Outlier detection for single particle analysis in Electron Microscopy

C.O.S. Sorzano[1,2], J. Vargas[1] J.M de la Rosa-Trevín[1], A. Zaldívar-Peraza[1], J. Otón[1], V. Abrishami[1], I. Foche[1], R. Marabini[3], G. Caffarena[2], and J.M Carazo[1]

[1] Biocomputing Unit, National Center of Biotechnology, CSIC. 28049 Madrid, Spain. {coss@cnb.csic.es}
[2] Bioengineering Laboratory, Univ. CEU-San Pablo. 28668 Madrid, Spain.
[3] Escuela Politecnica Superior, Univ. Autonoma de Madrid. 28049 Madrid, Spain.

Abstract. Electron Microscopy (EM) of macromolecular structures using a single particle approach normally involves a two-dimensional (2D) classification step as a exploratory data analysis in which conformational changes, contaminants, or damaged particles may be identified. This step is nowadays even more important as automatic acquisition procedures are routinely employed and hundreds of thousands or millions of images can be acquired at the electron microscope. Automatic particle picking algorithms have a non-negligible false positive rate (wrongly selected particles), and many times they unadvertedly pass through the 2D classification, thus contaminating the dataset employed for 3D reconstruction. In this article we present an algorithm to reduce the number of these contaminating images, generally referred to as outliers.

Keywords: Single particle analysis, 2D classification, outlier detection, one-class classification

1 Introduction

Electron Microscopy of single particles has become a valuable tool to visualize macromolecular structures at medium-high resolution. Currently, an increasingly high number of structures have reached a resolution level below 4Å [1–4]. This high resolution is achieved by combining a large number of homogeneous particles [5] and correcting for the aberrations introduced by the microscope. There have already been structural studies involving over 2 million particles [6]. This huge number of images has been possible thanks to the automation of micrograph acquisition [7–9] and particle picking [10,11]. Among other factors, the key to high resolution EM is the selection of a set of projection images as homogeneous as possible. This is done by fine tuning the biochemical purification of the complex being studied, by fixing as much as possible its conformational state, and by guaranteeing the homogeneity of acquisition conditions. If automatic particle selection algorithms are employed, an increasingly more frequent situation

in current structural studies, they may spoil all the special care dedicated to sample production, preparation and acquisition. However, manual selection of this huge amount of particles is totally unfeasible. For this reason, several image processing algorithms are starting to appear with the aim of mitigating this problem [12–15]. They are normally based in identifying outliers assuming that the majority of particles are correctly selected. All these algorithms are based on extracting a number of features on the particles, and then using a classification rule to distinguish between good and bad particles. Norousi *et al.* [12] use a classifier trained with examples of particles and non-particles. Their feature vector is rotationally invariant but it might be very much dependent on having selected the right particle center, which may not always be the case after particle picking. Their algorithm is also very much dependent on the set of non-particles selected for training. They must be representative of the kind of non-particles found among the chosen particles so that wrongly picked particles can be effectively identified. Langlois *et al.* [15] also use a Principal Component Analysis (PCA) projection of either the whole set of images or their bispectrum as a way of constructing a shift invariant feature. They handle rotations by introducing several rotated versions of each image. Interestingly, they introduce the concept of one-class classification, which is simply a threshold on the value of the projection of each image onto a 1-dimensional PCA subspace. Vargas *et al.* [14] has a rotationally and translationally invariant feature vector. The classification is performed by considering the whole set of picked particles, projecting them onto a low-dimensional space (computed by PCA), and calculating the Mahalanobis distance to the cloud centroid. The classification is then performed by setting a threshold on this distance so that images beyond a certain distance are classified as non-particles. Although designed with denoising purposes, [16] also designed a scheme in which a rotational invariant set of features are projected onto a PCA subspace. the work of Moriya *et al.* [13] is also addressing the problem of how to discard bad images. Their idea is to perform a 2D classification and discard all those images assigned to classes that do not converge to a stable centroid. They discard up to 50% of the input images.

All these classification schemes are hampered by the fact that the input dataset (the whole set of selected images) is too heterogeneous and, consequently, the feature vector has to deal with very different features and the PCA projects onto a linear subspace a too wide set of points. In this article, we propose to apply a similar classification strategy (identifying multivariate outliers in a low-dimensional subspace), but in a much more homogeneous environment, namely, the subset of images assigned to a 2D class during a 2D classification process. Ideally, all images assigned to a 2D class should be identical. In practice, a homogeneous class has a small variation around its centroid (what we call the class core), and outliers can be identified through their Mahalanobis distance to the centroid. In this regard, the approach is similar to that of Langlois *et al.* [15] although we are not limited by the dimension of the PCA subspace and we use raw images after being aligned, which solves most problems related to the variance introduced by shifts, rotations, different projection directions and .

Additionally, we propose to work with a hierarchical classification algorithm [17]. Then, we are able to further refine the class core by considering the subset of images that is basically classified together in the classification hierarchy. We will refer to this subset as the stable class core. We show that the core as well as the stable class core can effectively remove contaminating particles. In this way, the proposed methodology can be employed as an automatic refinement in a high-throughput EM environment.

2 Materials and methods

In this section we introduce our methodology to automatically reject some images in the set of selected particles. We first introduce the concept of class core, that can be applied to any 2D classification algorithm. Then, we introduce the stable class core, that is only defined for CL2D [17]. Both concepts start by assuming that a 2D classification algorithm has been applied to the whole set of selected images, and that the algorithm is such that it tries to produce as homogeneous classes as possible. This is particularly true for algorithms based on vector quantization such as K-means and CL2D. Let us assume that there are N images in the original dataset, and that they are divided into K disjoint classes. Let us represent as X_n the n-th image ($n = 1, 2, ..., N$), χ_k the k-th class representative ($k = 1, 2, ..., K$), and $\tilde{X}_n$ the n-th image aligned to its corresponding class representative. Let us refer to the number of images assigned to the k-th class as N_k, and the specific set of indexes of images assigned to that class as $\mathcal{C}_k$.

2.1 Class cores

K-means is an algorithm that tends to produce globular clusters of maximum radius ϵ (this property is shared by algorithms like CL2D). It has been recently shown that there is a strong connection between K-means and PCA [18] in such a way that PCA is a continuous solution of the cluster membership indicators in K-means. Given a set of images assigned to the same class and considering this property, it makes sense to measure the dispersion of the projection images assigned to this class in the PCA subspace. In this way, we can measure how heterogeneous the class is and easily identify outliers. The procedure is relatively simple. Given the set of aligned images $\tilde{X}_n$ assigned to a given class, we compute the class covariance matrix as $\hat{\Sigma}_k = \frac{1}{N_k - 1} \sum_{n \in \mathcal{C}_k} (\tilde{X}_n - \chi_k)(\tilde{X}_n - \chi_k)^t$, where $(\cdot)^t$ denotes the transpose operator. $\hat{\Sigma}_k$ is positive semidefinite by construction, and therefore, all its eigenvalues are real and non-negative. Let W_k^q denote the matrix whose columns are the q eigenvectors of $\hat{\Sigma}_k$ associated to the q largest eigenvalues, and D_k^q a diagonal matrix with the corresponding eigenvalues. Then, we can project the aligned images onto the PCA subspace by $\tilde{x}_n = (D_k^q)^{-\frac{1}{2}} (W_k^q)^t (\tilde{X}_n - \chi_k)$. The squared norm of these vectors, $\|\tilde{x}_n\|^2$ is strongly related to measuring the Mahalanobis distance of the aligned images to

the class centroid, only that it is better defined in case that the covariance matrix is singular. This is a standard multivariate outlier detection technique [19]. For a homogeneous class of images, one would expect that all of them are projected towards the center of the PCA subspace, and therefore multivariate outliers can be identified by setting a threshold on the squared distance. If the projected values, $\tilde{x}_n$, follow a multidimensional normal in the PCA subspace (which may not be necessarily the case in EM images), then it can be easily proven that $\|\tilde{x}_n\|$ follows a χ distribution with q degrees of freedom. In this case, it is custom to set the threshold at $\sqrt{q+3\sqrt{q}}$, that contains approximately 98% of the population. Note that since the EM images need not be projected in the PCA subspace as a multivariate normal, then this threshold does not necessarily correspond to 98% of the population. However, this number already gives a clue about reasonable values that the threshold can take. In our implementation, we normally work with $q = 2$, and therefore the threshold should be around 2.8. However, we let the user choose any desired value.

Summarizing, we define the class core as those images assigned to a class such that when they are projected onto a PCA subspace of dimension q, their distance to the subspace origin is smaller than a given threshold. Note that the PCA projection used for the class core is much more powerful than the PCA projection used by Vargas *et al.* [14] since the one used in this article is using the aligned image, while the one used by Vargas *et al.* reduces the image to a small feature vector, with the subsequent loss of information. Additionally, the PCA basis constructed in this article is performed on a much more homogeneous dataset (the 2D class) than that used by Vargas *et al.* in which all selected particles are used.

2.2 Stable class cores

We can further refine the class core by what we refer to as the stable class core and that is computed thanks to the hierarchical nature of CL2D. CL2D is a clustering algorithm that starts by clustering the images into a small set of classes. Then, these clusters are progressively subdivided till the desired number of classes is reached. This divisive approach has been shown to produce more stable, robust and accurate results than directly trying to divide the input dataset into the final number of classes [17]. In our approach, images are free to change class at any moment, that is, they are not restricted to be clustered into one of the two subclusters in which its parent cluster has been subdivided. We refer to the successive clustering steps as levels. At level 0, images are subdivided into N_0 clusters; at level 1, into $2N_0$ classes; at level 2, into $4N_0$ classes; this goes till the desired N_F classes are reached. In general, at level l there are $\min(2^l N_0, N_F)$ classes. Let $L = \text{floor}\left(\log_2 \frac{N_F}{N_0}\right)$ be the final number of levels.

Considering this hierarchical scheme, we say that two images belong to the stable class core if both of them have regularly being in the same class core through the different classification levels. Given two images, X_{n_1} and X_{n_2}, let us define the function $S(X_{n_1}, X_{n_2}, l)$ that is 1 if both images were classified in

the same class core at clustering level l, and 0 otherwise. Formally, two images belong the stable class core if $\sum_{l=0}^{L} S(X_{n_1}, X_{n_2}, l) \geq L - \eta$, where η is a tolerance parameter that takes a non-negative integer value $(0, 1, 2, ...)$. If $\eta = 0$, then the two images have always had to be in the same class core; if $\eta = 1$, they had to be always in the same class core except at 1 level, etc. In our implementation, η is a parameter that the user may choose and whose default value is 1.

3 Results

In the following sections we show how the class cores and stable class cores actually help to construct purer classes, we do that using simulated (where the ground truth is known) as well as experimental data. All results reported below are computed for $q = 2$ (i.e., the PCA subspace is two-dimensional), an class core threshold distance of 3, and a tolerance $\eta = 1$.

3.1 Simulated data

We simulated 10,000 projection images from the bacteriorhodopsin (PDB entry: 1BRD) at a sampling rate of 3.5Å. Each projection image had a completely random projection direction. We added noise to a Signal-to-Noise Ratio of 1/10 and performed a CL2D classification into 128 classes. For each class and each image pair within that class, we analyzed the angular distance between the projection directions associated to the two images in the pair. The average angular distance between images in the raw CL2D classes was 16.9° and 16.8° for the core classes (we refer to this value as the average intra-class angular dispersion). However, the stable class cores reduced this average intra-class angular dispersion to 15.2° (a Kolmogorov-Smirnov test on the two distributions (stable core and core) of angular distances showed that the two were different with a confidence level well above 99.9%).

We report the execution time for this first experiment since for the rest of the experiments the time proportion is similar. We used 16 cores of a cluster with 24 nodes with two Xeon E5405 2Ghz processors (4 cores/processor), 16 GB of RAM per node, and node interconnection at 1 Gb/s. CL2D took 291 minutes on a cluster. Computing the class cores took 4 minutes (1.4% of the execution time of CL2D) and computing the stable class cores 5 minutes (1.7%). Althogether, computing the cores and class cores added a 3% of extra execution time upon the CL2D execution time.

We added to the previous dataset 2,000 empty projection images with the same amount of noise. This situation is typical when some automatic particle selection algorithms are applied, they tend to pick a non-negligible percentage of images at local peaks of their identification function. We then divided the 12,000 images into 128 classes using CL2D. 23 of the 128 classes were contaminated by empty images; that is, most of the images in those 23 classes corresponded to bacteriorhodopsin projections although there also were many empty images

assigned to the same class. However, only one of the class cores and none of the stable class cores were contaminated by these empty images. The average intra-class angular dispersion was reduced from 21.5 in the CL2D classes to 18.1 in the stable class cores.

3.2 Experimental data

We used the 10,000 70S *E. coli* ribosome particles deposited at the EBI by J. Frank and normally used as a classification benchmark [20]. 5,000 of these particles correspond to the ribosome with EF-G bound, while the other 5,000 correspond to particles with absent EF-G. We divided them into 128 classes using CL2D. We evaluated the proportion of images in each class coming from each of the two structural conformations. It is said that a class is 100% pure if all its assigned images belong to the same class. For the raw CL2D classes, the average purity was 62.9%. The class cores increased the average purity to 63.4%, and the stable class cores increased this average purity to 64.5%.

We also tested our algorithm on the KLH blind benchmark dataset at http://i2pc.cnb.csic.es/3dembenchmark for particle picking. These images were acquired, using a Philips CM200 TEM equipped with a 2Kx2K CCD Tietz camera, as defocal pairs at a nominal magnification of 66,000x and a voltage of 120 KeV, using the Leginon system [21, 22]. Sampling rate was 2.2Å/pixel and the accumulated dose for each high magnification image area was 10 e/Å^2. We automatically selected 988 particles using the algorithm described in [23]. The precision, as defined in [24], of this dataset was 79.5%, meaning that this percentage of particles were particles that had been selected in the gold standard. Applying the algorithm of [14] with a threshold of 3, the precision was increased to 82.4%. Then, we analyzed the selected particles with CL2D dividing them into 16 clusters. The precision increased to 84.0% in case of the class cores, and 84.7% in case of the stable class cores. In terms of wrongly selected particles, the set of incorrect particles was reduced by more than 25% of its size (=(84.7%-79.5%)/(100%-79.5%)), that is, there are less than 25% of contaminants.

Finally, we used the images from 44,311 projection images of *Agrobacterium tumefaciens* VirE2 [25] at a sampling rate of 2.2 Å/pixel. Since this dataset is not part of any benchmark, we can only report qualitative results on it. Before entering in Xmipp CL2D, these images had been masked, aligned, and filtered by SPIDER. Figure 1 shows one of the 2D classes calculated by CL2D. It is the average of 1,077 projections. In the same figure we show some of these projections that did not progress to the stable core. It can be seen that some of the projections are empty, have a lower SNR, contamination or border artifacts.

4 Discussion and conclusions

As the number of particles used for structural studies is steadily increasing, thanks to the use of automated acquisition procedures at the microscope and automatic particle selection algorithms, there is a compelling need for accurate

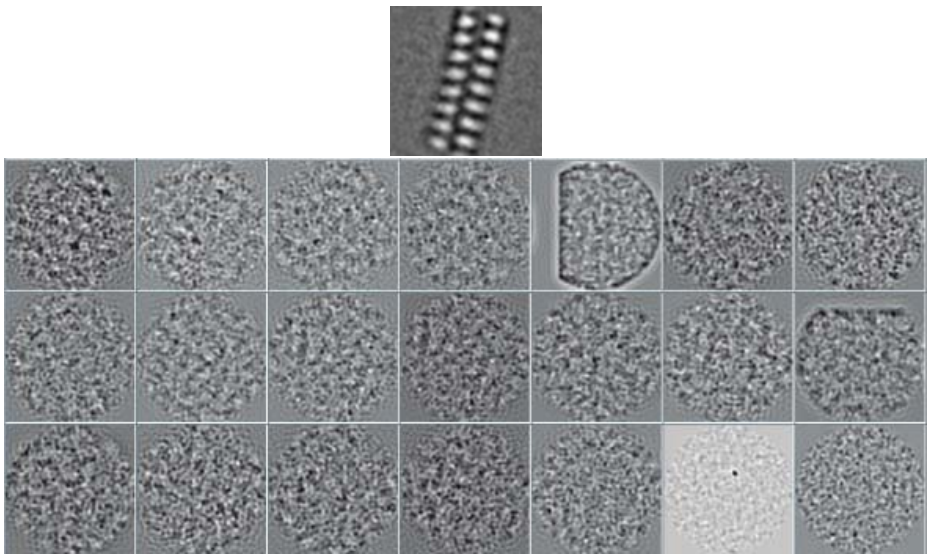

Fig. 1. Example of a 2D class of VirE2 and some of the images that belong to the CL2D class, but not to its stable core.

validation of the particles used during the analysis. This new requirement has been recently acknowledged and some works have already addressed this necessity [12,14,15]. All previous works dealt with the entire dataset trying to identify those particles not following the main trend (contaminants, damaged particles, electron dense regions, empty projections, ...). Because they were applied at a relatively early stage of the analysis, just after particle selection, they had to remove the heterogeneity caused by in-plane misalignments. They achieved this by computing some kind of rotational and/or translational invariant feature vector that either it is classified by a supervised classification algorithm that needs training [12], or they are automatically classified in a one-class classification scheme [14, 15] by choosing some threshold. The work of Vargas *et al.* [14] additionally offers the possibility of interactive selection of good particles, and this task is simplified thanks to a particle ranking according to a given score. The main drawback of using these feature vectors is that there is a great loss of information that cannot be used by the outlier detection algorithm.

The work presented in this paper is also a one-class classification scheme like [14] and [15]. Consequently, it does not need to be trained with negative examples (in fact, in Xmipp [26] this task is explicitly done by the automatic particle picking algorithm [23, 27]). This is convenient from the user point of view since she does not need to select the kind of outliers she wants to reject. Unlike [14] and [15], this outlier detection is performed after a 2D classification step. We, thus, avoid the heterogeneity introduced by misalignment as well as the heterogeneity introduced by different projection directions and conformational states. In this

situation, PCA can be directly applied to the aligned images and their projection onto a lower dimensional space performs an implicit denoising. Our algorithm can use any number of PCA dimensions (unlike [15] that is primarily aimed at 1D projections), and the outlier detection in this linear subspace is fully supported by standard multivariate outlier detection theory [19]. Additionally, when CL2D is used to perform the 2D classification, we can exploit its hierarchical nature to identify particles that are not stably classified. The assumption here is that a stable classification is a sign of good classification (something already put forward in [28] and [13]).

The numerical experiments performed support the hypothesis that class cores as well as stable class cores effectively reduce the variability of the 2D classes. This is beneficial to class representatives (that are the average of more homogeneous projections) and algorithms that may use them as their input (like all those constructing initial volumes based on common lines [29–31]). We may also use this outlier rejection step to prevent those images labeled as outliers to go into subsequent image processing steps like 3D reconstruction (although the effect of outlier rejection in 3D reconstruction is out of the scope of this article and needs to be further explored). Obviously, the price to pay for these outlier rejection algorithms is the number of good particles that are also discarded. However, this drawback loses importance in an era in which acquisition automation is the current trend and structural studies with up to 2 million projections have been carried out [6]. Overall, outlier rejection algorithms are beneficial as long as the number of correctly discarded particles is larger than the number of incorrectly discarded ones.

The methodology described in this paper is accessible through the CL2D protocol in Xmipp 3.0, http://xmipp.cnb.csic.es [32].

Acknowledgements

The authors would like to acknowledge economical support from the Comunidad de Madrid through grant CAM (S2010/BMD-2305), the NSF through Grant 1114901, the Spanish Ministry of Economy and Competitiveness through Grants AIC-A-2011-0638, BIO2010-16566, postdoctoral Juan de la Cierva Grants with references JCI-2011-10185 and JCI-2010- 07594, BES-2011-044096, JAEPre-09-01717. C.O.S. Sorzano is recipient of a Ramón y Cajal fellowship. This work was partially funded by Instruct, part of the European Strategy Forum on Research Infrastructures (ESFRI) and supported by national member subscriptions

References

1. H. Liu, L. Jin, S. B. S. Koh, I. Atanasov, S. Schein, L. Wu, Z. H. Zhou, Atomic structure of human adenovirus by cryo-em reveals interactions among protein networks., Science 329 (5995) (2010) 1038–1043.
2. X. Zhang, L. Jin, Q. Fang, W. H. Hui, Z. H. Zhou, 3.3 a cryo-em structure of a nonenveloped virus reveals a priming mechanism for cell entry., Cell 141 (2010) 472–482.

3. P. Ge, Z. H. Zhou, Hydrogen-bonding networks and rna bases revealed by cryo electron microscopy suggest a triggering mechanism for calcium switches., Proc. Natl. Acad. Sci. USA 108 (23) (2011) 9637–9642.
4. X. Yu, P. Ge, J. Jiang, I. Atanasov, Z. H. Zhou, Atomic model of cpv reveals the mechanism used by this single-shelled virus to economically carry out functions conserved in multishelled reoviruses., Structure 19 (5) (2011) 652–661.
5. X. Zhang, Z. H. Zhou, Limiting factors in atomic resolution cryo electron microscopy: no simple tricks., J. Structural Biology 175 (3) (2011) 253–263.
6. N. Fischer, A. L. Konevega, W. Wintermeyer, M. V. Rodnina, H. Stark, Ribosome dynamics and trna movement by time-resolved electron cryomicroscopy., Nature 466 (7304) (2010) 329–333.
7. U. Ziese, W. J. Geerts, T. P. Van Der Krift, A. J. Verkleij, A. J. Koster, Correction of autofocusing errors due to specimen tilt for automated electron tomography, J. Microscopy 211 (2003) 179–185.
8. C. S. Potter, J. Pulokas, P. Smith, C. Suloway, B. Carragher, Robotic grid loading system for a transmission electron microscope, J. Structural Biology 146 (2004) 431–440.
9. C. Suloway, J. Pulokas, D. Fellmann, A. Cheng, F. Guerra, J. Quispe, S. Stagg, C. S. Potter, B. Carragher, Automated molecular microscopy: the new leginon system, J. Structural Biology 151 (2005) 41–60.
10. Y. Zhu, B. Carragher, R. M. Glaeser, D. Fellmann, C. Bajaj, M. Bern, F. Mouche, F. de Haas, R. J. Hall, D. J. Kriegman, S. J. Ludtke, S. P. Mallick, P. A. Penczek, A. M. Roseman, F. J. Sigworth, N. Volkmann, C. S. Potter, Automatic particle selection: results of a comparative study, J. Structural Biology 145 (2004) 3–14.
11. C. O. S. Sorzano, E. Recarte, M. Alcorlo, J. R. Bilbao-Castro, C. San-Martín, R. Marabini, J. M. Carazo, Fast automatic particle selection from electron micrographs using machine learning techniques, J. Structural Biology 167 (2009) 252–260.
12. R. Norousi, S. Wickles, C. Leidig, T. Becker, V. J. Schmid, R. Beckmann, A. Tresch, Automatic post-picking using mappos improves particle image detection from cryo-em micrographs., J. Structural Biology 182 (2) (2013) 59–66.
13. T. Moriya, K. Mio, C. Sato, Novel convergence-oriented approach for evaluation and optimization of workflow in single-particle two-dimensional averaging of electron microscope images, Microscopy(in press).
14. J. Vargas, V. Abrishami, R. Marabini, J. M. de la Rosa-Trevín, A. Zaldivar, J. M. Carazo, C. O. S. Sorzano, Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques(submitted).
15. R. Langlois, J. T. Ash, J. Pallesen, J. Frank, Computational Methods for Three-Dimensional Microscopy Reconstruction, Springer, 2013, Ch. Fully Automated Particle Selection and Verification in Single-Particle Cryo-EM, (in press).
16. Z. Zhao, A. Singer, Fourier-bessel rotational invariant eigenimages, J. Optical Soc. America A 30 (2012) 871–877.
17. C. O. S. Sorzano, J. R. Bilbao-Castro, Y. Shkolnisky, M. Alcorlo, R. Melero, G. Caffarena-Fernández, M. Li, G. Xu, R. Marabini, J. M. Carazo, A clustering approach to multireference alignment of single-particle projections in electron microscopy., J. Structural Biology 171 (2010) 197–206.
18. C. Ding, X. He, K-means clustering via principal component analysis, in: Proc. 21st Intl. Conf. Machine Learning ,, 2004.
19. D. Peña, F. J. Prieto, Multivariate outlier detection and robust covariance matrix estimation, Technometrics 43 (2001) 286–310.

20. H. Gao, J. Sengupta, M. Valle, A. Korostelev, N. Eswar, S. M. Stagg, P. Van Roey, R. K. Agrawal, S. C. Harvey, A. Sali, M. S. Chapman, J. Frank, Study of the structural dynamics of the e coli 70s ribosome using real-space refinement., Cell 113 (6) (2003) 789–801.
21. C. S. Potter, H. Chu, B. Frey, C. Green, N. Kisseberth, T. J. Madden, K. L. Miller, K. Nahrstedt, J. Pulokas, A. Reilein, D. Tcheng, D. Weber, B. Carragher, Leginon: A system for fully automated acquisition of 1000 electron micrographs a day., Ultramicroscopy 77(3-4) (1999) 153–61.
22. B. Carragher, N. Kisseberth, D. Kriegman, R. A. Milligan, C. S. Potter, J. Pulokas, A. Reilein, Leginon: an automated system for acquisition of images from vitreous ice specimens, J. Structural Biology 132 (2000) 33–45.
23. V. Abrishami, A. ZaldÃvar-Peraza, J. M. de la Rosa-TrevÃn, J. Vargas, J. Otón, R. Marabini, Y. Shkolnisky, J. M. Carazo, C. O. S. Sorzano, A pattern matching approach to the automatic selection of particles from low-contrast electron micrographs(under review).
24. R. Langlois, J. Frank, A clarification of the terms used in comparing semi-automated particle selection algorithms in cryo-em., J Struct Biol 175 (3) (2011) 348–352. doi:10.1016/j.jsb.2011.03.009.
25. T. Bharat, D. Zbaida, M. Eisenstein, Z. Frankenstein, T. Mehlman, L. Weiner, C. O. S. Sorzano, Y. Barak, S. Albeck, J. A. G. Briggs, S. G. Wolf, M. Elbaum, Variable internal flexibility characterizes the helical capsid formed by agrobacterium vire2 protein on single-stranded dna, Structure(in press).
26. C. O. S. Sorzano, R. Marabini, , J. Velázquez-Muriel, J. R. Bilbao-Castro, S. H. W. Scheres, J. M. Carazo, A. Pascual-Montano, XMIPP: A new generation of an open-source image processing package for electron microscopy, J. Structural Biology 148 (2004) 194–204.
27. C. O. S. Sorzano, E. Recarte, M. Alcorlo, J. R. Bilbao-Castro, C. San-Martín, R. Marabini, J. M. Carazo, Automatic particle selection from electron micrographs using machine learning techniques., J Struct Biol 167 (3) (2009) 252–260. doi:10.1016/j.jsb.2009.06.011.
28. Z. Yang, J. Fang, J. Chittuluru, F. J. Asturias, P. A. Penczek, Iterative stable alignment and clustering of 2d transmission electron microscope images., Structure 20 (2) (2012) 237–247. doi:10.1016/j.str.2011.12.007.
29. M. van Heel, Angular reconstitution: *A posteriori* assignment of projection directions for 3D reconstruction., Ultramicroscopy 21 (1987) 111–124.
30. P. A. Penczek, J. Zhu, J. Frank, A common-lines based method for determining orientations for N¿3 particle projections simultaneously, Ultramicroscopy 63 (1996) 205–218.
31. A. Singer, R. R. Coifman, F. J. Sigworth, D. W. Chester, Y. Shkolnisky, Detecting consistent common lines in cryo-em by voting, J. Structural Biology 169 (2010) 312–322, (under review).
32. J. M. De la Rosa-Trevín, J. Otón, R. Marabini, A. Zaldívar-Peraza, J. Vargas, J. M. Carazo, C. O. S. Sorzano, Xmipp 3.0: one step forward in scientific computing for electron microscopy, J. Structural Biology(in press).