBAL IMAGE ALIGNMENT ALGORITHM FOR CRYOEM THROUGH VECTOR COMPRESSION

O.L. Zarrabeitia, E. Ramírez-Aportela, J.M. Carazo, C.O.S. Sorzano

Biocomputing Unit, Centro Nacional de Biotecnología (CNB-CSIC)

ABSTRACT

In cryo-electron microscopy (Cryo-EM), single-particle analysis involves aligning two-dimensional images of individual protein molecules. These particles are flash-frozen and imaged in various orientations. The goal of image alignment is to computationally determine the relative orientations and positions of the particles in each image. It is possible to reconstruct a high-resolution three-dimensional molecule model by doing so. This process is critical for understanding molecular structure and function. This study introduces a novel fast alignment algorithm designed explicitly for Cryo-Electron Microscopy (CryoEM) images, leveraging the efficiency of vector compression techniques such as PCA and IVF-PQ to accelerate the alignment process. Experimental evaluations on diverse CryoEM datasets demonstrate the algorithm's efficacy in improving alignment speed with a slight compromise in accuracy. Thus, we state that this algorithm is particularly interesting during the preliminary stages of 3D refinement, where computational efficiency is paramount, and the observed slight accuracy degradation is deemed acceptable.

***Index Terms*—** CryoEM, Image alignment, Vector compression, Fast image search

1. INTRODUCTION

Cryo-electron microscopy (Cryo-EM) is a cutting-edge imaging technique that has revolutionized structural biology, enabling scientists to visualize biological macromolecules at almost atomic resolution. Unlike traditional electron microscopy methods, Cryo-EM involves flash-freezing samples in vitreous ice, preserving biological specimens in their native state. The frozen samples are then captured using a Transmission Electron Microscope (TEM), and advanced computational algorithms are employed to elucidate the three-dimensional structure of the sample[1].

Image alignment plays a crucial role in CryoEM image processing, as it is used in many processes such as 2D classification, angular assignment, and 3D classification[2]. At the same time, it is computationally very expensive, requiring millions of image comparisons. Thus, a significant amount of time in CryoEM image processing is devoted to image alignment.

In the process of image alignment, experimental images are searched across a large set of reference images from which the parameters to be estimated are known. This involves considering all potential in-plane transformations. In this way, the unknown parameters of the experimental image can be inherited from the most similar reference image[3].

This paper presents a fast and reliable approach to global image alignment, which uses vector compression techniques to accelerate image comparisons. This allows us to obtain a very fast estimation of the particle's pose.

2. METHODS

Our study addressed image alignment as a nearest-neighbor problem, simplifying global 3D alignment to sampling projection parameters, including Euler angles and 2D in-plane shifts. We aimed to minimize a distance function over these sampled points. While this method seems straightforward, it is challenged by the exponential increase in sample points due to the high number of varied parameters.

We used a reduced set of Fourier coefficients to address this, focusing on those below 8Å resolution, as they carry the most alignment-critical information[4]. This approach lessens the computational load. Additionally, we applied a Wiener filter to correct for the Contrast Transfer Function (CTF) in experimental images and grouped similar CTFs for efficiency.

Our process for generating data is hierarchical, avoiding redundant computations. We first project the reference volume at various angles, apply in-plane rotations, compute the Fourier transform to extract low-frequency components, and finally apply in-plane shifts.

We have designed two different approaches to solve this image alignment problem: a Principal Component Analysis (PCA) and a vector database for efficient storage and retrieval to manage and compare many reference vectors. This dual strategy allows us to efficiently align each experimental image with its closest reference counterpart.

2.1. PCA-based approach

PCA is an essential tool for simplifying complex data. In this work, we have applied PCA to reduce the complexity

of vectors representing the frequencies in cryo-EM images. PCA identifies key combinations of frequencies in these vectors, enabling us to retain the most essential components and compress information without losing crucial details. This approach is fundamental for optimizing image storage and processing while maintaining their essential structure.

Our PCA methodology consists of two phases. In the first stage, we trained the model using data augmentation on reference images, enriching the diversity and quantity of data for a more comprehensive representation. In the second stage, we employed the principal components learned during training to represent both experimental and reference images. This strategy allows us to leverage the information obtained from reference images (PCA base) for a better compact representation of the experimental images.

Finally, the similarity between the compressed vectors of experimental and reference images is evaluated by calculating the Euclidean distance between each pair of vectors. This measure allows for comparing the structural proximity between all pairs of images.

2.2. Vector-database approach

FAISS is a library for efficient similarity search of dense vectors[5]. Using the previously mentioned vector representation of reference images, we can use this library to accelerate image searches and thus increase alignment performance. Indeed, FAISS is very efficient when dealing with large sets of high-dimensional vectors, which is the case of our problem, as it implements advanced vector compression and quantization techniques.

Several such techniques were explored in this project, but finally, Inverted File (IVF) and Product Quantisation (PQ) were selected. The former technique is designed to heuristically reduce the search space for a given query vector. The latter allows compression and quantification of vectors while providing fast distance estimates among a given pair of encoded vectors.

IVF uses k-means to segment the search space into Voronoi cells around cluster centroids. When querying a vector, this will first be searched across centroids. In this way, the exhaustive search can be delimited to a handful of cells.

PQ uses the residual vectors from the previous algorithm: this is the difference vector between the sample and its closest centroid. This increases the entropy so that quantization is more effective. Then, these vectors are divided into fixed-size chunks, each containing a handful of components. For each of these chunks, space is once again divided using k-means. This allows quantizing chunks to their closest centroid and encoding them with its index. Thus, a vector can be encoded by concatenating its quantized chunk indices.

One of the largest benefits of PQ is that vectors can be encoded using a few bytes, allowing the storage of millions of vectors in memory. Usually, each chunk is encoded with 1 byte (k=256). Therefore, if a vector is divided into 64 chunks, only 64 bytes are needed to encode the vector.

Moreover, this technique allows quickly computing the distance between a pair of encoded vectors. To do so, pairwise distance tables are computed for the centroids of each chunk. Then, it is a matter of looking up and accumulating the distances from those tables.

3. RESULTS AND DISCUSSION

3.1. PCA based approach

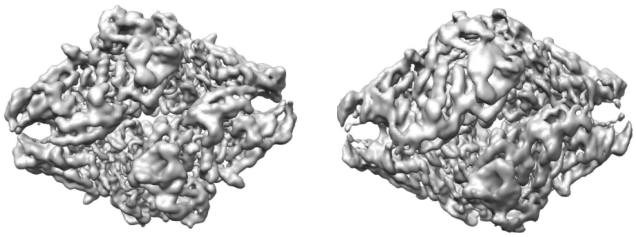
The proposed approach for particle alignment is defined as a global method. However, to achieve high resolution, it is crucial to complement our method with strategies that facilitate local search and evaluation. In this regard, we have combined our method with Relion's local search, considering the alignment previously obtained by our method. This way, the computation time can be reduced by utilizing vector compression with the PCA technique in the global search without compromising the alignment accuracy.

A first test was conducted using the Beta-galactosidase dataset (EMPIAR-10061). Figure 1 shows a comparison of the reconstructions obtained using our global method (Align-PCA, Fig. 1.a) and Relion (showing iteration 10, which is the last iteration before entering local search, Fig. 1.b). Subsequently, the angular assignment of particles determined by Align-PCA served as the starting point for local refinement in Relion. The resolution obtained by combining align-PCA and Relion in local mode was 3.0 Å, similar to that obtained using full-Relion. However, the execution times were significantly lower in the combined approach compared to full-Relion (Table 1), reducing from 3.5 hours to approximately 1 hour.

In a second test, the dataset corresponding to the TRPV5 structure (EMPIAR-10254) was used. Similar to the previous test, the effectiveness of AlignPCA as a global alignment method and initial point in local refinement through Relion was evaluated. Two reconstructions were carried out in parallel: one using Relion in its original form (full-Relion) and the other using the combination of AlignPCA and local Relion. In both cases, a similar resolution of 3.4 Å was achieved; however, a significant reduction in processing time was observed, with a decrease of over 70 % (Table 1). The time taken decreased from over 7 hours using full-Relion to approximately 2 hours with the combined methods.

3.2. Vector-database approach

At the time of writing, we have only compared this method against a normal exhaustive search. Nevertheless, these experiments give an insight into the performance gain, as most of the state-of-the-art alignment methods perform an exhaus-



(a) AlignPCA (b) Relion

Fig. 1. Reconstruction of the Beta-galactosidase dataset obtained from EMPIAR-10061. In (a), the reconstruction using the global AlignPCA method is shown, while in (b), Relion's reconstruction is shown. In the case of Relion, the reconstruction obtained after ten iterations corresponding to global alignment is depicted."

ID	Full [1] Relion	AlignPCA [2]	Local [3] Relion	AlignPCA + Local Relion[4]
10061	3h:30min	10min	48min	$\approx$ 1h
10254	7h:22min	43min	1h18min	$\approx$ 2h

Table 1. Execution times during the reconstruction of datasets obtained from EMPIAR 10061 and EMPIAR-10054. (1) It refers to the times obtained using Relion in its original form, including global and local searches. (2) Times obtained using the global method presented in this work. (3) Times obtained during local refinement with Relion, starting from the alignment obtained with alignPCA. (4) Total times for the combination of the presented method and local Relion.

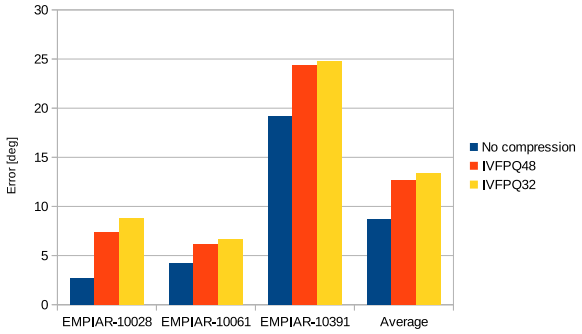
tive search. We have also measured the accuracy degradation induced by the vector compression.

We conducted an assessment of our algorithm by testing it across multiple publicly available datasets, specifically EMPIAR-10028[6], EMPIAR-10061[7], and EMPIAR-10391[8]. The experiments were executed on a workstation featuring dual Intel Xeon E5-X5647 processors and an NVIDIA Titan X GPU.

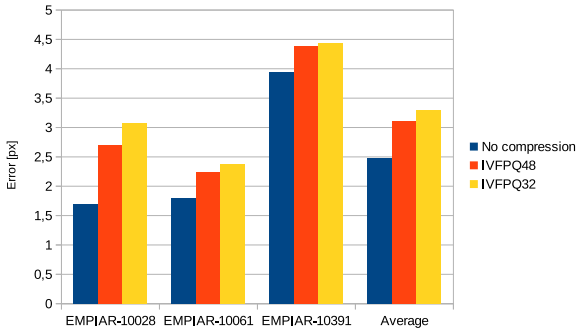
We employed a consensus derived from two Relion refinements to establish a reliable ground truth. It is crucial to note that the alignment errors presented here carry an inherent uncertainty associated with Relion's estimation, even after consensus.

As a basis, an alignment with no vector compression was carried out. Then, two additional alignments with varying vector compression ratios were assessed. They all used an angular sampling rate of 7.5 degrees and 2-pixel steps for shift exploration.

As shown in Figures 2 and 3, vector compression produces a discernible degradation in both angular and shift assignments. Accordingly, this translates into a slight resolution decrease. These degradation values vary across datasets, so we have averaged them for reference.



(a) Angular assignment error



(b) Shift assignment error

Fig. 2. Alignment errors with FAISS-based algorithm

Regarding the performance, Figure 4 shows alignment time regarding the particle count, also considering the database training and population time when applicable. Note that the axes of the graph are in logarithmic scale. Variations across datasets were insignificant, so only the average time was plotted.

The presented empirical outcomes affirm that this algorithm provides an extremely fast angular assignment, albeit with a discernible trade-off in accuracy. Indeed, for reasonably sized CryoEM datasets (more than 60,000 particles), a 10 times speedup is achieved.

Thus, we state that the vector database algorithm is suitable for the initial iterations of a refinement cycle, where angular assignment accuracy is not a priority. In such cases, it enables a considerable speed-up of the alignment process to resolve protein structures faster.

4. CONCLUSIONS

The study presented two algorithms designed to accelerate image alignment in Single Particle Analysis by CryoEM, a crucial step in determining molecular structures. By employing advanced vector compression techniques such as PCA and IVF-PQ, the algorithm efficiently speeds up the alignment process during the vital preliminary stages of 3D refinement.

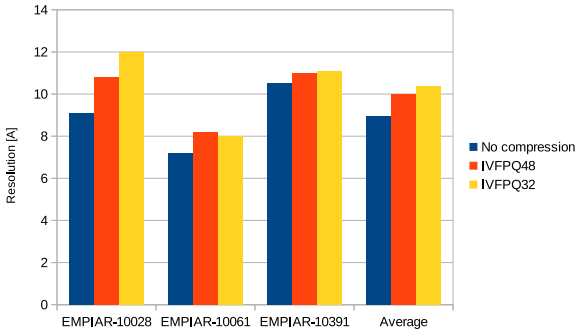


Fig. 3. Reconstruction resolution

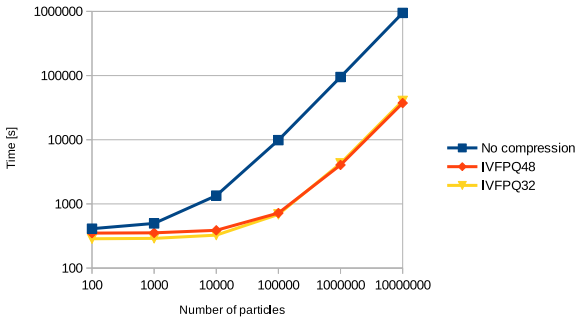


Fig. 4. Alignment performance with FAISS-based algorithm

While the focus is on computational speed, a slight reduction in accuracy is considered acceptable for these initial stages.

The research demonstrated the efficacy of this approach through tests on various datasets. The PCA-based method significantly reduced processing times without compromising the resolution of the reconstructed models. On the other hand, the vector-database approach, while faster, did introduce a noticeable decrease in alignment accuracy. This method proved particularly beneficial for initial refinement cycles, enabling researchers to quickly approximate structures, a crucial step in structural biology.

Overall, these innovative methods mark a significant advancement in CryoEM, promising to streamline the lengthy and computationally intensive process of molecular imaging, thus accelerating scientific discoveries in this field.

5. REFERENCES

[1] Robert M Glaeser, Eva Nogales, and Wah Chiu, Eds., *Single-particle Cryo-EM of Biological Macromolecules*, 2053-2563. IOP Publishing, 2021.

[2] Yu-Xuan Chen, Rui Xie, Yang Yang, Lin He, Dagan Feng, and Hong-Bin Shen, "Fast cryo-em image alignment algorithm using power spectrum features," *Journal of Chemical Information and Modeling*, vol. 61, no. 9, pp. 4795–4806, 2021, PMID: 34523929.

[3] Eva Nogales and Sjors H.W. Scheres, "Cryo-em: A unique tool for the visualization of macromolecular complexity," *Molecular Cell*, vol. 58, no. 4, pp. 677–689, 2015.

[4] Sjors H.W. Scheres, "Relion: Implementation of a bayesian approach to cryo-em structure determination," *Journal of Structural Biology*, vol. 180, no. 3, pp. 519–530, 2012.

[5] Jeff Johnson, Matthijs Douze, and Hervé Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[6] Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors HW Scheres, "Cryo-em structure of the *Plasmodium falciparum* 80s ribosome bound to the anti-protozoan drug emetine," *eLife*, vol. 3, pp. e03080, jun 2014.

[7] Shaoxia Chen, Greg McMullan, Abdul R. Faruqi, Garib N. Murshudov, Judith M. Short, Sjors H.W. Scheres, and Richard Henderson, "High-resolution noise substitution to measure overfitting and validate resolution in 3d structure determination by single particle electron cryomicroscopy," *Ultramicroscopy*, vol. 135, pp. 24–35, 2013.

[8] Yong Zi Tan, Lei Zhang, José Rodrigues, Ruixiang Blake Zheng, Sabrina I. Giacometti, Ana L. Rosário, Brian Kloss, Venkata P. Dandey, Hui Wei, Richard Brunton, Ashleigh M. Raczkowski, Diogo Athayde, Maria João Catalão, Madalena Pimentel, Oliver B. Clarke, Todd L. Lowary, Margarida Archer, Michael Niederweis, Clinton S. Potter, Bridget Carragher, and Filippo Mancia, "Cryo-em structures and regulation of arabinofuranosyltransferase aftd from mycobacteria," *Molecular Cell*, vol. 78, no. 4, pp. 683–699.e11, 2020.