

Automated Decision Making in DNA

BioXHIT/CCP4/e-HTPX
“Standards for Automation”
February 2005

Decision Making...

For instance, “given it's 10.34 pm do we have time to get to the pub?”

- Definition – I'd like a pint but it is getting late
- Information – pub closes in 26 minutes
- Judgment – since I'm ten minutes away I'm OK
- Execution – go get that drink!
- Verification – check your watch on the way

More sensibly...

Given that I need this data set to 1.5 angstroms, can I get decent data from this crystal in 30 minutes?

- Definition – you have a limited time to perform the experiment in
- Information – from indexing and strategy, it looks like data collection will take 26 minutes
- Judgment – time is adequate
- Execution – begin the data collection
- Verification – process the data as you go...

Decisions in DNA

- Quality of images (screening)
- Quality of indexing solution
- Determination of strategy
- Verification of quality of data
- Determination of point group, that is, checking the assumptions made from indexing are correct

Definition

Defining the problems in the scope of DNA is relatively easy:

- Is the crystal OK?
- What strategy do I need?
- Is this data set OK?

Information

Getting the information together is the hard bit:

- People are very good at recognising patterns in pictures, computers aren't
- Only input is the images recorded and the meta data from the beam line
- Given enough information most decisions are trivial, but getting the information can be hard

Judgement

- Again people are, in general, better at this than computer programs
- Given enough information the judgements can be OK – but can you gather enough information?
- The correct judgement is not cast in stone – ask ten crystallographers something and you will often get 11 answers

Execution

- Involves controlling the hardware
- Scope is relatively small -
 - Move detector to here
 - Record an exposure for this long, while moving the crystal thus
- May become more challenging as the abilities of beam lines become more sophisticated

Verification

Verification steps in DNA are exactly as described as “good practice” at the study weekend

- Process the data as you go
- Inspect the images and the results
- Try and solve the structure while the crystal is still available
- Look for signs of radiation damage

In DNA...

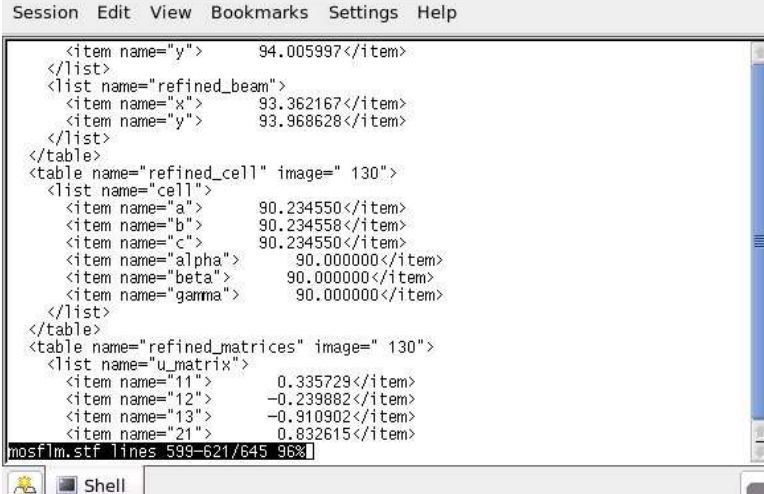
Information...

- Have had to develop a set of “eyes” to inspect the images with - DiffractionImage
- Have to have a reasonable idea of expectations on which to base the judgements – both “defaults” and dynamic expectations
- Have to get as much information as we can in a limited time – deciding on a “perfect” data collection strategy is much less optimal than coming up with a “good” one in one tenth the time

In DNA...

Information...

- Have to get data out of programs – for this modifications have been added to Mosflm, Scala, Truncate, Sortmtz to write out results as XML
- “Grep”ing works, but is not stable, and scales very badly for large amounts of output (ever read a mosflm.lp file?!)



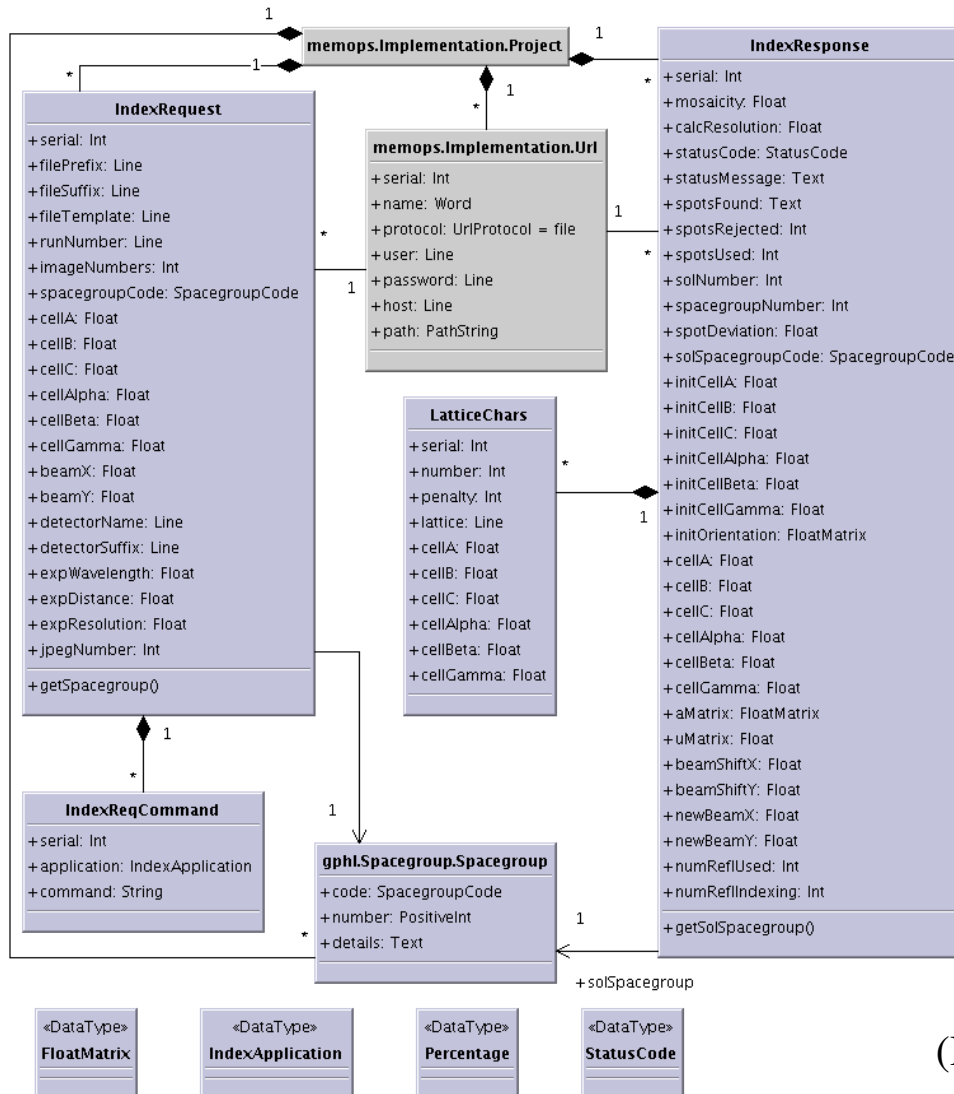
```
Session Edit View Bookmarks Settings Help
</item name="y"> 94.005997</item>
</list>
<list name="refined_beam">
  <item name="x"> 93.362167</item>
  <item name="y"> 93.968628</item>
</list>
</table>
<table name="refined_cell" image=" 130">
  <list name="cell">
    <item name="a"> 90.234550</item>
    <item name="b"> 90.234558</item>
    <item name="c"> 90.234550</item>
    <item name="alpha"> 90.000000</item>
    <item name="beta"> 90.000000</item>
    <item name="gamma"> 90.000000</item>
  </list>
</table>
<table name="refined_matrices" image=" 130">
  <list name="u_matrix">
    <item name="11"> 0.335729</item>
    <item name="12"> -0.239882</item>
    <item name="13"> -0.910902</item>
    <item name="21"> 0.832615</item>
  </list>
</table>
mosflm.stf lines 599-621/645 96%
Shell
```

In DNA...

Most of the problem is getting hold of enough information to make a decision, so

- Need a standardised way of describing the data
- Need a standardised way of handling the data
- Need a way to express the decision making process

Standard Data Formats...



- Standard DNA data model
- Allows interchange of data easily
- Probably the single most important part of DNA!

(Diagram supplied by Lorenzo Milazzo using CCPN software - Thanks!)



Comments

- Information gathering has dependencies – you need to autoindex before you can identify properly the strength of diffraction
- Speed is an issue, since people are waiting for the results
- Modularity is important
- Data management & communication are critical

Thanks to...

Everyone involved in
DNA!

Karen Ackroyd*, Alun Ashton, Gleb Bourenkov*, Gérard Bricogne, Sandor Brockhauser, Liz Duke, Eric Girard, Steve Kinder*, Ludovic Launer, Pierre Legrand, Andrew Leslie, Katherine McAuley, Sean McSweeney, Lorenzo Milazzo, Colin Nave, Venkataraman Parthasarathy, Alexander Popov*, Harry Powell*, Raimond Ravelli, Lucile Roussier, Darren Spruce*, Olof Svensson*, Andrew Thompson, Takashi Tomizaki, Graeme Winter*