# DNA: An e-Science Perspective

**Graeme Winter**, and the DNA team[1].
CCLRC Daresbury Laboratory, Keckwick Lane, Warrington, UK
MRC Lab for Molecular Biology, Hills Road, Cambridge, UK
European Synchrotron Radiation Facility, Grenoble, France

## Abstract

We present the DNA project, a collaboration between the ESRF, MRC LMB and CCLRC Daresbury Laboratory with the aim of fully automating the collection of X-Ray crystallography data. This project forms part of the e-HTPX high throughput protein crystallography project, and makes use of recent e-Science developments to allow remote collection and processing of X-Ray diffraction data, in particular allowing remote access to the experimental results, so that the user may guide the experiment from their home lab, rather than travelling to the synchrotron site.

[1] The DNA team are: Olof Svensson, Sean McSweeney and Darren Spruce at the ESRF, Harry Powell and Andrew Leslie at the MRC LMB and Graeme Winter, Steve Kinder, Karen Ackroyd and Colin Nave at Daresbury Laboratory.

## 1. Introduction

The DNA project [1] is a collaboration initially between the ESRF, the SRS and MRC LMB with the objective of automating the collection of X-Ray diffraction data from macromolecular crystals.
Macromolecular crystallography is a rapidly expanding area of science where biology and physics meet head on. In particular, biologists need to use the physics of X-Ray diffraction to solve protein structures. Much of the computational process involved in this structure solution has been addressed through software packages like Solve [2] and SHARP [3]. However, at the data collection stage the scientist is often left to fend for themselves, unless they have on hand an expert in the field. A significant part of the DNA project is to provide an expert system to assist the scientist in the data collection. In addition, this allows remote access to the experiment through a simplified interface.

## 2. DNA in e-HTPX

The e-HTPX[2] project is a BBSRC pilot project for automating the entire protein structure solution process, from bioinformatics and target identification through to the deposition of the structure in public databases. A core part of this is the automatic collection of X-Ray diffraction data, which makes use of the work done as part of DNA.
For e-HTPX, the objective is to allow completely userless data collection, operating in a similar way to the EPSRC National Crystallography Service. Remote access will be provided via a grid-portal interface, where the user may specify the experimental requirements, the objectives and the amount of processing required.

## 3. Automation Efforts in Macromolecular Crystallography

The recent advances in structural genomics have pushed the development of high throughput macromolecular crystallography. In particular, with programmes such as SPINE [4] and SPORT [5] aiming to solve a substantial number of protein structures in a short time, the use of automated systems has become critical. Protein manufacture and crystallisation technologies being developed at the OPPF will produce high volumes of protein crystals. These will need to be passed to high throughput crystallography facilities, promoting the development of high performance beamline hardware and software.

## 4. e-Science in DNA

From an e-Science point of view, the key ideas in DNA are based around providing remote user or userless data collection. In this way it includes many areas of e-Science, including expert systems, high performance computing, remote access and the use of GRID technologies for data storage.

### 4.1. Expert Systems

At the core of DNA is an expert system capable of analysing X-Ray diffraction data and making sensible decisions about how best to collect data. This makes use of both existing software for data processing and software developed specifically for the task.
The expert system itself is coded in Python, since this provides the flexibility for incorporating legacy programs with the object orientated strengths in data management. The modular architecture (Figure 1) enables rapid development of functionality, and assists the distributed development across a number of sites.

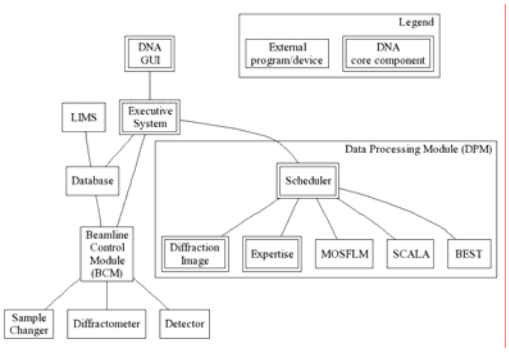


**Figure 1** A Schematic representation of the DNA architecture.

The full expert system is devolved throughout DNA. The beamline control system knows about the capabilities of the beamline, and the data processing module knows about data processing. The communication is mediated through an executive system, which simply knows how to ask for jobs to be done, but knows nothing of how they are done.
Although the DNA system includes a large number of components, a single user interface (Figure 2) is used to drive the system. This provides a simple interface to the experiment, given in terms that the scientist is familiar with, as well as sensible default values. With this system it is possible to collect a decent

[2] e-HTPX – HPC, Grid and Web-Portal Technologies in High Throughput Crystallography, this volume.

quality data set with a single button press, where this has traditionally required a substantial amount of decision making and transfer of information between different interfaces.

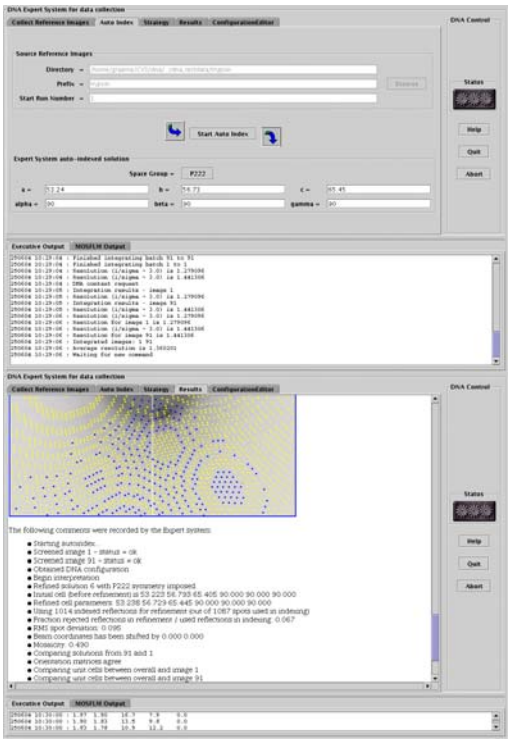

**Figure 2** The DNA user interface, after characterising a sample.

## 5. High Performance Computing

Although the processing involved with DNA does not require traditional supercomputer-style high performance computing, there is an emphasis on high throughput computing. The time taken for a decision to be made on the best approach to data collection is critical, even though it is only seconds. A substantial effort has been made to reduce this to the point where the rate limiting steps are the data collection, rather than any decision making or processing.

### 5.1. Remote Access

Although remote access is not the core concern of DNA, it is critical to the e-HTPX architecture. The plan is not to provide direct access to the experimental instruments, but instead to give the scientist the opportunity to give the experimental requirements in their terms prior to the experiment and translate these into specific requirements on the day.
Remote access will be provided through the e-HTPX grid portal, which will handle the safety authorisation and all authentication requirements through the use of grid certificates issued by the UK e-Science centre. The question of security is significant, since we plan to make the services available to industrial users, for whom the safety of the results are paramount.

### 5.2. Use of Grid Technologies

Once automatic data collection is in place, there should be no reason why data collection will not proceed on a 24 hour per day basis. With up to 4 GB of data per experiment, and one experiment taking half an hour, the total amount of data collected per beamline per day could potentially be collossal, and beyond the means of any one institution to archive reliably. We therefore plan to make use of central data storage facilities like ATLAS [6] for permanent archival of data, and only plan to keep maybe a days or a weeks worth on-line at any time. High performance file transfer, both to the central storage facilities and to the scientists data storage facilities will therefore be critical.
A second core use of grid technologies is the use of grid portal technology to provide access to the resources. As a part of e-HTPX, the objective is to provide a single grid portal from which to access protein manufacture, data collection, structure solution and deposition tools. Since these tools will be provided at a variety of locations, the use of third-party data transfers will be essential.

## 6. Future Plans

The DNA team will shortly be releasing version 1.0 of the software, which may be installed on all X-Ray sources from lab rotating anode generators to the highest brilliance beamlines in Europe. The future plans for DNA include extending this technology to include structure solution to provide real-time experimental feedback, and to optimise the use of experimental time. In addition, the use of Artificial Neural Networks as a tool for statistical analysis and data mining may be included in future releases, so that results from later in the structure determination pipeline may be used to train the system.

## 7. References

[1] Leslie, A. G. W., Powell, H. R., Winter, G., Svensson, O., Spruce, D., McSweeney, S., Love, D., Kinder, S., Duke, E. & Nave, C. (2002) Acta Cryst. D58, 1924-1928.
[2] Terwilliger, T.C. and J. Berendzen. (1999) Acta Cryst D55, 849-861.

[3] La Fortelle, E. de & Bricogne, G. (1997).
Methods Enzymol. 276, 472-494.
[4] http://www.spine-europe.org
[5] http://www.bbsrc.ac.uk/
science/initiatives/sport.html
[6] http://www.e-science.clrc.ac.uk/
web/services/datastore