Monte Carlo indexing with McMaille

A. Le Bail
Université du Maine
Laboratoire des Fluorures
CNRS – UMR 6010
FRANCE

alb@cristal.org
http://cristal.org/

Content

- Introduction : indexing in the powder diffraction maze

-The SDPD Round Robin 2002 indexing part – ultra brief report

- McMaille (pronounce « MacMy ») demonstrations

- « Simplicity » of the Monte Carlo algorithm in McMaille

- McMaille v2 : Gaussian profiles

- McMaille v3 : Columnar peak shape

- About the (accuracy + enlarged peaks) apparent contradiction

- « Refinement » and optimization

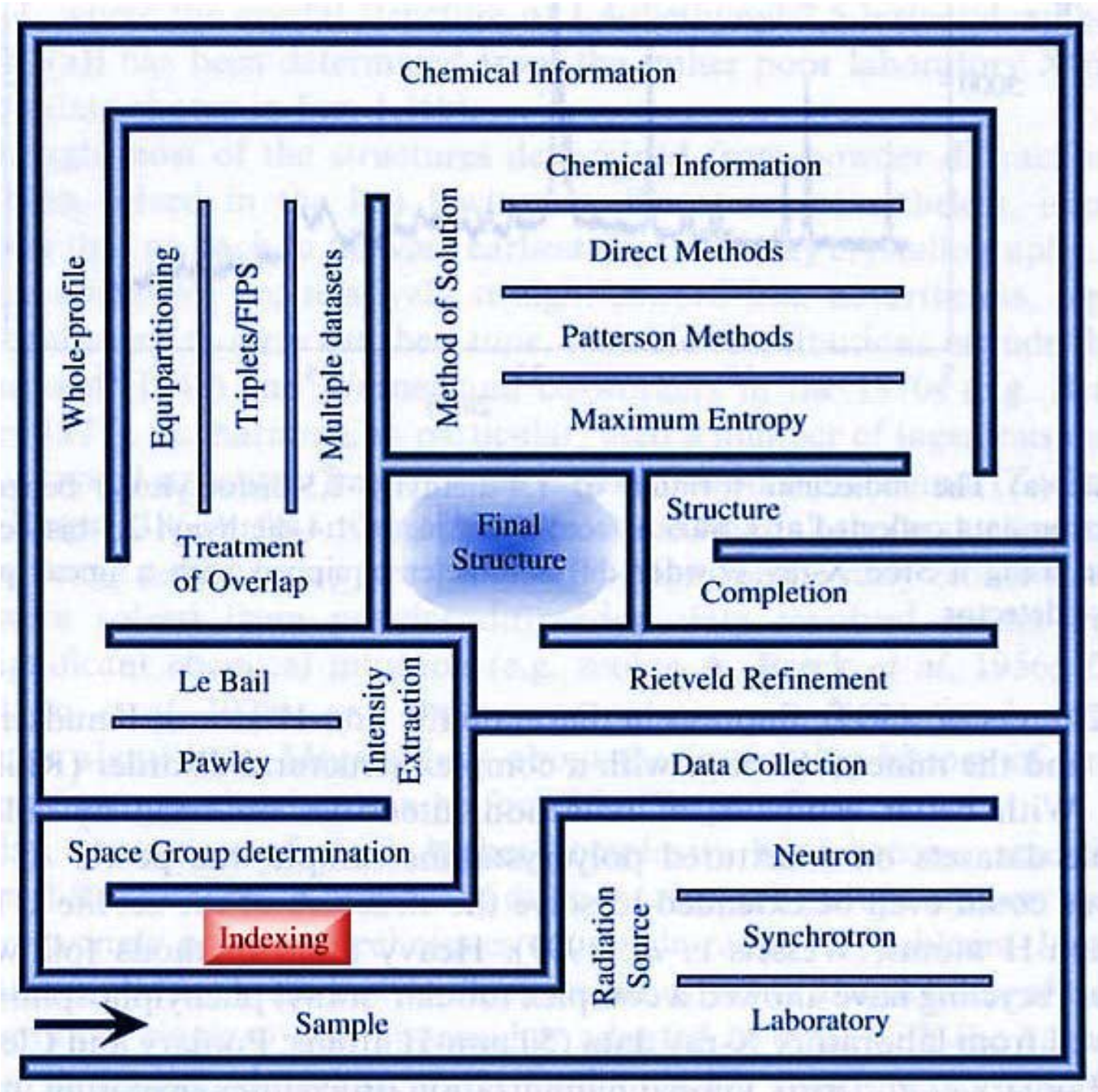
- Indexing multiphase patterns and impurity lines problems

- Conclusions

No way to escape the indexing bottleneck in the SDPD maze

(figure extracted from the IUCr monograph 13)

SDPD = Structure Determination by Powder Diffractometry



SDPD Round Robin 2 – September 2002 (indexing part)

8 powder patterns to index
100 participants having downloaded the data

6 answers received in due time (1 month)

	1	2	3	4	5	6	7	8
P1	X	X	X					
P2	X	X	X					
P3	X	X	X					
P4		X						
P5	X	X	X		X		X	X
P6	X	X	X					

CRYSFIRE (P1, P3), *DICVOL* (P2), *ITO* (P4), *Index* (P5) and *X-Cell* (P6).

Example : sample 3 of the Round Robin
in principle very simple : cubic

- Organizers: 18.881 18.881 18.881 90 90 90 (vol 6734 Å^3)
- P1: 13.349 13.349 9.439 90 90 90 (Tetragonal - 1638Å^3)
- P2: 18.878 18.878 18.878 90 90 90
- P3: 13.354 13.354 9.442 90 90 90 (Tetragonal - 1638Å^3)
- P4: no solution
- P5: 18.878 18.878 18.878 90 90 90
- P6: 18.88 18.88 18.88 90 90 90

50% successful only !

Possibly because of too small default maximum volume limit.
P1 and P3 provide the same (correct) subcell.

Well, the Round Robin conclusion is that indexing would not be that easy ?

Hence the need for more efforts in the less explored directions for indexing, using not only peak positions but also intensities, and why not the whole powder profile (that way was initiated by Kariuki et al., 1999, applying a genetic algorithm). McMaille follows the same route by using the Monte Carlo method in order to generate randomly cell parameters tested against an idealized powder profile.

Find more details about the SDPD-2002 Round Robin at :

Web site : http://sdpd.univ-lemans.fr/sdpdrr2/results/

Report : in the IUCr - CPD – Newsletter N°29 – July 2003

McMaille demonstrations

Program download (GNU Public Licence) :

http://www.cristal.org/McMaille/

Crystallographers want solutions fast !
Is that possible with McMaille ?
YES… if you consider 5-15 minutes being fast…

The first recommended approach with McMaille is to use the quite simple automated « black box » mode.

Peak positions and intensities can be extracted by using the WinPlotr program which is able to build the McMaille entry data file directly for the automated mode.

Demonstrations

1 – An easy task : peak hunting with WinPLOTR + Indexing by McMaille v3 in automated mode.

2 - Indexing in manual mode – same powder pattern

a – adding 6 impurity lines to 20 lines (23%) – total 26 lines

b – adding 12 impurity lines to 20 lines (37.5%) – total 32 lines

c – more difficult : 10 lines of 20 are impurity lines (50%)

3 - Indexing the Crysfire test file in automated mode, how long (~5mn) ?

4 - Indexing samples 1, 2 and 3 of the SDPD Round Robin (manual mode)

5 - Indexing a (simple) two-phases (both cubic) pattern (30 lines : 2x15)

6 - Seing the result of a more complex two-phases case (tetragonal + orthorhombic) needing more than 20 minutes of calculations.

A bit ambitious if McMaille is a slow indexing program…

What is examined in the automated « black box » mode ?

All symmetries in restricted volume and cell parameter ranges

(unrestricted in cubic)

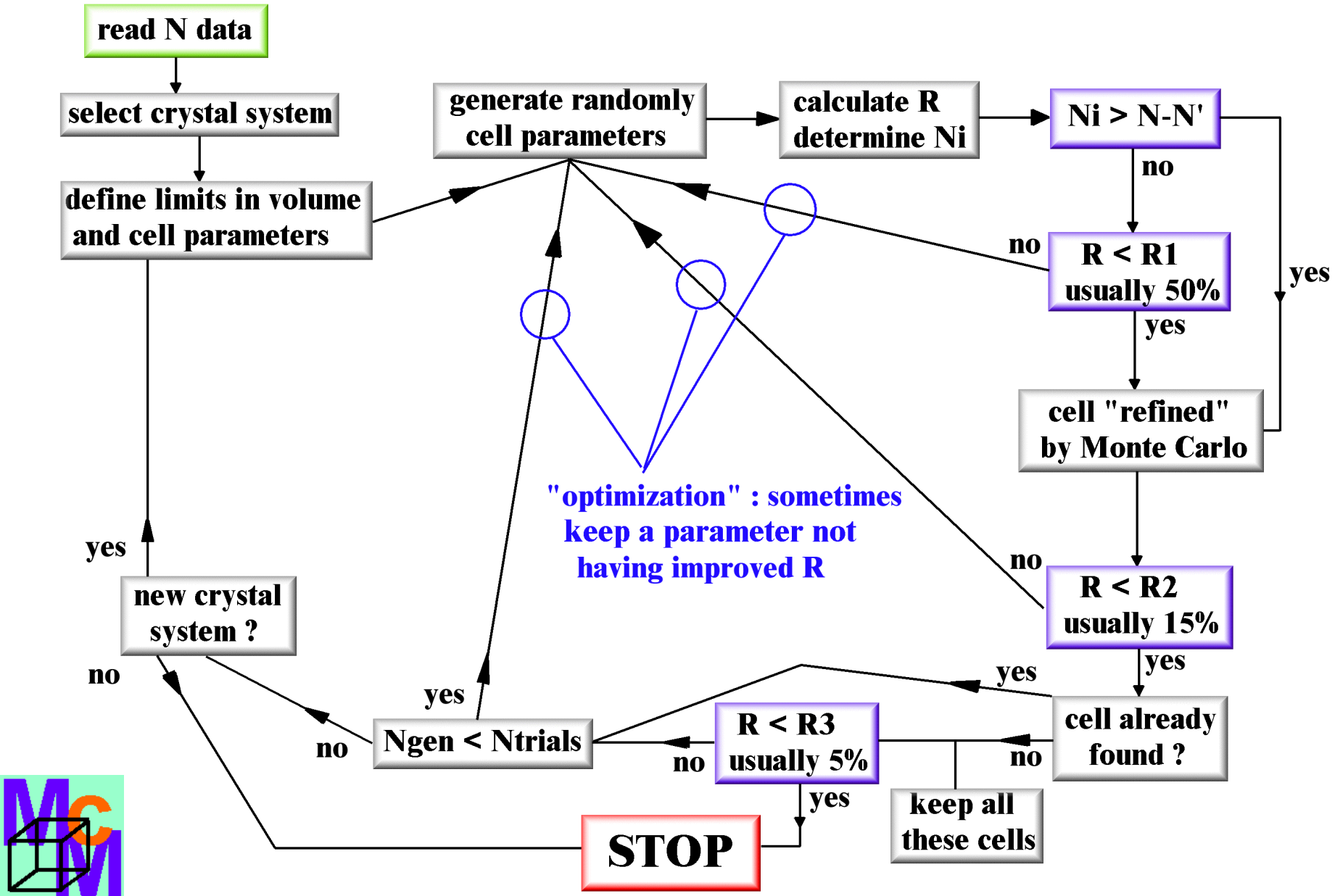
Symmetry	max MC events	Pmax	Vmax
cubic	V*0.5	3*dmax	(3*dmax)**3 - no limit
hex/rhomb/tetra	400000	30	4000
orthorhombic	4x1000000	20	500-1000-1500-2000
monoclinic	4x10000000	20	500-1000-1500-2000
triclinic	4x100000000	20	250-500-750-1000

Hence the need for exploring other ranges in manual mode if the automated mode was unsuccessful, or if you want to have a deepest look

Fortunately, McMaille produces a file ready for the manual mode at the end of an automatic examination

Now let us examine more closely all the McMaille « secrets »

« Simplicity » of the Monte Carlo algorithm in McMaille



Saving time with the hkl list

hkl Miller indices are predetermined (400 to 1000) for every crystal system and saved in files read once at the beginning.

They are only selected (not ordered which would be too long) according to the cell parameters trial and cut off at d(*hkl*)min.

If a calculated profile do not intercept any observed one, then the corresponding *hkl* set is considered as unobserved, not taken into account.

McMaille v1 and v2 : Gaussian profiles

- Choice was made of an idealized profile (Gaussian shape applied to extracted peak positions) rather than using the raw pattern – for velocity reasons. Fit by 3 iterations of the Rietveld decomposition formula (= Le Bail method).

-Version 1 worked only in cubic for studying the feasibility which was quite encouraging with 1000 tests per second by using a > 2GHz processor.

-Version 2 extended to all crystal systems, 300 tests per second in triclinic.

Not fast enough with low symmetries needing 10^8-10^9 tests…

McMaille v3 : Columnar peak shape

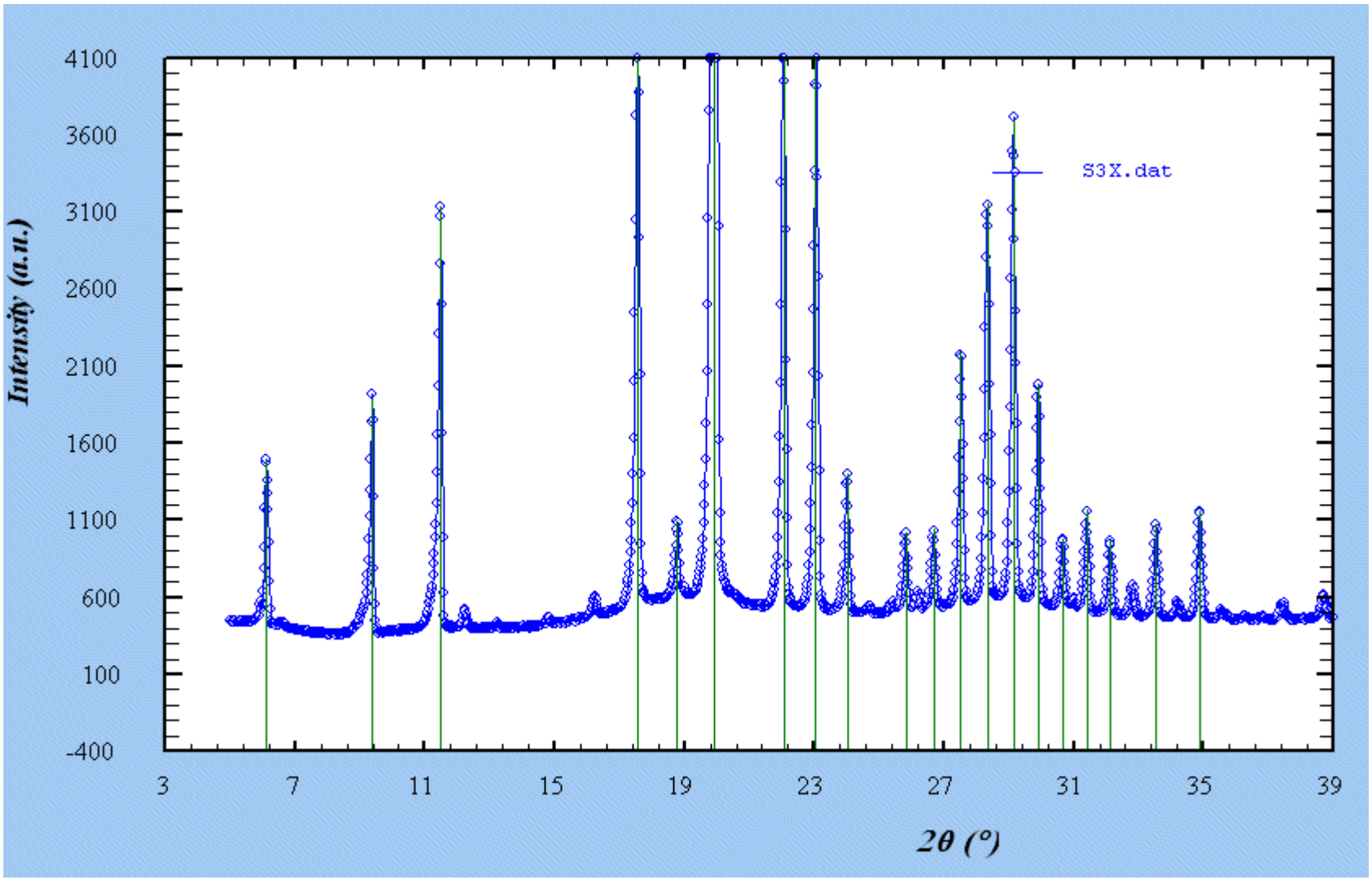
Speed increased by a factor 20…

20.000 tests per second in cubic, 6000 in triclinic.

No real need for a fit, the observed and calculated columns are given the same height and same width.

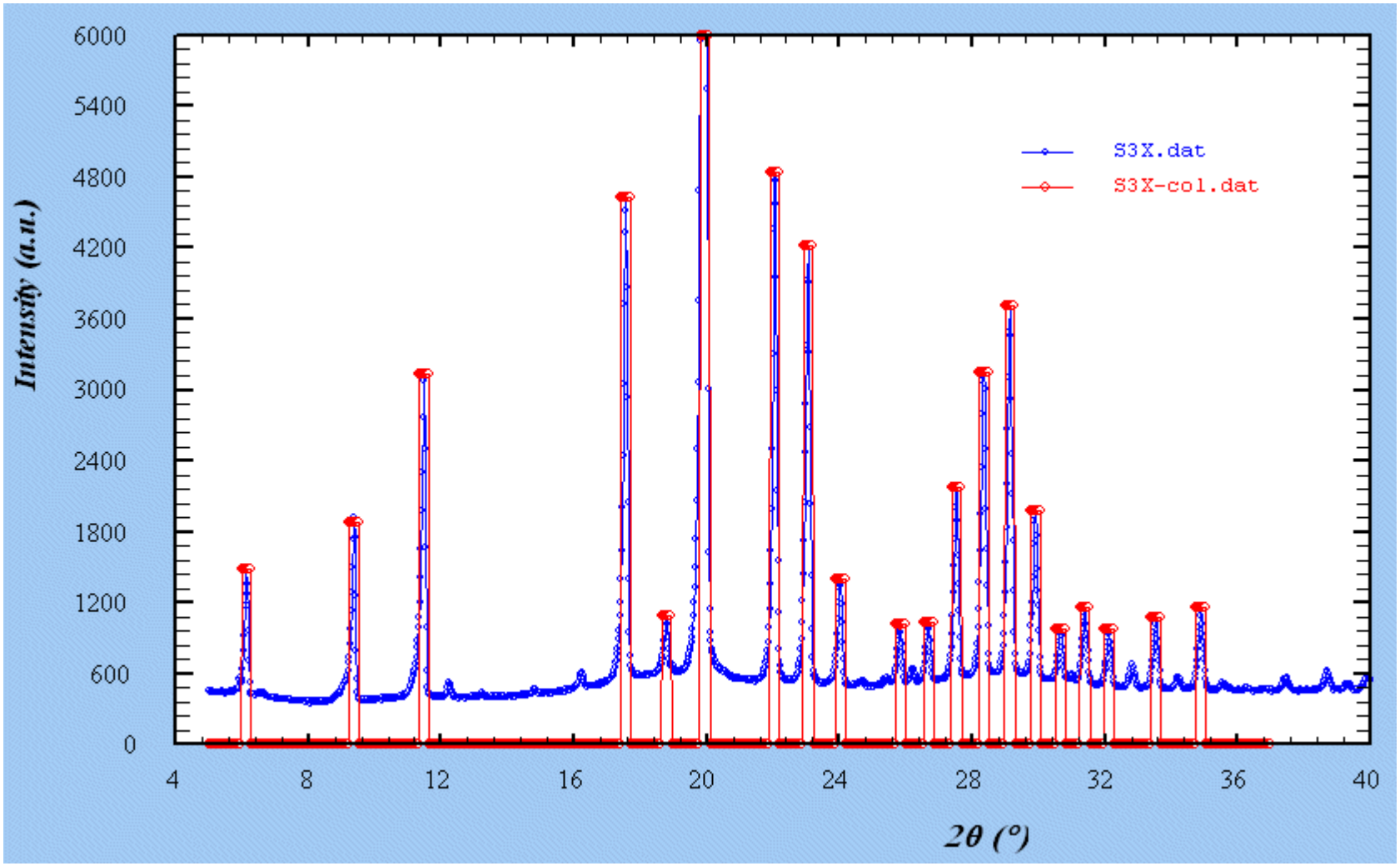
The R factor becomes function of the percentage of overlapping between observed and calculated columns.

SDPDRR2 Sample 3 – conventional X-rays

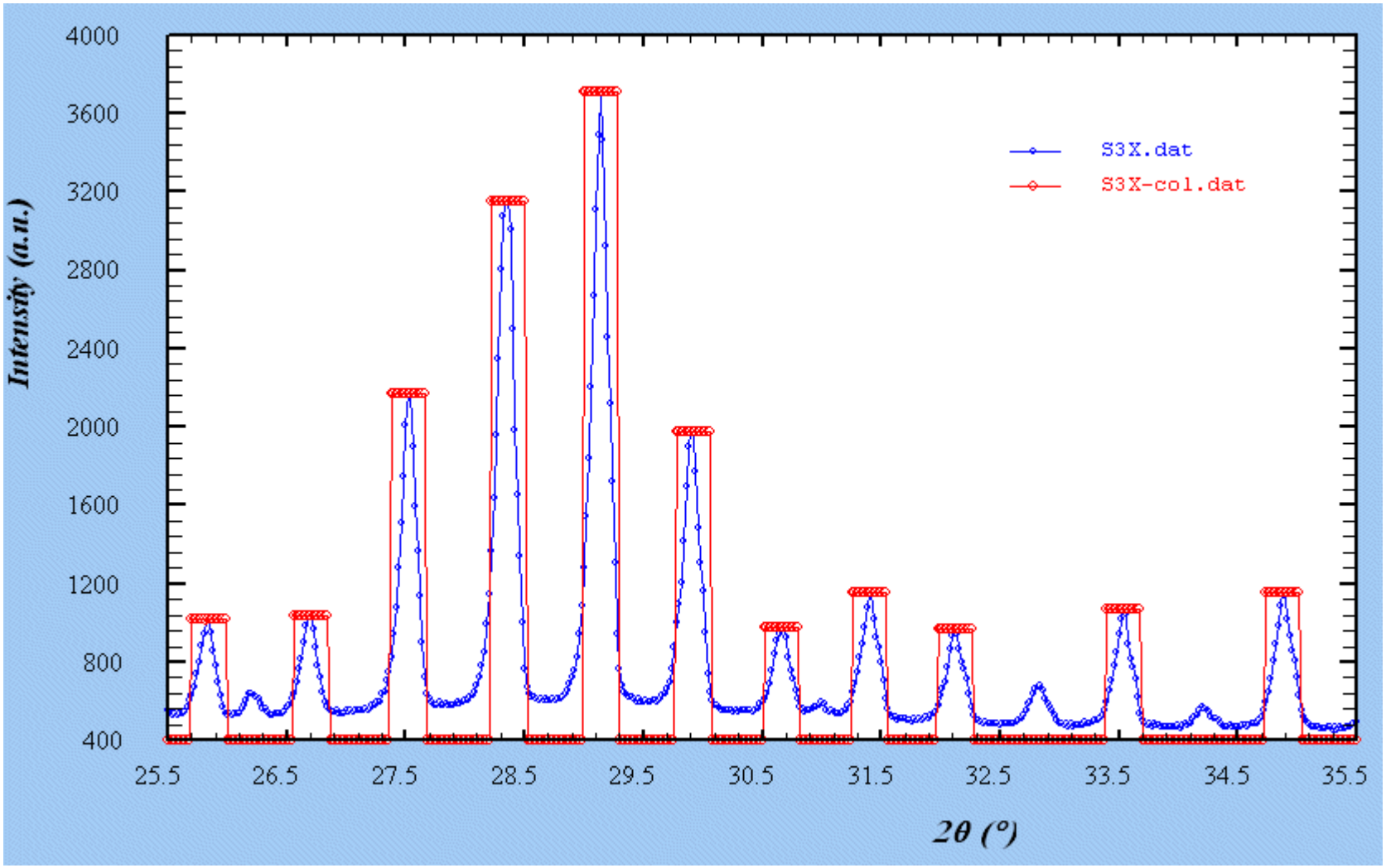


Peak positions extracted by WinPLOTR

The columnar peak shapes used by McMaille v3



Zooming on the last reflections

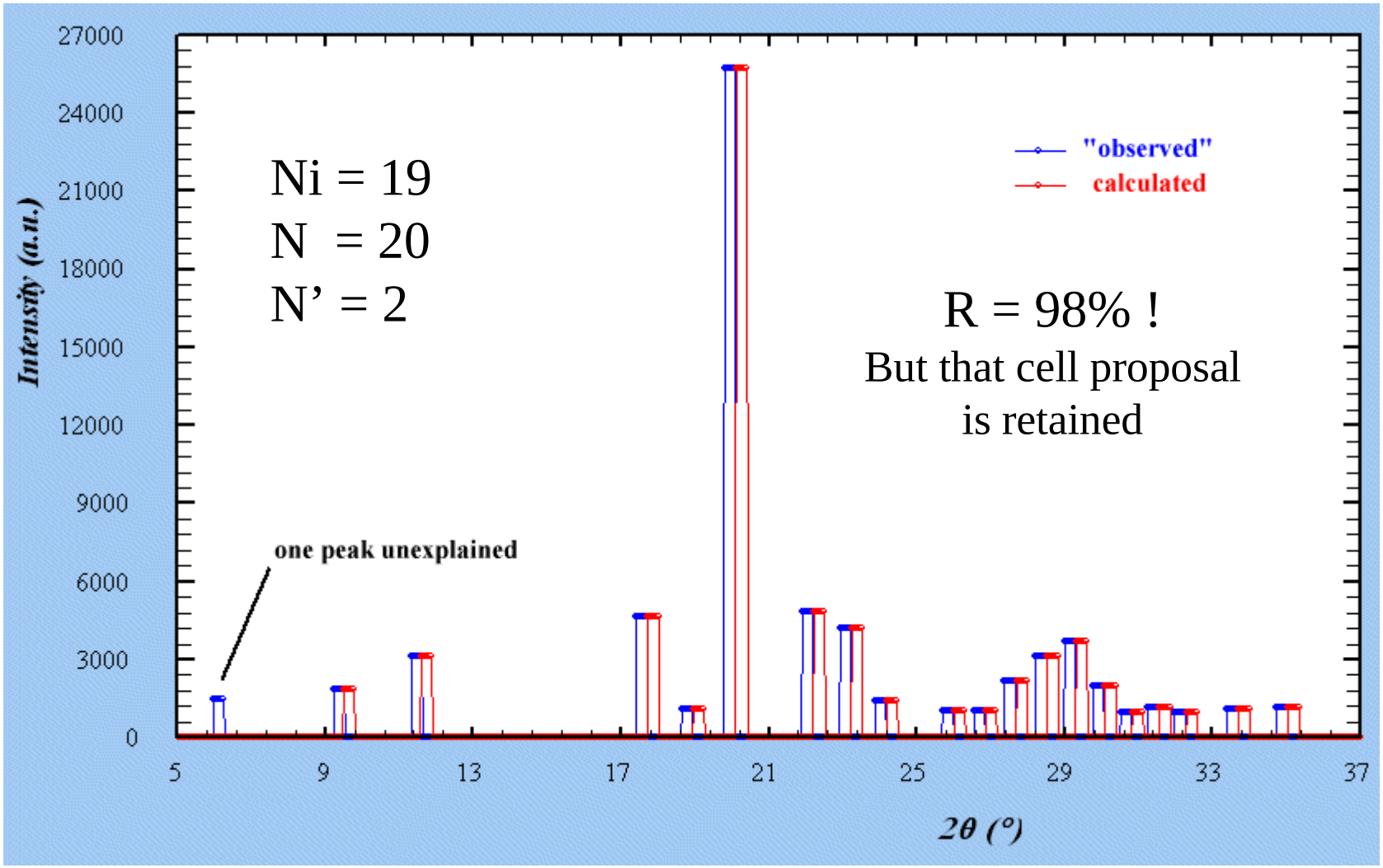


The width in the automated mode is calculated as :

$0.3 \times \lambda / 1.54056$

it depends on the user in manual mode

One of the 2 cases leading to « refine » a cell : Ni > N - N'



A peak is considered as indexed if some overlap occurs with a calculated one.
The second case leading to « refine » a cell is when R < R1 (usually 50%).

About (accuracy + enlarged peaks) apparent contradiction

The more the « observed » columns are large, the more you have chances to intercept them by the calculated columns.

Cubic example : A column at 10°(2θ) (d = 8.838 Å) will extend from 9.85 to 10.15 °(2θ) (d =8.972 to d = 8.707; λ = 1.54056 Å). So that if the peak is the 200 reflection, the range of ***a*** values leading certainly (if the accuracy is high) to the solution is [17.41-17.94 Å].
Any test in that interval larger than 0.5 Å is a winning test.

But at the « refinement » stage, it is mainly the position accuracy which is important. It will lead effectively to low R values (exact overlapping corresponds to R = 0.) allowing to distinguish the true solution from bad proposals.

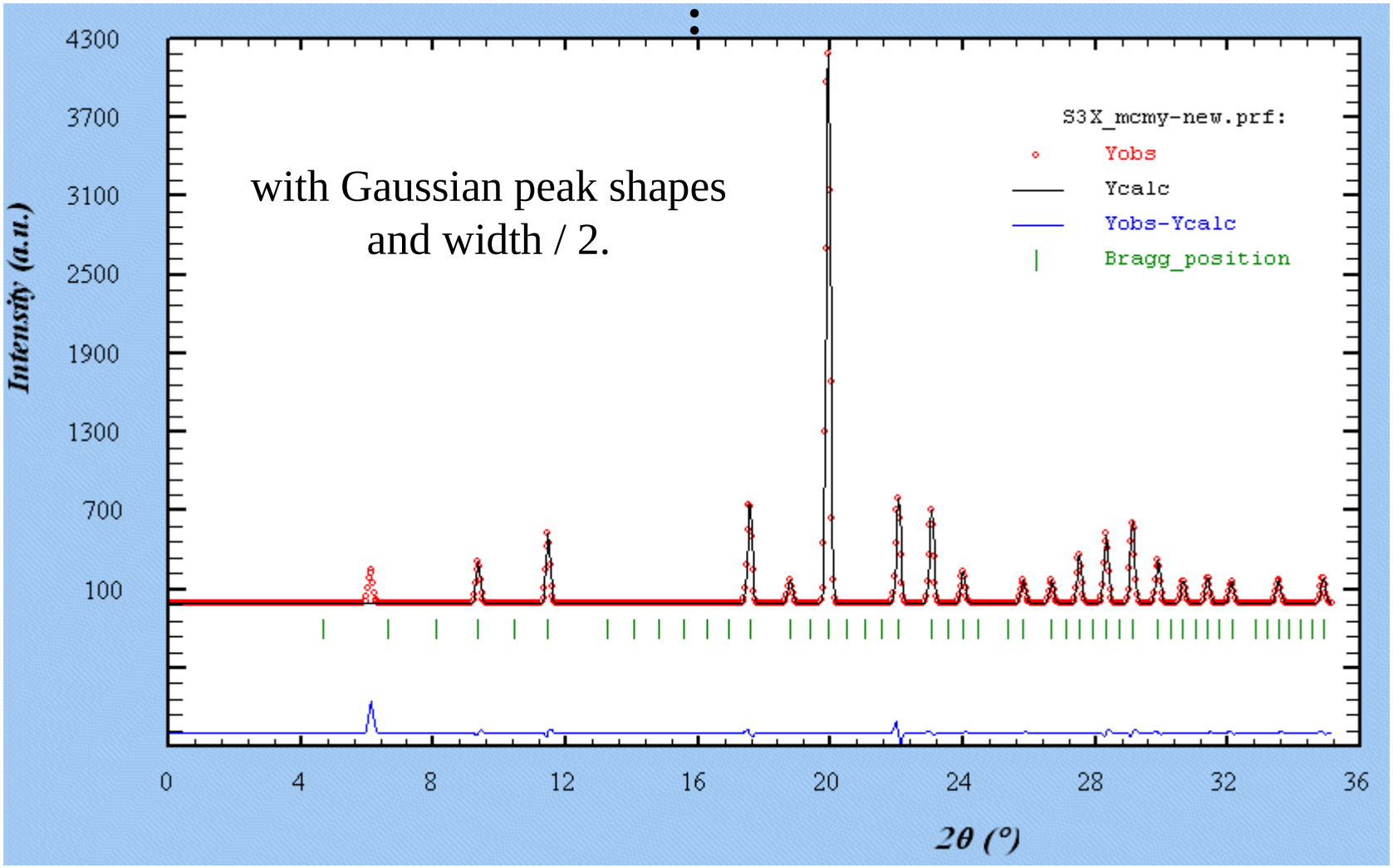
More on the cell « refinement » in McMaille

At R ~ 50%, no least square refinement is possible

So that the cell parameters are adjusted by Monte Carlo (200 steps in cubic to 5000 steps in triclinic) with small amplitude change (|delta max| = 0.02 Å). Similar to tempering ?

The best proposal at R < R3 ~ 5% is finally least-squared refined and classical figures of merit (M_{20}, F_{20}) are calculated.

Final plot produced by McMaille, displayed by WinPLOTR



Other software compatible with McMaille outputs :
CHEKCELL, CRYSFIRE

Optimization : not being trapped in a false minima

Accepting a parameter change even if the fit is not improved

Effects with various **Probability** values (probability to accept a new cell parameter if the fit is not improved) : number of times the correct answer is found for the same number of Monte Carlo steps :

P (%)	0	15	30	45	60	75	100
Test 1 – orthorhombic	41	45	32	27	15	6	1
Test 2 – rhombohedral	28	41	40	28	17	10	6
Test 4 – monoclinic	47	60	46	45	25	19	2
Test 6 – triclinic	36	42	36	24	18	12	12

The tendency is to work better with P ~ 15 %, as a mean

P : a value of 15 means that in 15% of the tests, a parameter change may be accepted even if that change does not lead to any R decrease or number of indexed reflections improvement (no change means that you keep the previous parameter unchanged)

P = 100 : always accepted even if it does not improve the fit

P = 0 : not accepted at all if it does not improve the fit

Relative insensitivity to impurity

The user decides by two control parameters :

N’ : number of unindexed lines.

R2 : consider only proposals with R < R2. Fixing it at 15% means that cell proposals explaining at least 85% of the peaks total intensity will be listed.

An impurity should not concern more than 15% of the total intensity, right ? But the number of (small) peaks belonging to the impurity can be high…

Preliminary results about impurity lines :

Provided the total intensity of the impurity lines is less than 15% of the total intensity :

With less than 35% (in number) of impurity lines, McMaille generally provides the correct cell in top position. However, the figures of merit decrease.

With 35-50% (in number) of impurity lines, McMaille may still propose the correct cell, but generally not in first position. Thus it is more difficult to locate it.

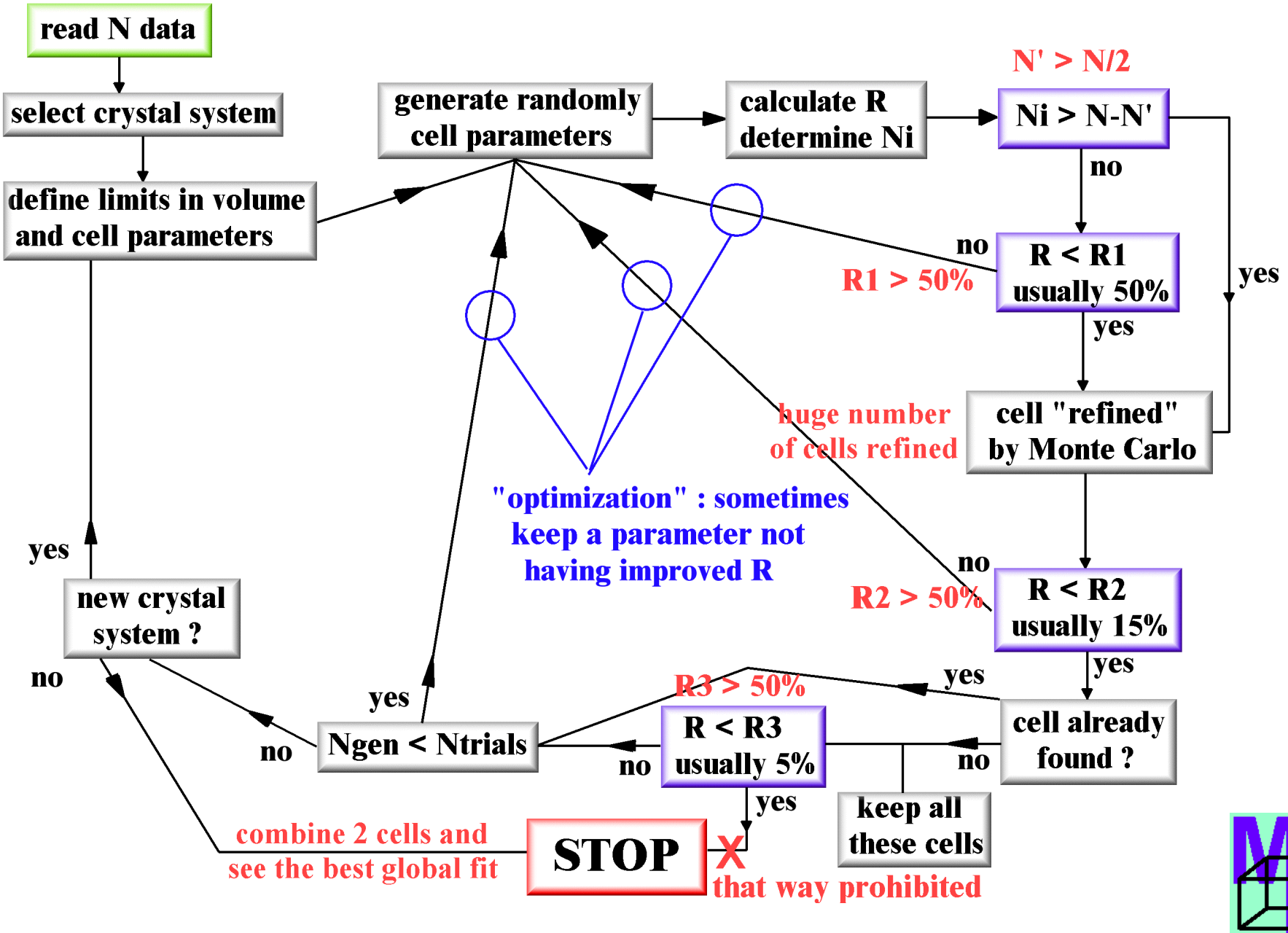
Beyong impurities, indexing multiphase patterns

Multiple synthesis in varying conditions should reveal the multiphase nature of the sample.

It is much better to adjust the synthesis conditions, and even if the phases cannot be prepared as pure phases, intensities variations should allow to define the peaks belonging to one or the other phase.

But if really you want to attempt indexing of a mixture, let us see the cost on the McMaille organization chart…

Indexing a 2-phases powder pattern with McMaille



Preliminary conclusions about two-phases indexing with McMaille

Provided at least 30 lines are examined with 13-17 lines belonging to each phases, and 40-60% of the total intensity distributed to each phase, then :

McMaille appears to be able to produce solutions in reasonable times (<1 hour) for combinations of two phases either cubic or hexagonal or tetragonal or orthorhombic.

Monoclinic and triclinic not examined (too long).

Example of 2-phases indexing (20 minutes) :
One mixture of a tetragonal with an orthorhombic phase

McMaille combines the cell proposals by couples and detects the best combinations indexing the largest number of peaks :

```
================================================================================
Double cells with largest number of peak indexed
================================================================================

WARNING - WARNING - WARNING - WARNING - WARNING
           This is the two-phase mode
    It could be better to go back to the lab
         and try and make a pure sample
```

Rp2	Vol	Vol/V1	Ind	Nsol	a	b	c	alpha	beta	gamma
0.259	1188.120	1.00	30	13	11.1880	11.1880	9.4919	90.000	90.000	90.000
0.106	378.244	1.00	15	4	10.0276	3.4206	11.0274	90.000	90.000	90.000
0.259	1188.120	1.00	29	13	11.1880	11.1880	9.4919	90.000	90.000	90.000
0.100	507.781	1.00	12	1	4.5923	10.0280	11.0265	90.000	90.000	90.000
0.259	1188.120	1.00	29	13	11.1880	11.1880	9.4919	90.000	90.000	90.000
0.125	788.734	1.00	14	1	11.0263	7.1331	10.0282	90.000	90.000	90.000
0.259	1188.120	1.00	29	13	11.1880	11.1880	9.4919	90.000	90.000	90.000
0.140	1344.555	1.00	16	4	12.1596	10.0281	11.0266	90.000	90.000	90.000
0.259	1188.120	1.00	29	13	11.1880	11.1880	9.4919	90.000	90.000	90.000
0.110	516.988	1.00	13	2	4.6750	10.0275	11.0282	90.000	90.000	90.000

Conclusions

- Promising (?) method…

- Already quite efficient if you have time and a fast computer.

- Needs some skills in manual mode, but nothing to do in « black box » mode (except finding the zeropoint).

- Improve it if you have some ideas (GNU Public Licence).

- Completely free access.

- Use cautiously the 2-phases mode…

A useful address for a distance learning course :

SDPD Internet Course

http://sdpd.univ-lemans.fr/course/

Some words about McMaille from Robin Shirley :

(in IUCr Computing Commission Newsletter No. 2, July 2003)

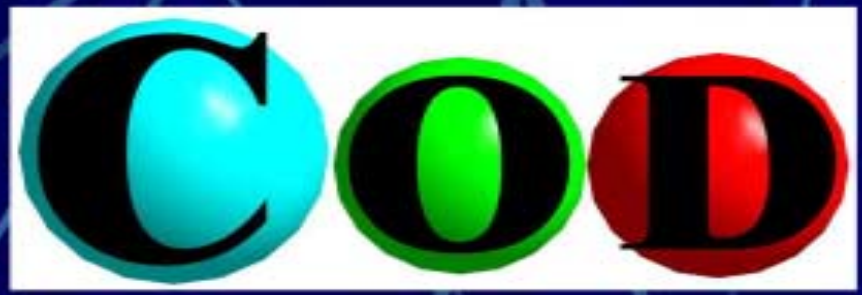
http://www.iucr.org/iucr-top/comm/ccom/newsletters/

« … the fact that it can work effectively at all shows how risky it can be to make negative predictions – less than a year ago at the Geneva Congress I predicted that it would be many years before computers became fast enough for whole-profile-based indexing to become feasible ! »

If finally you solve your problem with public licensed software, why not to deposit your results in the public domain ?

See the recommendations to IUCr journals authors :
§ 1.5 :
« The inclusion of material in an informal publication, e.g. a preprint server or a newsletter, does not preclude publication in an IUCr journal »

Crystallographers
join the
Crystallography Open Database



www.crystallography.net

Deposit your crystal data in the Public Domain
Thanks !