Single particle analysis integrated with microscopy: A high-throughput approach for reconstructing icosahedral particles

Xiaodong Yan [a,1], Giovanni Cardone [a,1], Xing Zhang [c], Z. Hong Zhou [c], Timothy S. Baker [a,b,*]

[a] Department of Chemistry & Biochemistry, University of California, San Diego, La Jolla, CA 92093-0378, United States
[b] Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093-0378, United States
[c] Department of Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095-7364, United States

ARTICLE INFO

ABSTRACT

In cryo-electron microscopy and single particle analysis, data acquisition and image processing are generally carried out in sequential steps and computation of a three-dimensional reconstruction only begins once all the micrographs have been acquired. We are developing an integrated system for processing images of icosahedral particles during microscopy to provide reconstructed density maps in real-time at the highest possible resolution. The system is designed as a combination of pipelines to run in parallel on a computer cluster and analyzes micrographs as they are acquired, handling automatically all the processing steps from defocus estimation and particle picking to origin/orientation determination. An *ab initio* model is determined independently from the first micrographs collected, and new models are generated as more particles become available. As a proof of concept, we simulated data acquisition sessions using three sets of micrographs of good to excellent quality that were previously recorded from different icosahedral viruses. Results show that the processing of single micrographs can keep pace with an acquisition rate of about two images per minute. The reconstructed density map improves steadily during the image acquisition phase and its quality at the end of data collection is only moderately inferior to that obtained by expert users who processed semi-automatically all the micrographs after the acquisition. The current prototype demonstrates the advantages of integrating three-dimensional image processing with microscopy, which include an ability to monitor acquisition in terms of the final structure and to predict how much data and microscope resources are needed to achieve a desired resolution.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Cryo-electron microscopy (cryo-EM), in combination with three-dimensional (3D) image processing techniques, is progressively becoming a reliable and efficient tool for determining the structures of biological macromolecular complexes. This is largely attributed to steady advancements in instrumentation and software (Chang et al., 2012; Glaeser and Hall, 2011; Grigorieff and Harrison, 2011; Zhou, 2011). The most recent technological breakthrough is the development of direct detection device (DDD) cameras whose performance at least matches and can even exceed that of traditional photographic film (Bammes et al., 2012; Campbell et al., 2012), thus surpassing the need for manual and lengthy digitization procedures. Furthermore, several software systems are now available to collect data automatically at the microscope such that thousands of images can now be acquired with minimal user intervention during a continuous session of up to several days (Korinek et al., 2011; Lei and Frank, 2005; Shi et al., 2008; Suloway et al., 2005). DDDs and automated acquisition software naturally combine to produce high-throughput systems that facilitate obtaining large numbers of particle images and therefore targeting high-resolution structures by single particle analysis. With these advances, the current limitation for analyzing such a potential plethora of data in a high-throughput manner lies with the determination of a 3D structure from the acquired images. This process involves several procedural steps, which include screening acquired images by their quality, estimating the defocus imposed at the microscope, locating and extracting particles from the images, determining their orientation and origin with respect to a common reference system, and calculating a 3D reconstruction from an optimal subset of particles. All these processing steps have been extensively analyzed and different methods have been

* Corresponding author at: Department of Chemistry & Biochemistry and Division of Biological Sciences, University of California, San Diego, 9500 Gilman Drive, MC-0378, La Jolla, CA 92093-0378, United States. Fax: +1 (858) 534 5846.
  E-mail address: tsb@ucsd.edu (T.S. Baker).
[1] Authors contributed equally.

proposed to automate some of them (e.g. see (Lyumkis et al., 2010) for a review).

Recent work has shown that high-resolution reconstructions can be obtained rapidly by combining all the necessary steps in an efficient way into a single pipeline (Sorzano et al., 2013). However, some level of user intervention is still required to adapt some of the individual computational procedures to the specific properties of each data set. Such intervention thwarts the implementation of an integrated approach that streamlines data analysis, and this generally forces most researchers to wait and only process the images after all microscopy is done. Consequently, 3D reconstructions are usually obtained only in days or weeks after acquisition is finished, with the total time required depending on the size of the data set and the computational resources available. This delay in the outcome of the experiments represents a critical bottleneck not only for high-throughput analysis, but also for validation of the experiments themselves.

Here, we explore the concept of integrating 3D image reconstruction with microscopy as a means to overcome these limitations. Specifically, we study the feasibility of a real-time, high-throughput, automated processing system that analyzes electron micrographs of icosahedral particles as soon as they are acquired at the microscope and integrates the results into a 3D reconstruction. The task of such a system is designed to provide the microscopist during a data collection session an electron density map that is regularly updated as more micrographs are recorded. For the purpose of this analysis, we have implemented a software prototype that gathers the processing steps that are fundamental to single particle analysis into a combined group of unsupervised processing pipelines running on multiple processors. The approach is similar to what we have recently proposed to generate a 3D reconstruction from particles located in a single electron micrograph. In that case, the goal was to obtain low resolution, 3D structural information from one micrograph and perform an initial evaluation of the sample. On the contrary, the current system we describe below aims at determining a final, single structure at high-resolution by integrating the information from all currently available micrographs. We present a strategy that we have implemented, and the algorithms selected to perform each step, to process in real-time images acquired at the microscope. We also report results based on simulated tests with three experimental data sets to demonstrate the potential of such an approach. Finally, we discuss the advantages of integrating image analysis with microscopy and the limitations of the current implementation.

2. Implementation

Our goal in this study was to verify how feasible it is to generate 3D reconstructions during a data acquisition campaign at the microscope, using the micrographs as soon as they are recorded. For this purpose, the software package Auto3DEM (Yan et al., 2007a) provides most of the basic functionalities needed, but without them being fully automated. For example initial preprocessing of the micrographs (CTF estimation, particle picking, and extraction) is performed manually. Also, the *ab initio* determination of the initial model and the iterative procedure for alignment and 3D reconstruction are only partially automated and therefore require decisions from the user at different steps of the analysis. Therefore, we have currently developed a prototype software package written in Perl language that monitors the status of the acquisition and oversees the entire processing scheme, which is mostly performed by components of Auto3DEM. The program, though still at the prototype stage of development, implements the basic functionalities of a workflow manager, adapted to our needs. Indeed, it monitors the availability of new input images, schedules tasks for

each computing step according to the resources available, verifies correct execution of tasks, and includes feedback control loops to adjust the input parameters of the computational step on the basis of how the reconstruction results progress. A detailed description of the workflow is provided in Supplemental Material and Fig. S1. Most of the tasks are distributed among multiple processors and exploit the MPI parallelism paradigm adopted by Auto3DEM, with the exception of the Contrast Transfer Function (CTF) estimation, which is performed by the CTFFIND3 (Mindell and Grigorieff, 2003) program that uses OpenMP. All the tasks running at any time during the processing are launched on a computer cluster through the PBS/TORQUE resource manager (Staples, 2006).

3. Experimental data sets

The approach we describe here was tested on three available, minimal-dose cryo-EM data sets that were previously processed in a semi-automatic manner by an expert user, and the resulting reconstructions were used for comparison. One data set consisted of 258 micrographs of grass carp reovirus (GCRV) (Zhang et al., 2010), as originally selected from the entire set according to the quality of their Fourier transforms. This data set was mostly included to evaluate the importance of rigorous micrograph screening. The cryo-images were recorded on Kodak SO-163 film in an FEI Titan Krios electron microscope operating at 300 kV and later digitized with a Nikon Coolscan 9000ED microdensitometer to a nominal pixel size of 1.075 Å. For the purpose of comparing the results from the current approach with those previously published, we used the calibrated pixel size of 1.1 Å to calculate the final resolution. A second data set included 4276 images of bacteriophage P22 (Lander et al., 2006; Tang et al., 2011) recorded on a Tietz 4kx4k CCD camera at a nominal pixel size of 1 Å in an FEI F20 Tecnai electron microscope operating at 200 kV, using the Leginon automated data collection system (Suloway et al., 2005). Additional details concerning these two data sets are available in the cited literature. A third, unpublished data set consisted of 2020 cryo-EM images of phage CUS-3 recorded on a Direct Electron DE12 4kx3k DDD camera at a nominal pixel size of 1.359 Å in an FEI Polara Tecnai electron microscope operating at 200 kV, using Leginon. Each DDD micrograph was obtained by discarding the first two recorded frames and combining the next fifteen. A drift correction algorithm was applied to the individual frames to align and then sum them together (Shigematsu and Sigworth, 2013). The numbers of particles extracted, either manually or automatically, from the micrographs in all three data sets are reported in Table 1. The two reconstructions obtained from each data set, one by a manual and the other by the automated approach, were compared by visual inspection and by calculation of the Fourier Shell Correlation, using the 0.5 cutoff criterion (van Heel and Schatz, 2005).

4. Test procedure

The acquisition of images at the microscope was simulated by regularly transferring single micrographs in a data set to a defined

Table 1
Acquisition results: particles extracted.

	No. of micrographs	No. of particles	
		High-throughput	Manual[a]
P22	4276	39,345	21,645 (2888)
CUS-3	2020	40,858	7766 (419)
GCRV	258	15,778	20,473 (247)

[a] The particles were extracted from a subset of the micrographs, which is indicated in parentheses.

file directory. In order to mimic the delay between successive image acquisition events, the interval time was fixed to 30 s for the P22 and CUS-3 data sets. Since the GCRV data set consisted of digitized micrographs with a linear size exceeding 8000 pixels, in this test the interval was set to 2 min. In all three instances, the simulated acquisition intervals represent realistic, albeit stringent, conditions of automated data collection on a well-aligned, stable microscope. All the simulated tests were performed on a dedicated Linux cluster composed of one front node and five compute nodes. Each node was equipped with two, 4-core, 8-thread, 2.67 GHz, Intel Xeon X5550 processors, and between 36 and 48 GB of memory. The Auto3DEM executables were generated with the compiler GCC 4.4.6 and relied on the openMPI library for parallelism. The number of processors used at each time of the processing depended on the amount of tasks being performed simultaneously, but by design it never exceeded 53 processors.

5. Results

5.1. A data-driven pipeline approach

In single particle analysis, a 3D reconstruction is the result of a sequence of processing steps that is applied to all images acquired at the microscope. The design of a high-throughput system integrated with the microscopy requires that these same steps be applied to each image independently, without any user intervention. Additionally, the system has to keep up with the micrograph acquisition rate at the microscope to avoid any delay or accumulation of unprocessed data. As a further complication, the origin and orientation parameters of each particle are commonly determined by using projection matching algorithms, which work in an iterative fashion since the accuracy of the alignments depend on the quality of the reference map, and *vice versa*. This aspect becomes even more critical in a high-throughput system because, initially the reference map is generated from the particles available in the few micrographs acquired thus far, and it is expected to improve rapidly as more images are acquired. As a simplified but optimal approach to satisfy these requirements, we implemented the system as a set of computational, asynchronous pipelines, *i.e.* chains of processing tasks arranged so that the output of each task is the input of the next (Fig. 1).

In a pipeline, the results of each task (either images or parameter files) are stored in a separate directory and a master program monitors these directories cyclically. As soon as new input data are available and the processing of the previous data is completed, the master program assigns a new job to the task. The use of pipelines permits optimization of the computational resources assigned to each task and allows all of them to be executed in parallel. This assumes that enough resources are available, with a minimum being one processor per task. As an additional advantage, this approach simplifies the synchronization between the sequential tasks to be performed, since the directories implicitly provide a buffer system. The entire processing is divided into two concatenated and independent pipelines, each one complying with different requirements. Both pipelines run in an unsupervised manner with the settings for each task being determined adaptively from the data. Before starting the system, the user needs to input only a few parameters: the microscope settings (voltage, spherical aberration coefficient of the objective lens, pixel size, and amplitude contrast), the diameter of the particle, and the name of the directory into which the acquisition system will store the micrograph data.

A first pipeline (Pipeline 1) reads each micrograph as soon as it is acquired and ultimately outputs a stack of particle images and a parameter file. The parameter file contains a specification of the defocus level, origin, and orientation of each particle, as determined by projection matching against the best reconstruction available at that time. A similar pipeline is executed one time at the beginning, on the very first micrographs acquired, to determine an initial template by the Random Model Computation method (Yan et al., 2007b). In order to speed up the computation without losing too much accuracy in the alignment, all the processing in Pipeline 1 is performed on the images after down-sampling them by a factor of two. A second, iterative pipeline (Pipeline 2) reads all the particle images output from the first pipeline, generates an updated reconstruction, and refines the orientation of each particle image against the new density map. Using this approach, we separate the slowest and most variable part of the processing (Pipeline 2) from the more constant processing of single micrographs and initial alignment of the particles extracted from that micrograph (Pipeline 1). In Pipeline 1, the latency (time required to obtain a set of aligned particles from a micrograph) is given by the sum of the time required for each task, whereas the throughput
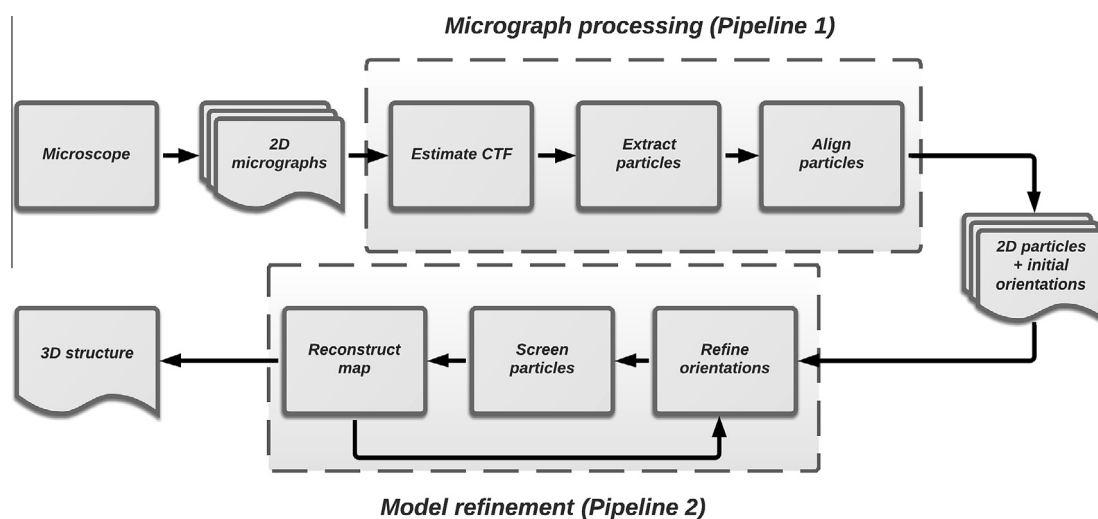
Fig.1. Data-driven, pipeline approach for real-time, high-resolution 3D reconstruction. The workflow highlights the two pipelines that the approach includes, and the major computational steps in each. Pipeline 1 processes micrographs separately, immediately after each is recorded, to compile a stack of particle images and determine their origin/orientation parameters. Pipeline 2, running in parallel with Pipeline 1, continuously updates the reconstruction and refines the origin and orientation parameters for all particles currently available, in an iterative manner.

rate (number of micrographs processed per time unit) is given by the slowest task in the pipeline. This means that, as long as the average processing time for the slowest task is comparable to or less than the average time between two acquisitions, the system can guarantee a throughput in line with the acquisition rate, thus avoiding any accumulation of micrographs. The results on experimental data show (see Section 5.4) that, in practical applications, assigning moderate computing resources to each task can easily satisfy this constraint.

In Pipeline 1, the basic tasks, each consisting of one or more processing steps, are: CTF determination; particle picking and extraction; normalization of particle images; centering; global alignment; and origin/orientation refinement (hereafter simply referred to as "refinement"). All these tasks are applied sequentially to each single micrograph and most are common to any conventional single particle analysis workflow. The task of centering, which refers to refining the location of the center of each particle in its box window before determining the orientation of the particle, is here required because the automatic picking procedure (Boier Martin et al., 1997) does not provide 1-pixel accuracy. The centering operation is implemented by performing a global search of the origin/orientation parameters by projection matching, and using the determined origin shift values to adjust the boxing coordinates and to re-extract the particles. Global alignment and refinement refer to the determination of the alignment parameters on coarse and local parameter grids, respectively, where the local grid is centered around the solution determined on the global one. All parameters are adaptively determined based on the quality (*i.e.* estimated resolution limit) of the reference map and the size of the particle (Cardone et al., 2013; Yan et al., 2007a).

Pipeline 2 applies to all the particles available at any time and is composed of just two tasks: map reconstruction and origin/orientation refinement. This pipeline is applied in a cyclic manner because the entire procedure is iterative. In this respect, it is equivalent to any iterative, projection-matching-based approach, with the main differences being that all settings are automatically determined and, at every iteration, the number of input particles increases with the acquisition of new images. In the current implementation, the computational resources are assigned statically to each task of the two pipelines, with most of them running on multiple processors. In Pipeline 1, the numbers of CPUs assigned specifically for each task are as follows: CTF determination, 2; particle picking and extraction, and normalization of particle images, 1; centering, 10; global alignment, 8; refinement, 16. In Pipeline 2, 16 CPUs are assigned to the map computation and particle origin/orientation refinement tasks.

5.2. Tests on experimental data sets

We tested our processing approach on experimental data sets to verify its ability to provide high resolution reconstructions in an automated and high-throughput manner. For this proof of concept, we used data that had been previously processed in a semi-automatic manner by expert users who obtained sub-nanometer resolution density maps. Specifically, we selected three sets of micrographs because of their good to excellent quality and because they were recorded on different media. These included images of unstained, vitrified samples of GCRV (Zhang et al., 2010) and bacteriophages P22 (Tang et al., 2011) and CUS-3 (Parent et al., unpublished). For tests with GCRV, we limited the analysis to only those micrographs actually used for the published reconstruction (258 out of 650). We included this reduced data set for the specific purpose of testing the high-throughput approach under ideal conditions of uniform, high quality images. For all tests, we simulated that the microscope acquired images at regular time intervals by copying the micrographs to a pre-defined directory at fixed delays.

We performed the tests under "stress" conditions by setting the interval time to a nominal value of 30 s, which can be challenging to achieve with most current microscopes and data acquisition systems. Since the images for GCRV were recorded on photographic film and thus covered larger field of views than would be obtained from all currently available digital cameras, for this specific data set we set the simulated acquisition interval to 2 min. From these simulations, we were able to gather information about the capabilities and the limitations of the proposed high-throughput processing system. The reconstructions obtained automatically at the end of the simulated acquisition job were compared with those obtained manually by expert users to verify the existing gap between manual, post-acquisition processing and automated analysis during data acquisition.

5.3. Generation of initial model

At the beginning of a data acquisition job, no reference model is available for determining the origin shifts and orientation angles of the particles in the micrographs. Therefore, the first task of the system is to generate an *ab initio* model from the initial micrographs acquired, as soon as there are enough particles to successfully employ the Random Model Computation method (Yan et al., 2007b). This step introduces an initial delay that results in an accumulation of micrographs that require processing. In the simulations with the P22, CUS-3, and GCRV data sets, there were delays of 34, 15, and 52 min, respectively (corresponding to 68, 30 and 26 micrographs). Once an *ab initio* model was available, all the pending micrographs were processed together via Pipeline 1, which took 4 (P22), 2 (CUS-3), and 20 (GCRV) min to finish. After that, all subsequent micrographs were treated independently.

5.4. High-throughput processing of single micrographs

One important requirement for a high-throughput, 3D reconstruction system is to be able to process micrographs in a time frame comparable to or faster than the acquisition rate. In our tests, we determined that downsampling the input micrographs by a factor of two and using 37 processors distributed among the different processing steps enabled us to obtain a stack of particle images from a micrograph along with a good estimate of the particle parameters in a time only slightly longer than the interval between successive acquisitions (<1 min for the P22 and CUS-3 data and <3 min for the GCRV data). This timing represents pipeline latency, *i.e.* an initial delay in the output after the very first micrograph is recorded, but it does not become a limiting factor in the throughput capabilities of the pipeline. In fact, it is more important that all processing steps in a pipeline perform their tasks before the next image is recorded, as this avoids accumulation of unprocessed images and a progressive increase in the original delay.

The first two processing steps in Pipeline 1, CTF determination and particle picking, generally run in a time span that is independent of the data set and, on average, take 10 and 1 s, respectively. Timings for the other three processing steps in Pipeline 1 (centering, global alignment, and refinement) are more variable (Fig. 2). We note, however, that the mean throughput rate is faster than the acquisition rate. In fact, the processing time per micrograph of any task was always shorter than the time interval between successive acquisitions except for a small percentage of the P22 micrographs. For this data set, the centering step for 165 of the 4276 micrographs took more than 30 s, with a maximum of 46 s, and the refinement of just five of the 4276 stacks of particles required more than 30 s and a maximum of 37 s. Regardless, the mean processing time was less than 20 s for each of these steps, and the entire data processing scheme quickly recovered from any delays caused by recalcitrant data. Variability in the processing times
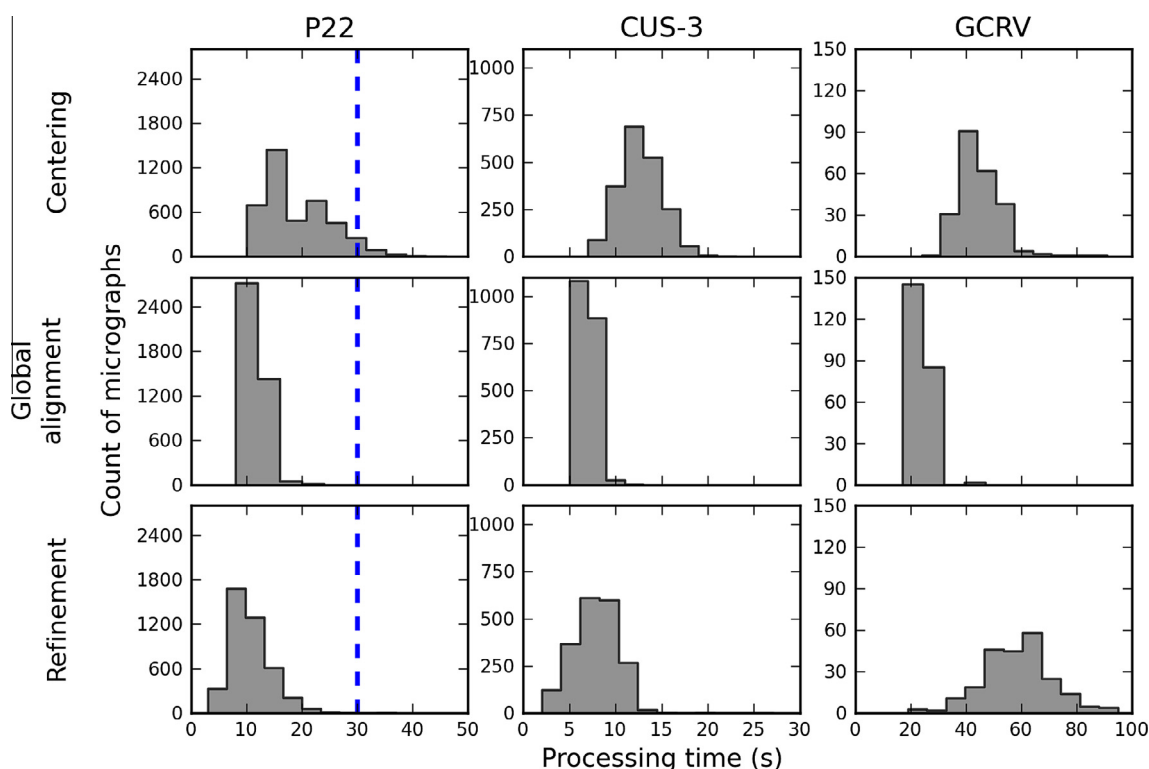
Fig.2. Processing times of single micrographs. Histograms plot the times required to perform specific computational steps in Pipeline 1 during the simulated acquisitions for each data set analyzed. Times are shown for the operations of centering (refinement of origin coordinates by projection matching and re-boxing of particles), global alignment (search of orientation parameters on a coarse grid), and refinement (search of parameters centered about the values determined by the global alignment). For the P22 tests, the vertical dashed line (blue) highlights the simulated acquisition time for these data (30 s), that is the time between an acquisition and the following one. For the other data sets the acquisition times fall at the limit of (30 s for CUS-3) or beyond (120 s for GCRV) the range displayed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for different micrographs is partially caused by the number of particles extracted from each micrograph, and this is more pronounced for the slower centering and refinement tasks and for larger particles (GCRV, unpublished data). It is notable that the centering and global alignment steps both perform coarse alignments of the particles using the same projection matching procedure based on the Polar Fourier Transform algorithm (Baker and Cheng, 1996). However, centering also involves re-boxing the particle images from the micrograph, and this operation contributes to making this step slower than the global alignment. Regarding global alignment, this step runs quickly for all the data sets, indicating that fewer computational resources can be assigned to this processing step without affecting overall performance.

5.5. Update of the 3D reconstruction

Since particles are extracted and their alignment parameters are determined while micrographs are being acquired, the system can generate a 3D reconstruction that is updated as new data become available. The reconstruction computed from all particle images available at any given moment is performed in Pipeline 2, which also contains a step in which the orientation parameters are refined provided the resolution of the map keeps improving. A new Pipeline 2 processing cycle is triggered whenever new particles are added (see Material and Methods). Consequently, during data acquisition, map quality improves as a result of particles being added to the data set and their origins/orientations being refined (Fig. 3).

All three data sets exhibit resolution improvement as a function of acquisition time that is primarily influenced by the corresponding linear increase in the number of particle images recorded.

Additionally, the progress of the reconstruction reflects and conveys information about the quality of each data set. For example, 2 h after the simulated acquisition sessions began, a total of 3768 (CUS-3), 1740 (P22), and 4179 (GCRV) particle images were collected and 3D density maps at 9.7, 9.4, and 6.3-Å resolution, respectively, were available for inspection and analysis. Of note, the number of particles included in any reconstruction is always less than the number of particles acquired at that time for two reasons: some particles are flagged as 'bad' according to the selection mechanism and there is latency in the initial processing of the micrographs. Moreover, an additional latency occurs owing to the time that is required to generate a 3D reconstruction from the set of selected particles. At the end of the acquisition session, *i.e.* after the last micrograph was recorded, the estimated resolution limit of the reconstruction for each of the three data sets was very close to the highest one attainable from the system, which was achieved after three further iterations of refinement (Table 2).

5.6. Comparison with reference results

Another important aspect in evaluating a high-throughput processing system that is tightly coupled to microscopy is the extent to which it can produce a final reconstruction that is comparable to that obtained via alternative methods of processing. For this reason, our tests were performed on data sets that were previously processed by an experienced user in a supervised, semi-automatic manner, as this provided a reliable reference map for comparison purposes. As expected, experts invested considerable effort to achieve the highest resolution possible for each set of image data. In each instance, the expert conducted a preliminary screening of
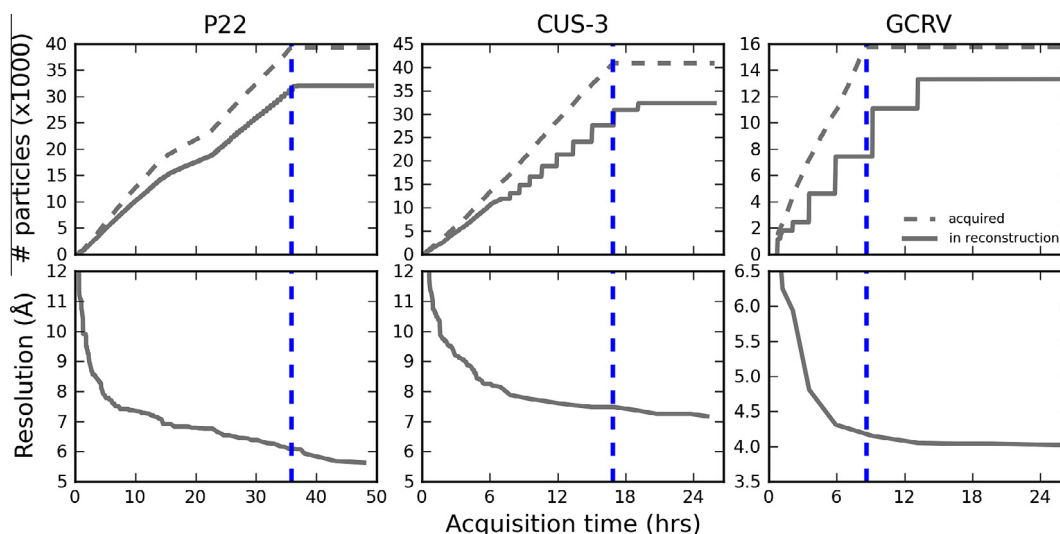
Fig.3. Progress of the reconstruction during data acquisition. Bottom row: for each data set, the resolution of the reconstruction is plotted as a function of the time during the simulated acquisition. Top row: as a reference, the total number of particles acquired (dashed line) and number used in the reconstruction (solid line) are plotted along the same time scale. Vertical dashed lines (blue) indicate the time points when data acquisition ended, after which the system performed only a few iterations of refinement. Note that some curves are plotted on different ordinate and abscissa scales. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Acquisition results: quality of final reconstruction.

	High-throughput		Manual	
	No. of micrographs/particles	Resolution (Å)[a,b]	No. of micrographs/particles	Resolution (Å)[a]
P22	3848/32,050	5.6 (6.1)	2888/18,602	5.4
CUS-3	1818/32,314	7.3 (7.5)	419/7766	6.8
GCRV	258/13,325	4.1 (4.4)	247/18,646	3.8

[a] The resolution is estimated by Fourier Shell Correlation using the cutoff 0.5 (van Heel and Schatz, 2005).
[b] The resolution achieved immediately at the end of the acquisition is indicated in parentheses.

the micrographs according to the quality of their Fourier transforms, and particle images were either picked manually (P22 and CUS-3) or automatically (GCRV), followed by visual particle screening. Manual processing of the data sets took 2 (GCRV), 10 (P22),

and 6 (CUS-3) weeks to complete, and this does not include the time it took to record the micrographs or, in the case of GCRV, to develop the films. When comparing the resolutions of the two reconstructions obtained from each data set by means of Fourier
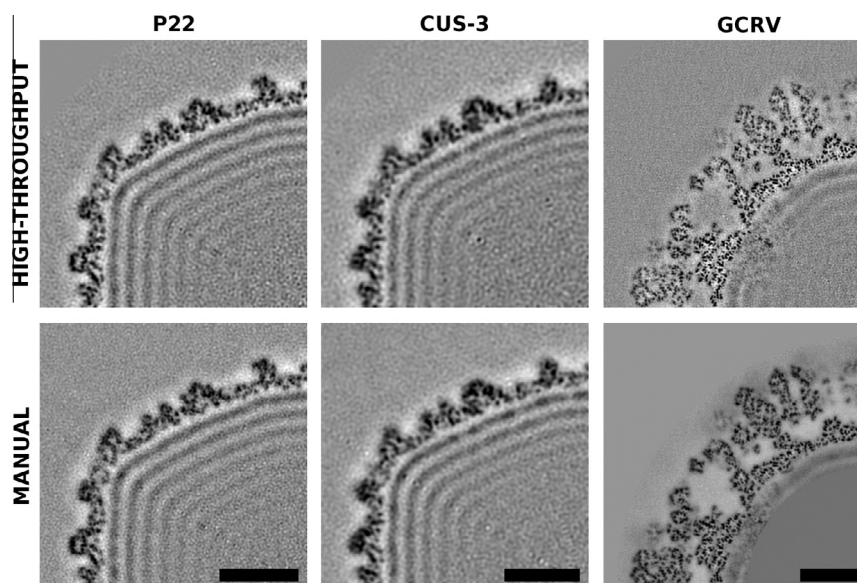
Fig.4. Comparison of cryo-reconstructions obtained by an experienced user ("manual") or by the high-throughput, real-time approach. Each panel shows a 1-pixel-thick, density projection of a quadrant of an equatorial section from the corresponding, final cryo-reconstruction of the P22, CUS-3, and GCRV samples. All bars represent 10 nm.
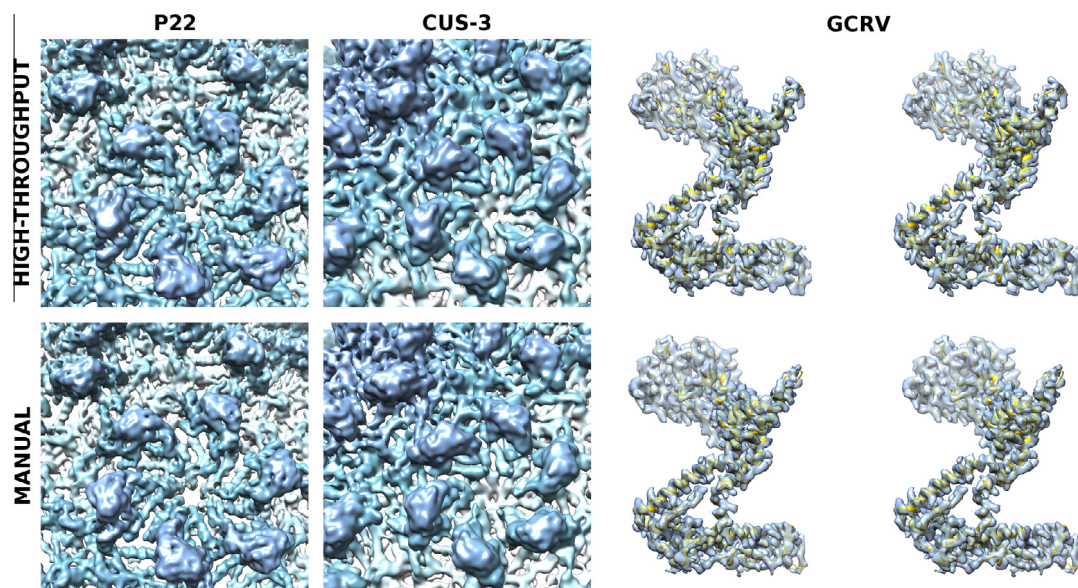
Fig.5. Comparison of results obtained by manual versus automatic, high-throughput approaches: surface renderings. Close-up views are shown of the outer surfaces of the final reconstructions obtained for the P22 (left panels) and CUS-3 (center panels) data sets. For GCRV, a segmented rendering of VP4, one of the outer capsid proteins, is shown in stereo overlapped with its atomic model (chain T of PDB model 3IYL).

Shell Correlation (Table 2 and Supplemental Fig. S2), the map from the automatic, high-throughput approach was quite close and only of slightly lower quality to the one determined manually. The resolution discrepancy ranged between a low of 0.2 Å for P22 (5.6 Å versus 5.4 Å) and a high of 0.5 Å for CUS-3 (7.3 Å versus 6.8 Å), and these differences in resolution estimates were confirmed by visual analysis of the reconstructions (Figs. 4 and 5). For CUS-3, the resolution difference between the two reconstructions is primarily reconciled by the variability in quality of the input micrographs. Indeed, many micrographs in this set were affected by severe specimen drift, and only 419 of the 2020 micrographs were included in the manual analysis. However, during the simulated acquisition, all micrographs were analyzed and only 10% of them were rejected based on the fitting score obtained by CTF estimation. As a comparison, in the P22 data set the quality of the micrographs was more homogenous, whereas in the GCRV simulation we used only the micrographs that were previously scored 'good' by the user, and in both cases the resolution difference between the maps obtained automatically and manually was smaller.

6. Discussion and conclusions

In this study, we present results that demonstrate the feasibility to develop, with currently available software and hardware, a high-throughput computing framework that can process low-dose electron micrographs of unstained, vitrified icosahedral particles during data acquisition. Such a framework would be ideally used in combination with an automated cryoEM data collection system to streamline all the image analysis. For the purposes of this study, we have implemented a proof-of-concept system to test the capabilities and the limitations of an approach that has to satisfy the strict constraints imposed by a real-time system, where processing needs to keep pace with the acquisition rate at the microscope. According to the proposed approach, micrographs are immediately analyzed after they are recorded, particle images are extracted from them and their alignment parameters are determined, and a regularly updated 3D reconstruction is provided, with all steps performed in a fully automated manner. Besides the microscope settings, the only additional information required to start the

process is the diameter of the particle and specification of a file directory for storing all the raw and processed data. To avoid template-induced bias, an initial model is obtained from the first micrographs acquired.

The results from simulated acquisition jobs demonstrate that it is possible to obtain from each micrograph a stack of particles properly aligned to a reference at the same rate at which images are acquired. The simulations were performed under the assumption of an acquisition rate of two images per minute, which is very stringent for most current microscopes and data acquisition software. While single micrographs are processed, a 3D map can be generated in parallel using a select set of all the currently available particles, so the microscopist gets frequent and nearly instant feedback about the quality of the reconstruction that can be achieved from those data. Most important, when acquisition ends the current version of the system provides a final reconstruction that, in terms of quality, is only slightly inferior to what a user could achieve manually and over a much longer time frame after all the data are collected. The strategy adopted by us to achieve these results relies on the use of two asynchronous pipelines to streamline and parallelize the processing, and requires a parallel computing platform to run.

The use of a pipeline to process single micrographs has two major advantages. First, the system does not need to complete all the processing by the time a new micrograph is collected, but only to ensure that the slowest step keeps up with the acquisition rate. Basically, this approach guarantees throughput at the expense of latency. Second, the allocation of multiple processors can be optimized among the different steps according to their individual computational requirements. This is particularly important in single particle analysis where there is large variability in the type of samples analyzed. Typically, the diameters of icosahedral virus particles range from ~20 to 200 nm, and the number of them included in each micrograph depends on the concentration of the sample and generally includes just a few to hundreds of particles. For each of the three test data sets in this study, we showed that the relative computational load of the single processing steps can be different (Fig. 2), and therefore it is desirable to be able to modulate the computing resources among those steps in order to achieve an optimal balance. It is notable that the requirements

for a dedicated, parallel computing platform to implement this approach are not excessive. For our tests, we used five server nodes from a cluster built in 2009, for a total of 53 processors statically assigned to the different tasks. Further optimizations and newer hardware can easily reduce these requirements and adapt the computing resources to the given sample and acquisition settings. All the knowledge acquired from the development of this proof of concept and the feedback derived from the test cases analyzed are currently being used to implement a software system, called SPRINT (Single Particle Reconstruction In No Time), that aims to contain all the elements of robustness and flexibility that are necessary for general use.

6.1. Advantages of a 3D microscope

The integration of image processing with microscopy has the potential to transform an electron microscope into a true 3D instrument, since the final output and even intermediate results from an acquisition session would be a density map ready for analysis. In this way, analysis and interpretation of the structure can start immediately after the end of the acquisition or even before that. Our proof of concept with icosahedral viruses represents a step forward in this direction and provides new and more efficient means to explore the structures and functions of such viruses.

A high-throughput system like the one proposed here eliminates any time lag between acquisition and 3D reconstruction, thus providing instantaneous feedback about the sample and the experimental settings, based on their ability to provide high-resolution structures. For example, if the quality of the sample was poor or suboptimal and unable to yield high-resolution structural data, this would become evident from the quality of the resulting maps obtained while acquiring images at the microscope, and the user would have the option to halt data collection and thereby be spared the need for post processing. Similarly, if the resolution achieved at some time point suffices to answer the biological question of interest, then no further acquisition would be needed and the microscope would be immediately available to examine other specimens.

We observed that the resolution potentially achieved from a particular sample could be predicted along with the number of particles required to achieve it. This capability is made possible given that, during acquisition, several consecutive reconstructions are generated, each one from a different but increasing number of particles. Theoretical considerations have demonstrated a relationship between the number of particles included in a reconstruction and the resolution that can be achieved (Glaeser, 1999; Rosenthal and Henderson, 2003), which also reflects the quality of the images and the experimental settings. Experimental studies (LeBarron et al., 2008; Liu et al., 2007) have verified this relationship by comparing the resolution of reconstructions from different subsets of particles in a data set, showing a linear behavior of $\ln(Nd)$ versus $1/d^2$, where N is the number of particles and d is the resolution achieved (FSC cutoff = 0.5). We observe a similar linear trend for our three data sets, using as data points the resolution values of the intermediate reconstructions obtained during the acquisition from the subset of particles collected thus far (Fig. 6). From the slope of the line, an apparent B factor can be derived that indicates the overall quality of the data and the accuracy of the alignment. In our tests, the apparent B factor agrees with our assessment of the quality of the data, with GCRV providing the lowest value and CUS-3 the highest. To compare these results with the previously published ones (LeBarron et al., 2008; Liu et al., 2007), we repeated the plot for reconstructions obtained by randomly selecting subsets of particles and using their final orientation/origin parameters, as determined at the end of the high-throughput processing. Again, the data follow a linear
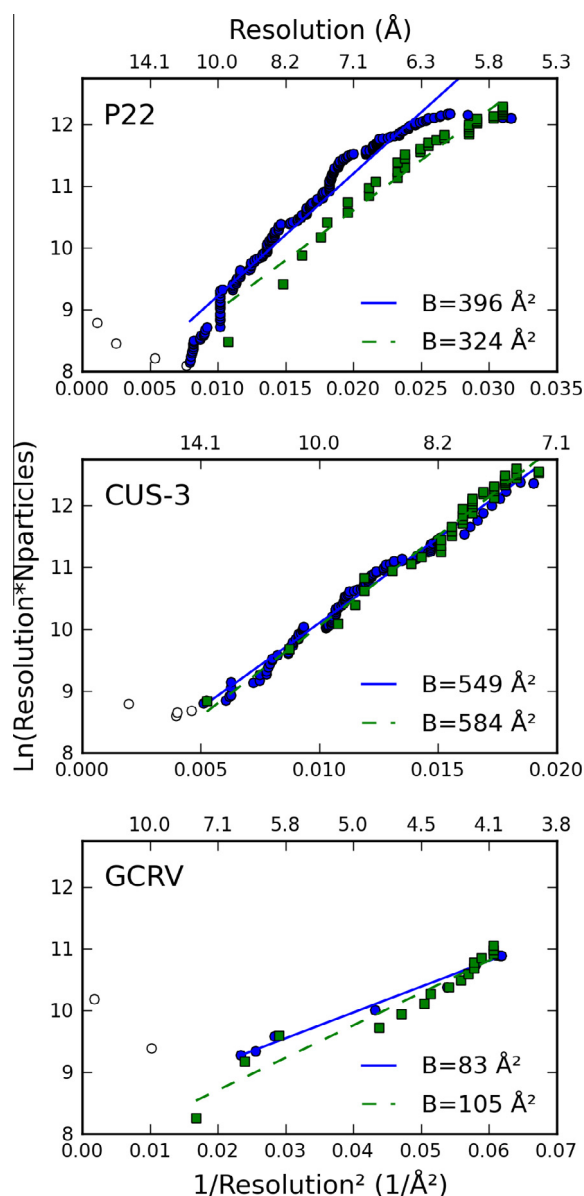
Fig.6. Relationship between number of particles and resolution. Each panel shows a plot for the tests with P22, CUS-3 and GCRV of the logarithm of the product between resolution and number of particles as a function of the reciprocal of the squared resolution. The points represented as circles (blue) were derived from the reconstructions obtained during the acquisition session with just those particle images available at particular times. The points represented as squares (green) were obtained by randomly selecting subsets of particle images (in multiples of 1000) from the entire data and estimating the resolution achieved by the reconstruction generated from these particles, using the latest values determined for their origins and orientations. The lines show linear fits to these points, and the apparent B factor displayed is determined by dividing by half the slope coefficient of the line as described (Rosenthal and Henderson, 2003). When acquisition begins, the quality of the first few reconstructions is not high enough to provide accurate estimates of the origin/orientation parameters of the particles extracted thus far, hence, the first three (GCRV) or four (P22 and CUS-3) points (open circles) were excluded from the linear fit analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

behavior, with only P22 showing a noticeable difference (apparent B factors: 396 versus 324 Å^2) between the two plots (Fig. 6). This general agreement between the two curves indicates that the origin and orientation determined for each particle immediately after its parent micrograph was recorded already closely match those eventually obtained after further cycles of refinement. Hence, the two curves provide consistent information, and this means that,

at any point during the acquisition, one can obtain a reasonable estimate of the number of particles required to achieve a higher target resolution based on a calculation of the linear regression of the curve generated so far. Consequently, the strategy and the time dedicated to a single acquisition session could be adjusted according to the specific needs of the project as determined by the quality of the reconstruction obtained so far and the time required to obtain further improvements. Overall, these monitoring and predictive capabilities of the system can prove to be quite useful in optimizing the use of expensive, multi-user microscopy resources.

6.2. Current challenges

6.2.1. 3D reconstruction

Our experiments have shown that it is possible to obtain aligned stacks of particles in a time comparable with the data acquisition. But, we have yet to achieve a complete, real-time system that includes the 3D reconstruction. In fact, each time when particles are added to the data set or their origins/orientations are refined, the latest 3D density map is discarded and the reconstruction algorithm is relaunched to compute a new map using all the available, updated information. In this regard, the reconstruction is a monolithic process, and even if only one new particle contributes to the map, all particles must be reprocessed. Consequently, while data are being acquired, the time needed to compute a 3D reconstruction increases with the size of the data set (Supplemental Fig. S3). For example, with the P22 data, only when there are fewer than 2000 particle images, reconstructions can be computed in less than 30 s and therefore in less time than that required to record a new micrograph. Reconstruction computation becomes even more costly as resolution improves and reaches the point where unbinned rather than binned particle images are used. This is dramatically true for large particles like GCRV, in which case only one processor could be used for generating the reconstructions, and towards the end of the acquisition session each reconstruction (~13,000 particles) takes ~55 min to compute. A complete, real-time processing system will be possible only after this reconstruction-calculation bottleneck is overcome. As explained above, though some advantages can be derived by minimizing the number of reconstructions to be performed, a comprehensive solution will require the development of novel algorithms.

6.2.2. Particle picking

All the different processing steps that comprise the proposed high-throughput system are based on software implementations (*i.e.* Auto3DEM and CTFFIND) that have been extensively tested and optimized by us and others. Except for the particle-picking procedure, the steps are all general in their applicability and robust under most common, practical situations. Location of particles in the micrographs is performed with an algorithm (Boier Martin et al., 1997) that provides satisfactory results with data sets of overall good quality, like the ones we used for testing, and is mostly suited for quasi-spherical particles. The performance of this algorithm is well documented in our previous analysis of a system for rapid reconstruction of 3D models from single micrographs (Cardone et al., 2013). In the current experiments we observed that the percentages of outliers and particles missed by the algorithm for the P22 and CUS-3 data sets were around 15% and 5%, respectively, which is consistent with previous results for micrographs of good quality. The number of particles missed in the GCRV data set was about 20%, and this is higher than in the other two sets because the GCRV sample was highly concentrated in the fields of view captured in the micrographs, and the algorithm excluded many particles in close contact with each other. Particle picking

is a critical step in a successful implementation of a real-time processing system, especially because the accuracy of the method poses a major limit to the resolution that can be achieved in the final reconstruction. To our knowledge, there are no unsupervised, particle-picking algorithms that can handle any sample quality condition in an automatic manner. A possible solution could come from employing several automatic picking methods simultaneously, and integrating their results by means of a consensus algorithm. Furthermore, picked particles could be screened by quality using a combination of descriptors as recently proposed (Vargas et al., 2013).

6.2.3. Quality control

A robust, automated processing system, in addition to being able to locate particles accurately in the micrographs, must screen the micrographs according to their quality. In a data set obtained with an automated acquisition system, it is typical to generate data sets with a significant fraction of images of suboptimal quality for a variety of reasons (*e.g.* as caused by large ice thickness, specimen heterogeneity and drift, high astigmatism, etc.). In the current experiments, we automatically excluded from the reconstruction all particles originating from micrographs whose fit correlation, as measured by the CTF estimation procedure, was lower than a fixed percentile. This metric, however, only accounts in a generic manner for micrograph quality, and more accurate measures are needed that can adapt to disparate experimental conditions. The difference in the quality of reconstructions obtained from the CUS-3 data using a manual versus a high-throughput approach clearly demonstrated that removal of inferior micrographs, along with a careful selection of particles, is pivotal for attaining the highest possible resolution from the data. In conjunction with quality indicators designed to aid the process of automatically screening micrographs and particles according to their influence in improving resolution in the reconstructed maps, there is a critical need in automated systems to include tools to validate the reconstructions themselves in a manner that surpasses simple resolution assessment (Henderson et al., 2012). Because of the intrinsic low signal-to-noise ratio of input micrographs, noise can adversely affect determination of particle alignment parameters and lead to sub-optimal 3D maps and an over-estimate of the resolution achieved (Stewart and Grigorieff, 2004). Our current software prototype implements a frequency-limited refinement approach (Scheres and Chen, 2012) that minimizes the risk of overfitting (Li et al., 2013). However, a more complete and robust solution will require the additional use of gold-standard refinement procedures (Scheres and Chen, 2012), which guarantee correct estimation of the resolution, and possibly other validation tools that monitor influence of noise in the reconstruction during the acquisition process.

6.2.4. Recording on direct electron detector cameras

The use of such devices is becoming a *de facto* standard to achieve reconstructions at near-atomic resolution with relatively low numbers of particles (Bammes et al., 2012; Campbell et al., 2012). In part, this is a consequence of their high frame rate and hence their ability to generate 'movies' that record beam-induced particle movements, allowing dedicated image processing procedures to compensate for particle image blurring by combining the information from properly aligned frames. This frame alignment and averaging procedure would be the first operation to be performed after acquisition and, in a system like the one proposed here, would correspond to the first task in Pipeline 1. Efficient solutions to accomplish this are available that run on graphic processing units capable of processing one set of frames in times that are potentially quicker than the rate of acquiring micrographs (Li et al., 2013). However, these solutions will need to be integrated with

algorithms that determine the optimal subset of frames to average, in order to maximize the high resolution frequency content.

6.3. Extension to asymmetric particles

The approach we propose here has been implemented specifically for particles with icosahedral symmetry, but most processing components of the system do not require such high particle symmetry to work. For example, Pipeline 2, which performs the reconstruction and iteratively refines the alignment parameters, can already process particles with reduced or no symmetry. On the contrary, some modifications would be required in Pipeline 1, which processes each micrograph separately. Specifically, the two computing steps that need modification and currently represent a challenge in processing images of asymmetric particles in a high-throughput manner are particle-picking and determination of an *ab initio* model. A universal procedure capable of locating particles in a micrograph, automatically and without supervision, still poses a significant challenge despite numerous efforts (Potter et al., 2004). The problem is currently further exacerbated with small (<500 kDa) molecular complexes that are difficult to discriminate in noisy, low-dose cryo-images. Successful results have been obtained by combining picking algorithms with multiple layers of classification (*e.g.* (Lyumkis et al., 2013)), but such approaches still rely on lengthy, manual user intervention.

Another major obstacle is being able to compute rapidly a reliable, *ab initio* model from the data. Recently, probabilistic approaches to tackle this problem (Elmlund et al., 2013; Sanz-García et al., 2010) show promise but require more extensive experimental analysis. Furthermore, regarding asymmetric particles, sample heterogeneity is more a norm rather than an exception, since particles may adopt a variety of different conformations or their components are incorporated with different stoichiometries (*e.g.* see (Fernández et al., 2013) where one structure was obtained from less than 3% of the full data set of images). This situation makes mandatory the introduction of an additional computational step that classifies the particles and segregates them among different models. Given these additional difficulties, the validation of 3D reconstructed maps is a sensitive issue and one that should be always be addressed (Henderson et al., 2012), and is especially critical for an automated system. Nevertheless, technological advances will make real time, 3D microscopy feasible in the foreseeable future.

Acknowledgments

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsb.2014.02.016.

References

Baker, T.S., Cheng, R.H., 1996. A model-based approach for determining orientations of biological macromolecules imaged by cryoelectron microscopy. J. Struct. Biol. 116, 120–130.

Bammes, B.E., Rochat, R.H., Jakana, J., Chen, D.-H., Chiu, W., 2012. Direct electron detection yields cryo-EM reconstructions at resolutions beyond 3/4 Nyquist frequency. J. Struct. Biol. 177, 589–601.

Boier Martin, I.M., Marinescu, D.C., Lynch, R.E., Baker, T.S., 1997. Identification of spherical virus particles in digitized images of entire electron micrographs. J. Struct. Biol. 120, 146–157.

Campbell, M.G., Cheng, A., Brilot, A.F., Moeller, A., Lyumkis, D., Veesler, D., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B., et al., 2012. Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. Structure 20, 1823–1828.

Cardone, G., Yan, X., Sinkovits, R.S., Tang, J., Baker, T.S., 2013. Three-dimensional reconstruction of icosahedral particles from single micrographs in real time at the microscope. J. Struct. Biol. 183, 329–341.

Chang, J., Liu, X., Rochat, R.H., Baker, M.L., Chiu, W., 2012. Reconstructing virus structures from nanometer to near-atomic resolutions with cryo-electron microscopy and tomography. In: Rossmann, M.G., Rao, V.B. (Eds.), Viral Molecular Machines. Springer, US, pp. 49–90.

Elmlund, H., Elmlund, D., Bengio, S., 2013. PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. Structure 21, 1299–1306.

Fernández, I.S., Bai, X.-C., Hussain, T., Kelley, A.C., Lorsch, J.R., Ramakrishnan, V., Scheres, S.H.W., 2013. Molecular architecture of a eukaryotic translational initiation complex. Science 342, 1240585.

Glaeser, R.M., 1999. Review: electron crystallography: present excitement, a nod to the past, anticipating the future. J. Struct. Biol. 128, 3–14.

Glaeser, R.M., Hall, R.J., 2011. Reaching the information limit in cryo-EM of biological macromolecules: experimental aspects. Biophys. J. 100, 2331–2337.

Grigorieff, N., Harrison, S.C., 2011. Near-atomic resolution reconstructions of icosahedral viruses from electron cryo-microscopy. Curr. Opin. Struct. Biol. 21, 265–273.

Henderson, R., Sali, A., Baker, M.L., Carragher, B., Devkota, B., Downing, K.H., Egelman, E.H., Feng, Z., Frank, J., Grigorieff, N., et al., 2012. Outcome of the first electron microscopy validation task force meeting. Structure 20, 205–214.

Korinek, A., Beck, F., Baumeister, W., Nickell, S., Plitzko, J.M., 2011. Computer controlled cryo-electron microscopy – TOM2 a software package for high-throughput applications. J. Struct. Biol. 175, 394–405.

Lander, G.C., Tang, L., Casjens, S.R., Gilcrease, E.B., Prevelige, P., Poliakov, A., Potter, C.S., Carragher, B., Johnson, J.E., 2006. The structure of an infectious P22 virion shows the signal for headful Dna packaging. Science 312, 1791–1795.

LeBarron, J., Grassucci, R.A., Shaikh, T.R., Baxter, W.T., Sengupta, J., Frank, J., 2008. Exploration of parameters in cryo-EM leading to an improved density map of the *E. coli* ribosome. J. Struct. Biol. 164, 24–32.

Lei, J., Frank, J., 2005. Automated acquisition of cryo-electron micrographs for single particle reconstruction on an FEI Tecnai electron microscope. J. Struct. Biol. 150, 69–80.

Li, X., Mooney, P., Zheng, S., Booth, C.R., Braunfeld, M.B., Gubbens, S., Agard, D.A., Cheng, Y., 2013. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat. Methods 10, 584–590.

Liu, X., Jiang, W., Jakana, J., Chiu, W., 2007. Averaging tens to hundreds of icosahedral particle images to resolve protein secondary structure elements using a Multi-path Simulated Annealing optimization algorithm. J. Struct. Biol. 160, 11–27.

Lyumkis, D., Moeller, A., Cheng, A., Herold, A., Hou, E., Irving, C., Jacovetty, E.L., Lau, P.-W., Mulder, A.M., Pulokas, J., et al., 2010. Chapter fourteen – automation in single-particle electron microscopy: connecting the pieces. In: Jensen, Grant J. (Ed.), Methods in Enzymology. Academic Press, pp. 291–338.

Lyumkis, D., Julien, J.-P., de Val, N., Cupo, A., Potter, C.S., Klasse, P.-J., Burton, D.R., Sanders, R.W., Moore, J.P., Carragher, B., et al., 2013. Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. Science 342 (2013), 1484–1490.

Mindell, J.A., Grigorieff, N., 2003. Accurate determination of local defocus and specimen tilt in electron microscopy. J. Struct. Biol. 142, 334–347.

Potter, C.S., Zhu, Y., Carragher, B., 2004. Automated particle selection for cryo-electron microscopy. J. Struct. Biol. 145, 1–2.

Rosenthal, P.B., Henderson, R., 2003. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J. Mol. Biol. 333, 721–745.

Sanz-García, E., Stewart, A.B., Belnap, D.M., 2010. The random-model method enables ab initio 3D reconstruction of asymmetric particles and determination of particle symmetry. J. Struct. Biol. 171, 216–222.

Scheres, S.H.W., Chen, S., 2012. Prevention of overfitting in cryo-EM structure determination. Nat. Methods 9, 853–854.

Shi, J., Williams, D.R., Stewart, P.L., 2008. A Script-Assisted Microscopy (SAM) package to improve data acquisition rates on FEI Tecnai electron microscopes equipped with Gatan CCD cameras. J. Struct. Biol. 164, 166–169.

Shigematsu, H., Sigworth, F.J., 2013. Noise models and cryo-EM drift correction with a direct-electron camera. Ultramicroscopy 131, 61–69.

Sorzano, C.O., Rosa Trevín, J.M., Otón, J., Vega, J.J., Cuenca, J., Zaldívar-Peraza, A., Gómez-Blanco, J., Vargas, J., Quintana, A., Marabini, R., et al., 2013. Semi-automatic, high-throughput, high-resolution protocol for three-dimensional

reconstruction of single particles in electron microscopy. In: Sousa, A.A., Kruhlak, M.J. (Eds.), Nanoimaging. Humana Press, Totowa, NJ, pp. 171–193.

Staples, G., 2006. TORQUE resource manager. In: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, ACM, New York, NY, USA, p. 8.

Stewart, A., Grigorieff, N., 2004. Noise bias in the refinement of structures derived from single particles. Ultramicroscopy 102, 67–84.

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, C.S., Carragher, B., 2005. Automated molecular microscopy: the new Leginon system. J. Struct. Biol. 151, 41–60.

Tang, J., Lander, G.C., Olia, A.S., Olia, A., Li, R., Casjens, S., Prevelige Jr., P., Cingolani, G., Baker, T.S., Johnson, J.E., 2011. Peering down the barrel of a bacteriophage portal: the genome packaging and release valve in P22. Structure 19, 496–502.

Van Heel, M., Schatz, M., 2005. Fourier shell correlation threshold criteria. J. Struct. Biol. 151, 250–262.

Vargas, J., Abrishami, V., Marabini, R., de la Rosa-Trevín, J.M., Zaldivar, A., Carazo, J.M., Sorzano, C.O.S., 2013. Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. J. Struct. Biol. 183, 342–353.

Yan, X., Sinkovits, R.S., Baker, T.S., 2007a. AUTO3DEM – an automated and high throughput program for image reconstruction of icosahedral particles. J. Struct. Biol. 157, 73–82.

Yan, X., Dryden, K.A., Tang, J., Baker, T.S., 2007b. *Ab initio* random model method facilitates 3D reconstruction of icosahedral particles. J. Struct. Biol. 157, 211–225.

Zhang, X., Jin, L., Fang, Q., Hui, W.H., Zhou, Z.H., 2010. 3.3 Å Cryo-EM structure of a nonenveloped virus reveals a priming mechanism for cell entry. Cell 141, 472–482.

Zhou, Z.H., 2011. Atomic resolution cryo electron microscopy of macromolecular complexes. Adv. Protein Chem. Struct. Biol. 82, 1–35.