CHART: Automating MAD and MIR Structure Determination
a.k.a.
Pandora's Black Box

Paul Emsley

March 14, 2000

Abstract

CHART[1] is an automated structure solution program built around SHELX[1] and the CCP4 Program Suite[2]. CHART is suitable for both MAD and MIR structure dermination.

With the advent of modern synchrotron data collection facilites and tracing algorithms, it is feasable that the entire structure solution for routine structures, from data collection to model building can be done quickly and for free[3].

1 Preamble

In May 1998 I went to a meeting in Prague and happened to attend the presentation of Thomas Terwilliger (in which he was talking about *SOLVE*, obviously). During and afterwards, it became apparent that the CCP4 suite had several programs that were equivalent[4] to the program (parts) about which Thomas was speaking and thus could do something equivalent to *SOLVE*. We (CCP4 users) didn't need several scoring methods to examine heavy atom sets[5] because the figures of merit from MLPHARE[3] were good enough.

"Hmmm...", I thought, "CCP4 just needs a bit of glue code so that structure solution can be automated - sounds easy - should have it finished by October[6]."

2 Introduction

The starting point for CHART is to take the data sets in the form of MTZs (output of TRUNCATE) or SCALEPACK `.sca` files for the native and possible derivative data sets and run through the the steps of a typical structure solution. At the end, density modification (DM) is performed and one can choose to subsequently run ARP/wARP[4] to place atoms and trace the chain.

As you may have realised, CHART is inspired by Thomas Terwilliger's '*SOLVE*'[5]. The CCP4 Program Suite[2] has some excellent programs - they just needed some engineering to communicate with each other in a more extensive and coherent manner: CHART is an attempt to do that.

[1]How is this name derived? It it the anglisization of χαρτη but it has also been suggested[2]that it means "Crystallographic Heuristic Automated Research Tool" - which I rather like.
[2]by Richard Pauptit
[3]for academics.
[4]at the time.
[5]neither do *SOLVE* users, I understand.
[6]October 1998, that is.

The user interface of CHART was designed with 2 things in mind:

- Minimum input[7].
- No intervention.
  - The crystallographer should not have to make any decisions.
    - Which can be good and bad...
    - *e.g.* your native data set is 50% complete. The program does not stop to ask you whether this is what you expect. It presumes you do and gets on with it.

Figure 1: Target audience for CHART. (Figure reproduced without permission.)

2.1 Intended Audience

- Synchrotron Users:
  - Those irritated with editting endless command files,
  - Tired, stressed, bored, hungry, thirsty... (Figure 1). I hate editting all this stuff - when pressed for time, chronically tired but attention to detail is paramount[8].

[7]If the program could work things out for itself (*e.g.* column labels) then it does so.
[8]"Use the GUI[9] then," EJD would say. I would agree, it certainly helps.
[9]CCP4i, obviously.

- Novices:
  - Because CHART can be viewed as a complete black box and requires the absolute minimum of effort, it is ideal for:
    - crystallographic novices,
    - the lazy.

3 GUI

I have included a GUI (Figure 2) for which work is progressing to interact smoothly with the CCP4 GUI, CCP4i. The GUI is often used to enter the data set file names[10], column labels (if necesary) and number of residues.

3.1 Menubar Options

- wARP parameter selection
  - "Fast", "Slow", "Rfree"
- Various parameter selection
- Save and Open data description files
- Solvent optimization
- Progress Bar.

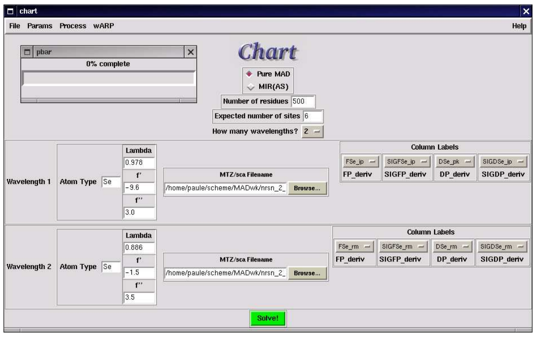

Figure 2: CHART GUI

3.2 Data Description

The interface between the GUI part of chart and the script that runs the structure solution is a simple text file, *e.g.*:

```
NRes 120
MIR
MTZ NATIVE  ~/data/caded_fhsc.mtz  F_nat SIGF_nat
MTZ Lu      ~/data/caded_fhsc.mtz  F_lu  SIGF_lu
MTZ U       ~/data/caded_fhsc.mtz  F_uo2 SIGF_uo2
```

In future, this file can be generated by data processing programs and the stucture solution can proceed from there with no user intervention.

[10] in SCALEPACK `.sca` or MTZ format.

4 Open the Box... Monsters!

However, the full horror of the command files and log files are available. Basically what CHART does is write `com` files, run the job, read the `log` files, decide what to do and then writes other command files, runs them... *etc* until the map is produced or it gives up in disgust at the quality of the data.

CHART is automating the CCP4 Suite because it strings together many of its components. The structure is divided into familiar components (programs, logs, coms) it is therefore easier to debug (if that's an issue).

It was clear from the initial design that I needed to be able to debug it. It was important therefore that all the log files, the command files, input and output coordinates, maps and data sets were easily available for later analysis. The flow of the structure solution could easily be monitored from watching how the output of one program is used in another because there is only one controlling program and it never[11] changes the names of files.

5 Refinement

When it comes to the heavy atom refinement, the philosophy is "Hold your nose and use MLPHARE". MLPHARE is not the best these days - but often it will do. Certainly, if you want to answer the question (while at the synchrotron): do the data I have collected provide with with an adequately interpretable map - and have that question answered in typically 20 minutes or so[12] - then the CCP4 Suite and CHART is an adequate solution. The Right Way, of course, is to use SHELX or *SOLVE*, and SHARP[6] and then some sort of ML-DM[7].

6 Scoring

6.1 SHELX scoring

Each solution set list written by SHELX is preceeded by several figures of merit. CHART uses the correlation coefficient to score SHELX solutions. There is atom seletion based on acceptable values of the patterson overlap function[13].

6.2 MLPHARE scoring

Scoring of the solution sets using the output of MLPHARE is at the heart of CHART. CHART uses the Cullis score, $\mathcal{C}$, (Figure 3) to score the solution set:

$$\mathcal{C} = r_w \sum_{i=1}^{Nbins} R_{cullis} - 1 \qquad \{R_{cullis} \leq 1\}$$

where:

$$r_w = \text{resolution bin width}$$

[11] well, almost never.
[12] and with minimal effort on your part.
[13] which is that PATSMF > 3.0.

$$
\begin{aligned}
\mathrm{Nbins} &= 10 \text{ (typically)} \\
R_{cullis} &= \frac{\Sigma_h \,||F_{PH} - F_P| - |F_{H_{calc}}||}{\Sigma_h \,||F_{PH}| - |F_P||} \\
|F_P| &= \text{the structure factor for the protein} \\
|F_{PH}| &= \text{the structure factor for the protein} \\
& \quad + \text{heavy atom} \\
|F_{H_{calc}}| &= \text{the calculated structure factor amplidude} \\
& \quad \text{forthe heavy atom solution set.}
\end{aligned}
$$

This[14] is the method by which different sets of solution (atom positions) are compared. The Cullis scores are multiplied by 1000 (for ease of reading) and written at the end of the mlphare log files.

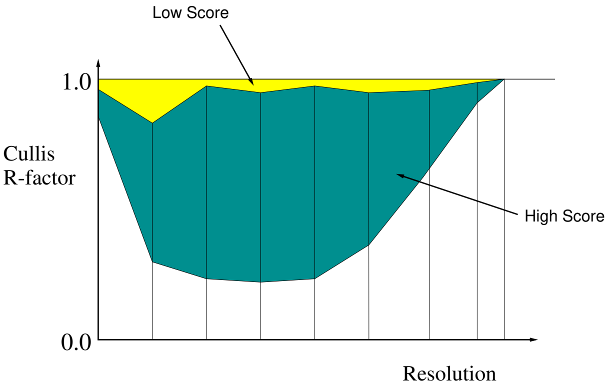

Figure 3: CullisArea score

6.3 Atom rejection

From where do the correct solutions to the heavy atom positions arrise? Initially, I thought that it would be a good idea to dig deep into the noise of potential solutions and refine each of them. CHART retains that posibility. However, in the tests it seems that the best solution is pretty much at the top of the list all the time, at all stages, *i.e.* post SHELX, post initial refinement, post additional (minor) atoms...

Therefore the new version of CHART chops the relatively poor-scoring SHELX solutions, using only the top dozen or so. In the case of MAD, only the top 5 phase sets are used for further refinement.

As I understand it, SOLVE takes great care to prepare only the right heavy atoms for subsequent refinement and phasing. CHART does not take such care - CHART relies on the MLPHARE refinement to judge the quality of atom solutions - poorly refining sites are rejected.

6.4 Enantiomorph Selection (Scoring the Maps)

Using the distribution of density in the map seems to be a useful way of scoring maps[8][15].

[14] EJD suggests that may not be the form of the Cullis R-factor calculated by MLPHARE.

[15] to be released in a later version (probably).

It used to be the case that the figures of merit from density modification programs were (badly) overestimated. In CHART, DM is used to determine the enantiomorph using the difference in figures of merit. This seems to work quite nicely, average differences of greater than 8% also have nice maps[16]. If the average figure of merit difference is less than 5 or 6% then even the best map is unsolvable (in my opinion)[17].

7 Execution Order

7.1 MIR

After the initial CAD, CHART puts the results in the following directories:

- `scale_1`
  - `reso_10_4`
    - SHELXS runs here
    - `mlphare`
  - `reso_12_6` *etc.*
- `scale_2` *etc*
- `diff-fouriers`
  - `mlphare/deriv_1`
    - `trial-combinations`
      - `mlphare-combined`
  - `mlphare/deriv_2` *etc.*
- `final-mlphare`
- `dm-ing`
  - `map`

7.2 MAD

- `revise`
- `rantan`
- `scale_1`
  - `reso_10_4`
    - SHELXS runs here
    - `mlphare`
- `mad-mlphare`
- `final-mlphare`
- `dm-ing`
  - `map`

8 ARP/wARP

Once CHART thinks that it has found an adequate map, it will, if requested then move on to ARP/wARPing the map. All you have to do is click the button on the GUI. Figure 4 is an example of the auto-tracing that ARP/wARP achieved (in about an hour or so):

[16] on the correct hand.

[17] but CHART is not so pesimistic, it says that "you may have solved the structure".

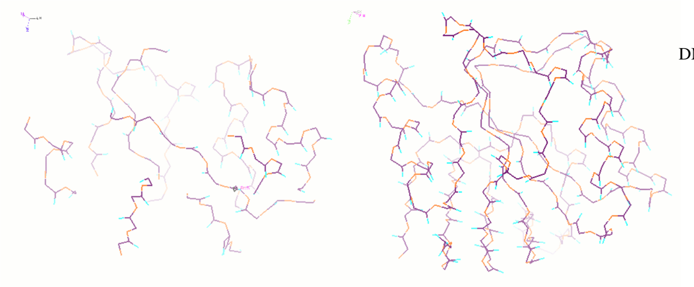

Figure 4: left: ARP/wARP traced coordinates (mode "FAST"); right: final refined coordinates (as deposited)[9].

9 Pop-Ups

To keep you amused, CHART prepares maps or shows graphs from log files for your edification. You can turn them off if you like. Occassionally they can be interesting.

10 MAD remote assignment

On the odd occasion, the data are mislabelled with the wrong wavelength. If you collect data

- from 4 (or more) wavelengths: then CHART can sort out the pseudo-native assignment for you.

- from 3 wavelengths: CHART makes a guess about what the wavelength assignment should be based on the isomorphous and anomalous differences. It then compares the differences there actually were to those expected from the f' and f'' values.

- from only 2 wavelengths: CHART has to rely on you giving it the right values of the wavelengths.

11 Optimization of Solvent

A recent addition is the "optimization" of solvent content for DM. Unusual packing of your protein means that the guess that CHART makes of the NCS[18] may be wrong. Therefore, you can choose to optimise the solvent content for DM by using the figures of merit to score the various phase sets. An example of the solvent content scores is shown in Figure 5.

12 Speed

Varies, of course. It depends on how many resolution ranges over which you try SHELX, the resolution and the number of sites.

Typically, However CHART takes 40 minutes to 20 hours per structure on a 300MHz Pentium II[19].

[18] non-crystallographic symmetry

[19] *i.e.* less than 20 minutes on today's computer for a typical 50kD protein (as you'll no doubt see).

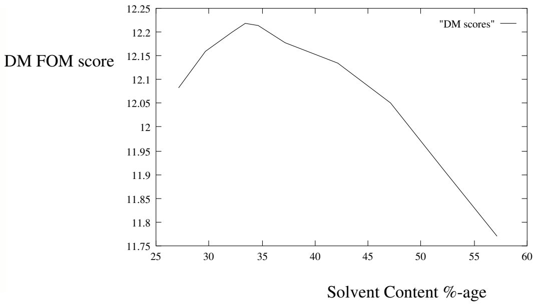

Figure 5: Total FOM score from DM *vs* solvent content percentage.

13 Will CHART solve my structure?

I don't know[20]. As far as I know, CHART can solve most solvable things. I use a suite of test data sets to check that the program works. New test datasets (especially MAD ones) are welcome.

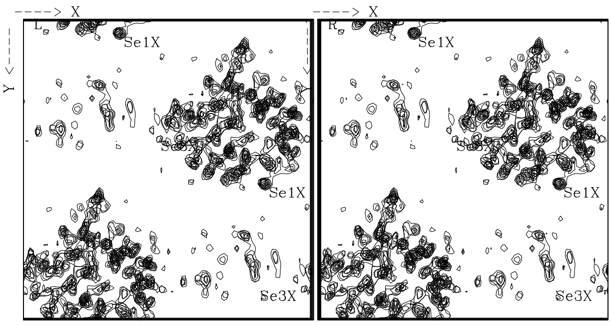

Figure 6: Post-DM map: Correct hand

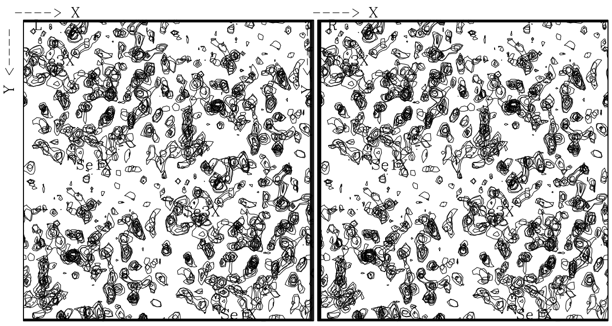

Figure 7: Post-DM map: Incorrect (*i.e.* relatively lower-scoring) hand. Density is lower in the "protein" region and higher in the solvent.

[20] I hope it will, of course.

14 Comparison vs. *SOLVE*?

- CHART (probably) doesn't solve structures as well as *SOLVE*.
- Free and easy to install.
- Faster (probably).
- CHART is not as careful in the atom selection as is *SOLVE*. *SOLVE* takes great care to get the heavy atom set correct before refinement - CHART does not. Instead it relies on incorrect atoms refining poorly in MLPHARE as the criterion for rejection.
- It is certainly useful to know that your sites are consistent through cross-phased difference fouriers.
- CHART will run ARP/wARP for you too.
- You can use CHART in the construction of nuclear weapons.

15 Expert Opinion

After you've spent countless hours designing, writing, debugging and testing your software, it's always nice when people write to you and tell you that your software worked and that they appreciated your efforts.

Here are some of the words of encouragement from colleagues and experts:

- "It's broken."
- "If it doesn't use DENZO format, we're not interested."
- "It's a pile of pants."
- "We'll have to buy '*SOLVE*."
- "No, no, no luvvie, you mustn't do that."
- "That's stupid."

16 More Details

16.1 Getting the Software

- CHART, the scheme interpreters and notes about compiling them can be downloaded from: `http://www.chem.gla.ac.uk/-paule/chart`
- Binaries for GNU/Linux and SGIs are provided for your convenience.
- But no Alpha version (for your inconvenience).
- Eventually, I will compile it so that you have to just download one executable.

16.2 Web interface?

- You upload the data sets.
- Coordinates[21] are mailed back to you.

16.3 Implementation

CHART is implemented in scheme[22]. It has turned out to be a good choice. Programming features, maths functions, easy type conversion, interpreted for speed of testing ideas, shell scripting. CHART can be optimized and compiled in the future (not that that will gain you much).

The initial design of CHART was to only write shell script, (*i.e.* write code that writes command files, runs them and analyses the output). This approach has got us where we are today - which is not so poor.

However, there are limitations. One of the crucial steps for automated procedures such as CHART is evaluating the quality of the map(s). Mostly today, the quality of experimental maps is judged by visual inspection, the figures of merit from programs such as DM are of little use. Simply put, there are no CCP4 programs that you can use to quantify the quality of maps. Recent publications[7, 8, 10] look interesting, I look forward to CCP4-compatible versions of these algorithms.

16.4 Mailing List

There is a mailing list should you wish to join: `chart-users@chem.gla.ac.uk`[23].

References

[1] G. M. Sheldrick. Phase Annealing in SHELX-90 - Direct Methods for Larger Structures. *Acta Crystallographica* **A**, 467 – 473 (1990).

[2] Collaborative Computing Project No. 4. A Suite of Programs for Protein Crystallography, distributed from Daresbury Laboratory, Warrington UK. *Acta Crystallographica Section D-Biological Crystallography* **50**, 760 – 766 (1994).

[3] Z. Otwinowski, In W. Wolf, P. R. Evans, and A. G. W. Leslie, editors, *Proceedings of the CCP4 Study Weekend 25-26 January 1991*, pages 80 – 85, Daresbury Laboratory, Warrington, UK., (1991). Science and Engineering Research Council.

[4] A. Perrakis, T. K. Sixma, K. S. Wilson, and V. S. Lamzin. Warp: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. *Acta Crystallographica Section D-Biological Crystallography* **53**, 448–455 (1997).

[21] and/or phases.
[22] a dialect of Lisp.
[23] mail `paule@chem.gla.ac.uk` and I will add you.

[5] T. C. Terwilliger and J. Berendzen. Automated MAD and MIR structure solution. *Acta Crystallographica Section D-Biological Crystallography* **55**, 849–861 (1999).

[6] E. de La Fortelle and G. Bricogne. *Methods in Enzymology*, volume 276, chapter Maximum-Likelihood Heavy-Atom Parameter Refinement for the Multiple Isomorphous Replacement and Multiwavelength Anomalous Diffraction Methods, pages 472–494. Academic Press, New York, (1997).

[7] T. C. Terwilliger. σ_R^2, a reciprocal-space measure of the quality of macromolecular electron-density maps. *Acta Crystallographica Section D-Biological Crystallography* **55**, 1174–1178 (1999).

[8] T. C. Terwilliger and J. Berendzen. Evaluation of macromolecular electron-density map quality using the correlation of local r.m.s. density. *Acta Crystallographica Section D-Biological Crystallography* **55**, 1872–1877 (1999).

[9] M. A. Walsh, I. Dementieva, G. Evans, R. Sanishvili, and A. Joachimiak. Taking MAD to the Extreme: Ultrafast Protein Structure Determination. *Acta Crystallographica Section D-Biological Crystallography* **55**, 1168–1173 (1999).

[10] T. C. Terwilliger and J. Berendzen. Discrimination of solvent from protein regions in native fouriers as a means of evaluating heavy-atom solutions in the MIR and MAD methods. *Acta Crystallographica Section D-Biological Crystallography* **55**, 501–505 (1999).