To be published in the EPDIC7 proceedings.
7th European Powder Diffraction Conference, Barcelona, 20-23 May 2000.

ESPOIR : A Program for Solving Structures by Monte Carlo from Powder Diffraction Data

Armel Le Bail

Université du Maine, Laboratoire des Fluorures, CNRS ESA 6010, Avenue O. Messiaen, 72085 Le Mans Cedex 9, France

Keywords: Monte Carlo, simulated annealing, molecule location, scratch, structure solution, powder diffraction

Abstract A direct-space program (*ESPOIR*) for the crystal structure solution of small molecules, from powder diffraction data, is developed under the GNU Public License. The program allows solving the structures of the two samples distributed during the Structure Determination by Powder Diffractometry Round Robin (SDPDRR) : the tetracycline hydrochloride and a cobalt-amine. Three calculation modes are possible, either locating up to 4 different molecule fragments, or starting from a set of randomly distributed atoms, or mixed approaches.

Introduction Powder diffraction is a theatre for the development of unconventional methods for stucture solution (different from classical Patterson and direct methods). However, less than 100 structures were solved from unconventional methods, and are reported in the SDPD (Structure Determination by Powder Diffractometry) Database [1]. Trying to locate a molecule in a crystalline cell while matching to either extracted "|Fobs|" or the full powder pattern or the Patterson function (etc), has become a very active topic since about 15 years. New direct-space methods use either systematic grid search, Monte Carlo (MC), simulated annealing (SA), genetic algorithm (etc), as already listed in recent review papers [2, 3]. Regretfully for the academic researcher, a problem is the lack of availability on the Internet of many of the computer programs involved in these SDPDs. Some single crystal programs are notable exceptions, and access to their source code is even possible. Unfortunately, they are unadapted to the powder diffraction handicap (overlapping). *DIRDIF* [4] allows the use of molecular fragments as input models, and then oriented by Patterson-vector-search-rotation functions. *PATSEE* [5] combines the merits of both Patterson and direct methods in order to position a fragment of known geometry in the unit cell. Indeed, *DIRDIF* and *PATSEE* were used in a few SDPD cases [1]. Other single crystal data programs are dedicated to molecular replacement (MR) methods for structure determination and are able to perform rotations and translations of macromolecules (for instance *MERLOT* [6], *AMORE* [7], *MOLREP* [8],...), none of these 3 programs was applied in any of the SDPD cases. The MR technique is far from being new, and a collection of papers is found in a 1972 book by Rossmann [9]. The more recent direct-space computer programs dedicated to *ab initio* solution of crystal structures from powder diffraction data are all commercial, with GUIs (Graphical User Interfaces) for MS Windows, and with the consequence that access to the source code is lost. This may confirm the high activity level in this topic. *POWDER SOLVE* [10] is the exclusivity of a pharmaceutical consortium through MSI, applying a combination of simulated annealing and rigid-body Rietveld refinement. *ENDEAVOUR*

[11], developped by Crystal Impact, uses "Pareto optimization" of the difference between the calculated and the measured diffraction pattern and of the potential energy of the system. The *DASH* (ex-*DRUID*) [12] release is announced for summer 2000 by the Cambridge Crystallographic Data Centre, and was one of the two regular winners in the solution of the SDPD Round Robin [13] sample II (the other winner used conventional methods through the *CSD* package). Other program names are *OCTOPUS* [14], *ROTSEARCH* [15], *MRIA* [16], *GAP* [17], *GAPSS* [18], *P-RISCON* [19], *GASPP* and *GULP* [20], *SNIFFER97* [21]... (and also one program without name [22], see the SDPD-Database [1] for references and Internet hyperlinks). No direct easy access to these programs has been found, but this does not mean that they could not be obtained by contacting their authors. Another approach makes use of packing considerations for molecule location and some programs may even be able to predict the cell parameters and space group : *PROMET* [23], *HARDPACK* [24], *PMC* [25], *UPACK* [26],... Special methods are dedicated to zeolites : *FOCUS* [27], *ZEFSAII* [28], and available with source code. That incredible number of new methods and recent programs (the list is certainly not exhaustive), if compared to the number of structures really determined up to now, could be mainly motivated by the satisfaction of pharmaceutical companies interests in the access to some crystal structures for which only powders are available. Those companies can certainly afford the commercial computer programs, but this is far from to be true in academia all over the world. It was thus believe timely to offer to academic researchers a free access to molecule location (ML) abilities through a new computer program following the open source code tradition, *ESPOIR*, also capable of much more.

Method in *ESPOIR* There is few if any new science inside those programs above, which essentially use algorithms built on the basis of previously developed concepts (Monte Carlo, simulated annealing, genetic algorithm, potential energy...) combined with crystallography rules. The new *ESPOIR* program makes no exception. It includes a basic Monte Carlo approach inspired from the *RMCA* code [29], generating random events through a pseudo-random-number subroutine. Those events may either be an atom move or a molecule or fragment rotation or translation. Dealing with overlapping is the important point with powder diffraction data, and here is the main "innovation" in *ESPOIR* : the solution retained is to reconstruct a pseudo-pattern from the previously extracted "|Fobs|". The advantage, if compared to the use of the real powder pattern, is in computer time. No background, Lorentz-polarisation, asymmetry or complex profile shape, reflection multiplicity, has to be considered. The program does not try to reproduce the profile shape in all its details, but grossly mimics the overlapping. A Gaussian shape (Ω) is thus selected for its fast calculation and short tails. The profile width is forced to follow the Cagloti law estimated at the structure factors extraction stage. Moreover, the number of profile steps is optimized by using only 3 to 5 points above the full width at half maximum (FWHM). The cost function depends on this reconstructed pattern $P(2\theta) = \Sigma\Omega|F_{obs}|$ through the equation : $R_{PF} = \Sigma |P_{obs} - K P_{calc}| / \Sigma P_{obs}$, where K is a scale factor. However, if the structure factor amplitudes are exempt of any overlapping problem (single crystal data, ideally), then *ESPOIR* has the option to work on a cost function (R_F) directly based on the |F| values, saving time by a factor 3 to 5. Simulated annealing (SA) is introduced through a tunable law reducing progressively the maximum atom move amplitude (or molecule translations). A parameter, also indexed on the SA law, allows to define random acceptance of MC events which do not necessarily decrease R_{PF}. This technique may allow to avoid being trapped in false minima. It is recommended to retain a 40% proportion of events not improving the fit. Nevertheless, at least 10 independent runs, starting at a different point of the pseudo-random number sequence, are needed for having chances of success. The more complex the problem, and the less is the success ratio. One hundred runs may be needed sometimes for attaining a really interesting minimum R_{PF}. In order to speed up the process, the many elements of the calculation which are not involved in the MC event are kept in large arrays in memory, avoiding useless recalculation. In spite of this, current computers are not fast enough for allowing the calculation of 10 to 100 independent runs in short time when dealing with

very complex problems in scratch mode. The ML mode is faster because a lower number of reflections is sufficient. Parallelizing the process would be quite interesting. Starting different runs on different processors would probably be the simplest and more efficient way to gain speed.

Scratch or molecule location modes When no fragment is previously known and if the classical approach (Patterson and direct methods) fails, then the crystallographer does not dispose of so many tools. Independent translation of dominant X-ray scatterers through the unit cell were attempted, for instance by systematic grid search [30, 31] as well as by Monte Carlo (with up to 2 different atoms in ref. [32]). Recent efforts in order to build larger models from scratch were done and implemented in few freely available software: *FULLPROF* [33] able for instance to locate Pb in $PbSO_4$ by Monte Carlo. It is to be noted that the user has to guess if atoms will occupy special positions. During tests, it became evident that *ESPOIR* runs much better in P1 than in any other space group. This is very probably because it is the only space group in which a truly random starting model can be built : the origin has no importance. It is thus recommended in scratch mode to run *ESPOIR* in P1, when possible, of course after extraction of the structure factor amplitudes in P1 too. Apart from the random moves, atoms are also allowed to permute randomly. Solving structures from "scratch" (random initial models) was proved to be possible for up to 15-30 independent atoms. In the ML mode, the random events are either rotations or translations of the molecule, analogously to the so-called "molecular replacement" method. A mixture of the "scratch" and ML possibilities is allowed with an upper limit of 4 "objects", each containing up to 50 independent atoms, either as a rigid body fragment or as a set of atoms randomly moving and permuting. The other main features of the program follow : X-ray or neutron diffraction data can be processed in any space group ; fine tuning is allowed on random moves, on permutations of atoms, on fixing of guessed special positions. Locating molecules in the cell is achieved by starting from models described either in crystalline cells or by Cartesian coordinates. Orientation disorder is accounted for (C60 molecule for instance) by using a global scattering factor. Restraints can be imposed on interatomic distances.

Examples of application Essentially, the (not so clever) strategy consists in trying again and again, jumping quickly to a new starting configuration if a model is frozen (false minima). Then, it is understandable that the main problem of this program (and some others), when dealing with the more complex cases, is computer time. The direct methods find 30-100 independent atoms (though 50 were never attained from powder diffractometry data till now) in a matter of minutes on a PC (100-1000 MHz), and less than 30 atoms in a matter of seconds. The millions of moves and atom permutations, necessary for finding 30 atoms with *ESPOIR* from a set of 300 *hkl* reflections, require one night, at least, if you are lucky. So that, testing for larger configurations was not already done, due to the lack of easy access to faster or parallel computer. Fortunately, the Moore's law is still expected to be applicable for many years, so that some hope may be placed on *ESPOIR* and the other above mentioned simulated annealing software. The program succeeds easily in the location of a whole molecule, with high success rates. The location of 2 objects when one is a molecule and the other is a mediumly heavy atom like Cl or S is also successfull with high success rate. Up to now, the problem of torsion angles is not adequately taken into account. However, the program can cope with a problem of up to 3 torsion angles by decomposing the molecule in 4 parts. In the series of examples shown on Table I, some structure factor amplitudes were obtained by applying a Le Bail method "$|F_{obs}|$" extraction to a simulated powder pattern. Note that the success rate is also dependent of some parameters that can stop the process if the R factor is not sufficiently low after a certain number of events (meaning that the success rate could even be higher). The R_F and R_{PF} values are the lowest of the series of runs, but a success was considered effective at $R_{PF} < 0.200$. Some examples are treated both in P1 space groups and in their true space group, in order to show the quite better efficiency of *ESPOIR* in P1. The 100% success rate (all with $R_{PF} < 0.07$) for Li_3RuO_4 [20] or $PbSO_4$ in P1 means obviously that the Ru or Pb atoms were correctly located, though many O and S atoms

are well positioned too (Li_3RuO_4 was originally solved by a general purpose structure prediction program by genetic algorithm using a Born model lattice energy minimization, without the need of the powder pattern intensities). The structure of the SDPD Round Robin [13] sample I, $[Co(NH_3)_5CO_3]NO_3 \cdot H_2O$ [34], for which no participant proposed a model, can be solved by *ESPOIR* (15 independent non-hydrogen atoms in $P2_1$ space group). In ML mode, the model consisted in 3 objects : one $[CoX_6]$ octahedra (X = O, N) with the C atom connected to the O atom, one NO_3 group, and three additional independent O atoms (two of them should complete the CO_3 group). For the SDPDRR sample II, tetracycline hydrochloride, 2 objects were used : the first being the Cl atom and the second was a model obtained from the CSD databank corresponding to the tetracycline hexahydrate, removing the water molecules and the H atoms. Those 2 examples are listed at the bottom of Table I. Obtaining a solution for these two cases in scratch mode was successful only when using good single crystal-like |F| values (see Table I). The user has to be conscious of the *ESPOIR* program limits, proposing a last-chance method, recommended if classical approaches fail.

Table I. Test examples delivered with the program.

	S.G.	S or ML	N obj	atom sites	DoF	hkl	R_F	R_{PF}	events $x10^{-3}$	success rate	time per run
Al_2O_3 *	$R\bar{3}c$	S	1	2	6	25	0.005	0.005	20	8/10	25 s
CaF_2 *	Fm3m	S	1	2	6	16	0.011	0.009	60	22/50	80 s
calcite *	$R\bar{3}c$	S	1	3	9	33	0.004	0.004	60	2/10	6 mn
aragonite *	Pmcn	S	1	4	12	107	0.076	0.043	60	8/10	24 mn
forsterite *	Pbnm	S	1	6	18	101	0.037	0.017	100	3/10	24 mn
Li_3RuO_4 *	P1	S	1	16	48	160	0.255	0.013	200	10/10	26 mn
Li_3RuO_4 *	P2/a	S	1	6	18	100	0.165	0.029	50	8/10	12 mn
$CuVO_3$ *	P1	S	1	10	30	120	0.001	0.001	100	7/10	24 mn
$CuVO_3$ *	$P\bar{1}$	S	1	5	15	120	0.001	0.001	100	2/10	15 mn
TeI *	P1	S	1	16	48	200	0.003	0.002	200	19/20	60 mn
TeI *	$P\bar{1}$	S	1	8	24	200	0.305	0.163	200	3/20	13 mn
C_{60} disordered *	Fm3m	S	1	1	3	50	0.062	&	20	10/10	2 mn
$PbSO_4$ #	P1	S	1	24	72	275	0.083	0.021	500	20/20	30 mn
$PbSO_4$ #	Pnma	S	1	5	15	83	0.001	0.001	100	9/10	10 mn
$Ba_2CdP_3O_{10}(OH)$ #	Im2m	S	1	9	27	130	0.081	0.067	60	5/10	24 mn
cobalt amine *	P1	S	1	30	90	300	0.078	&	2000	2/10	2 h
cobalt amine *	$P2_1$	S	1	15	45	150	0.037	&	2000	4/40	2 h
cobalt amine #	$P2_1$	S	1	15	45	150	0.193	&	2000	1/50	2 h
cimetidine #	$P2_12_12_1$	S	1	17	51	200	0.037	&	8000	1/50	7 h
cimetidine #	$P2_12_12_1$	ML	2	17	9	50	0.099	0.091	100	4/10	30 mn
pyrene *	$P2_1/a$	ML	1	16	6	50	0.110	0.073	80	10/10	13 mn
1-methylfluorene *	$P2_1/n$	ML	2	14	9	50	0.083	0.052	100	14/20	16 mn
tetracycline HCl #	$P2_12_12_1$	ML	2	33	9	50	0.207	0.161	200	2/50	23 mn
cobalt amine #	$P2_1$	ML	3	15	21	100	0.332	0.159	300	4/50	16 mn

The time per run corresponds to the use of a processor Intel Pentium II 266 or 333 MHz, possibly with several simultaneous calculations (up to 3). All tests are from X-ray data, though the program accepts neutron data. S or ML is for Scratch or Molecule Location. N obj. is the number of independent objects. DoF = Degree of

Freedom. (*) structure factor amplitudes extracted from a simulated powder pattern ; (&) fit directly on |F| ; (#) real powder data.
Both SDPDRR samples were difficult cases mainly because of medium resolution of the experimental powder patterns since the best pattern (tetracycline hydrochloride, synchrotron data) shows minimal FWHMs six time larger than those attainable at the best sources. This gives an idea of the feasibility limits by those unconventional methods : cases 6 times more complex can probably be undertaken in ML mode.

Entry and output files In the *ESPOIR* package, *PRESPOIR* is a small add on program which helps the user to build interactively the entry (.dat) file containing all parameters. This file can be quite short, as shown below for a typical scratch test :

```
Al2O3                          ! title
4.764 4.764 13.009 90.0 90.0 120.0        ! a    b    c  alpha   beta   gamma
R -3 C                         ! space group
1.54056 4 2 2 1 1 1            ! wavelength, data type, atom number, atom type, nob, ns, iprint
0.025  -0.046   0.030  3       ! U, V, W, step (from the structure factors extraction stage), if ns = 1
Al+3O-2                        ! atom or ion names (determines the atom-type order)
1                              ! code for constraint on distances (if 0, no constraint)
3.0 1.6 2.2                    ! shortest interatomic distances Al-Al, Al-O, O-O, in Angstroms
6. 6.                          ! maximum moves for each atom-type, in Angstroms
2.0 1.0 0.005                  ! anneal, sigma, reject (simulated annealing parameters)
5000 20000 20000               ! number of Monte Carlo events : for screen show, max, save
10000 0.25 2 10                ! events for stop, rmax, ichi, nruns (stop if R < rmax after 10^4 events)
1 10                           ! type for object 1, permutations tested every 10 MC events
1 1                            ! number of atoms of each type in object 1
1.0 1 0                        ! B overall, nocc, nspe (occupation and special position codes)
0.33333 0.5                    ! occupation numbers for each atom (if nocc = 1) in object 1
```

The second entry file (with .hkl extension) should contain the Miller indices and the structure factor amplitudes, extracted by either the Pawley or Le Bail methods (or coming from single crystal data measurement). As a result, the model characterized by the best R_{PF} or R_F factors in the series of independent runs has its atomic coordinates inserted in a *SHELX*-like .ins file and a .spf file, allowing drawing by the many programs reading those standard files (*WinORTEP*, *WinSTRUPLO*, *PLATON*...). The reconstructed pattern can be drawn on the PC screen by *WinPLOTR* or *DMPLOT*, or any program able to read a .prf file.

Availability and documentation The *ESPOIR* package including documentation, source code, executable for Windows 95/98/NT, and example files can be downloaded on the Internet at the URL: http://www.cristal.org/sdpd/espoir/. A link to a Linux version is given at this Web page. Additional example files are available at URL : http://sdpd.univ-lemans.fr/sdpd/espoir/examples/.

Conclusion The most interesting ability of direct-space ML methods is to provide solutions from very few data (50 reflections for one fragment corresponding to 6 DoF) if the model shape is sufficiently correct. Most of the structures determined, according to this method, and listed in the SDPD-Database [1], could not be refined without restraints on interatomic distances. This may sometimes pose problem of credibility if the final Rietveld R_F remains too high. Nevertheless, one can think that many Rietveld-method-based powder diffraction computer program packages, available in the public domain, will soon include unconventional structure solution as some of their multiple options. Most of them already have the possibility to extract "|Fobs|" (Le Bail method), so that performing tests from scratch or ML appears to be a natural evolution (scratch possibility is

already in *FULLPROF*, though not optimized yet for speed). If one considers Nature as a random process, it took geological times for building the first DNA molecule. The process in *ESPOIR* is theoretically also able to solve giant problems, but a human life will not be enough to see a correct result appearing, due to the current computer speed. Finding structures "by chance" is at the opposite side of the rational direct method or Patterson approach. The random principle behind *ESPOIR* explains its logo : a bottle containing a structure drawing, floating on the ocean. As a matter of fact, "espoir" (in french) means "hope" in english, suggesting that you should not lose it. Moreover, the source code (Fortran) is delivered with the package, allowing you to add your own stones to the building. How much the process can be improved, and can we expect to solve much larger structures than the current 15-30 atoms maximum in the scratch option have no easy answer.

Acknowledgment The option for locating several fragments was implemented thanks to funding from the DuPont Company.

References

[1] A. Le Bail, SDPD-Database, http://www.cristal.org/iniref.html (1995-2000).
[2] K.D.M Harris, M. Tremayne, Chem. Mater. 8 (1996), p. 2554.
[3] A. Le Bail, http://www.cristal.org/iniref/ecm18/ecm18.html (1998).
[4] H.M. Doesburg, P.T. Beurskens, Acta Cryst. A39 (1983), p. 368.
[5] E. Egert, G. M. Sheldrick, Acta Cryst. A41 (1985), p. 262.
[6] P. M. Fitzgerald, J. Appl. Cryst. 21 (1988), p. 273.
[7] J. Navaza, Acta Cryst. A50 (1994), p. 157.
[8] A.A.Vagin, A.Teplyakov, J. Appl. Cryst. 30 (1997), p. 1022.
[9] M. G. Rossmann (ed.), The molecular replacement method, Gordon and Breach, New York (1972).
[10] G.E. Engel, S. Wilke, O. König, K.D.M Harris, F.J.J. Leusen, J. Appl. Cryst. 32 (1999), p. 1169.
[11] H. Putz, J.C. Schön, M. jansen, J. Appl. Cryst. 32 (1999), p. 864.
[12] W.I.F. David, K. Shankland, N. Shankland, Chem. Commun. (1998), p. 931.
[13] A. Le Bail, L.M.D. Cranswick, http://www.cristal.org/SDPDRR/ (1998).
[14] M. Tremayne, B.M. Kariuki, K.D.M Harris, Angew. Chem., Int. Ed. Engl. 36 (1997), p. 770.
[15] J. Cirujeda, L.E. Ochando, J.M. Amigo, C. Rovira, J. Rius, J. Veciana, Angew. Chem., Int. Ed. Engl. 34 (1995), p. 55.
[16] V.V. Chernyshev, A.V. Yatsenko, V.A. Tafeenko, S.G. Zhukov, L.A. Aslanov, E.J. Sonneveld, H. Schenk, V.A. Makarov, V.G. Granik, V.A. Trounov and A.I. Kurbakov, Z. Kristallogr. 213 (1998), p. 477.
[17] K. Shankland, W.I.F. David, T. Csoka, Z. Kristallogr 212 (1997), p. 550.
[18] B.M. Kariuki, H. Serrano-Gonzalez, R.L. Johnston, K.D.M. Harris, Chem. Phys. Lett. 280 (1997), p. 189.
[19] N. Masciocchi, P. Cairati, F. Ragaini, A. Sironi, Organometallics 12 (1993), p. 4499.
[20] T.S. Bush, C.R.A. Catlow and P.D. Battle, J. Mater. Chem. 5 (1995), p. 1269.
[21] R.B. Hammond, K.J. Roberts, R. Docherty, M. Edmondson, J. Phys. Chem. B., 101, 33 (1997), p. 6532.
[22] S. Pagola, P.W. Stephens, D. Scott Bohle, A.D. Kosar, S.K. Madsen, Nature 404 (2000), p. 307.
[23] A. Gavezzotti, Acc. Chem. Res. 27 (1994), p. 309.
[24] R. Rudert, Acta Cryst. A52 (Supplement), C-94 (1996).
[25] D. Louër, M. Louër, V.A. Dzyabchenko, V. Agafonov, R. Ceolin, Acta Cryst. B51 (1995), p. 182.
[26] B.P. van Eijck, J. Kroon, J. Comput. Chem. 20 (1999), p. 799.
[27] R.W. Grosse-Kunstleve, L.B. McCusker, Ch. Baerlocher, J. Appl. Cryst.. 30 (1997), p. 985.
[28] M. W. Deem and J. M. Newsam, Nature 342 (1989), p. 260.
[29] R.L. McGreevy, Nucl. Instr. and Meth. in Phys. Res. A354 (1995), p. 1.
[30] J.P. Attfield, Acta Cryst. B44 (1988), p. 563.
[31] A.J. Mora, A.N. Fitch, J. Solid State Chem. 134 (1997), p. 211.
[32] M. Tremayne, Ph.D. Thesis, University of St. Andrews, Scotland (1995).
[33] J. Rodriguez-Carvajal, FullProf program, ftp://charybde.saclay.cea.fr/pub/divers/fullp/
[34] J.H. Zhu, H.X. Wu, A. Le Bail, Solid State Sciences 1 (1999), p. 55.