Tutorial on how to solve a Se-substructure using SHELXD

Thomas R. Schneider
Dept. of Structural Chemistry
University of Göttingen
trs@shelx.uni-ac.gwdg.de

July 4, 2002

1 Introduction

The Solution of the substructure during the course of a MAD-phasing experiment consists of four steps:

1. Evaluation of the data quality. In particular, it is important to determine the maximum resolution for which signficant anomalous signal is present
2. Derivation of substructure factors. These can be ΔF's or F_A's. Which of the two is the better choice depends on the circumstances and still is a matter of debate.
3. Solving the substructure. Different methods (Patterson and/or Direct Methods) can be used to find the anomalous scatterers.
4. Validation of the substructure sites. As some phasing algorithm require substantial computing resources it is advisable to check which of the sites found in the previous step are likely to be correct.

For steps 1 and 2, we like to use the program Bruker-AXS program XPREP (for information about this program contact Sue Byram at the the following email-address: `SByram@bruker-axs.com`). XPREP is started from the unix-command line by simply typing `xprep`.

For steps 3 and 4, we use the program SHELXD, for which information is available at: `http://shelx.uni-ac.gwdg.de/SHELX`.

Since early 2002, there is a new program available that will calculate phases once the substructure is solved. It is called SHELXE and you can find information about it at: `http://shelx.uni-ac.gwdg.de/SHELX/shelx_de.pdf`.

2 Some General Points

The following points apply to all examples:

- If possible, it is better to begin with scaled but unmerged data (e.g. produced by SCALEPACK using the `NO MERGE ORIGINAL INDEX`-keyword). Also the measurements for the systematic absences should be kept. This is important as most of the statistics that is useful for judging data quality can not be calculated anymore once the data are merged and/or systematic absences are removed.
- I prefer to use the data collected at the high energy remote wavelength as the reference, as data collected at this energy are the least susceptible to systematic errors caused by wavelength instabilities (the anomalous contribution to the structure factor stays more or less the same even if the wavelength drifts or fluctuates during the experiment).
- The program XPREP needs some starting values for f' and f'' in order to derive F_A-values. The set of starting values given in table 1 can be used for cases where no experimental values are available (or the experimental values where obviously wrong ...). These values are simple guesses, nevertheless they have worked for many cases.

- To learn the interpretation of a Patterson crossword table, you should read the following article: *G.M. Sheldrick (1997) Meth.Enzym. 276:628.*

λ	f'	f''
hrm	-3	2.5
pk	-6	5.0
ip	-7	3.5
lrm	-2	0.5

Table 1: Rough guesses for f' and f'' for data typically collected in a Se-Met MAD-experiment.

3 Example SS

To run this tutorial, you need to have two programs: SHELXD and XPREP and three files containing the data collected at the three different wavelengths: `sse1.sca`, `sse2.sca`, and `sse3.sca`.

3.1 Information available

SS is a protein with $\approx$ 200 residues and a molecular weight of $M_r \approx 22$ kDa. Crystals used where of space group $P4_122$ with the following unit cell parameters (taken from the hrm data set): $a = b$ =58.244 Å, c = 251.467 Å, $\alpha = \beta = \gamma$ = 90.0°. The asymmetric unit contains 2 molecules and the Se-substructure consists of 2 $\times$ 7 sites. Data were collected at three wavelengths and where stored as SCALEPACK .sca-files with symmetry related reflections merged.

λ [Å]	d_{min} [Å]	filename
pk	2.5	sse1.sca
ip	2.7	sse2.sca
hrm (13.2 keV)	2.7	sse3.sca

3.2 Data Quality and Substructure Structure Factors

Normally we would only scale and not merge the data during data processing (keyword `NO MERGE ORIGINAL INDEX` in SCALEPACK) in order to do a full analysis of the data using the program XPREP. This analysis would include the analysis of unit cell parameter, systematic absences and the determination of the space group.

In this case, the data have already been merged the space group $P4_122$ so that we can leave out the space group determination in XPREP. We have, in fact, to tell XPREP what the space group is as this decision can not be made on merged data.

After starting XPREP, we first read in the high energy remote data set (which we will use as the reference data set later). After moving through several menues, we finally arrive in the menu '`[D] Read, modify or merge DATASETS`'. By choosing the '`S`'-option we can generate various types of statistics. Following is a table for the high energy remote data with Friedel pairs kept separate. Note that the R_{int} column is empty as the data had been merged before.

Resolution	#Data	#Theory	%Complete	Redundancy	Mean I	Mean I/s	Rint	Rsigma
Inf - 7.20	1165	1229	94.8	0.95	752.4	54.11		0.0202
7.20 - 5.75	1146	1159	98.9	0.99	360.5	56.97		0.0175
5.75 - 5.00	1230	1236	99.5	1.00	421.8	41.37		0.0274
5.00 - 4.55	1159	1166	99.4	0.99	617.4	30.38		0.0319
4.55 - 4.20	1281	1286	99.6	1.00	569.2	29.84		0.0323
4.20 - 3.95	1215	1218	99.8	1.00	526.0	26.87		0.0358
3.95 - 3.75	1226	1234	99.4	0.99	454.8	26.28		0.0357
3.75 - 3.55	1514	1517	99.8	1.00	362.8	27.41		0.0334
3.55 - 3.40	1387	1390	99.8	1.00	285.4	26.16		0.0332
3.40 - 3.25	1641	1646	99.7	1.00	212.4	20.82		0.0408

```
3.25 - 3.15    1288   1293     99.6     1.00     148.3   18.36          0.0465
3.15 - 3.05    1433   1436     99.8     1.00     132.8   17.82          0.0478
3.05 - 2.95    1663   1667     99.8     1.00      99.5   15.36          0.0566
2.95 - 2.85    1897   1900     99.8     1.00      80.2   13.98          0.0633
2.85 - 2.75    2170   2174     99.8     1.00      64.3   11.85          0.0750
2.75 - 2.70    1244   1247     99.8     1.00      60.6   11.06          0.0810
2.70 - 2.70      14     39     35.9     0.36      34.9    5.19          0.1845
-------------------------------------------------------------------------------
2.80 - 2.70    2375   2405     98.8     0.99      62.1   11.24          0.0791
 Inf - 2.70   22673  22837     99.3     0.99     293.6   25.15          0.0342

Merged [S],  lowest resolution = 36.96 Angstroms,      0 outliers downweighted
```

The data are very complete, and have a good signal/noise (Mean I/s in the table) up to the highest resolution measured ($I/\sigma(I)$ is still larger than 10 at the high resolution limit). However, the $I/\sigma(I)$ given in this table has to rely on the sigmas attached to the data given to the program being correct - i.e. if the error model in SCALEPACK has not been adjusted correctly, the estimation of $I/\sigma(I)$ in this table will of course be wrong.

Typing RETURN after seeing this table will bring us back into the data treatment menue:

```
Index   # Data   Filename or Source of Data

   1    22673    sse3.sca  <- current dataset

[M] Sort-MERGE current data (no scaling)     [C] Change CURRENT dataset
[L] LEAST-SQUARES scale and merge datasets   [W] WRITE dataset to file
[I] INCLUDE Rfree flags from another file    [R] READ in another dataset
[S] Display intensity STATISTICS             [D] DELETE stored dataset
[F] FACE-indexed absorption corrections      [P] PSI-scan absorption corr.
[T] Copy file, TRANSFORM hkl and cosines     [A] MAD, SAS, SIR or SIRAS
[H] Apply HIGH/low resolution cutoffs        [N] NORMALIZE/scale sigmas
[G] Generate simulated powder diagrams       [E] EXIT to main menu
[Q] QUIT program
```

where we can use 'R' to read in the other two data sets. My personal system is to always read the wavelength in the order hrm, pk, ip, lrm (there is no reason for this apart from being less confusing). After reading the pk- and ip-data, the list of data sets looks like:

```
Index   # Data   Filename or Source of Data

   1    22673    sse3.sca
   2    28236    sse1.sca
   3    22431    sse2.sca  <- current dataset
```

Now we can analyse the data using the 'A'-option (for 'MAD, SAS, SIR or SIRAS' followed by the 'M'-option to do a multiple wavelength analysis. After chosing data set 1 as the reference, and providing estimated f' and f'' values for this wavelength (the values given have to be only roughly correct. I always use the numbers given in the Table on page 3), XPREP answers with a bit of statistics.

```
High resolution limit in Angstroms for this calculation [0.0]:

I/sigma threshold for rejecting (after merging) [0.5]:

Target number of reflections in local scaling sphere
(0 if no local scaling) [100]:
```



```
Number of remote or native dataset to which rest will be scaled [1]:

Enter f' and f" for this wavelength: -3 2.5

Anomalous signal/noise ratios (1.0 is random).  The first line is based on
input sigmas, the second on variances of F+ and F- (if not already averaged):
Inf - 8.0 - 6.0 - 5.0 - 4.3 - 4.1 - 3.9 - 3.7 - 3.5 - 3.3 - 3.1 - 2.9 - 2.7 A
   5.99  6.65  3.94  2.56  2.03  2.10  2.08  2.19  2.04  1.99  2.00  1.85

   81.8  Neighbors used on average for F+/F- local scaling
Rint(anom) =   0.0764  before and   0.0761  after local scaling
```

As the data had been merged before, the analysis of the signal/noise ratio for the anomalous difference again has to rely on the input sigmas being correct. If the data had been given unmerged, XPREP would print a second set of numbers related to the true variance of the various F^+ and F^- values. Normally it safe to use data for which the signal/noise ratio is larger than **1.5**. So, for this case, all data could be used.

The next step includes the same analysis for the peak data and also evaluates the correlation coefficient between the ΔF's for the two data sets. This correlation coefficient should be larger than 30 % for good data:

```
Number of next MAD dataset (<CR> if none): 2

Enter f' and f" for this wavelength: -6.0 5.0

Anomalous signal/noise ratios (1.0 is random).  The first line is based on
input sigmas, the second on variances of F+ and F- (if not already averaged):
Inf - 8.0 - 6.0 - 5.0 - 4.1 - 3.9 - 3.7 - 3.5 - 3.3 - 3.1 - 2.9 - 2.7 - 2.5 A
   4.29  4.67  4.73  3.98  3.79  3.25  3.09  2.96  2.84  2.86  2.27  1.75

   79.9  Neighbors used on average for F+/F- local scaling
Rint(anom) =   0.1403  before and   0.1378  after local scaling

   69.8  Neighbors used on average for local scaling to native/remote dataset
Rint =   0.0967  before and   0.0923  after local scaling

Anomalous correlation coefficients (%) against previous datasets
Inf - 8.0 - 6.0 - 5.0 - 4.1 - 3.9 - 3.7 - 3.5 - 3.3 - 3.1 - 2.9 - 2.7 - 2.5 A
   92.2  91.4  89.6  85.6  83.1  76.5  73.9  74.6  65.4  56.2  40.6  76.2
```

Again, all data are usable. Also, as expected, the signal/noise is larger for the anomalous differences measured at the peak (maximum f'') than at the high energy remote (intermediate f'').

Finally, we check the third wavelength, the inflection point:

```
Enter f' and f" for this wavelength: -7 3.5

Anomalous signal/noise ratios (1.0 is random).  The first line is based on
input sigmas, the second on variances of F+ and F- (if not already averaged):
Inf - 8.0 - 6.0 - 5.0 - 4.3 - 4.1 - 3.9 - 3.7 - 3.5 - 3.3 - 3.1 - 2.9 - 2.7 A
   4.24  4.75  3.08  1.96  1.60  1.50  1.63  1.77  1.86  1.72  1.76  1.72

   82.1  Neighbors used on average for F+/F- local scaling
Rint(anom) =   0.0592  before and   0.0587  after local scaling

   83.1  Neighbors used on average for local scaling to native/remote dataset
Rint =   0.0964  before and   0.0954  after local scaling

Anomalous correlation coefficients (%) against previous datasets
```

```
Inf - 8.0 - 6.0 - 5.0 - 4.3 - 4.1 - 3.9 - 3.7 - 3.5 - 3.3 - 3.1 - 2.9 - 2.7 A
   74.2  78.6  71.0  71.3  63.3  69.0  63.7  55.7  54.5  48.4  35.4  23.1
   81.4  80.2  73.8  75.7  71.6  74.3  67.3  65.1  60.5  60.9  47.6  36.0
```

Here we find the smallest signal/noise ratio for the anomalous signal. All correlation coefficients are larger than 30 % - the data do not need to be cut for the following steps.

The next step is the refinement of f' and f'' at the various wavelength in order to extract F_A-values that are as consistent as possible with all data collected. The refinement starts from the input values:

```
Number of next MAD dataset (<CR> if none):

Set      f'      Rf'        f"      Rf"     after MAD fit
 1    -3.000   0.0191     2.500   0.0420    sse3.sca
 2    -6.000   0.0761     5.000   0.0335    sse1.sca
 3    -7.000   0.0570     3.500   0.0477    sse2.sca

It is not possible to refine all f' and f" simultaneously, but with very
precise data it may be worth refining individual values against the rest
```

and normally converges after a few cycles.

```
 Enter N to refine f' and f" of set N or <CR> for no (further) refinement: 3

Set      f'      Rf'        f"      Rf"     after MAD fit
 1    -2.759   0.0063     3.524   0.0398    sse3.sca
 2    -6.527   0.0694     6.982   0.0229    sse1.sca
 3    -6.900   0.0631     2.453   0.0373    sse2.sca

Accept new estimates of f' and f" (A) or reinstate previous (R) [R]: a

Enter N to refine f' and f" of set N or <CR> for no (further) refinement:
```

The absolute values of the refined f' and f'' are not that important. As long as the final values make sense relative to each other (i.e. largest f'' for the peak, smallest f' for the inflection point etc.), we are on safe ground. If, for example, the inflection point data assume the highest f''-value after refinement, probably something went wrong during data collection or maybe filenames were mixed up (happens more often than people want to believe ...).

The following question can all be answered with RETURN, at some point the filename for the file containing the F_A-values has to be given. After the filename has been given, XPREP can also generate an input file to be used later in SHELXD (the number for the wavelength is again a rough estimate, 1.0 Å would work as well).

```
Write .ins file for SHELXD/XD (Y or N)? [Y]:

Filename [sse_fa1.ins]:

Element type for heavy atoms [Se]:

Number of unique heavy atoms [12]: 14

Wavelength [1.54178]: 0.98


File sse_fa1.ins set up as follows:

TITL sse_fa1 in P4(1)22
```



```
CELL 1.00000  58.2440  58.2440 251.4670  90.000  90.000  90.000
ZERR   16.00   0.0082   0.0082   0.0503   0.000   0.000   0.000
LATT -1
SYMM -X, -Y, 0.5+Z
SYMM -Y, X, 0.25+Z
SYMM Y, -X, 0.75+Z
SYMM -X, Y, -Z
SYMM X, -Y, 0.5-Z
SYMM Y, X, 0.75-Z
SYMM -Y, -X, 0.25-Z
SFAC SE
UNIT   224
PATS
FIND  14
MIND -3.5
HKLF 3
END

Enter <CR> to continue
```

We can now stay in the XPREP-session and also create a file containing the anamolous differences for the peak wavelength, ΔF_{pk}. First we have to change the active data set to be the pk-data, i.e. data set number 2. Then we choose the 'A'-option from the '`MAD, SAS, SIR or SIRAS`'-menue and the same statistics as before is shown.

```
Select option [E]: a

 [M] MAD (Multiple-wavelength Anomalous Dispersion)
 [I] SIR (Single Isomorphous Replacement)
 [A] SAS (Single-wavelength Anomalous Scattering)
 [R] SIRAS (Single Isomorphous Replacement with Anomalous Scattering)
 [E] EXIT to main menu
 [Q] QUIT program

 Select option [E]: a

 High resolution limit in Angstroms for this calculation [0.0]:

 I/sigma threshold for rejecting (after merging) [0.5]:

 Target number of reflections in local scaling sphere
 (0 if no local scaling) [100]:

 Anomalous signal/noise ratios (1.0 is random).  The first line is based on
 input sigmas, the second on variances of F+ and F- (if not already averaged):
 Inf - 8.0 - 6.0 - 5.0 - 4.1 - 3.9 - 3.7 - 3.5 - 3.3 - 3.1 - 2.9 - 2.7 - 2.5 A
    4.29  4.67  4.73  3.98  3.79  3.25  3.09  2.96  2.84  2.86  2.27  1.75

    79.9  Neighbors used on average for F+/F- local scaling
 Rint(anom) =   0.1403  before and   0.1378  after local scaling

 High resolution limit in Angstroms for saved data [0.000]:

 Enter effective B-value (e.g. 20) to normalize delta-F or Fa values,
 <CR> for no renormalization:

 Filename to write FA and phi(T)-phi(A) (<CR> for none): sse_df1.hkl

 Current dataset contains  11586 SAS delta(F)

 Write .ins file for SHELXD/XD (Y or N)? [Y]: y
```

If we type 'Y' here, the corresponding ins-file will be written as well.

If we want, we can still have a look at all the data sets:

```
Index   # Data   Filename or Source of Data

   1    22673    sse3.sca
   2    28236    sse1.sca
   3    22431    sse2.sca
   4    14575    MAD Fa-values -> sse_fa1.hkl
   5    11586    SAS delta(F) -> sse_df1.hkl  <- current dataset
```

There is a number of other ways to look at the data and the substructure structure factors in XPREP, one of the possibilities being to look at Patterson maps. In the case of F_A or ΔF-values, it can be very useful to compare the respective Patterson with respect to the level of noise and whether or not the maps are similar.

After playing a bit with our data, we finally quit the program.

3.3 Substructure Solution using default parameters

Normally, I deliberately cut the input structure factors at 3.0 Å even if the data are usable to a higher resolution (CC $>$ 30% and signal/noise for ΔF's $>$ 1.5). This is mostly to make the program run faster with less reflections (impatience ...).

To run this default scenario a line '`SHEL 999 3.0`' has to be inserted into the .ins-file. For this tutorial I also limit the number of tries to 100 ('`NTRY 100`') to avoid flooding of the workshop computers. The modified script looks like this:

```
TITL sse_fa1 in P4(1)22
CELL 1.00000  58.2440  58.2440 251.4670  90.000  90.000  90.000
ZERR   16.00   0.0082   0.0082   0.0503   0.000   0.000   0.000
LATT -1
SYMM -X, -Y, 0.5+Z
SYMM -Y, X, 0.25+Z
SYMM Y, -X, 0.75+Z
SYMM -X, Y, -Z
SYMM X, -Y, 0.5-Z
SYMM Y, X, 0.75-Z
SYMM -Y, -X, 0.25-Z
SFAC SE
UNIT   224
SHEL 999 3.0
PATS
FIND  14
MIND -3.5
NTRY 100
HKLF 3
END
```

Running SHELXD using this script takes ca. 17 minutes on a Pentium III at 800 MHz and produces 6 trys with correlation coeffecients CC larger than 40.0 (my personal lower limit for a solution). The best solution has the following statistics:

```
PSUM 28.57  PSMF Peaks: 48 47 47 37 32 31 31 29 29 26 26 25 24 21 20 20 19
Try 47:20  Peaks 99 98 98 97 88 74 72 67 61 58 58 52 49 48 40 39 36 33 33 31
R = 0.370, Min.fun. = 0.515, <cos> = 0.243, Ra = 0.484
Try    47, CC All/Weak 44.70 / 33.37, best 44.70 / 33.37, best PATFOM  4.22
```

The number of solutions per hour (6 solutions in 17 minutes $\rightarrow$ 21 solutions per hour) for this relatively small substructure is somewhat disappointing. This has two reasons: (1) The

job	d_{min}	E_{min}	$\#E$	# sol	CPU	CC	$sh_{1.0}$
sse_fa1a	3.0	1.5	1028	4	983	44.5/33.2	18
sse_fa1c	3.0	1.3	1785	8	1083	45.0/27.8	33
sse_fa1b	2.7	1.5	1381	14	1094	42.8/32.4	57
sse_fa1g	2.7	1.4	1841	17	1152	42.8/29.1	66
sse_fa1d	2.7	1.3	2377	23	1240	43.0/26.9	83
sse_fa1f	2.7	1.2	3028	27	1320	43.2/23.2	93
sse_fa1e	2.7	1.1	3742	19	1440	43.0/19.8	59
sse_pk1a	3.0	1.5	977	9	982	48.4/26.0	41
sse_pk1b	2.5	1.5	1674	7	2880	42.6/22.7	11

Table 2: Results of SHELXD-runs using different numbers of E-values. d_{min} is the high resolution cutoff applied in Å, E_{min} the smallest E-value used, $\#E$ number of E-values used, # sol number of solutions in 100 tries, CPU is CPU-time in seconds on a Pentium-III running at 800 MHz, CC is the correlation coefficient between observed and calculated E-values for strong (used) and weak (not used) reflections for the best try in percent, $sh_{1.0}$ is the number of solutions per hour normalized to a Pentium-III running at 1.0 GHz. Jobs with names ending in `_pk??` were run against ΔF_{pk}'s

calculation of full-symmetry Patterson minimum functions takes more computer time in high-symmetry space groups than in low symmetry space groups and (2) the number of Se-atoms per given volume of the asymmetric unit is abnormally high (in other words: we have many more Methionines in the sequence than we would normally expect) for this case. This results in an unfavorable ratio of the number of the structure factors to atoms to be found, or in other words bad parameter to observables ratio.

The correlation coefficient for the weak data (CC_{weak} = 33.47%) is also higher than normally expected. The reason for this is that the data are of unusually high quality.

3.4 Substructure Solution using non-default parameters

There are two ways to increase the number of E-values used in SHELXD: (1) the resolution cutoff can be moved to higher resolution (this is possible here, as the data are of high quality all the way to 2.7 Å). and (2) the lower limit for E-values, E_{min} can be lowered in order to use more reflections in the Fourier-recycling. The results for some different settings are shown in Table 3.4

In this case, the increase in CPU-time per try due to the larger number of reflections used is more than balanced by the higher probability of producing a solution. The best set of parameter (d_{min} = 2.7 Å and E_{min} = 1.2) produces solutions at a rate that is 5 times higher than what I used in my standard attempt (d_{min} = 3.0 Å and E_{min} = 1.5). The latter would however produce the first solution in less than five minutes, anyway.

Note that reducing E_{min} inevitably leads to worse values for CC_{weak} , as the E values in in this set become weaker and weaker.

3.5 Validation of the substructure

The crossword-table (Figure 1) can be interpreted in terms of 2×6 Se-sites related by non-crystallographic symmetry (the interpretation is shown in Figure 3.5, have a go yourself before looking at the solution !). The remaining two sites (B7 and A7) seem to be real as indicated by acceptable PMSF-values with the other 12 sites, but they do not follow the non-crystallographic symmetry. These sites may be located in a flexible N- (start-Met) or C-terminus which could be present in two different conformations in the crystal.

```
Minimum distances (top row, 0 if special position) and PSMF (bottom row)

Peak  self  cross-vectors

99.9  26.2
       9.6

96.8  36.4  19.8
      15.1  23.5

94.2  14.2  31.2 20.3
      10.4   9.4  7.5

93.0  34.9  15.6 23.2 19.9
      11.3   4.7 15.4  8.6

88.2   8.1  28.3 25.0 10.2 22.3
       1.8  12.5  5.1  0.0 16.0

87.1  21.0  25.2 20.2 23.6 11.0 29.6
      10.5  10.3 10.1  2.6  8.3  6.0

81.5  31.1  27.0 19.6 18.4 28.9 19.4 30.1
       2.5   9.0  7.1 11.3  9.9  0.0 10.7

80.1  17.7  10.7 23.3 39.8 25.1 35.7 33.7 32.2
       5.5  11.6 10.0  3.0  5.1  5.9  0.2 12.1

76.4  32.9  21.5 10.7 22.4 25.0 29.7 20.6 20.0 28.9
       1.8   2.6  5.7  3.4 11.5  9.3  3.7  3.3  5.3

76.4  29.9  10.8 19.9 34.8 25.5 32.2 35.7 22.7 10.5 25.0
       0.0   1.4  0.0  0.0  5.8  0.0  0.0  0.0  3.5  0.0

75.0  19.6  25.1 29.3 32.5 37.6 35.7 34.7 29.4 19.1 23.5 14.8
       0.0   5.7  4.3  0.0  0.0  0.0  7.6  0.0  0.0  3.6  0.0

72.2  37.2  24.2 28.5 20.2 10.6 24.6 10.1 22.1 30.3 24.7 33.4 30.1
       0.0   3.3 10.4  2.0  0.0  4.7  0.7  4.0  0.0  8.7  0.0  0.0

70.8  19.2  23.4 13.8 28.6 17.3 36.2 13.1 29.4 29.4 10.0 29.1 28.1 18.2
       0.0   0.0  8.5  1.6  2.3  4.8  0.9  0.1  0.0  0.0  0.0  2.4  0.0

69.3  31.5  25.5 36.2 29.4 24.8 35.1 18.5 23.4 19.6 31.7 25.2 27.5 14.6 24.6
       5.9   0.0  3.8  1.8  6.1  2.3  0.0  8.4  0.0  3.0  0.0  0.0  4.6  0.0

-------------------------------------------------------------------------------

59.8  40.4  18.8  4.3 20.0 27.0 23.2 24.4 16.7 21.7 13.1 16.9 25.6 32.7 17.7 36.8
       0.0   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0

58.6  20.9  26.5 16.6 27.6 26.0 27.3 27.5 29.6 23.6 27.3 23.3 24.2 35.0 27.2 23.6 15.8
       0.0   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
```

Figure 1: Crossword table for the try with the highest CC from run `sse_fa1a`. The columns containing fractional coordinates have been excluded for clarity.

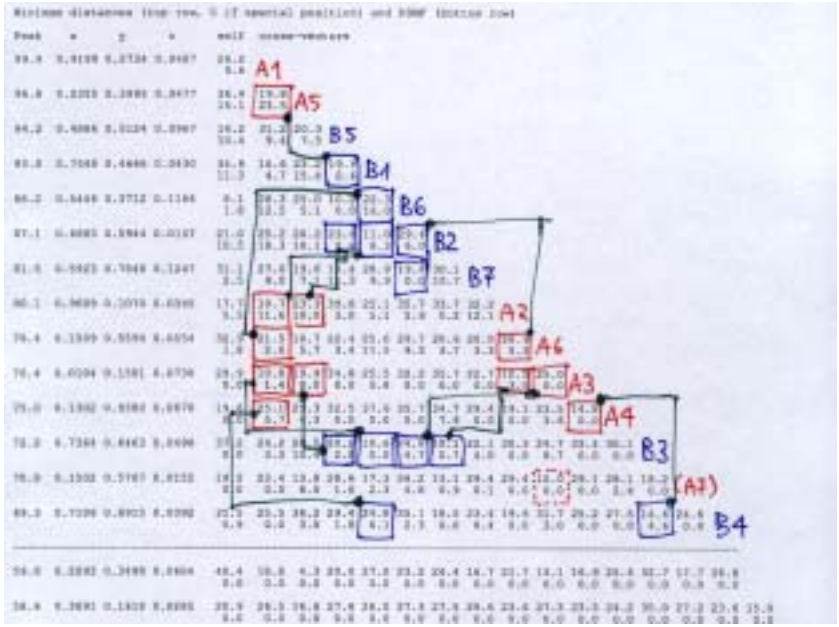

Figure 2: Crossword table for the try with the highest CC from run `sse_fa1a` with the hand-interpretation included.

4 Example JIA

For this tutorial, you need to have two programs: SHELXD and XPREP and three files containing the data collected at four different wavelengths: `jia_hrem.sca`, `jia_infl.sca`, `jia_lrem.sca`, and `jia_peak.sca`.

If you plan to extend the phases to the full resolution of 1.9 Å , you should also have the native data, `jia.hkl`. It may also be useful to have the pdb-entry 1C8U at hand to do some comparisons. For calculation phases you can use SHELXE , for displaying a map, XFIT (`http://www.sdsc.edu/CCMS/Packages/XTALVIEW/xtalview.html` for academics, `http://www.syrrx.com` for industrials) is a good program to use.

4.1 Information available

The JIA-dataset is a four wavelengths MAD-dataset used to solve the structure of *Escherichia coli* thioesterase II [1]. The data were kindly provided by Zbyzsek Dauter.

The final model (pdb-code 1C8U, S-Met) contained two molecules with 287 residues each. From the sequence there should have been 5 Met-residues per monomer.

The crystal used for the structure solution was of space group $C222_1$ with unit cell constants of: a = 96.000 Å, b = 120.000 Å, c = 166.130 Å, $\alpha = \beta = \gamma = 90.0^\circ$.

Date were stored as SCALEPACK .sca-files with symmetry related reflections merged

code	λ [Å]	d_{min} [Å]	filename
hrm	0.9747	2.5	`jia_hrem.sca`
pk	0.9787	2.5	`jia_peak.sca`
ip	0.9793	2.5	`jia_infl.sca`
lrm	0.9801	2.5	`jia_lrem.sca`

4.2 Data Quality and Substructure Structure Factors

We skip most of the XPREP-session (see the previous section if you are interested in details) and only look at the statistics for the anomalous data:

The signal/noise ratios for the anomalous differences are:

```
   Inf - 8.0 - 6.0 - 5.0 - 4.1 - 3.9 - 3.7 - 3.5 - 3.3 - 3.1 - 2.9 - 2.7 - 2.5 A
hrm   4.10  4.07  3.39  2.36  2.19  2.11  1.92  1.88  1.75  1.64  1.54  1.41
pk    5.31  4.86  3.88  2.84  2.64  2.50  2.29  2.25  2.08  1.91  1.75  1.57
ip    2.77  2.80  2.35  1.73  1.71  1.66  1.53  1.55  1.52  1.46  1.44  1.34
lrm   1.26  1.29  1.23  1.07  1.13  1.08  1.09  1.12  1.14  1.14  1.14  1.15
```

and the anomalous correlation coefficients:

```
    Inf - 8.0 - 6.0 - 5.0 - 4.1 - 3.9 - 3.7 - 3.5 - 3.3 - 3.1 - 2.9 - 2.7 - 2.5 A

pk-hrm  95.8  94.5  91.7  87.4  83.8  79.1  77.1  71.1  64.2  54.8  45.2  33.7

ip-hrm  91.2  88.3  84.4  76.2  70.4  67.4  61.2  56.2  44.9  36.4  33.9  19.0
```



```
ip-pk   92.9  89.3  87.6  81.1  75.8  69.0  67.9  60.9  51.6  43.4  36.9  28.5

lrm-hrm 62.1  53.0  44.1  38.4  25.6  31.6  22.4  19.7  16.0  11.4   8.3   6.5
lrm-pk  62.1  51.1  45.4  40.5  24.7  30.4  21.7  22.3  18.4  18.9  10.8  10.4
lrm-ip  59.7  49.9  47.1  38.1  25.7  25.5  21.1  22.5  17.8  11.5   4.0   5.9
```

Again, a very good data set. Based on the 30%-criterion, we could, in fact use almost all the data to 2.5 Åresolution. But, for convenience, we will only use the data to 3.0 Å.

4.3 Substructure Solution

As one of the Met-residues is N-terminal we will ask SHELXD for only 2 × 4 = 8 Se-sites. The corresponding ins-file looks like:

```
TITL jia_fa1 in C222(1)
CELL 0.98000  96.0000 120.0000 166.1300  90.000  90.000  90.000
ZERR   16.00   0.0192   0.0240   0.0332   0.000   0.000   0.000

LATT -7
SYMM -X, -Y, 0.5+Z
SYMM -X, Y, 0.5-Z
SYMM X, -Y, -Z

SFAC SE
UNIT   128
PATS
FIND   8
SHEL 999 3.0
MIND -3.5

HKLF 3
END
```

Please note, that the `SHEL 999 3.0` card had to be added manually to limit the resolution of the data.
Running this ins-file, we very quickly get a solution with very convincing figures of merit:

```
PSUM 21.61  PSMF Peaks: 76 72 64 61 60 59 48 44 35 31 31 30 29 28 28 27
Try 2:20  Peaks 99 93 93 90 84 82 76 68 22 18 17 17
R = 0.383, Min.fun. = 0.475, <cos> = 0.411, Ra = 0.381
Try     2, CC All/Weak 42.23 / 30.10, best 42.23 / 30.10, best PATFOM  1.05

PATFOM  18.71
```

There is a sharp drop between the heights of peaks number 8 and 9. Also, the cross-word table clearly indicates eight correct sites, for which the NCS-equivalent pairs can be easily determined (see Figure 3).

4.4 Calculating Phases and an Electron Density Map

The SHELXD/E-tutorial on the SHELX-website (`http://shelx.uni-ac.gwdg.de/SHELX` tells you how to use SHELXE to calculate phases that can be used to produce an electron density map.

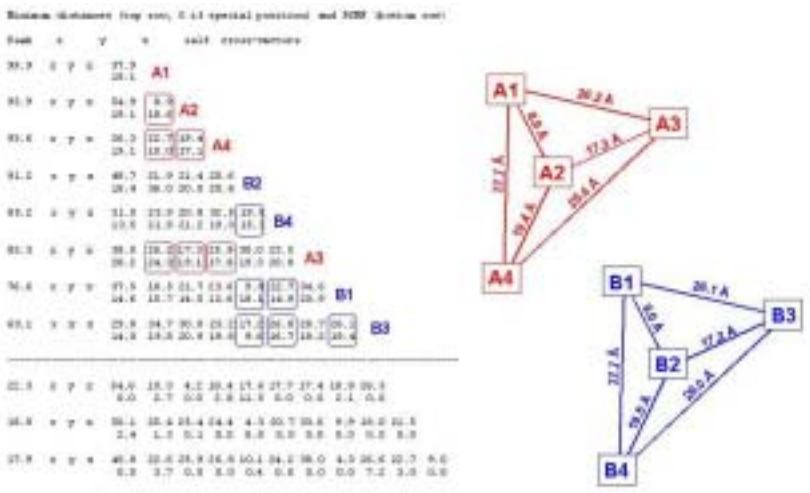

Figure 3: Crossword table and assignment of sites to NCS-related clusters for JIA.

5 Acknowledgements

Thanks go to Zbyszek Dauter (JIA) and Gordon Leonhard (SS) for providing the test data. The making of this tutorial was in part supported by an EU-funded project (AUTOSTRUCT, QLRI-CT-2000-00398).

References

[1] J. Li, U. Derewenda, Z. Dauter, S. Smith, and Z.S. Derewenda. Crystal structure of the escherichia coli thioesterase ii, a homolog of the human nef binding enzyme. *Nat.Stru.Biol.*, 7:555–559, 2000.