A comparison of some heavy-atom refinement and phasing programs in a test case

Kay Diederichs

Institut für Biophysik und Strahlenbiologie, Universität Freiburg, 79104 Freiburg, Albertstr. 23, Germany. e-mail: dikay@sun2.ruf.uni-freiburg.de

Abstract: The performance of some heavy atom refinement and phasing programs or packages (PROTEIN, VECREF/MLPHARE, HEAVY, DAREFI, PHASES, XtalView) was compared in a test case. As the success of heavy atom refinement and phasing heavily depends on the skills and experience of the investigator, and the resources for evaluation of more than one or two packages were not locally available, a proposal was posted in a Internet newsgroup (bionet.xtallography) and two mailing lists (CCP4 and O-info). Participants were given native and derivative diffraction data which had previously served to solve the structure, and used their favourite heavy atom refinement program to obtain the best possible heavy atom model and phases for the SIR and MIR case. The resulting phases were used to calculate the non-crystallographic symmetry (NCS) electron density correlation as a function of resolution (a measure that is already available during the heavy-atom refinement process once the NCS parameters are known) and were compared to the model phases; furthermore, the fit of the atomic model to the MIR maps was calculated. The results depend on both input to the programs and the programs themselves; with the data available these two factors of influence cannot be easily separated.

Introduction

Despite the increasing number of structures known and available for molecular replacement, the Multiple Isomorphous Replacement (MIR) method remains the only way to obtain an electron density map for a macromolecule with unknown structure. However, it is known that the joint refinement of heavy atom parameters (positions, occupancies and temperature factors) and the lack-of-closure errors that are associated with a given heavy-atom model is often ill-defined and the subject of theoretical as well as practical investigation (see, for example, contributions in the 1991 CCP4 Daresbury Study Weekend Proceedings). Still, there does not seem to exist a clear agreement about the 'right' way to calculate lack-of-closure errors and weighting of derivatives, if common sites are present. Also, scaling, choice of sigma cutoffs and inclusion of weak sites are parameters of influence. Therefore, all those of us who have tried to solve a macromolecular structure with the MIR method have at some point faced the question: what is the (in some sense) best heavy-atom refinement program (or package) and strategy to use?

In this paper, an attempt to answer this question for one case (see below) is described. It must be stressed that no evaluation of user friendliness, integration into an existing program package or special, unique features of some programs was attempted. Basically, only the agreement of phases and maps with those calculated from a refined atomic model and an internal indicator based on NCS were evaluated. Also, no attempt to understand the reasons for the different performance of the programs was made, as this would have required a thorough understanding of the programs and their underlying theory.

One way to overcome the problem of lack of installed programs and knowledge of operating these programs in my lab was to ask for help from the crystallographic community, utilising this wonderful means of communication, the Internet. After responding to a proposal posted to two Internet mailing lists (CCP4 and O-info) and one newsgroup (bionet.xtallography), five participants of the test copied the native and derivative diffraction data to their local computer site and ran the programs of their choice.

Materials and Methods

a) The test system

Granulocyte-Macrophage Colony-Stimulating Factor (GM-CSF) is a hormone of 127 amino acid residues that triggers the development of granulocyte and macrophage colonies from hematopoietic progenitor cells and can stimulate the immune response, similar to Interferon and Interleukin. Crystals grow in space group $P2_12_12_1$ (a=47.6Å, b=59.1Å, c=126.7Å) with two molecules per asymmetric unit. The structure was solved with MIR, non-crystallographic symmetry averaging and solvent flattening (Diederichs et al., 1991a). Iterative model building of poly-Ala/Leu chains and refinement enabled us to complete the structure and refine to 2.4 Å, R=20.5% (Diederichs et al., 1991b). The final model includes a few water molecules and can be obtained from the Protein Data Bank (accession code 1gmf; water molecules not included).

All diffraction data had been collected on a San Diego Multi-wire proportional Detector (Hamlin, 1985; Howard et al., 1985) and were of good quality (R_{sym} to 3Å between 5% and 6% at 3-fold redundancy) and completeness (90-99% to 3Å). The anomalous signal was about 60% complete, but rather bad.

b) Reasons why this is a good test case

The available data constitute a realistic test case because it is always easy to solve a structure with good derivatives but the bad derivatives seem to be rather the rule than the exception.

1. the heavy atom data are quite bad: the three Hg derivatives share the main sites and have low occupancy and substitution (R_{nat}=10% to 13%; higher soaking concentrations destroyed the crystals), and the one Pt derivative with high substitution (R_{nat}=29%), but anisomorphism. Consequently, the MIR-phased maps were of low quality and hardly interpretable.
2. the twofold non-crystallographic symmetry (NCS) can serve as a powerful internal indicator of phase/map quality.
3. the structure is known with relatively high accuracy so that the model phases can serve as an external reference; after several rounds of model-building any bias towards the initial MIR phases that were used to build a starting model is removed.
4. the space group is not polar so there are no possible shifts once the origin is fixed by choosing the main site of one derivative in a consistent way.
5. GM-CSF is a small protein which makes the problem easily tractable in terms of human and computer resources (reflection file size, CPU and wall clock time).

c) Rules of the game

The following data were passed to the testers:

1. native data to 2.4Å
2. 4 derivative data sets to be used out to 3.5Å, not scaled to the native data.
3. coordinates (no occupancies and temperature factors!) of the sites that were found in the original structure solution

The testers were asked to scale the data, check the original sites and given the option to come up with their own heavy atom model by difference fourier methods. Phase improvement by NCS averaging or solvent flattening should not be done, as this would have added a huge number of programs and parameters. To fix the origin, the main site of the first mercury derivative (p-chloromercuribenzenesulfonic acid) was to be chosen as in the original structure solution. SIR phases of the first Hg derivative (which allowed to find the sites of the other derivatives in the original structure solution) and MIRAS phases and corresponding heavy atom models should be returned to the author.

d) Programs/packages used for refinement and phasing

The following heavy atom refinement and phasing programs/packages were used by the testers, with the respective peculiarities indicated:

1. MLPHARE and VECREF version CCP4-2.5 (Collaborative Computational Project No. 4, 1994). Two alternatives were tested: MLPHARE for (both) refinement and phasing, and the combination: VECREF for refinement/MLPHARE for phasing (only). Furthermore, a second attempt with a slightly improved heavy atom model and better treatment of the anomalous signal was submitted by the same tester, again trying both alternatives. In all cases, the heavy atom model employed for phasing was changed with respect to the model given out.
2. HEAVY (Terwilliger and Eisenberg, 1983; Terwilliger, 1987). The heavy atom model employed for phasing used some but not all of the sites given to the testers. The anomalous signal of the Pt derivative, and only reflections in the 10-3.5Å resolution range were used.
3. PHASES (Furey and Swaminathan, 1990). The anomalous signal was not used for the MIR evaluation. The heavy atom model employed for phasing was not altered wrt. the one given to the testers.
4. PROTEIN Version 3.1-5d (Steigemann, 1991). The anomalous signal was not used for the MIR evaluation. The heavy atom model employed for phasing was not altered wrt. the one given to the testers. Only reflections in the 10-3.5Å resolution range were used.
5. XtalView 2.0 (McRee, 1993). Only the MIR results were returned. The heavy atom model employed for phasing was not altered wrt. the one given to the testers.
6. DAREFI - a modified version of a program according to Dickerson et al. (1968) whose roots are in Heidelberg (W. Kabsch) and is mainly used in the lab of G.E. Schulz and my lab. This was also the program that was used in the original structure solution. For this comparison, the heavy atom model was determined again, starting with the main site of the first mercury derivative, and had less sites than the model given out.

e) Methods and programs used for evaluation of results

As the classical indicators printed out by the programs (R-factors, figures of merit, phasing power etc.) are unreliable for a program-to-program comparison, the following protocol was used to assess the quality of the MIR phases without reference to the atomic model of GM-CSF.

- calculation of figure of merit (f.o.m.) weighted 'best' electron density map using reflections from ∞-6Å, 6-4.5Å and 4.5-3.5Å.
- calculation of the correlation of all electron density points in a 15Å sphere around fractional coordinates (0.3125, 0.1, -0.07) to those points related by NCS, namely in a sphere around (0.6074, 0.5479, -0.1728) after rotation by the Eulerian angles (43.8, 8.3, 357.5).

N.B. this procedure was also used during the original structure determination (Diederichs et al., 1991a) and proved very useful to help in choosing derivatives, picking up new sites from difference fouriers and monitoring the phasing process.

For the evaluation of figure-of-merit weighted phase differences to the phases derived from the atomic model, X-PLOR (Brünger, 1992) was used. Maps calculated in X-PLOR were used for correlating MIR electron density to the atomic model (real space correlation coefficient, using rsfit_all.o and 0.55, 0.85 for C and A0; only residues 10-122) in the program O version 5.9/5.10 (Jones et al., 1991). NCS density correlation was performed with NCSREF (see Diederichs, 1991a).

The inclusion of water molecules (not available in the Protein Data Bank coordinate set 1gmf.pdb) did not change the results significantly (data not shown). For the sake of reproducibility, all calculations involving the model were performed with 1gmf.pdb.

Results

The resolution-dependent internal and model-derived phase quality indicators are presented on figures shown on the next two pages. The first of these panels presents SIR evaluations, the second MIR data.

On each of the figures, the two plots of the bottom half compare the data obtained with VECREF/MLPHARE evaluations (as the upper plots would have become too crowded if they also had been included). Out of these, one curve (that using VECREF for the refinement and MLPHARE for phasing; first attempt) is also shown in the upper plots. The strategy of combining the two programs in this way seems to give somewhat better results than use of MLPHARE alone. However, the slightly changed heavy atom model and treatment of anomalous signal in the second attempt did not considerably improve the phases. The scatter in the curves of the bottom (VECREF/MLPHARE) plots is significantly lower than that of the plots comparing the other programs. This small amount of scatter gives an impression of how changes in the weak sites affect the phasing quality. In the following, the 'bottom plots' will not be discussed any further.

One should keep in mind that the quality of the SIR phases is crucial in the starting phase of a structure solution, as the success of difference fourier calculations (to find heavy atom sites of other derivatives) depends on them. Also, the SIR map might allow recognition of the NCS, which means a big leap forward as it allows NCS averaging and improves the solvent flattening process. On the other hand, the quality of the MIR phases determines the electron map interpretability. This means that it is not desirable to use a program that gives good SIR phases but does not use the whole information if more than one derivative is available. Vice versa, one will probably not arrive at a good MIR heavy atom model if the SIR phases and hence the difference fourier maps are bad.

a) Internal measures of phase quality: NCS correlation as a function of resolution

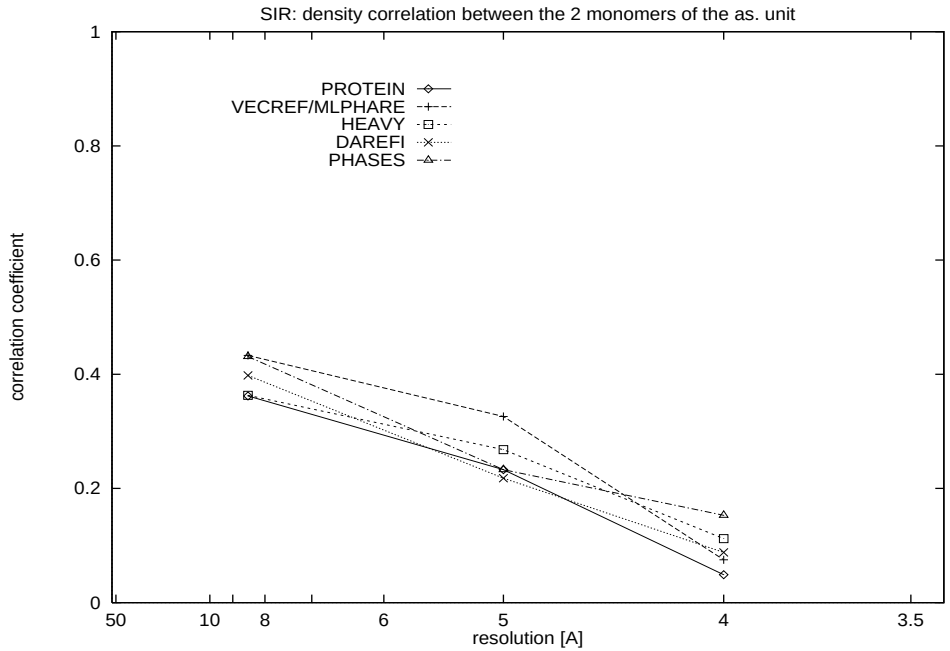
The correlation of the two crystallographically independent molecules of the asymmetric unit is shown on the left side of each panel. Of course, the curves that are **above** the others are the better ones. For the SIR case, the VECREF/MLPHARE combo outperforms the others at intermediate resolution and is the best, together with PHASES, at low resolution. PHASES shows the highest correlation at high resolution. For the MIR data, DAREFI is best, followed by PHASES.

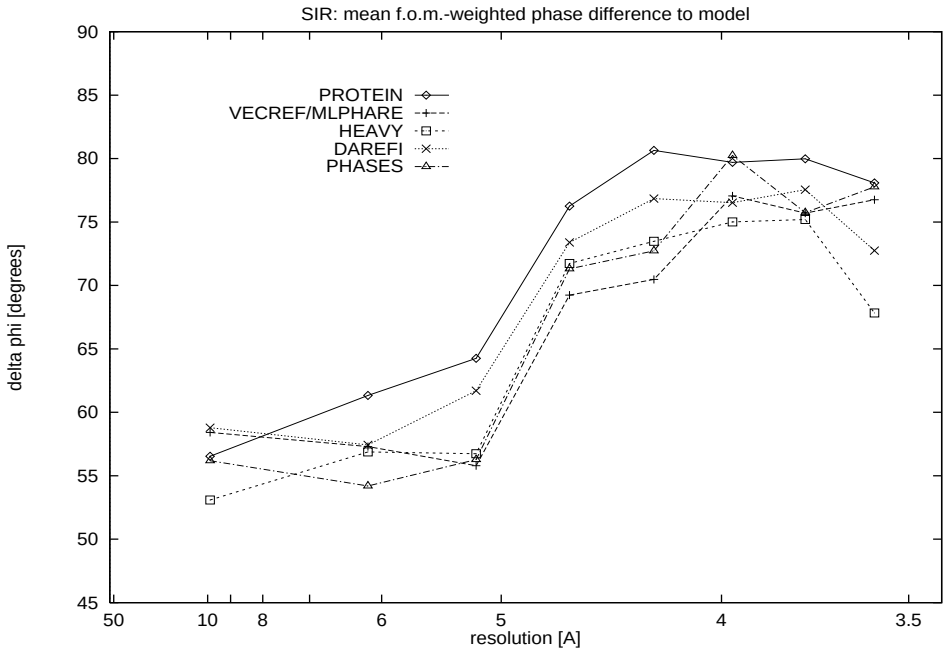
b) Comparison with external (model) phases

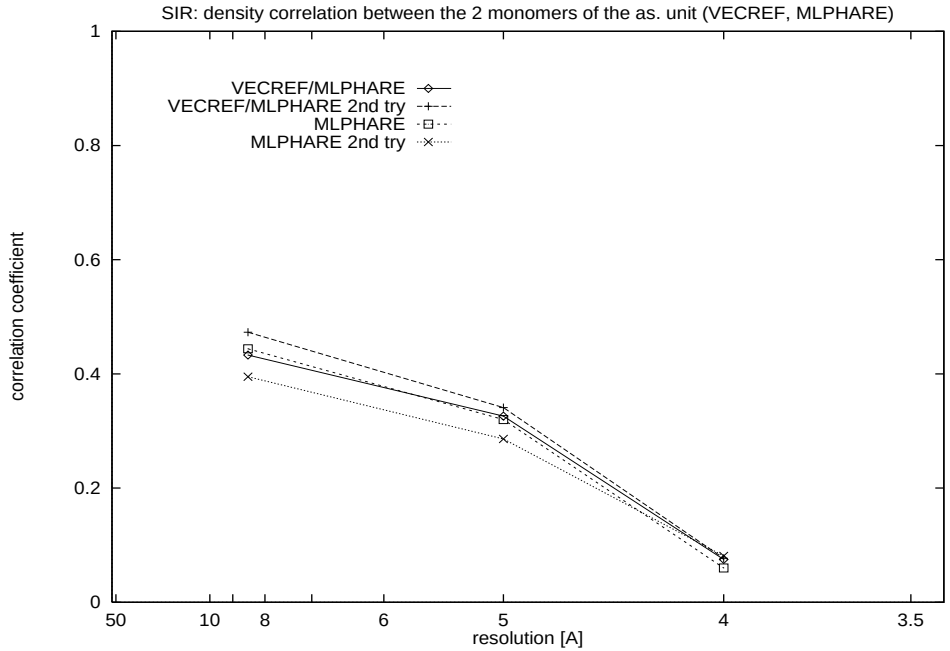
A comparison based on the model phases (1gmf.pdb) is shown on the right side of each panel. Here, the **lower** curves are the better ones. The deviation of figure-of-merit (f.o.m.) weighted mean phase difference among the programs is about 5 degrees, which is quite significant. SIR case: the phase differences beyond 4Å appear to drop which seems erratic. HEAVY, PHASES and VECREF/ MLPHARE perform about equally well. MIR case: PHASES, DAREFI and HEAVY yield good results.

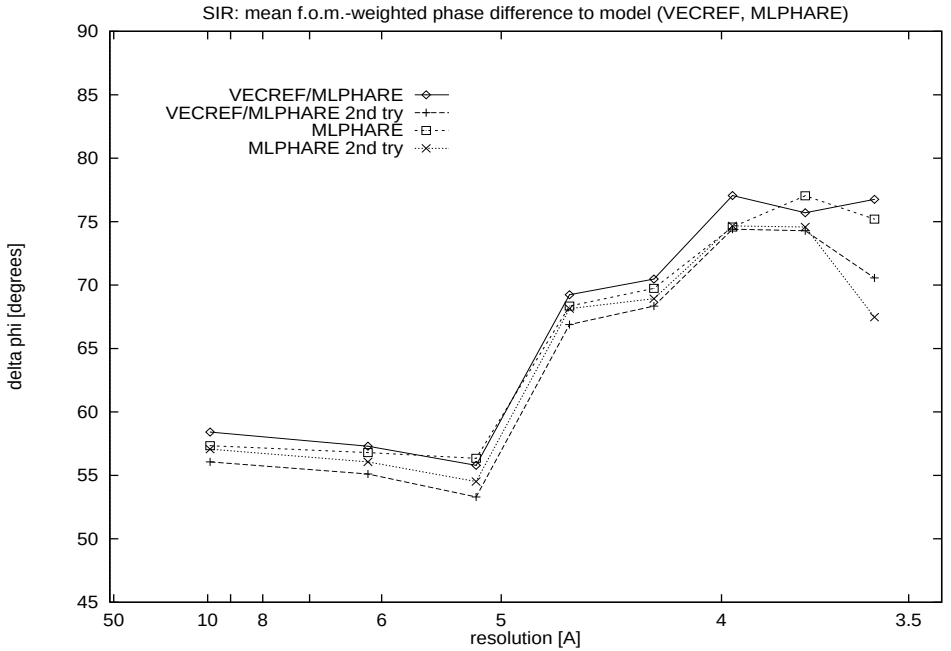
c) Real space correlation coefficient and other quantities

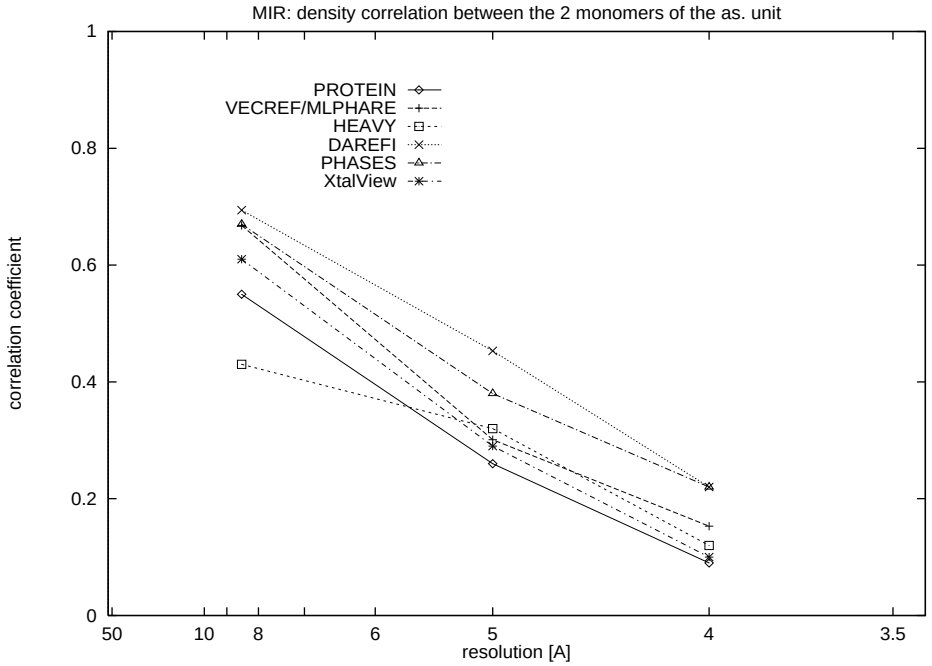
The MIR map interpretability is related to the agreement of the final atomic model with the MIR map, calculated as the mean value of the real space correlation coefficient which is computed in O on a per-residue basis. This quantity does **not** measure the same as the f.o.m. weighted phase difference, as into the density correlation the amplitudes of the reflections enter as well. It is given in the form of a table (after the figures). Here, DAREFI scores best, followed by VECREF/MLPHARE and PHASES. The standard deviation of this quantity is high, and a scatterplot (not shown) reveals that many residues have little or no density in all MIR maps.

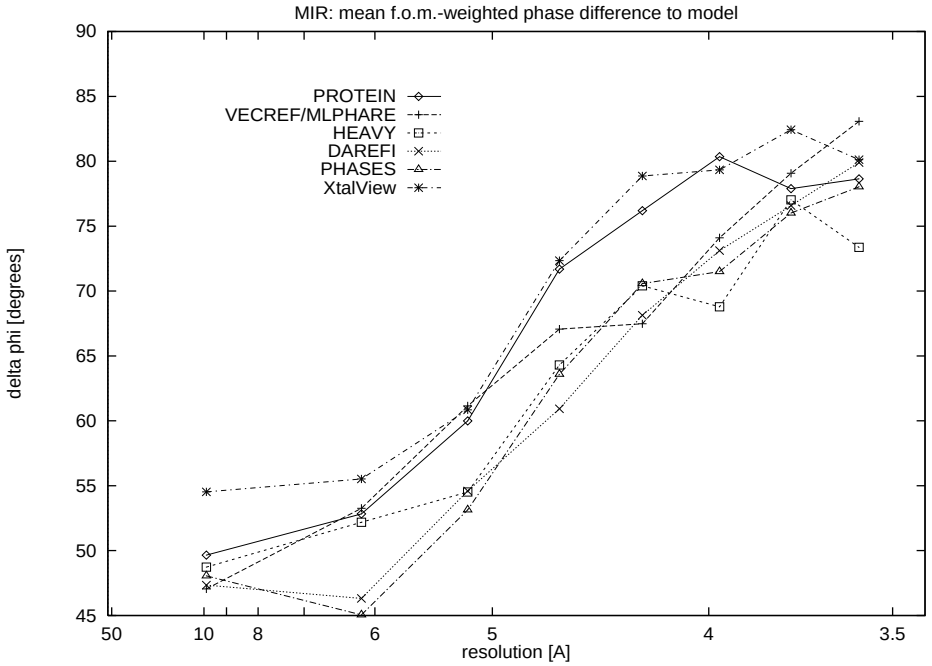

SIR: density correlation between the 2 monomers of the as. unit
PROTEIN
VECREF/MLPHARE
HEAVY
DAREFI
PHASES
correlation coefficient
resolution [A]
1
0.8
0.6
0.4
0.2
0
50
10
8
6
5
4
3.5

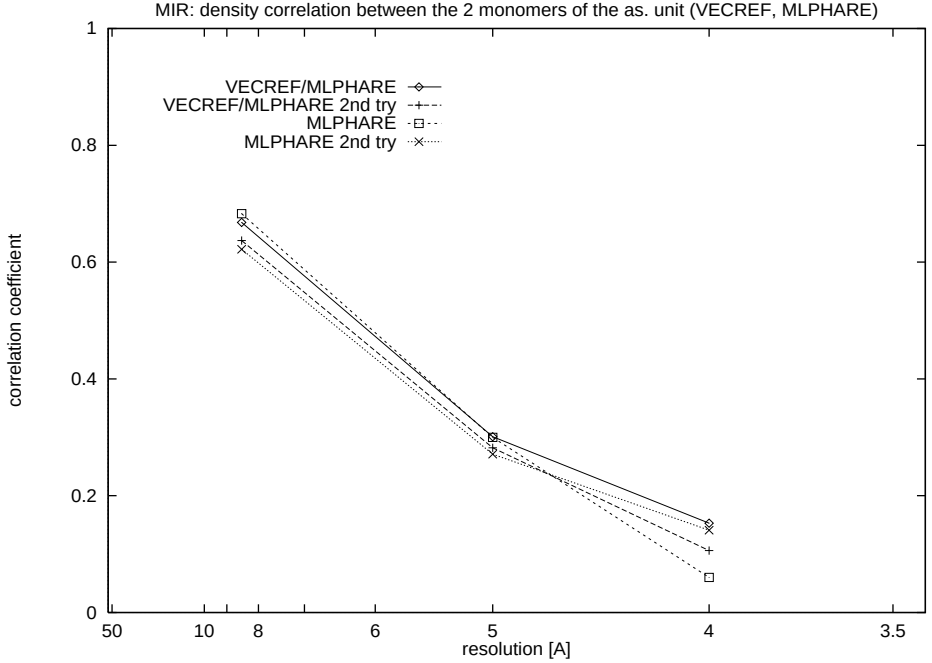

SIR: mean f.o.m.-weighted phase difference to model
PROTEIN
VECREF/MLPHARE
HEAVY
DAREFI
PHASES
delta phi [degrees]
resolution [A]
90
85
80
75
70
65
60
55
50
45
50
10
8
6
5
4
3.5

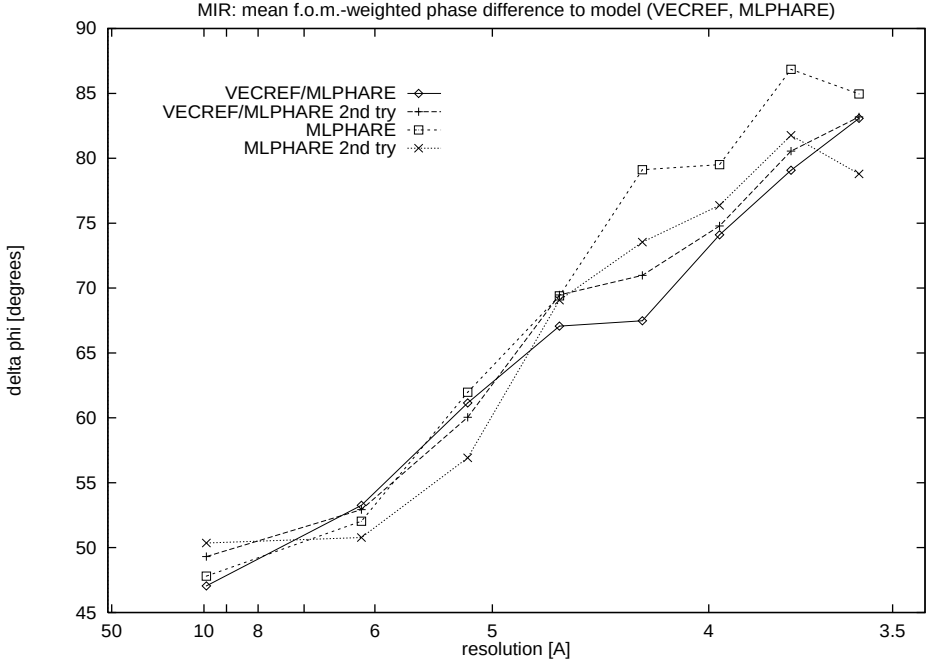

SIR: density correlation between the 2 monomers of the as. unit (VECREF, MLPHARE)
VECREF/MLPHARE
VECREF/MLPHARE 2nd try
MLPHARE
MLPHARE 2nd try
correlation coefficient
resolution [A]
1
0.8
0.6
0.4
0.2
0
50
10
8
6
5
4
3.5


SIR: mean f.o.m.-weighted phase difference to model (VECREF, MLPHARE)
VECREF/MLPHARE
VECREF/MLPHARE 2nd try
MLPHARE
MLPHARE 2nd try
delta phi [degrees]
resolution [A]
90
85
80
75
70
65
60
55
50
45
50
10
8
6
5
4
3.5


MIR: density correlation between the 2 monomers of the as. unit
correlation coefficient
resolution [A]
PROTEIN
VECREF/MLPHARE
HEAVY
DAREFI
PHASES
XtalView
1
0.8
0.6
0.4
0.2
0
50
10
8
6
5
4
3.5


MIR: mean f.o.m.-weighted phase difference to model
delta phi [degrees]
resolution [A]
PROTEIN
VECREF/MLPHARE
HEAVY
DAREFI
PHASES
XtalView
90
85
80
75
70
65
60
55
50
45
50
10
8
6
5
4
3.5


MIR: density correlation between the 2 monomers of the as. unit (VECREF, MLPHARE)
correlation coefficient
resolution [A]
VECREF/MLPHARE
VECREF/MLPHARE 2nd try
MLPHARE
MLPHARE 2nd try
1
0.8
0.6
0.4
0.2
0
50
10
8
6
5
4
3.5


MIR: mean f.o.m.-weighted phase difference to model (VECREF, MLPHARE)
delta phi [degrees]
resolution [A]
VECREF/MLPHARE
VECREF/MLPHARE 2nd try
MLPHARE
MLPHARE 2nd try
90
85
80
75
70
65
60
55
50
45
50
10
8
6
5
4
3.5

In the following table, further quantities are listed:

- the mean value of the f.o.m, as expected, has virtually nothing to do with the quality of the phases
- the 'correlation of the figure of merit with the phase error' (as printed out by X-PLOR) is a measure of the reliability of the figure of merit. If its absolute value is high, indeed the mean figure-of-merit weighted phase differences (<f.o.m.*Dphi>, for their dependence on resolution see the figures) are much lower than the non-weighted phase differences (<Dphi>).

		PROT-EIN	VEC-REF/ MLP-HARE	VEC-REF/ MLP-HARE 2nd try	MLP-HARE	MLP-HARE 2nd try	HEA-VY	DA-REFI	PHA-SES	Xtal-View
<real space correlation>	10-3.5 Å	0.338	0.481	0.464	0.419	0.453	0.390	0.532	0.489	0.389
	99-3.5Å	n.d.	0.566	0.546	0.501	0.529	n.d.	0.587	0.556	0.466
std. dev. of real sp.corr.	10-3.5Å	0.196	0.164	0.166	0.170	0.158	0.160	0.159	0.170	0.173
	99-3.5Å	n.d.	0.137	0.144	0.161	0.145	n.d.	0.149	0.153	0.165
<figure of merit>	10-3.5Å	0.506	0.446	0.429	0.526	0.442	0.429	0.600	0.556	0.853
	99-3.5Å	n.d.	0.457	0.439	0.532	0.450	n.d.	0.606	0.563	0.853
corr. (f.o.m., Dphi)	10-3.5Å	-0.163	-0.252	-0.239	-0.211	-0.216	-0.250	-0.237	-0.231	-0.101
	99-3.5Å	n.d.	-0.261	-0.250	-0.220	-0.223	n.d.	-0.238	-0.233	-0.107
<Dphi>	10-3.5Å	72.3	72.8	73.5	74.9	72.5	70.8	67.3	67.9	72.7
	99-3.5Å	n.d.	71.7	72.6	73.9	71.8	n.d.	66.8	67.0	71.3
<f.o.m.* Dphi>	10-3.5Å	67.0	63.1	64.1	68.1	64.5	61.6	61.3	61.6	70.9
	99-3.5Å	n.d.	61.7	62.8	66.8	63.5	n.d.	60.8	60.7	70.0

n.d. = not determined

For comparison, these are the results of refining and phasing in DAREFI with the heavy atom model sent out to the testers: (10-3.5Å) <real space correlation>=0.505, standard deviation of real space correlation=0.159, <figure of merit>=0.601, correlation (f.o.m.,Dphi)=-0.239, <Dphi>=67.9, <f.o.m.*Dphi>=61.7

The (10-3.5Å) values of those programs where the 99-3.5Å resolution range is present are simply calculated from the 10-3.5Å subset, without prior refinement.

Conclusions

Both internal and external indicators of phase quality agree very well, which means that an *a priori* determination of the usefulness of a heavy atom model and refinement strategy is possible. However, to the best of my knowledge, none of the refinement programs allows to monitor the NCS correlation during the refinement or to choose maximising the NCS correlation as a refinement target - a strategy which, of course, could only be applied if NCS is present and known. Also, as a proposal, I would suggest to integrate the concept of the free R-factor into the heavy atom refinement programs, to make the decision about whether to include weak sites easier.

Disclaimer: The statements in the following paragraph reflect the impressions of the author, based on this test case only. Other data sets and/or other testers might yield different results. As the results of the comparison probably depend as much on the person who chooses the refinement and phasing strategy as on the program, it is clear that it is hard to separate the influence of the persons from that of the programs based on the data available. In the following, the name of the program is assigned to a person using the program. It is not my intention to make anyone or a program look good or bad. The conclusions are meant to elicit further investigations and are not to be considered as final.

One thing this test certainly shows - there is quite a bit of variation in the results of a heavy atom refinement and phasing. As a summary, it appears that in this test case there appears to be a 'top group' consisting of PHASES, VECREF/MLPHARE and DAREFI. The good phase and map quality of PHASES was achieved without using the anomalous signal, which also demonstrates that the

anomalous signal of the derivative data does not carry much information. In the case of XtalView it seems to me that a problem consists in the lack of temperature-factor refinement. A surprise to me was that MLPHARE did not perform significantly better than some of the other programs; I had thought that the application of Maximum Likelihood theory would 'help the phases' more. Concerning the very good performance of DAREFI, one has to keep in mind that before the data were sent out, they had been selected for good electron density after phasing with DAREFI, so clearly the test is biased towards DAREFI and the program should give a reasonable result. It could be that some derivative data sets that were originally discarded would give better data when used with the other programs.

One weakness of the present study clearly is that the differences between the programs are to some extent obscured by different inputs to the programs, as is the case e.g. for the use of the anomalous signal or the resolution range used in the PROTEIN and HEAVY refinements - the results of these programs might be better if the refinement included the low-resolution reflections as well. This should be attributed to unclear statements about the tasks to perform and can be avoided in future studies.

It is planned to deposit the native and derivative structure factors with the Protein Data Bank, to facilitate comparisons with other or improved programs. This 'competition' could be considered as an ongoing test, with a follow-up - perhaps in the next CCP4 Newsletter?

Clearly, it is necessary and desirable to repeat the test with data sets from other projects.

Acknowledgements

The comparison described here would not have been possible without the help of Mingdong Huang (huangm@scripps.edu), Ingo Korndörfer (korndoerfer@genmic.biochem.mpg.de), Jay Pandit (pandit@pfizer.com), Thomas Terwilliger (terwill@prov2.lanl.gov) and Ian Tickle (tickle@mv3b.cryst.bbk.ac.uk). I want to thank them not only for their efforts which in some cases also resulted in removing bugs (features?) from the programs they used, but also for their patience in letting me write these things up after quite some time. Also, I would like to thank Eleanor Dodson for her efforts and discussion, and many people from the 'net' for encouragement and hints.

References

Brünger, A.T. (1992) X-PLOR Version 3.1 manual. Yale University Press, New Haven and London.

Collaborative Computational Project, Number 4 (1994) Acta Cryst. **D50**, 760-763.

Dickerson, R.E., Weinzierl, J.E. and Palmer, R.A. (1968) Acta Cryst. **B24**, 997-1003.

Diederichs, K., Jacques, S., Boone, T. and Karplus, P.A. (1991a) *J. Mol. Biol.* **221**, 55-60.

Diederichs, K., Boone, T. and Karplus, P.A. (1991b) *Science* **254**, 1779-1782.

Furey, W. and Swaminathan, S. (1990) PA33, American Crystallographic Association Meeting Abstracts, Series 2, Vol 18, p. 73.

Hamlin, R. (1985) in: Wyckoff, Hirs and Timasheff (eds.) *Methods in Enzymology* **114**, pp. 416-452.

Howard, A.J., Nielsen, C., Xuong, Ng.H. (1985) in: Wyckoff, Hirs and Timasheff (eds.) *Methods in Enzymology* **114**, pp. 452-472).

Jones, T.A., Zou, J.-Y., Cowan, S.W. & Kjeldgaard, M. (1991) *Acta Cryst.* **A47,** 110-119.

McRee, D.E. (1993) Practical Protein Crystallography. Academic Press, San Diego, California.

Steigemann, W (1991) Int. Union Cryst., Crystallogr. Symp. 5 (Crystallogr. Comput. 5), 115-125.

Terwilliger, T.C. (1987) *Acta Cryst.* **A43**, 6-13.

Terwilliger, T.C. and Eisenberg, D. (1983) *Acta Cryst.* **A39**, 813-817.