Segger – Docking Scores

Last updated: Oct. 11, 2015 (Segger v1.9.2, Chimera Version 1.11.0)

Overview

When docking (or fitting) a molecular model into a high resolution map (say, higher than 10Å), as was shown in the Fit to Segments tutorial, it can be quite clear by visual inspection if the fit is good, based on whether elements such as helices and beta sheets in the model match the higher density regions in the map.

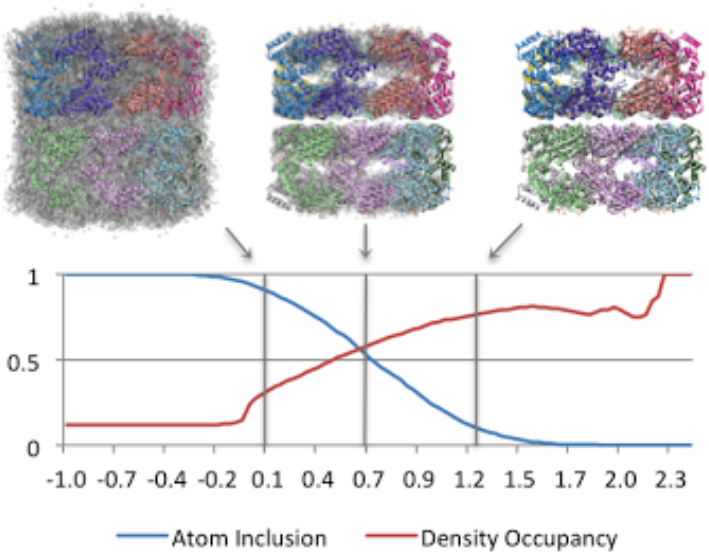
At lower resolutions, validation by visual inspection alone can be a bit harder. To evaluate whether a fit is good in such cases, Segger computes several scores:

- ***Density cross-correlation*:** This score is computed by first simulating a density map from the model, at approximately the same resolution as the density map. (In the Fitting To Segmentstutorial, it was shown how this map is generated and used during the fitting process). The density cross-correlation score is the sum of all the density values in the simulated map multiplied by the density value at the same position in the map.
- ***Atom inclusion*:** This score reflects what fraction of the atoms in the model are located at a position having density higher than a given density level.
- **Density occupancy:** This score reflects what fraction of the grid points in the density map with density above a given density level have atoms nearby (hence can be considered occupied).
- ***Clashes with symmetric copies*:** This score reflects how many atoms in the model clash with atoms from symmetric copies of the same model, in maps that have some type of symmetry. This score could also reflect clashes with other models fitted simultaneously within the map (e.g. as done by Multifit, or Chimera). Here the simpler case of fitting one structure at a time is considered.

The cross-correlation score is commonly used to pick out the "best" fit, for example while doing an exhaustive search (e.g. Situs). The other three scores (atom inclusion,

density occupancy, and clash score) were inspired and pioneered by the EMFit program.

Note that atom inclusion and density occupancy are both dependent on a contour level. While it is possible to increase/decrease the contour level so as to decrease/increase atom inclusion score, the same adjustment tends to have the opposite effect on density occupancy (see figure below with an example). Hence, ideally, the contour level chosen balances both scores.



The top 3 images shows the map of GroEL @4Å resolution (transparent surface), with docked models (ribbons with different colors for each protein), at 3 thresholds: 0.1, 0.7 and 1.25. The plot shows Atom Inclusion and Density Occupancy scores at evenly distributed contour levels. The contour level at which atom inclusion and density occupancy are roughly balanced in this case is ~0.7 (middle image).

To calculate a statistical significance for the fit, Segger computes a z-score for each of the scores (cross-correlation, atom inclusion, density occupancy, and clash score). The z-score is computed as follows:

$$z = \frac{S(1) - avgS(2..N)}{stdevS(2..N)}$$

In the above, S(1) is the top score obtained after a search (e.g. rotational search), avgS(2..N) is the average of all the other scores (N-1 in total), and stdevS is the standard deviation of all the other scores excluding the top score. The z-score reflects how much higher the top score is compared to the other scores; or, statistically speaking, how many standard deviations it is away from the mean.

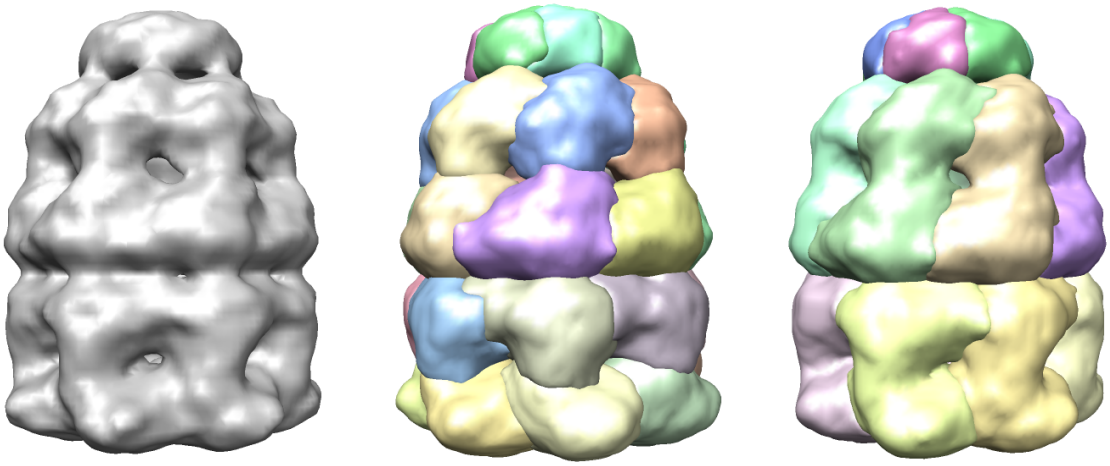
When this score is significantly higher, the z-score is also high. The computation of this score is described below.

Example: GroEL+GroES

As an example, we will use here the density map of GroEL at 23.5Å resolution. The map can be download from this link, or using File -> Fetch by ID... -> EMDB: 1046.

1. Segmentation

- First, the map was segmented using the **Segment Map** dialog at a threshold of **0.03**, with **3**steps of size **1**. The result is 35 regions.
- The regions were further grouped interactively to produce 21 regions.
- The density map is shown below on the left, the segmentation after smoothing and grouping is shown in the middle, and the regions after interactive grouping are shown on the right.

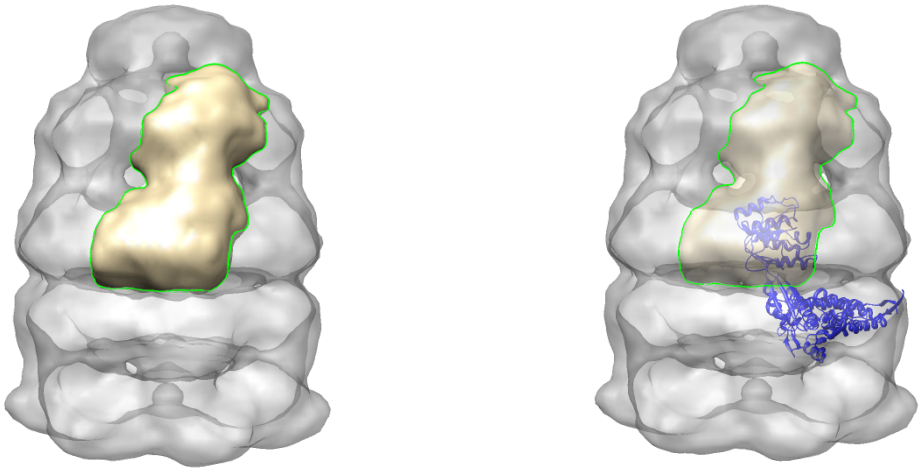


Fitting

- Separate Chain A from 1gru.pdb
  - this can be done by selecting it alone and saving it to a new pdb file, or
  - selecting all other chains and deleting them (type 'del sel') on Chimera command line (Tools -> General Controls -> Command Line to show it if it's not already shown).
- Select one of the segmented regions in the main window (Control+Left Mouse Button) as shown below left, and then select the following in the Fit To Segments dialog:

- **Structure to fit:** structure of isolated chain A
- **Density Resolution:** 23 (Å)
- **Which regions to use for fitting:** Combined selected regions (Default setting)
- **Alignment method:** Rotational search (try **100** evenly rotated fits)
- **Mask map with region(s) to prevent large drifts:** not checked (Default setting)
- **Optimize fits:** checked (Default setting)
- **Cluster fits that are < 5.0 Angstroms and < 3.0 degrees apart** (Default setting)
- **Add top 1 fit(s) to list** (Default setting)
- **Clashes with copies from symmetry:** not checked (Default setting)

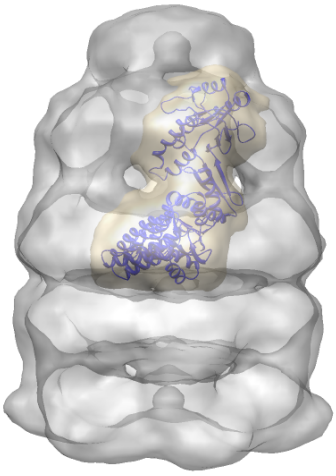
- Then press the Fit button.
  - The "fitted" model is shown below on the right.
  - Due to the optimization procedure being enabled, the structure has drifted towards the higher densities in the middle of the map.
  - Keeping in mind the structure of the entire complex, this is definitely not the right fit. (Note that depending on which of the 7 symmetric regions you have selected, the model may not have drifted as shown; if it hasn't, you may try another of the 7 regions).



Since the segmented region actually does match the protein, we can use it to better guide the fit.

- This is done by "masking" the map with the region.
- To enable this, make sure "Mask map with region(s) to prevent large drifts" is checked, then press the Fit button. This prevents the model from drifting outside the segmented region, since the densities outside the region are set to 0 (don't worry, the original map will not be affected; a copy of the map is masked, and after fitting, the masked version is discarded).

- The result from this is shown in the image below on the right. Now, the "right" fit is more certain to be found.



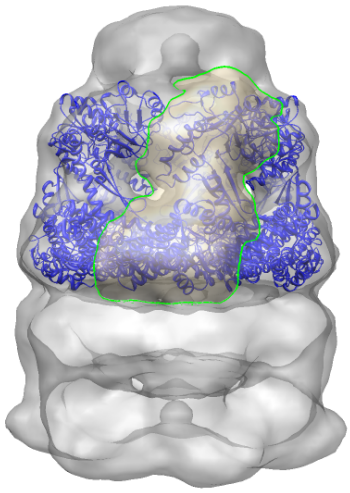
Note that the two fits done so far are both shown as entries in the "fits list" just below the "Structure to fit" field (in the Fit to Segments dialog). When doing multiple fits, each fit shows up in this list as a separate entry. Clicking on an entry will put the structure back in the corresponding position in the map (as long as the structure itself and the map are still open in your Chimera session).

Symmetry

This map has C7 symmetry for each of the 3 distinct protein structures in the GroEL+GroES complex. This means that we can take each one of the 3 proteins, apply the symmetry operations, which create 6 other copies of each of the proteins but rotated around the "ring" like shape of the complex. By doing so we would thus have recreated the whole complex. This is illustrated with one of the proteins (chain A from PDB:1GRU), which we just fitted, below:

Next to the option "Clashes with copies from symmetry:", press the **Show** button. This will do two things:

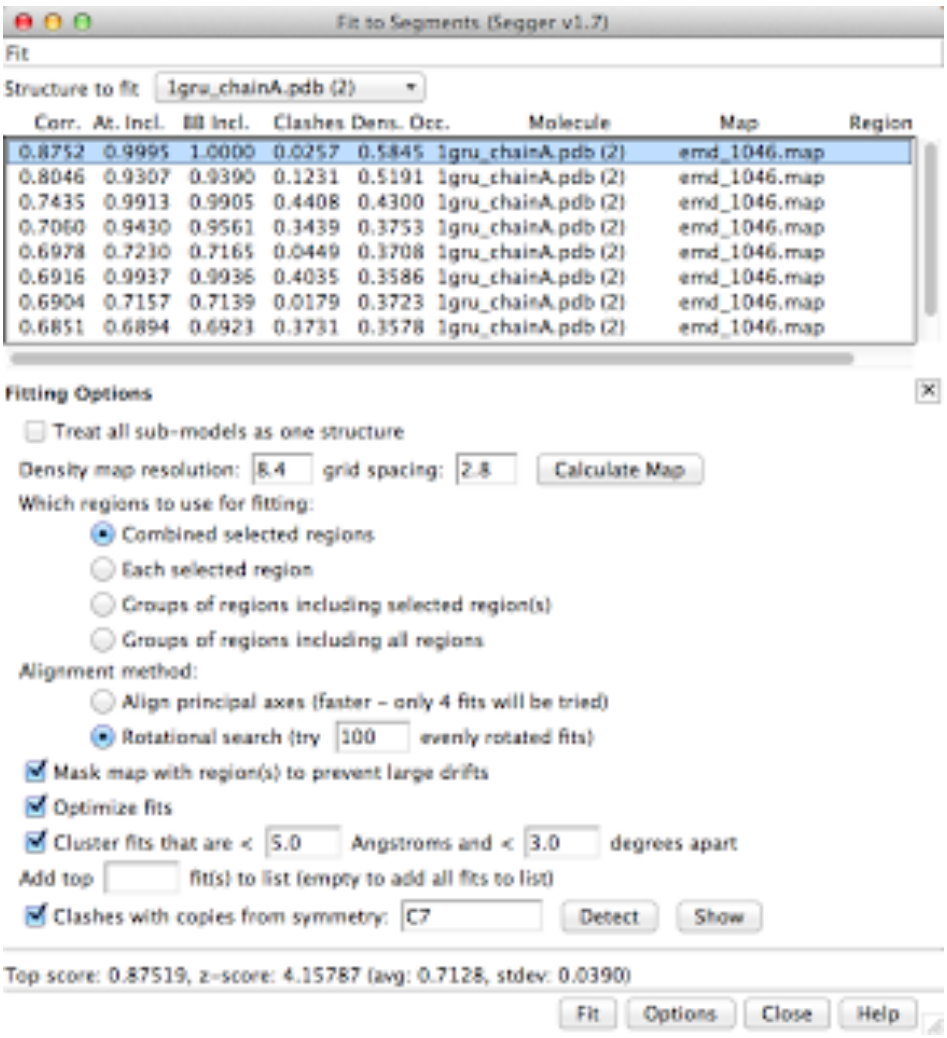
- Detect the symmetry of the map using the Chimera measure symmetry command.
- Place copies of the structure being fit based on this symmetry. In this case, the detected symmetry string "C7" is displayed, and the copies will be placed as shown below.



4. z-Scores

By default, each time you press the **Fit** button, after the search is done, whether it was by aligning principal axes or by rotational search, only one fit is added to the "Fits" list in the Fit to Segments dialog for each search done: the one with the highest density cross-correlation score.

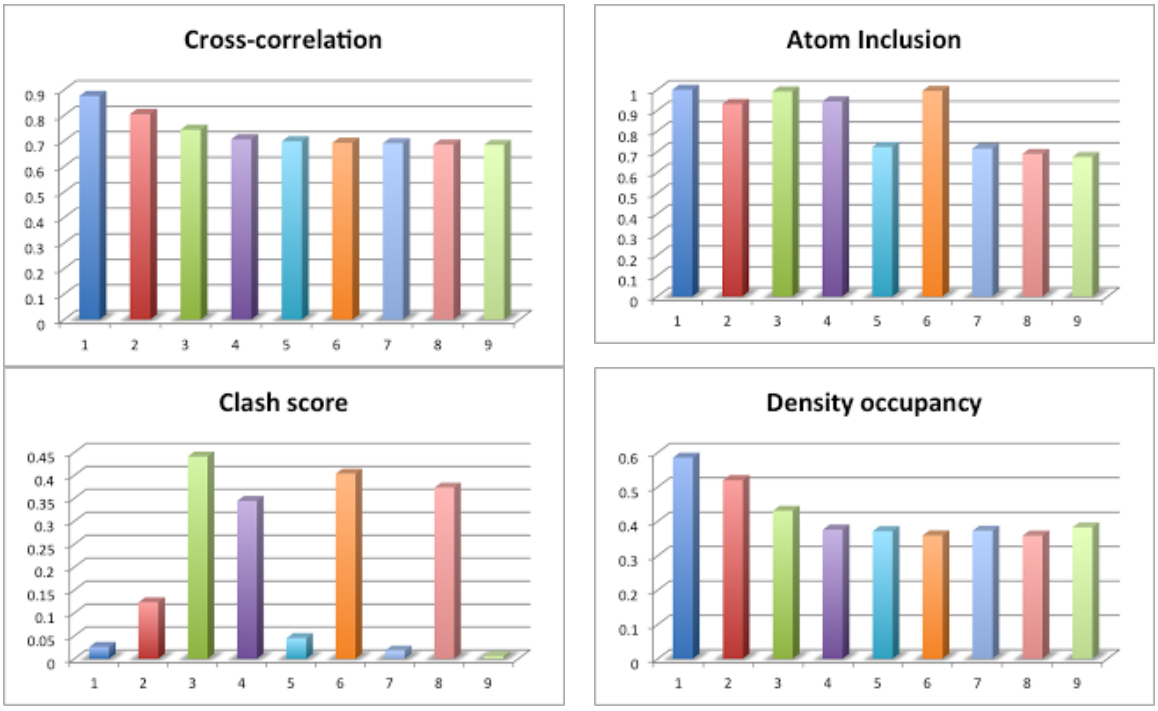
- Thus, after the two fit operations described above, you should see 2 entries in the fits results list. (If you did more than two fit operations, you will of course see more entries).
- To compute z-scores, the other fits produced during the search will have to be reported as well. To enable this remove from the box inside the option: **Add top [ ] fit(s) to list.** (Leaving the box empty for this option adds all the fits tried during the serach to the list).
    - Make sure the Rotational search option is selected. For computing z-scores, the more fits that are tried during the search, the more accurate the statistics will be.
    - To compute clash scores as well make sure that the box next to "Clashes with copies from symmetry" is checked.
    - Select "Delete ALL fits from list" from the Fit menu in the Fit To Segments dialog. This will remove all the fits produced so far. Clearing the list is important for computing z-scores, as the z-scores are computed amongst the all the entries in the list.
    - Then, press the Fit button. After some computation time, the list should contain 9 entries, as shown below.



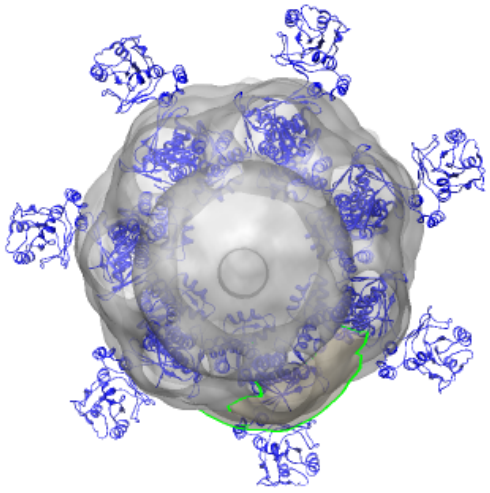
- Note that the number of fits tried was actually much higher (100), but after optimization, only 9 unique fits were identified. The entries in the fit list are sorted by cross-correlation.
- To export the scores to a text file, for plotting, further analysis, etc., select "Export fit scores" from the Fit menu at the top of the Fit to Segments dialog.
- The text file will have 4 columns, with each score in a different column, and each row representing a different fit.

Cross-correlation	Atom Inclusion	Backbone-Atom Inclusion	Clash score	Density occupancy
0.875192	0.999475	1	0.025729	0.584549
0.804556	0.93069	0.93897	0.123129	0.519053
0.74352	0.991336	0.990464	0.440798	0.430039
0.705959	0.94303	0.956135	0.343922	0.37526
0.697837	0.723024	0.716465	0.044894	0.370795
0.691633	0.993699	0.993643	0.403518	0.358589
0.690361	0.715673	0.713922	0.017852	0.372283
0.685095	0.68942	0.692308	0.373064	0.357845
0.683728	0.674718	0.680229	0.005513	0.381959

- The values in the table can be plotted with a plotting program, e.g. Excel. Below are the plotted values for each of the 9 fits.



- Note that the fit with highest cross-correlation score also has the highest atom inclusion scores (though not by much), and highest density occupancy score. However, it doesn't have the highest Clash score. Find this fit in the list and show the symmetric copies using the "Show" button to see why. Here's what it looks like:



- The z-scores are also included in the exported text file. The z-scores are quite low for this fit; typically z-scores are lower in low-resolution maps, since the difference between the top score and all the other scores is not quite that large.

	Z-score	Top score	Mean	STDev
Cross-correlation:	4.157869	0.875192	0.712836	0.039048
Atom Inclusion:	1.242762	0.999475	0.832699	0.134198
Density occupancy:	3.687111	0.584549	0.395728	0.051211
Clash score:	-1.097925	0.025729	0.219086	0.176112

Conclusions:

- The cross-correlation score is commonly used in assessing how good a fit is.
- Other scores, such as atom inclusion, density occupancy, and clash scores, can also give an idea of how good a fit is.
- Z-scores can be used to assess statistical significance of the results.