Tutorial on scoring methods for evaluation of electron density maps

Tom Terwilliger, Los Alamos National Laboratory 12-August-2005

What you will do in this tutorial

This tutorial will show you how you can combine scores from several criteria to evaluate the quality of an electron density map. You will be able to use real SAD data from a small protein (IF5a), compare the maps obtained with the correct, inverse, and random sites, and examine the scores obtained from analyzing features of the map such as its SD of local rms, its skew, or connectivity. You will be able to choose which scoring criteria are the most useful and to put these together into a simple scoring algorithm. You will then be able to apply your scoring algorithm to solve this structure and to test it on other structures.

Getting ready

Download the developmental version of PHENIX (www.phenix-online.org) and install it. Installation takes just a minute or two if /usr/local/ is where you want phenix to be installed and you use the command:

```
./install --prefix=/usr/local/ --no-textal-maps
```

Set up your environment: put the line

```
source /usr/local/phenix-1.21a/phenix_env
```
into your .cshrc file in your home directory and use the C-shell: just type

```
csh
```

and the phenix environment should be set up. You can test it with

```
iotbx.reflection_statistics
```
which should tell you about the iotbx analysis tool which you may wish to use to look at the statistics of a data file.

Copy over some data to work on

Make a new directory to work in and copy over the if5a data (called p9_sad here) to it. For example:

```
cd
mkdir phenix_tutorial
cd phenix_tutorial
cp -r $PHENIX/examples/p9_sad .
cd p9_sad
pwd
ls
```

Running AutoSol in PHENIX to phase and evaluate a map

Running AutoSol in PHENIX is easy. In this tutorial we will run it from a script. You control what AutoSol does with keywords in the file **Facts.list**. Edit the **Facts.list** that is in this directory to look like this:

```
force
create_scoring_table True
resolution 3.0
sg "I 4"
cell 113.949 113.949  32.474 90.000  90.000  90.00
read_sites True
ha_sites_file p9_sites.xyz
stop_after_scoring True
```

and also make a file **p9_sites.xyz** with the correct heavy-atom sites for p9:

```
xyz        0.680      0.388      0.012
xyz        0.314      0.237      0.478
xyz        0.336      0.207      0.418
```

Now you can run AutoSol which will phase with these sites (and it will also phase with the inverse solution) and evaluate the maps from each:

```
phenix.runWizard AutoSol restart |tee AutoSol.log
```
You need the **restart** command so that it will start over from the beginning every time. As it runs notice that it gives you an output listing of the scores for randomized maps for each of several criteria, and then the scores for the maps calculated from your sites and from the inverse. Then it calculates a Z-score for your map for each criteria, where the Z-score is the score for your map minus the mean score for the random maps, divided by the standard deviation of scores for the random maps. The overall score for your map is the sum of these Z-scores.

Part of the output in **AutoSol.log** might look something like this:

```
Evaluate_solution
Setting up new scoring table with  6  values
TEMP7/resolve.scores CC 0.1449478
TEMP7/resolve.scores RFACTOR 0.6076099
TEMP7/resolve.scores SKEW -0.0040215286
```
where the three criteria CC, RFACTOR an SKEW are being used to evaluate the map.

You can also look at the file **AutoSol_summary.dat** for a summary of the scoring on each of your maps (the one from the sites as you put them in and the inverse sites). Part of this file might look like:

```
Solution # 1   SCORE:58.8961885202 Dataset #1   FOM: 0.41 -----------
-----
```

```
Solution 1 read from file p9_sites.xyz Dataset #1

 Score type:          CC        RFACTOR        SKEW       NCS_OVERLAP
Raw scores:         0.473        0.501        0.122         0.000
Z-scores:          36.014       18.782        4.100         0.000

Refined heavy atom sites (fractional):
xyz        0.680        0.388        0.012
xyz        0.314        0.237        0.478
xyz        0.336        0.207        0.418
```

The main goal in this tutorial is to come up with the best set of criteria for combining with Z-scores. You can use a single criterion or a group of them.

Criteria for evaluating an electron density map

You have available to you several criteria for evaluating an electron density map. These are:

- SKEW -- the skew of the electron density map (related to rho**3)
- SD -- the standard deviation of local rms. This is the scoring criterion used in SOLVE. It is high if there are regions of low rms (solvent) and regions of high rms (protein), and low if the map has a uniform rms everywhere (random)
- RFACTOR -- the r-factor for 1 cycle of density modification. This reflects how well the observed amplitudes agree with those which would make the map conform to expectations (i.e., flat solvent).
- CC -- the correlation of the starting map and the map obtained with only map-probability phasing (phases from density modification not recombined with original phases). This criteria is closely related to RFACTOR.
- TRUNCATE -- The correlation of a map obtained by truncating the density in the map at a high level with the original map. A measure of how distinct the density is in the map.
- REGIONS -- The number of regions obtained if the map is truncated at a high level. A measure of the connectivity of the map.
- NCS_OVERLAP -- the overlap of NCS-related density. This is zero if no NCS is present or cannot be found from the heavy-atom sites.

Changing the criteria used for scoring

You can change the criteria used for scoring by adding a line to your **Facts.list** file specifying which criteria to use. Here is how you would choose SD and SKEW and RFACTOR:

```
score_type_list SD SKEW RFACTOR
```

Run AutoSol again, and see what scores you get for these criteria and how they add up. Now try all the scoring criteria. Decide how to put these together to get the best possible discrimination between the correct hand (the hand you put in) and the inverse.

Now reconsider what we are doing...is this the best way to optimize our scoring procedure? Perhaps we need to check it against a random set of sites? Perhaps the resolution matters? Perhaps we need also to check it against some different data?

Try a random set of sites by editing a new file **random.xyz** and putting some random coordinates in it. How does this fare in your scoring scheme?

Now try changing the resolution with the **resolution** command in your **Facts.list** file. The maximum resolution of this dataset is 2.1 A.

Trying your scoring criteria with a new set of data

Now try it with a new set of data. You can make a new directory for the sec17 data:

```
cd ../
mkdir sec17
cd sec17
cp $PHENIX/examples/sec17* .
pwd
ls
```

and a **Facts.list** something like this, except you will want to put in your own list of scoring criteria:

```
force
create_scoring_table True
resolution 3.0
read_sites True
ha_sites_file sec17.xyz
score_type_list TRUNCATE REGIONS CC RFACTOR SKEW SD
stop_after_scoring True
```

and a **sec17.xyz** like this:

```
xyz      0.3780      0.1961     -0.0047
xyz      0.1489      0.4676      0.0266
xyz      0.1411      0.4397      0.4504
```

Does your scoring scheme still work? Is it the best one for this dataset? Perhaps you might optimize your scoring scheme to simultaneously be good for both datasets.

Use your scoring scheme to solve the IF5a structure

Now that you have an optimized scoring scheme, use it to solve the IF5a structure. Make a new **Facts.list** that looks like this one, but with your scoring criteria:

```
force
create_scoring_table True
build True
```

```
ha_iteration False
resolution 2.5
sg "I 4"
cell 113.949 113.949  32.474 90.000  90.000  90.00
score_type_list CC RFACTOR SKEW
```

If you want to have the AutoSol Wizard iterate in finding additional heavy-atom sites and optimizing their positions using difference Fouriers after density modification, try adding the lines

```
ha_iteration True
max_ha_iterations 1
fix_xyz_after_denmod False
```

to your **Facts.list** file.

Now run AutoSol again now using the **restart** command as before:

```
phenix.runWizard AutoSol restart |tee AutoSol.log
```
You should obtain an **AutoSol_summary.dat** that lists the best solution found and the scores you obtained with it and where the model is located. You can look at the model with the command :

```
phenix.pymol
```

You can also look at all your results with a GUI if you type

```
phenix
```

and answer "yes" to the questions asked and then select **Wizards** and then double-click on **AutoSol** and then click on **RUN** in the lower right corner to reset the Wizard. You can then click on the **magnifying glass** icon at the top of the Wizard and it will give you some options of things to look at (phasing log, model, model and map, etc).