A fully automatic method yielding initial models from high-resolution electron cryo-microscopy maps

Thomas C. Terwilliger, Paul D. Adams, Pavel V. Afonine, Oleg V. Sobolev

Supplementary Table I

Supplementary Table I. Comparison of MAINMAST, Rosetta and Phenix procedures for model-building

				Coverage (%)			Rmsd (Å)		
EMDB entry	PDB entry	Chain ID	Resolution (Å)	MAIN-MAST	Ro-setta	Map to model	MAIN-MAST	Rosetta	Map to model
5376	3j17	D1	4.1	73	70	57	1.77	1.52	1.66
5185	3j06	A1	3.3	92	97	77	1.45	0.79	1.12
5925	3j6j	A	3.6	91	98	76	1.31	0.75	1.32
5764	3j4u	A1	3.5	84	62	75	1.56	1.46	1.32
8116	5ire	A1	3.8	71	83	67	1.75	1.34	1.55
3074	5a7a	A	4.1	73	78	75	1.65	1.34	1.50
3073	5a79	A	4.1	84	76	68	1.40	1.26	1.28
8011	5gam	d	3.7	81	16	82	1.70	1.62	1.45
2850	5aey	A	4.3	72	81	63	1.83	1.39	1.30
2364	4btg	A1	4.3	58	0	61	2.10		1.90
5778	3j5p	A	3.3	78	81	76	1.50	1.18	1.18
6374	3jb0	D1	2.9	97	82	87	0.80	1.00	0.60
3246	5foj	A1	2.8	66	44	63	1.92	1.86	1.75
2513	4ci0	A1	3.4	91	87	80	1.29	0.99	1.12
3063	5a6f	C	4.2	67	0	62	2.05		1.66
3231	5fmg	K	3.6	81	78	72	1.57	1.22	1.36
2867	4uft	B	4.3	81	81	59	1.59	1.32	1.48
6555	3jci	A1	2.9	94	42	90	1.02	1.57	0.72
6551	3jcf	A	3.8	86	85	78	1.44	1.10	1.25
8015	5gaq	A	3.1	94	30	87	1.13	1.95	1.26
5495	3j26	A1	3.5	91	44	84	1.27	1.79	1.30
6272	3j9s	A1	2.6	91	73	90	1.21	1.13	0.68
Mean				82	63	74	1.51	1.33	1.31

Legend to Supplementary Table I. Maps from the EMDB were segmented using the indicated model and chain from the PDB and only including the part of the map within 4 Å of an atom in the model as described[1]. Models for MAINMAST and Rosetta de novo model-building were generously provided by G. Terashi and D. Kihara. Models were built with Phenix using the present method ("map to model") fully automatically using a command such as, "phenix.map_to_model 6272_box.ccp4 resolution=2.6 seq_file=6272.seq quick=False nproc=24". Coverage is the percentage of C_α atoms within 3 Å of a C_α atom in the deposited structure. In the published analysis[1] no models were obtained for the Rosetta analysis of EMDB maps 2364 and 3063.

Reference

1. Terashi, G, Kihara, D. (2018). Nature Commun. 9, 1618. doi:10.1038/s41467-018-04053-718.

A fully automatic method yielding initial models from high-resolution electron cryo-microscopy maps

Thomas C. Terwilliger, Paul D. Adams, Pavel V. Afonine, Oleg V. Sobolev

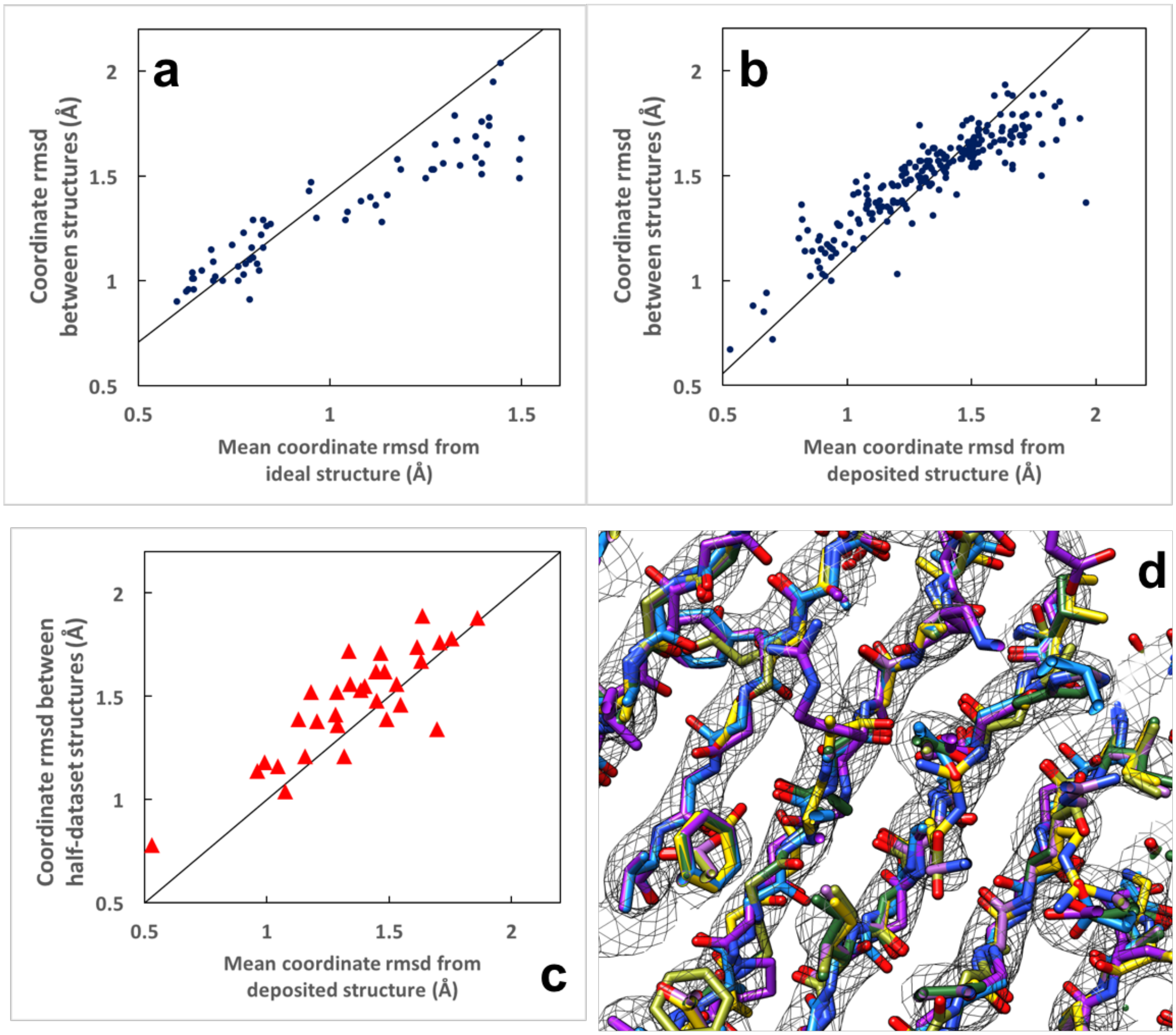
Supplementary Results

One of the benefits of automation of a process such as map interpretation is that the process can be repeated, varying random seeds or varying algorithms in order to generate an ensemble of possible interpretations. If all the interpretations are about equally consistent with the data and any prior knowledge, the variation amongst the interpretations can represent a lower bound on the uncertainty in these interpretations[1-4]. This type of approach has recently been applied to models representing maps from electron cryo-microscopy[2-4]. Local map resolution and model accuracy as well as model B-factors were evaluated using multiple molecular dynamics-based interpretations of a map[4]. Model accuracy and map resolution were evaluated by automated rebuilding of models[2,3]. It is important to note that if the errors in the models are independent, this variation can be an accurate estimate of errors, while if the models have correlated errors, it is rather a lower bound on the errors in these models[1]. A strength of our approach is that models are generated with several independent procedures, minimizing the correlation of errors in separate models.

We illustrate two applications of repetitive map interpretation to error estimation using our automated procedures. In the main text in Fig. 1D, three models were generated based on a sharpened cryo-EM map using different algorithms to maximize independence of the models.

The differences between overlapping parts of these models can be used as an estimate of the uncertainties in the models[1]. We first carried out a simulation to verify that this procedure would be expected to work in an ideal case. We generated maps with varying amplitude and phase errors based on chain A of PDB entry 5k0z at a resolution of 2.8 Å, and we then analyzed each map with our standard procedure. For each map, we calculated the true error in the models we generated based on their coordinate rmsd to the known true structure. We also calculated the coordinate rmsd between the two independent automatically-generated models, which would be expected to be about $\sqrt{2}$ times their individual errors if they are independent. In Fig. SR-1A we compare these uncertainty estimates with the actual coordinate differences, including only cases where at least 50% of the known structure was reproduced so as not to include very poor models. We find that they are similar to expected values (the line has a slope of $\sqrt{2}$), though with a small systematic difference that is consistent with a small correlation of errors in the automatically-generated models. In Fig. SR-1B we apply the same analysis to the models generated from data in the EMDB and compared with deposited models. Fig SR-1B shows that this relationship is very similar to the one shown in Fig. SR-1B, except that the slope is slightly different. The slope indicated by the line in Fig. SR-1B corresponds to that expected if the deposited models each had about half the rms error of the automatically-generated models. This analysis supports that idea that internal consistency of independently-generated models may be useful in creating estimates of model error[2-4].

Uncertainty estimates can also potentially be obtained from a comparison of models obtained from independent half-datasets as suggested recently[2,5,6], though the correlation of errors can be greater in this case if the same methods are applied in to each half-dataset. Fig. SR-1C illustrates such an analysis for 31 pairs of half-datasets with resolutions ranging from 2.2 to 4.5 Å. The rms



Supplementary Results Figure SR-1. Applications of automated map interpretation to error analysis. A. Simulation with known true structure, maps with simulated errors, and automatic interpretations of maps with errors. Comparison of rms coordinate differences between automatically-generated models and the known true structure (average of two values) with coordinate differences between the two automatically-generated models. The slope of the line (√2) represents the ideal slope if errors were random and equal for the two automatically-generated models. Models were generated by chain tracing followed by iterative coordinate randomization and refinement with automatically-determined secondary structure restraints, and by pattern recognition of secondary structure elements followed by iterative extension with short peptide libraries. B. Analysis of data from analysis of maps from the EMDB as in A, except the ordinate values correspond to the mean difference from the deposited structure. The line shown corresponds to the slope expected (approximately 1.1) if errors in the deposited model were half the size of those in the automatically-generated models. C. Comparison of rms coordinate differences between models built using independent half-datasets with rms coordinate differences between the known true structure and automatically-generated models built using the full dataset. D. Six models resulted from repetitive interpretation of sharpened map in Fig. 1D with varying random seeds. See text for details.

coordinate difference between models built from half-datasets is correlated with (and numerically similar to) the rmsd coordinate difference between deposited models and automatically-built models created using the corresponding full datasets. Note that the numerical similarity is in part fortuitous. Errors in automatically-built models come both from errors in the maps and from deficiencies in the modeling process. To see this, consider that two nearly perfect but very low-resolution half-maps, if analyzed by the same model-building procedure, might well yield two models that are very similar to each other but that are very different from the true structure. In Fig. SR-1C we use different random seeds in model-building to reduce the correlation of model-building errors, but this does not eliminate this correlation of errors. Consequently the rms coordinate difference between models built from two half-maps using the same method is largely an indication of the model errors arising from errors in the maps, and includes only a portion of the possibly more significant model errors coming from deficiencies in the model-building process.

Fig. SR-1D illustrates the use of repetitive map interpretation to estimate local uncertainties in our automatically-generated models as has been done recently by others[2-4]. The sharpened map in Fig. 1D was interpreted six separate times with *phenix.map_to_model*, each time with a different random seed for all steps such as refinement and fragment extension where randomization is used. The six models generated in this way are superimposed in Fig SR-1C, and it can be readily appreciated that the model shown in Fig. 1D was not the only possible interpretation of this map. This analysis gives an idea of the uncertainty in the model and potentially in each part of the model, with caveats as discussed previously in the context of crystallograpy[1] and cryo-electron microscopy[2-4].

In addition to providing mechanisms for error estimation in atomic models representing cryo-EM reconstructions, the full automation of map interpretation provides a path towards the vision of continuous re-interpretation of deposited cryo-EM reconstructions and improvement of the models that represent them[7]. In the crystallographic field, deposited structures are already being continuously improved with database-wide re-refinement procedures[8]. As improved procedures are developed for interpretation of cryo-EM maps, the new procedures can be incorporated into automated frameworks for re-interpretation of all existing data, yielding ever-improving representations of these structures. This re-interpretation could eventually be extended to begin with the original images obtained from cryo-EM, as is beginning to be done with X-ray data[9].

References for Supplementary Results

1. Terwilliger, T.C., et al. (2007). Acta Cryst. D63, 597-610.

2. Hryc, C.F., et al. (2017). Proc. Natl. Acad. Sci. USA 114, 3103-3108.

3. Herzik, M.A., Fraser, J.S., Lander, G.C. (2017). bioRxiv doi: https://doi.org/10.1101/128561.

4. Singharoy, A., et al., (2016). eLife 2016;5:e16105 DOI: 10.7554/eLife.16105

5. Brown A, et al., (2015). Acta Crystallogr D71:136-153.

6. DiMaio, F., et al. (2013). Protein Sci. 22:865-868

7. Terwilliger, T.C., Bricogne, G. (2014). Acta Cryst. D70, 2533-2543.

8. Joosten, R.P., Joosten, K., Murshudov, G.N., Perrakis, A. (2012). Acta Cryst. D68:484-496.

9. Grabowski, M., et al. (2016). Acta Cryst. D72, 1181-1193.