Open Journal of Statistics, 2015, 5, 820-836
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.57081

Structure Sorting of Multiple Macromolecular States in Heterogeneous Cryo-EM Samples by 3D Multivariate Statistical Analysis

Bruno P. Klaholz[1,2,3,4]

[1]Department of Integrated Structural Biology, Centre for Integrative Biology (CBI), IGBMC (Institute of Genetics and of Molecular and Cellular Biology), Illkirch, France
[2]Centre National de la Recherche Scientifique (CNRS), Illkirch, France
[3]Institut National de la Santé et de la Recherche Médicale (INSERM), Illkirch, France
[4]Université de Strasbourg, Strasbourg, France
Email: klaholz@igbmc.fr

Received 28 October 2015; accepted 27 December 2015; published 30 December 2015

Abstract

Heterogeneity of biological samples is usually considered a major obstacle for three-dimensional (3D) structure determination of macromolecular complexes. Heterogeneity may occur at the level of composition or conformational variability of complexes and affects most 3D structure determination methods that rely on signal averaging. Here, an approach is described that allows sorting structural states based on a 3D statistical approach, the 3D sampling and classification (3D-SC) of 3D structures derived from single particles imaged by cryo electron microscopy (cryo-EM). The method is based on jackknifing & bootstrapping of 3D sub-ensembles and 3D multivariate statistical analysis followed by 3D classification. The robustness of the statistical sorting procedure is corroborated using model data from an RNA polymerase structure and experimental data from a ribosome complex. It allows resolving multiple states within heterogeneous complexes that thus become amendable for a structural analysis despite of their highly flexible nature. The method has important implications for high-resolution structural studies and allows describing structure ensembles to provide insights into the dynamics of multi-component macromolecular assemblies.

Keywords

Heterogeneity, Structural Biology, Cryo Electron Microscopy, Particle Sorting, Multiple States, Macromolecular Complexes, Resampling, Jackknifing, Bootstrapping, Multivariate Statistical Analysis, 3D MSA, 3D-SC, Ribosome, RNA Polymerase

How to cite this paper: Klaholz, B.P. (2015) Structure Sorting of Multiple Macromolecular States in Heterogeneous Cryo-EM Samples by 3D Multivariate Statistical Analysis. *Open Journal of Statistics*, **5**, 820-836.
http://dx.doi.org/10.4236/ojs.2015.57081

1. Introduction

Most structural biology methods such as crystallography, nuclear magnetic resonance (NMR), small-angle X-ray or neutron diffraction (SAXS, SANS) and cryo electron microscopy (cryo-EM) rely on signal averaging. For example, X-ray diffraction results from the in-phase contribution of each molecule packed in a regular array, but structural variability of the molecules in the crystal will lead either to local disorder (multiple side-chain conformations, disordered loops or even domains in protein or single-stranded RNA/DNA) or in the worst case, will prevent crystallisation at all; on the other hand, crystallisation may also select only one state out of many functional states present in solution. Similarly, molecules in solution will contribute constructively to small-angle diffraction in SAXS/SANS or to unique chemical shifts in NMR experiments only if they are all in the same structural state. Multiple structures in SAXS data may make the *ab initio* shape determination of the molecule difficult; e.g. in the case of size variability, the larger molecules will dominate the SAXS data (see for example [1] [2]). Cryo-EM and 3D reconstruction of single molecules (e.g. [3] [4]) are also based on the assumption that all molecules to be averaged are identical. However, in many cases this assumption is only very approximate, thus limiting the resolution of 3D reconstructions to the common core features of the slightly different molecules. Heterogeneity is caused by particle populations that display structural variability, for example, at the level of composition or of conformational states: 1) a protein factor bound to a molecular machinery such as the ribosome or the RNA polymerase may have a sub-stoichiometric occupancy (and hence be hardly visible in experimental maps); or 2) a complex may adopt different conformational states, even with a full occupancy of all components. Structural variability between molecules may thus make correct map interpretation difficult or impossible due to ill-defined maps, or at least limit the obtainable resolution of the cryo-EM map.

The problem of sample heterogeneity can in principle be overcome by accounting for discrete sub-states through sorting of individual sub-populations. The visualization of distinct structural states within a biological sample requires the ability to separate the states present in the experimental data, based on some objective criterion or measured value, and to quantify the percentage of each state within the biological sample. The most convenient way will be to sort states based on their 3D variance, but this will require experimentally recorded 3D structures (e.g. by cryo electron tomography, but usually not possible for individual 2D projections) or possibly individual 3D states determined from different biochemically highly homogeneous and well-defined functional states. The difficulty of handling heterogeneous single particle cryo-EM data comes from the fact that the structure needs to be reconstructed from 2D images (transmission electron microscopy provides 2D projection images of the 3D object) rather than observing directly 3D objects. It therefore makes it difficult to link the structural state enclosed in a projection with that of a structure reconstructed from several hundred, thousands or hundred thousands of 2D images that may differ in their structural (and functional) state. Moreover, the relatively weak signal to noise ratio (SNR) of single particle images imposes some degree of averaging in order to enhance the signal, but this requirement stands in contradiction with the potential risk that the particle images may not describe a unique 3D object. Averaging occurs during image processing of single particle cryo-EM data at the level of 1) formation of class averages; 2) 3D reconstruction. Class averages 1) describe similar angular views of the projected object and are formed by merging and summing similar images into a single image according to their lowest intra-class variance by treating the data with classical two-dimensional multivariate statistical analysis, 2D MSA, and hierarchical ascendant classification [5]-[9]. The other step of image processing that involves averaging is the 3D reconstruction; 2) using back-projection algorithms [10]-[12]. The 3D reconstruction from single particle views assumes the same structure visualised in different (ideally random) orientations. Although preferential (non-evenly distributed) molecular views can be down-weighted to some extend by *weighted* back-projection algorithms [10] [11] [13], the problem of particle variability can be difficult to address [14]. Attempts have been made in the past to address this by using different approaches: 1) reference-based techniques resembling the method of molecular replacement in X-ray crystallography (structure determination by using a related, known structure as starting model; [15]); and 2) methods based on statistical analysis. In the first case, usually two known 3D structures are used as external references (templates) to sort particle images according to the best correlation with re-projections of each 3D structure (e.g. ref. [16], also termed “supervised classification” even though the method is template-based and not MSA-based); such multi-reference refinements can be significantly improved using a maximum-likelihood (ML) approach [17]-[19]) as illustrated by recent high-resolution structures [20]-[22]. Reference-based methods require independently determined structures (less though for maximum-likelihood methods), and they are intrinsically prone to be in some way biased against the starting structure

models, which may in fact be different from the actual structures in the sample.

Alternative particle sorting methods that do not require a set of distinct starting references have been developed that either rely on direct 3D reconstructions from tilted single molecules [23], perform unsupervised classification (using MSA and K-means clustering [24]; or graph cutting and common line methods [25] [26], or stochastic climbing [27] [28], or covariance analysis [29], or ML-derived variance maps [30] or merely make use of the statistical fluctuations of intensities that are often visible in restricted areas of 2D particle images. Indeed, large conformational or structural changes can be detected and localized quite easily within a set of class averages that describe similar angular views (e.g. first line in **Figure S1**). An approach has been described that consists of 2D MSA of selected, variable sub-regions of the molecular images (referred to as "local MSA" [31] or as "focused classification" [32]) using a re-projected mask for MSA ([31]; see **Figure S1**) or for cross-validation [32]. The application of local rather than only global MSA allows putting the attention of the analysis on the variable regions and thereby improves the classification quality, *i.e.* separates the issues of orientational and conformational classification. Using a small mask placed on the region of interest helps sorting particles with similar viewing angles, and thus the variability becomes even more apparent because different particle states get less mixed within a given class average. The efficiency comes from the fact that two successive classifications are performed: 1) the global MSA for classification according to particle orientations (*i.e.* classical class averages); 2) the local MSA for classification according to particle variability. It consists in a reclassification that allows distinguishing orientational and conformational classes [31] [33] (referred to as "double-MSA" by [34] [35]; global MSA can also be followed by K-means clustering [24]). Class averages describing similar molecular views are assigned to two (or more) different data sets (**Figure S1(b)**), and the set assignments are cross-validated by calculating the correlation coefficient between re-projections of the two structures obtained and the input images of the corresponding sets [31] [32]. Such analysis also provides direct evidence of inter-correlated movements inside the structure (**Figure S1(a)**). Local 2D MSA and related approaches have for example allowed separating particles of slightly different size [33], composition [36] [37] or of strikingly different conformation [31]. Such analysis has also worked out for other cases [32] [34] [38], but this approach may also harbour some intrinsic limitations: 1) it usually requires user-knowl- edge of the structure because some typical molecular views are needed to visually detect structural heterogeneity; 2) it harbours the problem of assigning a particle image to a precise group (*i.e.* one structural state or another) across different viewing angles (*i.e.* across the different lanes in **Figure S1(a)**) of the object (this can be addressed in part by automatic iteration of the above-mentioned cross-validation with re-projections [31] [33]); 3) importantly, the procedure is difficult to extend to more than two different states. Previously, an independently developed bootstrapping approach has been described for estimation of 3D variance [39] [40]. Here, a related approach is described that however extends beyond variance estimation and includes direct 3D classification and structure refinement, thereby addressing most of the above mentioned issues and allowing the analysis of structural or conformational variability of 3D structures. Indeed, a conceptually very powerful approach would be a direct separation of states based on statistical analysis of 3D structures, provided that these are already available in different states. Although this sounds like a "hen-and-egg" problem, there is a potential solution: rather than including all particles into a 3D reconstruction, only sub-ensembles are used. This allows a statistical analysis and classification of 3D structures generated from random, small sub-populations of the experimental data. The method described in the following is referred to as 3D sampling and classification (3D-SC) because it is based on random resampling (selection) of sub-ensembles of object populations and their sorting into 3D classes (**Figure 1** & **Figure 2**). The 3D-SC procedure has been successfully tested with heterogeneous test data (where the composition of the mixture is known, see below; **Figure 3** & **Figure 4**), it has served to process and sort the experimental data of a heterogeneous 30S ribosomal complex (**Figure 5**) and of other structures, and it is applicable to sorting of heterogeneous biological structures in general.

2. Results

The procedure described herein makes use of a fundamental principle in statistical analysis according to which resampling allows estimating statistical parameters from those of sub-populations (sub-samples [41]-[43]). Resampling (unrelated with the term "resampling" in image processing which refers to pixel size change such as coarsening) involves the selection of small subsets in statistical analysis, which, when applied to the case of 3D structures, allows describing the variability of 3D objects by resampling structures that are obtained from par-

ticle sub-ensembles. The implementation within the 3D-SC procedure relies on jackknifing (selection of small subsets) and bootstrapping (see concept in **Figure 1**; this includes a random selection of small subsets, part of

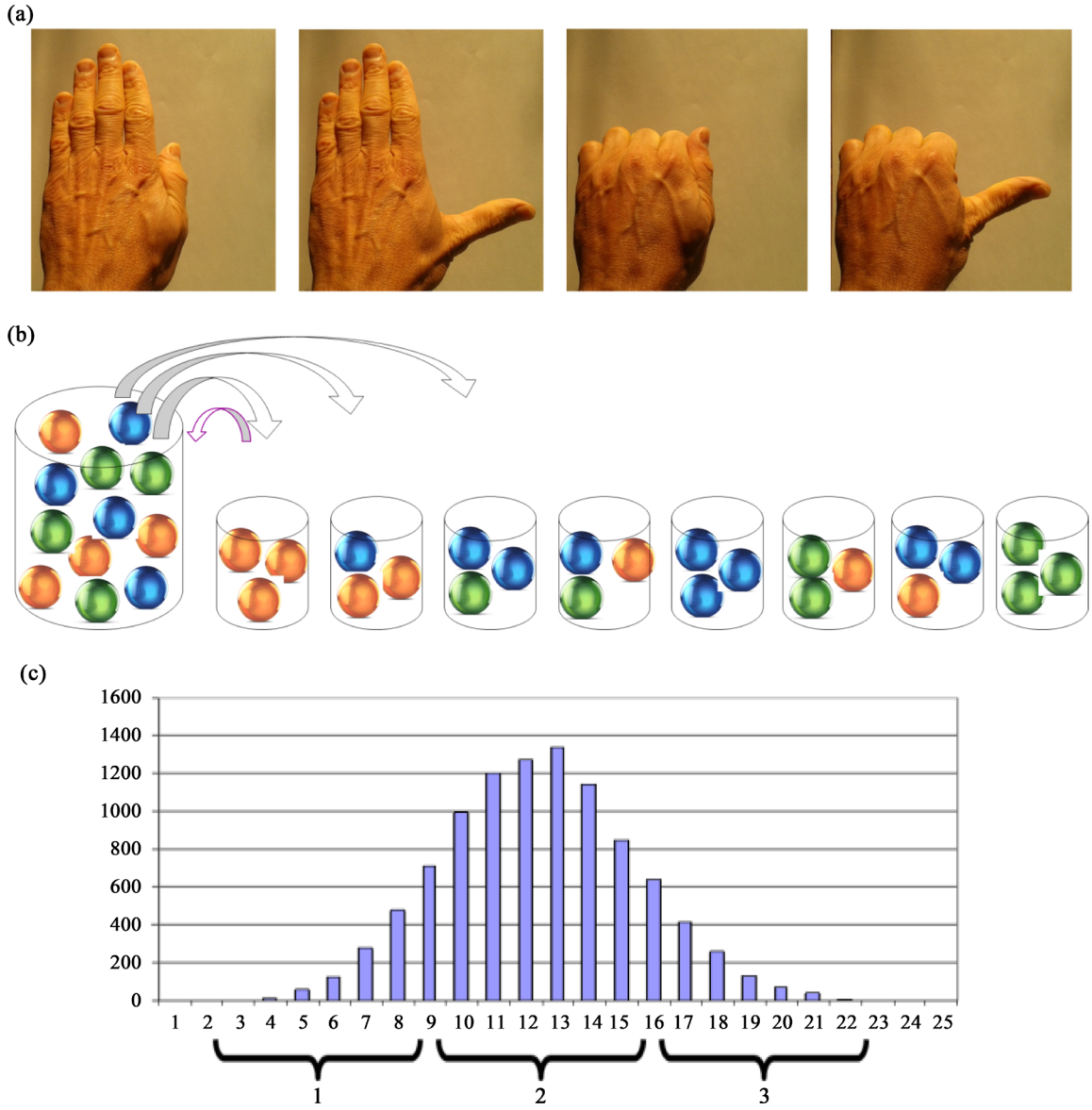

Figure 1. Variable structures and concept of resampling (jackknifing and bootstrapping). (a) Illustration of variable structures; the hand shows four different conformations. In an image data set, these different "structural states" would be mixed and require to be separated without *a priori* knowledge. Resampling allows identifying sub-populations more easily. (b) Concept of the resampling method: spheres with 3 different colours represent 3 different structures which are mixed inside a sample (big pot on the left). Selection of small subsets (jackknifing, little pots) comprises a random selection of spheres some of which can be re-selected, *i.e.* resampling with replacement (bootstrapping); the repeated random resampling is a Monte Carlo approach. The statistical distribution of spheres over the little pots will lead to individual compositions different from the average composition in the stock pot. (c) Histogram of the particles originating from one state within a mixed population (RNA polymerase model data set). Displayed is the number of images belonging to state 1 (the DNA-bound complex) when 50 images are selected randomly from a known mixed population containing 25% state 1 and 75% state 2 (abscissa) and their absolute occurrence (ordinate) when 10,000 such random selections are performed. The expected mean value for state 1 would be 12.5 images; however, the random selection (jackknifing/bootstrapping) produces some subsets (areas 1 and 3) that are significantly different from the average distribution (area 2), and that can be detected by 3D MSA (which is the key for particle sorting when the composition of the mixed population is not known, such as in experimental data).

which can be re-selected). Jackknifing and bootstrapping are part of resampling methods [41]-[46]. In 3D-SC, a sub-ensemble or subset of particle images is selected (jackknifing [41]), and the images can be selected more than once (**Figure 1(b)**), *i.e.* different subsets may contain some common or permutated images (but no re-selection within a set), which is resampling with replacement (bootstrapping [42] [43]). The 3D-SC approach thus combines two concepts: the resampling with replacements and the usage of random sub-ensembles, the repeated random resampling being a Monte Carlo approach. The random selection of particle images from a mixed population provides a statistical distribution in which some subsets are significantly different (areas 1 and 3 in **Figure 1(c)**, example for a pre-defined mixture) from the average distribution (central area in **Figure 1(c)**); these sets can be classified by 3D MSA and thus allow sorting of a heterogeneous data set into more homogeneous subsets.

The 3D-SC procedure comprises 4 steps (**Figure 2**) and consists in the generation of 3D reconstructions from small sets of randomly selected experimental molecular views, thus providing a 3D sampling of structure sub-ensembles; the full data set comprising all reconstructions is then analysed by 3D MSA, and similar 3D structures are clustered into different sets. Finally, the 3D-SC procedure is completed with a high-resolution refinement of the subsets (see **Figure 2(b)**):

1) Step I: numerous, but small subsets of particle images are selected randomly from the total data set, and a 3D reconstruction is calculated for each subset resulting in a large number of statistically variable 3D structures (~10^3 - 10^4; see scheme in **Figure 2(b)**).

2) Step II: 3D MSA is performed on the large set of variable 3D reconstructions generated in Step I, and the reconstructions are clustered into distinct groups by 3D classification and hierarchical ascendant classification, resulting in 3D class averages in which similar 3D structures are merged.

3) Step III: it consists of multi-reference alignment (MRA) of the original particle images using re-projections of the 3D class averages as references, and splitting into sets according to the best correlation with a given reference image; 3D reconstructions are calculated for the individual sets and subjected to a new round of 3D MSA and classification.

4) Step IV: it iterates Steps II and III 2 - 3 times, and the number of sets is gradually reduced from several hundred 3D reconstructions to several dozens and finally to a small number of 3D sets (e.g. 4 - 6, depending on the degree of variability within the sample, see below).

5) Step V: it consists of a high-resolution structure refinement of each individual subset.

Step I of the 3D-SC procedure allows sampling the particle images and therefore—in combination with 3D reconstruction—consider multiple 3D structures (much beyond two). Steps I-IV are usually run with coarsened image data (e.g. 3 - 10 Å/pixel) to speed up the computational processing, while the structure refinement of the individual particle subsets in step V uses less coarsened or uncoarsened data to preserve high frequency information and refine the structure to highest resolution. During Step I, the resampling with replacements will allow particle images to be selected several times, but placed within particle sets of different, random composition. The choice of the amount of particle images included per 3D reconstruction takes into account the following considerations: a) it should be as low as possible (ideally one, but conceivable only for some highly symmetric objects such as viruses) in order to avoid merging too many particles from different structural states; b) on the contrary, a reasonable distribution of Euler angles for a good quality 3D reconstruction can be obtained more easily when more particles are included. The requirement for a non-preferential angular distribution within the selected particles imposes a certain degree of averaging (e.g. 10 - 50 particle projections per 3D in the case of the asymmetric objects analyzed in this study) in order to obtain 3D reconstructions of sufficient quality for feeding into 3D MSA (Step II). 3D reconstructions with too few particles or too many preferential views will give rise to distortion artefacts in the maps, but this is not a problem because these will be detected easily in the 3D classification step and can be excluded from further refinement. To perform 3D MSA, consecutive 2D sections of a 3D reconstruction were represented as a single rectangular image (e.g. **Figure 3(a)**; see Experimental Procedures) that can be handled directly by standard 2D image processing routines.

Convergence of the 3D classification is obtained by combining MRA (which helps refining the individual structures) and 3D MSA (Step IV, iteration of Steps III and II): particle images are aligned against re-projections of all 3D class averages, split into sets according to the best correlation with a given reference, a 3D reconstruction is calculated for each set, and the 3D reconstructions from all sets together are subjected again to 3D MSA and classification (**Figure 2(b)**). During Step IV the number of sets is gradually reduced from several hundred 3D reconstructions to a smaller number that describes the multiple states best. The choice of the number of final conformers/structures is not a predefined value because it depends on the variability of the sample and the

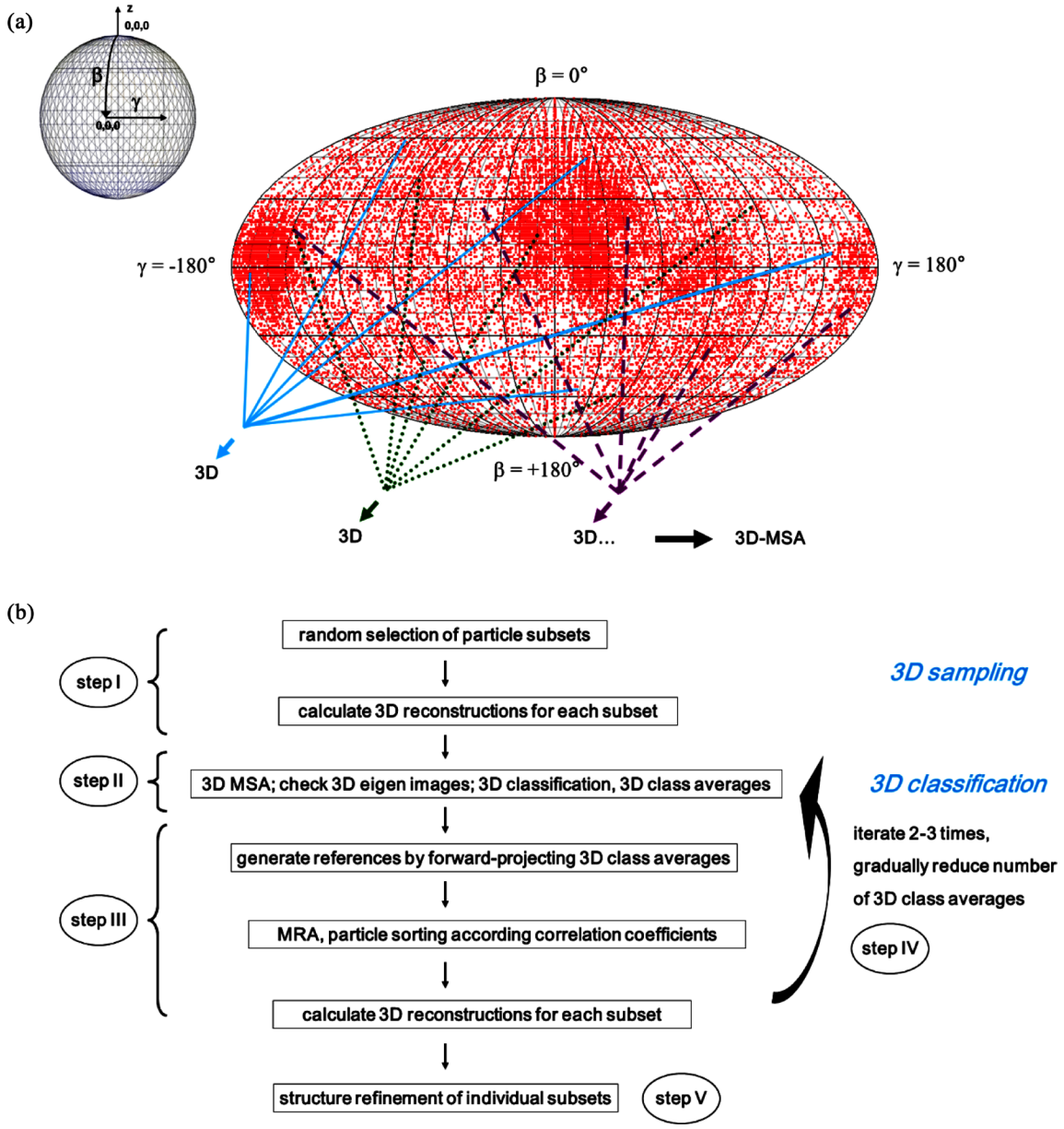

Figure 2. Principle of the 3D sampling and classification (3D-SC) approach. (a) Typical Euler angle distribution of particle images (example of a 30S ribosomal complex [47]; each red dot represents an image with Euler angles β and γ as defined in the top left insert) from which random subsets of 20 - 50 images are selected (blue, green and magenta lines). A 3D reconstructions is calculated for each subset and then subjected to 3D MSA. (b) Flowchart of the 3D-SC procedure. The 3D reconstructions generated from random particle subset selections in step I are subjected to 3D-MSA and 3D classification (Step II), and 3D class averages are formed. These 3D reconstructions (now with an enhanced signal to noise ratio) are used as references for multi-reference alignment (MRA) and particle separation based on correlation (Step III), and 3D reconstruction obtained for each subset are subjected together to 3D MSA. Iteration of Steps III and II allows gradually reducing the number of 3D class averages (see main text). In contrast to (local) 2D MSA (see **Supplementary** Figure S1), 3D-SC comprises 3D MSA and 3D classification on 3D structures and is performed on the entire map.

degree and resolution to which differences can be resolved.

To test the robustness of the 3D-SC approach we used both model and experimental data. Heterogeneous model data containing a pre-defined mixture were derived from the crystal structure of an RNA polymerase II complex [48] in which a 25% occupancy of the DNA ligand was simulated (the ligand represents only 0.7% of the total molecular mass of the complex; see Experimental Procedures). The 3D eigen images reveal a variability

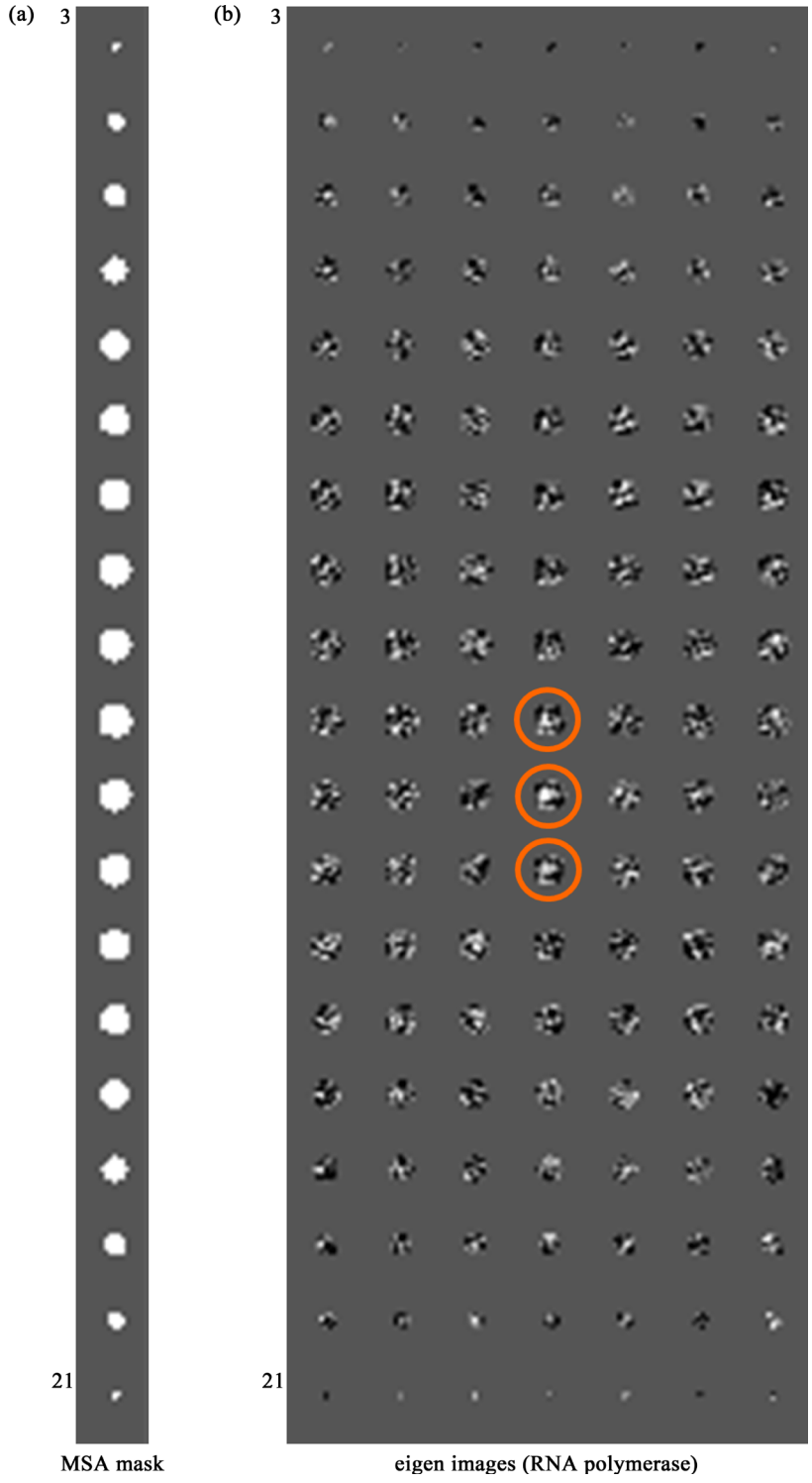

Figure 3. 3D MSA mask and 3D MSA eigen images of the model data set. (a) 3D MSA mask used for the 3D MSA analysis of the model data set (RNA polymerase structure in which heterogeneity is modelled with 25% of DNA occupancy; consecutive 3D sections are represented as single rectangular images; for clarity, only the core sections 3 - 21 out of 42 sections are shown), (b) 3D eigen images of the RNA polymerase heterogeneous model data set (mixed +/− DNA component, see Experimental Procedures; same sections as for (a); for clarity, only eigen images 14 - 20 are shown). Eigen image 17 reveals a variable area (marked with orange circles) in the position expected for the DNA ligand inside the RNA polymerase for which sub-stoichiometric binding is modelled.

located in the expected area of the model structure (**Figure 3** & **Figure 4**), visible already in round one of Step II of the 3D-SC procedure. This is found in a relatively high-order 3D eigen image because of the weak difference between the +/− DNA complexes. The detection of the weak ligand signal within the large RNA polymerase complex shows that 3D MSA is very sensitive for the detection of heterogeneity. Similarly, the eigen images produced from experimental data of the 30S ribosome complex provide several locations of variability (**Figure 5(a)**); accordingly, the 3D reconstructions can be clustered into groups (3D class averages with low intra-class variance) that describe distinct structures (presence/absence and conformational variability of components of the complex; **Figure 5(b)**). The low SNR of the individual 3D structures can be handled easily by 3D MSA, in fact even better than in the case of 2D MSA because the SNR of 3D structures is already much improved through the averaging of particles into the 3D reconstruction (through back-projection) compared to that of the individual input 2D images.

The variability in the structure (the 3D variance) can be addressed by statistical analysis of the subsets: the classification of classes (transition from round 2 to 3 of Step IV, e.g. the classification of 20 structures into 5 classes) provides 3D eigen images that directly illustrate and describe the 3D variability (**Figure 5(a)**). Furthermore, the 3D variance maps are then used right away to classify the 3D structures and obtain 3D class averages. The direct separation of states by 3D MSA is a powerful and reliable approach as illustrated by the results obtained from the model data. 3D-SC was able to detect the weak signal of the simulated DNA density, which represents less than 1% of the total mass of the complex. For real experimental data such as ribosome complexes, ligand-induced conformational changes of the overall structure will increase the 3D variability and thus facilitate the particle sorting procedure. Applied to the data sets of a 30S translation initiation complex this procedure allowed separating different states according to their statistical three-dimensional variability; the differences indeed reside in either the composition (presence or absence of translation initiation factor IF2 for example) or the conformation of components or of the entire ribosome (**Figure 5(b)**). The strength of the procedure is also illustrated by the improved resolution of each individual 3D reconstruction compared to a 3D reconstruction using all particle images as a single set, in spite of the smaller data size/particle number of each individual set as compared to the total data set.

3. Discussion

Compositional or conformational heterogeneity of biological samples is a major bottleneck towards their high-resolution structural analysis. This is particularly true in single particle cryo-EM where experimentally recorded 2D images of physically different molecules are averaged into a 3D reconstruction, implying a structural uniqueness which is never the case if one considers the multi-scale aspects that range from large conformational changes to atomic details such as alternative side-chain conformations. Here we present a method, called 3D-SC, which allows separation of states based on MSA and classification of 3D structures, following a random resampling (jackknifing and bootstrapping) of 2D image particle subsets into 3D structures. The concept is to, rather than including all particles into a 3D reconstruction, use small particle sub-ensembles which will differ statistically and thus allow performing particle sorting through classification of 3D structures. 3D-SC allows the *ab initio* separation of multiple structural states within a biological sample observed by cryo-EM imaging and thus can address the intrinsic structural heterogeneity and dynamics of large, multi-component macromolecular complexes directly in 3D. It provides starting structures straight from the biological sample that can then be refined separately (Step V of the procedure) without requiring initial reference structures. The reliability and robustness of the approach was confirmed by processing both model and experimental data, and has allowed studying the structure of the 30S initiation complex that was otherwise not amendable to structure determination because of its highly flexible and heterogeneous nature [47]. This method is now implemented in recent versions of the software suite IMAGIC-V [49]. A strength of the procedure is that full 3D maps are analysed without *a priori* knowledge rather than for example using pre-defined 2D or 3D areas. Moreover, the representation of 3D class averages and eigen images as rectangular 2D images is particularly convenient for displaying and comparing 3D structures and 3D eigen images (**Figures 3-5**). Finally, 3D-SC is not restricted to large data sets (which one would expect to benefit from better statistics) but it works also on relatively small data sets (few hundreds to thousands of particles) because resampling with replacements (Step I of the 3D-SC procedure) will allow particle images to be selected several times within the numerous small random particle sets. Similar resampling or bootstrapping methods have been used for image processing (for example [50] [51]), in particular for the 3D local-

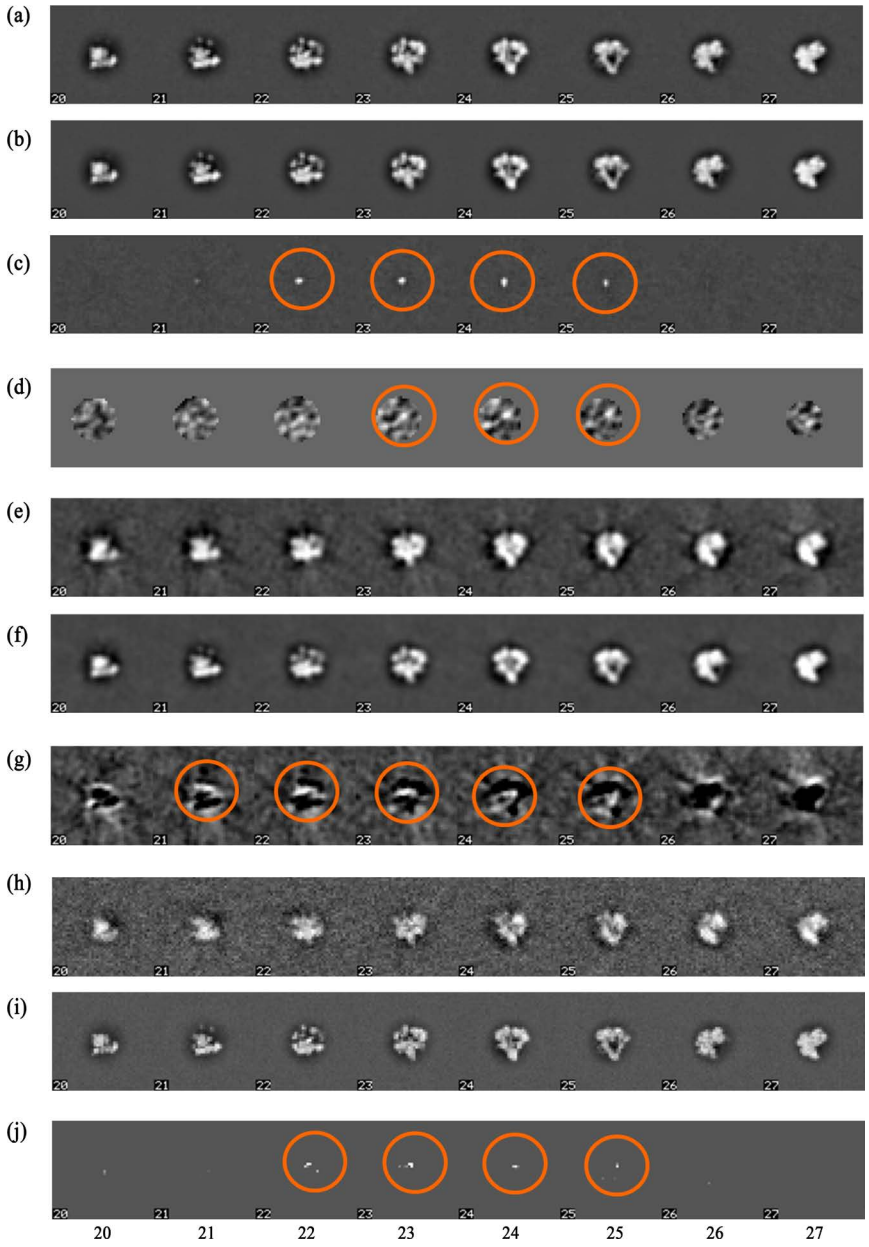

Figure 4. 3D sections of RNA polymerase model structures treated by 3D-SC. Sections are displayed in the horizontal direction (for clarity, only the core sections 20 - 27 out of 42 sections are shown). Model data set with random translations/rotations of the input projections (to imitate real data) and refinement of a heterogeneous structure (see Experimental Procedures) before starting 3D-SC. (a)-(c) Initial RNA polymerase maps derived from the crystal structure of the RNA polymerase II elongation complex (PDB code 1Y1W [48]) filtered to 10 Å resolution in which the DNA part (modelled as carbon atoms; see Experimental Procedures) was included (a) or removed (b). The density corresponding to the DNA part can be seen in the difference map between these structures (c), the sections with the density corresponding to the DNA part are marked with orange circles). Re-projections of the DNA-containing model complex (state 1) and the DNA-free model complex (state 2) were mixed 1:3 for generating a heterogeneous model data set (see Experimental Procedures). (d) 3D eigen image 46 of the first 3D MSA run (step II) reveals a variable area (in the position expected for the DNA). (e)-(g) First structures 1 (e) and 3 (f) obtained after a first run of step IV (from the 5 structures obtained, 3 were ill-defined and were therefore not kept for structure refinement); (g) Difference map between structures in (e) and (f) already shows some densities in the expected sections; (h)-(j) Structures after refinement (step V; refined in 128 pixel boxes, but scaled to 42 for displaying purposes). Structure 1 ((h), contains DNA; the structure is less well defined than structure 2 because of the model data set size which is 3x smaller) and structure 2 ((i), no signal for DNA part); (j) Difference map of the fully refined structures corresponding precisely to the area expected for the DNA part (compare with lane (c)).

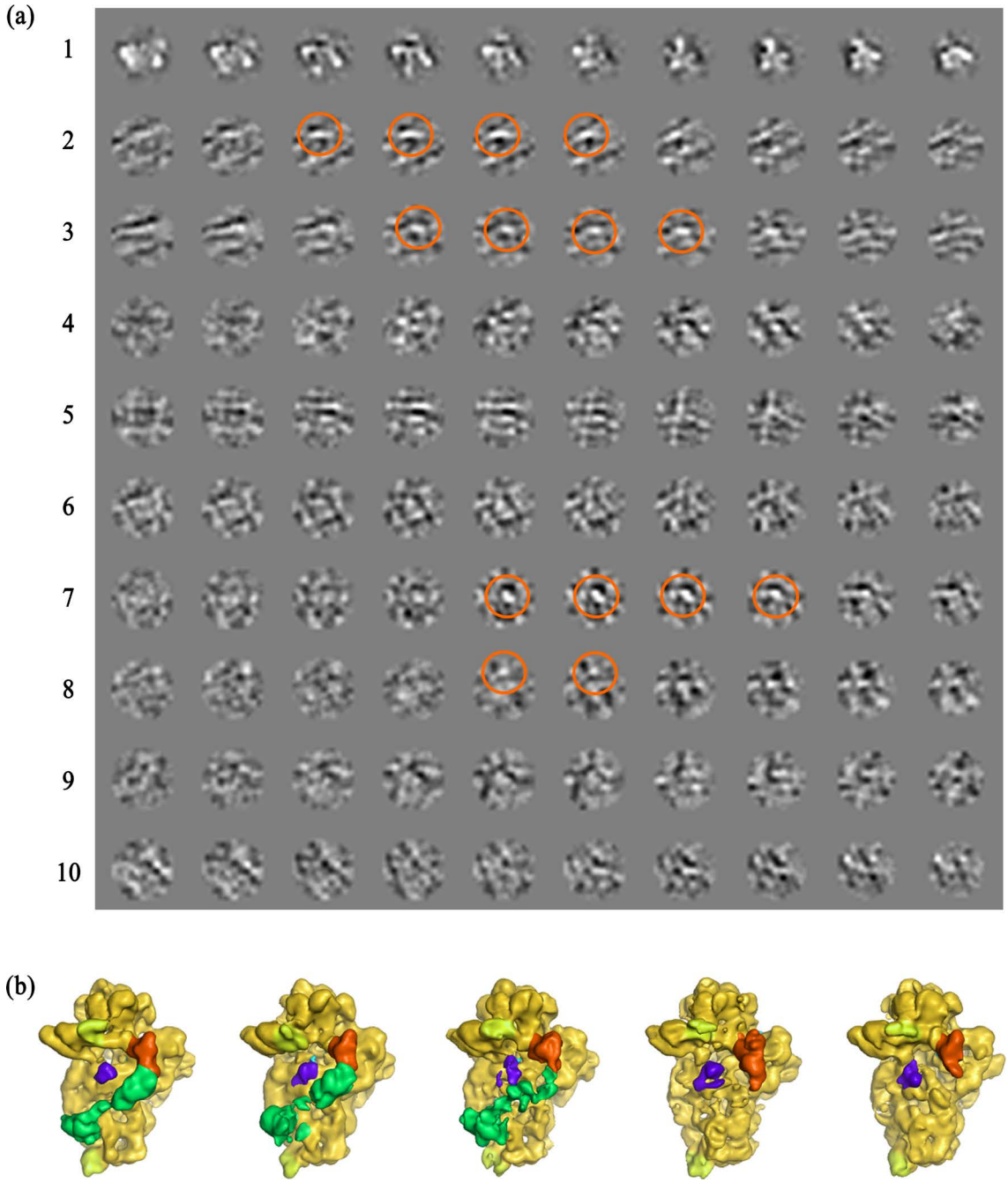

Figure 5. Analysis of an experimental data set (a 30S ribosomal subunit complex). (a) Sections of the 3D eigen images 1 - 10 obtained with the first 3D MSA, showing the performance of the 3D-SC procedure (step II, see **Figure 1**; for clarity, only the core sections 19 - 28 out of 42 sections are shown). The 3D eigen images directly highlight the 3D variability of the 30S initiation complex [47]. Variable areas are marked with orange circles; their positions correspond to regions in the structure with variable ligand occupancies (translation initiation factor IF2) or with conformational changes of the tRNA or the ribosome. (b) Surface representations of the multiple states of the 30S/IF1/IF2 initiation complex (refined structures [47]). Densities corresponding to 30S, fMet-tRNA^{fMet}, IF1 and IF2 are colored in orange, red, blue and green, respectively; conformational changes of the 30S subunit beak and toe area are highlighted in light green.

isation of variable areas [39] [40] but through the generation of notably fewer sub-ensembles (~100) while 3D-SC uses ~10^3 - 10^4 resampled structures to sort out many states simultaneously. 3D-SC includes both bootstrapping/resampling and 3D classification, *i.e.* resampling is used in direct combination with 3D MSA and 3D

classification of structures. Like for bootstrapping [39], 3D variance maps are also obtained from the 3D-SC procedure which are indeed useful for the identification of high-variance regions within a structure and which can be used for defining an MSA-mask for local MSA procedures [31] [32], a strategy useful for the analysis of inter-correlated conformational changes and which has recently been taken over for the ML-based focused high-resolution refinement of structures with flexible regions [52].

An interesting parallel between 3D-SC and normal 2D classification can be drawn in the following manner: when particle images are merged into 2D class averages, they can also be extracted from there to be re-classified, *i.e.* change their annotation to a given class. The statistical approach of jackknifing and resampling with replacement (bootstrapping) within the 3D-SC procedure thus allows direct re-classification by assigning particles to different subsets from which 3D reconstructions and eventually distinct 3D classes are formed. A potential limit however lies in the fact that even after some iteration of the classification (Step IV) sub-ensembles will not be entirely homogeneous due to the statistical nature of the procedure. However, the structure refinement of subsets (Step V) that include the high frequency data contributes to converging towards better sorted subsets, which can be refined individually to high resolution, and eventually reclassified through 3D-SC to refine the assignment of particles to specific sub-ensembles. In this context, it is important to note that 3D-SC is not limited by the number of different structural states in the sample. 3D-SC therefore opens the possibility to simultaneously study many multiple states which is very interesting because within a given biological sample these stand in equilibrium with each other. Thus, sample heterogeneity turns into an advantage by providing structure ensembles that describe multiple functional states. Provided that a sufficient amount of image data is available (which is only a matter of extensive data collection), individual subsets can be refined to high-resolution structures without being much limited by the homogeneity of the biological sample. This is illustrated as well by recent ML-based classification approaches [20] [53] during which random subsets are optimized and a low-resolution average structure is used as reference, *i.e.* resampling is used in combination with likelihood optimization [54]. Structure sorting has major implications for high-resolution studies and enables a more detailed structure, function and dynamics analysis of large macromolecular complexes. Applications of 3D-SC include the verification of sample homogeneity as done for a nuclear receptor DNA complex [55] and the structural sorting of transcription factor TFIID [56]. The concept of 3D-SC described here and of the general bootstrapping approach have been used for the sorting of numerous sub-states [47] [57] and for the estimation of equilibrium constants between different sub-populations [58], and for the sorting of ribosome states within poly-ribosomal samples [59]. 3D-SC may also be very powerful for the analysis of virus structures for which starting structures can in principle be determined by using a single or very few molecular views (*i.e.* essentially no merging in Step I), and thus different states would be easily detectable in Step II. 3D MSA could also be used for the classification of particles extracted from tomographic reconstructions, provided that the sub-tomograms have been aligned with respect to each other (in contrast to 3D-SC in which 3D reconstructions are *de facto* aligned because the Euler angles are defined for all particle images according to a unique coordinate system).

In the current prospect of high-resolution cryo-EM which can increasingly provide maps in the 3 Å resolution range or better at which atomic models can be built and refined like in X-ray crystallography, it will be interesting to see the role of particle sorting procedures such as 3D-SC and ML-based approaches. Typical examples of this are studies such as particle sorting of tens of sub-structures [58] or the structure determination of small sub-populations [53]. Once a general classification into major conformational populations has been obtained, 3D-SC can also be applied to subsets to search for finer differences. Eventually, fine details such as differences of side-chain conformations may become visible as for high-resolution crystal structures, with promising implications in the molecular and atomic analysis of macromolecular complex function. Beyond considerations within the electron microscopy field, the number of states in macromolecular complexes is conceptually very large; resolving these states depends only on how far structural differences can be distinguished: low-resolution/large-scale differences of composition or conformation at the level of 1) sub-complexes; 2) protein or RNA domains as seen by medium-resolution cryo-EM; 3) side-chain conformations for high-resolution crystal and cryo-EM structures. Therefore, the number of states that can be described is mostly limited by the resolution technically achievable during data collection and processing, either by cryo-EM or X-ray crystallography. It is interesting to mention in this context that in X-ray crystallography, NMR and molecular dynamics simulations, ensemble refinements with multiple copies are commonly used to address multiple states and model the experimental data more precisely [60] [61], just as multiple structures can describe the various structural states in cryo-EM image data, which implies that multiple atomic models can be derived accordingly.

4. Experimental Procedures

Model data were generated from the crystal structure of the full yeast RNA polymerase II elongation complex with bound DNA (PDB code 1Y1W [48]). Two models were generated including or excluding the DNA (when treated as C-α atoms, this represented 0.7% of the total mass of the complex, *i.e.* only 4 kDa within 587 kDa). The DNA part was handled as C-α atoms in order to obtain the same density as for the protein part (*i.e.* constant atom form factors; the purpose was to obtain a signal weaker than for normal nucleotides which contained phosphorus in order to test the sensitivity of the 3D-SC procedure). Coordinates were converted into density maps by using the IMAGIC-V software [49] with a sampling of 3 Å/pixel, and were filtered to 10 Å resolution. Model 2D image data were produced by projecting the maps along equally distributed directions with an angular sampling of 3˚ resulting in 4564 views for each set. 1000 images for the DNA-containing set and 3000 images for the empty complex were extracted randomly, normalized and covered with 10 σ Gaussian noise (final SNR = 0.1); the amount of empty complex was set to 75% in order to model typical experimental data in which ligands tended to dissociate. The two image sets were merged into a single, heterogeneous set (4000 images). A further set was generated by randomly rotating and translating these images in-plane; an average, heterogeneous structure was derived from this data set and was refined with AR and MRA following standard procedures [47] [55] [62] [63] in order to obtain a heterogeneous, initial average map. The data were coarsened 3x for Steps I-IV of the 3D-SC procedure in order to speed up the process, and un-coarsened data were used for high-resolution refinement with MRA and projection matching performed in IMAGIC-V. Before Step I, Euler angles were assigned to the projections by projection matching against re-projections of the average structure, and subsets of 50 particle images were extracted randomly (using a random number generator in IMAGIC-V with a changing seed for each set). 3D reconstructions were calculated for 10,000 random subsets with the BKPR program, a fast voxel-based back-projection algorithm [12] comprising an exact weighting scheme based on Voronoi-diagrams applicable to all point group symmetries [13]. In order to perform 3D MSA and 3D classification, the sections of the individual 3D reconstructions were converted into rectangular 2D images by changing the number of lines per image to the square of the number of pixels per line (header value 13 in IMAGIC-V). MSA and hierarchical ascendant classification of these 2D images (resulting in high inter-class variance and low intra-class variance) were done with the standard routines in IMAGIC-V. The concept of 3D-SC was now implemented in the IMAGIC-V software. In Step IV, the number of 3D class averages was reduced from initially 160 to 20 and finally 5; out of these five structures two were ill-defined (#4 and #5) because of their small data size and they were discarded, and one more (#3) did not refine well and hence was removed during the refinement process (Step V); the resulting two structures were virtually identical to the two initial model structures derived from the RNA polymerase II crystal structure (**Figure 3**). Experimental data of the *T. thermophilus* 30S/IF1/IF2/GTP/mRNA/ fMet-tRNA^{fMet} complex [47] were treated by 3D-SC in the same way, with the difference that 15.000 initial particle images were used for 3D-SC before extending the data set during the high-resolution refinement to 80.285 single particle images; about 32.000/22.000/8.000/6.000/11.000 particles form sets 1 to 5, respectively; the complex that contains all components represents ~40% of the data (experimental details on the complex formation and the full structural and functional analysis of this complex were described in [47]). The 3D reconstruction part of the 3D-SC procedure requires a data set of aligned particle images to which Euler angles have been assigned beforehand by using forward projections of the initial, low-resolution average structure as an anchor set for angular reconstitution [64] [65] or as a reference for projection matching. Computing was performed on a PC farm running under the Linux operating system; MRA jobs were run as parallel routines using MPI on the PC farm or at the IDRIS computing center.

Acknowledgements

B.P.K. would like to thank Igor Orlov, Alexander Myasnikov, Angelita Simonetti, Stefano Marzi and Massimiliano Maletta for interesting discussions. The original concept of random selection of sub-ensembles and 3D classification was first presented at the 3D-EM Gordon Research Conference 2006 (“Local multivariate statistical analysis in 2D and 3D to handle structural heterogeneity”). This work was supported by grants from the Agence National pour la Recherche (ANR), the European Molecular Biology Organization (EMBO) Young Investigator Programme (YIP), the Institut du Développement et des Ressources en Informatique Scientifique (IDRIS, France), Centre National pour la Recherche Scientifique (CNRS), the European Research Council (ERC Starting Grant N_243296 TRANSLATIONMACHINERY), and by the French Infrastructure for Integrated

Structural Biology (FRISBI) ANR-10-INSB-05-01, and Instruct as part of the European Strategy Forum on Research Infrastructures (ESFRI).

Conflict of Interests

The author declares no competing financial interests.

References

[1] Roblin, P., Potocki-Véronèse, G., Guieysse, D., Guerin, F., Axelos, M.A., Perez, J. and Buleon, A. (2013) SAXS Conformational Tracking of Amylose Synthesized by Amylosucrases. *Biomacromolecules*, **14**, 232-239. http://dx.doi.org/10.1021/bm301651y

[2] Kikhney, A.G. and Svergun, D.I. (2015) A Practical Guide to Small Angle X-Ray Scattering (SAXS) of Flexible and Intrinsically Disordered Proteins. *FEBS Letters*, 589, 2570-2577.

[3] Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M. and, Leith, A. (1996) SPIDER and WEB: Processing and Visualization of Images in 3D Electron Microscopy and Related Fields. *Journal of Structural Biology*, **116**, 190-199. http://dx.doi.org/10.1006/jsbi.1996.0030

[4] van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R. and, Schatz, M. (1996) A New Generation of the IMAGIC Image Processing System. *Journal of Structural Biology*, **116**, 17-24. http://dx.doi.org/10.1006/jsbi.1996.0004

[5] Benzécri, J.P. (1969) Methodologies of Pattern Recognition. Academic Press, New York, 35-74.

[6] van Heel, M. and Frank, J. (1981) Use of Multivariate Statistics in Analyzing the Images of Biological Macromolecules. *Ultramicroscopy*, **6**, 187-194.

[7] van Heel, M. (1984) Multivariate Statistical Classification of Noisy Images (Randomly Oriented Biological Macromolecules). *Ultramicroscopy*, **13**, 165-183. http://dx.doi.org/10.1016/0304-3991(84)90066-4

[8] Borland, L. and van Heel, M. (1990) Classification of Image Data in conjugate Representation Spaces. *Journal of the Optical Society of America A*, **7**, 601-610. http://dx.doi.org/10.1364/JOSAA.7.000601

[9] van Heel, M., Portugal, R. and Schatz, M. (2009) An Electronic Text Book: Electron Microscopy in Life Science. 3D-EM Network of Excellence.

[10] Harauz, G. and van Heel, M. (1986) Exact Filters for General Geometry Three Dimensional Reconstruction. *Optik*, **73**, 146-156.

[11] Radermacher, M. (1988) Three-Dimensional Reconstruction of Single Particles from Random and Nonrandom Tilt Series. *Journal of Electron Microscopy Technique*, **9**, 359-394. http://dx.doi.org/10.1002/jemt.1060090405

[12] Orlov, I.M., Morgan, D.G. and Cheng, R.H. (2006) Efficient Implementation of a Filtered Back-Projection Algorithm Using a Voxel-by-Voxel Approach. *Journal of Structural Biology*, **154**, 287-296. http://dx.doi.org/10.1016/j.jsb.2006.03.007

[13] Orlov, I.M. and Klaholz, B.P. Compensation of Preferable Orientations in 3D Particle Reconstructions Based on Voronoi Diagrams. (in preparation)

[14] Leschziner, A.E. and Nogales, E. (2007) Visualizing Flexibility at Molecular Resolution: Analysis of Heterogeneity in Single-Particle Electron Microscopy Reconstructions. *Annual Review of Biophysics and Biomolecular Structure*, **36**, 43-62. http://dx.doi.org/10.1146/annurev.biophys.36.040306.132742

[15] Rossmann, M.G. and Blow, D.M. (1962) The Detection of Sub-Units within the Crystallographic Asymmetric Unit. *Acta Crystallographica*, **15**, 24-31. http://dx.doi.org/10.1107/S0365110X62000067

[16] Gao, H., Valle, M., Ehrenberg, M. and Frank, J. (2004) Dynamics of EF-G Interaction with the Ribosome Explored by Classification of a Heterogeneous Cryo-EM Dataset. *Journal of Structural Biology*, **147**, 283-290. http://dx.doi.org/10.1016/j.jsb.2004.02.008

[17] Sigworth, F.J. (1998) A Maximum-Likelihood Approach to Single-Particle Image Refinement. *Journal of Structural Biology*, **122**, 328-339. http://dx.doi.org/10.1006/jsbi.1998.4014

[18] Scheres, S.H., Valle, M., Nuez, R., Sorzano, C.O., Marabini, R., Herman, G.T. and Carazo, J.M. (2005) Maximum-Likelihood Multi-Reference Refinement for Electron Microscopy Images. *Journal of Structural Biology*, **22**, 139-149. http://dx.doi.org/10.1016/j.jmb.2005.02.031

[19] Scheres, S.H. (2010) Classification of Structural Heterogeneity by Maximum-Likelihood Methods. *Methods in Enzymology*, **482**, 295-320. http://dx.doi.org/10.1016/S0076-6879(10)82012-9

[20] Lyumkis, D., Brilot, A.F., Theobald, D.L. and Grigorieff, N. (2013) Likelihood-Based Classification of Cryo-EM Images Using FREALIGN. *Journal of Structural Biology*, **183**, 377-388. http://dx.doi.org/10.1016/j.jsb.2013.07.005

[21] Amunts, A., Brown, A., Toots, J., Scheres, S.H. and Ramakrishnan, V. (2015) The Structure of the Human Mitochondrial Ribosome. *Science*, **348**, 95-98. http://dx.doi.org/10.1126/science.aaa1193

[22] Khatter, H., Myasnikov, A.G., Natchiar, S.K. and Klaholz, B.P. (2015) Structure of the Human 80S Ribosome. *Nature*, **30**, 640-645. http://dx.doi.org/10.1038/nature14427

[23] Ohi, M., Li, Y., Cheng, Y. and Walz, T. (2004) Negative Staining and Image Classification—Powerful Tools in Modern Electron Microscopy. *Biological Procedures Online*, **6**, 23-34. http://dx.doi.org/10.1251/bpo70

[24] Fu, J., Gao, H. and Frank, J. (2006) Unsupervised Classification of Single Particles by Cluster Tracking in Multi-Dimensional Space. *Journal of Structural Biology*, **157**, 226-239. http://dx.doi.org/10.1016/j.jsb.2006.06.012

[25] Herman, G.T. and Kalinowski, M. (2008) Classification of Heterogeneous Electron Microscopic Projections into Homogeneous Subsets. *Ultramicroscopy*, **108**, 327-338. http://dx.doi.org/10.1016/j.ultramic.2007.05.005

[26] Shatsky, M., Hall, R.J., Nogales, E., Malik, J. and Brenner, S.E. (2010) Automated Multi-Model Reconstruction from Single-Particle Electron Microscopy Data. *Journal of Structural Biology*, **170**, 98-108. http://dx.doi.org/10.1016/j.jsb.2010.01.007

[27] Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I. and Ludtke, S.J. (2007) EMAN2: An Extensible Image Processing Suite for Electron Microscopy. *Journal of Structural Biology*, **157**, 38-46. http://dx.doi.org/10.1016/j.jsb.2006.05.009

[28] Elmlund, H., Elmlund, D. and Bengio, S. (2013) PRIME: Probabilistic Initial 3D Model Generation for Single-Particle cryo-Electron Microscopy. *Structure*, **21**, 1299-1306. http://dx.doi.org/10.1016/j.str.2013.07.002

[29] Liao, H.Y., Hashem, Y. and Frank, J. (2015) Efficient Estimation of Three-Dimensional Covariance and Its Application in the Analysis of Heterogeneous Samples in Cryo-Electron Microscopy. *Structure*, **23**, 1129-1137. http://dx.doi.org/10.1016/j.str.2015.04.004

[30] Wang, Q., Matsui, T., Domitrovic, T., Zheng, Y., Doerschuk, P.C. and Johnson, J.E. (2013) Dynamics in Cryo EM Reconstructions Visualized with Maximum-Likelihood Derived Variance Maps. *Journal of Structural Biology*, **181**, 195-206. http://dx.doi.org/10.1016/j.jsb.2012.11.005

[31] Klaholz, B.P., Myasnikov, A.G. and van Heel, M. (2004) Visualization of Release Factor 3 on the Ribosome during Termination of Protein Synthesis. *Nature*, **427**, 862-865. http://dx.doi.org/10.1038/nature02332

[32] Penczek, P.A., Frank, J. and Spahn, C.M. (2006) A Method of Focused Classification, Based on the Bootstrap 3D Variance Analysis, and Its Application to EF-G-Dependent Translocation. *Journal of Structural Biology*, **154**, 184-194. http://dx.doi.org/10.1016/j.jsb.2005.12.013

[33] White, H.E., Saibil, H.R., Ignatiou, A. and Orlova, E.V. (2004) Recognition and Separation of Single Particles with Size Variation by Statistical Analysis of Their Images. *Journal of Structural Biology*, **13**, 453-460. http://dx.doi.org/10.1016/j.jmb.2003.12.015

[34] Elad, N., Clare, D.K., Saibil, H.R. and Orlova, E.V. (2008) Detection and Separation of Heterogeneity in Molecular Complexes by Statistical Analysis of Their Two-Dimensional Projections. *Journal of Structural Biology*, **162**, 108-120. http://dx.doi.org/10.1016/j.jsb.2007.11.007

[35] Orlova, E.V. and Saibil, H.R. (2010) Methods for Three-Dimensional Reconstruction of Heterogeneous Assemblies. *Methods in Enzymology*, **482**, 321-341. http://dx.doi.org/10.1016/S0076-6879(10)82013-0

[36] De Carlo, S., Carles, C., Riva, M. and Schultz, P. (2003) Cryo-Negative Staining Reveals Conformational Flexibility within Yeast RNA Polymerase I. *Journal of Molecular Biology*, **329**, 891-902. http://dx.doi.org/10.1016/S0022-2836(03)00510-2

[37] Cheng, A. and Yeager, M. (2007) Bootstrap Resampling for Voxel-Wise Variance Analysis of Three-Dimensional Density Maps Derived by Image Analysis of Two-Dimensional Crystals. *Journal of Structural Biology*, **158**, 19-32. http://dx.doi.org/10.1016/j.jsb.2006.10.003

[38] Myasnikov, A.G., Marzi, S., Simonetti, A., Giuliodori, A.M., Gualerzi, C.O., Yusupova, G., Yusupov, M. and Klaholz, B.P. (2005) Conformational Transition of Initiation Factor 2 from the GTP- to GDP-Bound State Visualized on the Ribosome. *Nature Structural & Molecular Biology*, **12**, 1145-1149. http://dx.doi.org/10.1038/nsmb1012

[39] Penczek, P.A., Yang, C., Frank, J. and Spahn, C.M. (2006) Estimation of Variance in Single-Particle Reconstruction Using the Bootstrap Technique. *Journal of Structural Biology*, **154**, 168-183. http://dx.doi.org/10.1016/j.jsb.2006.01.003

[40] Zhang, W., Kimmel, M., Spahn, C.M. and Penczek, P.A. (2008) Heterogeneity of Large Macromolecular Complexes Revealed by 3D Cryo-EM Variance Analysis. *Structure*, **16**, 1770-1776. http://dx.doi.org/10.1016/j.str.2008.10.011

[41] Quenouille, M.H. (1949) Approximate Tests of Correlation in Time Series. *Journal of the Royal Statistical Society*: *Series B*, **11**, 68-84.

[42] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**, 1-26.

http://dx.doi.org/10.1214/aos/1176344552

[43] Efron, B. (1981) Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods. *Biometrika*, **68**, 589-599. http://dx.doi.org/10.1093/biomet/68.3.589

[44] Simon, J.L. (1969) Basic Research Methods in Social Sciences: The Art of Empirical Investigation. Random House, New York.

[45] Simon, J.L. (1997) Resampling: The New Statistics. 2nd Edition, Thompson International, Duxbury.

[46] Good, P. (2005) Introduction to Statistics through Resampling Methods and R/S-PLUS. Wiley, Hoboken. http://dx.doi.org/10.1002/9780471722502

[47] Simonetti, A., Marzi, S., Myasnikov, A.G., Fabbretti, A., Yusupova, G., Yusupov, M., Gualerzi, C.O. and Klaholz, B.P. (2008) Structure of the 30S Translation Initiation Complex. *Nature*, **455**, 416-420. http://dx.doi.org/10.1038/nature07192

[48] Kettenberger, H., Armache, K.J. and Cramer, P. (2004) Complete RNA Polymerase II Elongation Complex Structure and Its Interactions with NTP and TFIIS. *Molecular Cell*, **16**, 955-965. http://dx.doi.org/10.1016/j.molcel.2004.11.040

[49] van Heel, M., Gowen, B., Matadeen, R., Orlova, E.V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M. and Patwardhan, A. (2000) Single-Particle Eletroncryo-Microscopy: Towards Atomic Resolution. *Quarterly Reviews of Biophysics*, **33**, 307-369. http://dx.doi.org/10.1017/S0033583500003644

[50] Haynor, D.R. and Woods, S.D. (1989) Resampling Estimates of Precision in Emission Tomography. *IEEE Transactions on Medical Imaging*, **8**, 337-343. http://dx.doi.org/10.1109/42.41486

[51] Maitra, R. (1998) An Approximate Bootstrap Technique for Variance Estimation in Parametric Images. *Medical Image Analysis*, **2**, 379-382. http://dx.doi.org/10.1016/S1361-8415(98)80018-2

[52] Voorhees, R.M., Fernández, I.S., Scheres, S.H. and Hegde, R.S. (2014) Structure of the Mammalian Ribosome-Sec61 Complex to 3.4 Å Resolution. *Cell*, **157**, 1632-1643. http://dx.doi.org/10.1016/j.cell.2014.05.024

[53] Bai, X.C., Fernandez, I.S., McMullan, G. and Scheres, S.H. (2013) Ribosome Structures to Near-Atomic Resolution from Thirty Thousand Cryo-EM Particles. *eLife*, **2**, e00461. http://dx.doi.org/10.7554/elife.00461

[54] Scheres, S.H., Gao, H., Valle, M., Herman, G.T., Eggermont, P.P., Frank, J. and Carazo, J.M. (2007) Disentangling Conformational States of Macromolecules in 3D-EM through Likelihood Optimization. *Nature Methods*, **4**, 27-29. http://dx.doi.org/10.1038/nmeth992

[55] Orlov, I., Rochel, N., Moras, D. and Klaholz, B.P. (2012) Structure of the Full Human RXR/VDR Nuclear Receptor Heterodimer Complex with Its DR3 Target DNA. *The EMBO Journal*, **31**, 291-300. http://dx.doi.org/10.1038/emboj.2011.445

[56] Papai, G., Tripathi, M.K., Ruhlmann, C., Layer, J.H., Weil, P.A. and Schultz, P. (2010) TFIIA and the Transactivator Rap1 Cooperate to Commit TFIID for Transcription Initiation. *Nature*, **465**, 956-960. http://dx.doi.org/10.1038/nature09080

[57] Simonetti, A., Marzi, S., Billas, I.M., Tsai, A., Fabbretti, A., Myasnikov, A.G., Roblin, P., Vaiana, A.C., Hazemann, I., Eiler, D., Steitz, T.A., Puglisi, J.D., Gualerzi, C.O. and Klaholz, B.P. (2013) Involvement of Protein IF2 N Domain in Ribosomal Subunit Joining Revealed from Architecture and Function of the Full-Length Initiation Factor. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 15656-15661. http://dx.doi.org/10.1073/pnas.1309578110

[58] Fischer, N., Konevega, A.L., Wintermeyer, W., Rodnina, M.V. and Stark, H. (2010) Ribosome Dynamics and tRNA Movement by Time-Resolved Electron Cryomicroscopy. *Nature*, **466**, 329-333. http://dx.doi.org/10.1038/nature09206

[59] Behrmann, E., Loerke, J., Budkevich, T.V., Yamamoto, K., Schmidt, A., Penczek, P.A., Vos, M.R., Bürger, J., Mielke, T., Scheerer, P. and Spahn, C.M. (2015) Structural Snapshots of Actively Translating Human Ribosomes. *Cell*, **161**, 845-857. http://dx.doi.org/10.1016/j.cell.2015.03.052

[60] DePristo, M.A., de Bakker, P.I.W. and Blundell, T.L. (2004) Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. *Structure*, **12**, 831-838. http://dx.doi.org/10.1016/j.str.2004.02.031

[61] Levin, E.J., Kondrashov, D.A., Wesenberg, G.E. and Phillips Jr., G.N. (2007) Ensemble Refinement of Protein Crystal Structures: Validation and Application. *Structure*, **15**, 1040-1052. http://dx.doi.org/10.1016/j.str.2007.06.019

[62] Klaholz, B.P., Pape, T., Zavialov, A.V., Myasnikov, A.G., Vestergaard, B., Orlova, E., Ehrenberg, M. and van Heel, M. (2003) Structure of the *Escherichia coli* Ribosomal Termination Complex with Release Factor 2. *Nature*, **421**, 90-94. http://dx.doi.org/10.1038/nature01225

[63] Marzi, S., Myasnikov, A.G., Serganov, A., Ehresmann, C., Romby, P., Yusupov, M. and Klaholz, B.P. (2007) Structured mRNAs Regulate Translation Initiation by Binding to the Platform of the Ribosome. *Cell*, **130**, 1019-1031. http://dx.doi.org/10.1016/j.cell.2007.07.008

[64] van Heel, M. (1987) Angular Reconstitution: A Posteriori Assignment of Projection Directions for 3D Reconstruction. *Ultramicroscopy*, **21**, 111-123. http://dx.doi.org/10.1016/0304-3991(87)90078-7

[65] Harauz, G. and Ottensmeyer, F.P. (1984) Direct Three-Dimensional Reconstructions for Macromolecular Complexes from Electron Micrographs. *Ultramicroscopy*, **12**, 309-320. http://dx.doi.org/10.1016/0304-3991(83)90245-0

Supplementary Figure

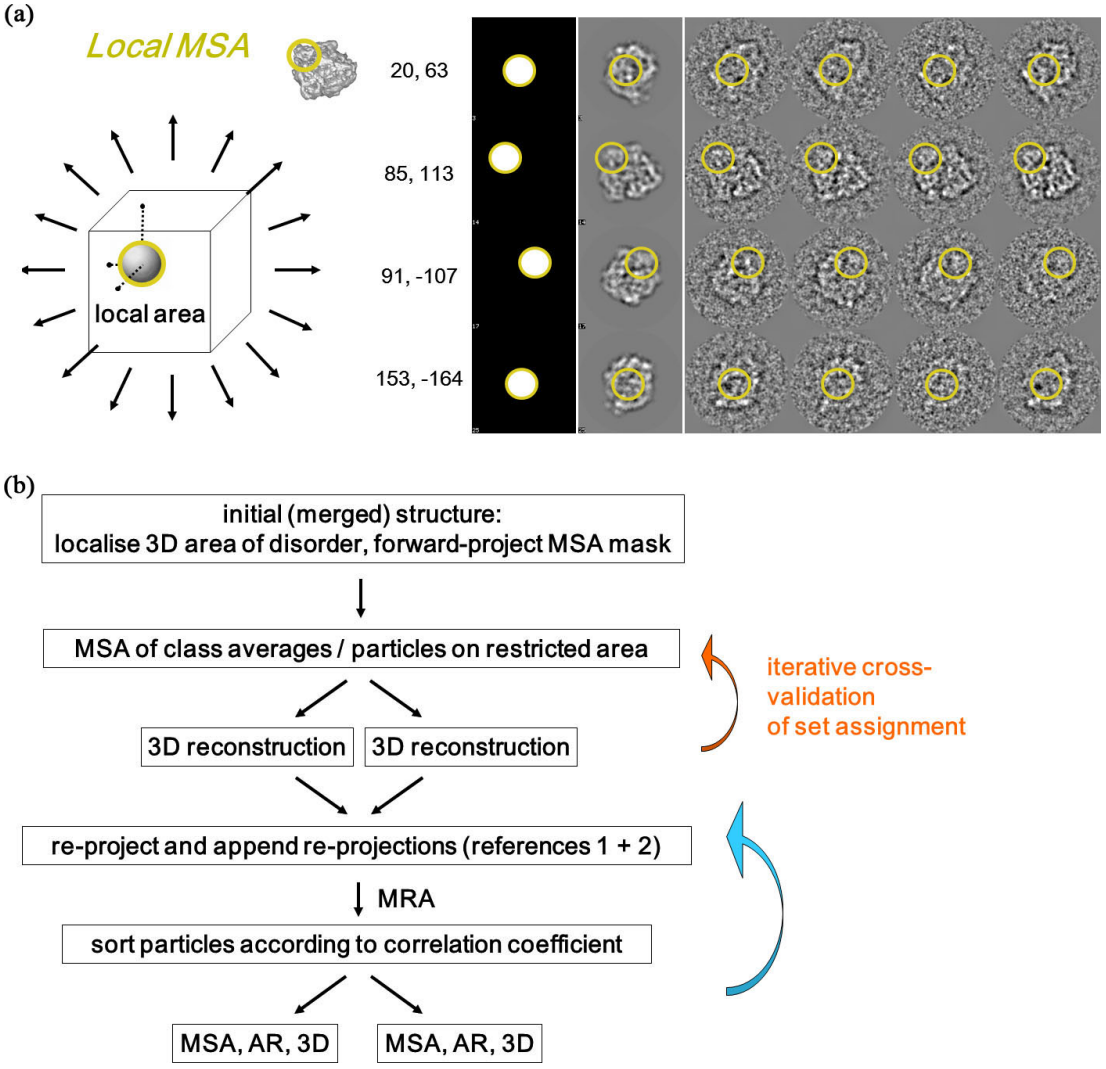

Figure S1. Details of the local 2D MSA procedure on standard class averages. The local 2D MSA procedure on standard class averages as described in [31] [32] includes the following steps. After localisation of a disordered region in a 3D cryo-EM map (the flexible 30S beak area of the 70S ribosome is labelled as an example, see insert), a 3D mask corresponding to this area (including a part of its direct neighbourhood) is re-projected in equidistant directions (e.g. 40˚ increments) such that MSA analysis will concentrate on the area of interest in each class average. For this, class averages (or particle images) have to have been sorted into views according to the corresponding Euler angles. The assignment of images into a set or another is achieved through an iterative cross-validation approach according to the best correlation with re-projections of the 3D reconstruction calculated from all images of a set. In a second step, fine-angle re-projections of the two 3D reconstructions are used as references, and iterative splitting into two sets is pursued according to the best correlation with either reference images; MRA, multi-reference alignment; AR, angular reconstitution. Moving the local MSA window reveals that long-distance conformational changes in the ribosome (P-, E-site tRNA, L1 stalk, beak or toe of the 30S subunit) correspond to concerted movements. Although proved to be quite powerful [31] [32] [34] [38] [63], the procedure harbours some limitations (see main text), which are addressed in the 3D-SC approach described in the present paper (see main figures). Local MSA can also be performed within the 3D-SC procedure and for high-resolution structure refinement (see main text).