DNA Meeting Minutes
7th May 2003

Agenda: (Times guessed)
11.00 Demo of DNA
11.15 Recent Advances in DNA
11.30 BEST
12.00 Links with other projects
(lunch)
13.30 Future directions 1: Real time data processing
15.00 Future directions 2: Sample ranking, changers and databases.
16.00 AOB and close.

Demonstration of DNA and Resulting Discussion.

Graeme gave a short offline demonstration of how the DNA interface is used, and what it can do. Initially the indexing and strategy determination slides were demonstrated, using previously collected data. Although the demonstration went smoothly there were some concerns about the speed of the operation, because the strategy determination stage took nearly a minute. On a high intensity beamline the concerns about the radiation damage mean that this could be too long.

In the discussion which followed, the possible criteria for success and actions in the event of failure were discussed. The criteria for success will be detailed later on, but are essentially:

If indexing fails on an image, this indicates a problem.
The value of the RMS error.
The fraction of reflections indexed.
The fraction of reflections rejected.

A default exposure time should be used. If too few reflections are observed then this should be doubled or quadrupled, and the crystal re-exposed. If the number of spots found is not substantially improved, then the crystal should be rejected as there will be little to be gained by increasing the exposure again.

In principle it would be easy to identify ice rings in the images, because there would be a large number of spots at the same radius. Something along these lines should be implemented.

From the developers, it was clear that the implementation of the CVS service by Alun Ashton was valuable. This has enabled the users to maintain coherency much more easily than before. Since users are generally not keen on reporting errors to the developers, it was thought that a system of automatically sending an email in the event of failure could be implemented.

BEST.

Sasha Popov presented the BEST program, which is used to compute an optimum data collection strategy and estimate the resulting statistics from such a strategy. The

actual presentation can be found on the DNA web site, so I will only try to provide a summary here.

The program requires information about the source, the detector, the crystal and the amount of time that you have available to collect data. These parameters are then used to determine an optimised data collection strategy for the time constraint, and give an estimate of the quality of the data resulting from this experiment. Currently the system makes use of Denzo for the data processing, to give reflection statistics, but details were given about how to make the necessary changes to Mosflm to include this functionality, and assistance in the form of subroutines was also offered.

The process used in BEST is based on the following assumptions:

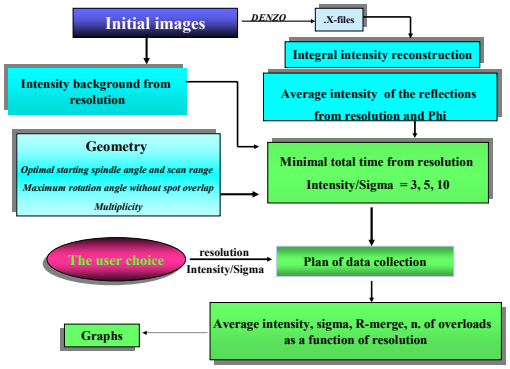
- All errors result from counting statistics.
- The instrument errors can be nonlinear.
- The probability density distribution for the intensities follows a Wilson distribution.
- All protein crystals have the same pattern of scattering intensity vs. resolution.
- The background doesn't change during rotation.
- The exposed volume of the crystal doesn't depend on the crystal/rotation???

The background intensity is obtained by finding and removing the spots from the diffraction image, then analysing the resulting intensity. The radial background is one of the requirements for the program, so that the effects of changing the exposure can be estimated.

As regards all crystals diffracting in the same way, the diffraction from a "large" number of crystals was inspected, and the mean was found to fit a standard distribution. The reflections from a single image therefore sample this distribution, so better information may be obtained. They therefore use experimental data gathered from a number of 1º images rather than a Wilson distribution.

Many large calculations were described, and I feel that there is little benefit in repeating them here.

The results of the program include a predicted Rmerge value, which would be anticipated from collecting and processing the suggested data set. There was a flow diagram for how BEST works, something like this (Taken from Sasha's Powerpoint slide)



This works as a command line program, with all of the required information being given on the command line. The program is then steered through standard input …. It appears to take about 30 seconds to run.

The suggested strategy may advise varying phi widths etc, to keep the statistics constant across the dataset. Alternatively, the statistics will vary if you have a constant phi width for the images.

A strategy was offered for the integration of BEST with DNA, using an extended Mosflm to perform the data processing tasks. There was some discussion as to which components could be developed to be "stand alone", and so work equally well with Mosflm and XDS, to prevent duplication of effort. At this time, no conclusion was reached and it was suggested that Sasha may attend a dna-dev meeting to take place in June.

Colin's discussion on links with other projects.

Colin thought that it would be sensible to have DNA as a focus for some of the forthcoming automation projects, as there has already been considerable progress, in particular using DNA to organise the interaction between

SPINE
BioXHit
Autostruct
Hamburg Developments, eg. BEST

There might be some sensitivity as some of these projects have deliverables which are being pursued by DNA or BEST. There should be no objection if they use the developments in DNA or BEST to achieve their aims, so long as the work is correctly acknowledged. There are other groups also doing similar work in Europe, and we should make an effort to find out who they are and what they are doing, for instance the Frencg CRG and at the ESRF. Pierre could be a link here.

Pierre on the inclusion of XDS.

Pierre is working on a scheduler add-on for XDS, based on the Mosflm.py wrapper. However, the operation of XDS is different to the way in which Mosflm works, in that the spots are indexed in P1, with the appropriate spacegroup being determined at

the integration stage. Once the reflections are integrated, they are reindexed into the appropriate space group. The question was asked as to how much information was needed, ie how many images. One image can be sufficient in some cases, but usually more are needed.

Andrew on Real Time Data Processing.

We are at phase 2 of the DNA plan, so it's time to consider a move into phase 3. We want to be able to measure the quality of data as we are collecting, and possibly use this information in the control of the data collection, for instance measuring the temperature factors. However, it is hard to predict when the radiation damage will affect results, and this is a point of some discussion. Perhaps a total user specified radiation dose on the crystal would be a good first step? A default value could be determined by the station manager.

The relationship between chemical damage to the crystal and the change in B factor, Rmerge etc can be argued, but the actual results are inconclusive. There is also the problem that the actual flux of the beamline is generally poorly calibrated, so we may have to be able to measure the dose in terms of "esrf-id14.4 exposure seconds", for instance.

Another unknown is the resolution limit to collect to, since this is designed to be a "fully automatic" system, some clever mechanism might be needed here.

As regards scaling, Andrew has a "command file" which will perform the scaling in background. It was agreed that this may make a good starting point to integrating the scaling into the DNA system, and it should be provided. One problem with the scaling is that it can be slow – taking a time comparable or greater than the integration, depending on the number of reflections. Given that we are aiming to return near real time scaling information, this speed issue could be a problem. Some discussion ensued regarding the parallelisation of Scala, but no firm conclusions were drawn. Given that the Scala is among the parallelisation targets of the e-HTPX project, we may benefit from this outside contribution.

Apparently someone at CHESS has looked at the parallelisation of scala, but no details came out.

Some discussion was made as to what should "run" Scala. Given that the program can be written with a small number of parameters, Graeme suggested that it wouldn't be too hard to integrate this with the "scheduler" mechanism, which currently drives Mosflm. All that would be needed to make progress on this would be a simple example of a Scala command file, which should give all of the required information.

One problem with this automation of the processing and scaling of the images is that at some point there will need to be a system which will be watching for "rogue" images and reflections. However, Colin suggested that we should fix this when it actually fails, so that we may learn from experience what may go wrong, rather than trying to predict a priori all of the possible failure situations.

More detailed discussion followed about how the scaling should be automated:

- Change in R factor should be 1-2%, but we have a few really bad images. We should then remove those images.
- We should scale in batches, so that the results can be returned while data collection proceeds.
- We should also watch for the parameters changing through the data set, for instance changes in the beam intensity.
- The Rbatch parameter should be "watched" by the Expert System.
- Should we process all data sets as if they contain anomalous scatterers?
- How do we know when to launch the integration process?
- Do we want to return the intensity information on a per-image basis?
- We should watch for changes in the cell dimensions as this could be an indicator of radiation damage.

Regarding the resolution limit, it was decided that a conservative default limit should be suggested, but made possible to override at the users suggestion. A default resolution limit could also be determined from the user defined dose limit using BEST. In principle, it should be possible to collect a reasonable data set using only the default values.

As regards the immediate work plan, we should aim to start real-time integration before worrying about using BEST for the strategy determination, since this can be included at a later stage at minimal cost. Most of the improvements from including BEST will be refinements, for example minimising the dose, and optimising the data collection statistics.

In order to trap errors in Scala, we will have to have enough information to decide on the spacegroup. The amount of images required for this will be determined by the lattice, and the scaling should not be started until at least this many images have been processed.

Initial data collection:

There was some discussion about how we should proceed with the initial data collection, since there is a requirement for images widely separated in phi for the cell refinement prior to integration. Andrew suggested the following strategy:

Data collection strategy

1. Collect images at phi = 0, 90 degrees with a standard detector distance and exposure time, distance and time selected from project requirements.
2. Autoindex both images, separately and together.
3. If the indexing is successful, integrate one or both images.
4. Use the integration data and indexing results to compute a data collection strategy, in terms of optimum rotation ranges, oscillation angles and exposure times.
5. Collect (say) 3 degrees around phistart and phistart + 90 degrees, to use for cell refinement.
6. Use this data to refine the cell parameters and orientation.
7. Collect data according to the strategy from (4), and integrate while the

data are collected. This may require distributed processing over several computers.
8. Scale and merge the data as the reflection files become available.

Scheme designed to minimise radiation damage. Required exposure time not known until after step 4. Delay between end of step 1 and step 5 less than 1 minute.

The criteria for a successful autoindexing are:
1. R.M.S. error in spot positions < index_spot_rms_error, typical value 0.15 mm.
2. Fraction of spots indexed > index_spot_frac_indexed, with a typical value of 95%.
3. Fraction of spots rejected < index_spot_frac_rejected, with a typical value of 5%.

All of these criteria should be satisfied for the indexing to be considered successful.

So it is worth collecting specific data for the preprocessing, since this can be used anyhow. All we have to ensure is that the names are unique, so that the data are not lost. We should aim to have the automatic integration included by the developers meeting on the 10th of June.

Other directions, Sample changers and databases.

Databases – we should send the URL for Joel Fillon's data model to the dna-dev mailing list.

One question to come up in these discussions was how to prioritise experiments. This may be an important thing in the fully-automated future, but is it a DNA problem? This should be a synchrotron-source policy thing. As an example, there is a small molecule structure service operating in Southampton – we should find out how they work.

The database interaction work will fall in part to the e-HTPX work, and perhaps this will be best accomplished there, allowing for use of for instance the ESRF database too. At the SRS there is currently now database, so this is something which will have to be addressed shortly.

There was some discussion about what should be stored. Colin agreed to circulate Sean McSweeny's thoughts on the matter. Another question related to the passing of data and the "harvest" files which are already implemented in CCP4 – is there an overlap? Also, is this needed for deposition etc, since the quality of the data will depend on instrumental parameters. The question as to what we want to store in the database will be an interesting and involved one.

Sample ranking:

How are the samples scored, and what should the figures of merit be? The strength of diffraction, in terms of $I/\sigma(I)$ at a given resolution or the resolution for a given $I/\sigma(I)$,

could be used although this would require some user input to decide on the required resolution.

How do we handle the following situations:

- Ranking individual crystals
- > 1 crystal per drop
- needing > 1 crystal for a dataset (matching up the sizes etc?)

The quality of the crystal should include information about the diffraction, the target cell and ice rings. Should we accumulate knowledge from previous experiments on the same project? This would make good use of a database.

Do we need to do fluorescence scans etc, searching for heavy atoms?…. Wavelength selection from fluorescence scans is a requirement for automation of anomalous scattering experiments, and has been implemented on some beamlines at the ESRF (???) and is an aim of the e-HTPX project.

The initial data collection for anomalous data should also be different, perhaps collecting an image at 180 degrees as well as 0, 90, to look for differences in the strength of diffraction. If we wish to treat the anomalous scatterers differently, then we should have a button on the DNA gui to suggest this.

ACTION: Miroslav should have a look at this. O(1month).

Crystal ranking.

Sample changers will be appearing all over Europe during the next couple of months. There's one at the SRS now, on 7.2, which holds 48 samples at once. This will probably be running properly at the end of the summer, when it will be transferred to 14.2. Similar timescales at the ESRF and Hamburg.

The control of the sample changing robot from the ES should be thought about, since this will probably be site dependant. -> a RCM (robot control module?)

There was a small discussion on the SPINE barcode proposal. This will probably come to pass, especially given the work in e-HTPX. There will be some work from other directions on this too….

Actions:

Andrew: the scala recipe.
Graeme: Implement this.
Olof: implement the data quality assessment
(Group of few) Implement use of BEST
Colin: circulate data model web address and Sean's thoughts about the storage of beamline information.
Colin: Obtain details of parallel processing developments for Scala being carried out at CHESS.
Olof, Graeme and Steve: Piccies + predictions on the dna GUI:
Six inches ^2, crosshair as prediction, nothing too fancy.

Colour code as Mosflm.

Graeme: add predictions etc to JPEGS.

Olof, Steve: If anomalous, different initial image collection.
Darren, Olof: BCM->ES “images are collected” XML

Aim to implement automatic integration by next developers meeting.

Date of next meeting? Will wait and see, but probably towards the end of the year.
It’ll mostly depend on what the developers want.