Commission on Crystallographic Computing
International Union of Crystallography
http://www.iucr.org/resources/commissions/crystallographic-computing
Newsletter No. 9, October 2008
This issue:
"Presentation notes from the IUCr Computing School, Kyoto, Japan, 18th to 23rd August, 2008"

Table of Contents

(Notes collated by Lachlan Cranswick)

(Warning – unless you want to kill 143 pages worth of forest – DO NOT press the "print" button. For hardcopies – you may like to only print out the articles of personal interest.)

THE IUCR COMMISSION ON CRYSTALLOGRAPHIC COMPUTING - TRIENNIUM 2008-2011

Chairman: Lachlan M. D. Cranswick
Canadian Neutron Beam Center (CNBC),
National Research Council of Canada (NRC),
Building 459, Station 18,
Chalk River Laboratories,
Chalk River, Ontario,
Canada, K0J 1J0
Tel: (613) 584-8811 (Ext. 43719)
Fax: (613) 584-4040
E-mail: lachlan.cranswick@nrc.gc.ca
WWW: http://neutron.nrc-cnrc.gc.ca/

Dr Benedetta Carrozzini
Institute of Crystallography (IC)
National Research Council (CNR)
Via G. Amendola, 122/o
70126 Bari - Italy
Tel. +39 080 5929 147
Fax +39 080 5929 170
E-mail: benedetta.carrozzini@ic.cnr.it
WWW: http://www.ic.cnr.it/carrozzini.php

Dr Ralf W. Grosse-Kunstleve
Lawrence Berkeley National Laboratory
One Cyclotron Road, BLDG 64R0121,
Berkeley, California, 94720-8118,
USA.
Tel: (510) 486-5929
Fax: (510) 486-5909
E-mail: rwgk@cci.lbl.gov
WWW: http://cctbx.sourceforge.net/
WWW: http://www.phenix-online.org/
WWW: http://cci.lbl.gov/~rwgk/

Dr Harry Powell
MRC Laboratory of Molecular Biology,
Hills Road, Cambridge, CB2 2QH, UK.
Tel: +44 (0) 1223 248011
Fax: +44 (0) 1223 213556
E-mail: harry@mrc-lmb.cam.ac.uk
WWW: http://www.mrc-lmb.cam.ac.uk/harry/

Prof Jordi Rius
Dpt. of Crystallography
Institut de Ciencia de Materials de Barcelona, CSIC
Campus de la UAB
08193-Bellaterra, Catalunya, Spain
Tel: +93 580 18 53
Fax: +93 580 57 29
E-mail: jordi.rius@icmab.es

Dr K. Sekar
Bioinformatics Centre
(Centre of Excellance in Structural Biology and Bio-computing)
Supercomputer Education and Research Centre
Indian Insitute of Science
Bangalore 560 012
India
Tel: 91-(0)80-23601409, 22933059, 22932469
Fax: 91-(0)80-23600683, 23600551
E-mail: sekar@physics.iisc.ernet.in and sekar@serc.iisc.ernet.in
WWW: http://www.physics.iisc.ernet.in/~dichome/sekhome/

Prof Peter Turner
Crystal Structure Analysis Facility,
School of Chemistry (F11),
University of Sydney,
Sydney, NSW,
Australia 2006.
E-mail: turner_p@chem.usyd.edu.au and p.turner@chem.usyd.edu.au
WWW: http://csaf.chem.usyd.edu.au/home_nographics.htm

Prof Björn Winkler
Inst. f. Geowissenschaften / FE Mineralogie / Abt. Kristallographie
Universitaet Frankfurt
Altenhoeferallee 1
D-60438 Frankfurt am Main, Germany
Phone: +49 69 798 40107 / 40108 (secretary)
FAX : +49 69 798 40109
E-mail: b.winkler@kristall.uni-frankfurt.de

Associate Prof. Dr. Min Yao
Laboratory of X-ray structural biology,
Faculty of Advanced Life Science,
Hokkaido University,
060-0810, Sapporo, Japan
Tel: +81-(0)11-706-4481
Fax: +81-(0)11-706-4481
E-mail: yao@castor.sci.hokudai.ac.jp
WWW: http://altair.sci.hokudai.ac.jp/g6/index-e.html

Consultants

Professor I. David Brown
Brockhouse Institute for Materials Research,
McMaster University,
Hamilton, Ontario, Canada
Tel: 1-(905)-525-9140 ext 24710
Fax: 1-(905)-521-2773
E-mail: idbrown@mcmaster.ca
WWW: http://www.physics.mcmaster.ca/?page=sw://lists/Minibio_2004.php4?ID=4

Dr Graeme Day
Department of Chemistry,
University of Cambridge,
Lensfield Road, Cambridge, CB2 1EW,
United Kingdom
tel: +44 (0)1223-336390
fax: +44 (0)1223-336362
E-mail: gmd27@cam.ac.uk
WWW: http://www.ch.cam.ac.uk/staff/gmd.html

Prof Santiago García-Granda
Faculty of Chemistry,
University Oviedo,
C/ Julian Claveria, 8
33006 Oviedo - Asturias, Spain
Tel. +34-985104061
Fax +34-985103125
Mobile: +34-690029092
E-mail: sgg@uniovi.es
WWW: http://www.uniovi.es/xray/

Consultants (cont'd)

Prof Alessandro Gualtieri
Università di Modena e Reggio Emilia,
Dipartimento di Scienze della Terra,
Via S.Eufemia, 19,
41100 Modena, Italy
Tel: +39-059-2055810
Fax: +39-059-2055887
E-mail: alessandro.gualtieri@unimore.it
WWW: http://www.terra.unimo.it/en/personaledettaglio.php?user=alex

Prof Bart Hazes
Department of Medical Microbiology and Immunology
1-41 Medical Sciences Building
University of Alberta
Edmonton, AB T6G 2H7
Canada
E-mail: Bart.Hazes@Ualberta.ca
WWW: http://www.ualberta.ca/~mmi/faculty/bhazes/bhazes.html

Dr James Hester
Bragg Institute, ANSTO,
PMB 1, Menai, New South Wales, 2234,
Australia
Tel: +61 2 9717 9907
Fax: +61 2 9717 3145
E-mail: jxh@ansto.gov.au and jamesrhester@gmail.com
WWW:
http://www.ansto.gov.au/research/bragg_institute/contacts/dr_james_hester.html

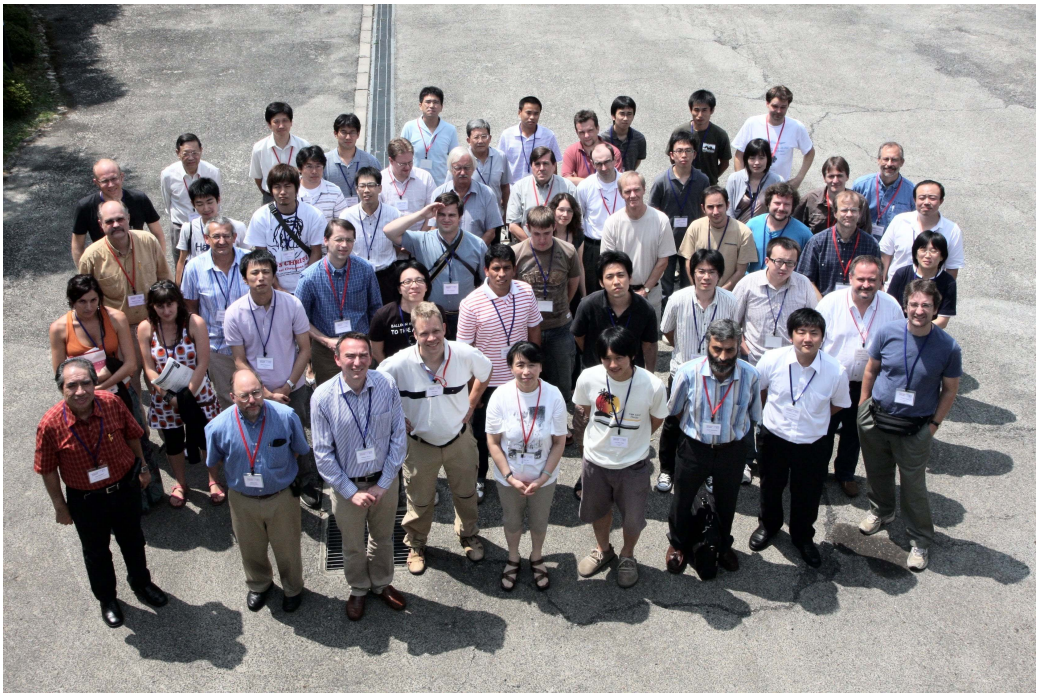
Prof Atsushi Nakagawa
Research Center for Structural and Functional Proteomics,
Institute for Protein Research, Osaka University,
3-2 Yamadaoka, Suita, Osaka, 565-0871
Japan
Tel: +81-(0)6-6879-4313
Fax: +81-(0)6-6879-4313
E-mail: atsushi@protein.osaka-u.ac.jp
WWW: http://www.protein.osaka-u.ac.jp/rcsfp/supracryst/

Dr Lukas Palatinus
Laboratory for Crystallography,
Swiss Federal Institute of Technology
BSP - Dorigny
CH-1015 Lausanne
Switzerland
Tel.: +41 (0)21 693 0639
Fax: +41 (0)21 / 6 93 05 04
E-mail: palat@fzu.cz

Dr Riccardo Spagna
Institute of Crystallography - CNR
Seat of Monterotondo
Area della Ricerca Roma 1
Via Salaria Km 29.3
00016 Monterotondo Stazione (Roma) Italy
Office tel.: +39.06 90672 614
Office fax : +39.06 90672 630
Cellular : +39 339 2766710
E-mail: riccardo.spagna@gmail.com and riccardo.spagna@ic.cnr.it
WWW: http://www.ic.cnr.it/ and http://www.ic.cnr.it/spagna.php

Prof. Dr. Anthony L. Spek
Director of National Single Crystal Service Facility,
Utrecht University,
H.R. Kruytgebouw, N-801,
Padualaan 8, 3584 CH Utrecht,
the Netherlands.
Tel: +31-30-2532538
Fax: +31-30-2533940
E-mail: a.l.spek@uu.nl
WWW: http://www.cryst.chem.uu.nl/spea.html
WWW: http://www.cryst.chem.uu.nl/platon/

Participants of the IUCr Computing School, Kansai Seminar House, Kyoto, Japan, Monday 18th - Saturday 23rd August 2008.



Organisers: Anthony L. Spek, Min Yao, Ralf Grosse-Kunstleve, Harry Powell, Atsushi Nakagawa, Lachlan Cranswick

Lecturers: Luc Bourhis, Hai-fu Fan, Chris Farrow, Ralf Grosse-Kunstleve, Rob Hooft, Garib Murshudov, Gábor Oszlányi, James Pflugrath, Harry Powell, Anthony L. Spek, Dmitri Svergun, Thomas Terwilliger, Brian Toby, Clemens Vonrhein, David Watkin, Min Yao

Registrants: Lucas Bleicher, Xim Bokhiml, Przemyslaw Dera, Santiago Garcia-Granda, Richard Gildea, Saulius Grazulis, Xuanzin Gu, Tomofumi Hashida, Teh Aik Hong, Kazuki Kawahara, Kengo Kitadokoro, Ivan Laponogov, Naohiro Matsugaki, Takanori Matsura, Marcus Meuller, Mathias Meyer, Noriko Nakagawa, So Nakagawa, Shota Nakamura, Terakazu Nogi, Kazuhiro Ohta, Chimari Okada, Jose Luis Pinto Camargo, Laura Roces Fernandez, Che-Hsau Shih, Teakeno Shinohara, Shinichi Takata, Jeremy Tame, Andrea Thorn, Laura Torre Fernandez, Lijie Wu, Akihiro Yamamura, Keitaroh Yamashita, Jian Yu, Tao Zhang, Yong Zhou

Sponsors:

IUCr2008 Osaka and International Union of Crystallography (IUCr)
Bruker AXS
Cambridge Crystallographic Data Centre
CCP4 - Collaborative Computing Project Number 4 in Protein Crystallography
Hampton Research
Hokkaido Wako
Infocom
Newtech
Oxford Diffraction
Phenix
Rigaku
SGI Japan, Ltd.

PHARE2009

Martina Franca (Italy) 15-24 April 2009

A MODULAR WORKSHOP ON GLOBAL PHASE RETRIEVAL

Crystallographic methods for the solution of crystal structures are of paramount importance for crystallography: their efficacy is responsible for the success or the failure of the research, and governs the economical aspects of the activity. The PHARE workshop will focus on crystallographic methods which have recently undergone an impressive acceleration along several different lines of research.

The **PHARE workshop** aims at:

- presenting the crystallographic bases of the methods;
- describing the latest advances obtained by the IC research group in Bari and Rome and by the colleagues from the University of Perugia;
- demonstrating the software packages that implement such advances by means of interactive computer sessions;
- obtaining a feedback from the users, so that the program functionalities can be adapted to the latest user demands.

PHARE will contribute to the general background of the participants and to their specialist expertise, in such a way that the computer programs, no more used as black boxes, can reveal all their potential.

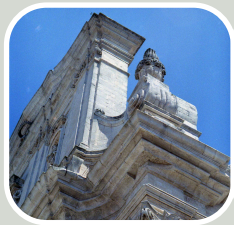
While the crystallographic phase problem is definitively solved in practice for small molecules, it is still a challenging problem in Powder and in Macromolecular Crystallography.

Peak overlapping, preferred orientations and background uncertainty in powder crystallography make uncertain the diffraction intensities. In the macromolecular area, the molecular complexity and the low data resolution make the AB INITIO crystal structure solution a very hard task. Molecular Replacement techniques are often used when similar structures are already known. Alternatively, additional (and expensive) efforts are necessary to prepare isomorphous derivatives and collect supplementary data, or to use anomalous dispersion effects to increase the experimental information available for the solution of the phase problem.

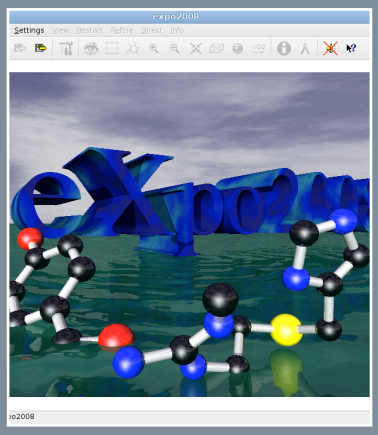
In all these cases a large amount of resources could be spent to attain the goal.

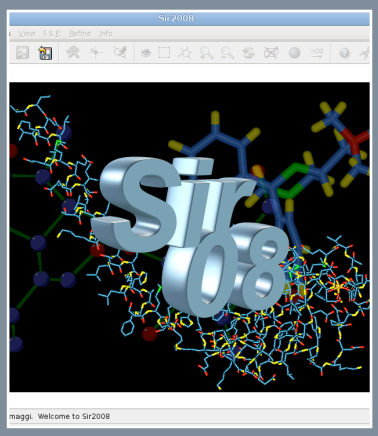
SIR (Single Isomorphous Replacement), MIR (Multiple Isomorphous Replacement), SAS (Single Anomalous Scattering), MIRAS (the combination of MIR with Anomalous Scattering) and MAD (Multiple Anomalous Dispersion) are the current techniques used to solve a completely unknown macromolecular structure.

PHARE2009 web site: http://phare.ic.cnr.it



IC SOFTWARE





Further information

Please refer to the PHARE2009 web site, http://phare.ic.cnr.it, for further information about the workshop, online registration and contact addresses.

SCIENTIFIC PROGRAM

PHARE will cover both powder crystallography and macromolecular area. It will consist of the following modules:

Module 1 (15-17 April 2009)

Crystal structure solution and refinement from powder data

Module 2 (20-22 April 2009)

Macromolecular crystal structure solution via AB INITIO, SIR-MIR, SAD-MAD and MOLECULAR REPLACEMENT techniques

Module 3 (23-24 April 2009)

Phasing small molecules

The IC staff is not sure that a module focused on small molecules (involving Direct Methods, Patterson methods, electron diffraction, Least Squares, Simulated Annealing techniques for locating flexible molecules, Fourier Syntheses, Hydrogen atom location, Minimization of the resolution bias) may be of interest for a sufficiently large number of the IC software users. Anyone interested to attend such a module should send the specific form to the secretariat as soon as possible. In case of positive answer, a third module on these topics will be organized just after the end of module 2.

PARTICIPATION

There will be no registration fee. Each participant can attend only one module. The attendance to two modules will be allowed only exceptionally.
The number of allowed participants for each module cannot exceed 50.

GRANTS

A number of grants is available for young scientists, to cover hotel costs.

Kyoto-2008
Crystallographic Computing School

Introductory Talk

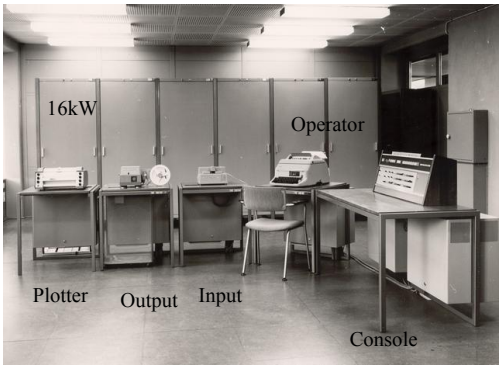
Ton Spek, Kyoto, 18-08-2008

WELCOME
Irasshaimase!

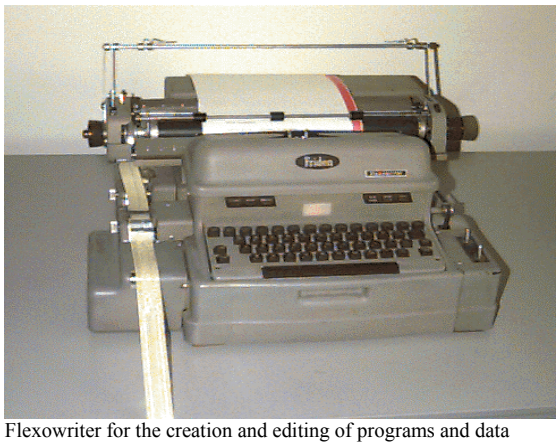
- Introductory Talk, History
  Ton Spek
- Overview of the Scientific Program
  Ralf Grosse-Kunstleve
- Practical Details,
  Min Yao

Some History

- Many of the older software developers, like me, have a background in Direct Methods for solving the phase problem that is today essentially solved.
- I started in the mid 60's, more than 40 years ago, at Utrecht University in the Netherlands as a small molecule crystallographer.
- At that time, I had to write my own Direct Methods program in the ALGOL-60 language in order to solve my structure.
- The next slide shows a university mainframe at that time (essentially single user, FG/BG).



~1966, Electrologica X8 ALGOL60 'Mainframe' (<1MHz)



Flexowriter for the creation and editing of programs and data

Direct Methods Meetings

- Many meetings and schools in the 70's were organized with Direct Methods (software and theory) to solve the phase problem as the central theme.
- Inspiring were the CECAM Direct Methods workshops in Orsay (France) bringing together experts in the field to work on current issues for 5 weeks ! around a (for that time) big European IBM-360 with lectures by Herbert Hauptman.
- Famous were the schools in Parma, York & Erice.
- Photo of the participants of the 1978 Erice School next :

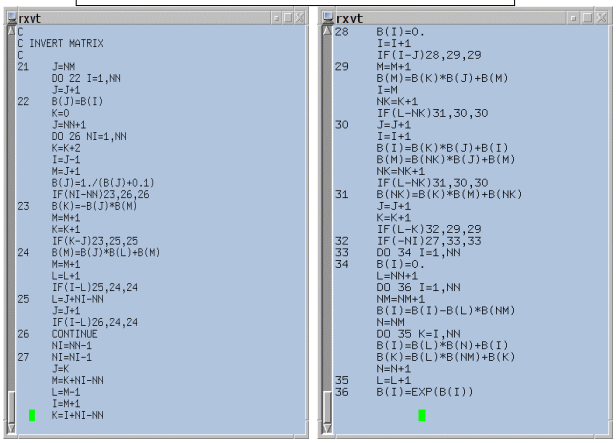
Tools and Platforms Changed

- MULTAN from the York group came out as the standard phase solution program for 15 years.
- Multiuser Mainframes, FORTRAN and PUNCHCARDS became the standard platform.
- MULTAN was eventually superseded in the 80's by the even more powerful SHELXS, SIR & DIRDIF software.
- In the 90's S&B, SHELXD entered the field, coming down from Macro-crystallography.
- Mainframe ➔ Lab Mini (MicroVAX) ➔ PC/WS

Software Languages

- Crystallographic software has been written in *machine language, assembly language, algol60, (turbo)basic, (turbo)pascal, Fortran, C, C++* and various scripting languages such as *python*
- 'Stone-age' *Fortran* based software is still ubiquitous in the small-molecule world (ORTEP, SHELX, CRYSTALS, SIR, DIRDIF, PLATON etc.)
- New (commercial) software development mainly in C++ and scripting languages.
- A project in the UK aims at a rewrite and documentation old Fortran based software to C++ (Durham, Oxford Age-problem project (David Watkin)).

SHELX76-STYLE FORTRAN



Current Hardware Platforms

- MS-Windows: (PC)
  - Small-Molecule Crystallography
  - Powder crystallography
- UNIX/LINUX/OSX: (PC or Cluster)
  - Macro Crystallography
  - (Small-Molecule Crystallography)

Current Computing Areas

- Bio-crystallography: Phasing, Building, Refinement, Graphics etc.
- Chemical Crystallography: Powder Diffraction, Charge Density Studies, Incommensurate Structures, Diffuse Scattering, Structure Analysis, Charge Flipping.
- General: GUI's, Data collection & Data Reduction, Databases, Validation, Automation.

IUCr Computing Schools

- ….. Mostly held jointly with IUCr Assemblies – Examples
- 195? - Pepinski
- 1963 - Rollett, Algorithms (black book)
- 1969 - Least-Squares & Absorption Correction (SHELX76 - code)
- 1978 - Program systems (SHELX, XTAL, NRCVAX etc.)
- 1996 - Macro-crystallography
- 1999 - Macro-crystallography
- 2005 - Siena (Small, Macro)
- Photo of Siena (It) school next ➔

Computing School Siena 2005

Outcome of Siena-2005

- The Siena lectures have been archived in the Sept-2005 Newsletter of the computing commission

http://www.iucr.org/iucr-top/comm/ccom/newsletters/2005sep/iucrcompcomm_sep2005.pdf

- Paper on the Hooft parameter based on discussions in Siena

Hooft et al. (2008). J. Appl. Cryst.,41, 96-103

Motivation for this Crystallographic Computing School

- A general feeling within the small-molecule community is '*The current generation of software developers is phasing out. Where is the new generation to keep things running in the future'*
- There exists a growing community of push-button users using Black Box and Proprietary Software
  - *What is not behind a button can not be done…*
  - *Lack of info about the algorithms used*
- A new generation of crystallographers should be trained to maintain, modify and develop Open Source software to secure continuity and scientific advance.

Issues to Consider

- A large FORTRAN code legacy
- Evolutionary Update of Current Software or Start from Scratch
- Documentation of currently Implemented Algorithms
- Toolboxes
- Funding of Software maintenance and development

Thanks to our Sponsors !

- Bruker-Nonius AXS
- Cambridge Crystallographic Data Center
- CCP4
- Hokkaido Wako
- IUCr-Osaka
- Infocom
- NTC
- Oxford Diffraction
- Phenix
- Rigaku
- Sgi-Japan

Integrated Crystallography

The CRYSTALS Experience

David Watkin
Chemical Crystallography Laboratory
Oxford

This introductory lecture on the first evening was intended to take the participants through the good and the bad things which had been done during the 40 years life of CRYSTALS. Like PLATON, CRYSTALS is a program in a continual state of change, with new ideas being incorporated regularly and new issues being released almost every year. This is in contrast to SHELX, which metamorphoses every 20 years or so to reappear with a quantum leap in features and facilities. The organic development path means that the authors have little opportunity to carry out major internal revisions of the data management or organisation of the program. For incremental development, the program has to be built on a well designed and well implemented plan otherwise it quickly descends into anarchy.

In this talk, emphasis was placed on the design of a flexible, extensible and well documented data management strategy. Part of this design is the development of a method for reliably, unambiguously and conveniently referring to atoms and parameters. The audience was encouraged to clarify their thoughts on the relationships between atoms, atom names, parameter names and parameter values. The idea of separating parameter values and operations on parameters was explored. Though this separation seems computationally logical, the popularity of SHELX has shown that it is not what users want.

The history of the GUI was traced, and in particular its relationship with the underlying crystallography. The link between the GUI and the crystallography is via a single pipeline Command-line instructions are passed down to CRYSTALS for execution. However, because the underlying program and the GUI have both been developed by the same team, increased functionality has been added to CRYSTALS to provide facilities required by the GUI writers.

The lack of good documentation for either users or programmers was lamented. For the users there is a substantial reference manual, but no training manual. For the programmers there is good documentation at the detailed code level, but no overviews or descriptions of the overall strategy or organisation of the program.

The talk ended with some comments and observations.

Conclusions

CRYSTALS has survived as an actively developing program because of the massive effort which went into the design of the overall structure of the program and into the unified data management.

The FORTRAN will probably work as long as compilers are available.

The internal data structure was sufficiently 'open' to enable developments which could never have been foreseen by the original designers, but has now been stretched to its limit. This restricts the possibility of including any major new developments to the program

The GUI, interfacing with rapidly changing graphics standards, is the most vulnerable part of the program. This is the part most likely to affect the long-term portability and usability of the package.

What have we learned?

- Users would prefer to press buttons rather than think about a problem – *the Microsoft syndrome.*
- Users do not want complicated environments – *you have to hide your cleverness from them.*
- Users will only read a manual as a last resort – *but a full reference manual must exist if only for your own sake.*
- Users do not understand the programmers problems – *don't expect sympathy when things go wrong.*
- Programmers do not understand the users problems – *but users probably don't understand them either.*
- Programmers relish complexity and compactness – *it is a symbol of how clever they are.*
- What is clear today will become obscure tomorrow – *write code that you will never have to re-visit, but just in case you do, comment it.*
- Do not re-invent but do improve the wheel – *stand on the shoulders of giants*.
- Remember that you are not the only expert in the world – *listen carefully to your colleagues, critics and users.*
- Ensure your sponsors will let you share the source code – *otherwise it will certainly die.*
- The internal data structure must be well defined and rigid – *ad hoc data definitions lead to duplication and confusion.*
- Avoid near-duplication of procedural functionality – *spend a little more time on generalising one function.*

Before we Start

The Golden Arrow

The British Crystallographic Association is trying to encourage speakers to use the golden arrow as a pointer rather than use a laser-pen.

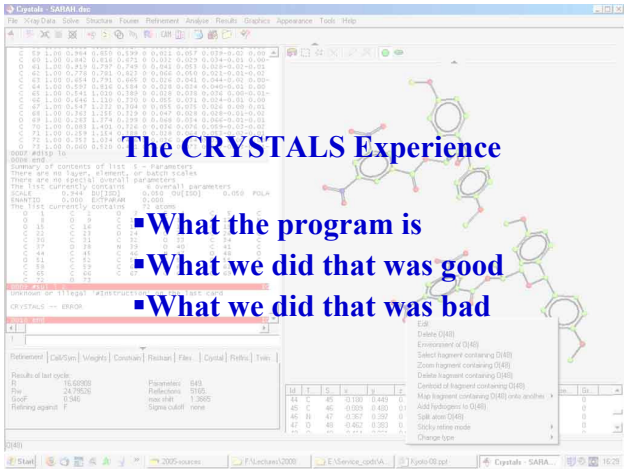
1

Integrated Crystallography

The CRYSTALS Experience

David Watkin
Chemical Crystallography Laboratory
Oxford

2



CRYSTALS

CRYSTALS is the product of many people, of whom the real founders were:

Keith Prout	raised the initial funding
John Rollett	understood the details of the crystallography
Bob Carruthers	was a brilliant programmer

4

CRYSTALS

Contributors

Contributor	Dates	Comments
JS Rollett	1967-1979	Deceased
CK Prout	1967-2001	Deceased
JR Carruthers	1967-1976	Out of Crystallography, Retired
R Spagna	1970-1975	Retired
DJ Watkin	1967-2007	About to Retire
P Betteridge	1982-1985	Out of Crystallography, Perhaps contactable
RI Cooper	1996-2004	Contactable
LJ Pearce	1991-1995	Out of Crystallography, Perhaps contactable
S Parsons	Various	Contactable
L Schroeder	1995-1996	Out of Crystallography, Perhaps contactable
A van der Lee	2005-2007	Contactable
Minor contributors (about 20)	Various	Whereabouts unknown

5

CRYSTALS

Volume of system.

Function	File type	bytes
Core FORTRAN	.fpp, .INC	5,363,007
GUI	.cc, .h	1,376,501
SCRIPTS	.scp, .sda	1,023,099
DATA	.ssr	83,951
Manual sources	.man	1,114,776
Processed Manual	.pdf, .html, .jpg	2,155,192
Website	.html, .jpg	3,746,662
Foreign Programs	.for	428, 818
Self-installing Distribution Kit		16,523,866

6

CRYSTALS

CRYSTALS is a program (system) for crystal structure analysis. It has evolved over 40 years in response to the needs of crystallographers, chemists and physicists.

'EDITION 1'	(Plain text data base definition) (Cruickshank, Freeman, Rollet, Truter, Sime, Smith and Wells, 1964)	
'NOVTAPE'	(In AUTOCODE)	(Hodder, Rollet, Prout and Stonebridge, Oxford, 1964)
'FAXWF'	(In ALGOL)	(Ford and Rollett, Oxford, 1967),
'CRYSTALS'	(In FORTRAN)	(Carruthers and Spagna, Rome, 1970)
'CRYSTALS'	(In FORTRAN, major re-write) (Carruthers, Prout, Rollet and Spagna, Oxford, 1975)	
'CRYSTALS'	Issue 2 (In FORTRAN, major re-write) (Carruthers, Prout, Rollet and Watkin, Oxford 1979)	
'CRYSTALS'	Issue 7 (in FORTRAN, VAX SMG user interface) (Betteridge, Prout and Watkin Oxford, 1983)	
'CRYSTALS'	Issue 11 (in FORTRAN with C++ GUI) (Watkin, Prout, Carruthers, Betteridge, Cooper, 1997)	

7

Machine Specifics

30 years ago there was a bewildering range of computers available. All had their own operating system and system libraries.

All had their own version of FORTRAN, with lots of 'useful extensions'

Jim Stewart promoted 'pidgin FORTRAN' – in evolved form still used by George Sheldrick and CRYSTALS

8

Machine Specifics

X-ray was re-written in RATFOR and RATMAC – home-made languages which could be translated in to machine specific FORTRAN

In CRYSTALS, machine specific code was collected into a single module. This dealt with such things as date, time, and most importantly, file manipulation.

9

CRYSTALS

Key Components:

1. Best available algorithms
2. Well designed global database
3. Hierarchical program structure
4. Consistent user-syntax
5. Users Documentation
6. Programmers Documentation

10

Primary User-interface

All the crystallography is available from the command line or a file of commands.

For the core program, only a limited number of I/O channels are permitted.

1. Reflection data
2. Commands, directives and other numerical data
3. Scratch files
4. Binary Database

11

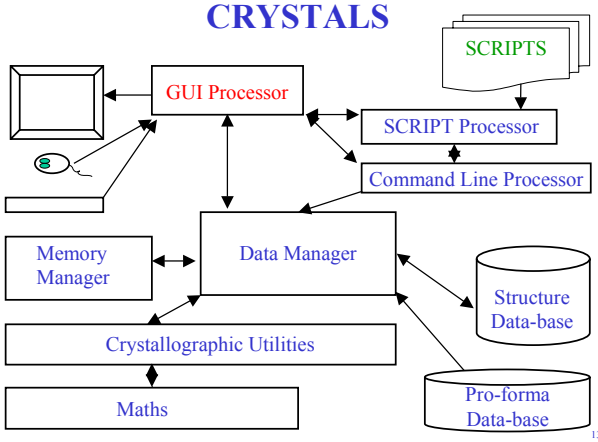
User-interface

CRYSTALS can be regarded as a subroutine library, in which the "subroutines" can be called directly by the user from a command line interface. Intermediate results are stored in a binary data base for subsequent use.

```
#command (optional parameters)
directive (optional parameters)
end

#fourier
map type=2fo-fc
end
```

12

CRYSTALS



13

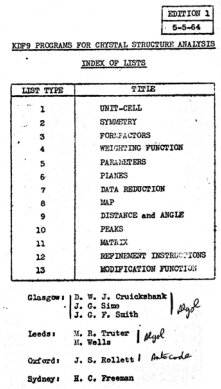
Data Management - the database

Unlike many programs from the 1960's, which intermixed data and commands and were developed in an ad-hoc fashion, the internal and external organisation of the data were based on an abstract model: *Edition 1*.

Cruickshank, D. W. J., Sime, J. G., Smith, J. G. F., Truter, W. R., Wells, M. Rollett, J. S. and Freeman, H. C. (1964). Computing Laboratory, Oxford University, Oxford.

14

Edition 1



EDITION 1
5-5-64

KPS PROGRAMS FOR CRYSTAL STRUCTURE ANALYSIS

INDEX OF LISTS

LIST TYPE	TITLE
1	UNIT-CELL
2	SYMMETRY
3	FORM FACTORS
4	WEIGHTING FUNCTION
5	PARAMETERS
6	PLANES
7	DATA REDUCTION
8	MAP
9	DISTANCE and ANGLE
10	PEAKS
11	MATRIX
12	REFINEMENT INSTRUCTIONS
13	MODIFICATION FUNCTION

Glasgow: D. W. J. Cruickshank, J. G. Sime, J. G. F. Smith | Mercury
Leeds: M. R. Truter, M. Wells | Mercury
Oxford: J. S. Rollett | Autocode
Sydney: H. C. Freeman

15

Data Management - the database

External Data Representation

The strategy of *Edition 1* was implemented in programs written in Autocode, Algol and FORTRAN, and is still evident today in both CRYSTALS and in CAOS (Cerrini S. & Spagna R.,1977).

Groups of related kinds of data are gathered together into '***lists***' – e.g. lists of atom parameters, lists of atomic properties, lists of reflection data etc.

16

Data and Memory Management - the database

Internal Data Representation

The well defined external '***list***' structure translates easily into an internal structure.

At that time, major programs (X-ray, XTAL, CRYSTALS, SHELX, SIR) allocated most of the memory as a single large array, with hand-crafted pointers indexing data in this array.

Lists and lists-of-lists meant that areas of memory could be allocated and purged cleanly.

17

CRYSTALS

External Representation

This clear, documented, data structure discouraged the proliferation and duplication of ad-hoc variables.

The well-organised memory-resident internal representation of the data maps easily onto an external direct access (binary) representation – a structural database.

18

CRYSTALS

Binary Database (bdb)

Advantages	Disadvantages
Quick startup for new tasks	
Start new task where previous finished	Can be confusing if a preceding task ended incorrectly
Possibility of 'undo' feature	
User cannot 'fiddle' with the data	Some kind of editor required for user-modifications to data
Large amount of information can be exchanged between tasks	

Curiously, while most scientists are happy with the bdb in Word or Photoshop, they seem to be worried about a crystallographic bdb

19

CRYSTALS

1970-76 Pro-forma Database

Formal definitions for:

- List-based data, *e.g.* atoms, reflections, symmetry, *etc*.
- Lexical (text) Data, *e.g.* constraints, restraints.
- Commands or directives.

Templates are held in a pro-forma database. These define both the order of parameters, and the keywords used to identify them. They also provide default values where applicable.

20

CRYSTALS

1970-76 Pro-forma Database

```
ATOMIC PARAMETERS
(1X,120A1/85X,36A1,12A1,48A1)
(1X,A4,F13.0,1X,8F12.5/79X,3F12.5,F8.1,3(10X,F2.0),A4)
    .TYPE              0   -1    1    0    0
    .SERIAL            0   -1    1    0    0
    .OCC               0    0    1    0    0 1.0
    .FLAG              0    0    1    0    0 1.0
    .X                 0   -1    1    0    0
    .Y                 0   -1    1    0    0
    .Z                 0   -1    1    0    0
    .U[11]             0    0    1    0    0 0.05
    .U[22]             0    0    1    0    0 0.0
    .U[33]             0    0    1    0    0 0.0
    .U[23]             0    0    1    0    0 0.0
    .U[13]             0    0    1    0    0 0.0
    .U[12]             0    0    1    0    0 0.0
    .SPARE             0    0    1    0    0 0.0
    .PART              0    0    1    0    0         0
    .REF               0    0    1    0    0         0
    .RESIDUE           0    0    1    0    0         0
```

21

ATOMS

Small molecule crystallographers work with atoms.

Devising a system-wide syntax for identifying atoms and atom parameters crucially affects the usability of a program for complicated situations.

Atoms need to be uniquely identified so that there is no confusion either by the user or the program.

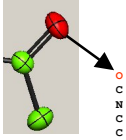
22

ATOMS

Lexical Data: the naming of atoms.

An **atom** is characterised by a set of data - the atomic parameters.

These include an identifying code, the element type, the positional coordinates, U^{ij} etc

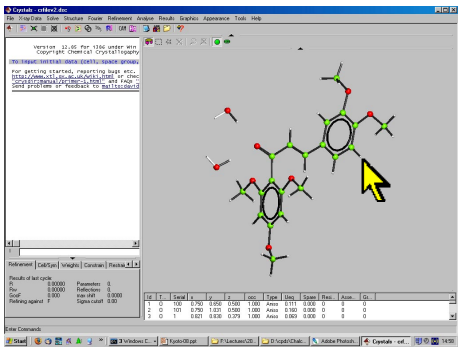


		Occ	x	y	z	U11	U22	U33	U23	U13	U12
O	1	1.00	0.477	0.607	0.184	0	0.025	0.050	0.055	0.00	0.00	0.00
C	2	1.00	0.236	0.616	0.165	0	0.029	0.025	0.037	0.00	0.00	0.00
N	3	1.00	0.039	0.591	0.224	0	0.027	0.037	0.031	0.00	0.00	0.00
C	4	1.00	0.104	0.547	0.309	0	0.027	0.034	0.033	0.00	0.00	0.00
C	5	1.00	0.238	0.631	0.371	0	0.040	0.032	0.034	0.00	0.00	0.00

23

ATOMS

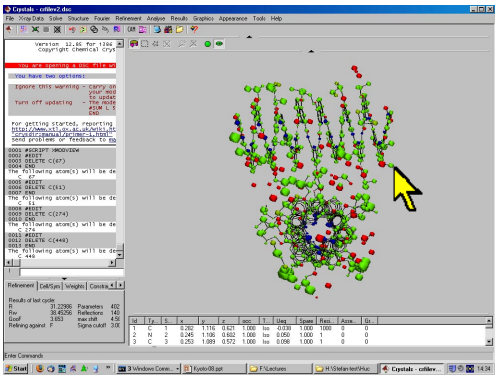
Atoms need to be uniquely identified so that there is no confusion either by the user or the program. We could use point-and-click …



24

ATOMS

We could use point-and-click …however, when the structure is large



25

ATOMS

To refer to an atom :

We could say '*the atom near .123, .234, .345*' which would be unique in the absence of disorder, but which could become meaningless after refinement.

We could say '*the 7th atom in the list of atoms*' but this could be misleading if atom 5 were deleted.

Bob Carruthers chose to combine the values of the parameters TYPE and SERIAL, *e.g.* Na(2)

26

ATOMS

Lexical Data: Atom parameters

An **atom** is characterised by a set of data - the atomic parameters.

An atomic **parameter** is characterised by a name and a value, *e.g.* '*z*' *and* .224

The atomic data can be visualised as a matrix indexed by row and column

A user may need to refer to a row, a column or a cell.

```
      Occ     x       y       z       U11     U22     U33   U23   U13   U12
O   1 1.00 0.477 0.607 0.184 0 0.025 0.050 0.055 0.00 0.00 0.00
C   2 1.00 0.236 0.616 0.165 0 0.029 0.025 0.037 0.00 0.00 0.00
N   3 1.00 0.039 0.591 0.224 0 0.027 0.037 0.031 0.00 0.00 0.00
C   4 1.00 0.104 0.547 0.309 0 0.027 0.034 0.033 0.00 0.00 0.00
C   5 1.00 0.238 0.631 0.371 0 0.040 0.032 0.034 0.00 0.00 0.00
```

27

Parameters

For maximum flexibility, parameters also need to be uniquely identified.

This was achieved by combining the atom name with the parameter name, *e.g.* Na(2,x).

This enables the user to do things to the ***parameter***:

`FIX Na(2,y)` constraint removes the y coordinate from the refinement

`FIX y` constraint removes all the y coordinates from the refinement

or to the ***parameter value***:

`ADD 0.5 Cu(123,x)` shifts the copper x coordinate by half a cell.

28

Higher level abstractions

Groups of contiguous **atoms** in the atom list

C(1) until N(6)

Contiguous groups of **atom parameters**

F(100,occ) until F(103)

All atoms, or given parameters for all the atoms

ALL(x)

Atoms can belong to RESIDUES, ASSEMBLIES and GROUPS (cif definitions).

29

Equivalent Atoms

For anything other than trivial situations, one needs to be able to access atoms both at the position given in the atom list, and in symmetry related equivalent positions.

The syntax chosen was

TYPE(SERIAL, ±S,L,Tx,Ty,Tz)

where S and L are indices into the tables of symmetry operators and centring operators, and T are whole cell translations.

e.g. C(14,-1,2,0,1)

The cif (X-Ray) syntax is more concise C14 3_756

30

Atom Editor

While it is sometimes convenient to use a text editor to modify parameter values, it is generally easier to use a crystallographic editor.

This is designed to perform crystallographic, lexical or logical operations on atoms or atom parameters

e.g.

```
SELECT U[ISO] .LE. 0.15

SHIFT 0.25 0.0 -.25 C(1) UNTIL C(17)
```

31

Keeping Track of Atoms

Atomic housekeeping is a major consideration for large program design, especially if the underlying program can also be accessed via a GUI.

What will the program do if the user

- Creates two atoms with the same identifier.
- Changes the order of atoms in the list.
- Deletes an atom which is also referenced in some other data structure (e.g. a restraint).
- Changes a parameter value which affect derived parameters (e.g. element type implies covalent radius and hence bonding).
- Wishes to refer to atoms related by symmetry operations.
- Shifts the whole or part of a structure by a symmetry operator.

A global solution to these and similar issues simplifies things for the user and the programmer.

32

Synchronisation

It is important that the consequences of changes in one part of the data structure are propagated.

e.g. Changing an atom type (*e.g.* C to N) will affect the bonding table (matrix) and the calculated structure factors.

Should this be done:

1. Automatically every time a change is made?
2. Under program control?
3. Under user control?

CRYSTALS uses methods 2 and 3 as appropriate.

33

Separating values and operations

Summary so far

CRYSTALS distinguishes between atoms, atomic parameters, parameter types and parameter values.

This means that the parameter values can be kept in a list which is quite separate from operations which act upon them.

In 1976 *modular* programming and data were in fashion. They were seen as real advantages for both the user and the programmer.

34

Separating values and operations

Advantages for the user:

A parameter value list - these are the ***actual values***:

```
# Punched on 30/04/08 at 16:42:36
#
#LIST      5
READ NATOM =     4
ATOM O       1.000000   1.000000   1.000000   1.956593   0.338930   0.525810
ATOM N       2.000000   1.000000   1.000000   1.900175   0.296523   0.442821
ATOM C       3.000000   1.000000   1.000000   1.831805   0.163828   0.430223
ATOM C       4.000000   1.000000   1.000000   2.103274   0.212306   0.397993
END
```

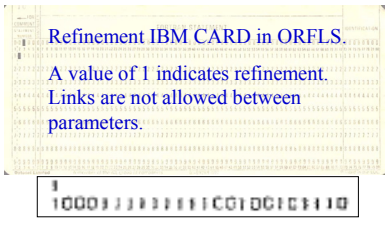
A lexical list - this is ***what to do*** with the parameters:

```
#LIST 12
FULL X'S U[ISO]
FIX F(18, U[ISO])
END
```

35

ORFLS (1962)

Separating parameter values and operators has a long history

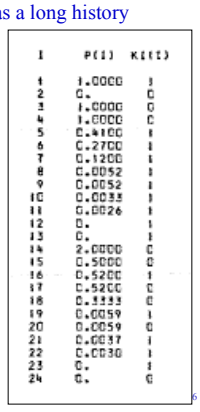


Refinement IBM CARD in ORFLS.

A value of 1 indicates refinement. Links are not allowed between parameters.

Right, output:

columns are parameter number, value and refinement flag.



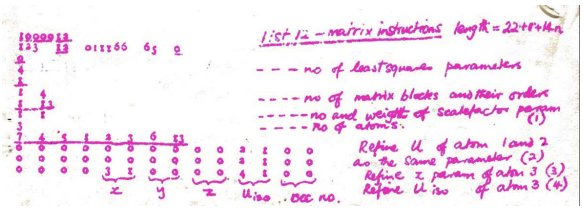
I	P(I)	K(I)
1	1.0000	1
2	0.	0
3	1.0000	0
4	1.0000	0
5	0.4100	1
6	0.2700	1
7	0.1200	1
8	0.0052	1
9	0.0052	1
10	0.0033	1
11	0.0026	1
12	0.	1
13	0.	1
14	2.0000	0
15	0.5000	0
16	0.5200	1
17	0.5200	0
18	0.3333	0
19	0.0059	1
20	0.0059	0
21	0.0037	1
22	0.0030	1
23	0.	1
24	0.	0

36

Nov-Tape

Semi-machine code precursor to CRYSTALS. Rollett et al, 1964

There is a row for each atom, and a pair of columns for each parameter.



This strategy permits easy implementation of special position constraints and disorder.

CRYSTALS

Symbolic representation of refinement directives:

Refine the x,y and z parameters of all non-H atoms, but allow H1 and H2 to ride with C1.

```
#list 12
full x's
ride c(1,x's) h(1,x's) h(2,x's)
end
```

It is not necessary, or useful, to know the parameter values themselves. 'Free variables' are allocated automatically by the program.

Separating values and operations

Advantages for the programmer:

```
IF(KHUNTR(5,0,IADDL,IADDR,IADDD,-1).NE.0)  CALL XFAL05
```

If the atom parameters (LIST 5) are not already loaded into memory, fetch them from the binary database.

25% of the FORTRAN code provides the infra-structure which helps the programmers to concentrate on the science.

Separating values and operations

In reality –

The advantages to programmers are very clear.

A substantial number of people, including students, have been able to contribute to the code of CRYSTALS, enabling it to rapidly evolve.

Separating values and operations

However –

For the users, the situation is less clear cut.

People coming from a background where they use text editors to prepare a task and then execute it, find it difficult to separate *the atoms and actions on the atoms* from a list of *parameter values*.

SHELXL

```
This is an example of AFIX (the atoms are isotropic because it is easier to see
what's going on that way).

P1    5   0.69271   0.36294   0.16280  11.00000   0.05000
AFIX  66
C1    1   0.62261   0.32910   0.07077  11.00000   0.05000
C2    1   0.64110   0.38758   0.01348  11.00000   0.05000
AFIX  43
H2    2   0.69568   0.44990   0.02075  11.00000  -1.20000
AFIX  65
C3    1   0.58205   0.35831  -0.05538  11.00000   0.05000
AFIX  43
H3    2   0.59534   0.40112  -0.09488  11.00000  -1.20000
AFIX  65
C4    1   0.50519   0.26851  -0.06648  11.00000   0.05000
AFIX  43
H4    2   0.46359   0.24953  -0.11379  11.00000  -1.20000
AFIX  65
C5    1   0.48736   0.20584  -0.01074  11.00000   0.05000
AFIX  43
H5    2   0.43482   0.16146  -0.01910  11.00000  -1.20000
AFIX  65
C6    1   0.54430   0.23418   0.05739  11.00000   0.05000
AFIX  43
H6    2   0.53115   0.18959   0.09628  11.00000  -1.20000
AFIX   0
```

The occupation factors are removed from the refinement by adding 10 to the actual value

Hydrogen atom treatment mixed up with values

Documentation

Programmers Documentation:

Ideally, should be hierarchical.

- Overview of system
  - Overview of principal components
    - Functionality or algorithms of main components
      - Description of modules
        - comment in code

Documentation

Programmers Documentation in CRYSTALS:

Overview of system – *only exists as superficial articles at computing schools*

Overview of principal components – *does not exist*

Functionality or algorithms – *sometimes good, mostly poor*

Description of modules – *sometimes good.*

comment in code – *generally very good*

Modifying an existing module is generally easy.

Creating totally new functionality is difficult.

Documentation

Comment in code – generally very good

The total FORTRAN source is 168,156 lines, of which COMMENT is 56,840 lines.

If you can find the module you need to edit, small changes are generally easy to make.

CRYSTALS FORTRAN

For ease of access, the source code is divided into 60 files based on functionality

absorb.fpp	empabs.fpp	lexical.fpp	modify5.fpp
accumula.fpp	execute.fpp	link.fpp	mtapes.fpp
alter5.fpp	fiddle.fpp	list11.fpp	normal.fpp
aniso.fpp	findfrag.fpp	list12.fpp	paramlist.fpp
anisotfs.fpp	foreig.fpp	list16.fpp	planes.fpp
calcul.fpp	fourie.fpp	list26.fpp	prcss.fpp
characte.fpp	fourier.fpp	list4.fpp	prcss6.fpp
control.fpp	genedit.fpp	list50.fpp	presets.fpp
crystals.fpp	geomet.fpp	list6.fpp	publsh.fpp
csdcode.fpp	geometry.fpp	lists1.fpp	punch.fpp
difabs.fpp	guibit.fpp	lists2.fpp	read.fpp
disc.fpp	hydrogen.fpp	lists3.fpp	read6.fpp
distan.fpp	input.fpp	lists4.fpp	reductio.fpp
distangl.fpp	input6.fpp	matrix.fpp	regular.fpp
	invert.fpp		

Documentation

Functionality or algorithms – sometimes good, mostly poor

Description of modules – sometimes good.

```
      SUBROUTINE XRDHI(IFIRST)
C--THIS ROUTINE READS A CARD FILE CONTAINING A SPECIFICATION OF EACH
C  '#' INSTRUCTION AND ITS ASSOCIATED DIRECTIVES AND PARAMETERS, AND
C  OUTPUTS THE RESULTS IN AN INTERNAL FORMAT TO THE DISC AS A LIST 50.
C  WITHIN THIS LIST, SEVERAL TYPES OF DATA RECORD ARE USED :
C
C  IFIRST   THE NUMBER OF THE FIRST COMMAND ADDED IN THIS RUN. THIS IS 1
C           FOR A COMPLETELY NEW FILE AND THE NUMBER OF THE FIRST NEW
C           COMMAND FOR AN APPEND.
C
C  50  RECORD TYPE 50 CONTAINS THE NAMES OF THE INSTRUCTIONS, EACH ONE
C      2  NAME 1, WORD 2.
C      FOLLOWING DIRECTLY AFTER THE ONE BEFORE IT. THE FORMAT IS
C      ONE CHARACTER PER WORD, AND MAY BE REPRESENTED AS :
C
C           0  THE NUMBER OF CHARACTERS IN THE COMPLETE NAME.
C      1  NAME 1, WORD 1.
```


Documentation

User Documentation: sadly, this is mostly poor.

The CRYSTALS Manual. This is a reference document, almost up to date, explaining every input function or parameter, *but with little advice about how to use the program*.

It is maintained as a text file with a markup language invented by GMS about 1972, though later extended by many people.

Programs exist (in FORTRAN and Perl) to convert it to postscript, HTML and LaTeX. One postscript version of the manual runs to 150 pages, with index and table of contents.

User Documentation - GUIs

These are a popular alternative to proper documentation.

With a suitable infrastructure, writing GUIs can be good fun, offers easy rewards, and lets you see results quickly.

They are an inescapable requirement for a modern system.

However -

It is too easy to make them idiosyncratic, trivial, obscure or misleading.

They can isolate the user from the underlying science, providing little training for what to do when they fail.

49

The Command Line User Interface

Commands typed at terminal, some results output on screen.

One needs to see some of the output all of the time and all of the output some of the time.

The screen output should contain enough information to confirm to the user that the work is progressing properly.

Simultaneous detailed output to a file should be copious for the analysis of difficult issues.

Some kind of filename generation numbers is useful.

VMS: **output.lis;1 output.lis;2 etc**

Windows: **output#00.lis output#01.lis etc**

50

Evolution of the CRYSTALS User Interface

- Punched paper tape in, paper tape + printer out
- Plain text files in and out (ICL 2900 VME/B) 1978
- Plain text file in and out (Vax VMS) (1980)
- **Commands typed at terminal, some results output on screen.**
- **SCRIPTS – question and answer sessions (1982)**
- Tektronix vector graphics for visualisation (1986)
- **VAX VMS SMG$ Menus (1990)**
- **MS VGA graphics for visualisation (1991)**
- MS Windows and ClearWin (Salford Software) (1996)
- **MS Windows, C++ and OpenGL (1997)**

51

CRYSTALS Scripts

In 1984 Bev Vincent, now at Rigaku, spent a few months in Oxford.

He wrote a small question-and-answer program to prompt a CRYSTALS user for the information needed to get started on a new structure.

52

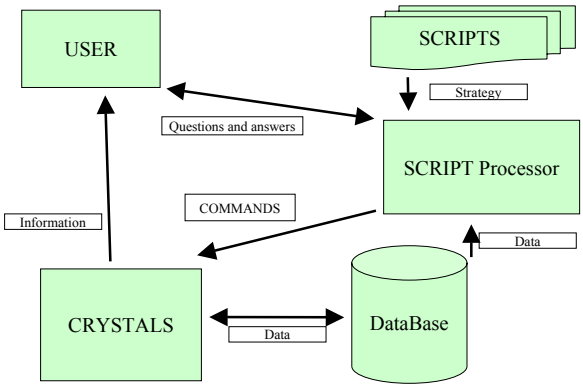
CRYSTALS Scripts

In 1985, Paul Betteridge started building a new I/O layer for CRYSTALS.

This processed a file of commands which instructed the program to interrogate the user, the structural data base and other files, and build up standard CRYSTALS commands for normal processing.

53

CRYSTALS Scripts



54

```
%SCRIPT INLIST1
%% CRYSTALS cell parameter input script
% BLOCK
%   ON ERROR REPEAT
%   ON END  TERMINATE
 Input of cell parameters (  a, b, c, alpha, beta, gamma )
                              use  END  to abandon input
%   QUEUE REWIND
%   CLEAR
%   INSERT 'REAL'
%   GET REAL 'a'
%   GET REAL 'b'
%   GET REAL 'c'
%   GET REAL 'Alpha ( degrees )' '90'
%   GET REAL 'Beta ( degrees )' '90'
%   GET FINAL REAL 'Gamma ( degrees )' '90'
%   QUEUE SEND
%   CLEAR
%   COPY '#LIST 1'
%   QUEUE PROCESS
%   COPY 'END'
% END BLOCK
%%
% FINISH
%END SCRIPT
```

CRYSTALS

Scripts

The full vocabulary includes looping, branching, testing, arithmetic, file handling, extracting and inserting data into the database.

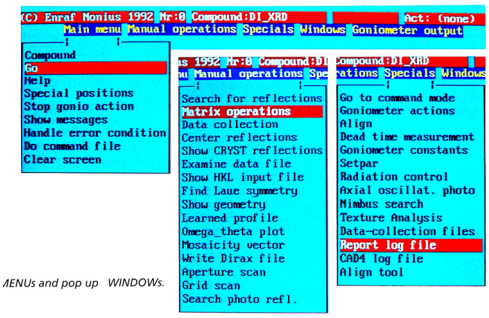
Later additions by Richard Cooper enabled the GUI to be constructed on the fly.

If we were building a new user interface, we would write the top layer in an established language.

55

The CRYSTALS User Interface

VAX VMS SMG$ Menus: Nonius CAD4 Express (1992)



MENUs and pop up WINDOWs.

56

CRYSTALS Graphics

MS DOS VGA graphics for visualisation (1991)

Brief was to determine the most useful features from

CHEM-X	Cosmic	DTMM	MacMolecule
MOLDRAW	Molview	ORTEP	Pluto
Shakal	Snoopi	STRUPLOT	XP.

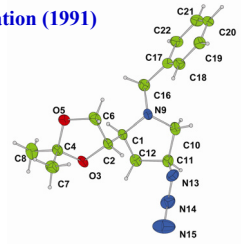
- *use CRYSTALS subroutines and data structure*

57

CRYSTALS Graphics

MS DOS VGA graphics for visualisation (1991)

The result was CAMERON



Unfortunately, **software-anarchy** lead to a new data structure being devised, and swathes of subroutines being taken from CRYSTALS and reproduced with small changes in CAMERON.

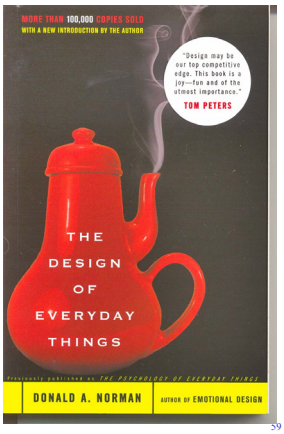
A maintenance nightmare

58

Issues for GUI designs

No one sets out to design a bad GUI, yet we are constantly exposed to programs which are difficult to begin to use.

Fortunately, it is usually easier to re-program users than to re-write the program



59

The CRYSTALS GUI

It is an extension of the SCRIPT processor.

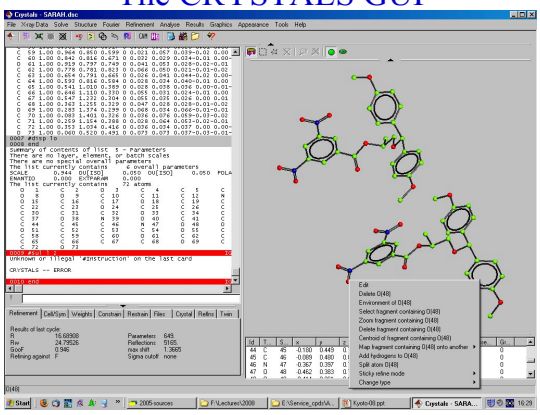
Design factors included:

•Intuition is unreliable – it depends upon experience
•Rescue procedures are needed to enable recovery from disasters & novel situations
•Teaching and training

A good GUI may help a novice get a simple task completed, but it is unlikely to be able to provide much help or support when things go seriously wrong.

60

The CRYSTALS GUI



The CRYSTALS GUI

Except for receiving atomic coordinates, unit cell parameters and symmetry operators from CRYSTALS, all communication between the GUI and CRYSTALS is via the command line interface.

This provides a consistent and well-defined context, and assures synchronisation.

The traffic can be watched and logged, and be edited and re-run without the GUI for debugging purposes.

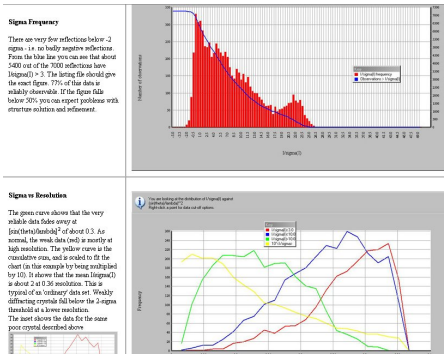
The experience gained with the VAX "**One needs to see some of the output all of the time and all of the output some of the time"** greatly helped in the design of the feedback to the users.

CRYSTALS

The CRYSTALS GUI

An html version of the manual is available from the GUI.

There are also html files which explain the various diagnostics



Interaction with 'foreign' code

"Foreign" code is code not written in the CRYSTALS style, and not based on the CRYSTALS infra-structure.

It may be library routines, a substantial part of another program, or an entire external program.

Interaction with 'foreign' code

Simplest is independent free-standing program, e,g, SHELXS, SIR92, superflip, PLATON.

Suitable when the foreign program can be driven from a plain text input file, and produces a succinct output file.

Interaction with free-standing programs

Advantages:

1. Foreign program is not altered in any way
2. Clean text interface simplifies debugging
3. Foreign program cannot interfere with CRYSTALS data base
4. New versions can be plugged in with minimal effort
5. Usually no licensing issues

Interaction with free-standing program

Disadvantages:

1. If the foreign program only produces verbose 'listing' files, finding data items can be tedious
2. CRYSTALS will need to be changed if the file formats change
3. CRYSTALS has no access to internal data or subroutines.

67

Co-compiling "foreign" code into CRYSTALS

Authors of a foreign programs have permitted us to integrate large chunks of their code with ours, e.g.

- NRCVAX. Derive symmetry operators from SG symbol
- DIFABS. Diagnostic for diffraction-geometry dependent trends in the LSQ residual
- MULTAN. Obtain Wilson Plot information as diagnostic for refinement difficulties (compare Wilson scale, LSQ scale and ΣFo/ΣFc)

68

Co-compiling "foreign" code into CRYSTALS

Advantages:

1. The original authors are probably experts in their field.
2. Small changes can be made to the code to facilitate parameter exchange or provide new features.

Disadvantages:

1. We may be unaware of bug-fixes.
2. We may need to re-integrate local modifications into new versions.
3. If we suspect that there is a bug, it may be difficult to trace.

69

Integration of libraries at code-level

e.g. LAPACK, BLAS,

Advantages:

- Avoid re-inventing the wheel.
- Take advantage of professional optimisation and numerical stability.

Disadvantages:

- Library may not do exactly what you want.
- Often IP issues with recent versions, and old versions are no longer supported.

70

Conclusions

CRYSTALS has survived as an actively developing program because of the massive effort which went into the design of the overall structure of the program and into the unified data management.

The FORTRAN will probably work as long as compilers are available.

71

Conclusions

The internal data structure was sufficiently 'open' to enable developments which could never have been foreseen by the original designers, but has now been stretched to its limit.

This restricts the possibility of including any major new developments to the program

72

Conclusions

The GUI, interfacing with rapidly changing graphics standards, is the most vulnerable part of the program. This is the part most likely to affect the long-term portability and usability of the package.

73

What have we learned?

Users would prefer to press buttons rather than think about a problem – *the Microsoft syndrome.*

Users do not want complicated environments – *you have to hide your cleverness from them.*

Users will only read a manual as a last resort – *but a full reference manual must exist if only for your own sake.*

Users do not understand the programmers problems – *don't expect sympathy when things go wrong.*

Programmers do not understand the users problems – *but they probably don't understand them either.*

74

What have we learned?

Programmers relish complexity and compactness – *it is a symbol of how clever they are.*

What is clear today will become obscure tomorrow – *write code that you will never have to re-visit, but just in case you do, comment it.*

Do not re-invent but **do improve** the wheel – *stand on the shoulders of giants.*

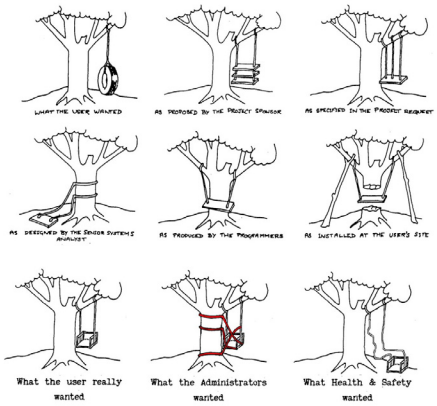
Remember that you are not the only expert in the world – *listen carefully to your colleagues, critics and users.*

Ensure your sponsors will let you share the source code – *otherwise it will certainly die.*

The internal data structure must be well defined and rigid – *ad hoc data definitions lead to duplication and confusion.*

Avoid near-duplication of procedural functionality – *spend a little more time on generalising one function.*

75



76

The End

77

CRYSTALS

Data Management – external representation

Possible models:

1. GUI based
2. Text based (batch mode, scripted, command line)
3. Hybrid

78

CRYSTALS

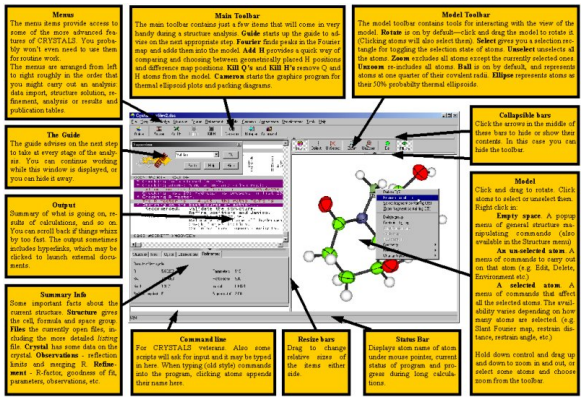
The GUI

The GUI 'writes' standard commands and data for the underlying program. This facilitates:

1. Data/command validation
2. User-readable error messages
3. GUI debugging
4. Replacement of the GUI by a different GUI

79

The CRYSTALS GUI



80

The CRYSTALS GUI

GUI Documentation

The html version of the manual is available from the GUI.

Some drop-down menus have links to dedicated html files explaining the operations.

All drop-down menus have a brief plain-text description.

A set of about 12 worked examples with notes are available from the GUI.

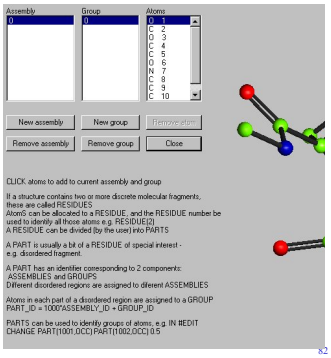
81

CRYSTALS

The CRYSTALS GUI

Part of the pane which enables users to divide structures into RESIDUES, PARTS, ASSEMBLIES and GROUPS

Useful for Z'>1 and disordered structures



82

Initialisation and tailoring

A real challenge for a programmer is not to force users into a particular pattern of working.

The *initialisation* and *tailoring* features of VMS brought the concept to the attention of all computer users, and is standard in CRYSTALS.

Almost all commercial software can now be customised

83

The CRYSTALS GUI

***MS Windows*, C++ and OpenGL (1997)**

By 1997 it was evident that the community was expecting a menu-driven GUI.

After experimenting with commercial GUI builders, Richard Cooper recognised that only a home-grown system would provide the close coupling with the underlying data structure needed for effective crystallography.

The actual GUI is constructed at run time by CRYSTALS SCRIPTS. Commands are passed down to the normal CRYSTALS command line processor and data is received back by the same mechanism. An asynchronous display model is updated whenever the CRYSTALS database is updated.

The graphical display is quite independent of CAMERON.

84

macHine valiDATion

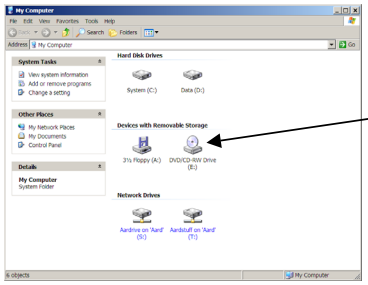
Word Processors have shown us both the *power and* **the *risks* of** machine validation of of routine tasks.

spelling & grammar checks, **automated** caPitalisation *etc.* filter out MANY errors, but it an exπerienced reader is still required to `review` the final documen T.

85

Screen real estate

The gratuitous dumbing-down of commercial user-interfaces poses serious problems for the developers of scientific interfaces, who may have to display a lot of material in a limited space



If a user needs this icon to indicate what a CD drive is, what chance do they have of making intelligent use of it?

86

CRYSTALS

Initialising Parameter Values –the bdb initialises all currently known values.

Additional data is usually input from a file.

Values associated by position, or by keyword, defaults inserted.

In the 1970's free format input was Rocket Science.

```
ATOM  c  1  0.5  1.0   .123  .234  .345  0.05
ATOM  c  1  occ=.5    x=.123  .234  .345
```

87

Richness & Complexity

Richness generally implies complexity.

e.g. Marquardt-type damping. The diagonal elements of the normal matrix are scaled up to increase the diagonal dominance.

SHELX: DAMP 0.7

s=(1+damp/1000) for all parameters.

CRYSTALS: LIMIT .01 C(3,X)

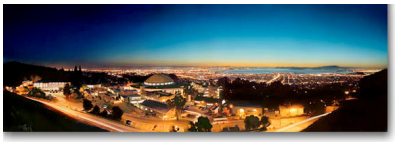
$s_{(3,x)}=1+ \Sigma(w\Delta^2)/((n-m)*0.01^2)$

LIMIT .02 Y

Shifts in individual parameters or parameter types may be restricted

88

Symmetry in crystallographic applications



Ralf W. Grosse-Kunstleve

Computational Crystallography Initiative
Lawrence Berkeley National Laboratory

IUCr 2008 Kyoto Crystallographic Computing School
"Sharing our knowledge": 18th to 23rd August

Background

- Share what I've learned about symmetry while working on:
  - SgInfo (1995) http://cci.lbl.gov/sginfo/
  - SgInfo2 (1999) never released
  - SgLite (2000) incl. in PyMOL
  - cctbx / Phenix (ongoing)

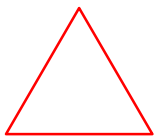
Crystal symmetry

- Crystal symmetry = Unit cell + Space group
- Int. Tab. Vol. A, chapter 8, sections 8.1.1 & 8.3.1:
  - Crystals are objects in the three-dimensional space in which we live.
  - A model for the mathematical treatment of this space is the so called point-space, which in crystallography is known as direct or crystal space.
  - In this space, the structures of finite real crystals are idealized as infinite perfect crystal structures.
  - Crystals are finite real objects in physical space which may be idealized by infinite periodic `crystal structures` in point space.
- Unit cell: model for the infinite periodicity
  - Strongest aspect of symmetry
  - Basis for definition crystallography
- Space group: model for symmetry "inside" the unit cell
  - Must be compatible with infinite periodicity
  - Decoration

Triangle experiment

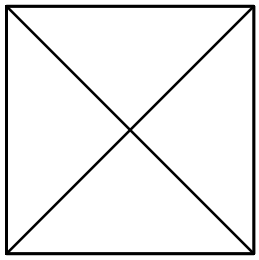
- Suggested by Michael Leyton
  - M. Leyton: A Generative Theory of Shape, Springer, 2001
- Take a piece of paper.
- Imagine you are sitting at your desk at home.
  - Please don't try to be funny.
- Draw a triangle.

Is this your triangle?

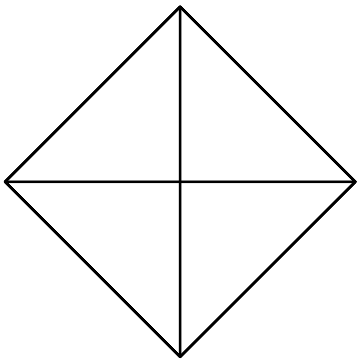


- Does your triangle have sides of equal length?
- Does your triangle point up?
- Why didn't you draw it this way?

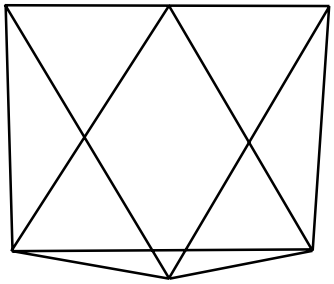
What do you see?



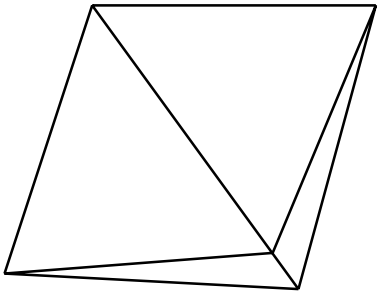
What do you see now?



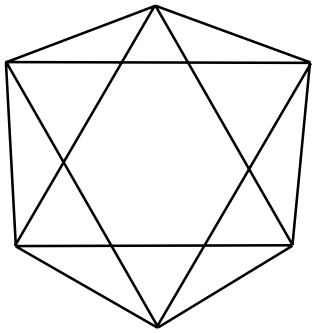
What do you see now?



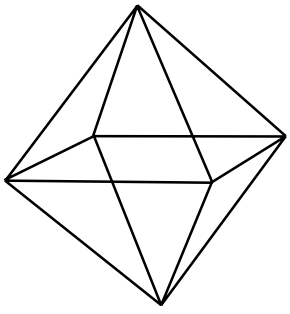
What do you see now?



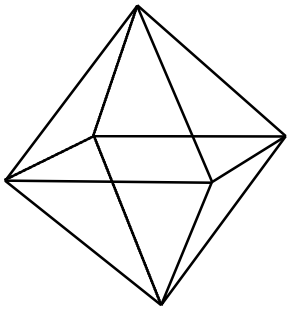
What do you see now?



Now you know!

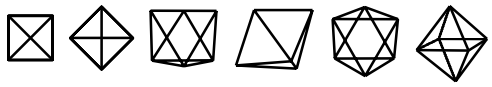


Now every child knows.



Mechanics of associations

- Viewing the same object "from different angles" can lead to very different impressions.



- More formally: the frame of reference leads to different associations.

- How does your brain work?
- You have seen something before.
- You have stored information about the things you have seen before.
- You need a "key" to recall the stored information.
- Your brain transforms the inputs into images it can use as keys.
- This brings back the things you have learned before.
  - Also known as "Ah!"

Crystal symmetry transformations

- Change-of-basis transformation.
  - Term adopted from
    - M.B. Boisen, Jr. & G.V. Gibbs; Mathematical Crystallography; Reviews in Mineralogy, Volume 15, Revised Edition; Mineralogical Society of America; Washington D.C. 1990
- Particularly simple class of transformation.
- For vectors: change-of-basis = linear transformation:

  $x' = M x$
- Where M is a "rotation-translation" matrix.
- Formation of change-of-basis matrices nicely explained in Boisen & Gibbs.
- Other objects have different change-of-basis laws:
  - symmetry operations
  - reciprocal-space vectors
  - tensors, e.g. anisotropic displacement parameters
- Change-of-basis laws nicely summarized in Giacovazzo et al. (1992), Fundamentals of Crystallography, Table 2.E.1.

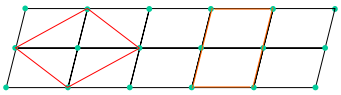
Table of change-of-basis laws

Table 2.E.1. Transformation relationships. In the table $\mathbf{M}$ is the matrix transforming $\mathbf{A} \equiv (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ into $\mathbf{A}' = (\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}')$. $\mathbf{G}$ and $\mathbf{G}'$ are the metric matrices of $\mathbf{A}$ and $\mathbf{A}'$ respectively, $\mathbf{G}^*$ and $\mathbf{G}'^*$ are the metric matrices of $\mathbf{A}^* \equiv (\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)$ and $\mathbf{A}'^* = (\boldsymbol{a}'^*, \boldsymbol{b}'^*, \boldsymbol{c}'^*)$ respectively. $\mathbf{C} \equiv (\mathbf{R}, \mathbf{T})$ is a symmetry operator ($\mathbf{R}$ is its rotational part, $\mathbf{T}$ its translational part): $\mathbf{C}$, $\mathbf{C}'$, $\mathbf{C}^*$, $\mathbf{C}'^*$ are symmetry operators defined in $\mathbf{A}$, $\mathbf{A}'$, $\mathbf{A}^*$, $\mathbf{A}'^*$ respectively. $\mathbf{Q}$ and $\mathbf{Q}^*$ are the quadratic forms of $\mathbf{A}$ and $\mathbf{A}^*$.

$\mathbf{A}' = \mathbf{MA}$	$\mathbf{A} = \mathbf{M}^{-1}\mathbf{A}'$
$\mathbf{X}' = (\mathbf{M})^{-1}\mathbf{X}$	$\mathbf{X} = \mathbf{MX}'$
$\mathbf{G}' = \mathbf{MGM}$	$\mathbf{G} = \mathbf{M}^{-1}\mathbf{G}'(\mathbf{M})^{-1}$
$\mathbf{Q}' = \mathbf{MQM}$	$\mathbf{Q} = \mathbf{M}^{-1}\mathbf{Q}'(\mathbf{M})^{-1}$
$\mathbf{R}' = (\mathbf{M})^{-1}\mathbf{RM}$, $\mathbf{T}' = (\mathbf{M})^{-1}\mathbf{T}$	$\mathbf{R} = \mathbf{MR}'(\mathbf{M})^{-1}$, $\mathbf{T} = \mathbf{MT}'$
$\mathbf{A}^* = \mathbf{G}^*\mathbf{A}$	$\mathbf{A} = \mathbf{GA}^*$
$\mathbf{X}^* = \mathbf{GX}$	$\mathbf{X} = \mathbf{G}^*\mathbf{X}^*$
$\mathbf{Q}^* = \mathbf{G}^*\mathbf{QG}^*$	$\mathbf{Q} = \mathbf{GQ}^*\mathbf{G}$
$\mathbf{R}^* = \mathbf{GRG}^* = (\mathbf{R})^{-1}$	$\mathbf{R} = (\mathbf{R}^*)^{-1}$
$\mathbf{A}'^* = (\mathbf{M})^{-1}\mathbf{A}^*$	$\mathbf{A}^* = \mathbf{MA}'^*$
$\mathbf{X}'^* = \mathbf{MX}^*$	$\mathbf{X}^* = \mathbf{M}^{-1}\mathbf{X}'^*$
$\mathbf{G}'^* = (\mathbf{M})^{-1}\mathbf{G}^*\mathbf{M}^{-1}$	$\mathbf{G}^* = \mathbf{MG}'^*\mathbf{M}$
$\mathbf{R}'^* = \mathbf{MR}^*\mathbf{M}^{-1}$, $\mathbf{T}'^* = \mathbf{MT}^*$	$\mathbf{R}^* = \mathbf{M}^{-1}\mathbf{R}'^*\mathbf{M}$, $\mathbf{T}^* = \mathbf{M}^{-1}\mathbf{T}'^*$
$\mathbf{Q}'^* = (\mathbf{M})^{-1}\mathbf{Q}^*\mathbf{M}^{-1}$	$\mathbf{Q}^* = \mathbf{MQ}'^*\mathbf{M}$

Unit cell transformations

- Fractionalization & orthogonalization matrices
  - Cartesian (orthogonal) frame: distance calculations, visualization
  - Fractional frame: compatible with symmetry operations in their "most intuitive frame"
  - Note: many possible orthogonalization conventions
  - Different applications may use different conventions!
  - Most macromolecular applications use the "PDB convention"
- Primitive setting vs. centered setting
  - Applicable to all space groups!
  - Any space group that is primitive in the conventional frame can be transformed to a centered setting.
  - Any space group that is centered in the conventional frame can be transformed to a primitive setting.



Unit cell transformations (cont.)

- Niggli-reduced cell
  - Basis vectors are shortest vectors of the lattice.
  - Rules for unique selection of angles:
    - Niggli, P. (1928). Handbuch der Experimentalphysik, Vol. 7, part 1, pp. 108-176.
    - Gruber, B. (1973). Acta Cryst. A29, 433-440.
    - Krivy, I. & Gruber, B. (1976). Acta Cryst. A32, 297-298.
    - Int. Tab. Vol. A, chapter 9.
  - Rules lead to "canonical" frame of reference
    - 28 Niggli-reduced cell types
  - Lookup table 9.3.1 in Int. Tab. Vol. A.
  - Example of transformation to "key" (one of the 28 types) that can be used to retrieve prior knowledge.

```
from cctbx.uctbx import unit_cell
given_cell = unit_cell((
  83.5591, 66.8846, 80.5753,
  44.7561, 18.3367, 26.4495))
print given_cell.niggli_cell()
(13.0023, 17.0009, 21.0007, 93.0049, 93.0084, 97.9861)
```

Space group transformations

- Primitive setting vs. centered setting
  - Applicable to all space groups!
  - Any space group that is primitive in the conventional frame can be transformed to a centered setting and vice versa.
  - e.g. primitive setting of Fm3m or C-centered setting of P4/mmm
- Centro-symmetric space groups: origin choices
  - e.g. Pnnn origin choice 1 or 2
- Monoclinic space groups: unique axes & cell choices
  - Int. Tab. Vol. A, Table 4.3.1: 13 monoclinic s.g. types but 105 settings (*8.1).
  - e.g. C1m1 (cell choice 1) or A1m1 (cell choice 2) or A11m (unique c)
- Orthorhombic space groups: permutation of axes
  - Int. Tab. Vol. A, Table 4.3.1: 59 orthorhombic s.g. types but 227 settings (*3.5).
  - e.g. $P222_1$ or $P22_12$

Space group transformations (cont.)

- Both monoclinic and orthorhombic:
  - Preferred symmetry operations vs. preferred unit cell parameters
  - Complex rules
  - Authoritative definition?
- Rhombohedral space groups
  - Centered setting: R 3 (hexagonal axes)
  - Primitive setting: R 3 (rhombohedral axes)

Determination of space group type

- Given a group of symmetry operations:
- What is the space group type?
  - I.e. what is space group number?
- What is the change-of-basis matrix to a reference setting?
  - Another example of transforming inputs to a "key" that can be used to retrieve prior knowledge.
- Outline of algorithm in Acta Cryst. (1999). A55, 383-395:
  - (a) change-of-basis: to primitive setting
  - (b) determination of point group type via classification of symmetry operations and counting (e.g. "how many twofold axes")
  - (c) change-of-basis: to "standard" setting for given point group type
  - (d) change-of-basis: special case adjustments for certain combinations of Laue groups and centering types;
    monoclinic and orthorhombic space groups: trial loop over change-of-basis matrices corresponding to cell choices and axis permutations
  - (e) Determination of origin shift. (This was the hardest part even though the change-of-basis is just a translation.)

Determination of space group type

- Development of this algorithm involved implementation of many tools that turned out to be useful in various other contexts.
- Outstanding examples:
  - change-of-basis algorithms for symmetry operation, space group, unit cell
  - Row-echelon reduction for the solution of linear equations
    - similar to Gaussian elimination, but also works for singular matrices

Determination of lattice symmetry

- Example that uses the previous algorithm
- The starting point are unit cell parameters, and optionally a centering symbol (A, B, C, I, R, F) or arbitrary lattice translation vectors.
- Question: what is the highest symmetry of the lattice?
  - Equivalently: what is the Bravais type?
  - Another example of transforming inputs to a "key" for the retrieval of prior knowledge.
- Additional input:
  - Tolerance for deviation from ideal symmetry.

Determination of lattice symmetry

- Outline of the algorithm:
  - Change-of-basis 1: to primitive cell
  - Change-of-basis 2: to reduced cell
    - "Minimum-lengths cell", Acta Cryst. (2004). A60, 1-6.
  - Search for twofold axes
    - Le Page (1982). J. Appl. Cryst. 15, 255-259.
    - Lebedev, Vagin & Murshudov, Acta Cryst. (2006). D62, 83-95.
    - CCP4 newsletter No. 44, Summer 2006.
  - Sort twofold axes using "Le Page delta", smallest to largest:
    - angle between original axis direction and symmetry-equivalent vector
  - Group multiplication, adding twofolds in sorted order to the group
    - Stop if adding a twofold generates a group of infinite order
    - Result is a point group, potentially in an unusual setting
  - Change-of-basis 3: to reference setting
    - Using the algorithm for the determination of the space group type
    - Result is a change-of-basis matrix to the reference setting of the point group with the highest symmetry compatible with the input unit cell parameters
  - Determination of subgroups of highest symmetry

Determination of subgroups

- Boisen & Gibbs (and many other texts):
  - Three symmetry operations are sufficient to generate any crystallographic space group
  - The same is true for point groups (since they are simply a subset of the space groups)
  - One of the three symmetry operations is the center of inversion
  - Two symmetry operations are sufficient to generate any acentric (not centro-symmetric) space group or point group
- These insights lead to a very simple algorithm:

```
for sym_op1 in highest_symmetry:
  for sym_op2 in highest_symmetry:
    subgroup = sgtbx.space_group()
    space_group.expand_smx(sym_op1)
    space_group.expand_smx(sym_op2)
```

- The result is a space group in a potentially very unusual setting.
- Change-of-basis: to reference setting
  - again via algorithm for the determination of the space group type
- Implementation: iotbx/command_line/lattice_symmetry.py

iotbx.lattice_symmetry example

```
Symmetry in minimum-lengths cell: F m -3 m (-x+y+z,x-y+z,x+y-z) (No. 225)
      Input minimum-lengths cell: (8.3359, 8.44603, 8.52071, 61.1613, 61.2992, 61.0215)
           Symmetry-adapted cell: (8.50892, 8.50892, 8.50892, 60, 60, 60)
            Conventional setting: F m -3 m (No. 225)
                       Unit cell: (12.0334, 12.0334, 12.0334, 90, 90, 90)
                 Change of basis: z,-x,-y
                         Inverse: -y,-z,x
      Maximal angular difference: 2.236 degrees

Symmetry in minimum-lengths cell: I 4/m m m (y-z,-x+z,x+z) (No. 139)
      Input minimum-lengths cell: (8.3359, 8.44603, 8.52071, 61.1613, 61.2992, 61.0215)
           Symmetry-adapted cell: (8.48528, 8.52071, 8.52071, 59.7251, 60.1374, 60.1374)
            Conventional setting: I 4/m m m (No. 139)
                       Unit cell: (8.48528, 8.48528, 12.1, 90, 90, 90)
                 Change of basis: -x+y,-x-y,z
                         Inverse: -1/2*x-1/2*y,1/2*x-1/2*y,z
      Maximal angular difference: 2.236 degrees

...

Symmetry in minimum-lengths cell: P -1 (No. 2)
      Input minimum-lengths cell: (8.3359, 8.44603, 8.52071, 61.1613, 61.2992, 61.0215)
           Symmetry-adapted cell: (8.3359, 8.44603, 8.52071, 61.1613, 61.2992, 61.0215)
            Conventional setting: P -1 (No. 2)
                       Unit cell: (8.3359, 8.44603, 8.52071, 61.1613, 61.2992, 61.0215)
                 Change of basis: -x-y-z,x-y+z,-x+y+z
                         Inverse: -1/2*x-1/2*z,-1/2*x-1/2*y,1/2*y+1/2*z
      Maximal angular difference: 0.000 degrees
```

More: reference setting → prior knowledge

- Simple reporting: universal Hermann-Mauguin symbol
  - `Symmetry in minimum-lengths cell: I 4/m m m (y-z,-x+z,x+z)`
  - Acta Cryst. (2008). D64, 99–107.
- Automatic derivation of subgroup/super-group relations
  - phenix.explore_metric_symmetry (Peter Zwart)
  - CCP4 newsletter No. 44, Summer 2006.
- Euclidean or affine normalizers ("Cheshire symmetry")
  - Int. Tab. Vol. A, chapter 15
  - Acta Cryst. (2003). D59, 1974–1977.
- Wyckoff positions
  - Acta Cryst. (2002). A58, 60–65.
- Direct-space asymmetric units
  - Newsletter of the IUCr Commission on Crystallographic Computing 2003, 2:10-16.
  - http://cci.lbl.gov/asu_gallery/
- Reciprocal-space asymmetric units

Other symmetry algorithms

- Twin-law enumeration and re-indexing laws from first principles
  - phenix.reflection_statistics
  - Newsletter of the IUCr Commission on Crystallographic Computing 2005, 5:69-91.
- Determination of site-symmetry
  - incl. constraints for coordinates and displacement parameters (U, B)
  - enumeration of symmetry-equivalent sites
  - Acta Cryst. (2002). A58, 60–65.
- Operations on Miller indices
  - systematic absences
  - phase restrictions
  - phases of symmetry-equivalent reflections
- Structure-seminvariant vectors and moduli from first principles
  - Acta Cryst. (1999). A55, 383-395.
- Handling of various search symmetries
  - to support structure solution
  - Acta Cryst. (2003). D59, 1974–1977.

Open Source

- Many more algorithms not mainly concerned with unit cells and space groups.
- Python code: 184k lines
  - a significant fraction of the code are unit tests (ca. 1/3 - 1/2)
- C++ code: 158k lines
- Yet: very easy to install, virtually no external dependencies
  - except Operating system and C/C++ compiler
  - C/C++ compiler free on all major platforms
- All algorithms mentioned in this talk are open source.

Conclusion

- Main theme of this talk: change-of-basis transformations
- The most important symmetry algorithms in the cctbx determine transformations to some type of reference:
  - Space group type
  - Point group type
  - Bravais type (lattice symmetry)
  - Niggli-reduced cell
- Change-of-basis matrices can be applied to all important cctbx types:
  - Unit cell
  - Space group
  - Miller arrays (re-indexing of reflection data)
  - Crystal structures (coordinates)
- ~~cctbx~~
- "Change-of-basis toolbox" - cbtbx

Acknowledgments

- Paul Adams
- Peter Zwart
- All Phenix developers (phenix-online.org)
- Syd Hall
- Lachlan Cranswick
- National Institute of Health (NIH)
- Phenix industrial consortium members
- Schweizerischer Nationalfonds (SNF)

Introduction to Bayesian methods in macromolecular crystallography

Tom Terwilliger
Los Alamos National Laboratory

Why use a Bayesian approach?

- **We often know how are measurements are related to our model...**
- **The Bayesian approach gives us the probability of our model once we have made a measurement**
- **It is useful for dealing with cases where there are errors (uncertainties) in the model specification (missing parts of model)**

Introduction to Bayesian methods in macromolecular crystallography

Basics of the Bayesian approach

- **Working with probability distributions**
- **Prior probability distributions**
- **How do we go from distributions to the value of "x"?**
- **Bayesian view of making measurements**
- **Example: from "400 counts" to a probability distribution for the rate**
- **Bayes' rule**
- **Applying Bayes' rule**
- **Visualizing Bayes' rule**

Marginalization: Nuisance variables and models for errors

- **How marginalization works**
- **Repeated measurements with systematic error**

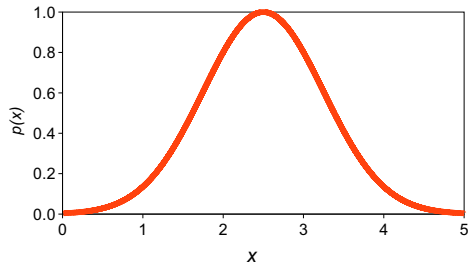
Applying the Bayesian approach to any measurement problem

Basics of the Bayesian approach

Working with probability distributions

Representing what we know about x as a probability distribution

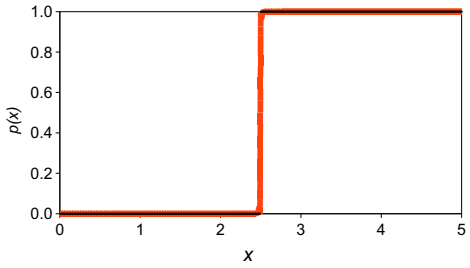
p(x) tells us the relative probability of different values of x



p(x) does not tell us what x is...
...just the relative probability of each value of x

Prior probability distributions

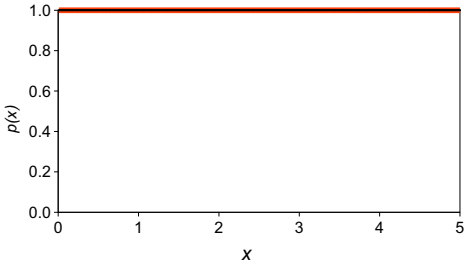
What we know before making measurements



I am sure *x* is at least 2.5

Prior probability distributions

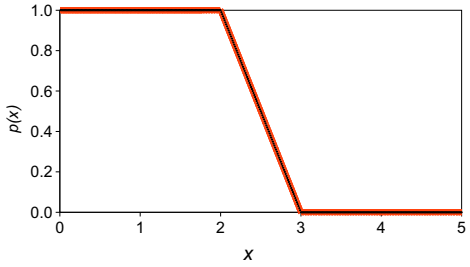
What we know before making measurements



All values of *x* are equally probable

Prior probability distributions

What we know before making measurements



x is less than about 2 or 3

Working with probability distributions

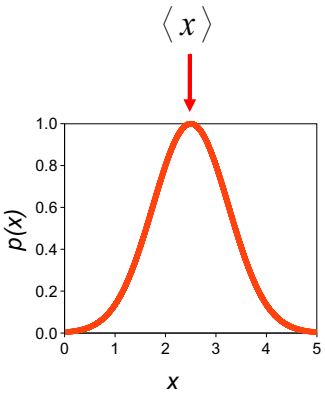
What is the "value" of *x* ?

We don't know exactly what "x" is...

but we can calculate a weighted estimate:

$$\langle x \rangle = A \int x\, p(x)\, dx$$

Weight each value of x — ***by its relative probability p(x)***



$$A = 1 / \int p(x)\, dx$$ ← ***A is normalization factor***

A Bayesian view of making measurements

A crystal is in diffracting position for a reflection
The beam and crystal are stable...

We measure 400 photons hitting the corresponding pixels in our detector in 1 second

What is the probability that the rate of photons hitting these pixels is actually less than 385 photons/sec?

A Bayesian view of making measurements

A crystal is in diffracting position for a reflection
The beam and crystal are stable...

We measure 400 photons hitting the corresponding pixels in our detector in 1 second : N_{obs} = 400

A good guess for the actual rate k of photons hitting these pixels is 400:

k ~ 400

What is the probability that k is actually < 385 photons/sec?

What is p(k<385 | N_{obs} = 400)

A Bayesian view of making measurements

Start with prior knowledge about which values of k are probable: $p_o(k)$

Make measurement N_{obs}

For each possible value of parameter k (385...400...)

Calculate probability of observing N_{obs} if k were correct: $p(N_{obs} \mid k)$

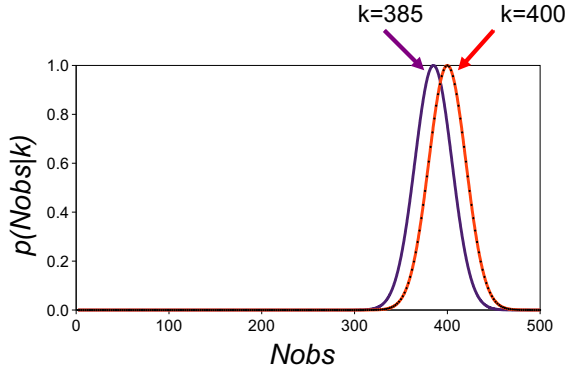
Use Bayes' rule to get p(k) from $p_o(k)$, N_{obs} and $p(N_{obs}|k)$:

$$p(k) \propto p_o(k)\, p(N_{obs} \mid k)$$

A Bayesian view of making measurements

What is the probability that we would measure N_{obs} counts if the true rate were k?

$$p(N_{obs} \mid k)$$



Bayes' rule

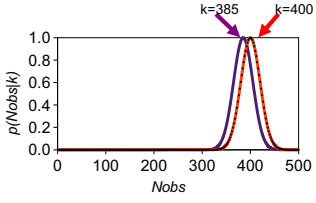
$$p(k) \propto p_o(k)\, p(N_{obs} \mid k)$$

The probability that k is correct is proportional to...

the probability of k from our prior knowledge

multiplied by...

the probability that we would measure N_{obs} counts if the true rate were k



Bayes' rule

$$p(k) \propto p_o(k)\, p(N_{obs}|k)$$

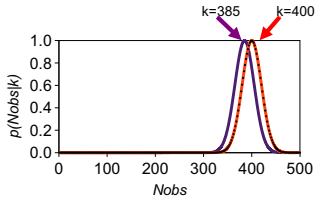
Prior — Likelihood

The probability that k is correct is proportional to...

the probability of k from our prior knowledge (prior)

multiplied by...

the probability that we would measure N_{obs} counts if the true rate were k **(likelihood)**



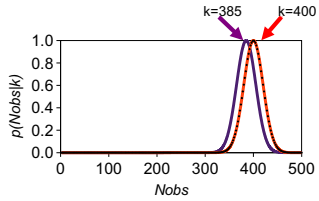
Application of Bayes' rule

$$p(k) \propto p_o(k)\, p(N_{obs}|k)$$

No prior knowledge: $p_o(k) = 1$

Poisson dist. for N_{obs} (large k)

$$p(N_{obs}|k) \propto e^{-[N_{obs}-k]^2/(2k)}$$



Application of Bayes' rule

Probability distribution for k given our measurement N_{obs} =400:

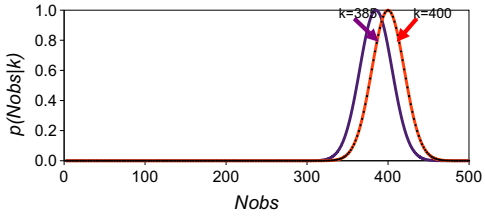
$$p(k) \propto e^{-[N_{obs}-k]^2/(2k)}$$

Probability that k < 385:

$$p(k<385) = A\int_{-\infty}^{385} p(k)\,dk$$

p = 22%

$$A = 1/\int_{-\infty}^{\infty} p(k)\,dk$$



Visualizing Bayes' rule

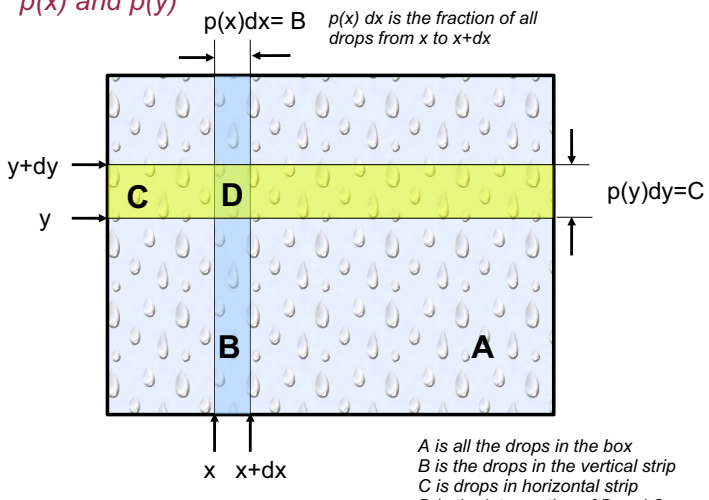
$$p(x|y_{obs}) \propto p_o(x)\, p(y_{obs}|x)$$

Where does Bayes' rule come from?

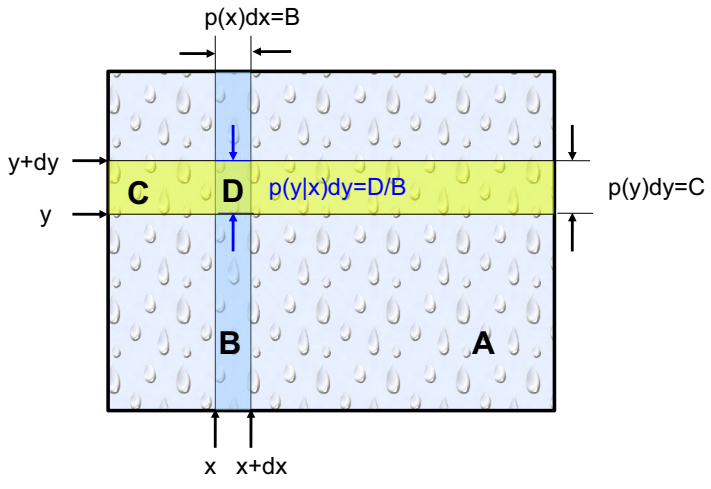
Using a graphical view to show how p(x|y) is related to p(y|x)

Visualizing Bayes' rule: *p(x) and p(y)*

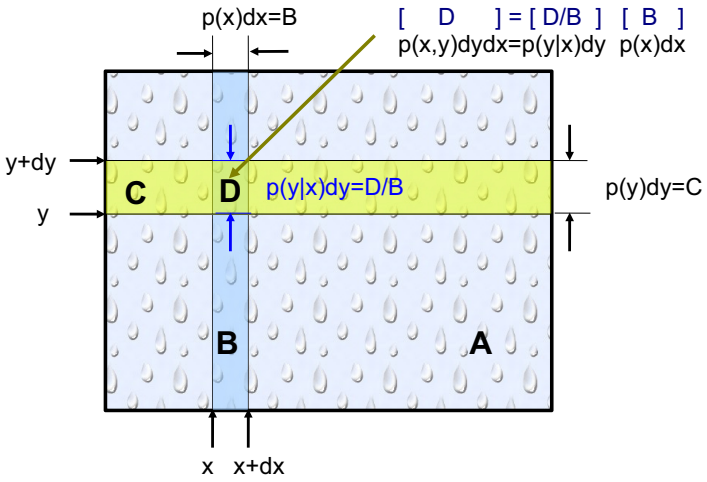
$$p(x|y_{obs}) \propto p_o(x)\, p(y_{obs}|x)$$



Visualizing Bayes' rule: *p(y|x) and p(x|y)*

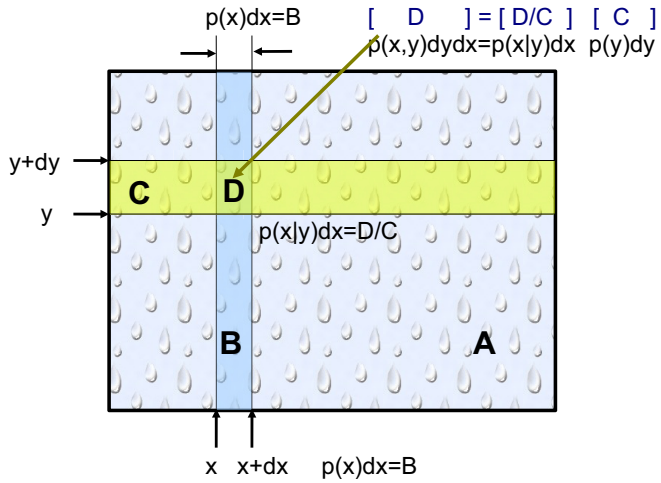


Visualizing Bayes' rule: *p(x,y)*



p(x,y)dxdy is the fraction of all drops inside the box from x to x+dx and y to y+dy

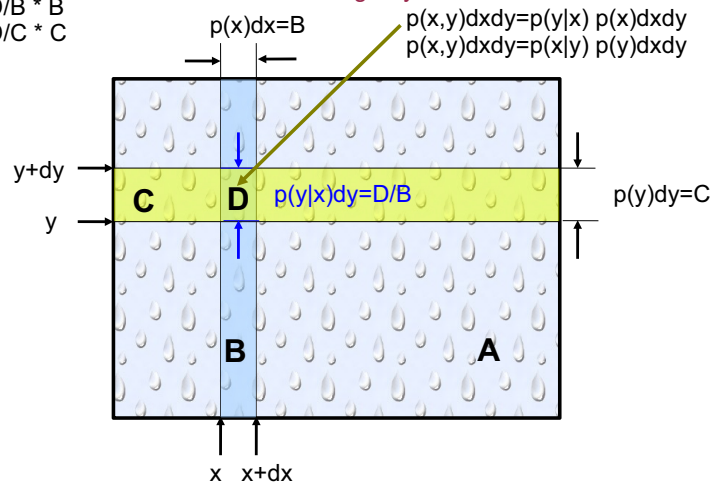
Visualizing Bayes' rule: *p(x,y)*



p(x,y)dxdy is the fraction of all drops inside the box from x to x+dx and y to y+dy

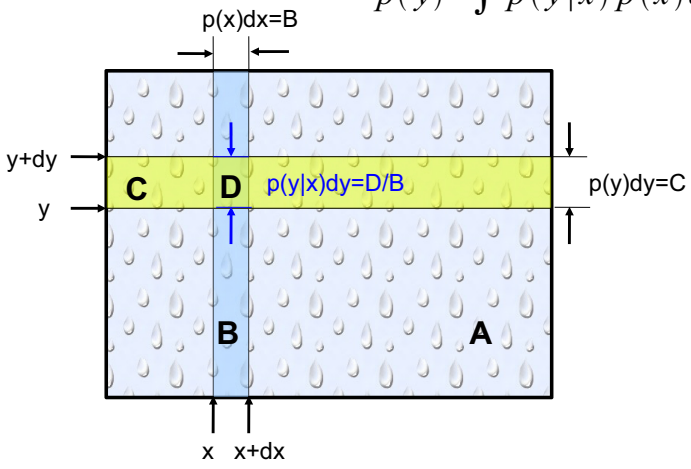
Visualizing Bayes' rule

D = D/B * B
D = D/C * C



An identity we will need now and later....

$$p(y)=\int p(y|x)\,p(x)\,dx$$



Visualizing Bayes' rule

p(x,y) written two ways $p(x|y)\,p(y)=p(y|x)\,p(x)$

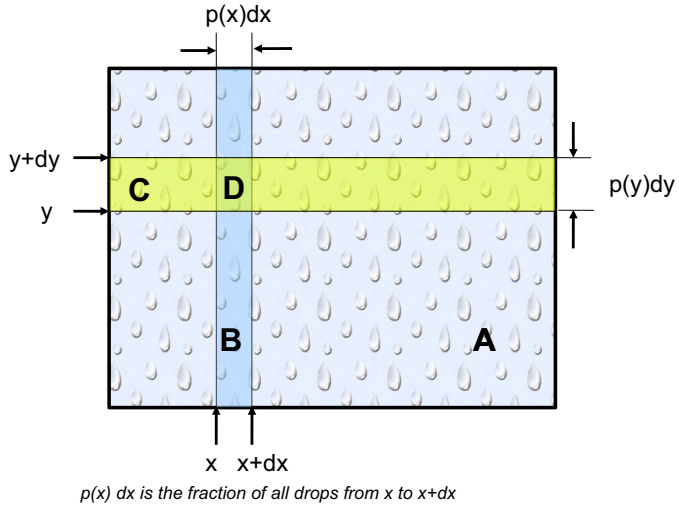
rearrangement... $p(x|y)=p(y|x)\,p(x)/p(y)$

An identity $p(y)=\int p(y|x)\,p(x)\,dx$

Substitution...Bayes' rule:

$$p(x|y)=p(y|x)\,p(x)/\int p(y|x)\,p(x)\,dx$$

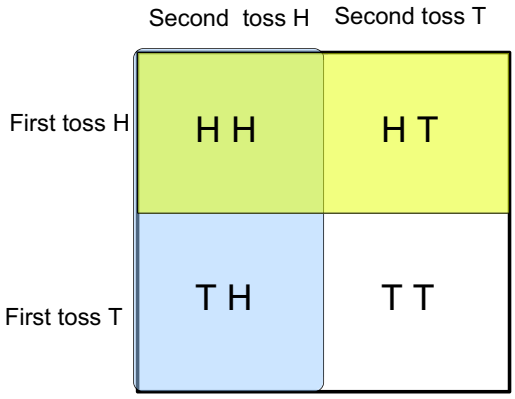
Bayes' rule as a systematic way to evaluate truth-tables



p(x) dx is the fraction of all drops from x to x+dx

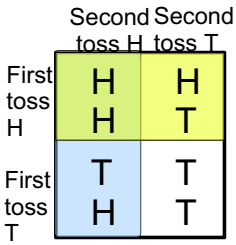
Bayes' rule as a systematic way to evaluate truth-tables

We toss a coin twice and get at least one "heads".
What is the probability that the first toss was a "head?"



Bayes' rule as a systematic way to evaluate truth-tables

We toss a coin twice and get at least one "heads".
What is the probability that the first toss was a "head?"



FS=head on first or second toss

H= heads first toss T= tails first toss

Bayes' rule:

$p(H)=A\, p_o(H)\, p(FS|H)$
$A = 1 /[\, p_o(H)\, p(FS|H) + p_o(T)\, p(FS|T)\,]$

$p_o(H)=1/2$
$p(FS|H)= 1$
$p(FS|T)=1/2$

$A = 1/[1/2 + 1/2 * 1/2] = 4/3$
$p(H)= 4/3 * 1/2 = 2/3$

Quick Review of Bayes' rule

$$p(x|y_{obs}) \propto p_o(x)\, p(y_{obs}|x)$$

$p(x|y_{obs})$ Probability of x given our observations

$p_o(x)$ What we knew beforehand about x

$p(y_{obs}|x)$ Probability of measuring these observations if x were the correct value

Marginalization

What if the observations depend on *z* as well as *x* ?
(Maybe z is model error)

$p(y_{obs}|x)$ What we want to use in Bayes' rule

$$p(y_{obs}|x) = \int p(y_{obs}|x,z)\, p(z)\, dz$$

"Integrate over the nuisance variable z, weighting by p(z)"

Marginalization

y_{obs}=observations

$$p(y_{obs}) = \int p(y_{obs}|z)\, p(z)\, dz$$ Identity we saw earlier

$$p(y_{obs}|x) = \int p(y_{obs}|z,x)\, p(z|x)\, dz$$ The whole equation can be for a particular value of x

$$p(y_{obs}|x) = \int p(y_{obs}|z,x)\, p(z)\, dz$$ If z does not depend on *x*, *p(z)=p(z|x)*

"Integrate over the nuisance variable z, weighting by p(z)"

Marginalization with Bayes' rule

We want to get *p(x)* using $p(y_{obs}|x)$ in Bayes' rule...

y_{obs} is an experimental measurement of *y*

$$p(y_{obs}|y) \propto e^{-(y_{obs}-y)^2/2\sigma^2}$$

y depends on x and z (perhaps z is model error)

$$y = y(z,x)$$

...then we can integrate over *z* to get $p(y_{obs}|x)$:

$$p(y_{obs}|x) = \int p(y_{obs}|y(z,x))\, p(z)\, dz$$

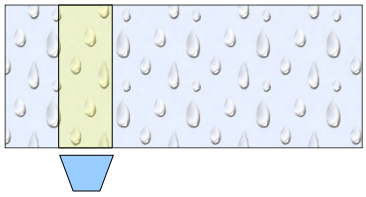
Repeated measurements with systematic error

We want to know on average how many drops D_{avg} of rain hit a surface per 100 cm^2 per minute.

The rain does not fall uniformly: $D(x)=D_{avg}+E(x)$ where the SD of $E(x)$ is e. However we only sample one place

We count the drops N falling in 1 minute into a fixed bucket with top area of 100 cm^2 m times (N_1, N_2...) with a mean of n.

What is the weighted mean estimate <D_{avg}>? What is the uncertainty in <D_{avg}>?



Repeated measurements with systematic error

We want to get $p(D_{avg})$ using $p(N_{obs}|D_{avg})$ in Bayes' rule...but the rate into our bucket D depends on D_{avg} and E:

$$D = D_{avg} + E$$

$$p(E) \propto e^{-E^2/2e^2}$$

N_{obs} is the number of drops we count with SD of $n^{1/2}$:

$$p(N_{obs}|D_{avg},E) \propto e^{-(N_{obs}-(D_{avg}+E))^2/2s^2}$$

Including all m measurements N_1, N_2...

$$p(N_1,N_2...|D_{avg},E) \propto e^{-\sum_i (N_i-(D_{avg}+E))^2/2s^2}$$

From previous slide

$$p(N_1,N_2...|D_{avg},E) \propto e^{-\sum_i (N_i-(D_{avg}+E))^2/2s^2}$$

$$p(E) \propto e^{-E^2/2e^2}$$

We have $p(N_1,N_2...|D_{avg},E)$. We want $p(N_1,N_2...|D_{avg})$. Integrate over the nuisance variable E:

$$p(N_1,N_2...|D_{avg}) = \int p(N_1,N_2...|D_{avg},E)\,p(E)\,dE$$

Yielding (where n is the mean value of N: <N_1,N_2...>)

$$p(N_1,N_2...|D_{avg}) \propto e^{-(D_{avg}-n)^2/2(e^2+s^2/m)}$$

Now we have $p(N_1,N_2...|D_{avg})$ and we are ready to apply Bayes' rule

We have the probability of the observations given D_{avg},

$$p(N_1,N_2...|D_{avg}) \propto e^{-(D_{avg}-n)^2/2(e^2+s^2/m)}$$

Bayes' rule gives us the probability of D_{avg} given the observations:

$$p(D_{avg}|N_1,N_2...) \propto p_o(D_{avg})\, e^{-(D_{avg}-n)^2/2(e^2+s^2/m)}$$

If the prior $p_o(D_{avg})$ is uniform:

$$p(D_{avg}|N_1,N_2...) \propto e^{-(D_{avg}-n)^2/2(e^2+s^2/m)}$$

$$\langle D_{avg}\rangle = n = \langle N\rangle \qquad \sigma^2 = e^2 + s^2/m$$

Summary: How to apply a Bayesian analysis to any measurement problem

1. Write down what you really want to know: $p(D_{avg})$

2. Write down prior knowledge: $p_o(D_{avg})=1$

3. Write down how the true value of the thing you are measuring depends on what you really want to know and any other variables: $D=D_{avg}+E$

4. Write down probability distributions for errors in measurement and for the variables you don't know: $p(N_{obs}|D)$ and $p(E)$

How to apply a Bayesian analysis of any measurement problem

5. Use 3&4 to write probability distribution for measurements given values of what you want to know and of nuisance variables: $p(N_1, N_2...|D_{avg},E)$

6. Integrate over the nuisance variables (E), weighted by their probability distributions p(E) to get probability of measurements given what you want to know: $p(N_1, N_2...|D_{avg})$

7. Apply Bayes' rule to get the probability distribution for what you want to know, given the measurements: $p(D_{avg}|N_1, N_2...) = p_o(D_{avg})\, p(N_1, N_2...|D_{avg})$

Applications of the Bayesian approach in macromolecular crystallography

- Correlated MIR phasing (errors due to non-isomorphism are correlated among heavy-atom derivatives)
- Correlated MAD phasing (errors in heavy-atom model are correlated among wavelengths)
- Bayesian difference refinement (errors in model of macromolecular structure correlated between two structures)
- Macromolecular refinement (phase unknown and model errors present)

Tutorials

- **Working through simple Bayesian exercises from handout in a group**
- **PHENIX demo and discussion**
- **Density modification and model-building theory and discussion**
- **Discussion of individual challenging examples and questions from students**

Exercise 1

1. Draw a probability distribution that means "I know that x is between 0 and 1." Draw another that means "I know that x is within 0.01 of being an integer."

Exercise 2

2a. A measurement system consists of a biased ruler that systematically reads 1 mm too high and that can be read with a precision of +/-0.5 mm. Suppose we measure the diameter of a pencil that is actually 2.0 mm across. Draw a probability distribution for these measurements.

2b. The Gaussian function $y = \exp -[(x-x_o)^2 /2s^2]$ has a maximum at x_o and a SD of s. Write an equation p(obs|D) for the probability distribution you have drawn in 2a.

Exercise 3

Consider the example in Exercise 2 (a ruler that always reads 1 mm too high and has an uncertainty in measurement of 0.5 mm). We now have a measurement d=3.0mm

Suppose we know in advance that the diameter of the pencil is at least 1.8 mm.

a. Draw this a priori probability distribution

b. Use Bayes' rule to write an expression for the probability distribution of the diameter D given a measurement d=3.0 mm made with our biased ruler.

c. Draw this probability distribution for D. Approximately what is the mean value of D?

Exercise 4 *Applying the Bayesian approach to a measurement problem without nuisance variables*

You make 10 measurements W_j of the weight of a ball bearing. You think your scale is unbiased and has a Gaussian distribution of errors with SD of s. You are willing to believe any value of the weight.

a. What is your probability distribution for the weight after making these 10 measurements (go through steps 1-7 in "How to apply a Bayesian analysis to any measurement problem, with no nuisance variables)? What is your best estimate of the weight <x>?

b. Now suppose you are absolutely certain that this ball bearing is heavier than a NBS calibrated standard with weight $M_o g$. Write down your a priori probability distribution for W. Now incorporate this into your expression for the probability of W given your measurements using Bayes' rule. How would you have deal with this information if you did not use a Bayesian approach?

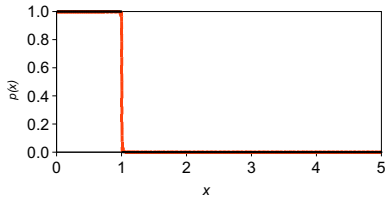
Exercise 5 *Applying the Bayesian approach to a measurement problem with nuisance variables*

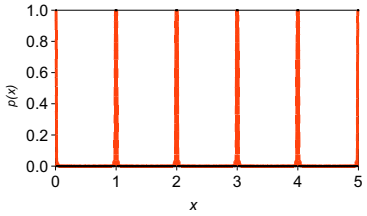
Suppose you expect that the scale used in the previous exercise as biased, reading systematically too low or too high. You don't know which, but you think this bias has a Gaussian distribution with a standard deviation of D. You have no prior knowledge about the weight.

Now what is your probability distribution for the weight after making the same 10 measurements made in the previous exercise? Don't bother to evaluate the integrals, just write them down.

Answer to Exercise 1

1. Draw a probability distribution that means "I know that x is between 0 and 1." Draw another that means "I know that x is within 0.01 of being an integer."

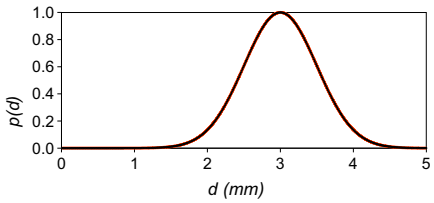




Answer to Exercise 2

2a. A measurement system consists of a biased ruler that systematically reads 1 mm too high and that can be read with a precision of +/-0.5 mm. Suppose we measure the diameter of a pencil that is actually 2.0 mm across. Draw a probability distribution for these measurements.

2b. The Gaussian function $y = \exp -[(x-x_o)^2 /2s^2]$ has a maximum at x_o and a SD of s. Write an equation p(obs|D) for the probability distribution you have drawn in 2a.



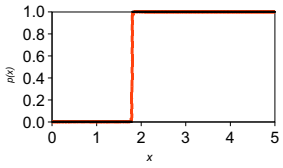
$$p(d_{obs}|D) \propto e^{-(d_{obs}-(D+1.0))^2/2\sigma^2}$$

Answer to Exercise 3

Consider the example in Exercise 2 (a ruler that always reads 1 mm too high and has an uncertainty in measurement of 0.5 mm). We now have a measurement d=3.0mm

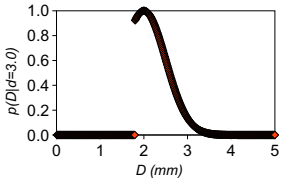
Suppose we know in advance that the diameter of the pencil is greater than 1.8 mm.

a. Draw this a priori probability distribution



b. Use Bayes' rule to write an expression for the probability distribution of the diameter D given a measurement d=3.0 mm made with our biased ruler.

$$p(D|d_{obs}) \propto p_o(D)\, e^{-(d_{obs}-(D+1.0))^2}$$



c. Draw this probability distribution for D. Approximately what is the mean value of D?

Answer to Exercise 4

You make 10 measurements W_j of the weight of a ball bearing. You think your scale is unbiased and has a Gaussian distribution of errors with SD of s. You are willing to believe any value of the weight.

a. What is your probability distribution for the weight after making these 10 measurements? (go through steps 1-7 in "How to apply a Bayesian analysis to any measurement problem, with no nuisance variables) What is your best estimate of the weight <x>?

1. Write down what you really want to know: p(W)

2. Write down prior knowledge: $p_o(W)=1$

3. Write down how the true value of the thing you are measuring depends on what you really want to know and any other variables: W=W (no nuisance variables)

Answer to Exercise 4, continuation 1

4. Write down probability distributions for errors in measurement and for the variables you don't know:

$$p(W_{obs}|W) \propto e^{-(W_{obs}-W)^2/2\sigma^2}$$

5. Use 3&4 to write probability distribution for measurements given values of what you want to know and of nuisance variables:

$$p(W_1, W_2 \ldots|W) \propto e^{-\sum_i (W_i-W)^2/2\sigma^2}$$

6. Integrate over the nuisance variables (E)... NONE

Answer to Exercise 4, continuation 2

7. Apply Bayes' rule to get the probability distribution for what you want to know, given the measurements:

$$p(W|W_1, W_2 \ldots) \propto e^{-\sum_i (W_i-W)^2/2\sigma^2}$$

What is your best estimate of the weight <x>?

Best estimate of weight is the weighted mean value

$$\langle W \rangle = \int W\, p(W|W_1, W_2 \ldots)\, dW$$

Answer to Exercise 4, continuation 3

b. Now suppose you are absolutely certain that this ball bearing is heavier than a NBS calibrated standard with weight $M_o g$. Write down your a priori probability distribution for W. Now incorporate this into your expression for the probability of W given your measurements using Bayes' rule. How would you have deal with this information if you did not use a Bayesian approach?

1. Write down what you really want to know: p(W)

2. Write down prior knowledge: $p_o(W)$= {0, W<$M_o g$; 1, W> $M_o g$}

3. Write down how the true value of the thing you are measuring depends on what you really want to know and any other variables: W=W (no nuisance variables)

Answer to Exercise 4, continuation 4

4. Write down probability distributions for errors in measurement and for the variables you don't know:

$$p(W_{obs}|W) \propto e^{-(W_{obs}-W)^2/2\sigma^2}$$

5. Use 3&4 to write probability distribution for measurements given values of what you want to know and of nuisance variables:

$$p(W_1, W_2 \ldots|W) \propto e^{-\sum_i (W_i-W)^2/2\sigma^2}$$

6. Integrate over the nuisance variables (E)... NONE

Answer to Exercise 4, continuation 5

7. Apply Bayes' rule to get the probability distribution for what you want to know, given the measurements:

$$p(W|W_1, W_2 \ldots) \propto p_o(W)\, e^{-\sum_i (W_i-W)^2/2\sigma^2}$$

What is your best estimate of the weight <x>?

Best estimate of weight is the weighted mean value. The prior is zero below $M_o g$ so we integrate from $M_o g$ to infinity

$$\langle W \rangle = \int_{M_o g}^{\infty} W\, p(W|W_1, W_2 \ldots)\, dW$$

Answer to Exercise 5

Suppose you expect that the scale used in the previous exercise as biased, reading systematically too low or too high. You don't know which, but you think this bias has a Gaussian distribution with a standard deviation of D. You have no prior knowledge about the weight.

Now what is your probability distribution for the weight after making the same 10 measurements made in the previous exercise? Don't bother to evaluate the integrals, just write them down.

1. Write down what you really want to know: p(W)

2. Write down prior knowledge: $p_o(W)$=1

3. Write down how the true value of the thing you are measuring depends on what you really want to know and any other variables: W=W+E

Answer to Exercise 5, continuation 1

4. Write down probability distributions for errors in measurement and for the variables you don't know:

$$p(W_{obs}|W, E) \propto e^{-(W_{obs}-(W+E))^2/2\sigma^2}$$

$$p(E) \propto e^{-E^2/2D^2}$$

5. Use 3&4 to write probability distribution for measurements given values of what you want to know and of nuisance variables:

$$p(W_1, W_2 \ldots|W) \propto e^{-\sum_i (W_i-(W+E))^2/2\sigma^2}$$

Answer to Exercise 5, continuation 2

6.Integrate over the nuisance variables (E). (We won't bother to evaluate the integral)

$$p(W_1, W_2 \ldots|W) \propto \int e^{-\sum_i (W_i-(W+E))^2/2\sigma^2}\, e^{-E^2/2D^2}\, dE$$

Answer to Exercise 5, continuation 3

7. Apply Bayes' rule to get the probability distribution for what you want to know, given the measurements:

$$p(W \mid W_1, W_2 \ldots) \propto \int e^{-\sum_i (W_i - (W+E))^2 / 2\sigma^2} e^{-E^2/2\mathrm{D}^2} dE$$

What is your best estimate of the weight <x>?

Best estimate of weight is the weighted mean value

$$\langle W \rangle = \int W \, p(W \mid W_1, W_2 \ldots) \, dW$$

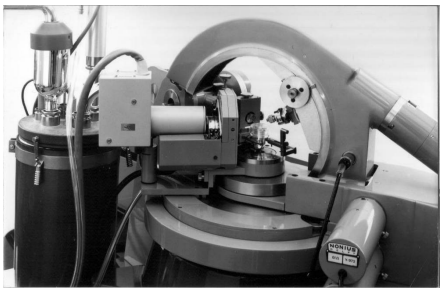
Software attached to Hardware

How our software development prepares us for the future.

Rob W.W. Hooft, Bruker AXS, Delft

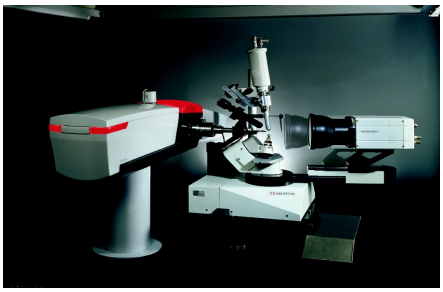
And why it is FUN!

1960s: Precision Instrument



1980s: Intermediate Complexity

2000s: Integrated Instruments



Why is so much software included?

- Research becomes technology
- Instruments are more complicated

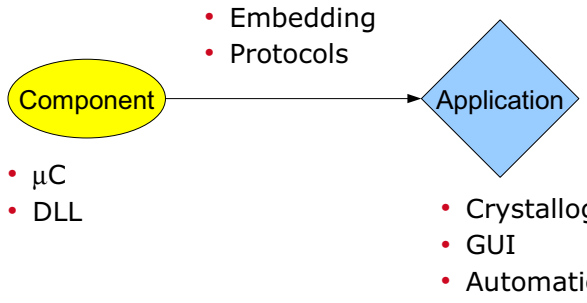
Research becomes technology

- Data collection strategy ?
- Data reduction ?
- Structure solution
- Structure refinement
- Twinned crystals
- Sample changing
- Incommensurate structures ?

Why is so much software included?

- Research becomes technology
- Instruments are more complicated

1980s instrument

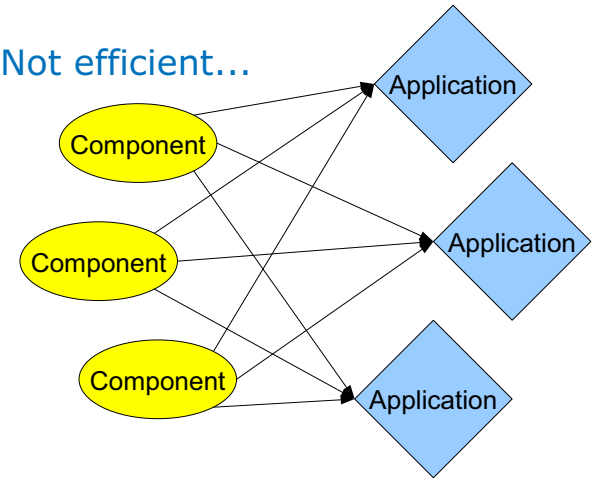


Instrument complexity

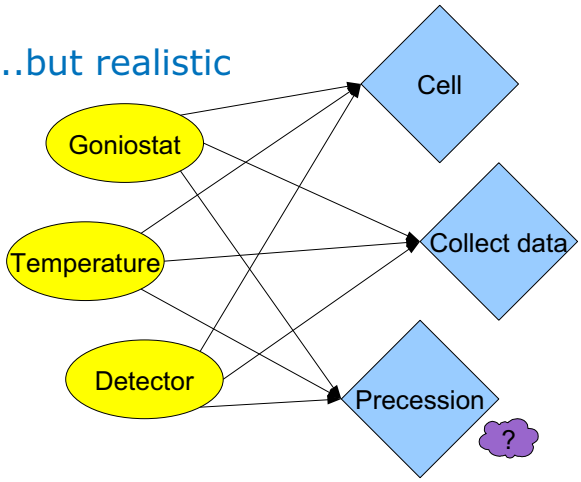
Components
- More components
- More complex components
- More alternatives
- Faster evolution

Applications
- More applications

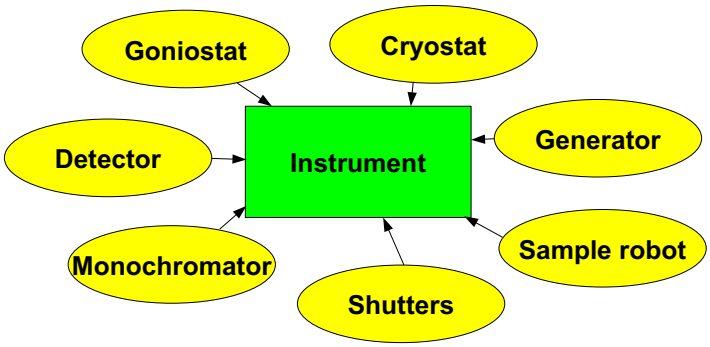
Not efficient...



...but realistic

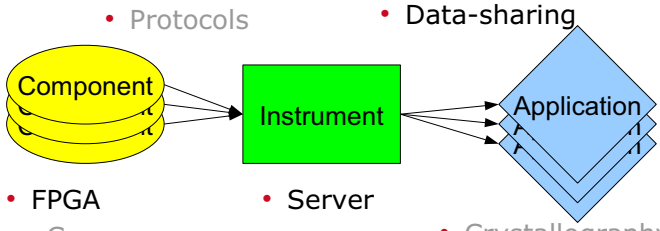


Complexity and variations



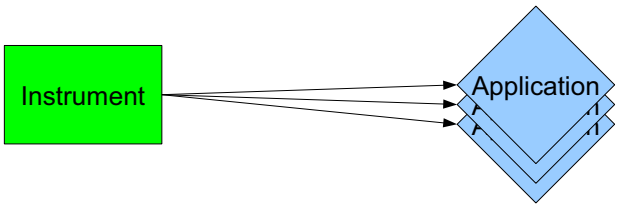
2000s instrument

- Embedding
- Protocols
- IP Protocols
- Security
- Data-sharing



- FPGA
- µC
- DLL
- Server
- Crystallography
- GUI
- Automation
- Databases

2000s instrument



- Hardware-specific
- Constants, Capabilities, Calibration ?
- Instrument-independent Application

2000s instrument

- Instrument-independent application
- One framework
- Common subsystems (e.g. frame handling, JWP)
- Thin GUI ?
- AI layer

How do we make that software?

- Pay attention to Longevity
- Pay attention to Speed
- Pay attention to Relationships

Why pay attention to longevity?

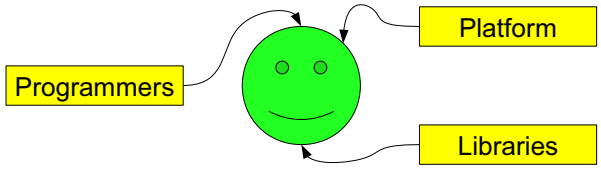
- Rewrites are expensive
- Continuity of support
- Continuity of news

How pay attention to longevity?

- Minimize external dependencies
- Assure maintainability
- Invest rather than hurry

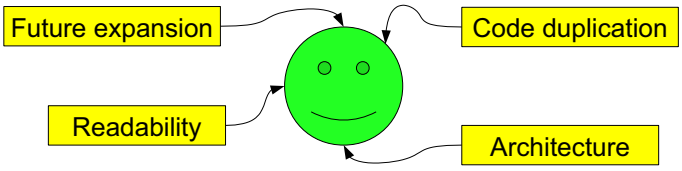
How pay attention to longevity?

- Minimize external dependencies
- Assure maintainability
- Invest rather than hurry



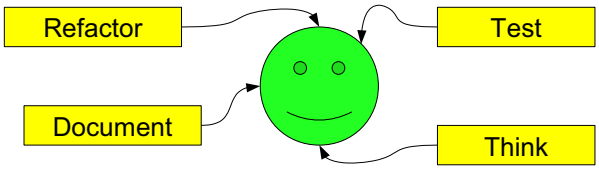
How pay attention to longevity?

- Minimize external dependencies
- Assure maintainability
- Invest rather than hurry



How pay attention to longevity?

- Minimize external dependencies
- Assure maintainability
- Invest rather than hurry



How do we make that software?

- Pay attention to Longevity
- Pay attention to Speed
- Pay attention to Relationships

Why pay attention to speed?

- Productivity
- Lifetime of bugs

How pay attention to speed?

- LEAN
- Broad expertise
- Programmer time vs. processing time
- Libraries

How do we make that software?

- Pay attention to Longevity
- Pay attention to Speed
- Pay attention to Relationships

Pay attention to relationships

- Testing
- Feedback
- New ideas

Why is it fun?

People
- Multidisciplinary
- International

Projects
- Contribution to an enterprise
- Work on future systems now
- Software will actually be used

References

LEAN
- http://en.wikipedia.org/wiki/Lean_software_development

Python coding standard
- http://www.python.org/dev/peps/pep-0008/

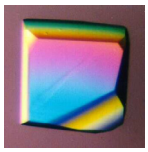
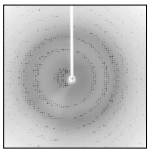
Programming at Python speed
- http://www.artima.com/intv/speed.html
- http://tinyurl.com/69c94m

Some software we use
- http://epydoc.sourceforge.net/
- http://trolltech.com/products/qt/
- http://www.riverbankcomputing.co.uk/software/pyqt

Data reduction: d*TREK Kyoto 2008

J.W. Pflugrath
Rigaku Americas

こうえん　の　ごしょうたい、　ありがとうございます。


Rigaku

Acknowledgements

Rigaku Americas
Robert Bolotovsky
Shijie Yao
Thad Niemeyer
Cheng Yang
Kris Tesh
Thom Hendrixson
Joe Ferrara

Others
Ed Westbrook
US Dept of Energy
Contract 943072401
R. Jacobson
Gerard Bricogne
EEC Workshops

Rigaku

Outline of the talk

- It's a COMPUTING WORKSHOP
- Some ancient history
- Outline of the problem
- Object oriented programming
- d*TREK internals
  - Could help you build your own program or give you ideas when building new programs
  - C++ classes and objects
- Some results

Rigaku

It's a COMPUTING WORKSHOP

- We can discuss software and algorithms
- Your chance to learn from other crystallographic computing experts
- We need more experts because …

  difficult to find and hire qualified crystallography knowledgable software engineers

Rigaku

Some Ancient History

- I was last in Kyoto in 1983 for my first IUCR computing workshop


Rigaku

Photos by A. Nakagawa

Some Ancient History

- Shortly thereafter in 1984 I began work on area detector data processing
- SCAN12, OSC, FILME, IDXREF, MOSFLM were FORTRAN packages
- XDS, buddha, xdisplayf, denzo, xengen, MADNES

Rigaku

Some Ancient History

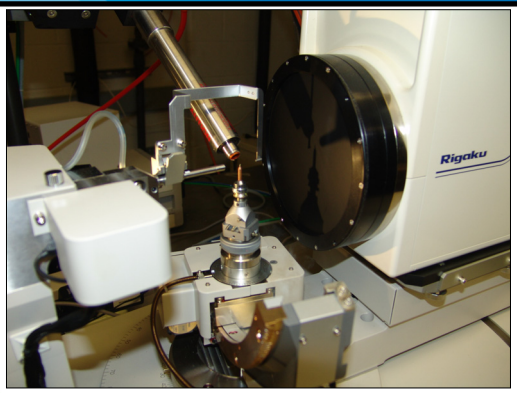
- In 1986 Gerard Bricogne organized the EEC Cooperative Workshop on Position-sensitive Detector Software
  - Very important in bringing together the developers of diffraction image processing software to share source code and ideas
  - Device-independence and shareability
  - MADNES was re-written with contributions from many to make it device independent

Some Old History

- In 1994 I moved to MSC and began work on d*TREK
  - Object-oriented
  - C++ language
  - X Windows / Motif
  - Truly device-independent

The Diffraction Experiment

- Pick up crystal in loop, plunge into LN_2
- Put crystal on magnet on goniometer head and align optically
- Take a diffraction image or two
- Look at image(s) and decide whether to proceed
- Collect images, index, integrate, scale



Another example

Thom Hendrixson

Before Object-oriented programming

- FORTRAN and C
  - Data structures and separate routines to act on that data.
  - Often fixed array sizes
  - /Common blocks/
  - Structured programming

Object-oriented programming

- C++
  - Data and methods were combined into CLASSES
  - Objects were instances of the classes
  - Dynamic array sizes
  - Encapsulation
  - Overloading
  - Polymorphism
  - Inheritance
  - ***Not C*** with new style

Rigaku

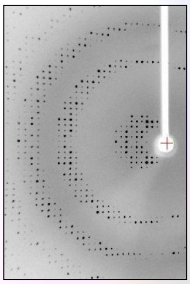
The Diffraction Experiment

- Naturally fits into Object-Oriented Programming
- Objects you can touch:
  - Source
  - Detector
  - Shutter
  - Goniometers
    - Crystal
    - Detector
  - Crystal

Rigaku

The Diffraction Experiment

- Objects you can't touch
  - Images
  - Spacegroup
  - Reflections
  - Array of Reflections
  - et al.



Rigaku

d*TREK internals

Example Class:
Cimage – encapsulate an image and methods to deal with the pixels

Member variables
- Raw data
  - Dimensions
  - Byte order
  - Data type
- Member Objects
  - Header
  - Scan info
  - Non-uniformity
  - Spatial distortion
- Compression info
- Saturated value

Methods (not all of them)
- Constructors
  - Cimage oImage(sFilename);
- Destructor
- nGetDimensions
- nRead
- nWrite
- fGetPixel
- fPutPixel
- nGetRect
- poGetHeader
- …

Rigaku

d*TREK internals

Example Class:
Cgoniometer – encapsulate an goniometer and methods to use it

Member variables
- Num values
- Names
- Vectors
- Units
- Rot values
- Trans values
- Hardware limits
- …

Methods (not all of them)
- Constructors
  - Cgoniometer(oHeader);
- Destructor
- nGetNames
- nUpdateHeader
- vCalcGetRotMatrix
- nGetRotVector
- nGetValue
- nSetValue
- …

Rigaku

d*TREK internals

Example Class:
Cgoniometer – 2 examples in a header

- Goniometer axes and angles
  - Single axis, Eulerian, Kappa, Other
  - Axes vectorially defined

```
CRYSTAL_GONIO_DESCRIPTION= Kappa goniostat;
CRYSTAL_GONIO_NUM_VALUES=3;
CRYSTAL_GONIO_NAMES=Omega Kappa Phi;
CRYSTAL_GONIO_UNITS=deg deg deg;
CRYSTAL_GONIO_VALUES=0.000 0.000 0.000;
CRYSTAL_GONIO_VECTORS= -1 0 0 0.6426 0 -0.7662 -1 0 0;
...
```

Rigaku

d*TREK internals

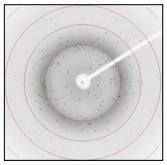
Example Class:
Cgoniometer – 2 examples in a header

```
CRYSTAL_GONIO_DESCRIPTION=Eulerian 3-circle;
CRYSTAL_GONIO_NAMES=Omega Chi Phi;
CRYSTAL_GONIO_NUM_VALUES=3;
CRYSTAL_GONIO_UNITS=deg deg deg;
CRYSTAL_GONIO_VALUES=0 0 0;
CRYSTAL_GONIO_VECTORS=1 0 0 0 0 1 1 0 0;

RX_GONIO_DESCRIPTION=R-AXIS motorized distance;
RX_GONIO_NAMES=RotAboutBeam 2Theta RotY XShift YShift Distance;
RX_GONIO_NUM_VALUES=6;
RX_GONIO_UNITS=deg deg deg mm mm mm;
RX_GONIO_VALUES=-1.6972 0.0239 0.0140 0.0486 -0.0712  111.4263;
RX_GONIO_VALUES_SIGMA=0.0049 0.0165 0.0239 0.0131 0.0131 0.0444;
RX_GONIO_VECTORS=0 0 1  1 0 0  0 1 0  1.0 0.0 0.0 0.0  1.0 0.0 0.0 0.0  -1.0 ;
```


d*TREK internals

Some Image properties

```
RX_NONUNF_INFO=beam.mask;
RX_NONUNF_TYPE=Simple_mask;
RX_SPATIAL_BEAM_POSITION=1496.70 1500.90;
RX_SPATIAL_DISTORTION_INFO=1496.70 1500.90 0.10000 0.10000;
RX_SPATIAL_DISTORTION_TYPE=Simple_spatial;
RX_SPATIAL_DISTORTION_VECTORS=0 1 -1 0;
```



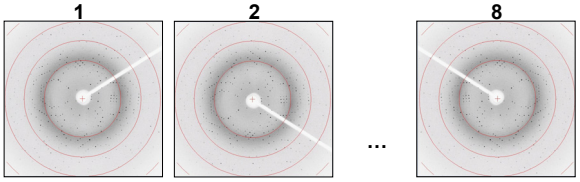
d*TREK internals

Image properties

```
RX_SPATIAL_DISTORTION_VECTORS=0 1 -1 0;
                             0 1  1 0;
                             ...
                             1 0  0 1;
```

8 possibilities!

1 2 … 8



d*TREK internals

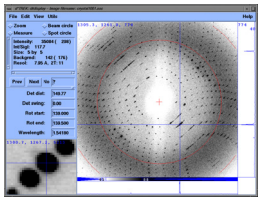
Image properties

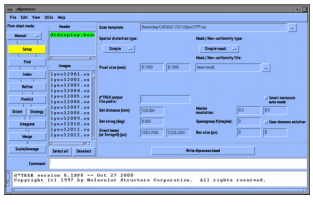
See also CBFlib web pages (Crystallographic Binary Format aka imgCIF)

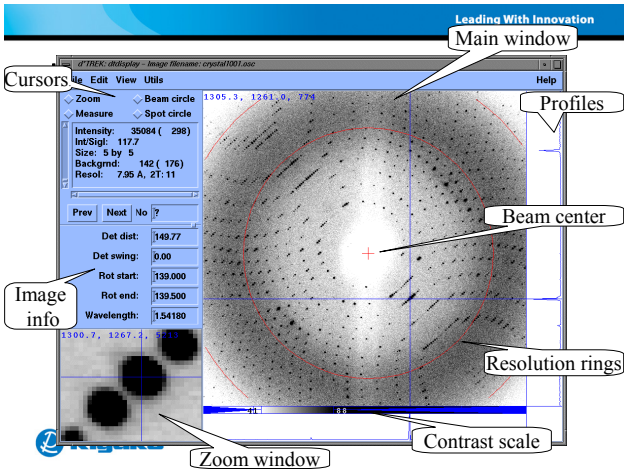
http://www.bernstein-plus-sons.com/software/CBF/

d*TREK tutorial

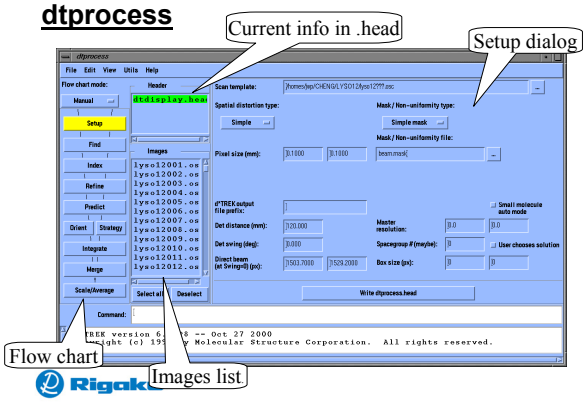
- New video tutorials, see:
  `${DTREK_ROOT}/doc/VIDEO`





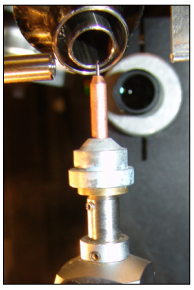


dtprocess



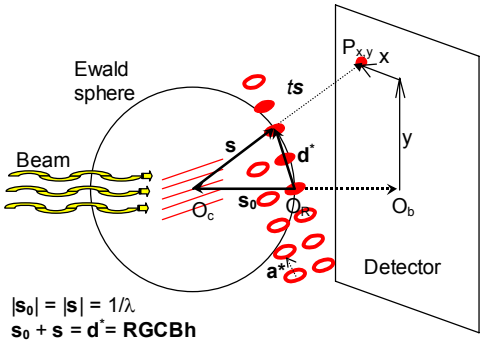
The Diffraction Experiment

- Pick up crystal in loop, plunge into LN_2
- Put crystal on magnet on goniometer head and align optically
- Take a diffraction image or two
- Look at image(s) and decide whether to proceed
- Collect images, index, integrate, scale



Reflections (from images)

- Find
  - X,Y, ϕ
    - Images -> Reflnlist
- Index
  - Crystal unit cell
  - Crystal orientation
- Refine
  - Crystal
  - Detector
  - Source
- Predict / Strategy
  - Rotation start, end
  - Completeness
- Integrate
  - *hkl*, Intensity, σ_I
  - Profile fitting
- Scale and Average
  - Rmerge
  - Rmeas
  - $|\chi^2|$



$|\mathbf{s}_0| = |\mathbf{s}| = 1/\lambda$

$\mathbf{s}_0 + \mathbf{s} = \mathbf{d}^* = \mathbf{RGCBh}$

$\mathbf{s} = f(P_{x,y}, \phi,$ Detector position)

Diffraction math

$\mathbf{s}_0 + \mathbf{s} = \mathbf{d}^* = \mathbf{RGCBh}$

h	Miller index (*h,k,l*)
B	Crystal orth. matrix (a,b,c,α,β,γ)
C	Crystal orientation matrix
G	Crystal goniometer matrix
R	Rotation axis matrix
d*	Reciprocal lattice vector
$\mathbf{s}_0$	Direct beam wavevector
s	Scattered beam wavevector

Refine

- $\mathbf{s}_o + \mathbf{s}_i = \mathbf{RGCBh} = \mathbf{d}^*_i$
- Min $\chi^2 = \Sigma w_i (\mathbf{s}_{i,obs} - \mathbf{s}_{i,calc})^2$
  $= \Sigma w_i[\mathbf{s}_{i,obs} - (\mathbf{RGCBh} - \mathbf{s}_o)]^2$
- Crystal (**B**, **C**): a, b, c, α, β, γ, Rot1, Rot2, Rot3
- Detector ($\mathbf{s}_{i,obs}$ = *f(Det, X, Y,* ϕ, ***R***))
  - Beam center, Distance, Rotations (2θ)
- Source ($\mathbf{s}_o$):Direction, wavelength

Integrate - Predict

d* = x_r = RGCBh **$x_r - s_0 = s \rightarrow P(x,y)$**

Rocking curve

$R_i = 2L[\Delta d^* \cos\theta + (\delta\lambda/\lambda) d^* \sin\theta]$

Lorentz factor

$L = | 1.0 / (\mathbf{x_r} \cdot (\mathbf{r_1} \times \mathbf{s_0})) |$

Polarization factor

$p = 1 - [P_n * (\mathbf{s} \cdot \mathbf{s_n})^2 + (1 - P_n) * (\mathbf{s} \cdot \mathbf{n_p})^2]$

Oblique incidence correction factors

$O_1 = (1 - \exp(f_{obl})) / (1 - \exp(f_{obl} / \cos\alpha)$

$O_2 = \exp(f_{obl})) / \exp(f_{obl} / \cos\alpha)$

Rigaku

Integrate - Predict

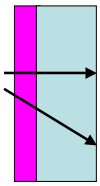
Oblique incidence correction factors

[See Zaleski, Wu & Coppens (1998) J. Appl. Cryst. 31, 302-304.]

$O_1 = (1 - \exp(f_{obl})) / (1 - \exp(f_{obl} / \cos\alpha)$ (ZWC)

$O_2 = \exp(f_{obl})) / \exp(f_{obl} / \cos\alpha)$

$f_{obl} = -\mu d$



Rigaku

Crystal Screening

- Expose a diffraction image or two
  - Pick a distance, pick an exposure time
- Try to "rank" or "score" the image
  - Run **dtranker**
  - If rank high enough, run dtmultistrategy
  - Collect the strategy

Rigaku

Optimize the Data Collection

- Screening and ranking of crystals
- Strategy
  - Rotation range
  - Image rotation increment
  - Exposure time
- Completeness and redundancy
- Anomalous differences

Rigaku

dtmultistrategy - Possible scans Generic: Any goniometer!

- Goniometer axes and angles
  - Single axis, Eulerian, Kappa, Other
  - Axes vectorially defined

Inputs

- Scan axis
- Rotation limits
- Maximum resultant scan range
- Collision limits or conditions
  - Knowledge of detector size, shape, position
  - XML file

Rigaku

dtmultistrategy - Possible scans

- Goniometer axes and angles
  - Single axis, Eulerian, Kappa, Other
  - Axes vectorially defined

```
CRYSTAL_GONIO_DESCRIPTION= Kappa goniostat;
CRYSTAL_GONIO_NUM_VALUES=3;
CRYSTAL_GONIO_NAMES=Omega Kappa Phi;
CRYSTAL_GONIO_UNITS=deg deg deg;
CRYSTAL_GONIO_VALUES=0.000 0.000 0.000;
CRYSTAL_GONIO_VECTORS= -1 0 0 0.6426 0 -0.7662 -1 0 0;
```

- Scan axis
- Rotation limits
- Maximum resultant scan range

Rigaku

Robert Bolotovsky

dtmultistrategy - Collision limits

- All objects described as 3D polygons
- Collision limits or conditions
  - Knowledge of detector size, shape, position
  - XML file

Thom Hendrixson

Orient the Crystal dtorient

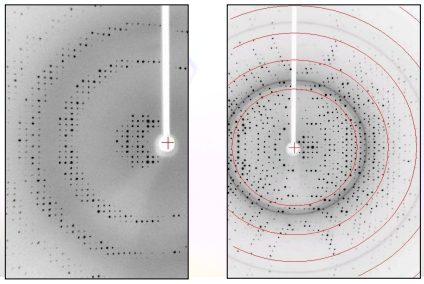
- Once crystal orientation is known
- And goniometer definition is valid
- Specify a crystal vector
  - a*, b, (1 1 0), …

to make parallel to a lab vector

- X, Z, source, …

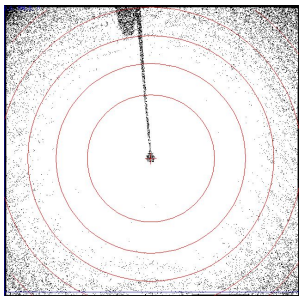
Phil Evans

Orient the Crystal dtorient



Do I need to mask out the beamstop shadow?

- Can I use low pixel value cutoff instead?
  - No, because often the pixel value in the shadow is higher at low resolution than non-shadowed pixel values at high resolution.

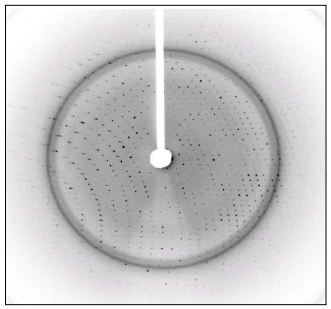


Do I need to mask out the beamstop shadow?

Solution use a recursive edge detection algorithm to mask out the shadow:

- Algorithm BeamMask:
  a) Select pixel at P_{xy} in shadow
  b) Move along directions and find edge (1st, 2nd derivative tests). Set value based on edge detection.
  c) Mask out pixels near P_{xy} that are below value.
  d) If any pixels masked out, then call BeamMask, else return.
- Benefits: Very localized value of the shadow

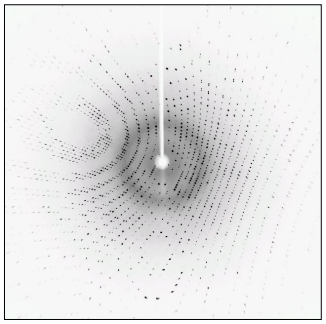
P142 – Remacle & Winter



Integrate

- How to select best
  - Integration box size
  - Integration spot shape

even when spot shapes vary?



Rigaku

Integrate

Box size

Where are neighboring reflections?

Known from prediction

Spot size

What are the spot sizes?

Known from fitting to ellipsoids

Rigaku

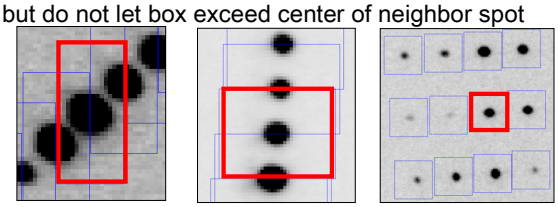
Integrate

Box size

So … set box size to ….

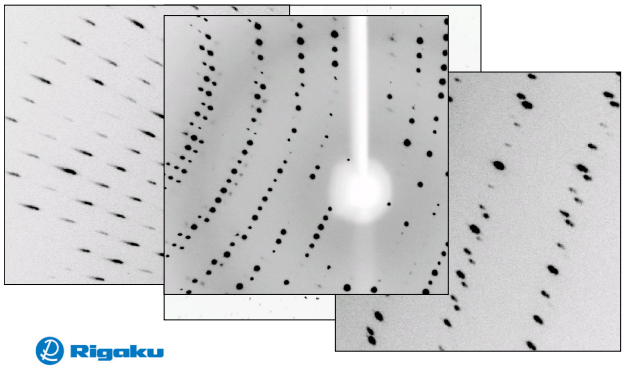
2.8 to 3 times the spot size in that direction,

but do not let box exceed center of neighbor spot



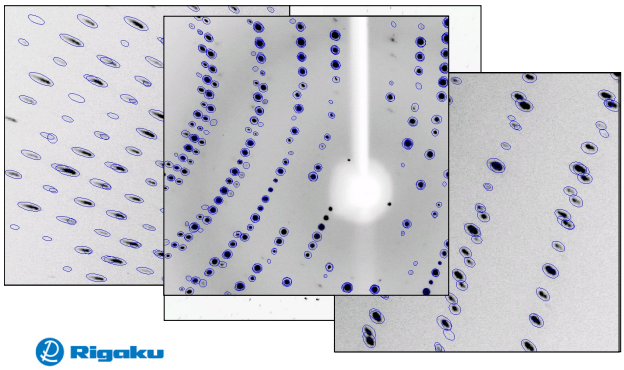
Rigaku

Spot shapes vary! An awful lot!



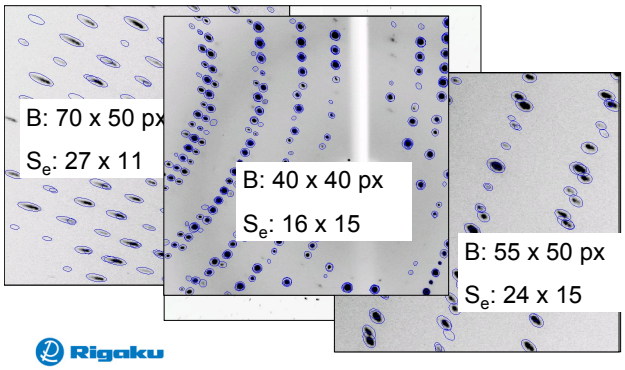
Rigaku

Solution: Empirical ellipsoid spot shapes



Rigaku

Solution: Empirical ellipsoid shapes



Rigaku

Scaling

- Correction of systematic errors
- Outlier rejection
- Validation of sigmas

$$\sigma_{adj}^2 = (\sigma_{in} E_{mul})^2 + (I_{in} E_{add})^2$$

Scaling

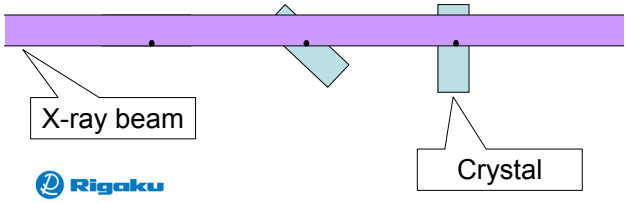
Correction of systematic errors

- different crystal volumes
- different exposure times
- different detectors
- radiation damage
- wavelength dependent factors
- different or fluctuating source intensities
- different absorption due to different paths through the crystal and other matter

Scaling

Correction of systematic errors

- Different crystal volume in the beam
  - Due to poor optical alignment
  - Due to poor beam and rotation axis alignment



Scaling

Two batches of 3 reflections each

h k l	Intensity	Batch
1 3 5	100	1001
1 3 6	200	1001
1 3 7	300	1001
1 3 5	200	2001
1 3 6	400	2001
1 3 7	600	2001

$$\chi^2 = \sum_h \sum_j w_{hj} (G_{hj} I_{hj} - \langle I_h \rangle)^2 = \sum_h \sum_j \frac{(G_{hj} I_{hj} - \langle I_h \rangle)^2}{\sigma_{hj}^2}$$

Scaling

Two batches of 3 reflections each

h k l	Intensity	Batch
1 3 5	104±21	1001
1 3 6	199±27	1001
1 3 7	311±36	1001
1 3 5	198±31	2001
1 3 6	604±21	2001
1 3 7	590±26	2001

$$\chi^2 = \sum_h \sum_j w_{hj} (G_{hj} I_{hj} - \langle I_h \rangle)^2 = \sum_h \sum_j \frac{(G_{hj} I_{hj} - \langle I_h \rangle)^2}{\sigma_{hj}^2}$$

Scaling

- Simple K and B
- Local scaling
- Absorption correction
  - Incident + Scattered beam directions
  - Spherical harmonics
  - Fourier series

Scaling

- Reduced $|\chi^2|$

$$\left|\chi^2\right| = \frac{\sum_{h}^{N_{uniq}} \sum_{i}^{n_h} \frac{\left(I_{hi} - \langle I_h \rangle\right)^2}{\sigma^2_{hi,adj}}}{\sum_{h}^{N_{uniq}} (n_h - 1)} = \frac{\sum_{h}^{N_{uniq}} \sum_{i}^{n_h} \frac{\left(I_{hi} - \langle I_h \rangle\right)^2}{\sigma^2_{hi,adj}}}{N_{tot} - N_{uniq}} \approx 1$$

Rigaku

Scaling

$$Rmerge = \frac{\sum_h \sum_i |I_{hi} - \overline{I_h}|}{\sum_h \sum_i \overline{I_h}}$$

$$R_{meas} = \frac{\sum_h [N/(N-1)]^{1/2} \sum_i |I_{hi} - \overline{I_h}|}{\sum_h \sum_i \overline{I_h}}$$

$$R_{pim} = \frac{\sum_h [1/(N-1)]^{1/2} \sum_i |I_{hi} - \overline{I_h}|}{\sum_h \sum_i \overline{I_h}}$$

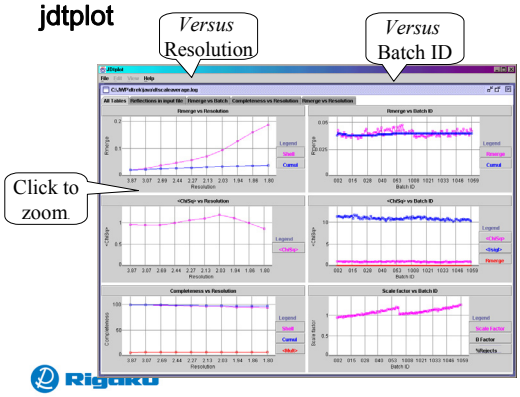
Weiss & Hilgenfeld (1997) *J. Appl. Cryst.* **30**, 203-205.

Weiss, M. (2001) *J. Appl. Cryst.* **34**, 130-135.

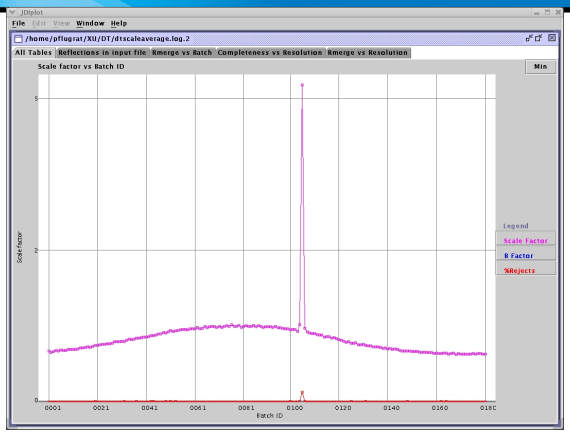
Diederichs & Karplus (1997) *Nature Struct. Biol.* **4**, 269-274.

Rigaku

jdtplot



Rigaku



What to do if scaling bad?

- Exclude reflns on bad images
- Restrain batch scale factors
- Change resolution cutoffs
- Check spacegroup hypothesis
- Radiation damage?
- Use another crystal

Rigaku

Output statistics

```
Summary of data collection statistics
---------------------------------------------------------------
 Spacegroup                  P6
 Unit cell dimensions        130.41   130.41   65.57
                             90.00    90.00    120.00
 Resolution range            29.29 - 1.80     (1.86 - 1.80)
 Total number of reflections 636461
 Number of unique reflections 59057
 Average redundancy          10.78            (10.37)
 % completeness              99.8             (98.9)
 Rmerge                      0.097            (0.543)
 Reduced ChiSquared          0.99             (0.80)
 Output <I/sigI>             15.9             (3.6)
---------------------------------------------------------------
  Note: Values in () are for the last resolution shell.


  641271 reflections in data set
       0 reflections rejected (|ChiSq| >  50.00)
    4810 reflections total rejected (  0.75%  |Deviation|/sigma >  13.69)
  123625 reflections excluded from scaling/absorption (I/sig <= 5.00)
```

Rigaku

Does $|\chi^2|$ or σ affect Rmerge?

$$\left|\chi^2\right| = \frac{\sum_{h}^{N_{uniq}} \sum_{i}^{n_h} \frac{\left(I_{hi} - \langle I_h \rangle\right)^2}{\sigma^2_{hi,adj}}}{\sum_{h}^{N_{uniq}} (n_h - 1)} = \frac{\sum_{h}^{N_{uniq}} \sum_{i}^{n_h} \frac{\left(I_{hi} - \langle I_h \rangle\right)^2}{\sigma^2_{hi,adj}}}{N_{tot} - N_{uniq}} \approx 1$$

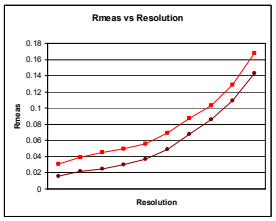
$$Rmerge = \frac{\sum_h \sum_i | I_{hi} - \overline{I_h} |}{\sum_h \sum_i \overline{I_h}}$$

Rigaku

Scaling

Summary of data collection statistics

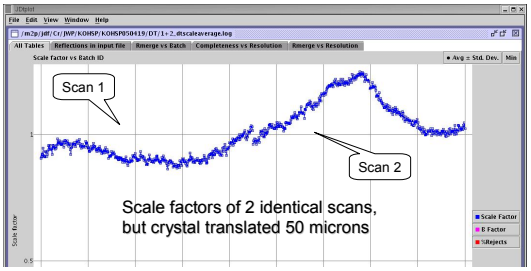
Spacegroup	$P2_12_12_1$		
Unit cell dimensions	29.95 90.00	55.94 90.00	72.08 90.00
Mosaicity	0.22		
Resolution range	13.52 - 1.50	(1.55 - 1.50)	
Total number of reflns	87054		
Number of unique refns	19815		
Average redundancy	4.39		(4.15)
% completeness	98.9		(99.5)
Rmeas (I+ = I-)	0.046		(0.168)
Rmeas (I+ != I-)	0.029		(0.143)
Ras	2.4		(1.9)
Output <I/sigI>	43.0		(12.6)



Rigaku

Integrate & Scaling

- Do I need to center my crystal in the beam?



YES!

Rigaku

The Diffraction Experiment

- The data that gets you the atomic coordinates
- All subsequent steps are just massaging the data
- Pay attention to details as you may not have a chance to get better data
- There is no more data after this experiment!

- Use software like dtranker & dtmultistrategy

Rigaku

Rigaku Americas
Robert Bolotovsky
Shijie Yao
Thad Niemeyer
Cheng Yang
Kris Tesh
Thom Hendrixson
Joe Ferrara

Others
Ed Westbrook
R. Jacobson
US Dept of Energy
Contract 943072401
Gerard Bricogne
EEC Workshops

Thanks!

こうえん　の　ごしょうたい、　ありがとうございました。

Rigaku

MRC Laboratory of Molecular Biology

Data reduction: Mosflm

Harry Powell
MRC Laboratory of Molecular Biology, Cambridge, UK

Kansai Seminar House, Kyoto 20th August 2008

Data reduction - three main steps

- Indexing - to locate spots on image and obtain approximate cell parameters
- Refinement - to optimise the initial values
- Integration - to perform the actual measurement of the spot intensities

Indexing

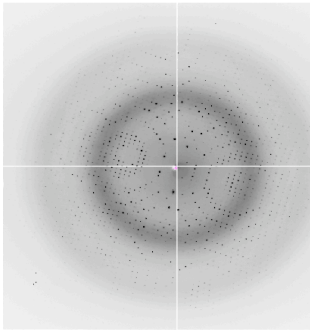
- Find spots on the image
- Convert 2D co-ordinates (image) to 3D co-ordinates (reciprocal space)
- Index
- Cell reduction
- Apply Bravais lattice symmetry
- Pick a putative solution
- (Estimate mosaic spread)

Indexing

After locating the spots on an image, we can convert the 2D image co-ordinates to scattering vectors that correspond to lattice points in a (distorted) 3D reciprocal lattice by means of the relationship

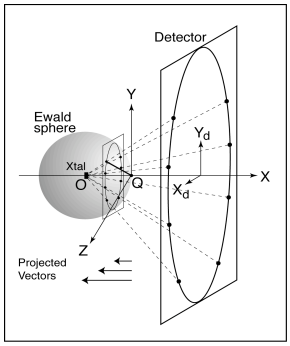
$$s = \begin{pmatrix} D/r - 1 \\ X_d/r \\ Y_d/r \end{pmatrix}$$

$$r = \sqrt{D^2 + X_d^2 + Y_d^2}$$



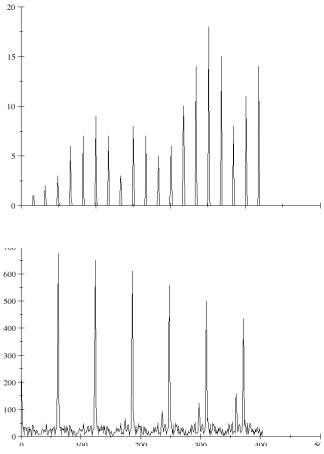
Indexing

If the scattering vectors calculated from the 3D R.L. co-ordinates are projected along a real space axis direction (such as a, b or c) all the projected vectors for spots in the same reciprocal space plane will have the same length, as will all those spots in the next plane etc. This will give a large peak in the Fourier transform.



Indexing

The first large peak in the Fourier transform corresponds to a real space cell edge length. In this case, ~56Å.



Refinement

- Optimise the fit of observed to predicted spot positions
- Improve estimates of:
  - Crystal parameters
  - Instrument parameters
- Can be performed by either (or both):
  - Positional refinement using spot co-ordinates on detector
  - Post-refinement using intensity measurements of partial reflections.

Positional refinement and post-refinement

- Positional refinement
  - since it uses the spot positions on each image, it can be done for each image without reference to the others.
- Post-refinement
  - needs intensity measurements for spots which are partial across at least two images
  - ∴ needs at least two adjacent images (and probably more for fine-phi slicing, where the mosaic spread is more than twice the oscillation angle)
- Mosflm uses both to refine different parameters

Positional refinement

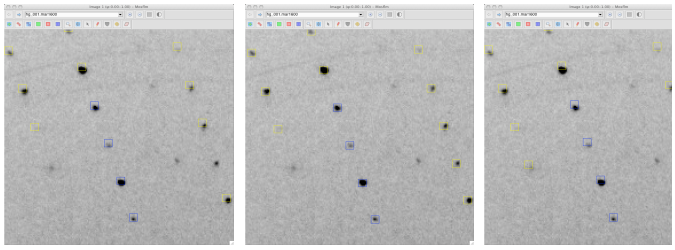
Minimises

$$\Omega_1 = \sum_{i=1}^{n} w_{ix}\left(X_i^{calc} - X_i^{obs}\right) + w_{iy}\left(Y_i^{calc} - Y_i^{obs}\right)$$

n.b. i) rotation of crystal about phi axis has no effect on this residual so it can't be refined.
ii) cell dimensions and other parameters (*e.g.* crystal to detector distance) may be strongly correlated.

Positional refinement

Can be visualised – the prediction boxes (X^{calc}, Y^{calc}) are displaced from th diffraction spots (X^{obs}, Y^{obs}) because the crystal to detector distance is 1% too long in the left-hand image, and the cell edges are 1% too short in th right-hand image.



Post-refinement

Minimises the angular residual δ (see next slide) via

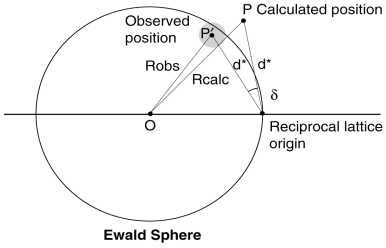
$$\Omega_2 = \sum_{i=1}^{n} w_i \left[\frac{(R_i^{calc} - R_i^{obs})}{d_i^*}\right]^2$$

n.b. we need:
i) a reasonable knowledge of intensities for this, so it can only be done after integration - hence "*post-refinement*"
ii) a model for the "rocking curve"

And we can refine either the mosaicity or the beam divergence

Post-refinement

We can visualise this in the Ewald sphere construction, minimising the angular residual δ. A suitable model for the rocking curve allows us to determine the "observed" position (P').



Integration

- Place measurement boxes over spots
- Optimise measurement boxes
- Account for the background counts
- Measure the intensity of each spot
- Take the individual detector characteristics into account
- Calculate error estimates

Currently two common methods:

- Summation integration
- Profile fitting

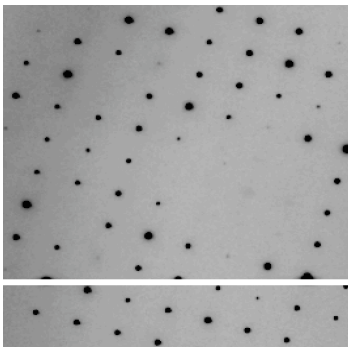
Most other (less common) methods can be viewed as "improvements" on these, *e.g.* "seed-skewness".

Summation integration

- In the absence of background, just add the pixel counts in the spot region together - but there is always background!
- Need to define spot and background regions - we cannot measure background directly under the spots, so we calculate a local background plane and slope from nearby non-spot pixels
- Use this to remove the background under the spots
- Weak spots may have their shoulders under the background

Profile fitting integration - background

- Based on the assumption that spots corresponding to fully recorded reflections in the same region of the detector (and on images nearby in phi) have similar profiles.



Profile fitting integration

use a profile determined empirically from well-measured reflections to measure the intensity of weak reflections (whose shoulders disappear below the background), simply by minimising *R to* optimise a scale factor (*K*):

$$R = \sum_{peak\ pixels} w\ (X_i - KP_i)^2$$

- requires accurate (sub-pixel) placement of the profile
- improves variance estimates for weak reflections
- should reduce random error (weak reflections)
- may increase systematic error (strong reflections)

Profile fitting integration - partials

- in 2D integration, the profiles are normally developed by analysing fully recorded reflections (the act of building up profiles from partials is the basis of 3D integration). It can be shown that there is some validity in this approach, but there is some disagreement about this, so it will not be examined further here.
- in fine phi slicing, where there are no fully recorded reflections, the situation is more complex!

Acknowledgements

MRC Laboratory of Molecular Biology

Andrew Leslie
Phil Evans

CCP4
Medical Research Council

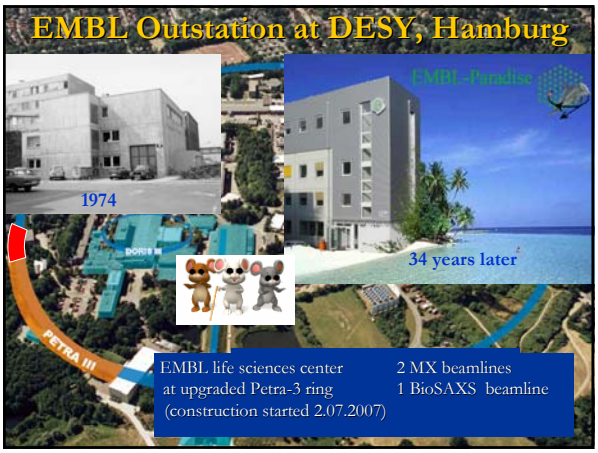
Everyone at the 1986 LURE Area Detector Workshop (which I did not attend!)

Small-angle scattering from macromolecular solutions

Dmitri Svergun
EMBL, Hamburg Outstation, and
Institute of Crystallography, Moscow

EMBL

EMBL Outstation at DESY, Hamburg



EMBL life sciences center at upgraded Petra-3 ring (construction started 2.07.2007)

2 MX beamlines
1 BioSAXS beamline

Biological SAXS @ EMBL-HH

Group leader: D. Svergun
Project leader, Petra-3: M. Roessle
Staff: M. Petoukhov, D.Franke
Postdocs: P. Konarev, T. Ikonen, H.Mertens, A.Zozulya, W.Shang, M.Gajda
Predocs: E. Mylonas, A. Kikhney, A.Shkumatov

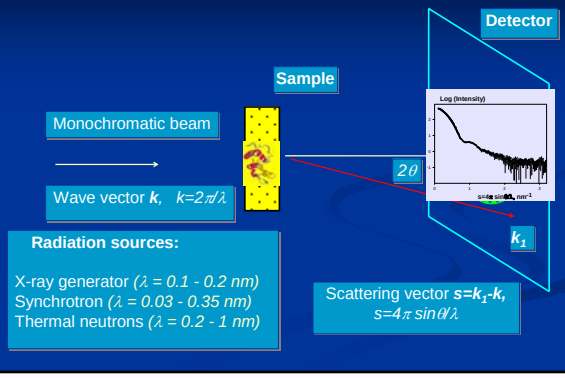
Major tasks

- Running EMBL SAXS beamline X33
- User support and collaborative projects
- Development of data analysis methods
- Education and training (including regular EMBO courses)
- BioSAXS beamline @ Petra-3
- SAXIER: Small-angle X-ray scattering at high brilliance European synchrotrons for bio- and nano-technology

When crystallographers learn SAS (trivial case)

SAXSMAN © A.Kikhney

Small-angle scattering: experiment

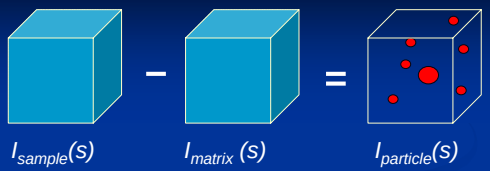


Radiation sources:

X-ray generator (λ = 0.1 - 0.2 nm)
Synchrotron (λ = 0.03 - 0.35 nm)
Thermal neutrons (λ = 0.2 - 1 nm)

Scattering vector $s=k_1-k$, $s=4\pi \sin\theta/\lambda$

Small-angle scattering: contrast



$I_{sample}(s)$ $I_{matrix}(s)$ $I_{particle}(s)$

- To obtain scattering from the particles, matrix scattering must be subtracted, which also permits to significantly reduce contribution from parasitic background (slits, sample holder etc)
- **Contrast** $\Delta\rho = \langle\rho(r) - \rho_s\rangle$ where ρ_s is the scattering density of the matrix, may be very small for biological samples

X-rays versus Neutrons

Addition of sucrose or salts | *Isotopic H/D substitution*

RNA, 550 e/nm^3

60% sucrose, 430 e/nm^3

Protein, 410 e/nm^3

H_2O, 344 e/nm^3

D-Protein, 130% D_2O

D-RNA, 120% D_2O

D_2O, 6.38×10^{10} cm^{-2}

H-RNA, 70% D_2O

H-Protein, 40% D_2O

H_2O, -0.59×10^{10} cm^{-2}

Contrast variation (H.Stuhrmann, 1965)

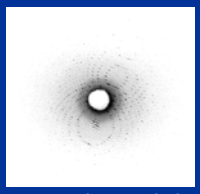
Scattering from dilute macromolecular solutions (monodisperse systems)

$$I(s) = 4\pi \int_0^D p(r) \frac{\sin sr}{sr} dr$$

The scattering is proportional to that of a single particle averaged over all orientations, which allows one to determine size, shape and internal structure of the particle at low (*1-10 nm*) resolution.

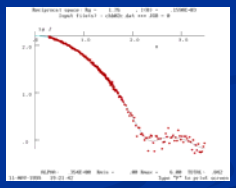
Crystal versus solution

- Thousands of reflections
- 3D, high resolution



- Data undersampled, $\Delta s = 2\pi / D$

- A few Shannon channels
- 1D, low resolution

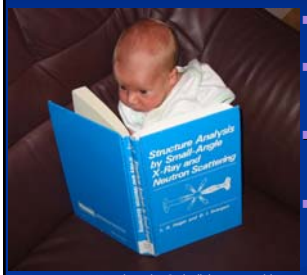


- Data oversampled, $\Delta s << \pi / D$

Crystal versus solution

Crystal versus solution

SAS renaissance in structural biology



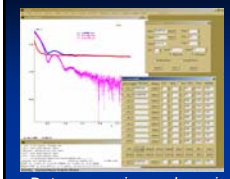
A young NMR talent (E.Gabel) interested in SAS (courtesy of F.Gabel, IBS, Grenoble)

- SAS requires neither crystals nor special sample preparation
- SAS is not limited by molecular mass and is applicable under nearly physiological conditions
- SAS permits quantitative analysis of kinetic processes like (dis)assembly
- SAS allows to study structural transitions and conformational changes on which biological function often relies

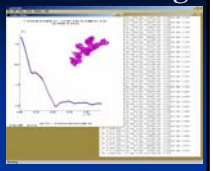
The Major Problem of SAS

- Sample preparation
- Experiment
- Data processing
- Unambiguous interpretation
- Changing conditions
- Relation to function

Methods development at EMBL-Hamburg

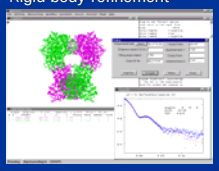


Used in more than 500 laboratories worldwide



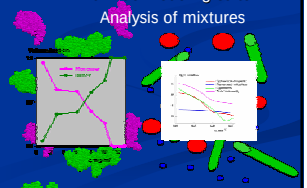
Data processing and manipulations

Rigid body refinement



Ab initio modeling suite

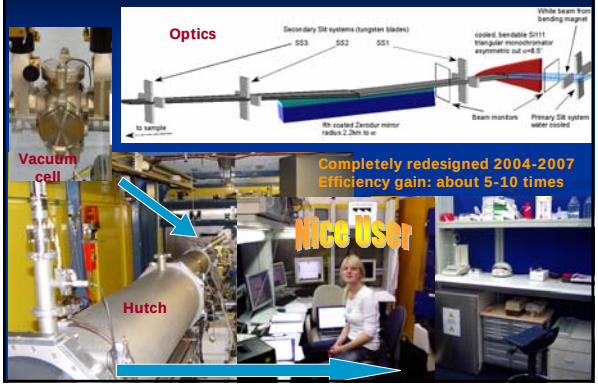
Analysis of mixtures



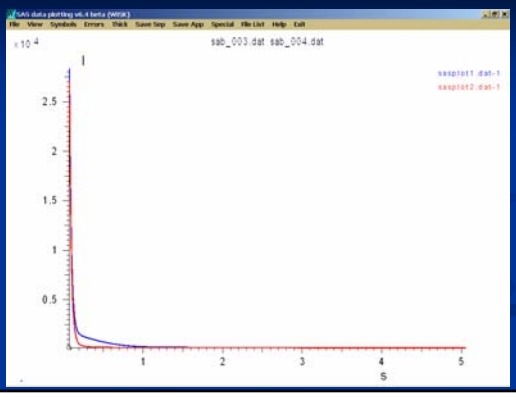
Konarev, P.V., Petoukhov, M.V., Volkov, V.V. & Svergun, D.I. (2006). *J. Appl. Cryst.* **39**, 277

Experiment and data processing

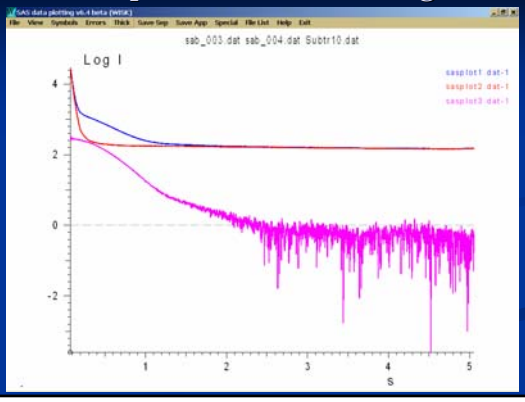
EMBL BioSAXS beamline X33, 2007



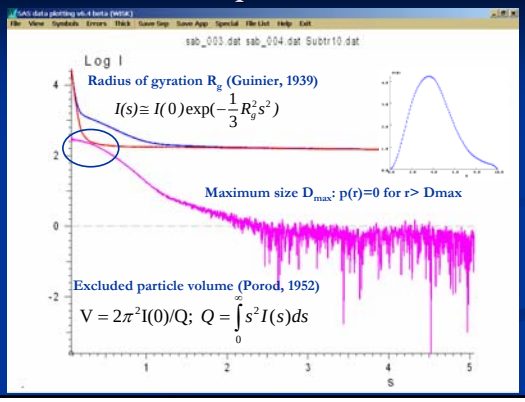
Sample and buffer scattering



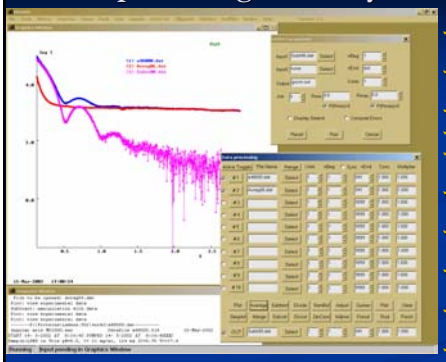
Sample and buffer scattering



Overall parameters



Data processing and analysis: PRIMUS



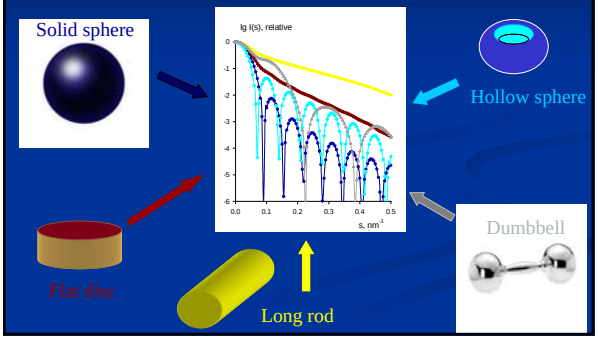
- ✓ Calibration and normalization
- ✓ Raw data processing and radial averaging
- ✓ Data manipulations
- ✓ Merging and splicing
- ✓ Concentration series analysis
- ✓ Computation of invariants
- ✓ Indirect Fourier transformation
- ✓ Simple bodies modelling
- ✓ Singular value decomposition
- ✓ Analysis of mixtures
- ✓ Peak analysis

Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. (2003). *J. Appl. Cryst.* **36**, 1277

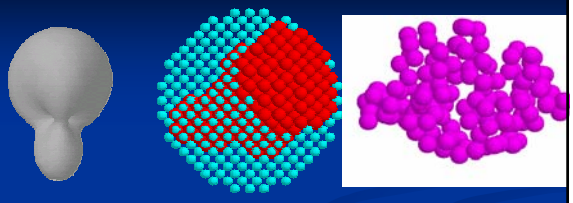
Ab initio methods

Advanced methods of SAS data analysis employ spherical harmonics (Stuhrmann, 1970) instead of Fourier transformations

The scattering is related to the shape (or low resolution structure)



Models for *ab initio* methods



Envelope function

Stuhrmann, H. B. (1970) *Z. Physik. Chem. N.F.* **72**, 177

Svergun, D.I. *et al.* (1996) *Acta Crystallogr.* **A52**, 419

Bead models

Chacón, P. *et al.* (1998) *Biophys. J.* **74,** 2760

Svergun, D.I. (1999) *Biophys. J.* **76**, 2879

Dummy residues model

Svergun, D.I., Petoukhov, M.V. & Koch, M.H.J. (2001) *Biophys. J.* **80**, 2946-2953.

All the methods minimize Discrepancy[Data] + Penalty[Additional info]

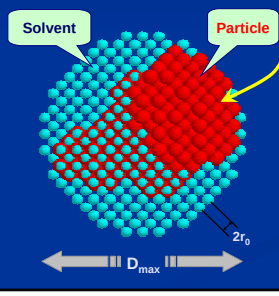
Bead (dummy atoms) models

A sphere with diameter D_{max} is filled by densely packed beads of radius $r_0 << D_{max}$. A configuration vector X indicates whether the j-th atom belongs to the particle or to the solvent.

Vector of model parameters:

$$Position\ (j) = X(j) = \begin{cases} 1 & if\ particle \\ 0 & if\ solvent \end{cases}$$

(phase assignments)



The number of model parameters $M \approx (D_{max}/r_0)^3 \approx 10^3$ is too large for conventional minimization methods.

A Monte-Carlo type search starting from a random X can be employed to find a configuration that yields the calculated scattering curve fitting the experimental data

Chacón, P. *et al.* (1998) *Biophys. J.* 74, 2760-2775.

Svergun, D.I. (1999) *Biophys. J.* 76, 2879-2886

Why/how do *ab initio* methods work

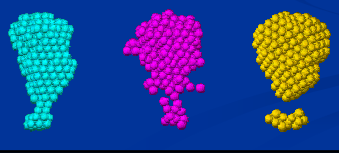
The 3D model is required not only to fit the data but also to fulfill (often stringent) physical and/or biochemical constrains

Ab initio program DAMMIN

Using simulated annealing, finds a compact dummy atoms configuration X that fits the scattering data by minimizing

$$f(X) = \chi^2[I_{exp}(s), I(s,X)] + \alpha P(X)$$

where χ is the discrepancy between the experimental and calculated curves, *P(X)* is the penalty to ensure compactness and connectivity, $\alpha>0$ its weight.



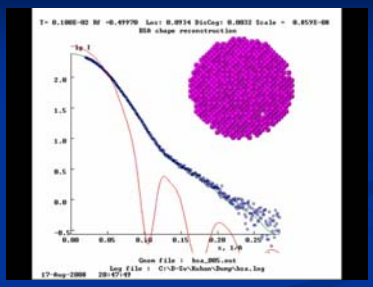
compact

loose

disconnected

A test *ab initio* shape determination run

Program DAMMIN

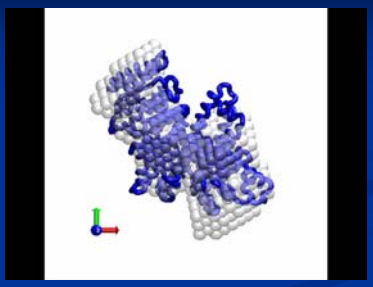


Slow mode

Bovine serum albumin, molecular mass 66 kDa, no symmetry imposed

A test *ab initio* shape determination run

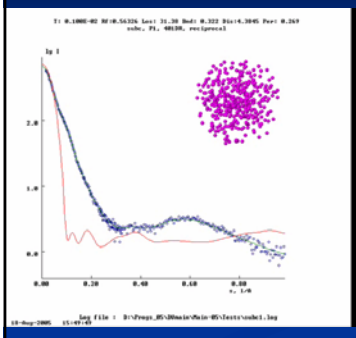
Program DAMMIN



Slow mode

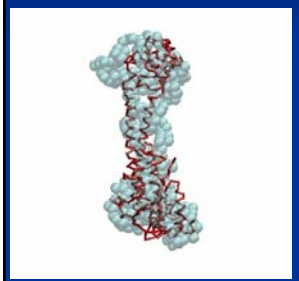
Bovine serum albumin: comparison of the *ab initio* model with the crystal structure of human serum albumin

GASBOR run on C subunit of V-ATPase



Starting from a random "gas" of 401 dummy residues, fits the data by a locally chain-compatible model

GASBOR run on C subunit of V-ATPase



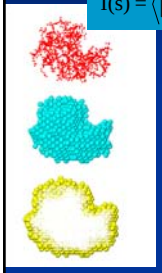
Beads: Ambruster *et al.* (2004, June) *FEBS Lett.* **570,** 119

C_α trace: Drory *et al.* (2004, November), *EMBO reports,* **5**, 1148

Joint use of SAS with high resolution methods

Computation of scattering from atomic structures

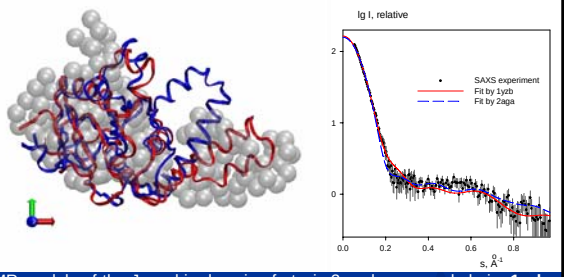
$$I(s) = \left\langle |A(s)|^2 \right\rangle_\Omega = \left\langle |A_a(s) - \rho_s A_s(s) + \delta\rho_b A_b(s)|^2 \right\rangle_\Omega$$



- $A_a(s)$: atomic scattering in vacuum
- $A_s(s)$: scattering from the excluded volume
- $A_b(s)$: scattering from the hydration shell

***CRYSOL* (X-rays):** Svergun et al. (1995). *J. Appl. Cryst.* **28**, 768
***CRYSON* (neutrons):** Svergun et al. (1998) *P.N.A.S. USA*, **95**, 2267

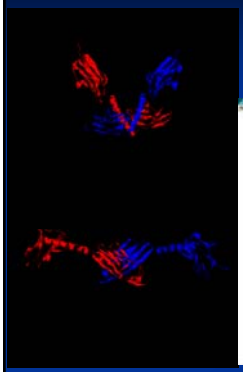
Validation of high resolution models



NMR models of the Josephin domain of ataxin-3: red curve and chain: **1yzb**, Nicastro et al. (2005) *PNAS USA* **102**, 10493; blue curve and chain: **2aga**, Mao et al. (2005) *PNAS USA* **102**, 12700.

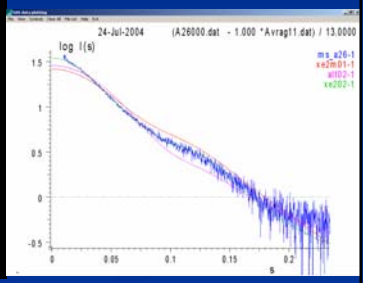
Nicastro, G., Habeck, M., Masino, L., Svergun, D.I. & Pastore, A. (2006) *J. Biomol. NMR*, **36**, 267.

Identification of biologically active oligomers



Biologically active dimer of myomesin-1

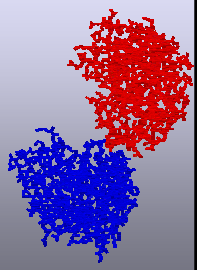
Experiment started: 24-07-2004 at 21:09
Final result obtained: 24-07-2004 at 21:48



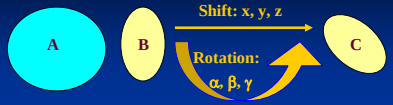
Pinotsis, N., Lange, S., Perriard, J.-C., Svergun, D.I. & Wilmanns, M. (2008) *EMBO J.* **27**, 253-264

Rigid body refinement

- ‘Post-genomic’ era brought unprecedented amount of structural information about individual macromolecules
- Large macromolecular complexes are more difficult to study by high resolution methods
- High resolution models of subunits can be used to model the quaternary structure of complexes based on low resolution methods



Principle of rigid body modelling



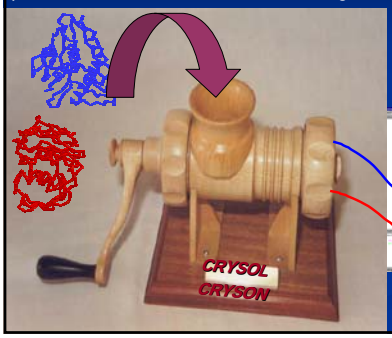
Using spherical harmonics, the amplitude(s) of arbitrarily rotated and displaced subunit(s) are analytically expressed *via* the initial amplitude and the six positional parameters: $C_{lm}(s) = C_{lm}(B_{lm}, \alpha, \beta, \gamma, x, y, z)$.
The scattering from the complex is then rapidly calculated as

$$I(s) = I_A(s) + I_B(s) + 4\pi^2 \sum_0^\infty \sum_{-l}^{l} \mathrm{Re}\left[A_{lm}(s) C_{lm}^*(s)\right]$$

Svergun, D.I. (1991). *J. Appl. Cryst.* **24**, 485-492

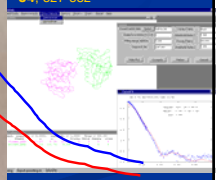
Interactive and local refinement

Scattering amplitudes of the subunits are pre-computed and positional parameters are refined to fit the scattering from the complex



Kozin & Svergun (2000). *J. Appl. Cryst.* **33**, 775-777

Konarev, Petoukhov & Svergun (2001). *J. Appl. Cryst.* **34**, 527-532

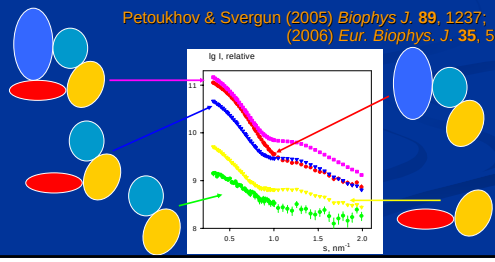


- ***MASSHA*** (Windows PC)
- ***ASSA*** (SUN/SGI/DEC)

Global rigid body modelling (SASREF)

- Fits (multiple X-ray and neutron) scattering curve(s) from partial constructs or contrast variation using simulated annealing
- Requires models of subunits, builds interconnected models without steric clashes
- Uses constraints: symmetry, distance (FRET or mutagenesis) relative orientation (RDC from NMR), if applicable

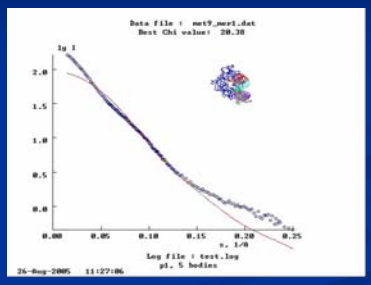
Petoukhov & Svergun (2005) *Biophys J.* **89**, 1237; (2006) *Eur. Biophys. J.* **35**, 567.



A global refinement run with distance constraints

A tyrosine kinase MET (118 kDa) consisting of five domains

Program SASREF



Single curve fitting with distance constraints: C to N termini contacts

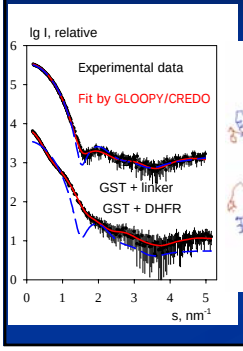
Gherardi, E., Sandin, S., Petoukhov, M.V., Finch, J., Youles, M.E., Ofverstedt, L.G., Miguel, R.N., Blundell, T.L., Vande Woude, G.F., Skoglund, U. & Svergun, D.I. (2006) *PNAS USA*, **103**, 4046.

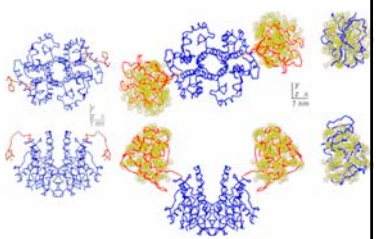
Addition of missing fragments

- Flexible loops or domains are often not resolved in high resolution models or genetically removed to facilitate crystallization
- Tentative configuration of such fragments are reconstructed by fixing the known portion and adding the missing parts to fit the scattering from the full-length macromolecule.
- Moreover, the domains or subunits can be moved as rigid bodies (BUNCH)

Addition of missing loops and fragments

- Dimeric fusion protein: GST+DHFR
  - ✓ addition of a 21 residue long linker
  - ✓ addition of the DHFR domain (159 residues

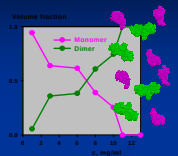




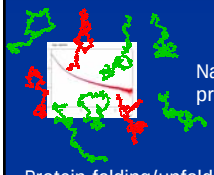
Petoukhov, M.V., Eady, N.A.J., Brown, K.A. & Svergun, D.I. (2002) *Biophys. J.* **83**, 3113

Life is more complicated: mixtures and processes

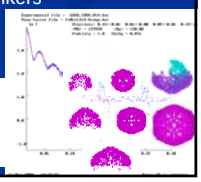
Equilibrium oligomeric mixtures



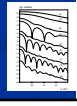
Stoichiometry and complex formation



Natively unfolded proteins and multidomain proteins with flexible linkers



Protein folding/unfolding kinetics



Assembly/disassembly processes

Scattering from mixtures

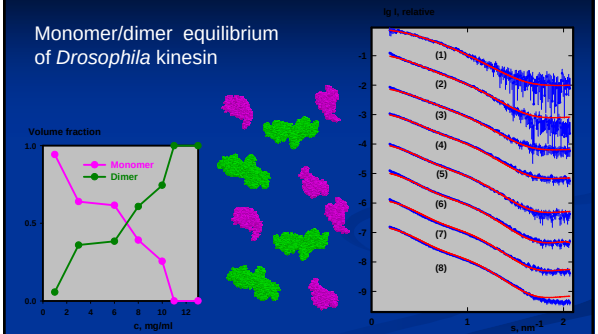
$$I(s)=\sum_k v_k I_k(s)$$

The scattering is proportional to that of a single particle averaged over all orientations, which allows one to determine the number of components and, given their scattering intensities $I_k(s)$, also the volume fractions

For equilibrium and non-equilibrium mixtures, solution scattering permits to determine size, shape and internal structure of the particle at low (1-10 *nm*) resolution.

Oligomer content in mixtures

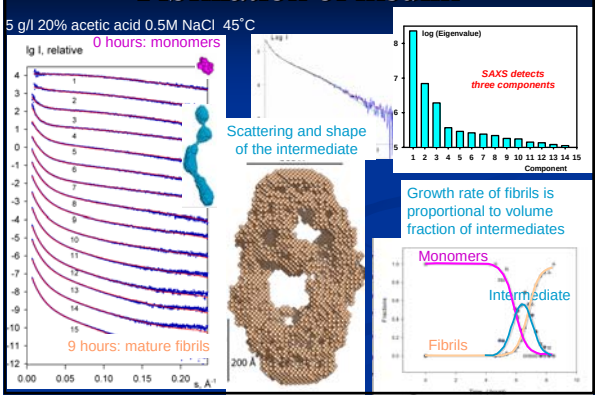
Monomer/dimer equilibrium of *Drosophila* kinesin



Kozielski, F., Svergun, D.I., Zaccai, J. Wade, R.H. & Koch, M.H.J. (2001) *J. Biol. Chem.* **276**, 1267

Fibrillation of insulin

5 g/l 20% acetic acid 0.5M NaCl 45˚C



Fibrillation of insulin

Oligomers are fibrillation nuclei and potential targets against amyloidosis

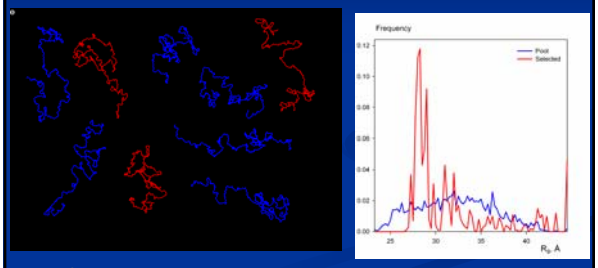
Assembly of protofilaments from the helical precursors (5-6 units)

Formation of mature fibrils from intertwinning protofilaments

Vestergaard, B., Groenning, M., Roessle, M., Kastrup, J.S., de Weert, M.V., Flink, J.M., Frokjaer, S., Gajhede, M. & Svergun, D.I. (2007) *PLoS Biol.* **5**, e134

Characterization of flexible systems: Ensemble Optimization Method (EOM)

Flexible systems, e.g. **intrinsically disordered proteins or multidomain proteins with flexible linkers** are analyzed not in terms of a single structure, but as an ensemble of structures. These are selected from a large initial pool by a genetic algorithm.

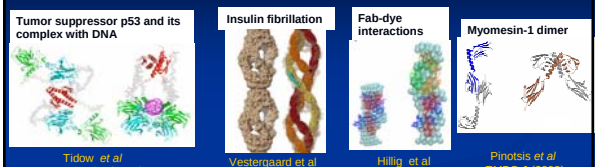


Bernadó, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., & Svergun, D. I. (2007) *J. Am. Chem. Soc.* **129**, 5656-5664.

Recent SAXS applications at X33

Complexes and assemblies

Domain and quaternary structure



Tidow *et al* PNAS USA (2007)

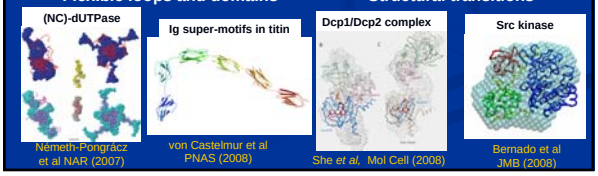
Vestergaard et al PLoS Biol (2007)

Hillig et al JMB (2008)

Pinotsis *et al* EMBO J (2008)

Flexible loops and domains

Structural transitions



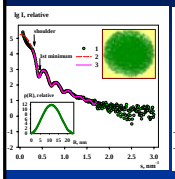
Németh-Pongrácz et al NAR (2007)

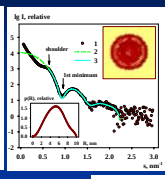
von Castelmur et al PNAS (2008)

She *et al*, Mol Cell (2008)

Bernado et al JMB (2008)

Magnetic Iron Oxide Nanoparticles Encapsulated by Phospholipids with PEG Tails

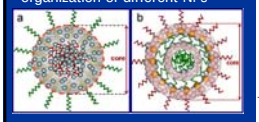


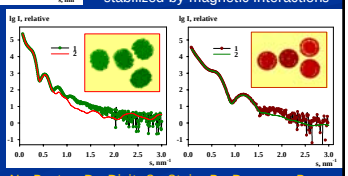


Highly monodisperse NPs are prepared by thermal decomposition of iron compounds including oxygen-containing ligands in boiling surfactants. The NPs must be coated to become soluble.

Rigid body analysis reveals equilibrium clusters of the NPs stabilized by magnetic interactions

Ab initio analysis: peculiarities of organization of different NPs





Shtykova, E.V, Huang, X., Remmes, N., Baxter, D., Dixit, S., Stein, B., Dragnea, B., Svergun, D. I. & Bronstein, L. M. (2007) J. Phys. Chem. C. **111**, 18078-18086

Current hard/software developments at X33

A novel 500K PILATUS pixel X-ray detector (PSI, Villigen)

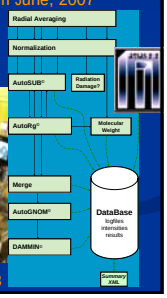
In user operation on X33 from 22.11.2007

Automated sample changer to run the beamline in a 'high throughput' mode

In user operation on X33 from 09.09.2007

Automated data analysis and Web access

Available for the users from June, 2007



All developments are also pilot projects for BioSAXS@Petra-3

What does SAXS tell about biological macromolecules

High throughput SAXS

- Nothing known: *ab initio* low resolution structure
- Incomplete high resolution structure known: probable configuration of missing portions
- Complete high resolution structure known: validation in solution and biologically active oligomers
- High resolution structure of domains/subunits known: quaternary structure by rigid body refinement
- Mixtures/assemblies: volume fractions of components

Synchrotron SAXS: from 1-2 mg purified material, concentration from 0.5-1 mg/ml, exposure times a few seconds/minutes

Acknowledgments

EMBL-Hamburg: D.Franke, T.Ikonen, A. Kikhney, P. Konarev, E. Mylonas, M. Petoukhov, M. Roessle, A. Round, P. Bernado
Institute of Crystallography, Moscow:
M. Kozin, E. Shtykova, A. Sokolova, V. Volkov
http://www.embl-hamburg.de/ExternalInfo/Research/Sax

Collaborative projects at the X33 beamline:

- **V-ATPase C-subunit:** G.Grueber (Saarland University, Homburg)
- **Importins/exportins:** E.Conti (EMBL, Heidelberg), P.Timmins (ILL, Grenoble)
- **Josephin:** A.Pastore (National Institute for Medical Research, London)
- **MET/HGF:** E.Gherardi (MRC Centre, Cambridge)
- **Myomesin-1:** M.Wilmanns (EMBL, Hamburg)
- **PrrB:** P. Tucker (EMBL, Hamburg)
- **p53:** A. Fersht (MRC Centre, Cambridge)
- **Fab:** R. Hillig (Bayer Schering Pharma AG, Berlin)
- **Insulin:** B.Vestergaard (Pharmaceutical Uni Copenhagen)
- **Titin:** O. Mayans (University of Basel)
- **Dcp1/2:** H.Song (Institute of Molecular and Cell Biology, Singapore)
- **Src kinase:** P. Bernado (Institute for Research in Biomedicine, Barcelona)
- **Nanoparticles:** L. Bronstein (Indiana University, Bloomington)

Internet SAS resources

- **ATSAS Home page:**
  http://www.embl-hamburg.de/ExternalInfo/Research/Sax/software.html
- **A textbook on SAS:**

Feigin, L.A. & Svergun, D.I. (1987) Structure analysis by small-angle X-ray and neutron scattering. New York: Plenum Press, 335 pp.

is available for download from

http://www.embl-hamburg.de/ExternalInfo/Research/Sax/reprints/feigin_svergun_1987.pdf

- **A forum on ATSAS programs:**
  www.saxier.org/forum

The PLATON Toolbox

Ton Spek
National Single Crystal
Service Facility,
Utrecht University,
The Netherlands.

Kyoto, 20-Aug-2008

Overview of the Talk

1. What is PLATON?
2. PLATON Tools (General)
3. Selected Examples/Details on:
   1. ADDSYM
   2. TWINNING DETECTION
   3. VOID DETECTION & SQUEEZE

What is PLATON

- PLATON is a collection of tools for single crystal structure analysis bundled within a single SHELX compatible program.
- reads/writes .ins, .res, .hkl, .cif, .fcf
- The tools are either unique to the program or adapted and extended versions of existing tools.
- The program was/is developed over of period of nearly 30 years in the context of and the needs of our National Single Crystal Service Facility in the Netherlands.

DESIGN HISTORY

- PLATON started out as a program for the automatic generation of an extensive molecular geometry analysis report for the clients of our service.
- Soon molecular graphics functionality was added (ORTEP)
- Over time many tools were included, many of which also require the reflection data.

DESIGN FEATURES

- As hardware independent as possible
- Limited dependence on external libraries
- Single routine for all graphics calls
- Single routine for all symmetry handling
- Sharing of the numerical routines by the various tools
- Single Fortran source, simple compilation
- Small C routine for interface to X11 graphics

PLATON USAGE

- Today, PLATON functionality is most widely used in its validation incarnation as part of the IUCr checkCIF facility.
- Tools are available in PLATON to analyze and solve the issues that are reported to need attention.
- PLATFORMS:
  UNIX/LINUX, MAC-OSX, MS-WINDOWS

Selected Tools

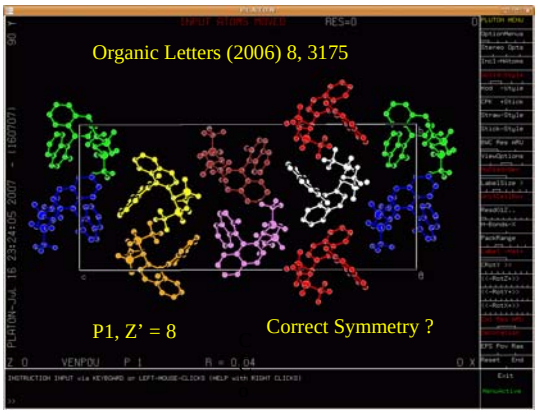
- ADDSYM – Detection and Handling of Missed (Pseudo)Symmetry
- TwinRotMat – Detection of Twinning
- SOLV – Report on Solvent Accessible Voids
- SQUEEZE – Handling of Disordered Solvents in Least Squares Refinement (Easy to use Alternative for Clever Disorder Modelling)
- BijvoetPair – Post-refinement Absolute Structure Determination (Alternative for Flack x)
- VALIDATION – PART of IUCr CHECKCIF
- ORTEP & PLUTON – Molecular Graphics
- CONTOUR – Contoured Fourier Maps

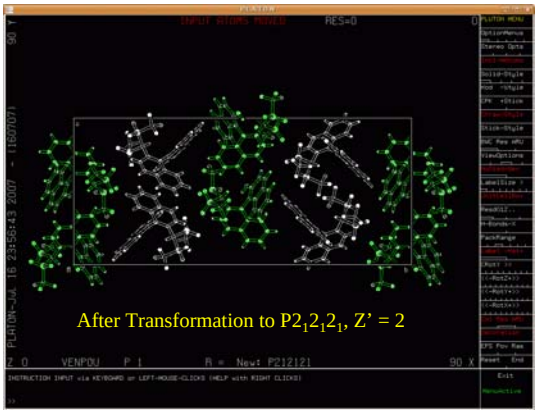
OTHER PLATON USAGE

- PLATON also offers guided/automatic structure determination and refinement tools for routine structure analyses from scratch (i.e. the 'Unix-only' SYSTEM S tool and the new FLIPPER/STRUCTURE tool that is based on the Charge Flipping Ab initio phasing method).
- Next Slide: Main Function Menu ➔

ADDSYM

- Often, a structure solves only in a space group with lower symmetry than the correct space group. The structure should subsequently be checked for higher symmetry.
- About 1% of the 2006 & 2007 entries in the CSD need a change of space group.
- E.g. A structure solves only in P1. ADDSYM is a tool to come up with the proper space group and to carry out the transformation (➔ new .res)
- Next slide: Recent example of missed symmetry



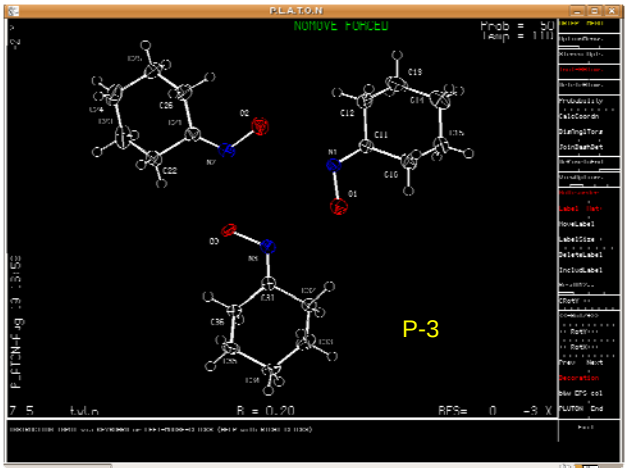


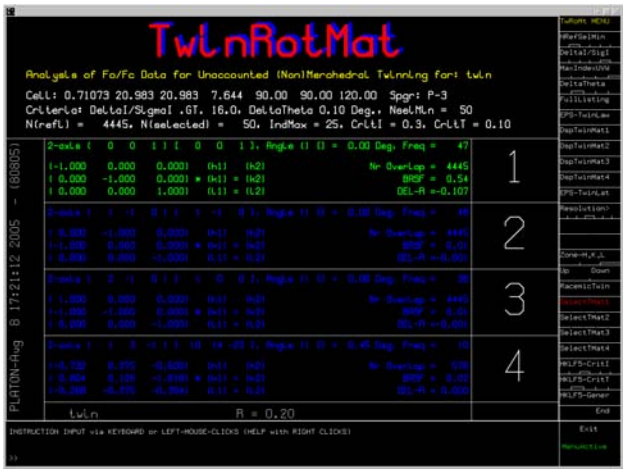
(Pseudo)Merohedral Twinning

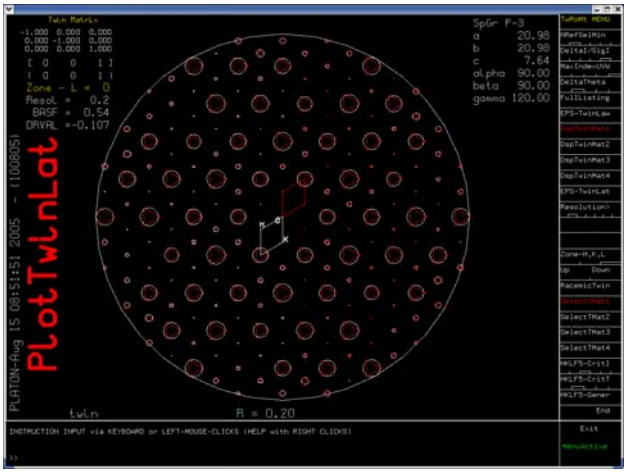
- Options to handle twinning in L.S. refinement available in SHELXL, CRYSTALS etc.
- Problem: Determination of the Twin Law that is in effect.
- Partial solution: coset decomposition, try all possibilities (I.e. all symmetry operations of the lattice but not of the structure)
- ROTAX (S.Parson et al. (2002) J. Appl. Cryst., 35, 168. (Based on the analysis of poorly fitting reflections of the type F(obs) >> F(calc))
- TwinRotMat Automatic Twinning Analysis as implemented in PLATON (Based on a similar analysis but implemented differently)

TwinRotMat Example

- Originally published as disordered in P3.
- Correct Solution and Refinement in the trigonal space group P-3 ➔R= 20%.
- Run PLATON/TwinRotMat on CIF/FCF
- Result: Twin law with an the estimate of the twinning fraction and the estimated drop in R-value
- Example of a Merohedral Twin ➔



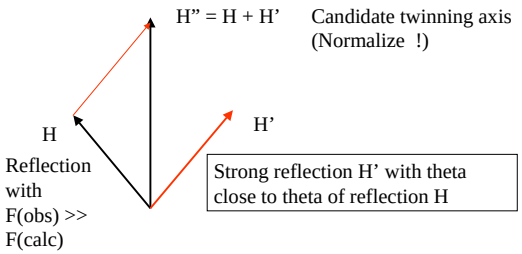




Ideas behind the Algorithm

- Reflections effected by twinning show-up in the least-squares refinement with F(obs) >> F(calc)
- Overlapping reflections necessarily have the same Theta value within a certain tolerance.
- Generate a list of implied possible twin axes based on the above observations.
- Test each proposed twin law for its effect on R.

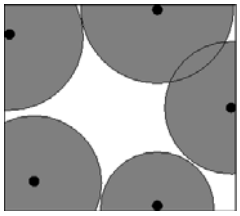
Possible Twin Axis



Solvent Accessible Voids

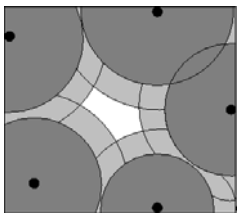
- A typical crystal structure has only in the order of 65% of the available space filled.
- The remainder volume is in voids (cusps) in-between atoms (too small to accommodate an H-atom)
- Solvent accessible voids can be defined as regions in the structure that can accommodate at least a sphere with radius 1.2 Angstrom without intersecting with any of the van der Waals spheres assigned to each atom in the structure.
- Next Slide: Void Algorithm: Cartoon Style ➔

DEFINE SOLVENT ACCESSIBLE VOID



STEP #1 – EXCLUDE VOLUME INSIDE THE VAN DER WAALS SPHERE

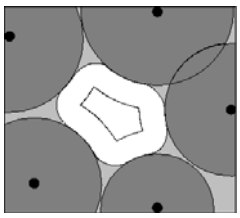
DEFINE SOLVENT ACCESSIBLE VOID



White Area:
Ohashi Volume.
Location of possible
Aton centres

STEP # 2 – EXCLUDE AN ACCESS RADIAL VOLUME TO FIND THE LOCATION OF ATOMS WITH THEIR CENTRE AT LEAST 1.2 ANGSTROM AWAY

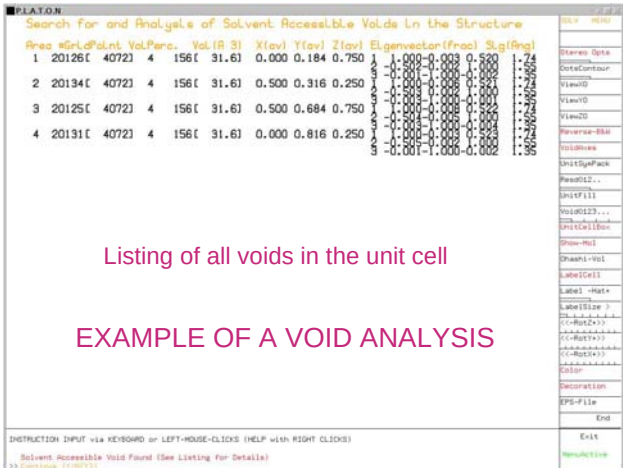
DEFINE SOLVENT ACCESSIBLE VOID



STEP # 3 – EXTEND INNER VOLUME WITH POINTS WITHIN 1.2 ANGSTROM FROM ITS OUTER BOUNDS

VOID SEARCH ALGORITHM

- Move a probe with radius 1.2 Ang over a fine (0.2 Ang) grid through the unit cell.
- Start a new void when a gridpoint is found that is at least 1.2 Ang outside the van der Waals surface of all atoms.
- Expand this void with connected gridpoints with the same property until completed.
- Find new starting gridpoint for the next void until completion.
- Expand the 'Ohashi' volumes with gridpoints within 1.2 Angstrom to surface gridpoints.



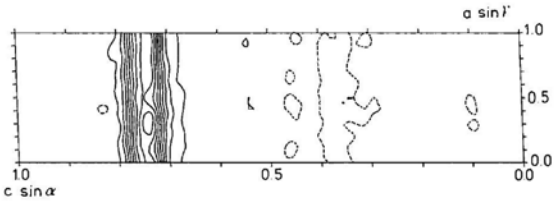
Listing of all voids in the unit cell

EXAMPLE OF A VOID ANALYSIS

VOID APPLICATIONS

- Calculation of Kitaigorodskii Packing Index
- Determination of the available space in solid state reactions (Ohashi)
- Determination of pore volumes, pore shapes and migration paths in microporous crystals
- As part of the SQUEEZE routine to handle the contribution of disordered solvents in a crystal structure.

Structure Modelling and Refinement Problem for Salazopyrine structure



Difference Fourier map shows disordered channels rather than maxima

How to handle this in the Refinement ?

SQUEEZE !

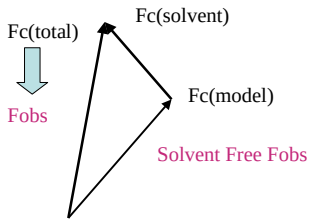
SQUEEZE

- Takes the contribution of disordered solvents to the calculated structure factors into account by back-Fourier transformation of density found in the 'solvent accessible volume' outside the ordered part of the structure (iterated).
- Filter: Input shelxl.res & shelxl.hkl
  Output: 'solvent free' shelxl.hkl
- Refine with SHELXL or Crystals
- Note:SHELXL lacks option for fixed contribution to Structure Factor Calculation.

SQUEEZE Algorithm

1. Calculate difference map (FFT)
2. Use the VOID-map as a mask on the FFT-map to set all density outside the VOID's to zero.
3. FFT^{-1} this masked Difference map -> contribution of the disordered solvent to the structure factors
4. Calculate an improved difference map with F(obs) phases based on F(calc) including the recovered solvent contribution and F(calc) without the solvent contribution.
5. Recycle to 2 until convergence.

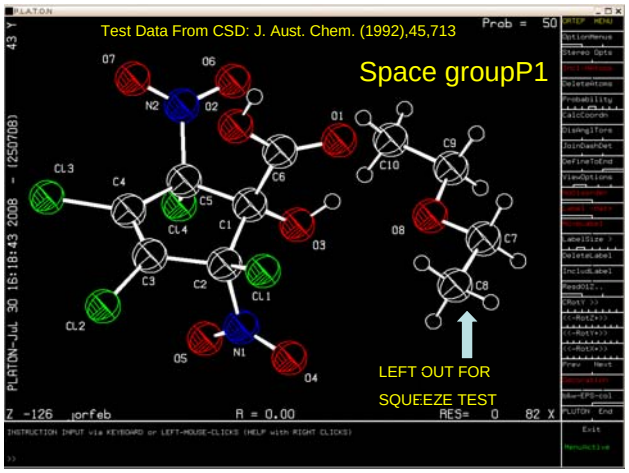
SQUEEZE
In the Complex Plane

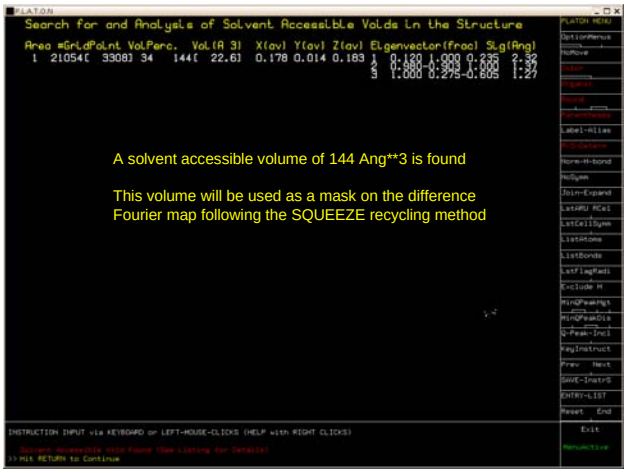


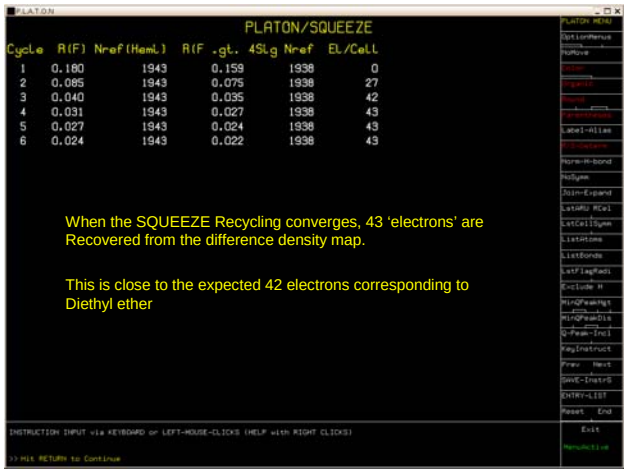
Black: Split Fc into a discrete and solvent contribution
Red: For SHELX refinement, temporarily substract recovered solvent contribution from Fobs.

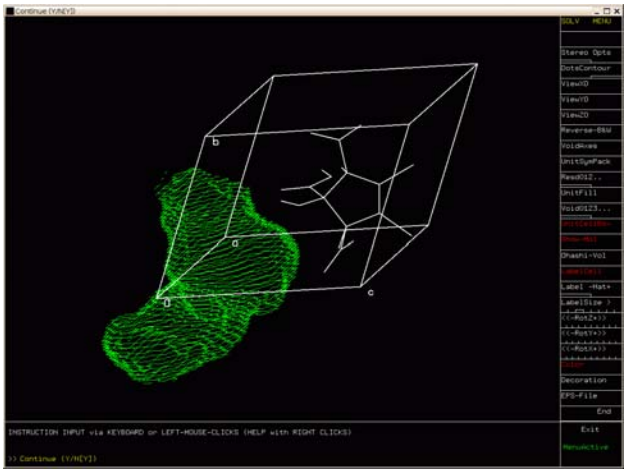
Comment

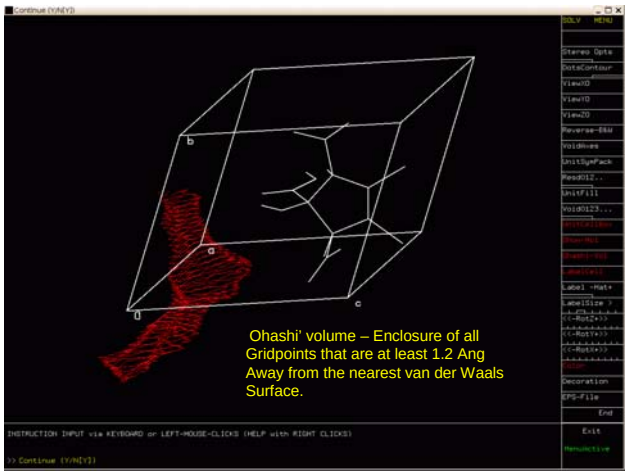
- The Void-map can also be used to count the number of electrons in the masked volume.
- A complete dataset is required for this feature.
- Ideally, the solvent contribution is taken into account as a fixed contribution in the Structure Factor calculation (CRYSTALS) otherwise it is substracted temporarily from F(obs)^2 (SHELXL) and re-instated afterwards with info saved beyond column 80 for the final Fo/Fc list.











Additional Info

http://www.cryst.chem.uu.nl

(including a copy of this powerpoint presentation)

Thanks

for your attention !!

Small Molecule Refinement

Small Molecule Toolbox

Small molecule field: multidimensional

- material: single crystal / powder
- property: electron density / magnetic moment
- symmetry
  - real space: periodic / aperiodic
  - how many Miller indices per reflection

How small is small?

- what matters is the asymmetric unit content
  - biggish small molecule (100 atoms)
  - 10 differently conformed copies in a.s.u.

  = 1000 atoms
- pretty much the upper bound

Solving a structure

- Solving in protein, aperiodic crystals, powder = from hard to nearly impossible
- For small molecule, most often trivial:

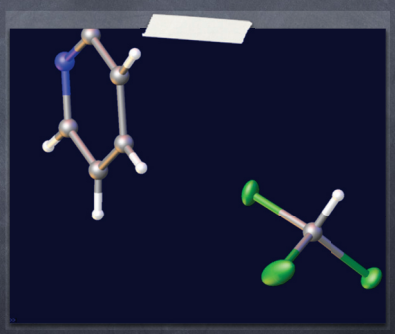
  1. direct methods
  2. drop inversion centre, then direct methods
  3. [rare] Dual-space methods, charge flipping

Electron density model (mainstream)

- site, harmonic displacements, s.o.f.
- isotropic scattering factor shared by all atoms of the same type
- frequent use of constraints:
  - s.o.f. in model of disordered parts
  - special positions, idealised Hydrogen atoms

Disorder

- Cl_3CH is in one position in a fraction x of the unit cells
- in another position in a fraction 1-x
- model as a superposition with s.o.f. x and 1-x
- equality constraint on ADP of corresponding atoms



Electron density model (fringes)

- anharmonic displacements:
  - toroidal distribution for $-CH_3$ and the like
  - Gram-Charlier (generic)
- scattering factors = refined multipolar expansion, one per scatterer

charge density

Refinement (minimisation)

- always least-squares
- always full matrix (Gauss-Newton)
- weighting schemes: $T = \Sigma\ w(h)\ (\ K\ F_c(h) - F_o(h)\ )^2$
  - $w = f(\ F_c,\ F_o,\ \sigma_{F_o}\ ;\ a,\ b,\ c,\ \ldots)$
  - adjust a,b,c,... s.t. $w\ (\ K\ F_c - F_o\ )^2$ does not show any trend w.r.t. F_c

Mixed optimisation

- continuous parameters: sites, ADP's, etc
- discrete parameters: atom types
- consequence: for most of a refinement, optimum continuous parameters = rubbish (moreover: only part of the structure)
- thus: not much need for sophisticated continuous optimisation techniques

Gauss-Newton

- small residuals => good Hessian approximation computable from ∇F_c at each step => Newton

$\Sigma\ \nabla F_c(h)\ \nabla F_c(h)^T\ \delta = -\Sigma\ \nabla F_c(h)\ (K\ F_c(h) - F_o(h))$

Normal matrix

- Stabilisation
  - add $-\lambda\ \delta$ to the l.h.s. before solving
  - shift limiting restraint: only apply $\lambda\ \delta$ to parameters instead of δ after solving

Floating origin problem

- the normal matrix is singular iff there is a shift δ s.t. $\nabla F_c(h)\ .\ \delta = 0$ for every h
  => can't solve normal equations
- axial space-groups: each scatterer shifted by the same amount along one of the axis
  => F_c does not change
- best solution: restraint to nail down the barycentre of the sites x_s => + $(\Sigma\ w_s\ x_s)^2$

Runtime costs

- n scatterers, p parameters, m reflections

$O(n \times m)$ compute $F_c(h)$ and $\nabla F_c(h)$ for each h

$O(p^2 \times m)$ accumulate normal matrix $\Sigma\ \nabla F_c(h)\ \nabla F_c(h)^T$

$O(p^3)$ solve for shifts (Cholesky)

Runtime costs

- n scatterers, p parameters, m reflections
- O(n×m) compute $F_c(h)$ and $\nabla F_c(h)$ for each h
- O(p×m) accumulate normal matrix $\Sigma \nabla F_c(h) \nabla F_c(h)^T$
- $< p^2$ solve for shifts (Cholesky)
- sparsity: $\Sigma \partial_i F_c(h) \partial_j F_c(h)$ negligeable unless i,j are parameters from 2 nearby scatterers

Runtime costs

- n scatterers, p parameters, m reflections
- compute $F_c(h)$ and $\nabla F_c(h)$ for each h
- FFT computation: O(n+m log m)
- O(p×m) accumulate normal matrix $\Sigma \nabla F_c(h) \nabla F_c(h)^T$
- $< p^2$ solve for shifts (Cholesky)
- sparsity: $\Sigma \partial_i F_c(h) \partial_j F_c(h)$ negligeable unless i,j are parameters from 2 nearby scatterers

Runtime costs

- n scatterers, p parameters, m reflections

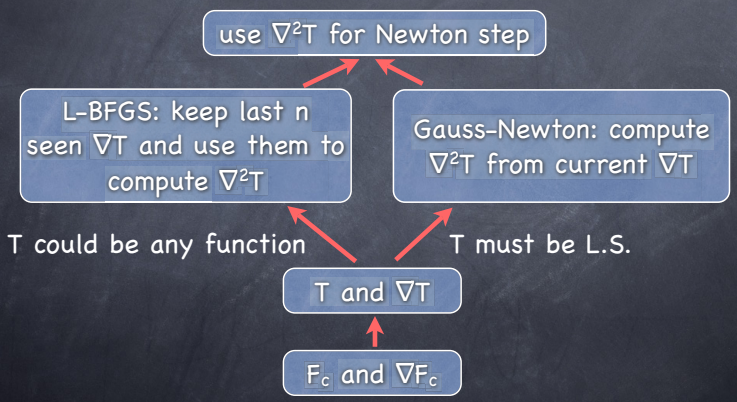
Jelsch C., Acta Cryst., 2001, A57, 558.
Crystals implementation
(D. Watkin, R. Cooper, S. Pantos)

- solve for shifts (Cholesky)
- sparsity: $\Sigma \partial_i F_c(h) \partial_j F_c(h)$ negligeable unless i,j are parameters from 2 nearby scatterers

Why always L.S. with Gauss-Newton?

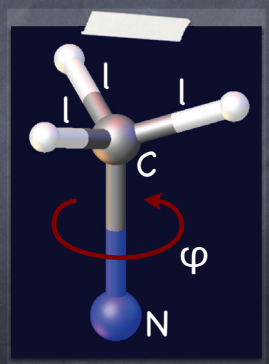
- statistical property of L.S.:
  normal matrix inverse = variance-covariance
  - diagonal terms -> e.s.d.
  - non-diagonal terms -> correlations
  - all necessary for publications
- But then normal matrix only needed once at the very end of refinement

Gauss-Newton vs L-BFGS



-CH3

- all lengths C-H_i equal
- all angles N-C-H_i and H_i-C-H_j tetrahedral
- H sites (9 parameters): function of x_N, x_C, φ, l



Constrained Minimisation

- min T(X)
- subject to $f_1(X) = 0$, $f_2(X) = 0$, ...
- 3 + 6 such constraints for -CH3:
  - $||x_H - x_C|| - l = 0$, ...
  - $(x_{H1} - x_C).(x_{H2} - x_C) - l^2 \cos 109.4° = 0$, ...

Quadratic penalties (60's)

- min T(X) s.t. $f_1(X) = 0$, $f_2(X) = 0$, ...
- going back to unconstrained:

  min $T(X) + \Sigma\ \mu\ f_i(X)^2$
- as $\mu \to \infty$, $f_i(X) = 0$ gets better enforced
- but e.g. LBFGS ill-conditioned for too large μ

Modern developments

- Augmented Lagrangian (70's till 90's):

  min $T(X) - \Sigma\ \lambda_i\ f_i(X) + \Sigma\ \mu\ f_i(X)^2$
- Sequential Quadratic Programming (SQP)

  king of the hill since 2000
- J. Nocedal and S. J. Wright. Numerical Optimization. Springer, 1999

Constraints in crystallography

- methods seen so far:
  - constrained -> unconstrained
  - work the same for any constraint
  - ~~crystallography~~
- so what do crystallographer do then? ch3

Reparametrization (-CH3)

- before: $Fc(X) = Fc(x_{H1}, x_{H2}, x_{H3}, x_C, x_N, ...)$
- new parameters l, φ that x_{H_i} is function of
- after: Fc(X) function of Y = (l, φ x_C, x_N, ...)

  $Fc(X) = Fr(Y)$
- new derivatives: chain rule

Code

1. start from some value of l, φ, x_C, x_N, ...
2. place Hydrogen's, i.e. compute x_{H1}, x_{H2}, x_{H3}
3. compute $Fc(x_{H1}, x_{H2}, x_{H3}, x_C, x_N, ...)$ and its gradient [cctbx]
4. compute gradient of Fr(l, φ, x_C, x_N, ...)
5. pass that gradient to minimiser
6. minimiser returns shifts δl, δφ, δx_C, δx_N, ... bringing us closer to minimum
7. update parameters: φ += δφ, x_C += δx_C, ...
8. Loop back to 2.

Code

1. start from some value of l, φ, x_C, x_N, ...
2. place Hydrogen's, i.e. compute x_{H1} , x_{H2} , x_{H3}
3. compute Fc(x_{H1} , x_{H2} , x_{H3}, x_C, x_N, ...) and its gradient

geometry distortion

4. compute g
5. pass that
6. minimiser bringing
7. update parameters: φ += δφ, x_C += δx_C, ...
8. Loop back to 2.

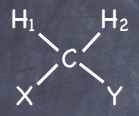
alternative

8. δl, δφ, δx_C ➡ δx_{H1} , δx_{H2} , δx_{H3} using derivatives of xs w.r.t. l, φ,... then loop back to 3.

Software

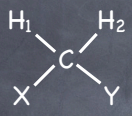
- mainstream: ShelXL, Crystals, ε
- ShelXL rulez the world
- they do the job
- monolithic Fortran 77 code only maintainable by their creator
- much knowledge buried in unreusable code

Example: secondary ideal CH_2



- ShelXL documentation says that
  - angle X-C-H_1 and Y-C-H_2 equal
  - angle H_1-C-H_2 depends on angle X-C-Y:

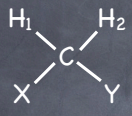
    latter wider => former narrower



H_1-C-H_2 = θ

X-C-Y = Φ

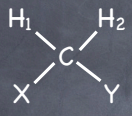
```
      U=-FB(2)-FB(7)
      V=-FB(3)-FB(8)
      W=-FB(4)-FB(9)
      Q=U**2+V**2+W**2
      IF(Q.LT.0.01)GOTO 79
      Q=R/SQRT(Q)
      IF(NA.NE.4)GOTO 106
      FB(17)=U*Q
      FB(18)=V*Q
      FB(19)=W*Q
      GOTO 113
106   XX=FB(2)-FB(7)
      YY=FB(3)-FB(8)
      ZZ=FB(4)-FB(9)
      S=XX**2+YY**2+ZZ**2
      IF(S.LT.0.01)GOTO 79
      X=V*ZZ-W*YY
      Y=W*XX-U*ZZ
      Z=U*YY-V*XX
      P=X**2+Y**2+Z**2
      IF(P.LT.0.01)GOTO 79
      S=1.0376-.0349*S
      Q=Q*COS(S)
      P=R*SIN(S)/SQRT(P)
      DO 107 NJ=17,22,5
        FB(NJ)=U*Q+X*P
        FB(NJ+1)=V*Q+Y*P
        FB(NJ+2)=W*Q+Z*P
      P=-P
```



H_1-C-H_2 = θ

X-C-Y = Φ

θ = 1.0376 - 0.0349×2(1-cos Φ)

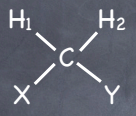


H_1-C-H_2 = θ

X-C-Y = Φ

normalised C->X = (FB(2), FB(3), FB(4))

normalised C->Y = (FB(7), FB(8), FB(9))



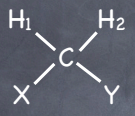
H_1-C-H_2 = θ

X-C-Y = Φ

```
      U=-FB(2)-FB(7)                    IF(S.LT.0.01)GOTO 79
      V=-FB(3)-FB(8)
      W=-FB(4)-FB(9)
      Q=U**2+V**2+W**
      IF(Q.LT.0.01)GOTO
      Q=R/SQRT(Q)
      IF(NA.NE.4)GOTO 106
      FB(17)=U*Q                        S=1.0376-.0349*S
      FB(18)=V*Q                        Q=Q*COS(S)
      FB(19)=W*Q                        P=R*SIN(S)/SQRT(P)
      GOTO 113                          DO 107 NJ=17,22,5
  106 XX=FB(2)-FB(7)                      FB(NJ)=U*Q+X*P
      YY=FB(3)-FB(8)                      FB(NJ+1)=V*Q+Y*P
      ZZ=FB(4)-FB(9)                      FB(NJ+2)=W*Q+Z*P
      S=XX**2+YY**2+ZZ**2                 P=-P
```

C->H_1 = (FB(17), FB(18), FB(19))

C->H_2 = (FB(22), FB(23), FB(24))



H_1-C-H_2 = θ

X-C-Y = Φ

```
      U=-FB(2)-FB(7)                    IF(S.LT.0.01)GOTO 79
      V=-FB(3)-FB(8)                    X=V*ZZ-W*YY
      W=-FB(4)-FB(9)                    Y=W*XX-U*ZZ
      Q=U**2+V**2+W**2                  Z=U*YY-V*XX
      IF(Q.LT.0.01)GOTO 79              P=X**2+Y**2+Z**2
      Q=R/SQRT(Q)                       IF(P.LT.0.01)GOTO 79
      IF(NA.NE.4)GOTO 106               S=1.0376-.0349*S
                                        Q=Q*COS(S)
                                        P=R*SIN(S)/SQRT(P)
      FB(19)=W*Q                        DO 107 NJ=17,22,5
      GOTO 113                            FB(NJ)=U*Q+X*P
  106 XX=FB(2)-FB(7)                      FB(NJ+1)=V*Q+Y*P
      YY=FB(3)-FB(8)                      FB(NJ+2)=W*Q+Z*P
      ZZ=FB(4)-FB(9)                      P=-P
      S=XX**2+YY**2+ZZ**2
```

distance d(C,H_i) = R

Example(2): toroidal displacements

```
SUBROUTINE XRING (..., M6RI, ...)
...
INCLUDE 'STORE.INC'
...
REFLH=STORE(M6RI)
REFLK=STORE(M6RI+1)
REFLL=STORE(M6RI+2)
...
C-C-C-CALCULATE FINAL FACTOR FOR RING s
...
```

Here goes the juicy meat!

Example(2): toroidal displacements

```
SUBROUTINE XRING (..., M6RI, ...)
...
INCLUDE 'STORE.INC'
...
REFLH=STORE(M6RI)
REFLK=STORE(M6RI+1)
REFLL=STORE(M6RI+2)
...
C-C-C-CALCULATE FINAL FACTOR FOR RING s
...
```

Crystals STORE

Here goes the juicy meat!

But can't be reused without the STORE

smtbx goals

- dig out gems from ShelXL and Crystals
- dig out useful algorithm from literature: e.g. hindered rotor
- don't reinvent the wheel => cctbx
  - reuse as much as possible (fixing and improving if need be)
  - contribute back any new feature

Acknowledgements

- EPSRC ("Age Concern" grant)
- University of Durham

Introduction to the charge flipping method

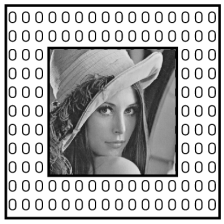
Gábor Oszlányi & András Sütő

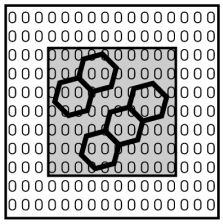
Research Institute for
Solid State Physics and Optics
Budapest, Hungary

- History
- Basic principle
- Main properties
- Algorithmic improvements
- Applications
- Use of known information
- User programs
- Crystallographic teaching

Non-periodic object: known support, surrounded by zero density

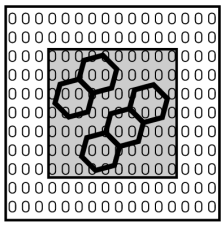
Non-periodic object: known support, surrounded by zero density

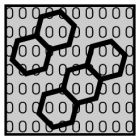




Sparse non-periodic object: support is much smaller than confining box

Non-periodic object: known support, surrounded by zero density





Periodic crystal: unknown support, zero density in the cell

History

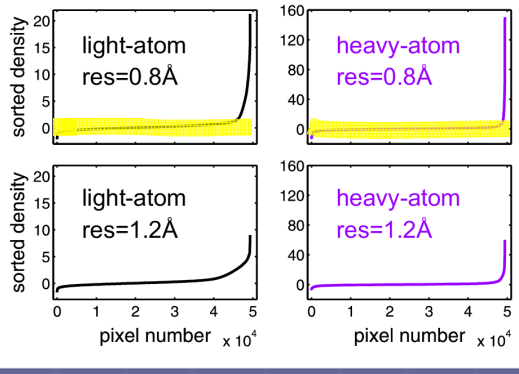
⇨ • Phase retrieval of non-periodic objects
Gerchberg-Saxton 1972 & Fienup 1982

- Low-density elimination
Shiono-Woolfson 1992
- Solvent flattening / Solvent flipping
Wang 1985, Abrahams-Leslie 1996
- Use of the space group P1
Sheldrick-Gould, Pavelcik, Burla et al.
- All dual-space methods
`SnB,ShelxD,ACORN,SIR,XLENS,Diff.Map` …

Principle

- The unit cell is mostly empty.
  (small oscillations at finite resolution)
- At sufficiently high resolution
  low electron density can be utilized
  to develop high (atomic) density.

  Instead of more constraints
  reduce the size of the search space
  by iterative perturbations.



True electron density samples sorted in ascending order

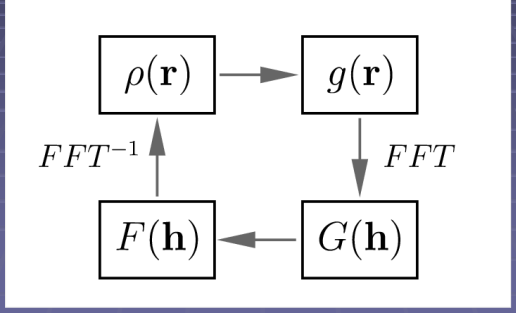
What is required ?

- Diffraction data
  - Type (preferably single crystal)
  - Dimension (1,2,3,…)
  - Resolution (<1Å)
  - Completeness (>90%)
- Electron density
  - Grid (0.2–0.4Å spacing)
  - Zero plateaus
  - Constraint of positivity

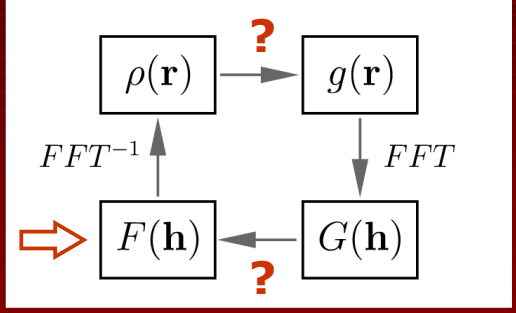
What is not required ?

- Atom types
- Chemical composition
- Data normalization (E's)
  (not required but useful)
- Space group symmetry
- Probabilistic phase relations
- Minimization of a cost function

Fourier recycling scheme



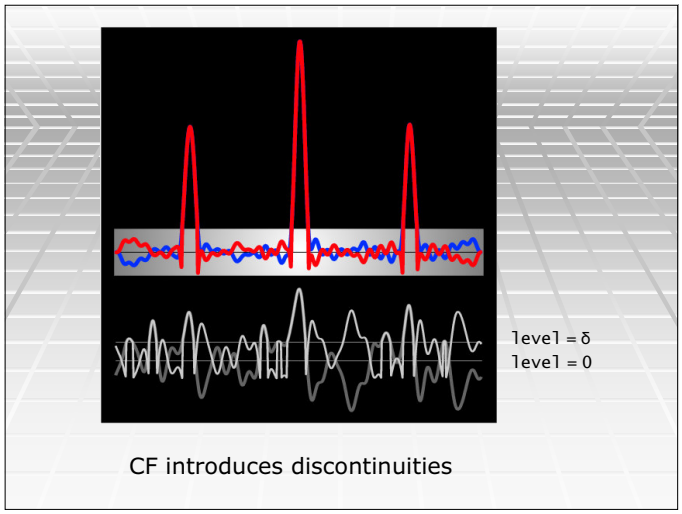
Variants of this scheme



The choice of charge flipping

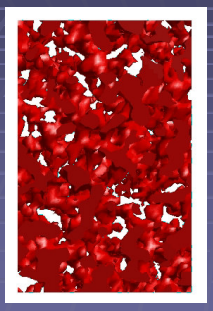
- Real space
  - if $\rho<\delta$ then $g=-\rho$ (flip)
  - if $\rho\geq\delta$ then $g=\rho$ (accept)
- Reciprocal space
  - if $h\leq H$ then $F=|F_{obs}|\cdot\exp(i\phi_G)$
  - if $h>H$ then $F=0$
  - if $h=0$ then $F=G$

Oszlányi & Sütő, Acta Cryst. A (2004)

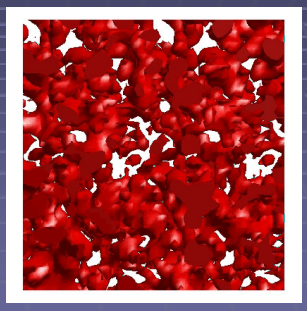


CF introduces discontinuities

Progress in real space



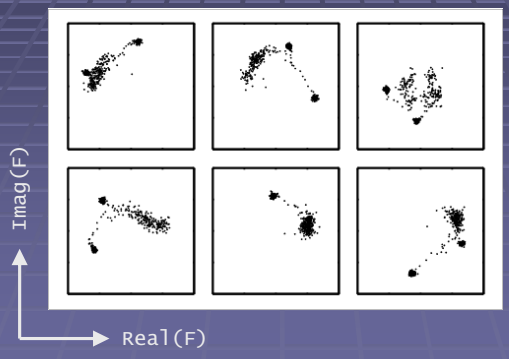
Pbca



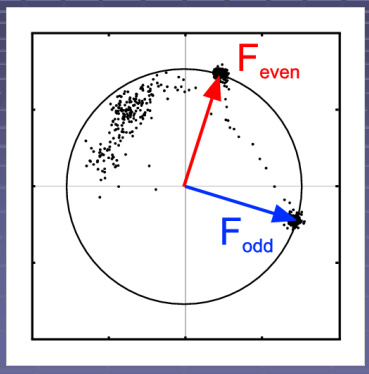
Ccmm

Progress in reciprocal space

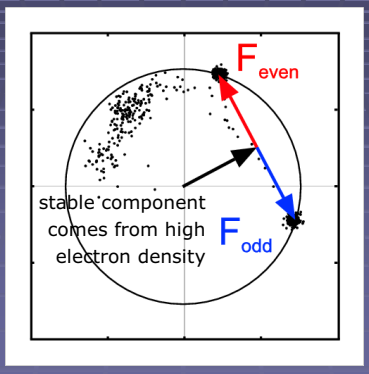
(evolution of strong reflections in the complex plane)



The solution is a limit cycle



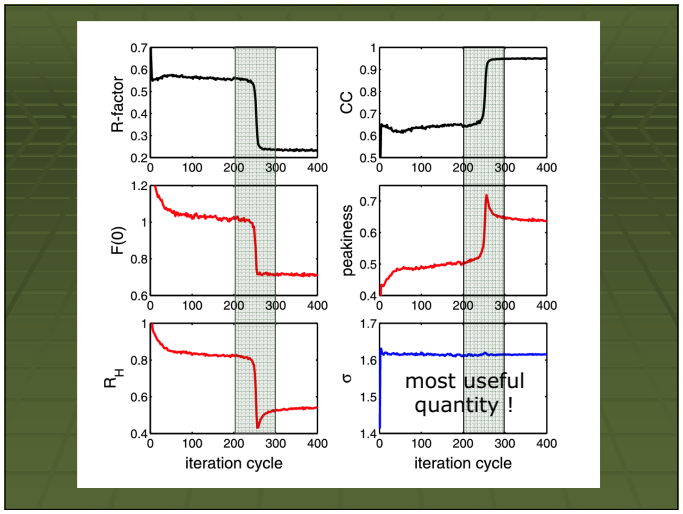
The solution is a limit cycle



When should we stop the iteration process?

Follow some indicators:

- R-factor
- CC correlation coefficient
- F(0) total charge
- $\sum\rho^3$ peakiness/skewness
- $R_H=\sum|sort(\rho)-sort(\rho_{ref})|/\sum|sort(\rho_{ref})|$

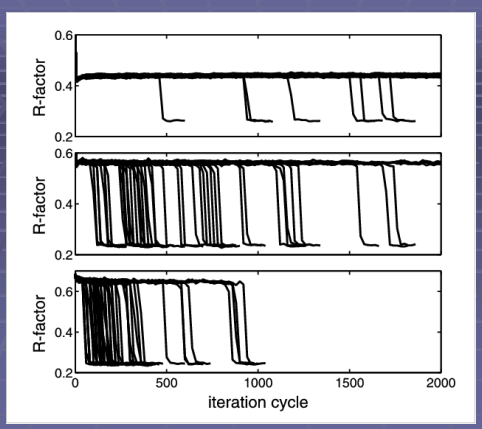


The (only) parameter δ

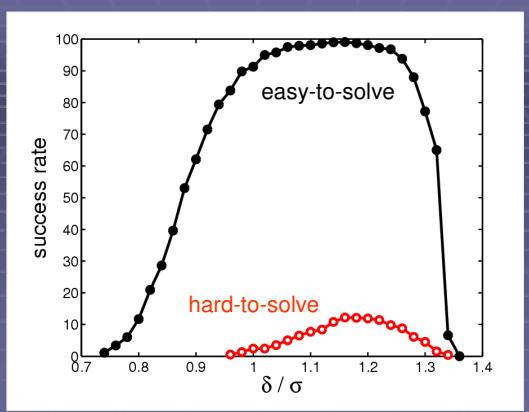
- Its choice is a delicate problem
  constant δ, variable Φ ($N_{\rho<\delta}/N_{grid}$)
  constant Φ, variable δ
  $\delta = k\cdot\sigma$, where k=1-1.2, σ=std(ρ)
- Determines the speed of a solution
  success rate = n/N, cost = $\sum x_i/n$
- Helps to improve the solution
  cleanup: 1. set ρ<1.5·δ to zero
  2. complete the cycle

The (only) parameter δ

- Its choice is a delicate problem
  constant δ, variable Φ ($N_{\rho<\delta}/N_{grid}$)
  constant Φ, variable δ
  $\delta = k\cdot\sigma$, where k=1-1.2, σ=std(ρ)
- Determines the speed of a solution
  success rate = n/N, cost = $\sum x_i/n$
- Helps to improve the solution
  cleanup: 1. set ρ<1.5·δ to zero
  2. complete the cycle



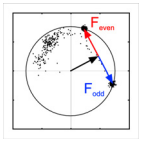
Different runs start from different random phase sets



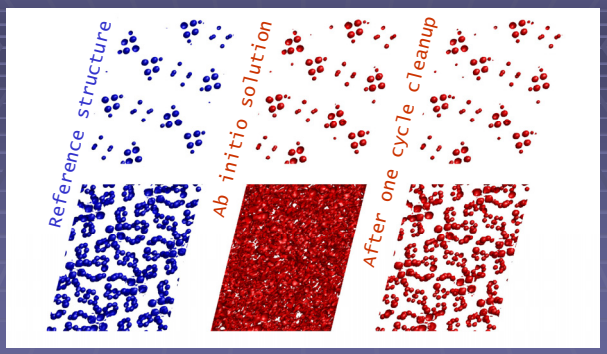
Statistics is needed to compare structures and algorithm variants. One point comes from 1000 structure solutions

The (only) parameter δ

- Its choice is a delicate problem
  constant δ, variable Φ ($N_{\rho<\delta} / N_{grid}$)
  constant Φ, variable δ
  $\delta = k \cdot \sigma$, where k=1-1.2, σ=std(ρ)
- Determines the speed of a solution
  success rate = n/N, cost = $\sum x_i / n$
- Helps to improve the solution
  1. set ρ<1.5·δ to zero
  2. complete the cycle



Isosurface at 1.0·σ (I,Re,P,Cl)



Isosurface at 0.2·σ (plus C,O)

Add more perturbations

and parameters ...

In reciprocal space

- weak reflections: set F = 0
- weak reflections: $\phi_F = \phi_G + \pi/2$ (shift)
  $|F| = |G|$ (float)
- all reflections: F_o+ΔF synthesis
  (positive feedback)
  $F = F_o + \beta \cdot (F_o - |G|) \cdot \exp(i\phi_G)$

In real space

- use memory: $g^{n+1} = g^n + \beta \cdot (\rho^n - \rho^{n-1})$
  (another positive feedback)

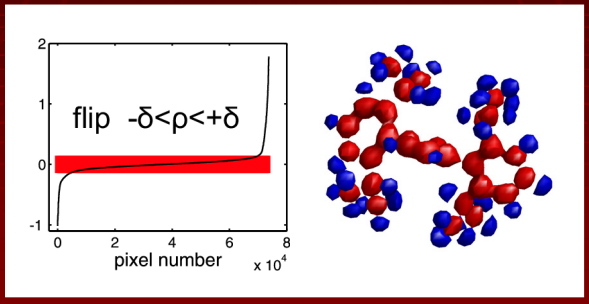
Efficiency of closely related algorithm variants differ by many orders of magnitude

	F at 0.8Å	E at 0.8Å	F at 1.0Å	E at 1.0Å
basic charge flipping	$> 8 \cdot 10^6$	32000	$> 8 \cdot 10^6$	600000
weak=0, α=0.20	$8 \cdot 10^6$	11000	$> 8 \cdot 10^6$	210000
weak=0, α=0.40	$4 \cdot 10^6$	1700	$> 8 \cdot 10^6$	32000
weak=0, α=0.60	$3 \cdot 10^6$	500	$8 \cdot 10^6$	5400
$\Delta\varphi = \pi/2$, α=0.20	190000	1500	$1 \cdot 10^6$	10000
$\Delta\varphi = \pi/2$, α=0.40	140000	1300	790000	2800
$\Delta\varphi = \pi/2$, α=0.60	210000	4400	$8 \cdot 10^6$	2000
$F_o + \Delta F$, w=0.25	59000	750	190000	3400
$F_o + \Delta F$, w=0.50	42000	1600	120000	900
$F_o + \Delta F$, w=∞	48000	2600	130000	650
flip-mem, β=0.6	$1 \cdot 10^6$	1400	$8 \cdot 10^6$	10000
flip-mem, β=0.8	220000	2500	$4 \cdot 10^6$	3400
flip-mem, β=1.0	40000	9700	460000	3100

Oszlányi & Sütő, Acta Cryst. A (2008)

Remove positivity constraint

(band flipping solves negative scattering density)



Oszlányi & Sütő, Acta Cryst. A (2007)

Previous schemes as constraints + perturbations

C_A := constraint of electron density
C_B := constraint of data

$\rho^{(n+1)} = C_B C_A \cdot \rho^{(n)}$ (susceptible to stagnation)

$\rho^{(n+1)} = C_B(C_A + \Delta_A) \cdot \rho^{(n)}$

$\rho^{(n+1)} = (C_B + \Delta_B)(C_A + \Delta_A) \cdot \rho^{(n)}$

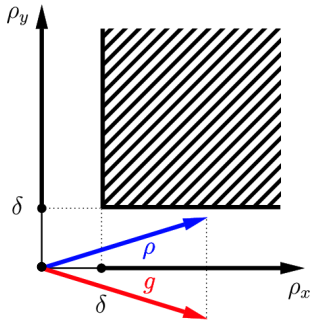
CF may increase the distance from the set of fixed points

("two-pixel phase problem")

$\rho = (\rho_x, \rho_y)$

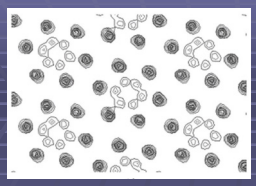
$\rho_x > \delta$ or $\rho_x = 0$

$\rho_y > \delta$ or $\rho_y = 0$



Applications

- Periodic crystals
  pseudosymmetry, unknown space group
- Aperiodic crystals
  modulated structures, quasicrystals
- Powder diffraction
  intensity repartitioning of overlapping reflections
- Macromolecules
  ab initio and substructure determination
- 2D Projections
- Non-periodic objects

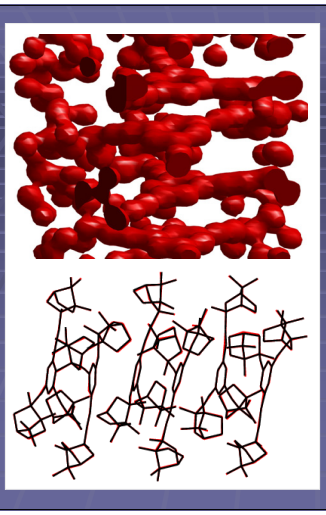


Structure solution using experimental data of known structures

Wu et al, Acta Cryst. A (2004)

Structure solution of unknown structures

Wardell et al, *Acta Cryst. E* (2007a, 2007b)
van der Lee & Astier, *J. Solid State Chemistry* (2007)
Rocha et al, *Acta Cryst. E* (2007)
Nylen et al, *J. Solid State Chemistry* (2007)
Piao et al, *Inorganic Chemistry* (2008)
Yang et al, *Inorganic Chemistry* (2007, 2008)
Bernot et al, *J. Am. Chem. Soc.* (2008)
…



3-fold pseudo-symmetry
271 atoms, 6 molecules
space group P1

Oszlányi, Sütő, Czugler, Párkányi
JACS (2006)

$Mo_2P_4O_{15}$
441 atoms, 21× supercell

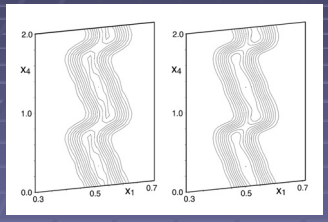
Evans et al. (2007)

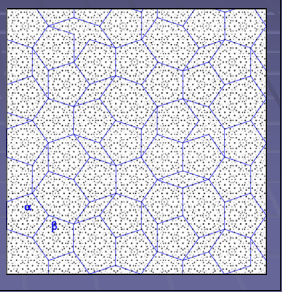
Superspace version of charge flipping

Palatinus, Acta Cryst A (2004)

Structure solution of unknown modulated structures

Palatinus et al. (2006)
Zuniga et al. (2006)





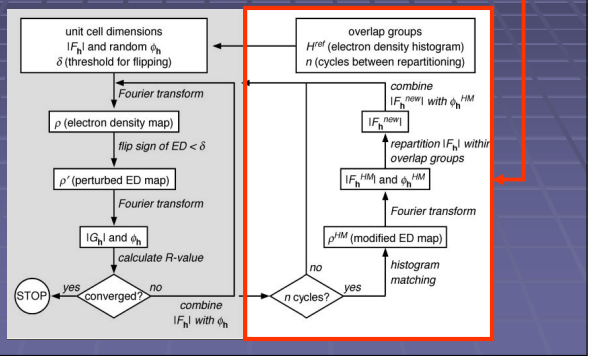
Structure solution of an Al-Ir-Os decagonal quasicrystal

Katrych et al. (2007)

Powder diffraction: handle overlapping reflections
two different implementations

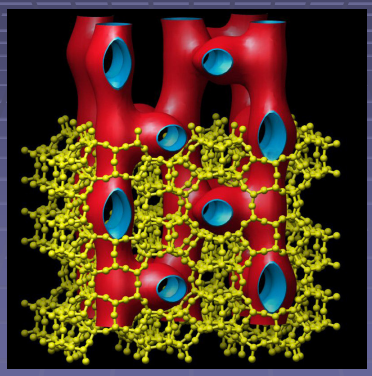
Wu, Leinenweber, Spence, O'Keeffe, *Nature Materials* (2006)

Baerlocher, McCusker, Palatinus, *Z. Krist.* (2007)



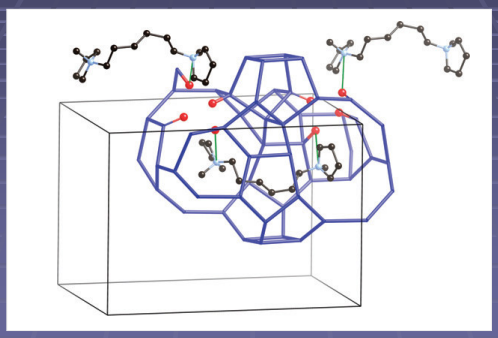
Structure solution of the zeolite catalyst IM-5
combing powder diffraction and electron microscopy

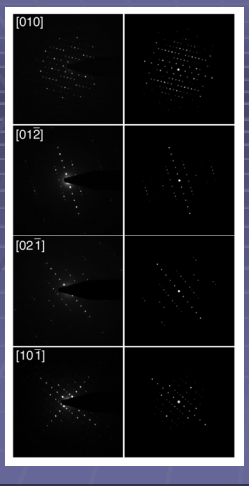
Baerlocher et al. Science (2007)



Ordered silicon vacancies in the framework structure
of the zeolite catalyst SSZ-74

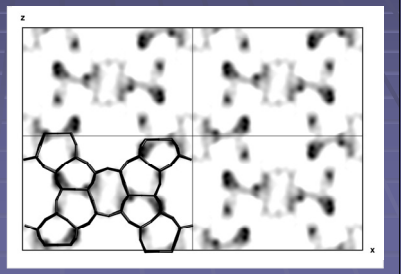
Baerlocher et al. Nature Materials (2008)





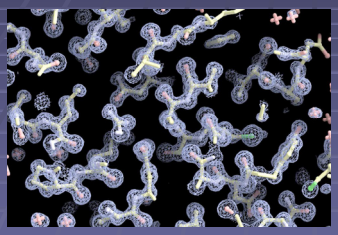
Combining PED + XPD
to facilitate structure solution
(CF used in two stages)

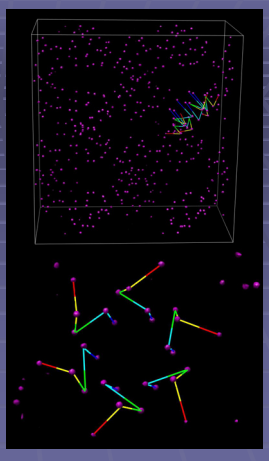
Xie, Baerlocher & McCusker
submitted to JAC



Macromolecular structure
solution by charge flipping

Dumas & van der Lee
Acta Cryst. D (2008)

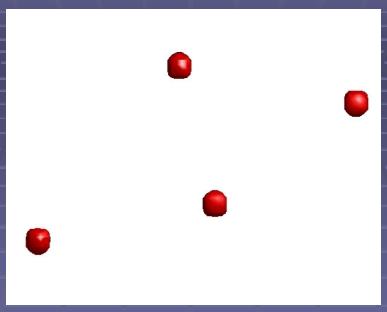




Mode: both ab initio and
heavy atom substructure

Utilize more information

- More data
  - resolution, sharpening, anomalous scattering
- Better starting point
  - a small fraction of ρ in real space
- Constraint in every cycle
  - known phases in reciprocal space
- New algorithm variants
  - balance of constraints and perturbations



Structure completion of vitamin B12
(10 cycles starting from 4 Co positions)

Constraints and perturbations

(fine balance is needed: may depend on the problem)

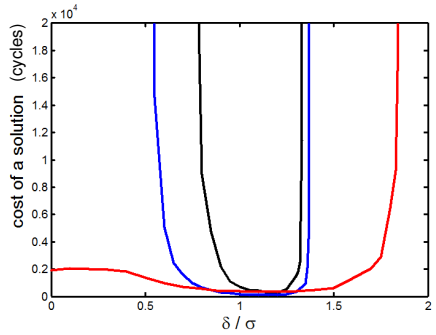
$$\rho^{(n+1)} = C_B C_A \cdot \rho^{(n)}$$

$$\rho^{(n+1)} = (C_B+\Delta_B)(C_A+\Delta_A) \cdot \rho^{(n)}$$

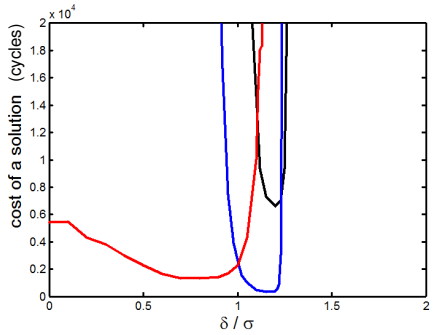
$$\rho^{(n+m)} = C_B C_A \ \ldots \ (C_B+\Delta_B)(C_A+\Delta_A) \cdot \rho^{(n)}$$

constraint once

perturbation m times



basic charge flipping
π/2 variant of charge flipping
1:1 alternation of π/2 and histogram matching



basic charge flipping
π/2 variant of charge flipping
1:1 alternation of π/2 and histogram matching

User programs

- BayMEM
  Smaalen, Palatinus & Schneider 2004
- SUPERFLIP
  Palatinus & Chapuis 2006
- Platon's FLIPPER module
  Ton Spek 2006
- TOPAS (Academic & Bruker)
  Alan Coelho 2007
- smtbx / cctbx / OLEX2
  Bourhis / Grosse-Kunstleve / Dolomanov et al
- connection to other programs
  e.g. CRYSTALS, Jana2000

Crystallographic teaching

- Simplicity matters …
- Do it yourself
  Perhaps written in two stages.
  Pick a high-level language with good graphics and speed up later.
- Ready-to-use Java applet
  Schoeni & Chapuis
  http://escher.epfl.ch/flip/

Summary

- There is a simple structure solution method around, user programs exist
- Try it !
- Can be used for teaching
- Can be used for everyday work
- Can be used for testing new ideas
- Limitations may eventually disappear

Acknowledgements

- **Scientific discussions**
  Gyula Faigel, Gábor Bortel, Miklós Tegze
- **Editors of Acta A and referees**
  Edgar Weckert, Dieter Schwarzenbach
- **Early users and program developers**
  Jinsong Wu, John Spence, Michael O'Keeffe
  Lukas Palatinus, Gervais Chapuis, Walter Steurer
  Christian Baerlocher, Lynne McCusker, Dan Xie
  Ton Spek, Alan Coelho, Luc Bourhis
  Arie van der Lee, Christian Dumas

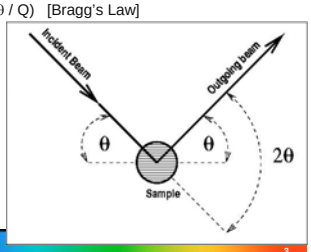
Introduction to: Powder Diffraction, Rietveld, GSAS & EXPGUI

Brian H. Toby

Physics of Single Crystal Diffraction

Diffraction from single crystals

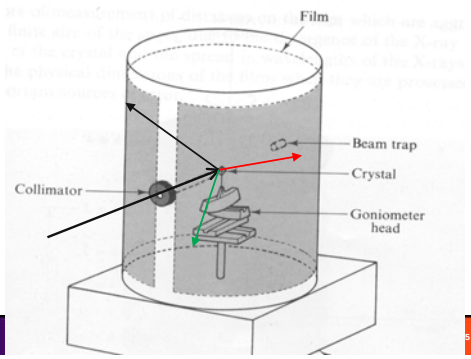
- Diffraction occurs when the *reciprocal lattice planes* of a crystal are aligned at an angle θ with respect to the beam and the wavelength of an incident beam satisfies:
  - $\lambda = 2\ d \sin\theta$ (or better, $\lambda = 4\ \pi \sin\theta / Q$) [Bragg's Law]
  - $d = 1/|\mathbf{d^*}| = 1/|h\mathbf{a^*} + k\mathbf{b^*} + l\mathbf{c^*}|$



Single Crystal Diffraction Intensities

- The Intensity of a diffracted beam, I_{hkl} is related to a imaginary number called the ***structure factor***, F_{hkl}
  - $I_{hkl} \propto |F_{hkl}|^2$
- The structure factor is determined by summing over all atoms in the crystal:
  - $F_{hkl} \propto \sum f_i \exp[2\pi i(hx_i + ky_i + lz_i)] \exp(-U_i Q^2/2)$
  - f_i represents the scattering power of an atom
  - U_i represents the average displacement of an atom from its ideal site

Single crystal intensities are collected by orienting the crystal in multiple orientations with a detector to measure scattered intensities

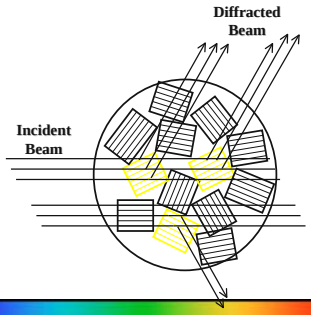


Physics of Powder Diffraction

Diffraction from random polycrystalline material

In a sufficiently large, randomly oriented polycrystalline sample (e.g. a powder), there are a number of crystallites in every possible orientation.

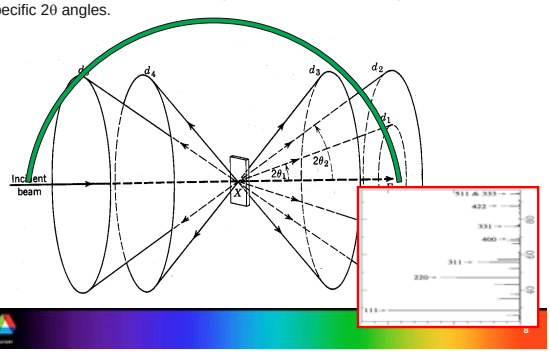
A beam impinging on the sample will diffract yielding all possible reflections



Argonne 7

Collecting powder diffraction intensities

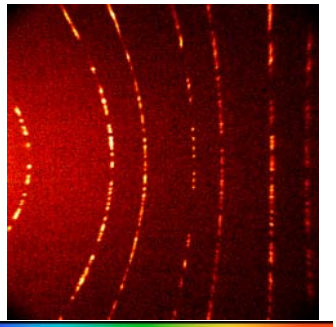
§ Reflections occur at discrete 2θ angles. Reflections fall in rings at these specific 2θ angles.



Argonne 8

What if we don't have an infinite number of crystals?

§ When the number of crystals is too small, the pattern becomes "grainy" -- diffraction from individual crystals dominates
§ *Diffraction intensities become unreliable*
  – Increase sample size
  – Grind the sample to decrease domain size
  – Oscillate or rotate the sample
    • *Spinning does not help much*
  – Use area detection & integrate the entire ring.



Argonne 9

What if two materials are present?

§ When a sample is composed of two or more crystalline phases the powder diffraction contains the weighted sum of the diffraction patterns from each phase
§ Mixtures can be identified
§ Peak areas allow amounts of each phase to be quantified

Argonne 10

What if the crystals do not have random orientation?

If some crystal orientations are over- or under-represented, the intensities of lines will be increased or decreased
  – in extreme cases, classes of lines can disappear
§ ***Preferred orientation***
  – Can be desired for engineering properties
  – Occasionally beneficial for structure solution
  – Usually problematic for Rietveld & quantitation

Argonne 11

What about "bad crystals"?

§ Crystals smaller than <<1 μm can show broadening
§ Twinning: not a problem
§ Stacking faults: can add intensity in weird ways (see Diffax program)
§ Poorly ordered materials: intensity falls off quickly (~ like extra-large Debye-Waller)

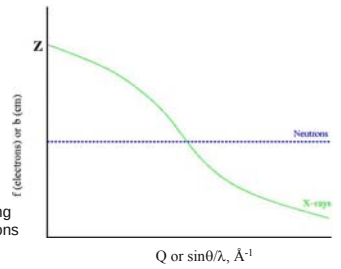
Argonne 12

X-ray vs. Neutron Diffraction

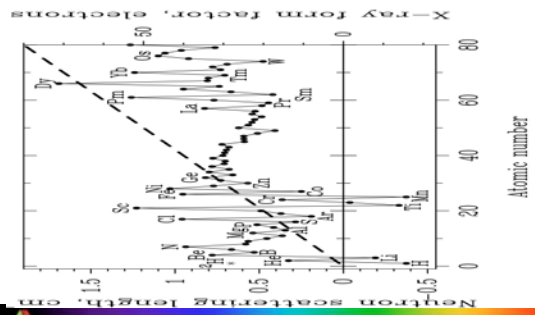
Atomic Scattering Power

- The scattering power (form factor, *f*) of an atom for X-rays depends on the number of electrons in the atom and Q
- The scattering power (scattering length, b) of an atom for neutrons depends on the isotope and is independent of Q

$$F_{hkl} \propto \sum f_i \cos[2\pi(hx_i + ky_i + lz_i)] \exp(-U_i Q^2/2)$$

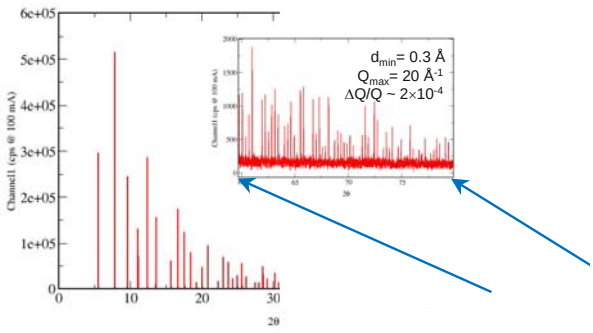


Comparison of Neutron and X-ray Atomic Scattering



11-BM: Exquisite data for the most complex problems

NIST SRM 660a (LaB_6) @30 keV (0.4Å)



Crystallographic Analysis from Powder Diffraction

How does one learn where the atoms are? Single Crystal Version

- Obtain single crystal (often hard)
- Begin data collection (usually easy)
- Determine unit cell (very easy)
- Complete data collection & reduce to structure factors (automatic)
- "Solve structure" -- determine approximate structure (usually very easy)
- Optimize structure to fit data -- "refinement" (requires considerable care)

How does one learn where the atoms are? Powder Diffraction Version

- Obtain powder sample (often only thing available)
- Collect diffraction pattern (usually easy)
- Determine unit cell -- autoindexing (can be hard)
- "Solve structure" -- determine approximate structure (can be even harder)
- Optimize structure to fit data -- "Rietveld refinement"
  - requires considerable care
  - fewer observations than from single crystals -- less detailed models
  - Easier to see if fit is providing a good match to data

Argonne 19

Crystallography from powder diffraction: before Rietveld

How did crystallographers use powder diffraction data?

- Avoided powder diffraction
- Manually integrate intensities
  - discard peaks with overlapped reflections

  Or
  - rewrote single-crystal software to refine using sums of overlapped reflections

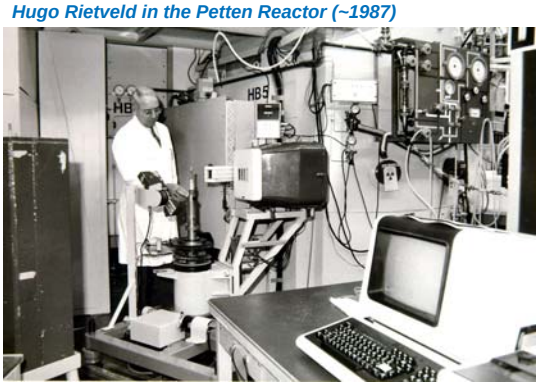
Simulation of powder diffraction data was commonly done

- Qualitative reasoning: similarities in patterns implied similar structures
- Visual comparison between computed and observed structure verifies approximate model
- Fits, where accurate (& precise) models were rarely obtained

Error propagation was difficult to do correctly (but not impossible)

Argonne 20

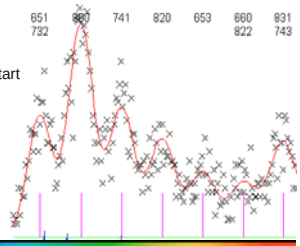
Hugo Rietveld in the Petten Reactor (~1987)



Argonne 21

Hugo Rietveld's technique

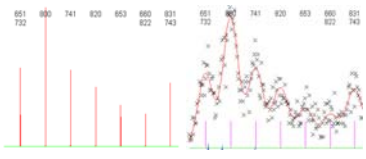
- Hugo Rietveld realized that if a pattern could be modeled, the fit between a computed pattern and observed data could be optimized.
  - Similar to single-crystal diffraction, except that now "experiment dependent parameters" must now be fit as well.
    - *Background*
    - *Peak broadening*
    - *Lattice constants*
  - Must have approximate model to start
  - Fewer data are available (usually)



Argonne 22

Calculation of Powder Diffraction: Graphical Example

hkl	mult	D-space	F_{hkl}	phase
6,5,1	48	1.548	0.29	0
7,3,2	48	1.548	1.709	180
8,0,0	6	1.5236	29.45	0
7,4,1	48	1.5004	2.327	0
8,2,0	24	1.4781	3.703	0
6,5,3	48	1.4569	1.27	0
6,6,0	12	1.4365	0.242	180
8,2,2	24	1.4365	2.086	0
8,3,1	48	1.417	0.22	180
7,4,3	48	1.417	1.827	180



1) Generate reflection list

2) Compute F_{hkl} from model

3) Peak heights are generated from $|F_{hkl}|^2$*multiplicity

4) Convolute peaks & add background

5) Optimize model, peak widths, etc. to improve fit

Argonne 23

Single crystal fitting

Powder data fitting

Minimize equation $\Sigma\, w_i[y_i - Y(x_i,\mathbf{p})]^2$ where

Data: $y_i = F_{hkl\ (obs)}$

y_i = observed powder diffraction intensities

Model: $Y(x_i,\mathbf{p}) = F_{hkl\ (calc)}$

$Y(x_i,\mathbf{p})$ = computed diffraction intensities from ($F_{hkl\ (calc)}$, background model, profile convolution, preferred orientation correction…

Parameters (p_1, p_2, p_3… p_m): atomic coordinates, displacement (T) factors

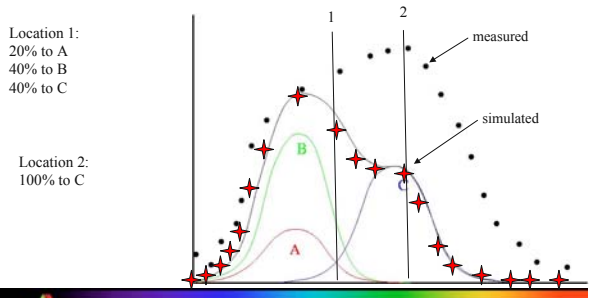
+ lattice parameters

+ "experimental" parameters for peak shapes, background…

Argonne 24

Hugo Rietveld's other breakthrough

§ Based on intensities from the model, estimates for F_{hkl} can be made, even when reflections are completely overlapped:



Rietveld Applications

§ Crystallographic structure determination

§ Quantify amounts of crystalline phases
  – (Amorphous content too, with neutrons)

§ Engineering properties
  – Residual stress
  – Preferred orientation

§ Lattice constant determination

Disadvantage of Rietveld: Many parameters need to be fit

§ Background
  – fixed
  – functions

§ Peak shape
  – "fundamental parameters"
  – functions

§ Lattice constants
  – zero correction
  – flat plate terms

§ Scaling
  – Phase fractions

§ Structural parameters
  – atom positions
  – occupancies
  – displacement parameters

§ Preferential Orientation

§ Absorption

Powder diffraction offers fewer observations and worse peak-to-background than single crystal diffraction

Rietveld Software

There are dozens of Rietveld codes, many have advantages for particular types of data or types of analysis

§ Multi-data type, multi-dataset Rietveld Codes
  – **FullProf**:
    • *uses somewhat "ugly" .PCR file*
    • *Most sophisiticated treatment of magnetism*
  – **Topas**
    • *Commercial*
    • *Fundamental parameters*
    • *Very fast, powerful (in expert hands)*
    • *Poor documentation*
  – **GSAS** (& EXPGUI)
    • *Most general: all types of data*
    • *Easiest for new users to learn*
    • *75% "market share" (DANSE survey)*

What is GSAS?

GSAS (General Structure and Analysis System)

§ Software package for fitting atomic structural models ("crystal structures") to single crystal and powder diffraction data.

§ Use virtually any type of neutron or x-ray diffraction data as input

§ Wide range of constraints and other features useful for complex problems.

§ GSAS includes a number of plotting and utility tools.

§ GSAS runs on Windows, Linux and Macintosh.

What is EXPGUI?

EXPGUI is a graphical user interface to GSAS.

- Intuitive access to only a small range of the GSAS capabilities,
  - much of what is needed for Rietveld analysis
  - Full range of GSAS capabilities still available through command-line type (EXPEDT) interface.
- EXPGUI also provides many useful utilities for viewing fits and refinement results.
- Distributions of EXPGUI include GSAS to simplify installation
- Available for Windows, Linux and Macintosh.

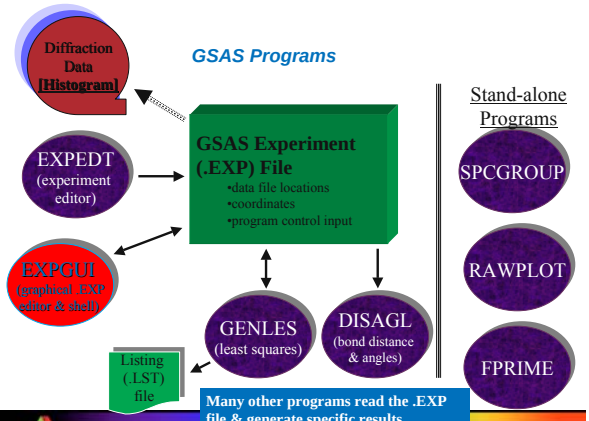
What does GSAS do?

- GSAS supports all types of x-ray & neutron crystallography instruments
  - TOF & CW neutron
  - Lab & synchrotron & energy dispersive x-ray
  - X-ray & neutron single crystal
- GSAS has special features not found in most packages:
  - Rigid bodies, distance & compositional constraints
  - Several profile & preferred orientation functions
  - Protein refinements
- Supported: Bugs get fixed (usually quickly)

GSAS: capsule history

- GSAS – conceived in 1982- 1983
  - by A. C. Larson & R. B. Von Dreele
- 1st version released in Dec. 1985
  - Only TOF neutrons (& buggy)
  - Designed to run only on VAX computers
    - *later for SGI, then MS-DOS (now windows) & then LINUX*
      - (At one time ULTRIX & HP Unix versions existed)
    - *2003 ported to Mac OS X*
- Designed from the start for:
  - multiple datasets & multiple phases
  - single crystal & TOF powder
- Later – CW neutron & CW x- ray powder data
- Basic structure & user interface is essentially unchanged from 1980's

GSAS Programs



Core GSAS programs

- Use EXPEDT to set up Experiment file
  - (or EXPGUI, to be covered later)
- Use POWPREF to map reflections to data
- Use GENLES to compute pattern/refine

rerun POWPREF when reflection positions must be updated, due to changes in:
- cell parameters/zero correction
- profile/sample displacement
- data range
- phases or data sets are added

Partial list of other GSAS programs

- POWPLOT: plot observed vs. computed pattern
- FOURIER, FORSRH, FORPLOT: compute Fourier map, find peaks & plot
- BIJCALC: analyze anisotropic displacement parameters
- ORTEP, VRSTPLOT: plot structures
- DISAGL: distances & angles
- HSTDUMP, TCLDUMP, REFLIST: list powder & reflection intensities
- PUBTABLE: tabulate results
- GSAS2CIF: create CIF's with results (rewritten!)
- RCALC: compute reflection R-factors (RF & RF2)
- GEOMETRY: compute from fitted parameters: Rigid bodies, L-S planes, error estimates on parameter sums.
- PC-GSAS: a windows program used to invoke other GSAS programs. N.B. also possible to use right-click on .EXP files

Powder Diffraction Data files for GSAS

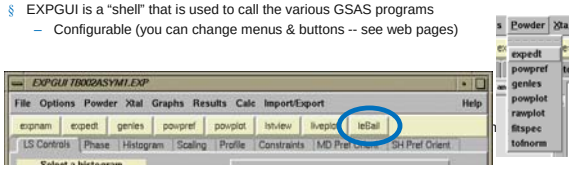
Two input data files are needed:

- Raw Histogram (contains intensities & optionally 2θ, etc. or s.u.).
  - *Many formats: read manual*
  - *GSAS is pretty picky on data formatting: read manual*
- Instrument parameter file: defines data type (λ, x-ray/neutron, expected profile, ...)
  - *Defines data type (x-ray/neutron, α_1 α_2, TOF, etc.)*
  - *Defines starting wavelength*
  - *Defines starting profile type & parameters*
  - *Create/edit with INSTEDIT in EXPGUI*

Argonne 37

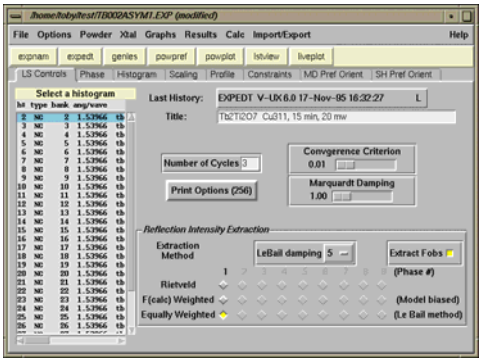
A Tour of EXPGUI

§ EXPGUI is a "shell" that is used to call the various GSAS programs
- Configurable (you can change menus & buttons -- see web pages)



Argonne 38

Least squares controls

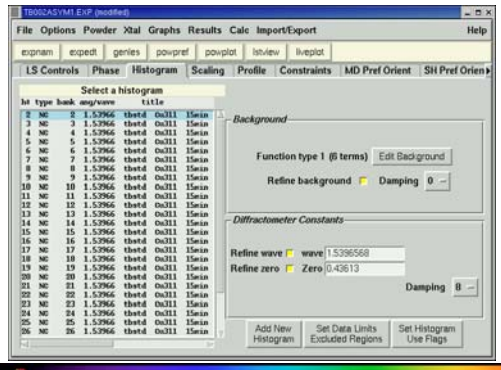


Argonne 39

Phase and atom parameters

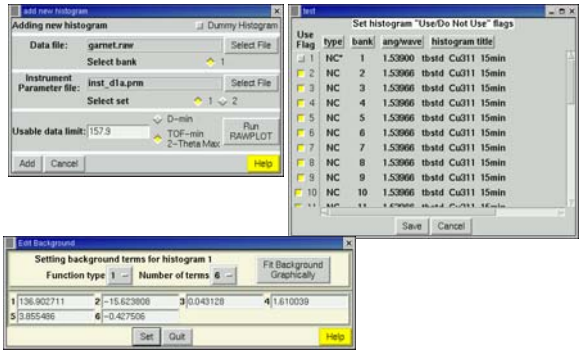
Argonne 40

Diffraction Data (Histogram) Parameters



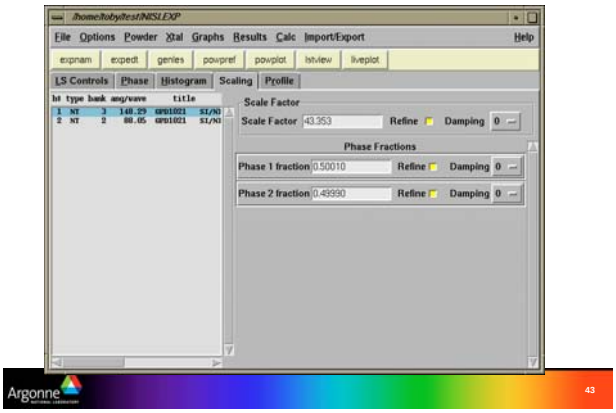
Argonne 41

Histogram parameter submenus

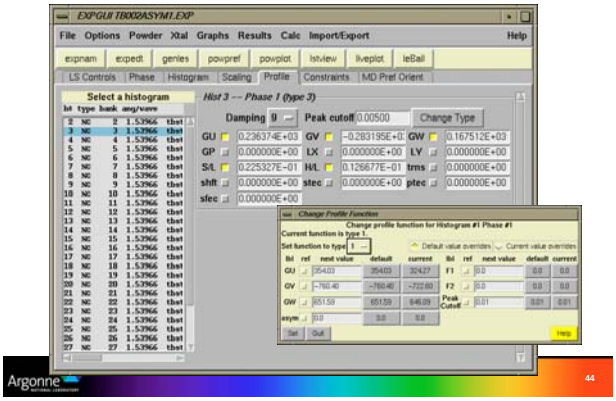


Argonne 42

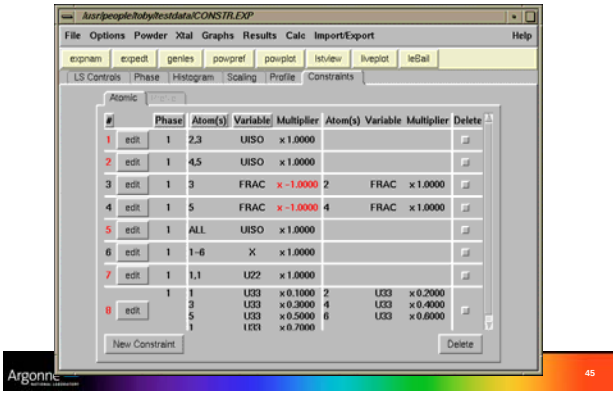
Phase & Histogram Scaling



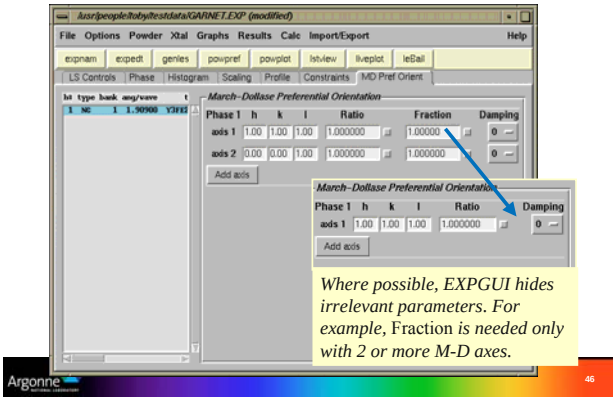
Profile terms



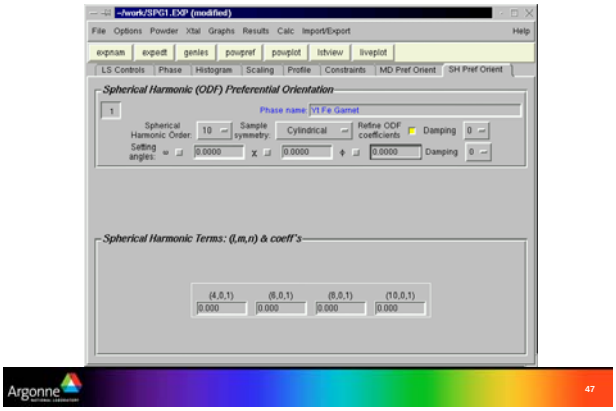
Atom/Profile Constraints



Preferred Orientation:March-Dollase



Preferred Orientation: Spherical Harmonics



EXPGUI design features

- Unlike EXPEDT, edits are made in memory
  - file is not changed until File/Save is used (or a GSAS program is run)
- Invalid values are highlighted in red and are not saved

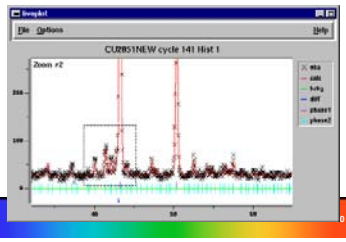
EXPGUI Utilities

Programs that add features not in GSAS.

- LIVEPLOT
  - Plot fit
- BKGEDIT
  - Manual background fit
- EXCLEDT
  - Edit data range & excluded regions
- Import/Export routines
  - GSAS2CIF
  - Import & write coordinates in multiple formats
- WIDPLT
  - Plot peak widths vs 2θ
- LSTVIEW
  - Examine GSAS output listing (.LST file)
- FillTemplate & CIFSelect
  - Add descriptive info to CIF
  - Select publication flag for distances & angles
- INSTEDIT
  - edit instrument parameter files

LIVEPLOT

- "Zoomable" plot of obs/calc/**bkg**/diff
- tickmarks, label with *hkl* labels
- Optional: overlay with unit cell or ICDD entry
- Export to GRACE & .csv
- (obs-calc)/σ
- Cumulative χ^2
- Shortcut keys



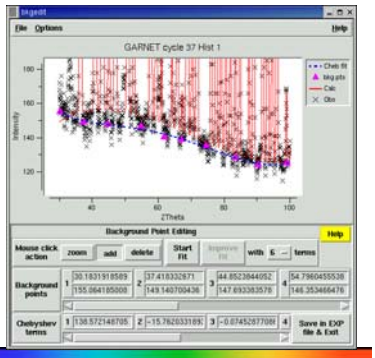
BKGEDIT

Background can be hard to fit, particularly when the structure is poorly modeled or when using a LeBail fit.

- Use of fixed background points is problematic

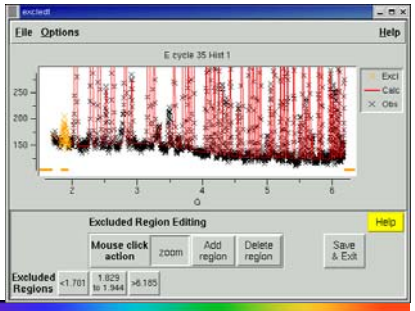
BKGEDIT:

- Fit most background functions to points input by user
- Coefficients can be refined once model is complete
- Compare functions (*type #1 is best -- BHT*)



EXCLEDT

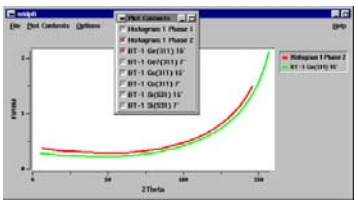
- Change data limits
- Add/delete excluded regions
- Work graphically in 2θ, TOF, Q, (even d-space)



WIDPLT

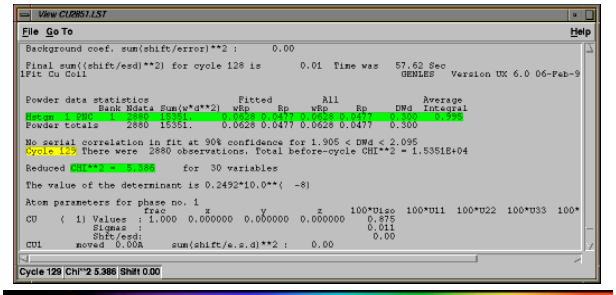
WIDPLT shows the FWHM as a function of 2θ, Q,… for a set of profile coefficients.

- Examine fit results to see that they make sense
- Compare values to predefined sets
- Look for hints of symmetry lowering



LSTVIEW

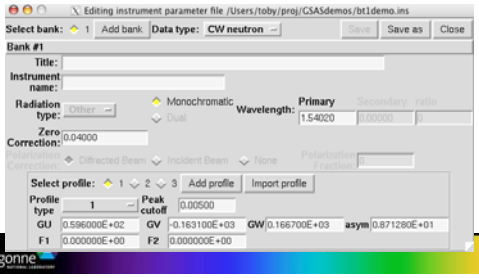
- A Scrollable viewer for .LST file



INSTEDIT

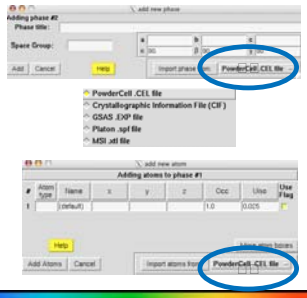
Allows instrument parameter files to be edited

- Save yourself some time:
  - make a file specific to your instrument or even each mode



Import/Export coordinates

Note that ***you*** can create import & export routines



EXPGUI: other features

- Detachable menus
  - (click on dashes)
- Archive/revert .EXP file
- Sort atoms/histograms
- Multiple histogram selection
  - Global changes
- CIF support
  - Generate & customize powder submissions (pdCIF)
  - Import coordinates
  - Import powder scans (latest version)

What Can't EXPGUI do?

EXPGUI has only a small fraction of the features of GSAS, but most of the commonly-used powder features

Use EXPEDT for:

- set up single crystal refinements
- set up macromolecular phases
- soft constraints (perhaps someday)
- All other features not present in EXPGUI (*e.g.* Fourier maps, rigid bodies)

Note that EXPGUI & EXPEDT coexist well: EXPEDT can always be called to access features not implemented in EXPGUI

Other free software from Brian Toby

- CMPR
  General purpose powder diffraction tool (graphics, reflection generation, indexing, peak fitting, ICDD search)
- CIFTOOLS
  CIF editor & pdCIF viewer (please use when refereeing!)

Also see web tools @ www.ncnr.nist.gov/xtal

Pair Distribution Function: DiffPy

Kyoto 2008: Crystallographic Computing School
August 21, 2008

Christopher L. Farrow
Department of Physics and Astronomy
Michigan State University
farrowch@msu.edu

MICHIGAN STATE UNIVERSITY

Outline

- DANSE and DiffPy
- The PDF method
- PDFfit2 and PDFgui
- SrRietveld
  - RefinementAPI
  - RietveldAPI
- Complex modeling and SrFit

DANSE and DiffPy

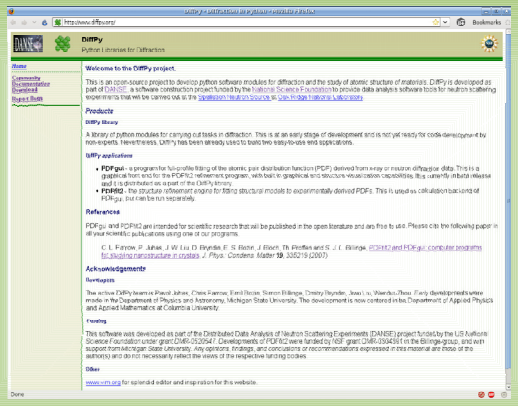
Distributed data analysis for neutron scattering experiments http://danse.us

- Data analysis for SNS, Oak Ridge National Lab, Tennessee
- NSF-funded software-development project (5 years, $12M)
- Centered at Caltech (P.I. Prof. Brent Fultz) with 5 sub-groups
  - Diffraction, Columbia University and MSU
  - Small-angle scattering, NIST
  - Reflectometry, NIST
  - Engineering Diffraction, ISU
  - Inelastic Scattering, Caltech
- Open-source data analysis modules and applications
- Developed in Python or compiled languages with Python bindings

DiffPy - Diffraction in Python http://www.diffpy.org

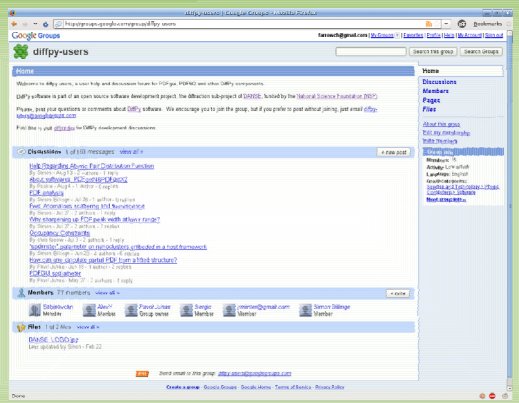
- Byname given to all DANSE-diffraction software
- Python library encompassing python modules for diffraction
  - e.g. `> from diffpy import pdffit2`

DiffPy resources



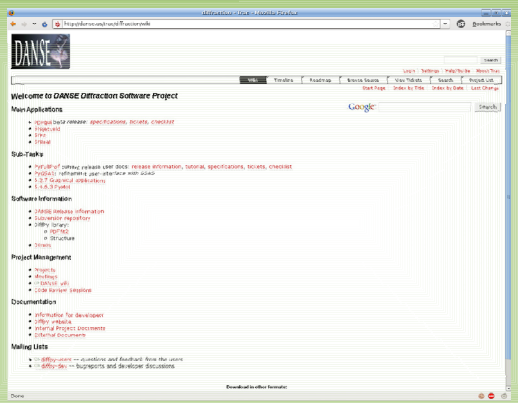
http://www.diffpy.org

DiffPy community



http://groups.google.com/group/diffpy-users

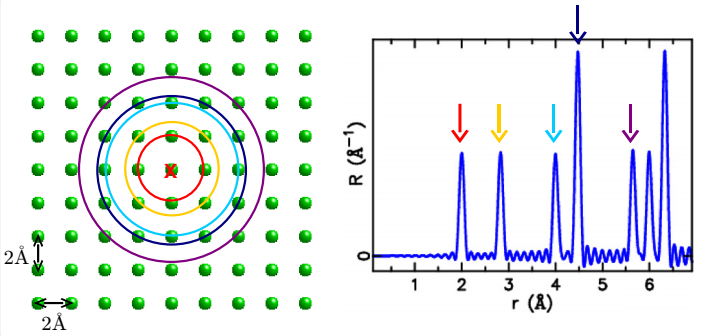
DiffPy development



http://danse.us/trac/diffraction/wiki

What is the PDF?

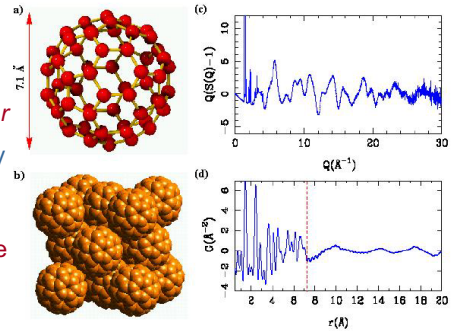
T. Egami and S. J. L. Billinge, Underneath the Bragg Peaks, Pergamon Press, 2003



The Atomic Pair Distribution Function (PDF) gives the probability of finding atom pairs separated by distance *r*.

What is the PDF?

- Sit on an atom and look at your neighborhood
- *G(r)* gives the probability of finding a neighbor at a distance *r*
- PDF is experimentally accessible
- PDF gives instantaneous structure
- PDF is intuitive

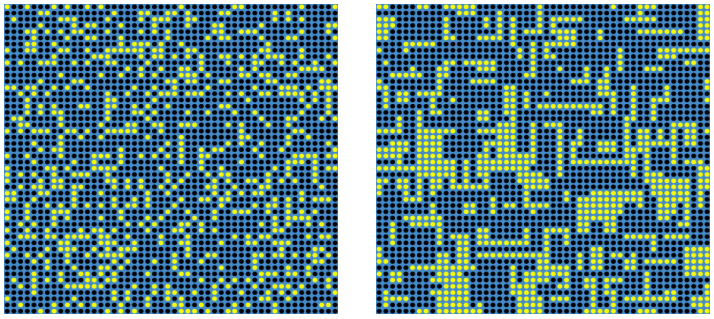


PDF is a total-scattering technique

Why diffuse scattering?

Underneath the Bragg peaks

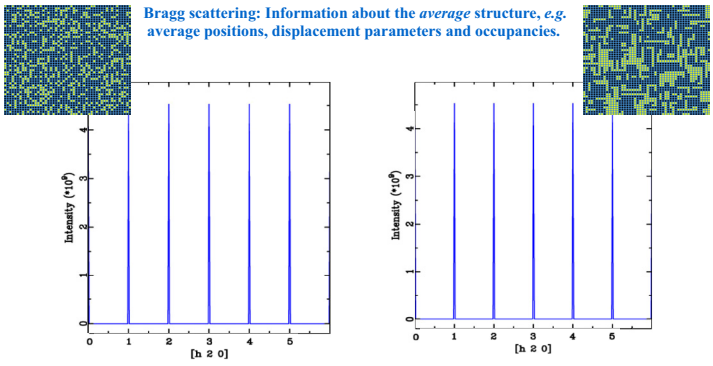
Simulations courtesy of Thomas Proffen



Cross section of 50x50x50 unit cell model crystal with 70% black atoms and 30% *vacancies*

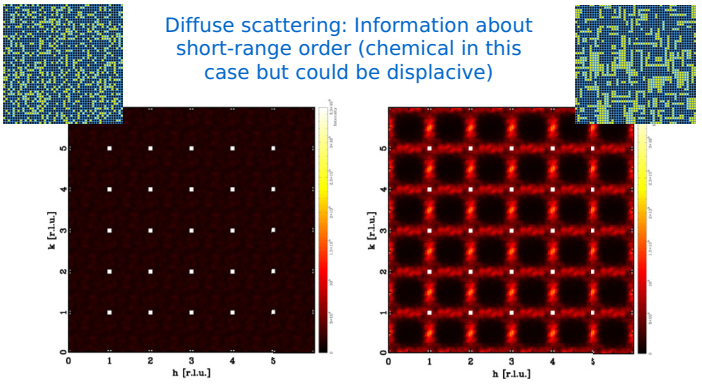
Bragg peaks give average structure

Simulations courtesy of Thomas Proffen

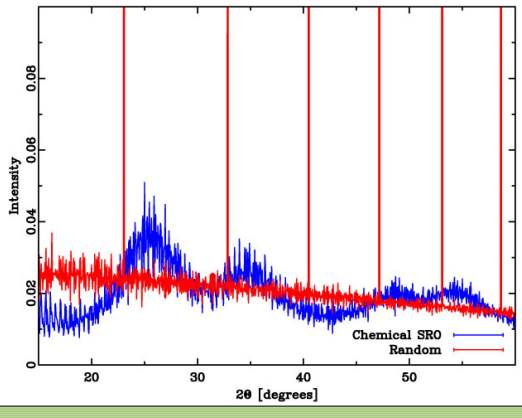


Diffuse scattering gives local structure

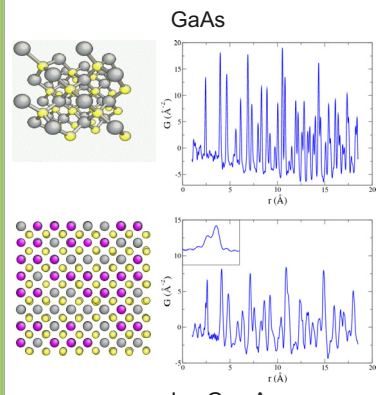
Simulations courtesy of Thomas Proffen



How about powder diffraction ?

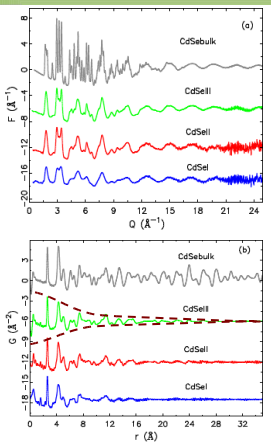


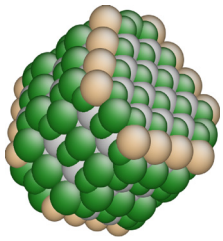
Local vs. Average: InGaAs



- (Crystallograpy) nearest-neighbor distance follows Vegard's law
- (Nanocrystallography) Local structure reveals distinct Ga-As and In-As bonds

CdSe Nanoparticles

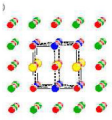




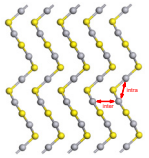
- Decay of PDF gives nanoparticle shape and size
- Shift in first peak gives inhomogeneous strain

Other PDF studies

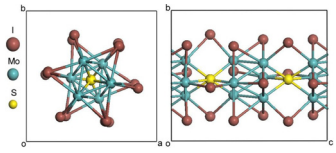
- Nanostructured bulk materials: $Pb_mAgSbTe_{m+2}$
  - Lin *et al.* Phys Rev B. (2006)



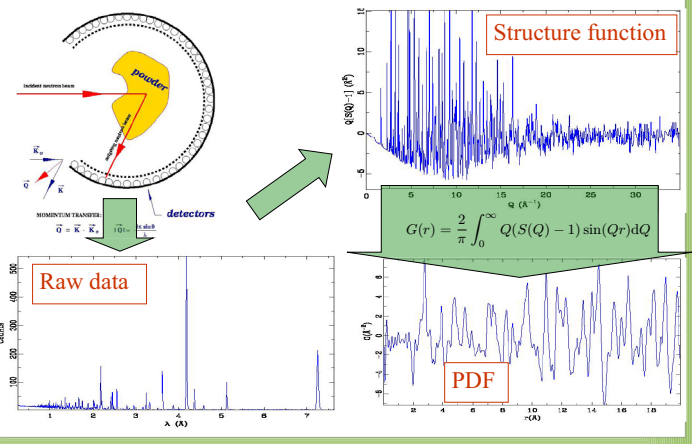
- Intercalants in host nanoporous materials
  - Shatnawi *et al.* J. Am. Chem. Soc. (2007)



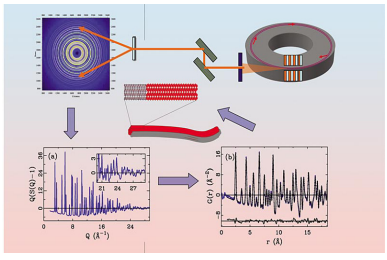
- Nanowires: $Mo_6S_xI_{10-x}$
  - Paglia *et al.* Chem. Mater. (2006)

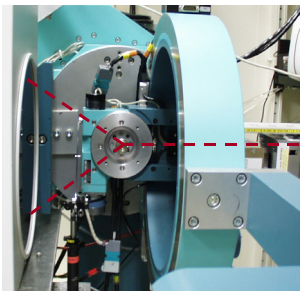


Measuring the PDF



Rapid Acquisition PDF



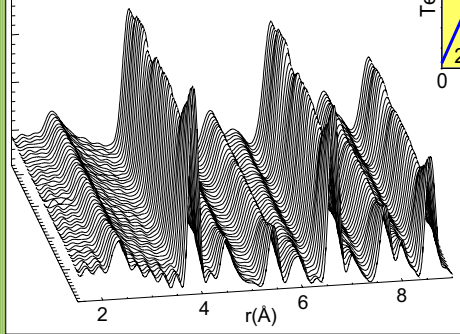


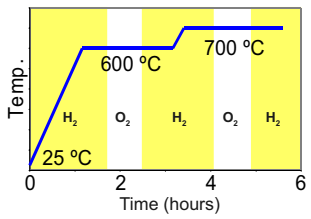
- Developed by a Billinge-group/Grey-group/BNL/APS collaboration in 2003
- Has revolutionized the applicability of PDF studies:
  - From a niche method for esoteric physics problems to a powerful and broadly applicable method for materials scientists and chemists to study complex nanostructured materials
  - Opens the door for in-situ and time-resolved studies

Chupas et al, J. Appl. Cryst. 36, 1342-1347 (2003)

Time-resolved studies

Study of nano-ceria demonstrated by Pete Chupas, Clare Grey and Jon Hanson
Chupas et al. JACS (2004)

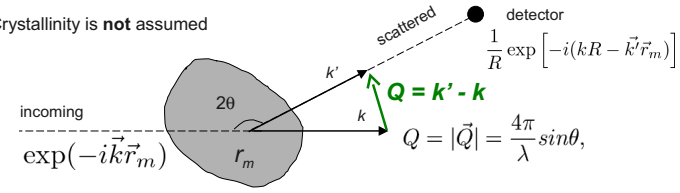




- Reduction-Oxydation cycles
- It is possible to see oxygen move in and out of the structure.

Scattering experiment

Crystallinity is **not** assumed



$$\psi(\vec{Q}) = \frac{e^{-ikR}}{R}\sum_m b_m \exp(i\vec{k}'\vec{r}_m)\exp(-i\vec{k}\vec{r}_m) = \frac{e^{-ikR}}{R}\sum_m b_m \exp i\vec{Q}\vec{r}_m$$

$$I(\vec{Q}) = \frac{(current\ at\ \mathrm{d}\Omega)}{N\mathrm{d}\Omega\ (flux\ in)} = \frac{R^2\mathrm{d}\Omega\ v\ |\psi(Q)|^2}{N\ \mathrm{d}\Omega\ 1\ v} =$$

$$= \frac{1}{N}\sum_{m,n} b_m b_n \exp i\vec{Q}(\vec{r}_m - \vec{r}_n) = \frac{1}{N}\sum_{m,n} b_m b_n \exp i\vec{Q}\vec{r}_{mn}$$

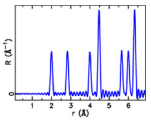
Scattered intensity depends on inter-atomic separations r_{mn}.

Theoretical highlights

constant — structure info

$$I(\vec{Q}) = \frac{1}{N}\sum_{m,n} b_m b_n \exp i\vec{Q}\vec{r}_{mn} = \langle b^2\rangle + \frac{1}{N}\sum_{m\neq n} b_m b_n \exp i\vec{Q}\vec{r}_{mn}$$

pair density function: $\rho(\vec{r}) = \frac{1}{N}\sum_n\sum_{m\neq n}\frac{b_m b_n}{\langle b\rangle^2}\,\delta(\vec{r}-\vec{r}_{mn})$



$$I(\vec{Q}) - \langle b^2\rangle = \langle b\rangle^2 \int \rho(\vec{r})\exp i\vec{Q}\vec{r}\,\mathrm{d}V$$

total scattering structure function: $S(Q) = \frac{I(\vec{Q}) - \langle b^2\rangle + \langle b\rangle^2}{\langle b\rangle^2}$

$$S(\vec{Q}) - 1 = \int [\rho(\vec{r}) - \rho_0]\exp i\vec{Q}\vec{r}\,\mathrm{d}V$$

Theoretical highlights

powder averaging → all possible orientations of *Q*: $\langle \exp i\vec{Q}\vec{r}\rangle = \frac{\sin Qr}{Qr}$

$$F(Q) = Q\,[S(Q) - 1] = \int_0^\infty 4\pi r\,[\rho(r) - \rho_0]\sin Qr\,\mathrm{d}r$$

inverse Fourier transformation:

$$G(r) \equiv 4\pi r\,[\rho(r) - \rho_0] = \frac{2}{\pi}\int_0^\infty Q\,[S(Q) - 1]\sin Qr\,\mathrm{d}Q$$

pair distribution function

$$\rho(r) = \frac{\mathrm{d}N}{4\pi r^2\mathrm{d}r} = \frac{1}{4\pi r^2}\frac{1}{N}\sum_n\sum_{m\neq n}\frac{b_m b_n}{\langle b\rangle^2}\,\delta(r - r_{mn})$$

The PDF is like the Patterson function, but it includes both Bragg and diffuse scattering.

$$G(r) = \frac{1}{Nr}\sum_{m\neq n}\frac{b_m b_n}{\langle b\rangle^2}\,\delta(r - r_{mn}) - 4\pi r\rho_0$$

Theoretical highlights

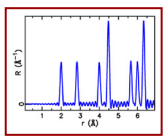
Debye scattering equation: $I(Q) = \frac{1}{N}\sum_m\sum_n \frac{\sin(Qr_{mn})}{Qr_{mn}}$

Total scattering structure function: $S(Q) = \frac{I(Q) - \langle b^2\rangle + \langle b\rangle^2}{\langle b\rangle^2}$

Atomic pair distribution function:

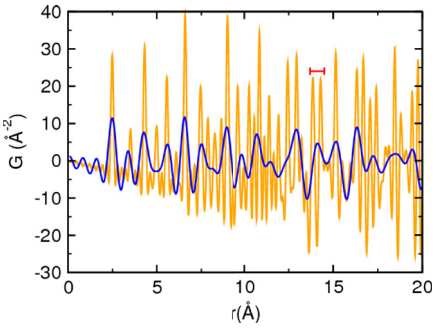
The PDF is like the Patterson function, but it includes both Bragg and diffuse scattering.

$$G(r) \equiv 4\pi r\,[\rho(r) - \rho_0]$$
$$= \frac{2}{\pi}\int_0^\infty Q\,[S(Q) - 1]\sin Qr\,\mathrm{d}Q$$



Can I use a lab source?

$\Delta r = 2\pi/Q_{max}$



Q_{max} = 25 Å^{-1}

Q_{max} = 8 Å^{-1}

PDF vs. conventional structure analysis

- Conventional analysis:
  - build real-space structure model
  - *calculate intensities from structure model (Q-space)*
  - compare observed and simulated intensities
- PDF analysis:
  - *apply Fourier transformation on measured intensities to get G_{obs}*
  - build real-space structure model
  - calculate G_{sim} for the model (real-space)
  - compare G_{obs} with G_{sim}
- Where is the difference?
  - conventional analysis assumes crystal lattice → intensities are evaluated only close to Bragg reflections, where $Q = 2\pi / d_{hkl}$
  - conventional methods see only the average structure, local structure features are filtered away together with non-Bragg intensities

PDF Modeling

- Least-squares refinement of parameters to minimize χ^2

$$\chi^2(\mathbf{p}) = \sum_n \frac{(G_{obs}(r_n) - G_{sim}(r_n;\mathbf{p}))^2}{\sigma_n^2}$$

1. G_{sim} calculated from current parameters **p** = (*p1, p2, ...*)
2. χ^2, $\nabla_{\mathbf{p}} \chi^2$ calculated
3. Parameters adjusted in downhill direction
   - $\mathbf{p} \leftarrow \mathbf{p} - \epsilon \nabla_{\mathbf{p}} \chi^2$ (ϵ is a suitably small number)
4. Continue until $\nabla_{\mathbf{p}} \chi^2 = 0$

PDF modeling

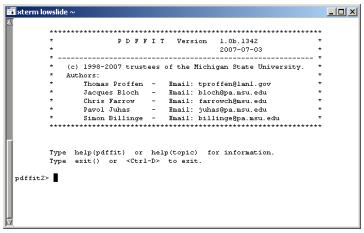
$$G_p(r) = \frac{1}{r} \sum_{ij} \frac{c_i c_j b_i b_j}{\langle b \rangle^2} \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left[-\frac{(r - r_{ij})^2}{2\sigma_{ij}^2}\right] - 4\pi r \rho_0$$

$$G(r) = \Lambda \exp\left(-\frac{1}{2}(\sigma_Q r)^2\right) \sum_p \lambda_p G_p(r) \quad \rightarrow \quad \textit{FFT to get ringing}$$

$$\sigma_{ij}^2 = (\sigma_i^2 + \sigma_j^2)\left(1 - \frac{\delta_1}{r_{ij}} - \frac{\delta_2}{r_{ij}^2} + (\alpha_Q r_{ij})^2\right) \phi(r_{ij})$$

Structural (mostly)				Non-Structural	
r_{ij}	- pair distance	b_i	- scattering power	Λ	- amplitude
σ_{ij}^2	- Debye-Waller factor	λ_p	- phase concentration	σ_Q	- peak damping
δ_1	- Low-T correlation	ρ_0	- number density	α_Q	- peak broadening
δ_2	- High-T correlation	c_i	- atomic concentration		

PDFfit2



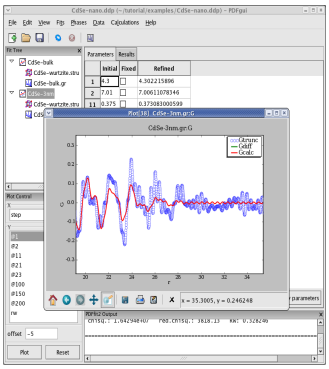
PDFfit2

- Real space structure refinement program, upgrade of PDFfit
  - Least squares refinement of structure model to experimental data

Features:

- Written in C++, optimized for speed
- Dynamic memory allocation, structure size is limited only by physical memory
- Flexible constraint system with analytic derivatives calculation
- Wrapped as Python library
  - Cross-platform
  - Command line and scripting interface
  - Can interface with any other Python module

PDFgui



PDFgui

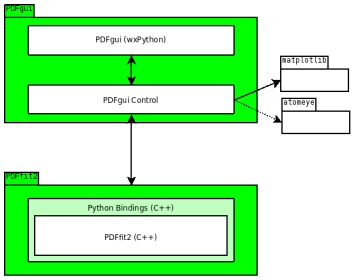
- Friendly GUI interface to PDFfit2
- Can organize multiple related fits in a single project file
- Powerful visualization facilities
  - Live plotting of refined PDF curves
  - Parametric plots of variables from multiple fits
  - 3D structure visualization
- Structure model manipulation
  - supports xyz, PDF, CIF and PDFfit formats
  - Supercell expansion
  - Expansion of asymmetric unit
  - Generation of symmetry constraints for coordinates and thermal factors
- Wizards for T-series, doping-series, r-series (smart extraction of meta-data from files)
- Built-in bug-reporting

Refinement of LaMnO3 temperature series with PDFgui

PDFgui Design Goals

- Easy to use
  - We want to recruit new PDF users!
  - We want non-specialists to use the software
- Maintainable
  - Needs to outlast its initial developers
  - Available and thorough developer's documentation
  - Simple design
- Extensible
  - Must be easy to add new functionality
- Reusable
  - Want to incorporate into future software projects

PDFgui Design



- PDFfit2 is the engine
  - Heavy lifting done in C++
- Control mediates communication
- GUI interfaces with user

- Benefits of Design
  - Engine works without GUI
  - Can update or change engine and keep GUI
  - Can update GUI and keep engine
  - Engine specialized for computation, stable
  - GUI specialized for user-interaction, rapid evolution
  - More code, but its easier to maintain

We we would change

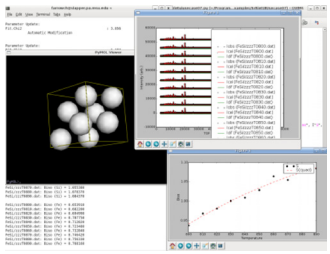
- Only one minimization scheme
  - No global optimizers
  - No restraints (soft constraints)
- Minimizer entangled with data structures
  - Good minimizer, but can't be reused
  - Hard to add new theory/parameters
- Only works with crystal structure model
  - Hard to study amorphous materials
- PDFGui control layer is hard to extend
- Some GUI functionality could exist in the engine

SrRietveld

- Resource for Rietveld refinement
- Parametric refinement
- Guided refinement
- Global optimization

- Fast distributed computing
- Read FullProf or GSAS input
- FullProf or GSAS backend
- 10-line refinement setup

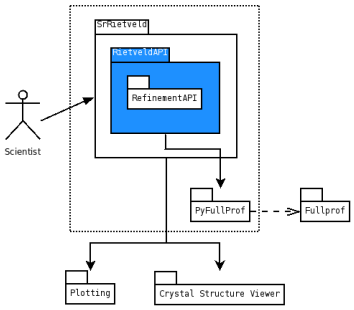
```
myfit = SimpleFit(
  data = "mydata.dat",
  instrument = "IPNS_Backscattering",
  model = ["Si.cif"])

autoRefine(myfit)

save(myfit)

plot(myfit)

visualizeStructure(myfit)
```



The 10 line script

- Routine refinements in 10 lines (or less!)
- Input data file, structure files and instrument name, e.g. “IPNS_Backscattering” or “POWGEN”
- Structural parameters set using structure file
- Peak shape, polarization and background set from tabulated instrument information
- Refinement follows heuristic or instrument-recommended progression
  - Instrument and heuristic information provided by beamline scientists
- Refinement can be tweaked after running

Design Overview



- SrRietveld
  - Powerful Rietveld refinement application
- RietveldAPI
  - Rietveld application interface for developers
- RefinementAPI
  - Refinement application interface for developers
- PyFullProf
  - Python scripting interface for FullProf

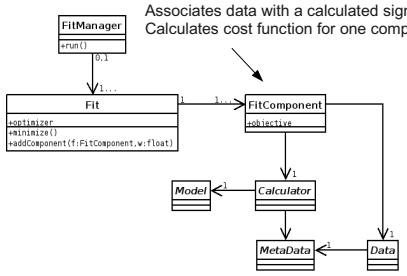
RefinementAPI

- Used to build refinement applications
- Built-in optimizers, objective functions, restraints and constraint handler
- Programmer/Scientist specifies calculator, model and data representation

$$\chi^2_{\text{total}} = \sum_i w_i \chi^2_i$$

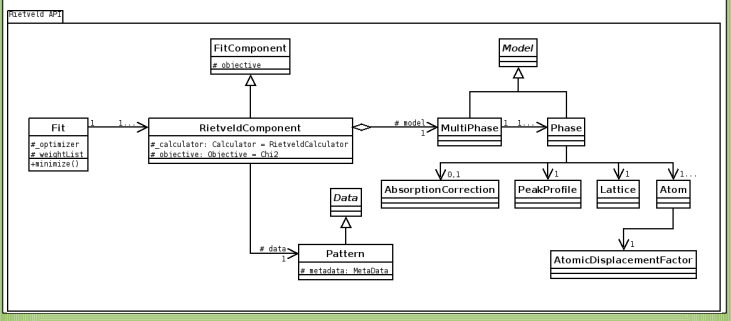
Total cost function is sum of cost functions from each component.

Associates data with a calculated signal. Calculates cost function for one component.



RietveldAPI

- Minimal application programming interface for Rietveld fitting
- Implements Rietveld Calculator (calls on PyFullProf), model representation and data representation
- Allows for multiple back-ends

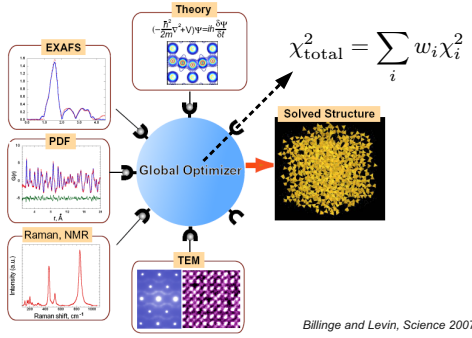


What does this give us?

- SrRietveld
  - Powerful and user-friendly Rietveld refinement application
  - Global optimization and parametric refinement → New Science!
- RietveldAPI
  - Minimal, reusable infrastructure for Rietveld refinement
  - Code sharing
    - Want community input on the API
    - SrRietveld works with any program using the API
- RefinementAPI
  - Reusable infrastructure for refinement applications
  - New optimizers accessible to any application using the API
  - Infrastructure for *complex modeling*
    - Multiple data types (PDF, Rietveld, EXAFS) can contribute to a fit

Complex Modeling

- Complex modeling mixes experiment and theory in a coherent computational framework



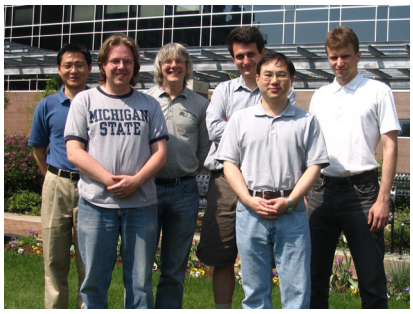
Billinge and Levin, Science 2007

SrFit

- Complex modeling application builder
  - Combines multiple data types into a single refinement
    - First public release will include PDF and Rietveld
  - Structural restraints on bond-lengths and angles
  - Multiple structural representations
  - Global optimization
- Built on RefinementAPI
  - uses RietveldAPI (and PDFAPI)
  - Enhancements in RefinementAPI benefit SrRietveld (and the next generation PDF application)

Acknowledgments

- DANSE Diffraction team
  - Simon Billinge
  - Emil Bozin
  - Pavol Juhas
  - Jiwu Liu
  - Wenduo Zhou
- DANSE team
- NSF



Direct Methods in Protein Crystallography

Haifu Fan & Yuanxin Gu
Beijing National Laboratory for Condensed Matter Physics
Institute of Physics, Chinese Academy of Sciences
P.R. China

2008 Kyoto
IUCr Crystallographic Computing School

Contents

The Phase Problem

$$\rho(x,y,z) = \frac{1}{V}\sum_h\sum_k\sum_l F(h,k,l)\,e^{-i2\pi(hx+ky+lz)}$$

$$F(h,k,l) = |F(h,k,l)|\,e^{i\alpha(h,k,l)}$$

$$\alpha(h,k,l) = ?$$

From the direct-method point of view:

Phases are not missing but just hidden in the magnitudes!

What is a Direct Method ?

It derives phases directly from the magnitudes.

$$\{|F(h,k,l)|\} \Rightarrow \{\alpha(h,k,l)\}$$

Why it is possible ?

$$|F(h,k,l)|\cos\alpha(h,k,l) = \sum_{j=1}^{N} f_j \cos 2\pi(hx_j + ky_j + lz_j)$$
$$|F(h,k,l)|\sin\alpha(h,k,l) = \sum_{j=1}^{N} f_j \sin 2\pi(hx_j + ky_j + lz_j)$$

Each reflection is accompanied by an unknown phase, but yields two simultaneous equations. Hence in theory, a diffraction data set of 3n reflections can be used to solve a structure with n independent atoms (assuming 3 parameters per atom).

That is to say, the phases may, at least in theory, be derived from a large enough set of magnitudes given the known quantities of atomic scattering factors.

But don't try to solve such kind of simultaneous equations really, It is impossible in practice!

We should go other ways round.

Sayre's Equation

$$F_{\mathbf{h}} = \frac{f}{f^{sq}V}\sum_{\mathbf{h}'} F_{\mathbf{h}'}F_{\mathbf{h}-\mathbf{h}'}$$

$F_{\mathbf{h}}$ – *structure factor with reciprocal vector* h
f – *atomic scattering factor*
f^{sq} – *scattering factor of the squared atom*
V – *volume of the unit cell*
Σ – *sum over the whole reciprocal space*

Sayre's equation defines the relationship among structure factors (magnitudes and phases)

Sayre's equation lies on the assumption of resemblance between the real and the squared structure.

Hence, the real structure can be derived by first solving the squared structure.

This leads to conditions for Sayre's equation to be valid:

1. ***Positivity*** – *electron densities should everywhere positive;*
2. ***Atomicity*** – *data resolution should better than about 1.2Å;*
3. ***Equal-atom structure*** – *the unit cell should consists of only one kind of atoms.*

It is rare that all the conditions are satisfied!

In case that one or more conditions are not satisfied, the result of Sayre's equation would tend to the squared structure rather than to the real one.

We should know what is the effect:

- ***Negative atoms in neutron diffraction analysis***

  The squared structure have coordinates consisting with that of the real structure, but all atoms will be positive. With prior chemical information we can figure out which atom is originally negative, what we need to do is simply flip densities of that atom into negative.

- ***Unequal-atom structure***

  Squaring the structure will make heavy atoms relatively heavier and light atoms lighter. This doesn't affect the result too much. Just keep in mind the effect and accept it.

- ***Low resolution data***

  This could cause real problems. But in our experience, Sayre's equation can work at a resolution down to 3.5Å or even lower.

The tangent formula

$$\tan\varphi_{\mathbf{h}} = \frac{\sum_{\mathbf{h}'} \kappa_{\mathbf{h},\mathbf{h}'}\sin(\varphi_{\mathbf{h}'} + \varphi_{\mathbf{h}-\mathbf{h}'})}{\sum_{\mathbf{h}'} \kappa_{\mathbf{h},\mathbf{h}'}\cos(\varphi_{\mathbf{h}'} + \varphi_{\mathbf{h}-\mathbf{h}'})}$$

$\varphi_{\mathbf{h}}$ – *phase of the structure factor* $F_{\mathbf{h}}$
$\Sigma_{\mathbf{h}'}$ – *sum over* h' *for those having a known* $\varphi_{\mathbf{h}'} + \varphi_{\mathbf{h}-\mathbf{h}'}$ *value*
$\kappa_{\mathbf{h},\mathbf{h}'} = 2\sigma_3\sigma_2^{-3/2}|E_{\mathbf{h}}E_{\mathbf{h}'}E_{\mathbf{h}-\mathbf{h}'}|$
$\sigma_n = \Sigma_j (Z_j)^n$, Z_j – *atomic number of the* j^{th} *atom*
$E_{\mathbf{h}}$ – *normalized structure factor with reciprocal vector* h

Tangent formula defines the relationship among phases (NOT including magnitudes) of structure factors

The tangent formula can be derived by maximizing the Cochran distribution at φ_h

$$P(\varphi_h) = [2\pi I_0(\alpha)]^{-1} \exp[\alpha \cos(\varphi_h - \beta)]$$

$$\alpha \sin\beta = \sum_{h'} \kappa_{h,h'} \sin(\varphi_{h'} + \varphi_{h-h'})$$

$$\alpha \cos\beta = \sum_{h'} \kappa_{h,h'} \cos(\varphi_{h'} + \varphi_{h-h'})$$

$$\alpha = \left[\left(\sum_{h'} \kappa_{h,h'} \sin(\varphi_{h'} + \varphi_{h-h'}) \right)^2 + \left(\sum_{h'} \kappa_{h,h'} \cos(\varphi_{h'} + \varphi_{h-h'}) \right)^2 \right]^{1/2}$$

Splitting Sayre's equation into real and imaginary parts and dividing the imaginary part with the real part, we obtain

$$\tan\varphi_h = \frac{\sum_h F_h F_{h-h} \sin(\varphi_h + \varphi_{h-h})}{\sum_h F_h F_{h-h} \cos(\varphi_h + \varphi_{h-h})}$$

This is formally equivalent to the tangent formula. In this sense, the tangent formula may be regarded as the angular portion of Sayre's equation.

On the other hand, Sayre's equation is an exact equation, while the validity of tangent formula is evaluated from probabilistic theory. The summation in Sayre's equation should cover the whole reciprocal space, but that of tangent formula may include only a few terms. This makes the tangent formula much easier to use in ab initio phasing.

Inputting no phases into the right-hand side of either Sayre's equation or the tangent formula, no phase information will come out from the left. On the other hand, by putting some starting phases into the right-hand side, improved phase information may return from the left. The larger is the starting phase set, the better will be the result.

This means that direct methods are more efficient in phase extension than in ab initio phasing.

Another obvious property of direct methods is that, the "sign problem" (making choice between 0 or π) will be much easier and more accurately to solve than the "phase problem" (estimating a phase value within the range of 0 to 2π).

Our applications of direct methods in protein crystallography belong to the category of phase extension. Phase problems in our applications are all reduced to a sign problem.

Use of direct methods in Protein Crystallography

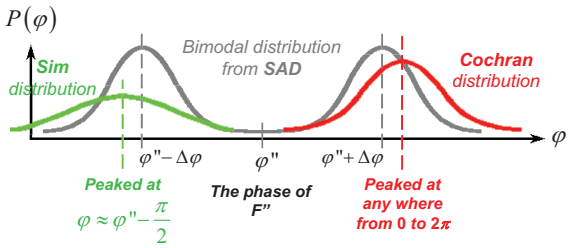
- *Locating heavy atoms*
- *Ab initio phasing of protein diffraction data at a resolution of 1.2Å or higher*

These are implemented by the program SnB, SHELXD or ACORN

- *Direct-method aided SAD/SIR phasing and partial-model extension*

These are implemented by the program OASIS

Phase information available in SAD



$$F'' = \sum_{j=1}^{N_{ano}} i\Delta f_j'' \exp\left[i2\pi\left(hx_j + ky_j + lz_j\right)\right]$$

Two different kinds of initial SAD phases

1. Combining bimodal SAD phase distribution with phase information from the anomalous-scattering substructure

2. Combining bimodal SAD phase distribution with phase information from the anomalous-scattering substructure and phase information from direct methods

Breaking the SAD/SIR phase ambiguity

The P_+ formula

$\varphi_h = \varphi'_h \pm |\Delta\varphi_h|$ *Reducing the phase problem to a sign problem*

$$P_+(\Delta\varphi_h) = \frac{1}{2} + \frac{1}{2}\tanh\left\{\sin|\Delta\varphi_h| \times \left[\sum_{h'} m_{h'} m_{h-h'} \kappa_{h,h'} \sin(\Phi'_3 + \Delta\varphi_{h',best} + \Delta\varphi_{h-h',best}) + \chi \sin\delta_h\right]\right\}$$

Breaking the SAD/SIR phase ambiguity by the Cochran distribution incorporating with partial structure information

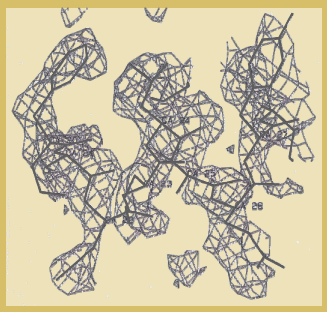
Acta Cryst. A40, 489-495 (1984)
Acta Cryst. A40, 495-498 (1984)
Acta Cryst. A41, 280-284 (1985)

Breaking the SAD/SIR phase ambiguity

The first application: Direct-method phasing of the 2Å experimental SAD data of aPP

Avian Pancreatic Polypeptide
Space group: C2
Unit cell:
a = 34.18, b = 32.92, c = 28.44Å;
β = 105.3°
Number of protein atoms in AU: 301
Resolution limit: 2.0Å
Anomalous scatterer:
Hg (in centric arrangement)
Wavelength: 1.542Å (Cu-Kα)
Δf" = 7.686
Locating heavy atoms &
SAD phasing: direct methods

Acta Cryst. A46, 935 (1990)



Data courtesy of Professor Tom Blundell

Further developments

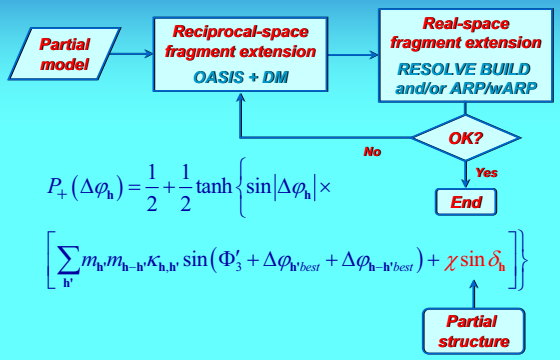
- *Direct-method SAD/SIR phasing combined with density modification*

OASIS + DM; OASIS + RESOLVE

- *Direct-methods aided dual-space structure-model completion*

ARP/wARP/REFMAC + OASIS; PHENIX + OASIS

Dual-space fragment extension



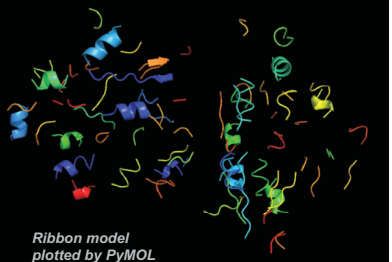
$$P_+(\Delta\varphi_h) = \frac{1}{2} + \frac{1}{2}\tanh\left\{\sin|\Delta\varphi_h| \times \left[\sum_{h'} m_{h'} m_{h-h'} \kappa_{h,h'} \sin(\Phi'_3 + \Delta\varphi_{h',best} + \Delta\varphi_{h-h',best}) + \chi \sin\delta_h\right]\right\}$$

Dual-space fragment extension

TTHA1634 from Thermus thermophilus HB8

Space group: $P2_12_12$
Unit cell:
a = 100.57,
b = 109.10,
c = 114.86Å
Number of residues in the AU: 1206
Resolution limit: 2.1Å
Multiplicity: 29.2
Anomalous scatterer: S (22)
X-ray wavelength:
λ = 1.542Å (Cu-Kα)
Bijvoet ratio: <|ΔF|>/<F> = 0.55%

Phasing & Model building:
Cycle 0:
OASIS + DM + RESOLVE



Ribbon model plotted by PyMOL

RESOLVE found the NCS and 381 of the total 1206 residues, 10 of which have side chains.

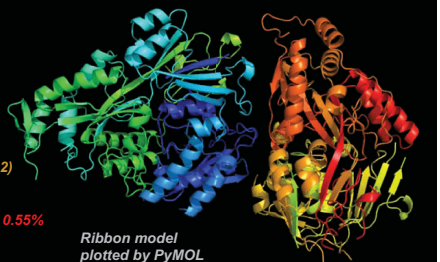
Data courtesy of Professor Nobuhisa Watanabe
Department of Biotechnology and Biomaterial Chemistry, Nagoya University, Japan

Dual-space fragment extension

TTHA1634 from Thermus thermophilus HB8

Space group: $P2_12_12$
Unit cell:
a = 100.57,
b = 109.10,
c = 114.86Å
Number of residues in the AU: 1206
Resolution limit: 2.1Å
Multiplicity: 29.2
Anomalous scatterer: S (22)
X-ray wavelength:
λ = 1.542Å (Cu-Kα)
Bijvoet ratio: <|ΔF|>/<F> = 0.55%

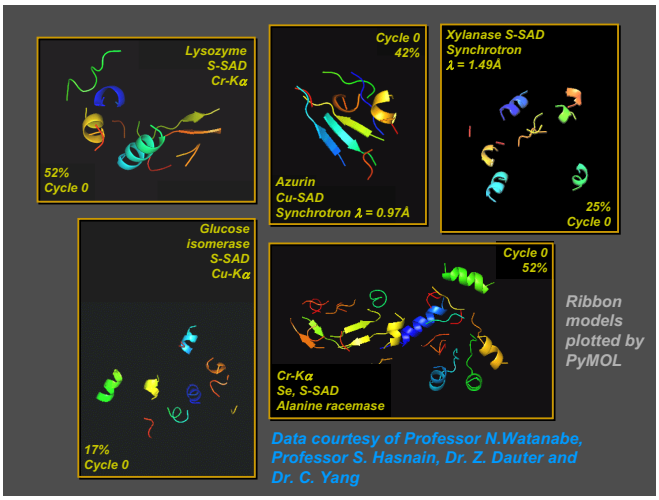
Phasing & Model building
Cycle 1:
OASIS + DM + ARP/wARP

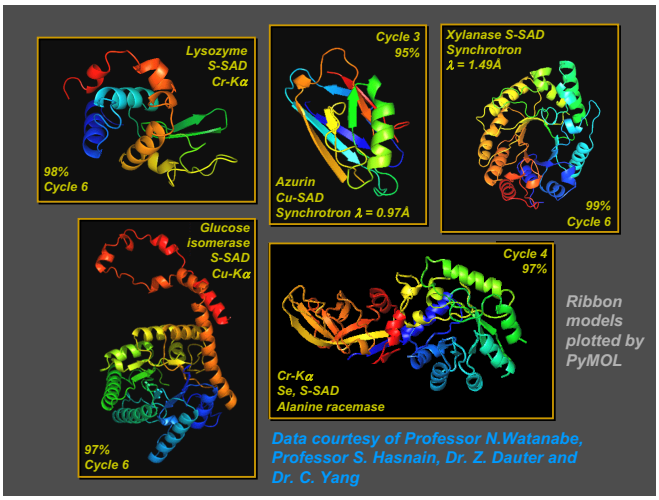


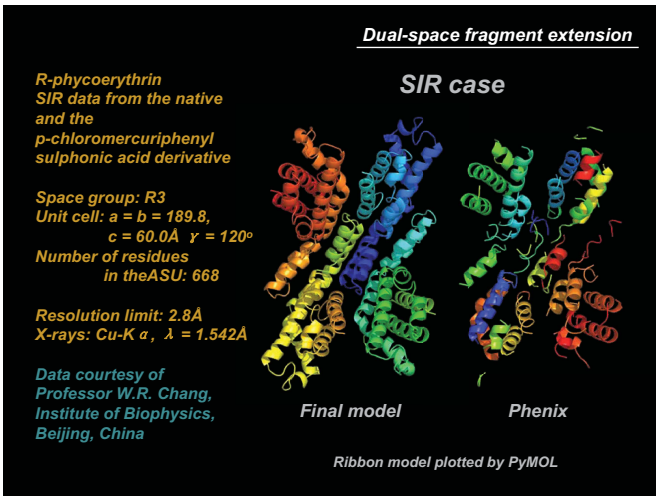
Ribbon model plotted by PyMOL

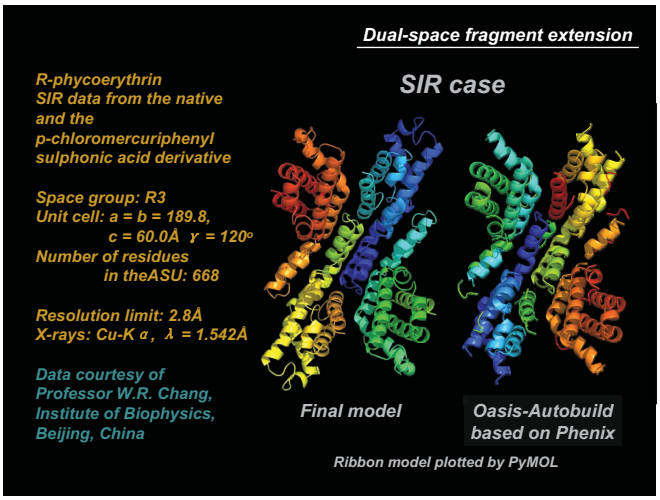
ARP/wARP/REFMAC found 1124 of the total 1206 residues, all docked into the sequence.

Data courtesy of Professor Nobuhisa Watanabe
Department of Biotechnology and Biomaterial Chemistry, Nagoya University, Japan



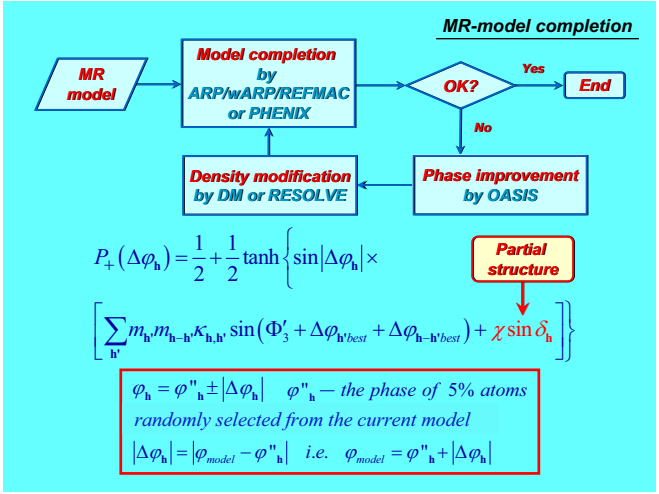






Dual-space fragment extension without SAD/SIR information

Direct-method aided MR-model completion



MR-model completion

$P_+ > 0.5$

$\varphi'' + \Delta\varphi = \varphi_{model}$

$<|\Delta\varphi|> \sim 90^o$

$<2|\Delta\varphi|> \sim 180^o$

$|\Delta\varphi|$ φ'' $|\Delta\varphi|$

$P_+ < 0.5$

$\varphi'' - \Delta\varphi$

$= \varphi_{model} - 2|\Delta\varphi|$

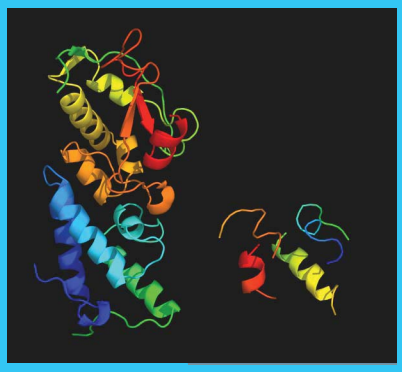
MR-model completion

E7_C/Im7_C (1UJZ)

Space group: I222
Unit cell: a = 62.88, b = 74.55, c = 120.44Å
Residues in the ASU: 215 (87 in mol. A; 128 in mol. B)
Resolution limit: 2.1Å
X-rays: Synchrotron

Molecular replacement & Model completion:

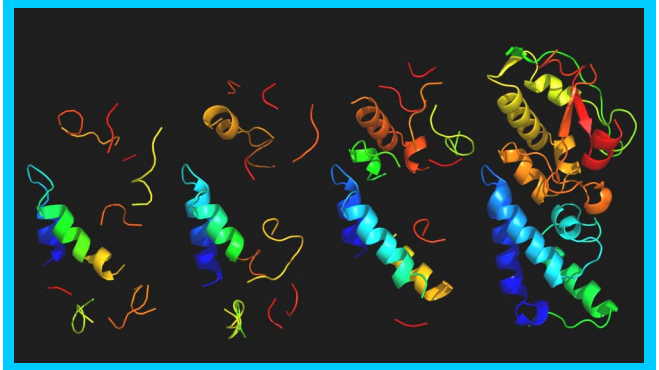
Searching model: Molecule A of 1BXI pruned using CHAINSAW
MR model searching: PHASER
Iterative Model completion: ARP/wARP/REFMAC-OASIS-DM



Final model Starting model

E7_C/Im7_C (1UJZ)

MR-model completion



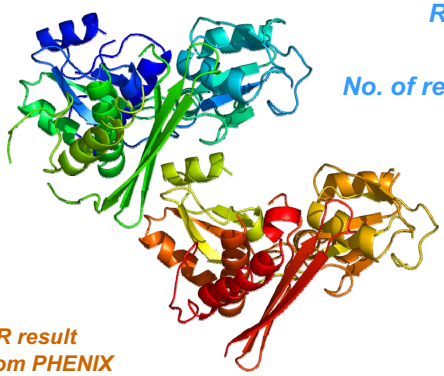
Cycle 1-4 Cycle 2-2 Cycle 6-3 Cycle 10-1

MR-model completion

Applying to an originally unknown protein

Space group: $P2_12_12_1$
a=71.81, b=81.40, c=108.95Å
Number of residuals in AU: 728
Solvent content: 0.37
Resolution limit: 2.5Å

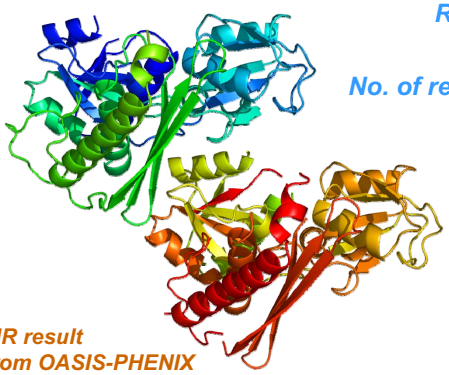
MR-model completion



R-factor: 0.33
R-free: 0.40
No. of residuals: 503

MR result from PHENIX

MR-model completion



R-factor: 0.24
R-free: 0.30
No. of residuals: 597

MR result from OASIS-PHENIX

OASIS-2008

to be released

Institute of Physics
Chinese Academy of Sciences
Beijing 100080, P.R. China

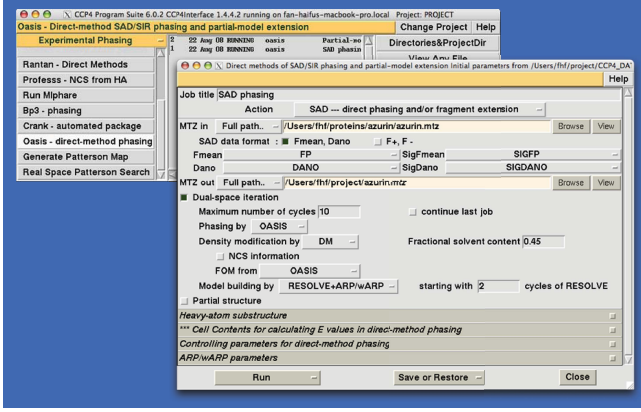
http://cryst.iphy.ac.cn

What's New?

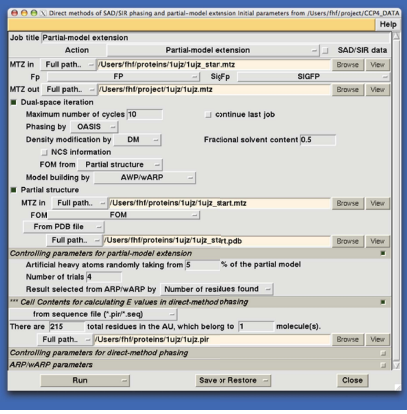
- *Running under Linux or MAC OSX*
- *Phasing up to 150,000 reflections in one data set*
- *HL coefficients for experimental SAD/SIR bimodal phase distribution listed in output.mtz*

  Thanks are due to Dr. T.C. Terwilliger for permission of migrating subroutines from SOLVE
- *A new CCP4 compatible GUI for automation of iterative dual-space phasing/fragment-extension and for integration of OASIS with DM, SOLVE/RESOLVE, ARP/wARP/REFMAC & PHENIX*

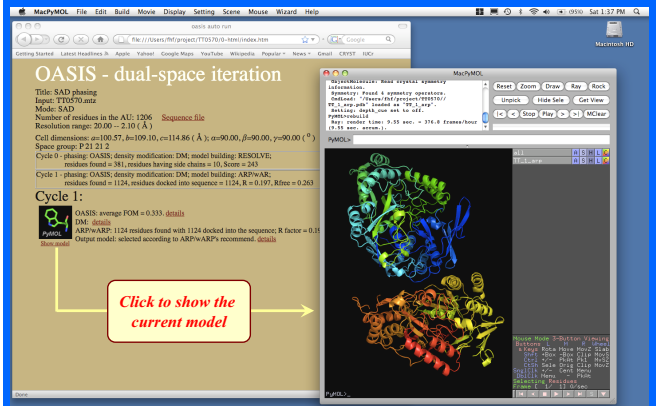
OASIS-2008: CCP4 Compatible GUI
Running for ab initio SAD phasing



OASIS-2008: CCP4 Compatible GUI
Running for partial model extension without SAD/SIR signals



OASIS-2008: CCP4 Compatible GUI
Online monitoring



References

Basic direct methods

[1] Sayre, D. (1952). *The squaring method: a new method for phase determination. Acta Cryst.* **5**, 60-65.
[2] Cochran, W. (1955). *Relations between the phase of structure factors. Acta Cryst.* **8**, 473-478.
[3] Karle, J. & Hauptman, H. (1956). *A theory of phase determination for the four types of non centro-symmetric space groups 1P222, 2P22, 3P12, 3P22. Acta Cryst.* **9**, 635-651.
[4] Fan, H.F. (1998). *Sayre equation, tangent formula and SAYTAN, in《Direct Methods for Solving Macromolecular Structures》* Ed. by S. Fortier, Kluwer Academic Publishes, The Netherlands, 1998, pp. 79-85. (*Full text in "Publications" at http://cryst.iphy.ac.cn*)

Ab initio direct methods in protein crystallography

[5] DeTitta, G.T., Weeks, C.M., Thuman, P., Miller, R. & Hauptman, H.A. (1994). *Structure solution by minimal function phase refinement and Fourier filtering: theoretical basis. Acta Cryst.* A**50**, 203-210.
[6] Sheldrick, G.M., & Gould, R.O. (1995). *Structure solution by iterative peaklist optimization and tangent expansion in space group P1. Acta Cryst.* B**51**, 423-431.
[7] Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Yao, J.X. and Zheng, C.D. (2000). *A flexible and efficient procedure for the solution and phase refinement of protein structures. Acta Cryst.* D**56**, 1137-1147.

Breaking the phase ambiguity intrinsic in SAD/SIR experiments

[8] Fan, H.F. & Gu, Y.X. (1985). *Combining direct methods with isomorphous replacement or anomalous scattering data III. The incorporation of partial structure information. Acta Cryst.* A**41**, 280-284.

References (*continued*)

Dual-space fragment extension with SAD/SIR information

[9] Wang, J.W., Chen, J.R., Gu, Y.X., Zheng, C.D. & Fan, H.F. (2004). *Direct-method SAD phasing with partial-structure iteration -- towards automation. Acta Cryst.* D**60**, 1991-1996.

[10] Yao, D.Q., Huang, S., Wang, J.W., Gu, Y.X., Zheng, C.D., Fan, H.F., Watanabe, N. & Tanaka, I. (2006). *SAD phasing by OASIS-2004: case studies of dual-space fragment extension. Acta Cryst.* D**62**, 883-890.

[11] Yao, D.Q., Li, H., Chen, Q., Gu, Y.X., Zheng, C.D., Lin, Z.J., Fan, H.F., Watanabe, N. & Sha, B.D. (2008). *SAD phasing by OASIS at different resolutions down to 3Å and below. Chinese Physics* B**17**, 1-9.

Dual-space fragment extension without SAD/SIR information

[12] He, Y., Yao, D.Q., Gu, Y.X., Lin, Z.J., Zheng, C.D. & Fan, H.F. (2007). *OASIS and molecular-replacement model completion. Acta Cryst.* D**63**, 793-799.

Acknowledgements

Professor Zhengjiong Lin[1,2]

Institute of Biophysics, Chinese Academy of Sciences, Beijing, China

Drs Y. He[1], D.Q. Yao[1], J.W. Wang[1], S. Huang[1], J.R. Chen[1], Prof. T. Jiang[2], Mr. T. Zhang[1], Mr. L.J. Wu[1] & Prof. C.D. Zheng[1]

[1] Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, China
[2] Institute of Biophysics, Chinese Academy of Sciences, Beijing China

The project is supported by the Chinese Academy of Sciences and the 973 Project (Grant No 2002CB713801) of the Ministry of Science and Technology of China.

Macro-molecular structure solution and refinement: SHARP and BUSTER

Clemens Vonrhein

Global Phasing Ltd.

Overview - 1

- data collection is last experiment
- result of experiment is electron density
- model is interpretation of electron density
- phase problem:

$$\rho(\vec{r}) = \sum_{\vec{h}} F(\vec{h}) \cdot e^{-i2\pi \vec{h} \cdot \vec{r}}$$

$$F(\vec{h}) = |F(\vec{h})| e^{i \cdot \varphi(\vec{h})}$$

Overview - 2

- underlying "phasing picture":
  - data collection
  - structure solution
  - model building
  - refinement
- goal: improving signal-to-noise ratio
- quality of final phases:
  - how well can they show mistakes and (yet) unmodelled parts
  - Fo-Fc difference Fourier maps
  - better Fo through better experiments (and data processing)
  - better Fc through better models (parametrization) and refinement

Outline

- SHARP and BUSTER
- phases and amplitudes
- heavy atom parameter refinement and phasing
- density modification
  - improving starting phases
  - cycling between real space and reciprocal space
- initial model building
  - automatic model building
- refinement
  - correct model parametrization
  - choice of observations
  - external phase information

SHARP and BUSTER

- SHARP
  - Heavy atom parameter refinement
  - Phase calculation

G. Bricogne, C. Vonrhein, C. Flensburg, M. Schiltz, and W. Paciorek (2003). Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. Acta Crystallogr D59,2023-30

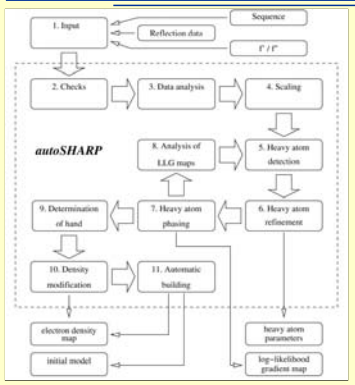
- BUSTER
  - model refinement
  - some unique features

E. Blanc, P. Roversi, C. Vonrhein, C. Flensburg, S. M. Lea, and G. Bricogne (2004). Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. Acta Crystallogr D60, 2210-21

- autoSHARP
  - automated pipeline centred around SHARP
  - uses external programs (SHELXD, SOLOMON, ARP/wARP, CCP4)
  - data analysis and HA detection
  - density modification
  - automated model building

C. Vonrhein, E. Blanc, P. Roversi, and G. Bricogne (2007). Automated structure solution with autoSHARP. Methods Mol Biol, 364, 215-30.

Structure solution process



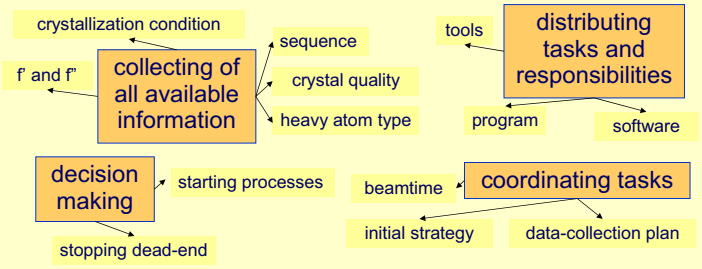
input

different steps, different programs, different methods ...

several levels of results

Non-automated structure solution

- Each method/tool has its own set of assumptions, which are met (automation) or not (hand-tuning)
- crystallographer = “structure solution manager”:

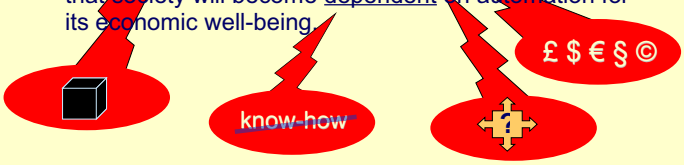


Why automation?

- **Definition:** [...] automation can be defined as a technology concerned with performing a process by means of programmed commands combined with automatic feedback control to ensure proper execution of the instructions.
- **Advantages:** [...] higher production rates and increased productivity, more efficient use of materials, better product quality, improved safety, shorter workweeks for labour [...]. Also, increased process control makes more efficient use of materials, resulting in less scrap.

Why automation?

- **Disadvantages:** [...] possibility that workers will become slaves to automated machines, that the privacy of humans will be invaded by vast computer data networks, that human error in the management of technology will somehow endanger civilization, and that society will become dependent on automation for its economic well-being.



- http://search.britannica.com/eb/article?eu=117180

Example (Data analysis) 1/3

- ... something went wrong

H	K	L		FMID	SMID	DANO	SANO	ABS(DANO)/FMID
-3	1	3	w2	4.09	1.58	7.92	29.96	1.94
-3	1	12	w1	203.27	5.43	398.98	7.68	1.96
			w3	212.89	6.19	418.12	8.75	1.96
-2	2	12	w1	79.28	2.86	-150.88	4.05	1.90
-1	1	2	w3	187.55	6.28	-358.76	8.87	1.91

I(h,k,l) >> I(-h,-k,-l)

- huge signal?
- I(-h,-k,-l) wrongly integrated?
- radiation damage: I(-h,-k,-l) measured late?

Example (Data analysis) 2/3

- ... something went wrong

H	K	L	Reso	w1	w2	w3
1	1	1	22.60	98.59	132.40	26.30
2	2	0	13.90	585.57	586.64	84.49
-1	1	2	17.51	28.28	146.59	-
-2	0	4	10.75	126.22	137.03	2.31*
-3	1	3	12.15	3.12	-	581.69*

low resolution bad?

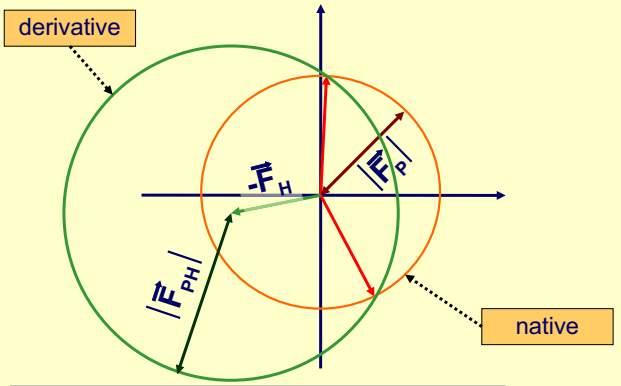
w3 bad (radiation damage)?

can't really tell

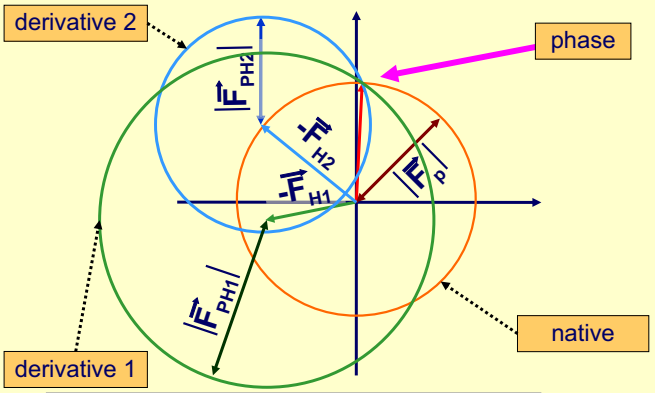
Example (Data analysis) 3/3

- removing reflections:
  - 2 sites (CC=0.319)
  - fom [acen/cen] = 0.59/0.57
  - phasing power (peak/ANO) = 1.74
  - original / inverted hand: CC = **0.29 / 0.46**
- keeping reflections:
  - 2 sites (CC=0.415)
  - fom [acen/cen] = 0.46/0.42
  - phasing power (peak/ANO) = 1.12
  - original / inverted hand: CC = **0.15 / 0.16**

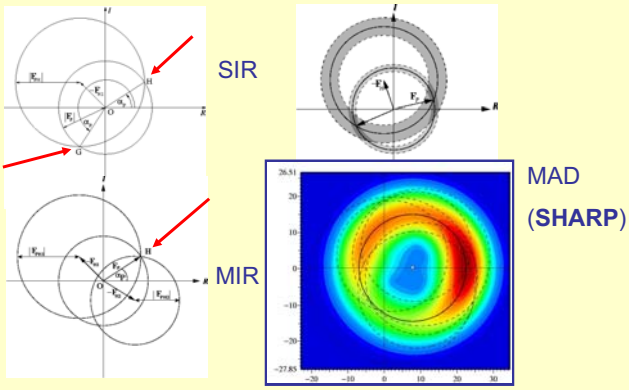
SIR



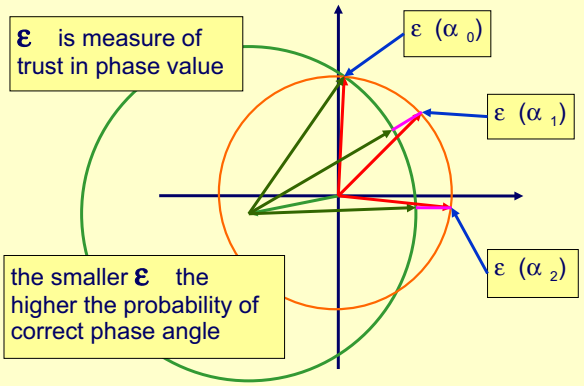
MIR



Real world



LOC cont.



Phase combination

$$P(\alpha_i) = N \cdot e^{-\frac{[\varepsilon(\alpha_i)]^2}{2E^2}}$$

$$P(\alpha_i) = N \cdot e^{(A\cdot\cos\alpha_i + B\cdot\sin\alpha_i + C\cdot\cos 2\alpha_i + D\cdot\sin 2\alpha_i)}$$

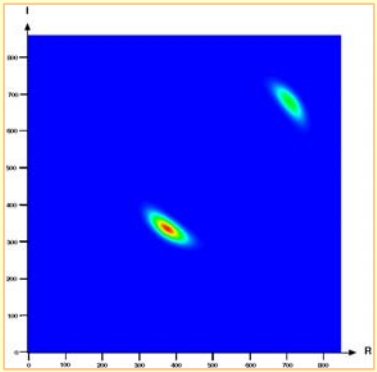
Hendrickson-Lattman coefficients

probabilities are multiplied

→

HL coefficients are added

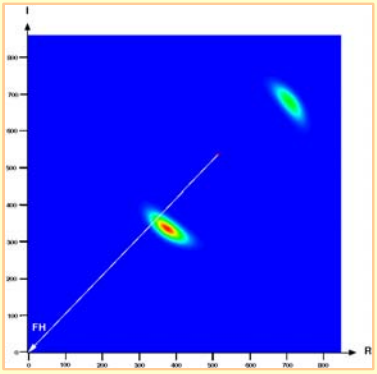
SHARP: 2D phase probability



- 1GXT (Rosano et al., 2002)
- SAD (Hg)

- HKL = (5,5,0)
- complex plane
- probability at each point

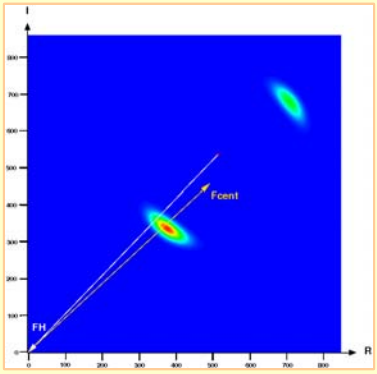
SHARP: 2D phase probability



- calculated FH
- from heavy atom model

- X,Y,Z
- B
- f'/f"
- scale
- Lack-of-isomorphism (LOI)

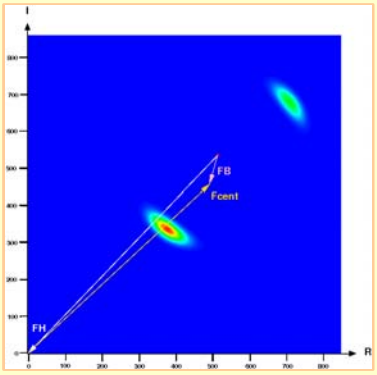
SHARP: 2D phase probability



- centroid structure factor
- best map

- Fcent = light atoms only

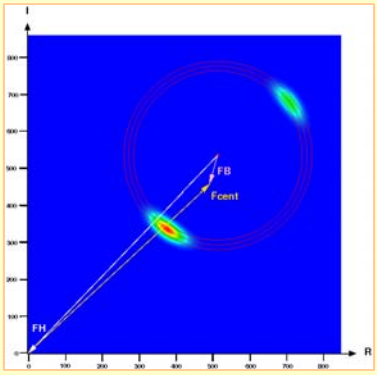
SHARP: 2D phase probability



- centroid structure factor
- best map

- FB = all atoms (incl HA)

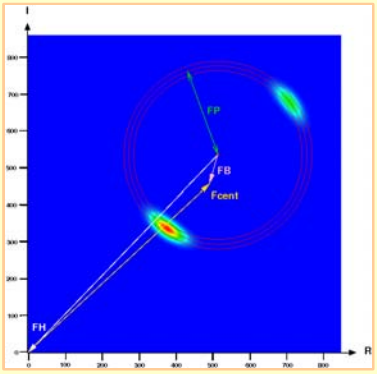
SHARP: 2D phase probability



- fitting HL description into picture

- calculating HLs on this circle

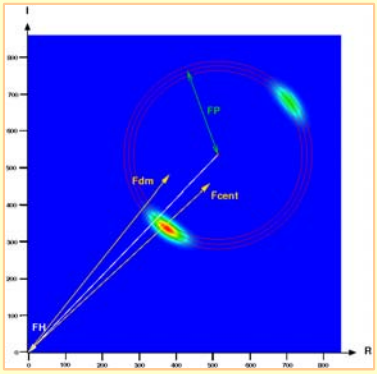
SHARP: 2D phase probability



- FP/SIGFP
- amplitude and sigma

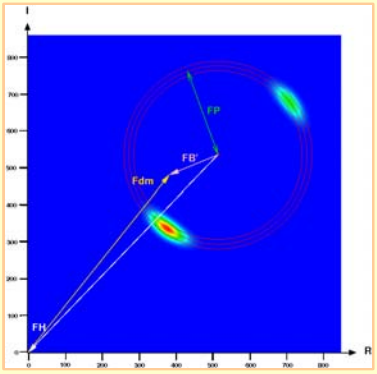
- HLs in sync with FP/SIGFP

SHARP: 2D phase probability



- calculating map Fcent
- density modified map Fdm

SHARP: 2D phase probability

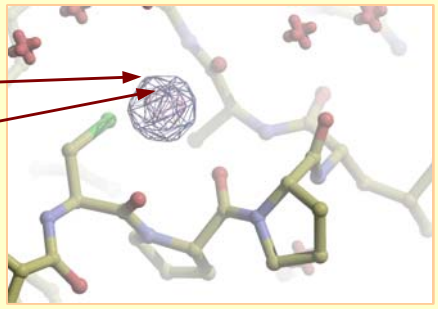


- calculating map Fcent
- density modified map Fdm

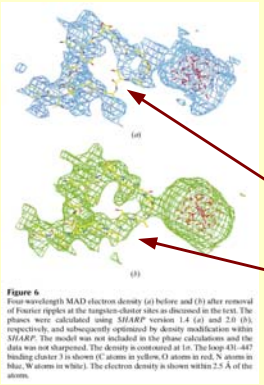
- shift of origin (FH) = FB'
- phase combination with HLs

Density modification with offset

- Hg site at 15sigma
- with offset
- without offset



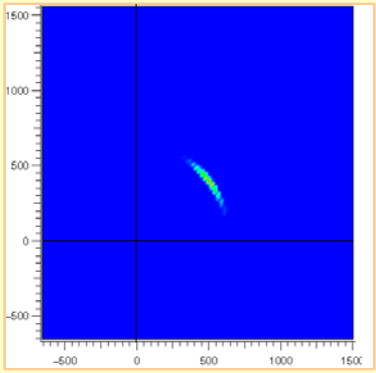
Density modification with offset



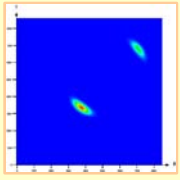
G. Rudenko, L. Henry, C. Vonrhein, G. Bricogne, and J. Deisenhofer. 'MAD'ly phasing the extracellular domain of the LDL receptor: a medium-sized protein, large tungsten clusters and multiple non-isomorphous crystals. Acta Crystallogr D Biol Crystallogr, 59(Pt 11):1978-86, 2003.

- breaks
- connectivity
- 4A resolution
- W-clusters

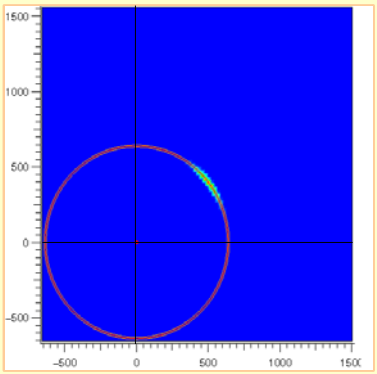
SHARP: 2D phase probability



- SIRAS (nat + Hg)
- single peak in 2D



SHARP: 2D phase probability

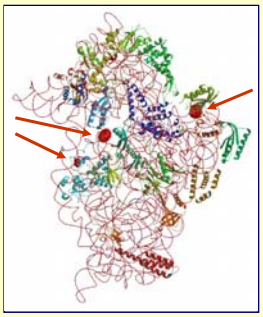


- SIRAS (nat + Hg)
- FP centred at (0,0)

- Fcent == FB

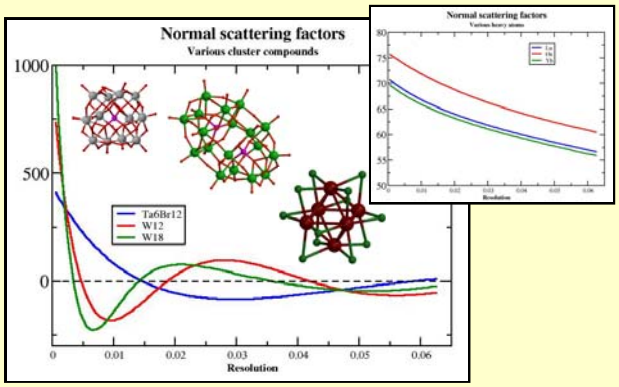
Phasing with HA clusters

Os - 93 sites

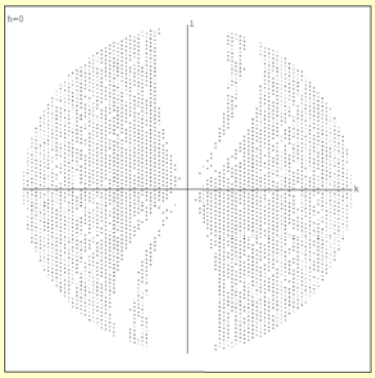


W18 - 3 sites

Spherical averaged form factors



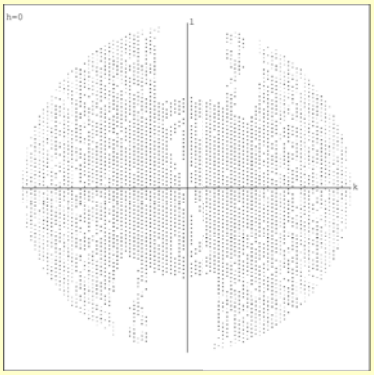
Completeness – "merging" in SHARP



- SIRAS (nat + Hg derivative)
- native data
- wedge missing
- **87%** completeness

- density modification
- SOLOMON
- optimising solvent content
- scores: CC(E), contrast

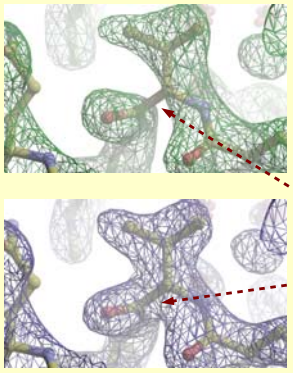
Completeness – "merging" in SHARP



- SIRAS (nat + Hg derivative)
- native + Hg
- FP from SHARP
- **98%** completeness (2.25A)

- density modification
- SOLOMON
- optimising solvent content
- scores: CC(E), contrast

Completeness – "merging" in SHARP

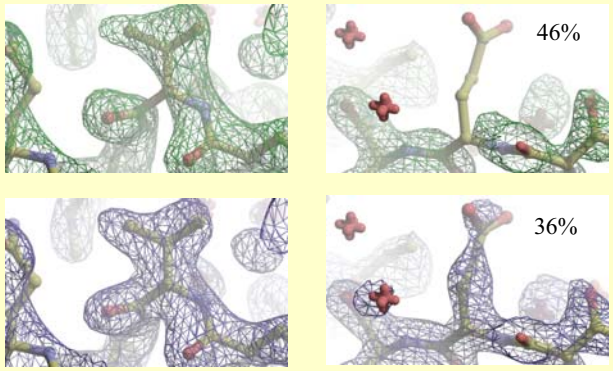


- SIRAS (nat + Hg derivative)
- phasing (SHARP)
- density modification (SOLOMON)
- 20 – 2.25 A

HKL for which native data measured:
- gaps in main-chain

all HKL (nat or Hg measured):
- connectivity

Completeness – "merging" in SHARP



Refinement

- ❑ model
- ❑ measured data: e.g. F_nat, F_hg
- ❑ experimental phase information: Hendrickson-Lattmann coefficients from SHARP

- ❑ But: HLs are only really in sync with FP/SIGFP
- ❑ BUT: what reflection file should I submit to PDB?
- ❑ **BUT:** what should we put into Table 1 to describe data?
- ❑ **BUT:** what will reviewer say?

- ❑ use of correct form factor for atoms at given wavelength!

Refinement

- 1GXT (Rosano et al, 2002)
- native (20 – 1.25 A)
- Hg derivative (20 – 2.25 A)

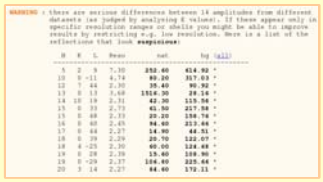
- SHARP: SAD or SIRAS

Refinement – choice of amplitude/phase

run	amplitude	phases	R	Rfree
01	F_Hg	SAD	0.147	0.189
02	F_Hg	-	0.145	0.197
03	F_nat	SAD	0.162	0.220
04	F_nat	-	0.152	0.233
05	**FP**	**SAD**	**0.147**	**0.189**
06	FP	-	0.145	0.197
07	F_nat	SIRAS	0.161	0.208
08	F_nat	-	0.151	0.237
09	FP	SIRAS	0.181	0.231
10	FP	-	0.171	0.248

- SAD:
  - FP == F_Hg
  - FP better than F_nat
- SIRAS:
  - FP = F_nat/F_Hg
  - F_nat better than FP

Refinement – choice of amplitude/phase



- SIRAS: comparing F_nat and F_Hg:
  - differences in E values > 10
  - R-factor > 30%
- scaling should be done on unmerged data

1/resol^2	Res	NRef	<FP**2>	Sc_kraut	SCALE	RFAC	RF_I	Wted_R	<diso>	max(diso)
0.016	7.9	90	4654371.	0.951	0.859	0.359	0.657	0.165	684.1	2159
0.048	4.6	421	3211461.	1.011	0.901	0.338	0.577	0.138	524.3	2403
0.080	3.5	583	3820756.	0.991	0.916	0.283	0.493	0.144	487.0	1914
0.112	3.0	738	2397294.	0.997	0.901	0.325	0.553	0.123	439.1	1938
0.144	2.6	852	1124225.	1.005	0.877	0.369	0.665	0.087	345.9	1625
0.176	2.4	985	818771.	1.002	0.882	0.360	0.651	0.081	288.2	1377
0.208	2.2	646	638860.	0.993	0.858	0.398	0.710	0.094	277.6	1500
THE TOTALS		4315.	1841172.	0.996	0.895	0.339	0.577	0.108	382.0	2403.

WARNING : the low resolution bin shows at least one very high R-factor for data from nat vs hg. This could be either real signal, severe non-isomorphism or a problem with the low-resolution data of one of the two datasets.

Conclusions

- automation:
  - especially helpful for repetitive tasks and creating template
  - needs to provide good feedback (especially about problems and ill-defined steps)
- data analysis
  - fast
  - lots of feedback (if presented nicely)
- structure factors, maps, phase combination, HLs
  - use correct combination
  - decision can be task specific
- heavy atom model
  - important for phasing (obviously)
  - visible in density modification
  - part of model (and data) during refinement

Acknowledgement

Global Phasing, Cambridge (UK):

- Gérard Bricogne
- W. Pacziorek, C. Flensburg, M. Schiltz
- O. Smart, T. Womack, M. Brandl, P. Keller
- Previous group-members: E. Blanc, P. Roversi, G. Evans, R. Morris

http://www.globalphasing.com/

SHARP/autoSHARP (free for academics): SIR, MIR, SAD, MAD, MAD+native, MAD+native+derivative,…

Automated refinement for protein crystallography

2008 Kyoto Crystallography Computing School

Min Yao

Laboratory of X-ray structural biology
Faculty of Advanced Life Science,
Hokkaido University, 060-0810, Sapporo, Japan

Outline

1. What dose LAFIRE do
2. Partial model building in LAFIRE
   - How do we make mathematic model from actual problem
   - Algorithms in LAFIRE
     - Tree/Binary Tree
     - Dynamic programming
3. Future of LAFIRE

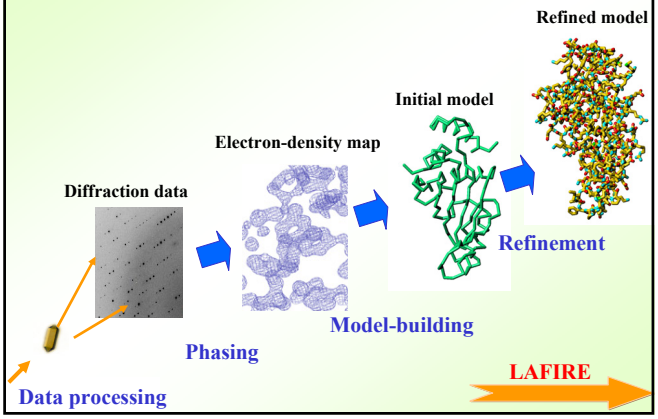
Yao, et al., Acta Cryst. D62, 189-196 (2006) *Zhou, et al., J. Appl. Cryst. 39, 57-63 (2006)*

Outline

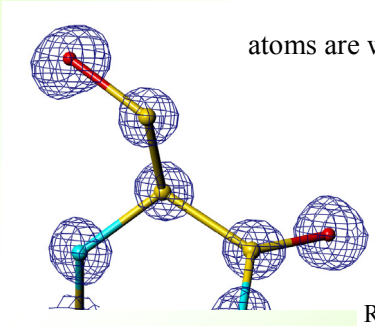
1. What does LAFIRE do
2. Partial model building in LAFIRE
   - How do we make mathematic model from actual problem
   - Algorithms in LAFIRE
     - Tree/binary tree
     - Dynamic programming
3. Problems and future of LAFIRE

Yao, et al., Acta Cryst. D62, 189-196 (2006) *Zhou, et al., J. Appl. Cryst. 39, 57-63 (2006)*

X-ray Crystallography (after data collection)



Electron density of small-molecule crystals

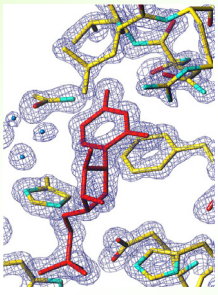


atoms are well separated

Resolution: 0.54 Å
1EJG

Electron density of protein crystals

$\rho(\mathbf{r}) = (1/V)\Sigma|\mathbf{F(h)}|_{obs}i\alpha \exp 2\pi i(-\mathbf{hr})$



Atoms are not separated
Low resolution data

Noisy
Error in phases

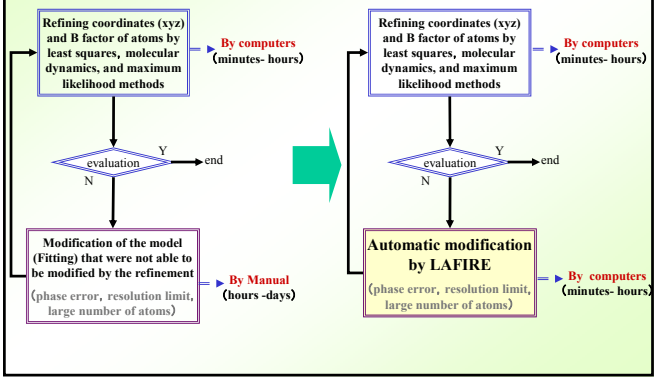
Initial model:
Model is traced but not complete
Mosaic structure with accurate and less accurate regions

Refinement:
Low parameter-to-observation ratio

◆Using stereochemical information, such as bond lengths, bond angles, torsion angle, and secondary structures

◆Using graphics to detect the problematic regions and adjusd them based on the electron

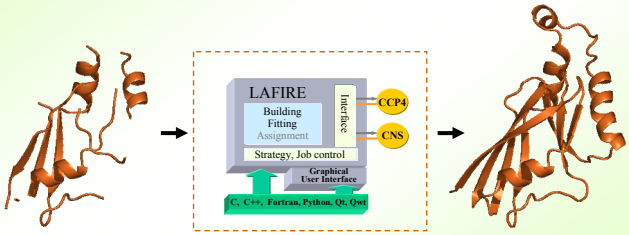
Refinement of protein structure



LAFIRE

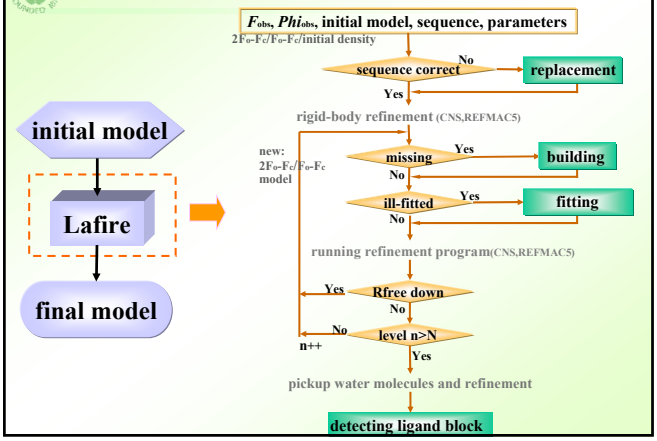
(Local-correlation-coefficient-based Automatic FItting for Refinement)

LAFIRE is essentially a simulation program that automatically performs the refinement process usually performed by experienced crystallographers.



1. **Building partial model**
2. **Fitting ill-match residues**
3. **Interface with refinement program and job control**

Refinement flowchart in LAFIRE

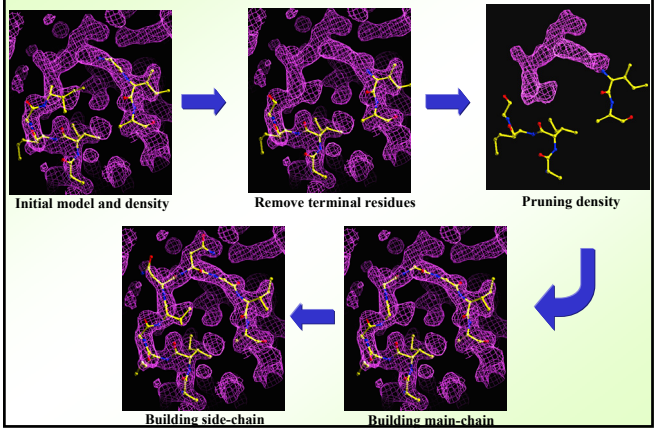


Outline

1. What does LAFIRE do
2. Partial model building in LAFIRE
   - How do we make mathematic model from actual problem
   - Algorithms in LAFIRE
     - Dynamic programming
     - Tree/Binary Tree
3. Future of LAFIRE

Yao, et al., Acta Cryst. D62, 189-196 (2006) Zhou, et al., J. Appl. Cryst. 39, 57-63 (2006)

Example of building by LAFIRE



How do we make mathematic model from actual problem

Partial model building

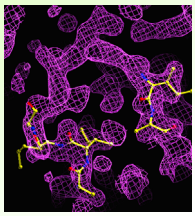
1. Simplify the problem to make it describable by a mathematic model and solvable

   Consider our problems as a problem of image processing and pattern recognition

   3D -> 2D -> 1D

2. Find algorithms to solve problems in a polynomial time

Partial model building



1. Pruning Map

$P := \{x \mid \rho(x) > \text{threshold}\}$

$P_u := P_1 \cap P_2 \cap \ldots \cap P_n$

$P_E := \{P_i \mid P_i \text{ contains existing model}\}$

2. Locating begin/target points

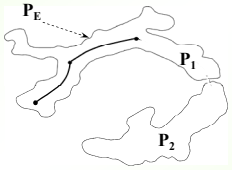
C_α^{i-1}, C_α^{j+1}

$P_u(C_\alpha^{i-1}) = P_u(C_\alpha^{j+1})$

$P_u(C_\alpha^{i-1}) \neq P_u(C_\alpha^{j+1})$

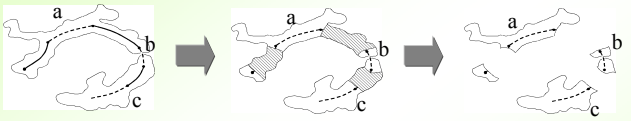
$P_u(C_\alpha^{i-1})$

$P_u(C_\alpha^{j+1})$



Partial model building

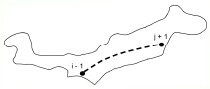
solid line: existing model: dotted line: missing parts



Locating begin and the target points

a

Missing part lies in one map segment



$P(C_\alpha^{i-1}) = P(C_\alpha^{j+1})$

x and *y* are linkable

$x, y \in G$

b, c

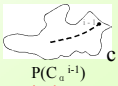
Missing part lies in more than one map segment

Missing part is terminus

$P(C_\alpha^{i-1}) \neq P(C_\alpha^{j+1})$

x and *y* are unlinkable

$x \in G1, y \in G2$



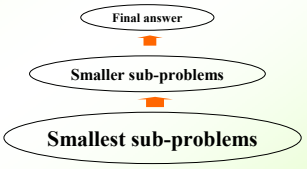
$P(C_\alpha^{i-1})$

no y in the segment

$x \in G$

Dynamic programming

- A problem is divided into a collection of sub-problems.
- Tackling of the smallest ones helps figure out larger ones.
- The final answer is obtained when all sub-problems are solved.

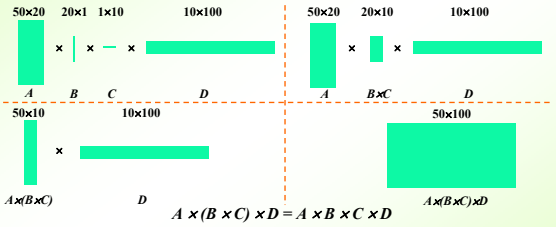


Reduce complexity
Solve problem in a polynomial time

Example : Matrix multiplication

Matrix multiplication is associative but not commutative.

$A \times (B \times C) = (A \times B) \times C$ $\quad A \times B \neq B \times A$



$A \times (B \times C) \times D = A \times B \times C \times D$

Since multiplying an m×n matrix by an n×p matrix takes mnp multiplications.

Multiplication order	Multiplications	Total
A×((B×C)×D)	20×1×10+20×10×100+50×20×100	120,200
(A×(B×C))×D	20×1×10+50×20×10+50×10×100	60,200
(A×B)×(C×D)	50×20×1+1×10×100+50×1×100	7,000

Solution by dynamic programming

Input : Matrices $A_1, \ldots, A_n$ with dimensions $m_0 \times m_1, \ldots, m_{n-1} \times m_n$

Output : Optimal order for computing $A_1 \times \ldots \times A_n$

Problem : C(i, j) ≡ minimum cost of multiplying $A_i \times \ldots \times A_j$

Sub-problem : $A_i \times \ldots \times A_k$ and $A_{k+1} \times \ldots \times A_j$

$C(i, j) = \min\{C(i, k) + C(k + 1, j) + m_{i-1} \times m_k \times m_j\}$ $(i \leq k < j)$

Following this function, a program of pseudocode can be made

```
for i = 1 to n: C(i, i) = 0
for s = 1 to n – 1:
    for i = 1 to n – s:
        j = i + s
        C(i, j) = min(C(i, k) + C(k + 1, j) + m_{i-1} × m_k × m_j : i ≤ k < j}
return C(1, n)
```

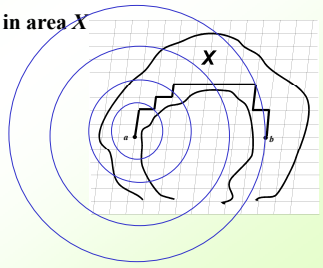
Calculate the length between two grids in the case b and c of Partial model building

Dynamic programming method is applied

Problem: $d_X(a, b)$ in area *X*

Sub-problem: $d_X(a, *) < d_X(a, b)$ in area *X*

1. Initially, $d_X(a, a) = 0$, $d_X(a, *) = \infty$
2. Add *a* to set *S*
3. Find $c \in S$ to make $d_X(a, c)$ minimum.
4. If $c = b$, output $d_X(a, b)$ and finish.
5. Otherwise
   Assign the $d_X(a, d)$ where *d* is the neighbour of *c*, $d = c \pm \Delta$
   $d_X(a, d) := \min\{d_X(a, c) \pm \Delta, d_X(a, d)\}$
6. Remove *c* from *S*
7. Loop from 3

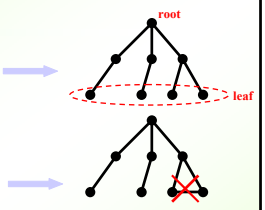


$path_X(a, b)$: the path all lies in set *X*

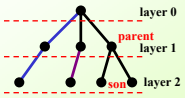
$d_X(a, b) := \min(d(path_X(a, b)))$

Tree

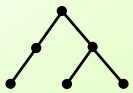
- Tree is a data structure composed of vertices (V) and edges (E).
- A vertex is defined as root and vertices at the extreme are defined as leaves.
- Tree is a non-cycle structure, and there is only one path linking two vertices.



- With the definition of the root and leaves, the tree can be divided into several layers.
- The vertices linking the same vertex of the upper layer are defined as the sons of this vertex.

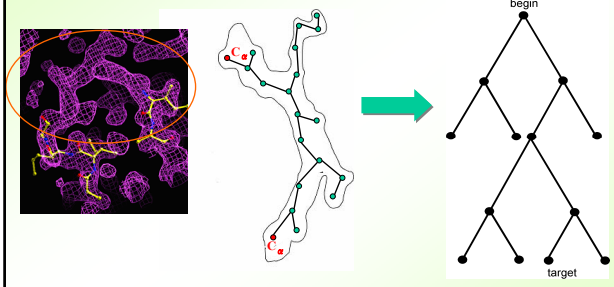


- A tree in which each vertex has at most two sons is defined as binary tree.



Trace main-chain using tree search

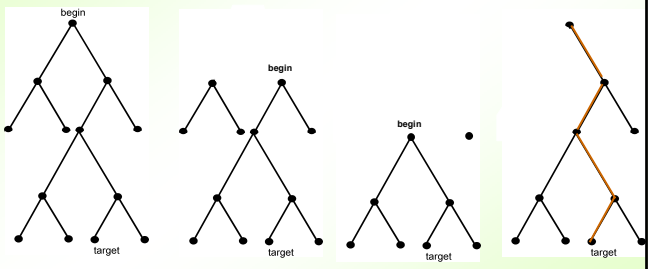
Electron-density map **2D project** **Tree**



- trace main-chain between known terminals
- a tree with begin and target points
- locate the path between the begin and target point

Problem resolution

(Locate a path as backbone in a tree)

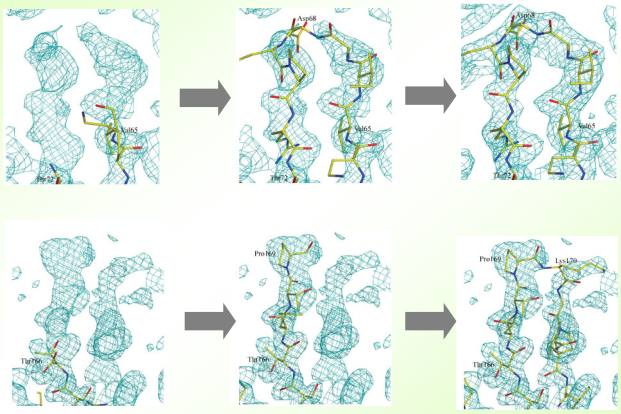


Determine begin and target points

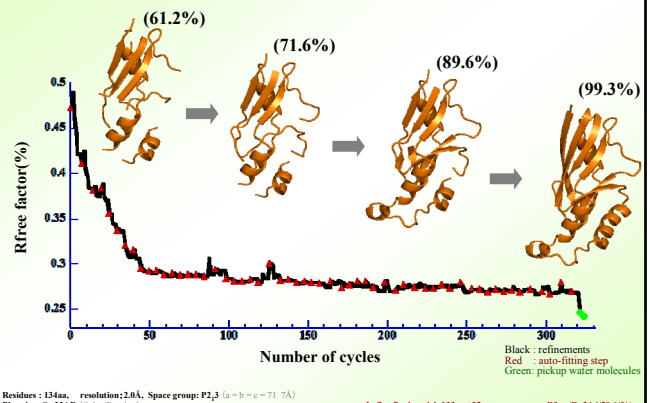
- Remove the begin point
- Two tree are generated
- Find the tree include the target point

- Locate an edge
- Shift the begin point

Results



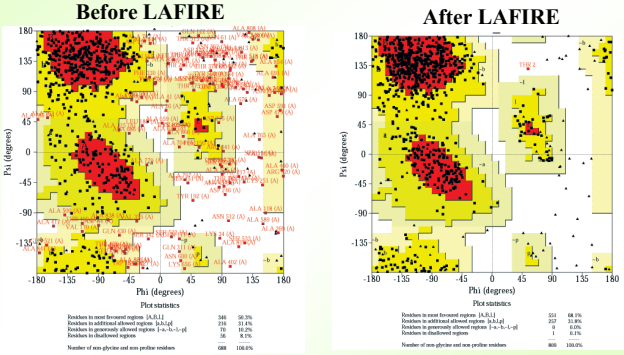
An example of refinement using LAFIRE



Dihedral optimization (Ramachandran Plot, 3.0 Å)

Manual Building

Before LAFIRE **After LAFIRE**



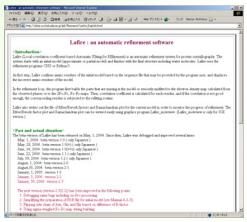
Otto, *et. al.*, *J. Biol. Chem.*, **280**, 17339 (2005)

LAFIRE 2.6 (SGI, Linux) is released

The versions of Windows and Mac with GUI are coming

http://altair.sci.hokudai.ac.jp/g6/Research/Lafire_English.html

http://altair.sci.hokudai.ac.jp/g6/Research/Lafire_Japanes.html



lafire@castor.sci.hokudai.ac.jp

Outline

1. What does LAFIRE do
2. Partial model building in LAFIRE
   - How do we make mathematic model from actual problem
   - Algorithms in LAFIRE
     - Tree/binary tree
     - Dynamic programming
3. Future of LAFIRE

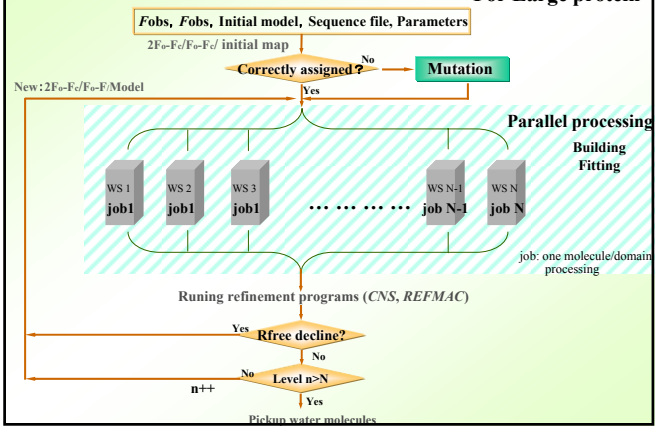
Yao, et al., Acta Cryst. D62, 189-196 (2006) *Zhou, et al., J. Appl. Cryst. 39, 57-63 (2006)*

Problems and future of LAFIRE

- **Calculation time**
  Parallel processing
- **Ligand**
  Ligand
  DNA/RNA
- **Graphics User Interface (GUI)**

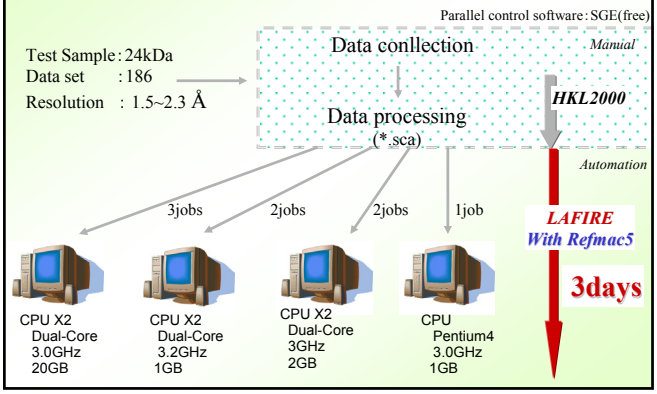
Parallel processing of LAFIRE(1)

For Large protein



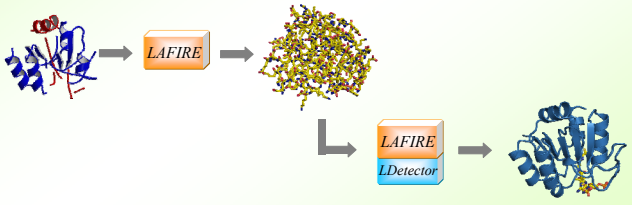
Parallel processing of LAFIRE(2)

For SBDD/FBDD



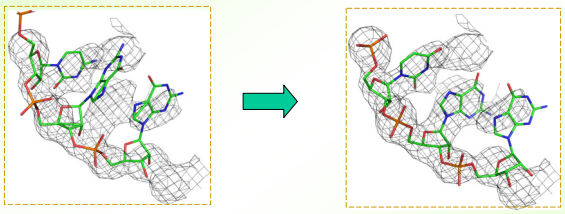
Ligand Fitting of LAFIRE

After refinement, the blocks that may be ligand will be detected and the ligand will be fitted in the blocks.



LDetector: Ligand Detector

Fitting DNA/RNA (under development)



*R*free = 38.64%

*R*free = 36.64%

GUI of LAFIRE

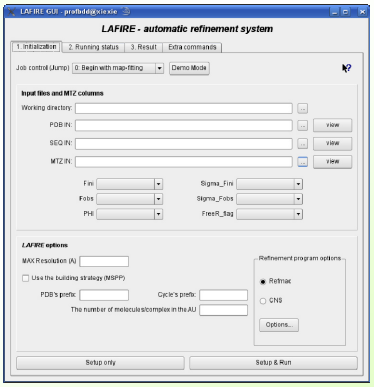
Using Python, PyQt, PyQwt

Main window

Setup
- working directory
- initial model
- amino acid sequence
- diffraction data
- options for refinement are

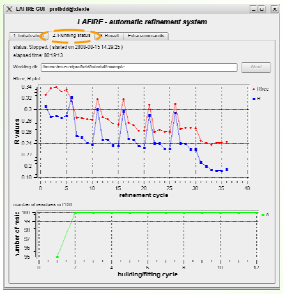
Check
- all information

Running
- the refinement process

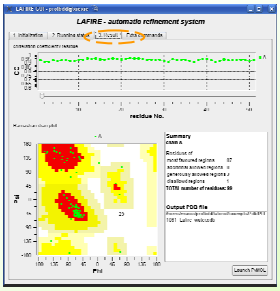


GUI of LAFIRE

Real time monitor
- R and FreeR factor
- The number of built residues of each molecule in an ASU



Final result
- The correlation coefficient of the main chain
- Ramachandrant plot
- the final structure and maps can be shown through PyMol



Acknowledgment

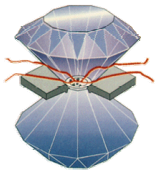
Prof., Isao Tanaka
PhD., Yong Zhou
MC student, Keitaro Yamashita

All users

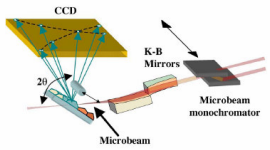
Hokkaido University

GeoSoilEnviroCARS
The University of Chicago
Argonne National Laboratory • Advanced Photon Source

Development of novel synchrotron single-crystal XRD techniques for high-pressure science



Przemyslaw Dera
GeoSoilEnviro CARS,
The University of Chicago



Kyoto Crystallographic Computing School, Kyoto, Japan, August 23, 2008

Main fields of application of high pressure research

Mineral physics
simulating conditions of the Earth and planetary interior

Physics
metallization
superconductivity

Chemistry
supercritical solvents, topochemical reactions, new material synthesis

Pharmacy
polymorphism of pharmaceutical compounds

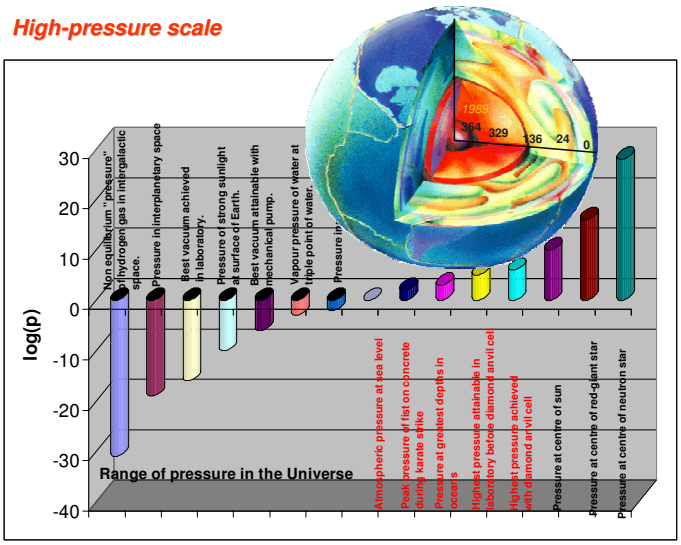
Material science
superhard materials
nanostructured materials
stress/strain

Biology
protein polymorphism,
Xe derivatives,
lowering B,
"ordering" – decreasing mosaic spread,
macromolecular "mechanics"

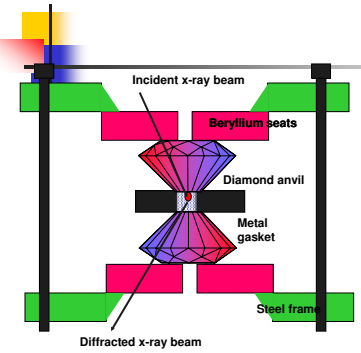
Technological applications
food processing

High-pressure science recognized as one of the future directions of focus at ESRF, APS, NSLS)

High-pressure scale



High pressure apparatus



- Sample is immersed in hydrostatic liquid, which freezes at some point during compression (usually below 10 GPa).
- Diffraction pressure calibrant is placed in the sample chamber along with the sample.
- Both incident and diffracted beam travel though diamonds, Be disks, pressure medium, and sample.

High-pressure crystallography goals

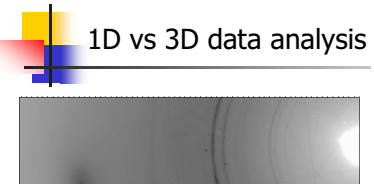
- Provide **precise** (<0.1% error) information about the changes of unit cell parameters as a function of pressure (compressibility).
- Provide **precise** information about the structural changes in a continuous compression regime, to allow studying compression mechanisms (changes in bonding).
- Provide a way of determining (**new**) structures of unquenchable high-pressure phases, which exist only at high pressure (no prior knowledge of unit cell, space group, structure factors).

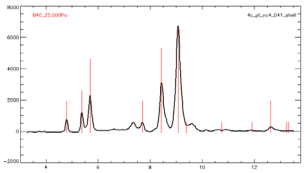
High-pressure crystallography challenges

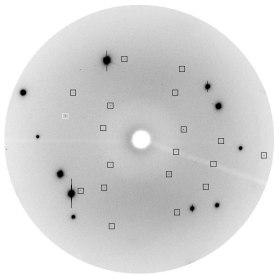
Experimental challenges:
- Very small sample (<0.01mm)
- Angular access restricted (low completeness)
- Absorption limits the incident energy to >15 keV
- Absorption and extinction affect intensity measurement
- High background (scattering)
- Multiple SXD diffraction signal
- Contamination by PXD signal
- Poor sample quality (strain, multi-grain assemblages)

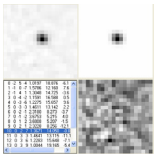
- **Current high-pressure "mind-set" on crystallography:**
  - Dominating tool (95% studies) is powder diffraction. It is sufficient for simple solids, but usually fails for more complex/interesting cases.
  - Use of polychromatic radiation limited to powder energy-dispersive experiments. Main reasons – lack of customized software and "fear" of intensity calibration.

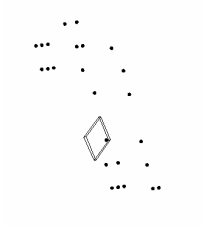
1D vs 3D data analysis



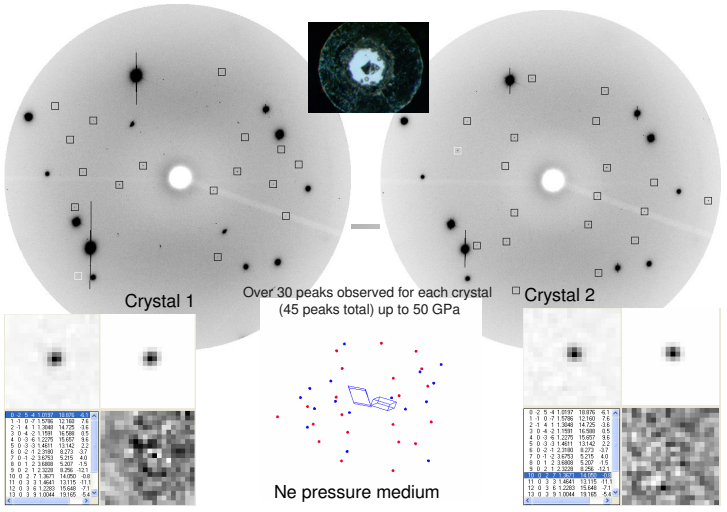


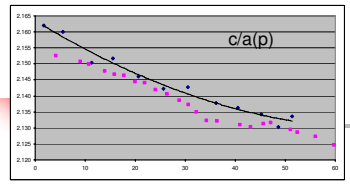




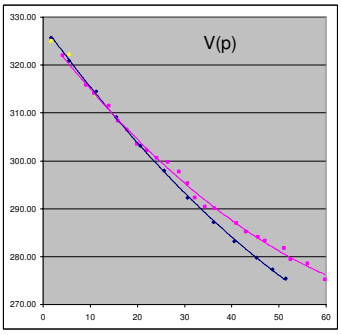


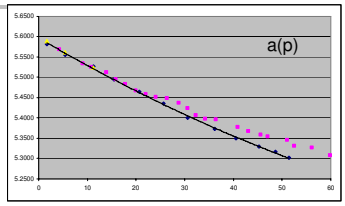
Single-crystal XRD study of B_4C to 50 GPa

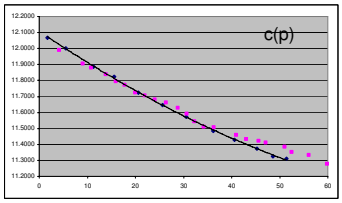




Single-crystal XRD study of B_4C to 50 GPa







What is an SXD experiment

- **Diffraction experiment performed with sample that contains one or more single-crystal grains with size of >0.001mm**
- **Data collection involves measurement of directions and intensities of individual diffracted beams and corresponding crystal orientations for a large population of reciprocal space vectors.**
- **Reciprocal space is reconstructed in three dimensions, allowing for unambiguous indexing**
- **Measured peak intensities are free from spatial and energy overlaps (or efficient ways to deal with these overlaps are available)**

GOAL 1: make mSXD possible at any monochromatic high-pressure synchrotron station equipped with area detector and rotation stage (instrument control and data analysis)

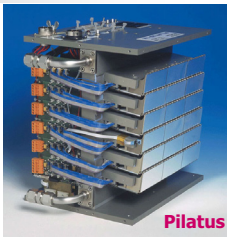
GOAL 2: Create custom (hardware and software) SXD solutions optimized for high-pressure experiments

mSXD with area detector

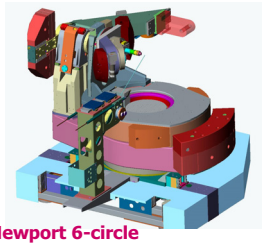
- **Center the sample on rotation axis and with the beam**
- **Collect diffraction images while rotating the sample**
- **Determine detector coordinates and sample orientations for each diffraction peak**
- **Reconstruct the reciprocal space in 3-d**
- **Determine the orientation matrix (index)**
- **Predict peak positions in recorded diffraction images**
- **Retrieve peak intensities (structure factor amplitudes)**
- **Solve/refine the structure**

mSXD with Pilatus

- **6-circle kappa geometry**
- **All axes have DC-servo motors with speeds up to 16 degrees/second, allowing on-the-fly scanning.**



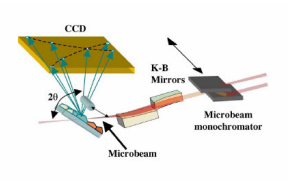
Pilatus

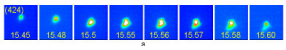


Newport 6-circle

- **Fixed-energy monochromatic beam at 10, 15 and 30 keV**
- **The expected focal spot size 0.025x0.025 mm.**
- **Super fast PILATUS detector (200 Hz data collection capabilities)**
- **Programmable energy threshold**

vmSXD with area detector



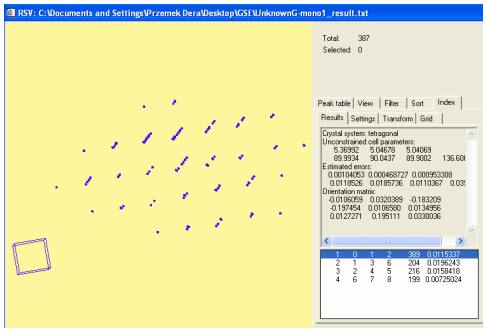


Ice, Dera et al. J. Synchr. Rad. (2005)

1. Scan the sample with white beam to find the most promising grain (Laue)
2. Perform series of monochromatic exposures at varied energies that cover whole or part of the incident spectrum
3. Assign peak energies using appearance and disappearance of peaks

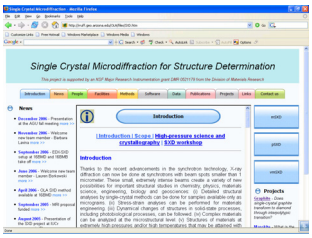
Challenge: The method was designed to study strain in known phases, requires knowledge of the unit cell

vmSXD with area detector: reciprocal space reconstruction



MRI project scope

- Budget of ~$0.8M for three years
- Two synchrotrons, three beamlines
  APS (GSECARS + HPCAT), ALS (CALYPSO)
- Three main techniques mSXD, vmSXD, pSXD
- New hardware (dedicated goniometers, area detectors)
- Customized and uniform software (IDL, EPICS)

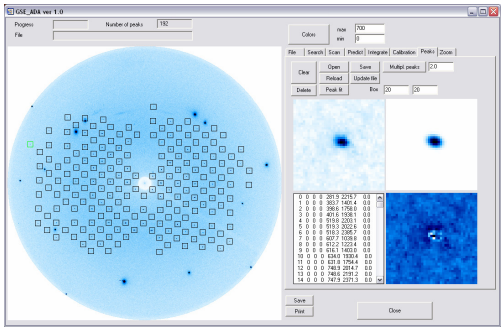


For more information, please visit:
http://rruff.geo.arizona.edu/OLA/files/SXD.htm

Custom SXD software: goals

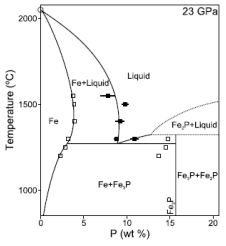
- Goals:
  - Flexible and custom-designed for high-pressure experiments
    - High background
    - Signal from multiple crystallites
    - Strain
    - Unknown phases
    - Poor quality
    - Diffuse scattering
  - Fast – real time data analysis to guide the experiment (single experiment is usually composed of 10-20 pressure points)
  - Object oriented
  - Open source
  - User friendly/intuitive
  - Multi format (for diffraction images)
  - Multi platform (IDL)
  - Combining analysis of monochromatic, polychromatic and variable monochromatic data
  - "Parametric" approach - take advantage of the information that changes little in pressure-series experiments
  - Compatible with EPICS instrument control protocol

Custom SXD software: GSE_ADA



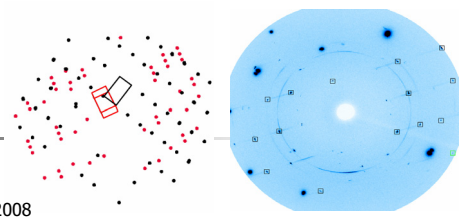
Light element phases in the core

- Evidence:
  - seismology
  - meteorites (cosmic abundances)
- Fe, Ni, Co
- Si, S, C, P, N, H
- Fe-P system
  - Fe_3P schreibersite
  - Fe_2P barringerite/allabogdanite
  - FeNi ironphodsphide

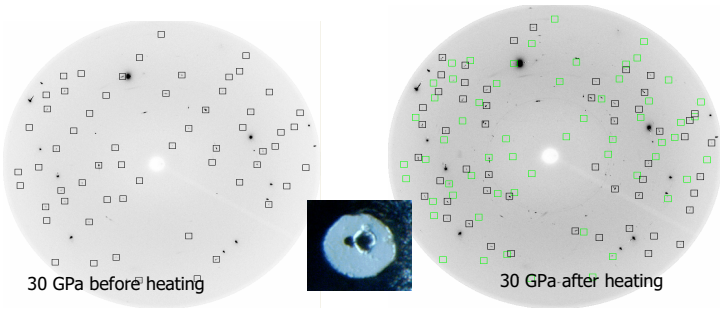


From Stewart and Schmidt, 2007

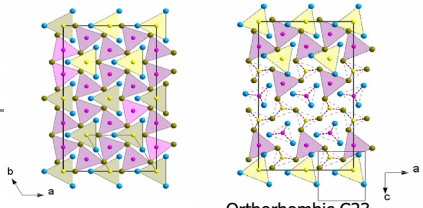
Fe_2P

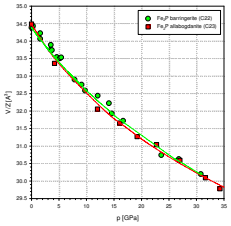


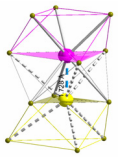
Dera et al. Geoph. Res. Lett, 2008



Fe_2P







Dera et al. Geoph. Res. Lett, 2008

Fe_2P and core density deficit

	Barringerite (C22) Fe_2P	Allabogdanite (C23) Fe_2P
a_0 [Å]	5.8764(3)	5.7768(4)
b_0 [Å]	5.8764(3)	6.645(3)
c_0 [Å]	3.4494(1)	3.5937(4)
V_0 [Å^3]	103.16(1) (Z=3)	137.95(3) (Z=4)
K_0 [GPa]	174(7)	177(3)
K' []	4.0 (fixed)	4.0 (fixed)

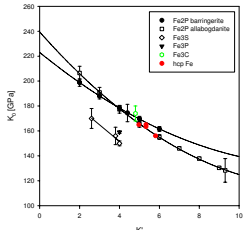
Conclusions:

•Phase transition from C22 to C23 type structure occurs on heating above 1000K at pressures above 8 GPa

•Both C22 and C23 phases can be metastably compressed/decompressed by at least 30 GPa

•The metastable orthorhombic phase heated at ambient pressure converts to hexagonal

•Occurrence of the orthorhombic phase in natural samples provides a very useful geobarometer/geothermometer



Summary

- SXD method is a very powerful technique that can be used to reveal details of even very complicated crystal structures at high pressure, complementary and usually superior to powder diffraction.
- SXD can provide much more precise information for EOS studies
- The data collection process can be as simple as that of powder experiments.
- All of GSECARS experimental stations offer now routine SXD capabilities.
- SXD experiments in 50-100 GPa range are now routinely performed at GSECARS
- At GSECARS SXD can be combined with heating (laser or resistive) and on-line spectroscopy (Brillouin and Raman)

- **SXD data analysis software developed within the framework of the MRI grant freely available to high pressure community.**

Acknowledgements

MRI postdocs
L. Borkowski
B. Lavina

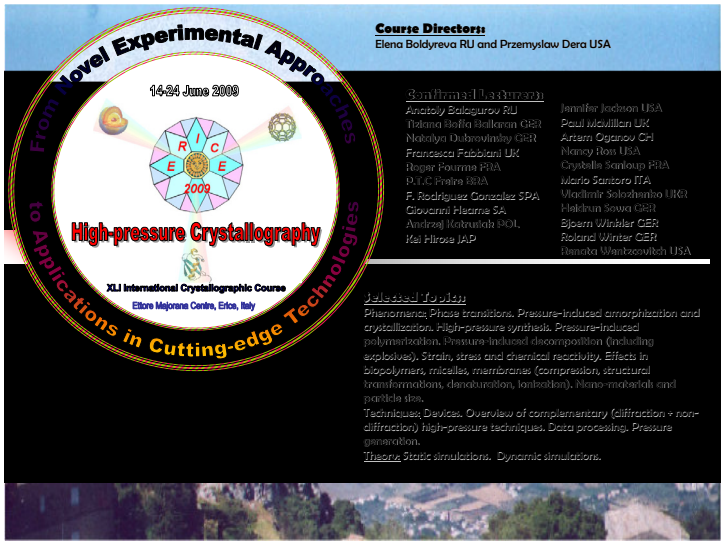
MRI grant
M. Nicol
R.T.Downs

COMPRES
NSF DMR MRI program

HPCAT:
G. Shen, P. Liermann, W.Yang

GSECARS:
M. Rivers
W. Prakapenka
Y.Wang

UNLV:
O.Tschauner



A Tale of COD

Saulius Gražulis
Kyoto, Osaka 2008

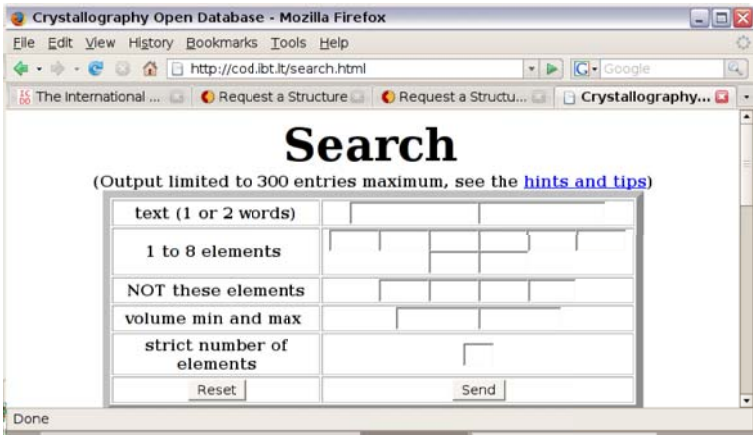
Crystal structure databases

- PDB, NDB
  - open access
  - free to be copied, under certain conditions
  - serve as base for numerous derived databases
- CSD, ICSD, ICDD/PDF, CRYSMET
  - proprietary, subscription based
  - copying is not permitted
  - contrast the situation with PDB...

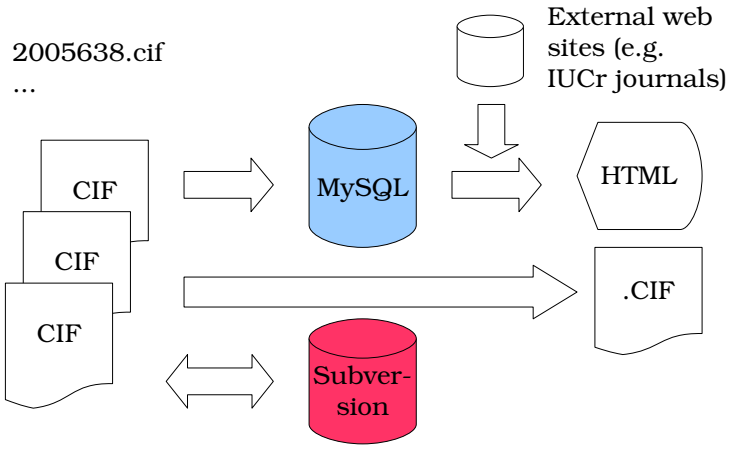
The COD way

- "All data on this site have been placed in the **public domain** by the contributors" (http://www.crystallography.net)
- Updated daily: 71250 entries in the COD
- Collect published structures from peer-reviewed journals and donations by well-established crystallography labs
- Provide high quality data: syntax clean CIF files; increasingly stringent validation tests

COD search interface



COD organisation



COD data sources

- Donations:
  - Mineralogical Society of America
  - Mineralogical association of Canada
  - Laboratoire de Cristallochimie et Physicochimie du Solide
  - Laboratoire de Cristallographie et Sciences des Matériaux CRISMAT
  - Laboratoire des Oxydes et Fluorures, Institue de physique del la Matière Condensée
- Donations by journals:
  - IUCr journals
- Collections by volunteers from the peer-reviewed journal supplementary data.

COD contents

- Structure data in CIFs, 1 structure/CIF
- Crystallographic data & coordinates (of course ;)
- Bibliography, chemical information, back-reference (original file)
- Empty, unrecognised, irrelevant, copyrighted tags are excluded

COD data deposition & quality checks

- Check syntax
- Check semantic consistency
- Check duplicates
- Split structures into separate files
- Add missing information (bibliography, etc.)
- Insert into the COD Subversion repository and into the SQL database

Problems with published data

- Impression: ~40% of published CIF files contain syntax errors; even syntactically correct files of recent years occasionally contain semantic problems...
- "These [i.e. syntactic and missing data] problems, which affect about 40% of incoming CIFs" (Frank H. Allen, The Cambridge Structural Database: ..., Acta Cryst. B, 2002, **58**, 380 – 388)

Access to COD data

- COD
  - http://www.crystallography.net
  - http://cod.ibt.lt/
  - svn://cod.ibt.lt/cod
  - rsync://cod.ibt.lt/cif
- PCOD
  - http://www.crystallography.net/pcod
  - svn://cod.ibt.lt/pcod

Community based effort?

- Wiki (Wikipedia) – like review, error correction, annotation, linking with other resources
- Invited editors and reviewers?
- Self-registered editors and reviewers?

COD applications

- Source of ligands for macromolecular crystallographers
- Collecting statistics (representative data subset?)
- Search-match software
- Teaching
- Software testing and validation

COD financing sources

- Volunteers and contributors
- COD Advisory Board
- Lithuanian Research Council
- Government funding?
- Private granting agencies?
- Company donations?

COD prospects

- Docking
- Software testing (after inclusion of Fobs data)
- Crystallographic publication validation and review
- Rational drug design
- QSAR
- Materials research

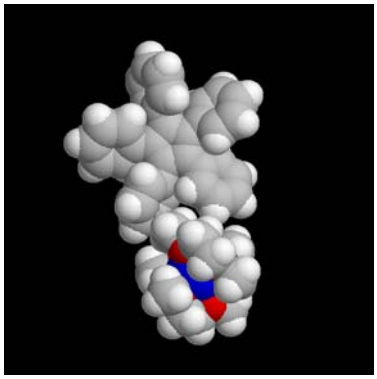
Data to knowledge

- Database is not yet knowledge...
- ... but it is nowadays an important prerequisite!
- Semantic webs?
- Automatic inference?

Acknowledgements

- Volunteers – data collectors
- Daniel Chateigner, Xiaolong Chen, Marco Ciriotti, Robert T. Downs, Armel Le Bail, Luca Lutterotti, Yoshitaka Matsushita, Peter Moeck, Miguel Quirós Olozábal, Hareesh Rajan, Alexandre F.T. Yokochi (COD Advisory board)
- Elena Manakova, Justas Butkus (IBT), Patrick Ducrot (ENSICAEN)
- Peter Strickland, Michael Dacombe (IUCr)
- IUCr for permission to automatically download the published CIFs
- In part financed by: Lithuanian Science Council Student Research Fellowship Award

Questions?



COD 2200075.cif: Holl et al. (2001), Acta Cryst. **E57**, m31-m32

Copyright issues

- Copyright covers works of authorship (novels, verse, sci. papers, computer programs)
- Copyright covers **only** the expression of ideas
- Copyright **does not** cover:
  - Ideas
  - (scientific) facts
  - Simple forms (i.e. ones that do not contain individual's “trace of the hand”)

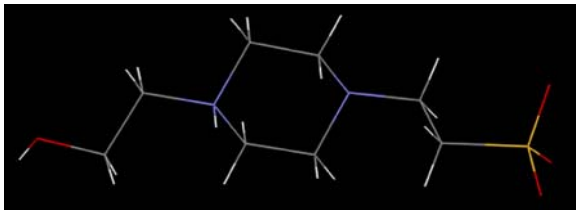
COD copyright policy

- Include data:
  - _atom_site_fract_x 0.333
- Exclude potentially copyrighted text:
  - _publication_text
    ;
    Introduction

    We have solved ...
    ;

The problem?

- Model glycerol/HEPES/MES/Tris into my protein structure:
  - need ideal(ised) coordinates:



COD 2005129.cif: Wouters et al (1996), Acta Cryst. C52, 1687-1688

Where are my (epps, HEPES) coordinates?

- Sources of coordinates:
  - Quantum mechanical calculations
    - tricky and time consuming
    - need verification
  - Idealised geometry
    - only well-known compounds
    - need verification
  - **X-ray diffraction experiment**
    - precise and accurate
    - time consuming, but ~500 000 molecules solved
    - **need access to published experimental data**

Crystallographic databases

- Structural information is scattered in several databases:
  - CCDC for organic molecules
  - ICSD for inorganic
  - ICDD/PDF for powder data
- All these databases are proprietary, subscription based "products"
- Contrast the situation with PDB or NDB...

Obtaining data from CCDC?

- "... CCDC provided a web form for data retrieval, which **requires** you to enter brief literature **citation** details and the CCDC **Deposition Number** (CCDCnnnnnn) which should appear in the paper"
  (http://www.ccdc.cam.ac.uk/products/csd/request/)
- Individual CIF data sets are provided ... on the understanding that they are used for bona fide research purposes only. They ... **may not be copied or further disseminated in any form**
  (http://www.ccdc.cam.ac.uk/products/csd/request/request.php4)

driving X-ray innovation

oxford diffraction
Now a part of Varian, Inc.

Q..

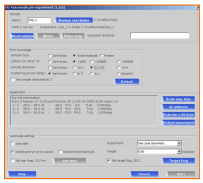
Data collection strategy

Mathias Meyer

Data collection strategy

- What are key parameters influencing data collection strategy?
- How advances in software and hardware can help in more efficient data collection?

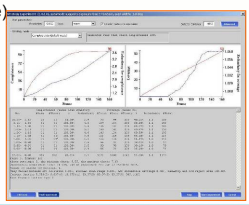
Diffraction experiment



- Pre-exp determines:
  - Unit cell and orientation
  - Laue class and extinctions
  - Diffraction power
- Crystallographer: How to (efficiently) collect the data?

Type of experiments

- Fastest possible
  - 'Complete data' (exploit symmetry)
  - Geometric objects (full, hemisphere)
- Charge density/absolute structure
  - Target redundancy
- 'I have 1hr'
  - Time constraint
- Absorption correction
  - Mixture of Complete data and geometric objects



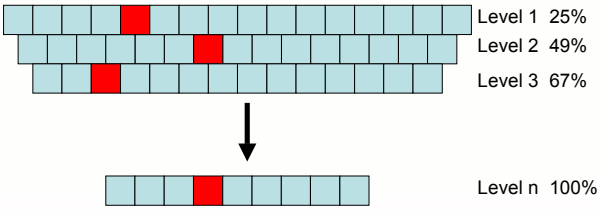
The strategy computation question

user constraints

distance, detector, collisions

Pool of runs
Typically 150-2000

The strategy computation question 2



oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

What is efficient?

- 'Fastest possible'

Completeness (under Laue symmetry) / Coverage (under P1)

Res	#Data	#Theory	%	Redundancy	#Total	#Data	#Theory	%	Redundancy	#Total
18.50- 1.74	409	415	98.55%	2.7	1108	644	828	77.78%	1.7	1123
1.74- 1.37	415	415	100.00%	2.6	1084	637	828	76.93%	1.7	1077
1.37- 1.19	415	415	100.00%	2.3	942	593	828	71.62%	1.6	946
1.19- 1.09	415	415	100.00%	2.0	850	585	828	70.65%	1.4	845
1.09- 1.01	415	415	100.00%	1.9	776	530	828	64.01%	1.5	775
1.01- 0.95	415	415	100.00%	1.7	719	546	828	65.94%	1.3	719
0.95- 0.90	415	415	100.00%	1.6	680	522	828	63.04%	1.3	677
0.90- 0.86	415	415	100.00%	1.5	623	500	828	60.39%	1.2	624
0.86- 0.83	415	415	100.00%	1.5	619	507	828	61.23%	1.2	617
0.83- 0.80	417	417	100.00%	1.4	576	496	830	59.76%	1.2	574
18.50- 0.80	4146	4152	99.86%	1.9	7977	5560	8282	67.13%	1.4	7977

runs: 7, frames: 520
theta settings: 2, abs minimum theta: 25.31, abs maximum theta: 25.93, max resolution: 0.734

Laue: 2/m; Friedel off, Target 0.800Ang

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Practical answer: OD strategy 3rd generation

- Cost function: Balancing new data intake vs. price for adding new runs
- Efficient computing kernel allowing 50-200 seeds/sec
- 'Deep Blue' simulation possible: Within 10% of optimal solution when evaluating 1000+ seeds at level 1-3.

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Curious questions: detector distance

2 times detector distance
=
2 * frames

- When considering completeness
- Side remarks:
  - Background $1/r^2$
  - dd has to be appropriate to allow lattice resolution
  - Collisions

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Curious questions: detector size

Detector	Unique speed	Observation speed
Eos	1	1
Atlas	1.3 – 1.6	1.6 – 1.8
Titan	1.4 – 1.8	2.0 – 2.2

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Curious questions: goniometer

- 3-circle vs. 4-circle
  - 3-circle incomplete at high resolution
  - 4-circle up to 30% more efficient due to extra degree of freedom in run pool

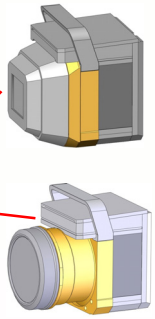
oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Curious questions: wavelength

- Cu vs. Mo
  - Same resolution, detector type, detector distance, completeness
  - Cu experiment has **4-5 times** more frames than Mo experiment

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Detector technology



- Eos & Atlas
  - Sensitivity
    - Eos: most sensitive – 330e^-/X_{Mo}
    - Atlas: 180e^-/X_{Mo}
  - Fast readout and low noise
  - High dynamic
  - Full 18-bit digitalization

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Instrumentation software

- Very efficient processing pipeline using 10+ threads of different priorities

↓

overhead close to theoretical minimum

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Through-put

- Experiment 360deg, dose time 7200s, correlated, detector with 90mm diameter

	Overhead	Lost time Relative to total	Total
Inefficient style 2-2.5s per readout, 0.3deg	5600s	44%	3h33
OD style (Sapphire3) <3s per readout, 1deg frames	2100s	23%	2h35
OD style (Eos) <0.75s per readout, 1deg frames	540s	7%	2h09

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Through-put using larger detector

- 90mm vs. 140mm

	Overhead	Lost time Relative to total	Total
Inefficient style 2-2.5s per readout 0.3deg	5600s	44%	3h33
OD style (Atlas) 250 deg = 5000s dose <0.75s per readout, 1deg frames	300s	6%	1h28
OD style (Atlas) Uncorrelated, 200 deg = 5000s dose <0.22s per readout, 1deg frames	44s	1%	1h24

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Through-put using larger detector

- 90mm vs. 140mm and Cu wavelength

	Overhead	Lost time Relative to total	Total
Inefficient style 2-2.5s per readout 0.3deg	5600s	44%	3h33
OD style (Atlas) 250 deg = 5000s dose <0.75s per readout, 1deg frames	300s	6%	1h28
OD style (Atlas) uncorr. Cu Nova 1250 deg = 1250s dose <0.22s per readout, 2deg frames	138s	10%	0h23

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Reliable multi theta data

- Enabled through
  - Efficient geometric strategy
  - Integration advances
  - Scaling + absorption correction improvements

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Time saving: A universal currency

- Shorter experiments
- More redundant experiment, i.e. higher data quality
- Smaller samples

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Summary

- Cu can be key element for efficient experiments
- ODs strategy & data reduction software have matured
- Size does matter
- Fast hardware and processing are key

oxford diffraction
Now a part of Varian, Inc.
driving X-ray innovation

Thank you

Call for Contributions to the Next CompComm Newsletter

The 10th issue of the Compcomm Newsletter is expected to appear around March of 2009 with the expected primary theme of 'Age Concern'. If no-one is else is co-opted, the newsletter will be edited by Lachlan Cranswick.

Contributions would be also greatly appreciated on matters of general interest to the crystallographic computing community, e.g. meeting reports, future meetings, developments in software, algorithms, coding, historical articles, programming languages, techniques and other news.

Please send articles and suggestions directly to the editor.

Lachlan M. D. Cranswick
CNBC
National Research Council of Canada
Building 459, Station 18,
Chalk River Laboratories,
Chalk River, Ontario,
Canada, K0J 1J0
E-mail: lachlan.cranswick@nrc.gc.ca
WWW: http://neutron.nrc.gc.ca/peep.html#cranswick