# Combined optimization using cultural and differential evolution: application to crystal structure solution from powder diffraction data

**Samantha Y. Chong and Maryjane Tremayne***

*Received (in Cambridge, UK) 28th June 2006, Accepted 23rd August 2006*
*First published as an Advance Article on the web 12th September 2006*
**DOI: 10.1039/b609138e**



**The principles of social and biological evolution have been combined in a Cultural Differential Evolution hybrid global optimization technique and applied to crystal structure solution.**

Evolutionary algorithms are being increasingly used to solve a variety of global optimization problems in chemistry, nanoscience and bioinformatics.[1] These powerful techniques are inspired by natural evolutionary processes, and mimic the principles of biological evolution and survival of the fittest to explore parameter space. However, the biological evolution of natural systems can be a slow process, especially compared to the rate of cultural evolution in a society when adapting to changing social environment. Cultural algorithms[2] have been developed to model behaviour based on the principles of human social evolution, and can be used to bias the search process by passing experience and knowledge of behavioural traits of a population from one generation to the next. In simple terms, this cultural information can be used to reduce the search space of a standard biological evolutionary algorithm, improving both performance and efficiency of the global optimization process.[3] In this paper, we report modification of the Differential Evolution (DE) global optimization algorithm, by incorporation of the concept of Cultural Evolution, with the aim of increasing the efficiency of DE when applied to crystal structure solution from powder diffraction data.

Although DE is a relatively new evolutionary algorithm,[4] it has proved highly effective in a range of chemical contexts, including X-ray scattering,[5] crystal growth epitaxy,[6] optimization of clusters,[7] protein crystallography,[8] molecular docking,[9] disordered crystal structures[10] and the direct-space crystal structure solution of organic molecules from powder diffraction data.[11] The direct-space approach to structure solution[12] involves generation of trial crystal structures in real space, by placing a structural model of the molecule inside the unit cell, independent of the diffraction data. A calculated powder diffraction profile is then compared to the experimental pattern to assess the 'fitness' of each structure. Global optimization techniques, such as Monte Carlo,[13] simulated annealing[14] or evolutionary algorithms[11,15] are used to find the minimum point on the fitness landscape (or hypersurface), corresponding to the correct crystal structure.

In this work, the DE algorithm operates by generating a randomly distributed population of trial structures that is mated and mutated over a number of generations until the global minimum is located. Child structures are created using information from current members of the population by carrying out recombination and mutation in a single step.[4] The child structure is then directly compared to each parent such that the fitter of the two is retained, constantly updating the population and adapting to the fitness hypersurface. The values of $K$ and $F$ (the rates of recombination and mutation respectively), and the population size $N_p$, are chosen to achieve a balance between optimal fitness of the solutions and the time taken for the calculation to converge.[11]

The dimensionality of the hypersurface, and hence the complexity of the optimization problem, is determined by the number of variables used to define the structural model for crystal structure solution (*e.g.* position ($x$, $y$, $z$), orientation ($\theta$, $\varphi$, $\psi$) and torsion angles to describe molecular conformation ($\tau_1 \ldots \tau_n$)). In DE, each of these variables has associated minimum and maximum boundary values which are used to reset each variable to a point between the boundary and parent values, if the child structure exceeds the corresponding limit. For a general case allowing unconstrained molecular movement, this would typically involve boundary values of [0,1] for fractional position and [0,360°] for overall molecular orientation and intramolecular torsion angles.

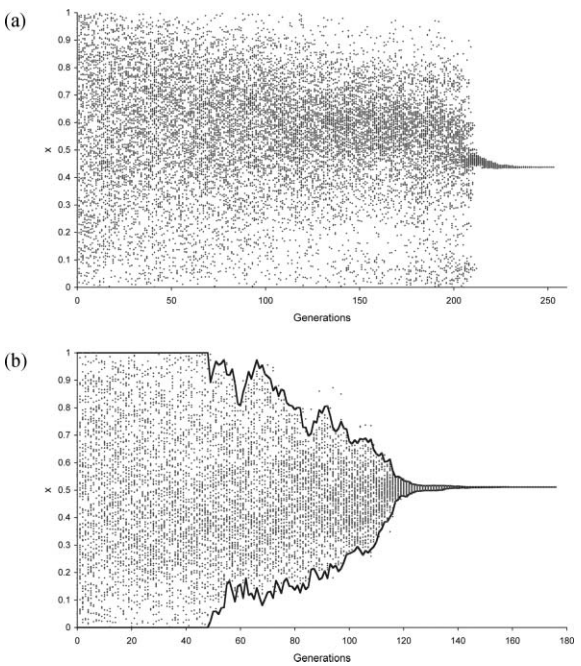



**Fig. 1** Distribution of the $x$ variable of the child structures (circles) generated during a DE calculation, (a) with static boundaries and (b) dynamic boundaries (the solid line indicates the boundary position).

*School of Chemistry, University of Birmingham, Edgbaston, Birmingham, UK B15 2TT. E-mail: m.tremayne@bham.ac.uk; Fax: 0121 414 4403; Tel: 0121 414 3201*

These inherent boundary conditions can also be used to restrict the DE search to specific regions of the hypersurface, allowing incorporation of structural constraints such as limits in molecular conformation, without disrupting natural optimization pathways. In the original DE algorithm, these boundary values remain constant throughout the structure solution calculation. However, examination of the child structures generated during a DE calculation with static boundaries shows that the random initial distribution in the values of a variable develops clustering as the calculation progresses (Fig. 1a). This distribution provides us with information that can be used, by incorporation of the concept of Cultural Evolution, to guide the DE search itself. In our implementation of the Cultural Differential Evolution hybrid algorithm (CDE), the behaviour of previous generations is used to influence subsequent generations using dynamic boundaries to restrict the search to low-lying regions of the hypersurface, and effectively prune population space. This mechanism of combining the two approaches is simple to implement and interpret in 'real-world' applications where parameter distribution and the resulting boundary conditions can also have physical meaning. It differs from that of Becerra *et al.*[16] in which the DE static boundaries are retained, and cultural evolution is used instead to influence the DE variation operator. The use of dynamic boundaries has a dramatic effect on the distribution of structure variables in the DE calculation. In the example given, the $x$ variable is allowed to take values $0 \leqslant x \leqslant 1$ throughout the conventional DE calculation (Fig. 1a), whereas in the CDE, the search was restricted to $0.2 \leqslant x \leqslant 0.7$ after only 90 generations (Fig. 1b).

In our CDE algorithm, the original boundary conditions are initially maintained to avoid encouraging premature convergence to local minima (*e.g.* 50 generations, Fig. 1b). After this period, a virtual histogram of child parameters is constructed for each variable within a generation (Fig. 2). This detects any clustering, and identifies at what values the dynamic boundaries are set by imposing an 'underpopulation threshold' at the maximum and minimum ends of the distribution. By defining the dynamic boundaries by exclusion of 'outliers' rather than inclusion of 'popular' values, potential problems with multi-modal distribution and over-aggressive culture-based pruning are avoided. Fig. 2 shows distribution of the $x$ variable over a population of 70 structures, divided over 22 histogram bins. In this case, the underpopulation threshold ($N_{ut}$) at each end of the distribution is set to four structures, and the end bins removed until the threshold

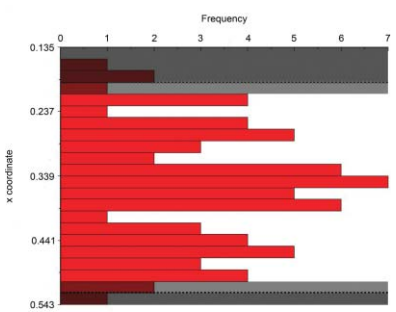
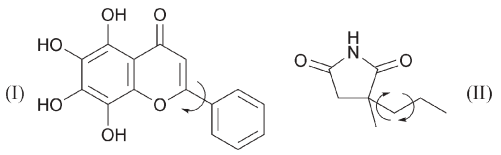



**Fig. 2** Bins are removed from the histogram until the underpopulation threshold is reached (shaded regions), but one bin at each extreme is reinstated to give the new boundaries (dotted lines).




**Scheme 1** Structural models of baicalein (I) and α-methyl-α-propyl succinimide (II). Arrows show variable torsion angles.

is reached. In order to allow the possible expansion of the boundaries with successive generations, one bin is then reinstated at each extreme, and the maximum and minimum of these remaining categories used as the new dynamic boundary values for the next generation. These boundaries are then invoked in the DE calculation as described earlier.

In this paper, performance of the CDE is illustrated by the structure solution of (i) a test case, baicalein[17] (I) and (ii) an unknown crystal structure, α-methyl-α-propyl succinimide (II). In both cases, the structure solution used a model comprising the whole molecule (excluding hydroxyl, methyl and amide hydrogens where applicable), and allowed translation throughout the unit cell, rotation in all directions and intramolecular rotation defining molecular conformation (Scheme 1). Baicalein was studied initially using the DE approach (static boundaries) with five DE runs performed for each combination of control parameters $K = 0.99$, $N_p = 105$ and $F = 0.4$, 0.5 and 0.6 (Fig. 3). Corresponding sets of CDE calculations were then performed using $N_{ut} = 3,4,\ldots,7$, with the same DE control parameters (Fig. 3). These results show a significant and consistent gain in the efficiency of the calculation with the CDE algorithm converging up to 54% quicker when averaged over the five runs for each set of parameters (*i.e.* $F = 0.6$, $N_{ut} = 7$). Similar results have been obtained for other values of $N_p$, and for other test structures.[18] It is also clear from these results that the optimal choice of $N_{ut}$ is, as expected, dependent on the combination of the other DE control parameters.

In the case of (II), the powder diffraction pattern was indexed as monoclinic, $P2_1/c$, with $Z = 1$. The CDE calculation was run several times with the optimal control parameters $K = 0.99$, $F = 0.5$, $N_p = 80$ and $N_{ut} = 4$ until convergence was reached. The best solution had $R_{wp} = 13.2\%$ (mean $R_{wp} \approx 33\%$) (Fig. 4), and this structure was used as the starting point for successful Rietveld refinement† (Fig. 5). This structure was also identified as the global minimum by a subsequent set of conventional DE calculations, using the same optimization control parameters, but requiring significantly longer convergence, *i.e.* the optimum CDE calculation converged after only 461 generations, whereas the optimum DE calculation needed 988 generations for convergence (Fig. 4).

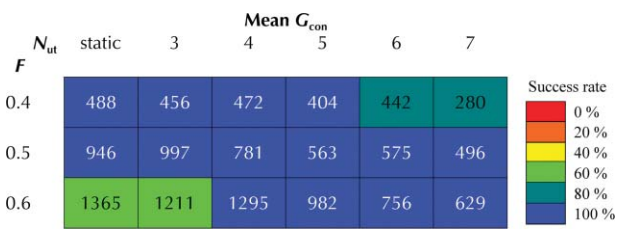



**Fig. 3** The mean rate of convergence for successful DE and CDE runs (those converged to the global minimum). The colour of each box denotes the % of successful runs for each set of control parameters.

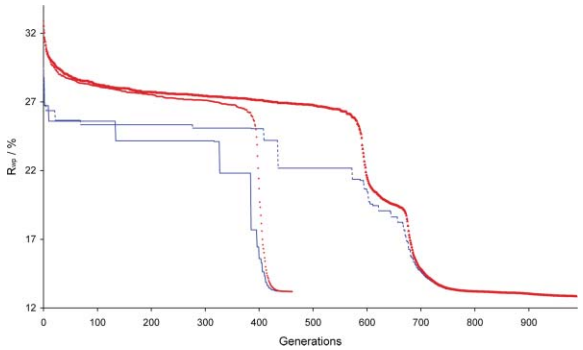



**Fig. 4** DE progress plot showing the best $R_{wp}$ (circles) and mean $R_{wp}$ (line) for each generation in (i) the optimum CDE calculation and (ii) the optimum DE calculation for (II).

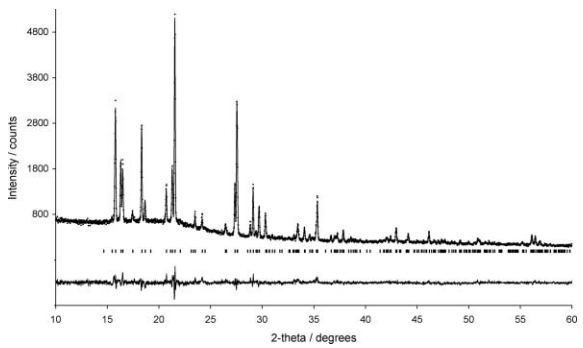



**Fig. 5** Final observed (circles), calculated (solid line) and difference (below) X-ray powder diffraction profile for the final Rietveld refinement of (II). Reflection positions are also marked.

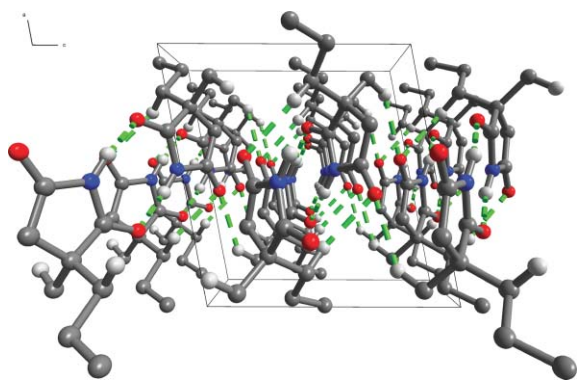


**Fig. 6** Crystal structure of (II). Only H atoms involved in hydrogen bonding (indicated by dashed lines) are shown. Selected intermolecular distances: N(H)···O, 2.80(1) Å; C(H)···O, 3.49(1) and 3.49(1) Å.

The crystal structure of (II) (Fig. 6) contains stacks of centrosymmetric N–H···O(=C) $R_2^2(8)$ dimers with C–H···O interactions between molecules in adjacent dimers forming additional $R_2^2(12)$ motifs within the stacks. A further C–H···O hydrogen bond produces a C(6) spiral running in the [010] direction between adjacent stacks. Combination of these motifs results in the formation of a hydrogen-bonded layer parallel to (100) with only weak hydrophobic interactions between the layers.

We have demonstrated that the Cultural Differential Evolution hybrid algorithm shows significantly quicker convergence on the global minimum when applied to crystal structure solution (an average of 40% improvement in tests[18]). The use of dynamic boundaries which are allowed to expand or contract with successive generations is an essential feature of our implementation, ensuring that the process does not become too restrictive. It allows the algorithm to follow population clustering that, while unlikely to expand in terms of parameter range, may shift in terms of absolute parameter values as the population evolves. Our work describes the first application of the concept of cultural evolution in a chemical or crystallographic context, and demonstrates the major gains in optimization efficiency that can be achieved by combining the dictates of biological and social evolution.

We thank the Royal Society (URF to MT) and the University of Birmingham for their support.

## Notes and references

† *Crystallographic data for (II)* $C_8H_{13}NO_2$: $M_r$ = 155.20, $a$ = 12.301(2), $b$ = 6.0698(4), $c$ = 11.6656(4) Å, $\beta$ = 100.958(4)°, $V$ = 855.1(2) Å$^3$, $P2_1/c$ (no. 14), $Z$ = 4, $D_c$ = 1.2055(2) g cm$^{-3}$, $T$ = 293 K.

*Data collection and Rietveld refinement*: Sample purchased from Aldrich, powder diffraction data ($10 \leqslant 2\theta \leqslant 60°$ in 0.020° steps over 1 h) collected on a Bruker-AXS D5000 using Ge-monochromated Cu-K$\alpha_1$ radiation and a linear PSD Rietveld refinement of all atom positions (except methyl and amide H in calculated positions) using geometric soft restraints (weighting factor of 0.001 for bond distances, 0.005 for geminal non-bonded distances), isotropic displacement parameters (non-H only) constrained by atom type, preferred orientation along [100]: ratio = 1.589. Final refinement gave $R_{wp}$ = 6.33%, $R_p$ = 5.08%, $\chi^2$ = 1.085.

CCDC 613003. For crystallographic data in CIF or other electronic format see DOI: 10.1039/b609138e


1 *Applications of Evolutionary Computation*, ed. R. L. Johnston, in *Struct. Bonding*, **110**, Springer, Berlin/Heidelberg, 2004.
2 A. P. Engelbrecht, *Computational Intelligence: An Introduction*, John Wiley and Sons Ltd., Chichester, UK, 2002, p. 171.
3 C. A. C. Coello and R. L. Becerra, *Eng. Opt.*, 2004, **36**, 219.
4 K. V. Price, in *New Ideas in Optimization*, ed. D. Corne, M. Dorigo and F. Glover, McGraw-Hill, London, UK, 1999, p. 77.
5 M. Wormington, C. Panaccione, K. M. Matney and D. K. Bowen, *Philos. Trans. R. Soc. London, Ser. A*, 1999, **357**, 2827.
6 C. C. Seaton and N. Blagden, *ACA Trans.*, 2004, **39**, 90.
7 N. Chakraborti, P. Mishra and S. Erkoc, *J. Phase Equilib. Diffus.*, 2004, **25**, 16.
8 D. E. McRee, *Acta Crystallogr., Sect. D*, 2004, **60**, 2276.
9 R. Thomsen and M. K. Christensen, *J. Med. Chem.*, 2006, **49**, 3315.
10 O. Oeckler, T. Weber, L. Kienle, H. Mattausch and A. Simon, *Angew. Chem., Int. Ed.*, 2005, **44**, 3917.
11 C. C. Seaton and M. Tremayne, *Chem. Commun.*, 2002, 880; M. Tremayne, C. C. Seaton and C. Glidewell, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 823.
12 K. D. M. Harris, M. Tremayne and B. M. Kariuki, *Angew. Chem., Int. Ed.*, 2000, **35**, 3523; *Structure Determination from Powder Diffraction Data*, ed. W. I. F. David, K. Shankland, L. B. McCusker and C. Baerlocher, Oxford University Press, Oxford, UK, 2002.
13 K. D. M. Harris, M. Tremayne, P. Lightfoot and P. G. Bruce, *J. Am. Chem. Soc.*, 1994, **116**, 3543; M. Tremayne and C. Glidewell, *Chem. Commun.*, 2000, 2425.
14 W. I. F. David, K. Shankland and N. Shankland, *Chem. Commun.*, 1998, 931; Y. G. Andreev, P. Lighfoot and P. G. Bruce, *Chem. Commun.*, 1996, 2169.
15 E. Y. Cheung, E. E. McCabe, K. D. M. Harris, R. L. Johnston, K. M. P. Raja and P. Balaram, *Angew. Chem., Int. Ed.*, 2002, **41**, 494.
16 R. L. Becerra and C. A. C. Coello, *Comput. Methods Appl. Mech. Eng.*, 2006, **195**, 4303.
17 M. Rossi, R. Meyer, P. Constantinou, F. Caruso, D. Castelbuono, M. O'Brien and V. Narasimhan, *J. Nat. Prod.*, 2001, **64**, 26.
18 S. Y. Chong and M. Tremayne, manuscript in preparation.