

Automation in DNA/e-HTPX

Automation in Data Collection,
Processing &
Structure Solution

Part I



Areas of interest

- Automated data collection, processing & analysis
 - properly collecting data
 - properly processing data
 - making decisions about data sets
 - characterisation of what you have
- Anomalous diffraction phasing (SAD/MAD) – at the moment just to test the automated processing
- Molecular replacement (ball dropped on this one, in process of picking up again)

DNA – automateD collectioN of datA

- DNA (obviously) connects the beamline control systems with data processing software, via an “executive system”
- includes DNA “scheduler”, which does
 - pre-screening
 - indexing
 - strategy calculation and interpretation
 - integration

The logo for the DNA system, consisting of the lowercase letters 'dna' in a bold, sans-serif font. The letters are dark grey with a subtle, grainy texture.

What does it do?

- Allow characterisation of your crystal with a single “click”, giving:
 - unit cell & lattice
 - resolution estimate for crystal
 - strategy with exposure time
- A second click will collect and process data according to the suggested strategy, or you can change it and collect something different

Collect Images

DNA Expert System for data collection

Collect Reference Images

Auto Index

Strategy

Results

ConfigurationEditor

Setup for collecting two reference images 90 degrees apart by Expert System

Directory = /tmp/beamline

Prefix = hep_cat_low

Keep PUBLIC copy of reference images = ☒ Enabled

Start Run Number = 5

Exposure Time = 1.0

Oscillation = 1.0

Resolution = 3.5

Bravais Lattice =

Collect and Characterize

Executive Output

MOSFLM Output

040105 10:57:22 :
040105 10:57:22 :
040105 10:57:22 : Startup of the Executive System
040105 10:57:22 : \$Id: ExecutiveSystem.py,v 1.49 2004/11/12 14:32:04 svenasson Exp \$
040105 10:57:22 : DNAHOME = /home/dna/dna/dna-1.0.0
040105 10:57:22 : DNAHOME = gw56port
040105 10:57:22 : BCM host = gw56port, port = 1992
040105 10:57:22 : BCM communication OK.
040105 10:57:22 : Switching to online mode.
040105 10:57:22 : Waiting for new command

DNA Expert System for data collection

Collect Reference Images

Auto Index

Strategy

Results

ConfigurationEditor

Source Reference Images

Directory = /tmp/beamline

Prefix = hep_cat_low

Start Run Number = 5

Browse

Start Auto Index and Integrate

Expert System auto-indexed solution

Space Group = unknown

a =

b =

c =

alpha =

beta =

gamma =

Executive Output

MOSFLM Output

040105 10:57:22 :
040105 10:57:22 :
040105 10:57:22 : Startup of the Executive System
040105 10:57:22 : \$Id: ExecutiveSystem.py,v 1.49 2004/11/12 14:32:04 svenasson Exp \$
040105 10:57:22 : DNAHOME = /home/dna/dna/dna-1.0.0
040105 10:57:22 : DNAHOME = gw56port
040105 10:57:22 : BCM host = gw56port, port = 1992
040105 10:57:22 : BCM communication OK.
040105 10:57:22 : Switching to online mode.
040105 10:57:22 : Waiting for new command

DNA Control

Status

Help

Quit

DNA Control

Status

Help

Quit

Abort

Analysis, Autoindex & Integrate

DNA Expert System for data collection

Collect Reference Images Auto Index Strategy Results ConfigurationEditor

Source Reference Images

Directory =

Prefix =

Start Run Number =

Expert System auto-indexed solution

Space Group =

a = b = c =

alpha = beta = gamma =

DNA Control

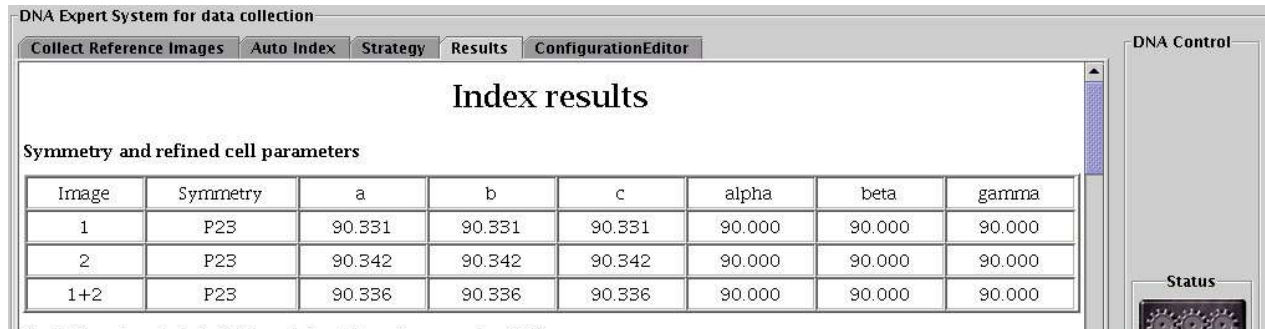
Status



Executive Output MOSFLM Output

```
040105 11:00:43 : Finished integrating batch 1 to 1
040105 11:00:43 : Integrating image batch 2 to 2
040105 11:00:43 : Copying log files to /tmp/beamline/hep_cat_low_5_dnafiles
040105 11:00:48 : Highest resolution recorded as 2.80
040105 11:00:48 : Finished integrating batch 2 to 2
040105 11:00:48 : Copying log files to /tmp/beamline/hep_cat_low_5_dnafiles
040105 11:00:48 : 0 out of 2 images had "too many" bad spots
040105 11:00:48 : That is 1 spots out of 2138 were bad
040105 11:00:50 : Integration results - image 1
040105 11:00:50 : Integration results - image 2
040105 11:00:50 : Resolution for image 1 is 2.80
040105 11:00:50 : Resolution for image 2 is 2.80
040105 11:00:50 : Integrated images: 1 2
040105 11:00:50 : Average resolution is 2.80
040105 11:00:50 : Calculated resolution: 2.80
040105 11:00:50 : Waiting for new command
```

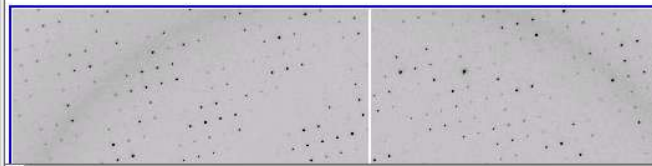
Results (1)



Spots found, rejected, RMS spot deviation, beamcentre shift

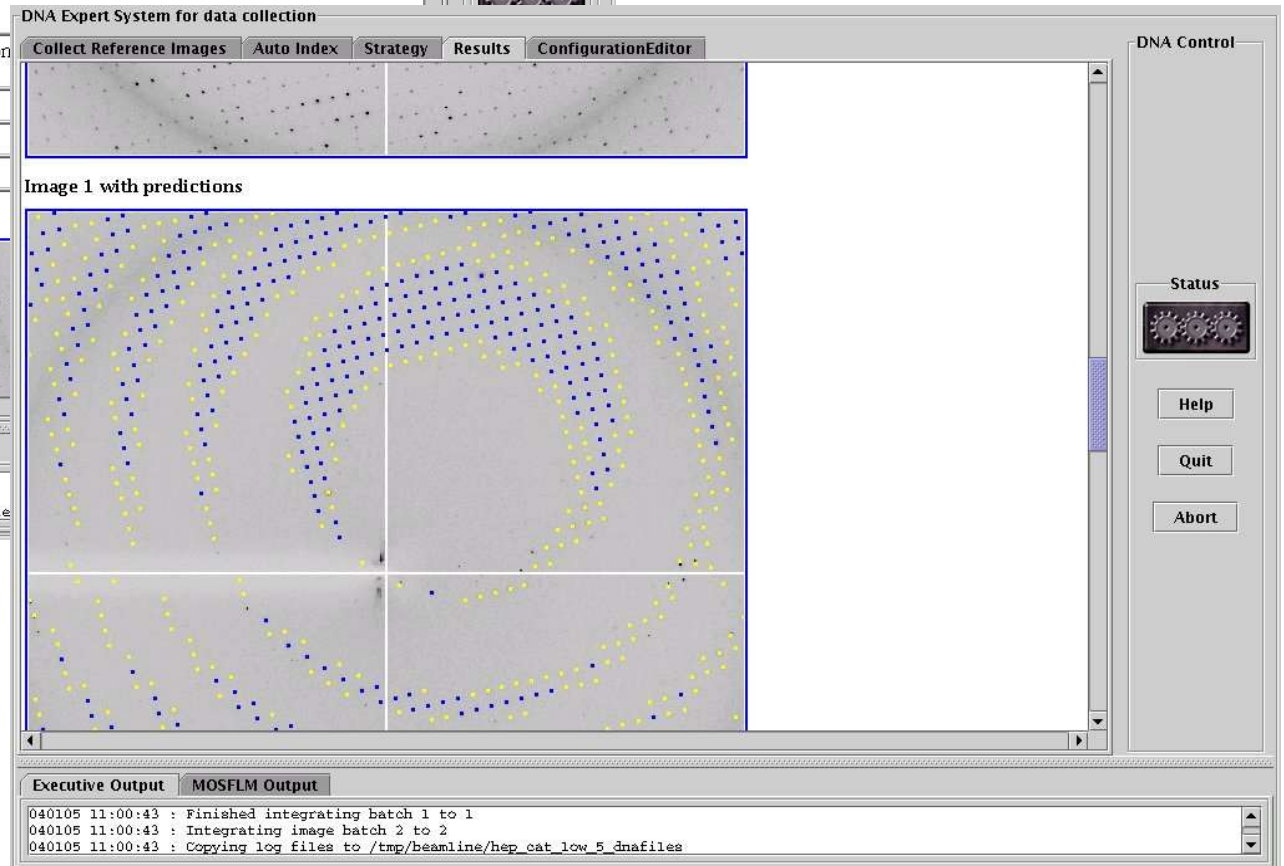
Image	Spots used in refinement	Spots used in indexing	Fraction rejected from refinement
1	325	355	0.085
2	285	308	0.075
1+2	614	663	0.074

Image 1: /tmp/beamline/ref-hep_cat_low_5_001.img



Executive Output MOSFLM Output

```
040105 11:00:43 : Finished integrating batch 1 to 1
040105 11:00:43 : Integrating image batch 2 to 2
040105 11:00:43 : Copying log files to /tmp/beamline/hep_cat_low_5_dnafile
```



Strategy

DNA Expert System for data collection

Collect Reference Images Auto Index Strategy Results ConfigurationEditor

DNA Reference Strategy

Phi Start	Increment	Time	Images	1st Imag...	Resolution
172.0	0.5	1.0	90	1	2.7987...

Data Collection Strategy

Phi Start	Increment	Time	Images	1st Imag...	Resolution
172.0	0.5	1.0	90	1	2.7987...

Calculate Strategy Collect Data Collect & Integrate Data

DNA Control

Status



Help

Quit

Abort

Executive Output **MOSFLM Output**

```

040105 11:03:41 : 5.27 5.12 74.7 2.3 0.0
040105 11:03:41 : 5.12 4.97 74.4 2.3 0.0
040105 11:03:41 : 4.97 4.84 75.4 2.3 0.0
040105 11:03:41 : 4.84 4.72 75.4 2.3 0.0
040105 11:03:41 : 4.72 4.61 76.4 2.3 0.0
040105 11:03:41 : 4.61 4.50 75.3 2.3 0.0
040105 11:03:41 : 4.50 4.40 76.4 2.3 0.0
040105 11:03:41 : 4.40 4.31 74.3 2.3 0.0
040105 11:03:41 : 4.31 4.23 74.4 2.3 0.0
040105 11:03:41 : 4.23 4.14 74.2 2.3 0.0
040105 11:03:41 : 4.14 4.07 72.0 2.3 0.0
040105 11:03:41 : 4.07 3.99 72.8 2.4 0.0
040105 11:03:41 : Overall
040105 11:03:41 : Min Max I/sig R Overload
040105 11:03:41 : 12.00 3.99 75.0 2.3 0.0
040105 11:03:42 : Waiting for new command
    
```

Collect & Integrate

DNA Expert System for data collection

Collect Reference Images Auto Index Strategy Results ConfigurationEditor

DNA Reference Strategy

Phi Start	Increment	Time	Images	1st Imag...	Resolution
172.0	0.5	1.0	90	1	2.7987...

Data Collection Strategy

Phi Start	Increment	Time	Images	1st Imag...	Resolution
172.0	2.0	1.0	30	1	2.7987...

Calculate Strategy Collect Data Collect & Integrate Data

DNA Control

Status



Help

Quit

Abort

Executive Output MOSFLM Output

```
040105 11:03:41 : 5.27 5.12 74.7 2.3 0.0
040105 11:03:41 : 5.12 4.97 74.4 2.3 0.0
040105 11:03:41 : 4.97 4.84 75.4 2.3 0.0
040105 11:03:41 : 4.84 4.72 75.4 2.3 0.0
040105 11:03:41 : 4.72 4.61 76.4 2.3 0.0
040105 11:03:41 : 4.61 4.50 75.3 2.3 0.0
040105 11:03:41 : 4.50 4.40 76.4 2.3 0.0
040105 11:03:41 : 4.40 4.31 74.3 2.3 0.0
040105 11:03:41 : 4.31 4.23 74.4 2.3 0.0
040105 11:03:41 : 4.23 4.14 74.2 2.3 0.0
040105 11:03:41 : 4.14 4.07 72.0 2.3 0.0
040105 11:03:41 : 4.07 3.99 72.8 2.4 0.0
040105 11:03:41 : Overall
040105 11:03:41 : Min Max I/sig R Overload
040105 11:03:41 : 12.00 3.99 75.0 2.3 0.0
040105 11:03:42 : Waiting for new command
```

Results (2)

DNA Expert System for data collection

Collect Reference Images Auto Index Strategy **Results** ConfigurationEditor

DNA Reference Strategy

Phi Start	Increment	Time	Images	1st Imag...	Resolution
172.0	0.5	1.0	90	1	2.7987...

Data Collection Strategy

Phi Start	Increment	Time	Images	1st Imag...	Resolution
172.0	2.0	1.0	30	1	2.7987...

Calculate Strategy Collect Data Collect & Integrate Data

DNA Control

Status

Help

Quit

Abort

Executive Output **MOSFLM Output**

```
040105 11:07:59 : 21 24.53 7.28 0.34
040105 11:07:59 : 22 25.87 11.61 0.08
040105 11:07:59 : 23 25.14 29.59 0.26
040105 11:07:59 : 24 24.98 9.33 0.50
040105 11:07:59 : 25 24.88 7.42 0.16
040105 11:07:59 : 26 24.68 14.23 0.25
040105 11:07:59 : 27 23.87 18.53 0.08
040105 11:07:59 : 28 23.56 6.59 0.17
040105 11:07:59 : 29 23.19 14.08 0.16
040105 11:07:59 : 30 21.81 15.32 0.24
040105 11:07:59 : Highest resolution recorded as 2.80
040105 11:07:59 : Finished integrating batch 21 to 30
040105 11:07:59 : Copying mtz /tmp/dna-dna/integrate_21_30/hep_cat_low_5_2005-01-04_process_21_30.mtz to /tmp/beamline/hep_cat_low_5_dnafil
040105 11:08:00 : Copying log files to /tmp/beamline/hep_cat_low_5_dnafiles
040105 11:08:20 : Waiting for new command
```

How does it do it?

- Using PXGen++/ProDC to collect the images
- Using DiffractionImage to “screen” the images
- Using Mosflm to autoindex
- Using Mosflm & BEST to compute data collection strategies
- Using Mosflm to integrate the data & refine the unit cell and so on

Where is it?

- Release 1.0.(nearly)1 installed on beamlines at SRS & ESRF
- Release 1.0.0 installed on MAR345 here in York
- Soon to be installed at NSLS for RapiData 2005

Aside: A Definition of Automation

- “From frames to an unrefined structure, with little or no user input, on a laptop” - ho ho ho ho, no really..
- Starting with frames minimizes chance of user error, and handles some of the more tedious things (like reading CCP4 study weekend proceedings!)
 - Decisions could be best made by actually looking at the data (suggestion, not statement!)

Automated Data Processing

- Same technology as DNA Scheduler
- Does more – includes scaling, space group (not just point group) estimation:

Message (1085): Considering correct spacegroup options!

Message (1085): Possible spacegroups: p43212 p41212

- Can evaluate the “quality” of data and figure out what to do with images
- Precursor to full “XIA” automation – images to structure without intervention

The logo for XIA, featuring the letters 'xia' in a stylized, lowercase font. The letters are dark grey with a subtle, grainy texture, giving them a three-dimensional or metallic appearance. The 'x' and 'i' are connected, and the 'a' is separate.

To be continued...

Automation in DNA/e-HTPX

Automation in Data Collection,
Processing &
Structure Solution

Part II



Images to Structure?

- Example:
 - Rummage (decide what images you have – group in to data sets)
 - Process data sets (properly)
 - Analyze (decide what's what – e.g. peak, inflection, high & low energy remote)
 - Phase & perform appropriate transformations (using SHELXC/D/E because it's easy!)
 - Build using Arp/wArp

xia

```
graeme@ehhttpx-pc3:/> xia.sh
```

```
Message (    0): #####
```

```
Message (    0): Running Processor.py for template s1_3_###.mccd
```

```
Message (    0): Selected images 1, 91 for autoindex
```

```
Message (    0): Starting autoindex...
```

```
...
```

```
Message (    0): #####
```

```
Message (    0): Running Processor.py for template s1_2_###.mccd
```

```
... etc ... (done 3 times) - can take more than a couple of minutes!
```

```
Message (    5): Analysis of ./s1_2_001.mccd.dir/scale/scaled.mtz:  
wave=0.979300 signal=0.051300
```

```
Message (    8): Analysis of ./s1_1_001.mccd.dir/scale/scaled.mtz:  
wave=0.979050 signal=0.082200
```

```
Message (   13): Analysis of ./s1_3_001.mccd.dir/scale/scaled.mtz:  
wave=0.885570 signal=0.075800
```

```
Message (   15): Performing MAD experiment
```

```
Message (   15): Peak: ./s1_1_001.mccd.dir/scale/scaled.mtz
```

```
Message (   15): Infl: ./s1_2_001.mccd.dir/scale/scaled.mtz
```

```
Message (   15): HRem: ./s1_3_001.mccd.dir/scale/scaled.mtz
```

```
Message (   16): Calculated nmol based on method by Kantartdjieff and  
Rupp (2002)
```

```
Message (   16): Best nmol/asu found: 3 (p = 0.869175)
```

The logo for 'xia' is displayed in a large, bold, sans-serif font. The letters are dark grey with a subtle, grainy texture, giving it a slightly metallic or digital appearance. The 'x' is particularly prominent, with its two strokes clearly defined.

Message (16): Running SHELXC
Message (21): Running SHELXD
Message (26): 1 => [99] [97] [91] [88] [88] [85] [78] [78] [77] [64] [55]
[51] <7> <4> <4> <3> <1>
Message (26): All heavy atoms located
Message (30): 2 => [99] [96] [91] [89] [88] [85] [78] [77] [77] [64] [55]
[52] <7> <5> <4> <4> <1>
Message (30): All heavy atoms located
Message (30): Running SHELXE
Message (55): cycle 1 => 0.138
Message (79): cycle 2 => 0.150

...

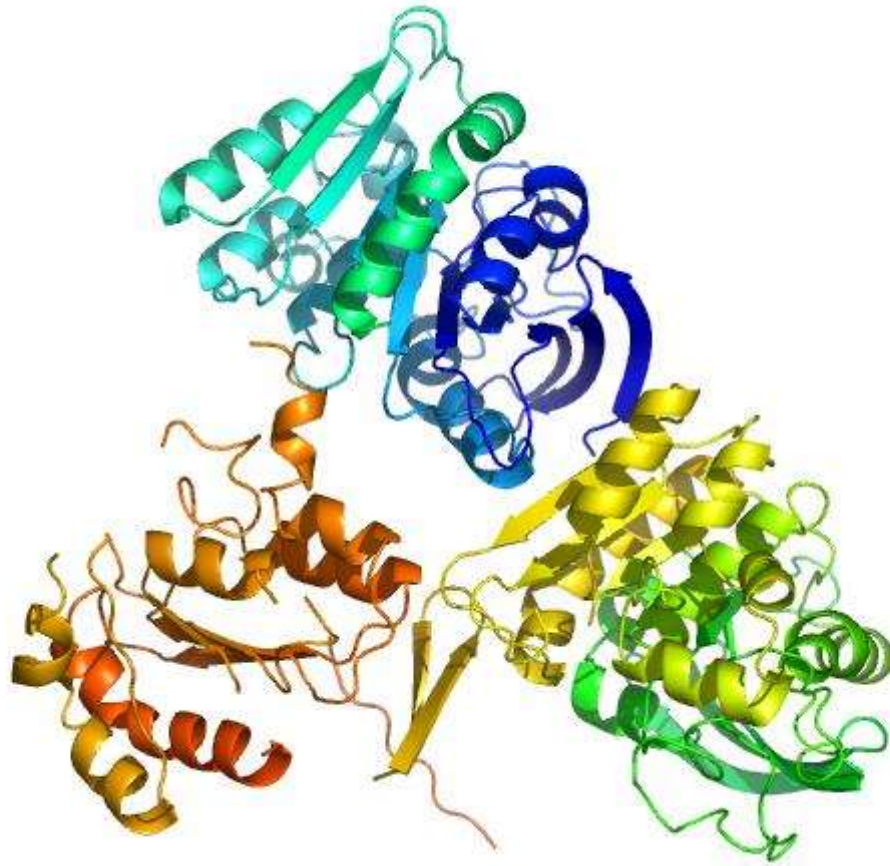
Message (413): cycle 16 => 0.344
Message (437): cycle 17 => 0.344
Message (437): Stopping density modification...
Message (461): cycle 18 => 0.345
Message (484): Running SHELXE
Message (509): cycle 1 => 0.180
Message (533): cycle 2 => 0.328

...

Message (794): cycle 13 => 0.626
Message (818): cycle 14 => 0.626
Message (818): Stopping density modification...
Message (842): cycle 15 => 0.625
Message (842): Stopping density modification...
Message (865): Running F2MTZ
Message (871): Running CAD

xia

After a little Arp/wArp...
(This is an easy example...)



xia

Thoughts...

- Bolting together data processing & structure solution can be very powerful
- Data management is an issue
- User interface is non-existent (command “xia.sh”
-> 1000 odd lines of logfile output)
- It is available (in XIA CVS) but is a little untested (testing is in progress though, and a “small scale” release may be in order)

Plans...

- Enhance data processing
 - Include XDS to give better processing of fine phi and 0-dose extrapolation
- Work on the UI – need a better “look for n zinc sites” than writing a “.nha” file
- Improve the HA location and phasing
- Improve data analysis & processing
- Incorporate automated MR – bmp
- Put out some kind of release..