# Segger Tutorial – SegFit Scores and Z-score

Last updated: Jan. 31, 2021 (Segger v2.5.4, Chimera Version 1.13)

Greg Pintilie
gregdp@stanford.edu

## Overview

When fitting a molecular model into a high resolution map (say, higher than 10Å), as was shown in the SegFit Dialog 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.

The cross-correlation score is commonly used to pick out the "best" fit, for example while doing an exhaustive search.

To calculate a statistical significance for the fit, SegFit computes a Z-score using the top N scores obtained during PCA search or by aligning centers plus # rotations. 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 higher, it can also mean the fit with the top score is more ‘correct’.

**Note** also that the N is important, and the higher the N, the more meaningful the statistics. Also **note** that in SegFit, N is the number entered in “Add top [N] fit(s)” option.

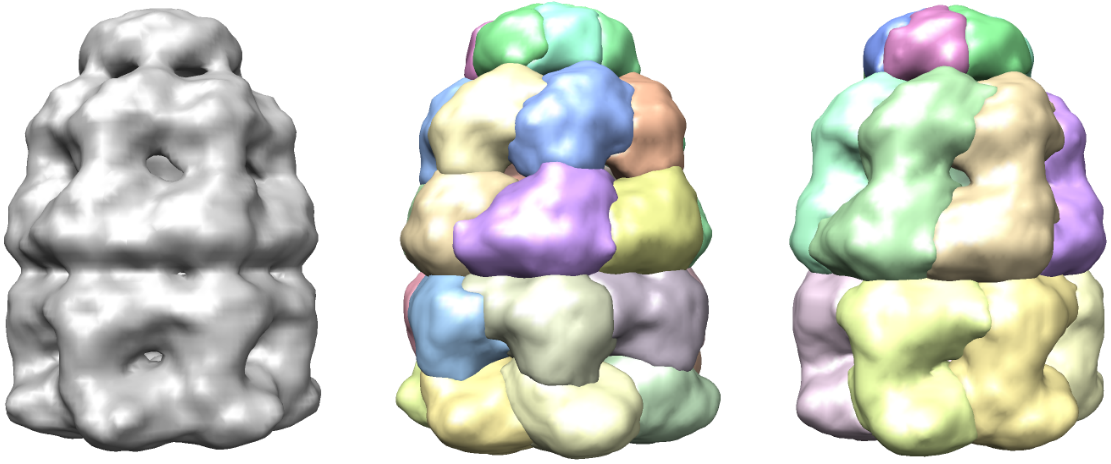
## Example: GroEL+GroES

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 **Segger** 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.

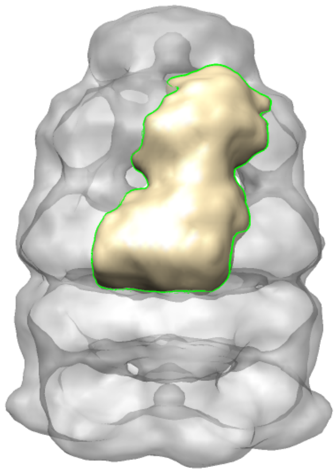


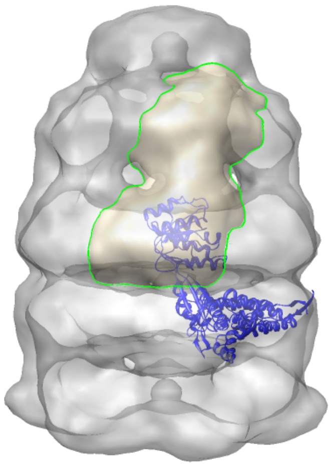
## Fitting

- 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 SegFit dialog:
  - **Structure to fit:** structure of isolated chain A
  - **Density Resolution:** 23 (Å)
  - **Fit to:** Combined selected regions (Default setting)
  - **Fit by:** 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 using 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.

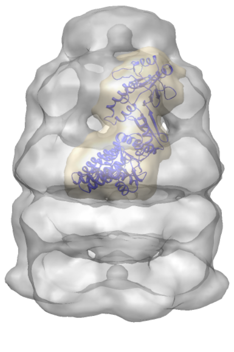
- Keeping in mind the structure of the entire complex, this is definitely not the right fit. (Note that depending on which of the 7 symmetric regions you have selected, the model may not have drifted as shown; if it hasn't, you may try another of the 7 regions).





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 SegFit 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).

## Scores

The scores shown in the SegFit list are:

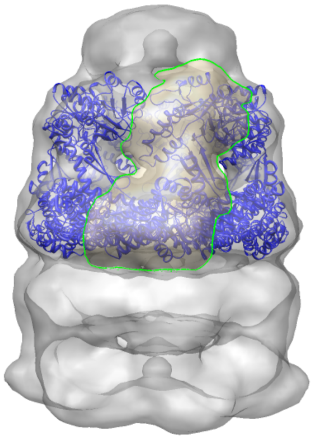
- **Corr.:** Cross-correlation between the model-generated map and the density map fitted to. Note that using “Mask map…” option can affect this score.
- **At. Incl.:** Atom Inclusion – % of atoms inside shown contour level of map
- **BB Incl.:** Backbone atoms Inclusion - % of backbone atoms within shown contour level
- **Clashes:** When the “Clashes using symmetry” option is checked, this will show %atoms clashing with symmetric copies.
- **Dens. Occ.:** % of the volume in the region(s) fitted to that are occupied by atoms in the fitted model (the higher, the better the model fits)

## 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 using symmetry:", press the **Show** button. This will do two things:

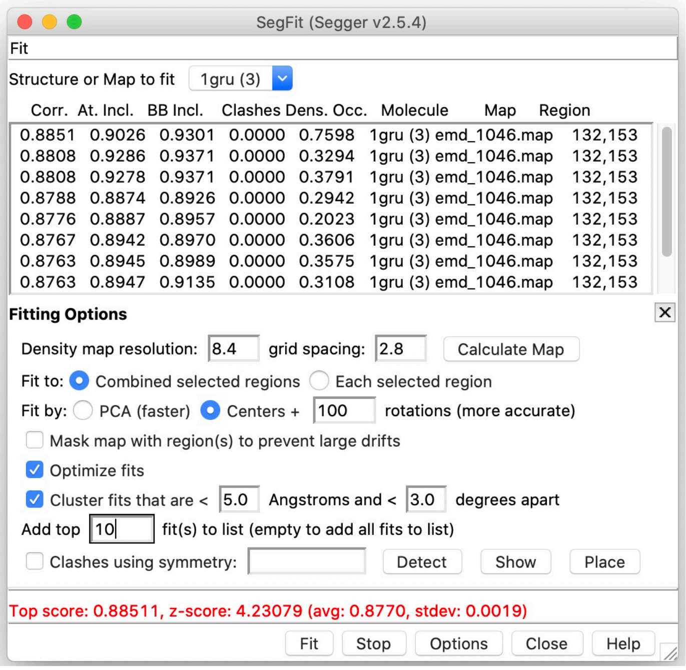
- Detect the symmetry of the map using the Chimera measure symmetry command.
- Place copies of the structure being fit based on this symmetry. In this case, the detected symmetry string "C7" is displayed, and the copies will be placed as shown below.



## 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 SegFit 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 using symmetry" is checked.
    - Select "Delete ALL fits from list" from the Fit menu in the SegFit 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.



Structure or Map to fit: 1gru (3)

| Corr. | At. Incl. | BB Incl. | Clashes | Dens. Occ. | Molecule | Map | Region |
|---|---|---|---|---|---|---|---|
| 0.8851 | 0.9026 | 0.9301 | 0.0000 | 0.7598 | 1gru (3) | emd_1046.map | 132,153 |
| 0.8808 | 0.9286 | 0.9371 | 0.0000 | 0.3294 | 1gru (3) | emd_1046.map | 132,153 |
| 0.8808 | 0.9278 | 0.9371 | 0.0000 | 0.3791 | 1gru (3) | emd_1046.map | 132,153 |
| 0.8788 | 0.8874 | 0.8926 | 0.0000 | 0.2942 | 1gru (3) | emd_1046.map | 132,153 |
| 0.8776 | 0.8887 | 0.8957 | 0.0000 | 0.2023 | 1gru (3) | emd_1046.map | 132,153 |
| 0.8767 | 0.8942 | 0.8970 | 0.0000 | 0.3606 | 1gru (3) | emd_1046.map | 132,153 |
| 0.8763 | 0.8945 | 0.8989 | 0.0000 | 0.3575 | 1gru (3) | emd_1046.map | 132,153 |
| 0.8763 | 0.8947 | 0.9135 | 0.0000 | 0.3108 | 1gru (3) | emd_1046.map | 132,153 |

Density map resolution: 8.4 grid spacing: 2.8

Fit to: ● Combined selected regions ○ Each selected region

Fit by: ○ PCA (faster) ● Centers + 100 rotations (more accurate)

☐ Mask map with region(s) to prevent large drifts

☑ Optimize fits

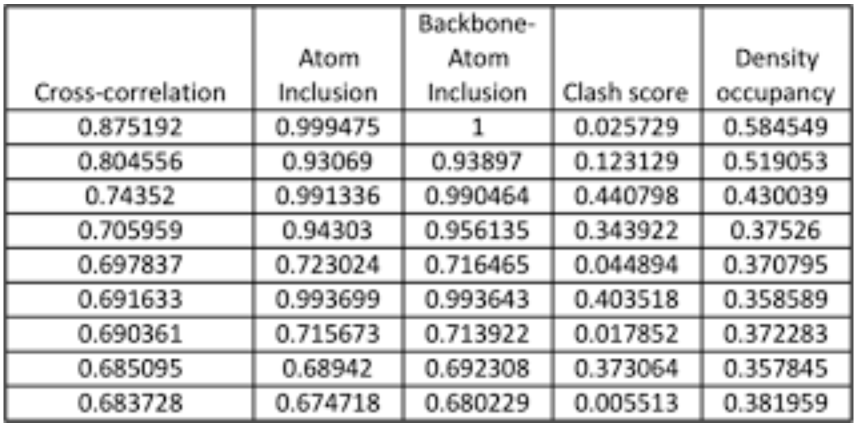
☑ Cluster fits that are < 5.0 Angstroms and < 3.0 degrees apart

Add top 10 fit(s) to list (empty to add all fits to list)

☐ Clashes using symmetry:

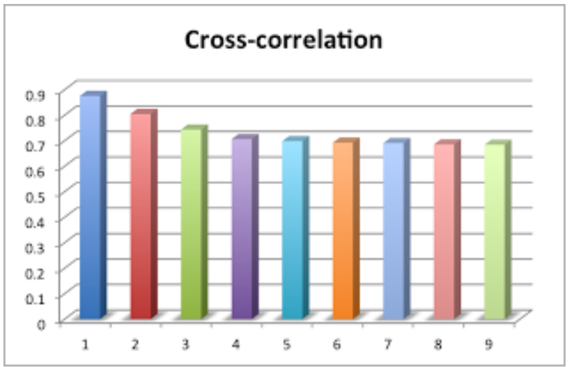
Top score: 0.88511, z-score: 4.23079 (avg: 0.8770, stdev: 0.0019)

- Note that the number of fits tried was actually much higher (100), but after optimization, 10 unique fits were added to the list. 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 SegFit dialog.
- The text file will have 4 columns, with each score in a different column, and each row representing a different fit.

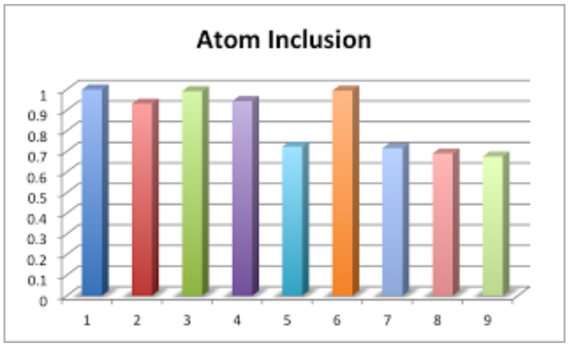


| Cross-correlation | Atom Inclusion | Backbone-Atom Inclusion | Clash score | Density occupancy |
|---|---|---|---|---|
| 0.875192 | 0.999475 | 1 | 0.025729 | 0.584549 |
| 0.804556 | 0.93069 | 0.93897 | 0.123129 | 0.519053 |
| 0.74352 | 0.991336 | 0.990464 | 0.440798 | 0.430039 |
| 0.705959 | 0.94303 | 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 |

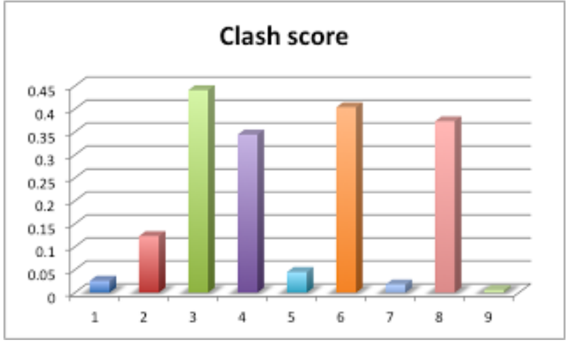
- The values in the table can be plotted with a plotting program, e.g. Excel. Below are the plotted values for each of the 9 fits.



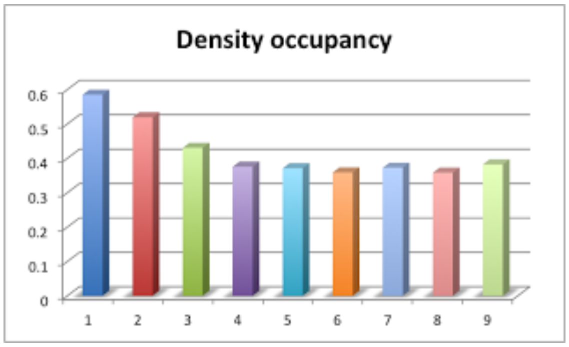
Cross-correlation



Atom Inclusion



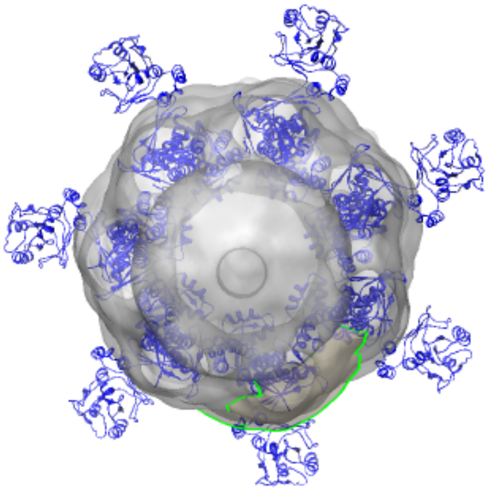
Clash score



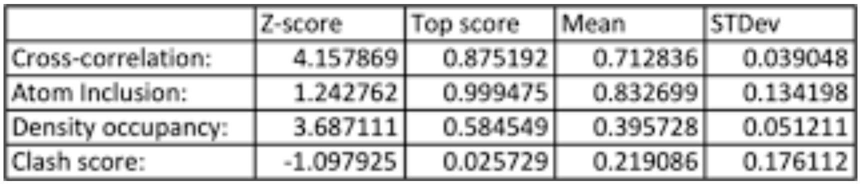
Density occupancy

- 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. A Z-score for Cross Correlation of ~4 at this resolution tends to be significant.



| | 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.