OptiCat: a versatile open-source optimisation platform for experimental design

F. Clerc, D. Farrusseng, C. Mirodatos

OptiCat: A versatile open-source optimization platform for experimental design

Frédéric Clerc, David Farrusseng *, Claude Mirodatos

Institut de Recherches sur la Catalyse et l'Environnement de Lyon , IRCELYON, UMR 5256 CNRS/Université de LYON - 2, Av. A. Einstein - F-69626 Villeurbanne, France

A new open-source software for experimental design is presented. It consists at a platform which enables to design specific optimization algorithms by simple drag and drop from a toolbox. Complex workflows can be implemented and tested within a few minutes by experimentalists. OptiCat can integrate MS Excel©, MatLab© and Statistica© functionalities for batch calculations. Optimization results on a custom benchmark are presented as examples. Four different algorithms are studied: genetic algorithm, evolutionary strategy, Tabu Search and simulated annealing. Advantages to use OptiCat in the frame of parallel experimentation in Material Science and Catalysis are addressed.

1. Introduction

The so-called High Throughput Screening (HTS) approach is readily expanding to many research domains such as polymer coating, electroluminescent materials, homogeneous and bio-catalysis [1]. This screening methodology consists at synthesizing and testing a collection of samples (called library) at the same time using parallel and fully automated devices. Today's technology enables to screen several dozen to several hundreds of samples in a day. However, when all possible combinations are considered, the high number of parameters to investigate lead to a combinatorial explosion of samples to be prepared and tested. The systematic investigation of the entire parameter space would result in an inefficient optimization that may take years despite HT technology.

In drug discovery, combinatorial optimization and experimental design use data mining in order to efficiently reduce the experimental effort [2]. Among data mining techniques, metaheuristic are Artificial Intelligent based methods which are generally applied to problems for which there is no satisfactory problem-specific algorithm or heuristic; or when it is not practical to implement such a method. Metaheuristics are usually targeted to combinatorial optimization problems, but of course can handle any problem that can be recast in that form. The goal of combinatorial optimization is to find a solution (such as a bit string) that maximizes an arbitrary function specified by the user [3]. Genetic Algorithms [4], Evolutionary Strategy [5], Hill Climbing [6], Simulated Annealing [7,8] and Tabu Search [9] are some of the most studied and applied metaheuristics. Despite numbers of successful applications of metaheuristic algorithms in diverse domains such as scheduling, e-purchasing or industrial design the use for discovery and optimization in Material Science and Catalysis is still scarce [10].

A first explanation for not using this strategy is the lack of confidence in the robustness and reliability of these algorithms. Indeed, a project for discovery of a new formulation can imply the screening of several hundreds to a few thousands of samples which can take months. Secondly, another aspect which hinders chemist to use metaheuristics for library design is the lack of software available on the market which would enable to build custom metaheuristics and to encode variables from complex studies. Indeed, metaheuristics are usually coded either in programming languages (Java, C++,..) or MatLab© platforms which make them not user friendly and oblige the assistance of informatics expert on site.

This paper presents the main functions of an optimization platform, named OptiCat, designed for chemists/physicists who are seeking for user-friendly tools for design of experiments and who are neither expert in programming nor in advanced statistics. OptiCat is an open-source software which can be used according CeCILL license [11,12]. As a proof of concepts, we show the implementation of different metaheuristic algorithms for solving a custom made problem as a mathematical surface response. This surface response is designed on chemical knowledge for catalytic CO oxidation in gas phase. The shape of the benchmark's response surface is defined by a set of mathematical functions mimicking the “real world”. The chemical knowledge used for defining the benchmark derives both from literature and local expertise based on real experiments.

* Corresponding author. Tel.: +33 4 72 44 53 65; fax: +33 4 72 44 53 99.
E-mail address: david.farrusseng@catalyse.cnrs.fr (D. Farrusseng).

2. Materials and methods

2.1. OptiCat description

OptiCat is a software which enables to design data treatment workflows easily. OptiCat consists at a library of data treatment tools (called nodes) such as for instance the GA operators which can be assembled for building custom made algorithms by simple drag and drop from the component tree (Fig. 1a). The Fig. 1b illustrates a Genetic Algorithm data workflow. The "Initialization" node provides a graphical user interface to specify the number of variables, their types and encoding, and the population size. Stopping criteria, such as the maximum iteration number can be recorded in the "Loop" node. The "Evaluation" node calculates fitness from a user defined function. Population data records and fitness visualization can be specified in the "Historian" node. The three other remaining nodes correspond to classical Evolutionary Algorithm operators, namely "Selection", "Crossover" and "Mutation". For each of these operators, several methods have been implemented in OptiCat.

Hundreds of different data workflows can be designed in OptiCat by combining nodes from the component tree, regardless options available in most of the nodes. Similarity/diversity measurements between individuals such as indices are also included which enable to monitor the distance between individuals for Simulated Annealing, Tabu Search algorithms as well as for sharing options in evolutionary algorithms. Besides, additional modes developed in purpose by the user can be inserted into the toolbox which makes OptiCat a very flexible platform for metaheuristic algorithm design.

This software is mainly intended to experimentalists who are seeking for stochastic experimental design processes and who are not programmers. In the "real" operating mode, the node "ES manual" allows the user to enter one response value (the fitness) for every individuals at each iteration (the yield of a catalyst for instance). OptiCat provides a $(n+1)$ population from a nth population associated with corresponding fitness. For experimentalists, it enables to generate a library of compounds based on the results of the previous generation i.e. a new optimized experimental planning. In addition, a "simulation mode" enables to assess the efficiency of user made algorithms on mathematical surface responses. This mode acts as a function solver which provides an ensemble of solutions for an implemented function. It is adapted to find optima of complex multivariate functions. Several published surface responses which are commonly considered as benchmarks such as Keane, Fintdec, Schwefel or Himmelbau functions are implemented in OptiCat. Those benchmarks are complex multivariate functions for which maxima are hard to find and which are used to compare algorithm efficiency and reliability [13,14]. Depending on the user expertise in programming two other alternatives are also provided (i) custom mathematical expressions can be integrated from Delphi© or MS Visual Basic©, (ii) alternatively, response function computations can be performed directly using classic MS Excel© functions or Matlab©. In addition, Opticat can integrate MatLab© libraries and Statistica© functions such as PCA or Artificial Neural Networks. This feature allows designing complex workflows by combining existing cutting-edge modelling tools in a flexible optimization platform. Finally, hundreds on runs can be performed within a few minutes starting either from a first randomly designed generation or from an initial library computed elsewhere. OptiCat operates on single core processors under Windows 2000/XP systems. Tutorials, videos and associated training files can be downloaded free of charge [11].

Table 1
Global and local maxima of the benchmark

Name	Optimum	Noble metal					Transition metal	T (C)	Support	Yield
		Pt (%)	Pd (%)	Au (%)	Ru (%)	Cu (%)	Sum (%)			
$PtPdAl_2O_3$	Global	1.5	1.5	0	1	0	0	220	Al_2O_3	1.0
$CuCeO_2$	Global	0	0	0	0	4	16	250	CeO_2	1.0
$AuTiO_2$	Global	0	0	4	0	0	0	260	TiO_2	1.0
$AuTiO_2$	Global	0	0	0.31	0	20	0	466	TiO_2	1.0
ZrO_2	Local	0	0	0	0	0	20	550	ZrO_2	0.8

2.2. Benchmark description

An experiment consists at measuring the yield (i.e. fitness) of catalysts under different process conditions. A catalyst is described by

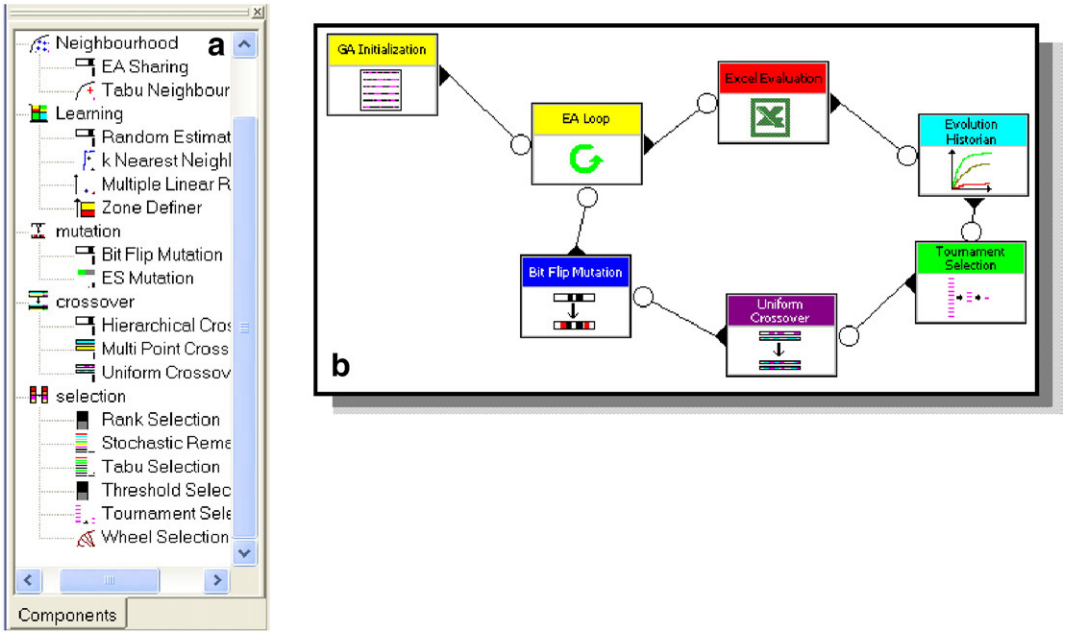

Fig. 1. Screenshots of OptiCat: a) the component tree showing data treatment nodes and b) evolutionary strategy data workflows.

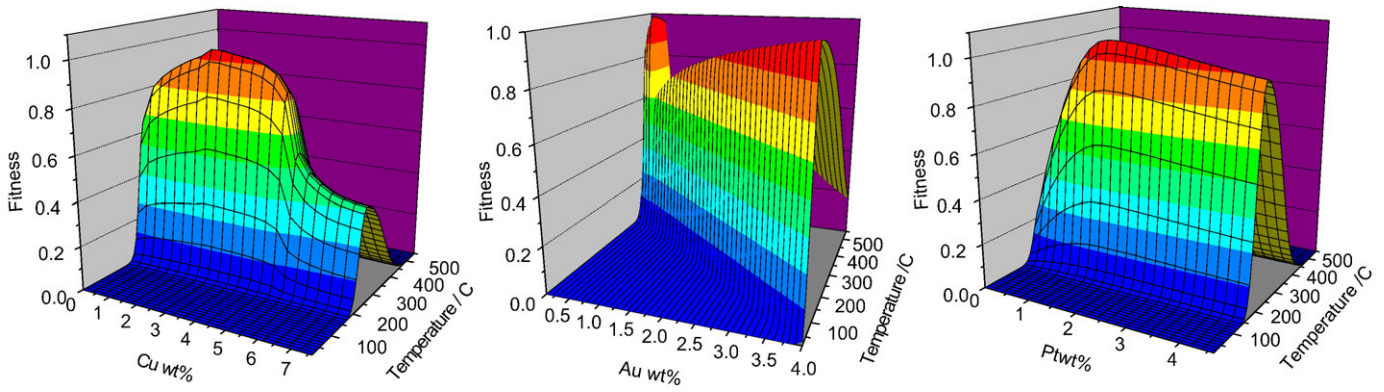

Fig. 2. Shapes of the three global maxima, a) Cu/CeO_2, b) $PtPd/Al_2O_3$, c) Au/TiO_2.

12 continuous variables accounting for the mass composition (wt.%) of respective elements (4 noble metals Pt, Pd, Au or Ru and/or 8 transition metals Cu, Co, Cr, Ni, Zn, V, Mo and Fe) and 1 categorical variable of 4 modalities which accounts for the support (TiO_2, Al_2O_3, CeO_2 or ZrO_2). The mass composition for noble metal and transition metals can range from 0 to 4% and 0 to 20%, respectively while the temperature (T) can vary from 20 to 550 C. The process condition is defined by the temperature at which the measured is carried out.

The mathematical expression of the benchmark is reported in Supporting Data. It is a product of two normalized expressions (0 to 1) named X and B, respectively. The expression X accounts for the yield according to catalyst composition while B expresses the range of temperature for which the catalysts is active and at which it extends.

$$\text{Yield} = X(\text{wt.\%, support}) \times B(f(\text{wt.\%, support}), T)$$

The function encompasses three distinct global maxima with a response value (fitness value) equal to 1 and large local maxima at 0.8 when the support ZrO_2 is set. The optima are: reported in Table 1.

Response surfaces of the three maxima along the highest relevant axis are shown in Fig. 2. Because of the smooth shape of the Cu/CeO_2 surface response, it is expected that this optimum would be easy to find by metaheuristics. On the other hand, the PtPd system should be more difficult to discover because of the deep fall near the global optimum. The most difficult global optimum to find would be for Au/TiO_2 for which the optimum peak is very sharp and because of the "deceptive" features. Deception cases make difficulty to metaheurisics because in the functions lower-order schema information is misleading, thereby causing an optimization algorithm to get attracted to sub-optimal solutions [13]. In order to evaluate the probability to find a global maximum by a random manner, a statistical study was performed. The fitness of 100,000 randomly generated data points in the search space was computed. The Fig. 3 indicates the number of individuals exhibiting fitness above selected threshold values. The shape of the curve resembles to an exponential decay with 80% of the individuals that have a fitness value below 0.1 while 0.04% have a fitness value above 0.95. Moreover, individuals were counted with respect to the local/global maximum they belong to. The results are shown in the Fig. 4 and represent the relative "size" of maxima. More precisely, it is an evaluation of the "hyper" surface area at different fitness value levels. It shows that the probability to discover the Au/TiO_2 maximum (fitness above 0.95) is much greater than Cu/CeO_2 on a random manner.

2.3. Algorithm description

We used OptiCat for implementing four different optimization methods. Namely, Simulated Annealing (SA), Tabu Search (TS), Genetic Algorithm (GA) and Evolutionary Strategy (ES). For each method, 50 runs were performed on the benchmark described above starting from a different randomly generated generation. For the sake of comparison, the optimization process is stopped at the same number of individual evaluations (i.e. 400). For GA and ES, the optimization proceeds in 10 iterations of 40 individuals whereas for the TS 80 iterations are performed with 5 individuals and 400 iterations for the SA. All four algorithms do not have specific features and were not

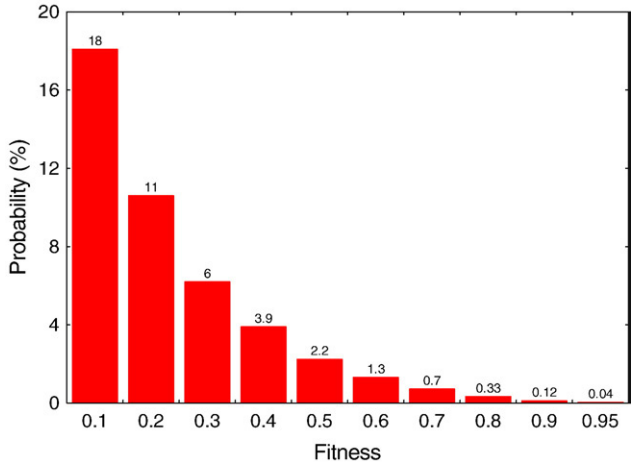

Fig. 3. Distribution of the fitness function.

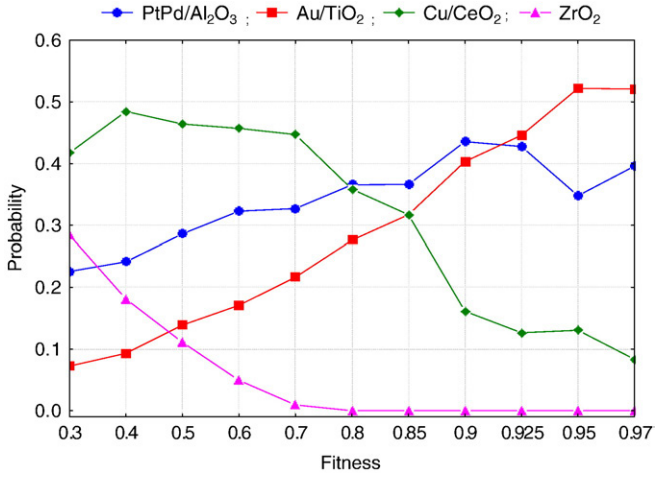

Fig. 4. Size of the different global and local maxima as function of the fitness function.

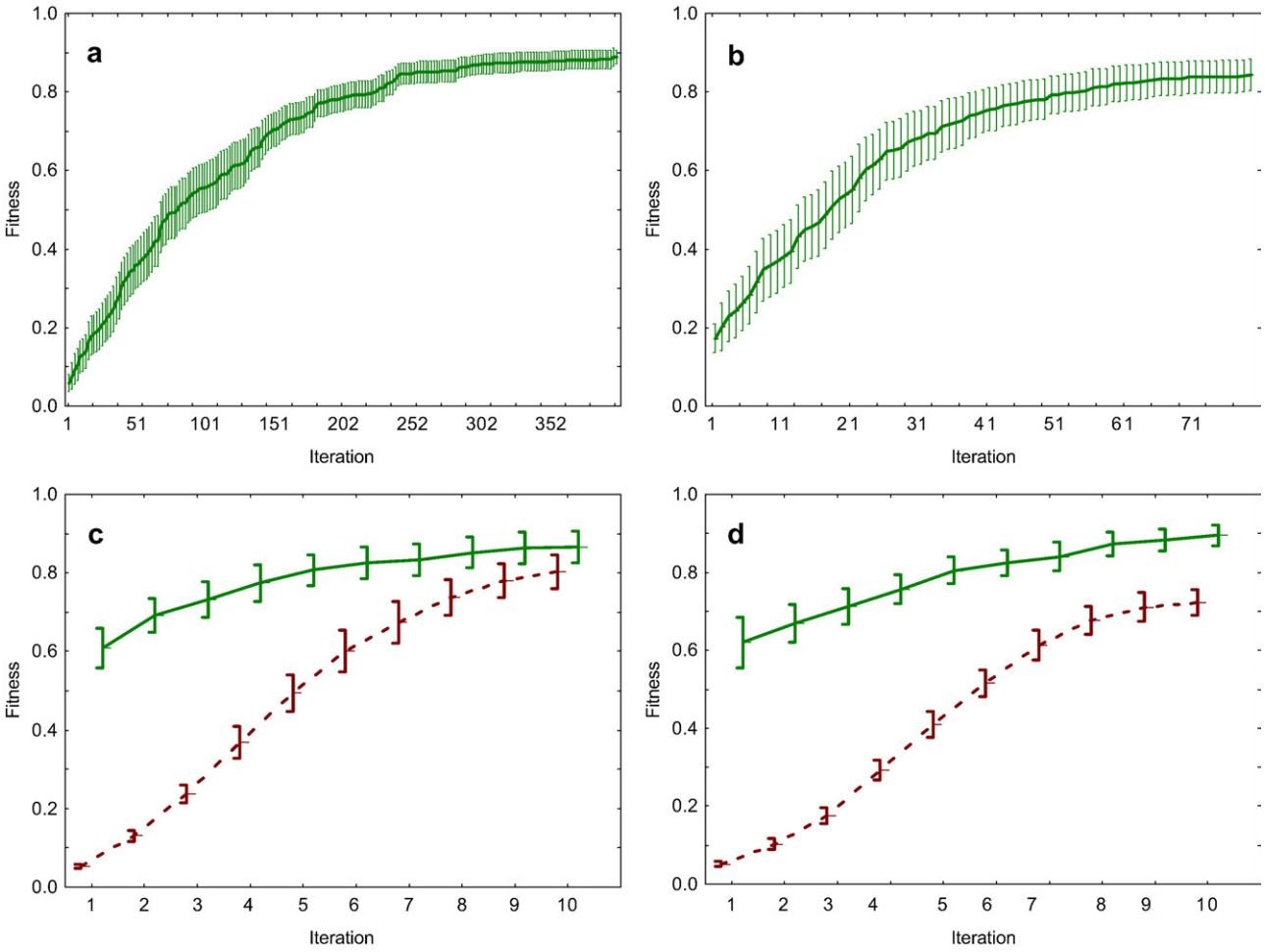

Fig. 5. Results of the optimization on the benchmark for a) Simulating annealing, b) Tabu Search, c) Genetic Algorithm and d) Evolutionary Strategy. The plain line is the average on the 50 runs of the best individuals a each iteration while the dotted line is the average on the 50 runs of the average fitness at each iteration. The bars are the 95% confidence intervals calculated on 50 independent runs.

optimized for this particular benchmark. Details on the algorithmic settings are given in supporting information.

3. Results and discussions

Results of the optimization study using Simulated Annealing, Taboo Search, Genetic Algorithm and Evolutionary Strategy are presented in Fig. 5.

All four algorithms tend to converge towards global maxima without reaching them however. Global optima are found at greater iteration number and population sizes. For GA and ES typically 20 iterations and 100 as population size enable to get one of the optima (fitness >0.97) with 95% of probability. The 95% confidence interval is reasonably narrow for all algorithms while smaller for the SA. This indicates a very high success rate when an optimization process is undertaken whatever algorithms and starting libraries. The robustness and reliability of an optimization process is of prime importance in practice because high costs of R&D do not allow duplicating experiments. For a screening of 400 catalysts on a random basis there is 1.3% of probability to obtain individuals exhibiting fitness above 0.8 whereas using a ES there is 95% of probability to get at least one individual between 0.85 and 0.92. The ES shows slightly better performances than GA especially at high iteration numbers. It might be explained by higher diversity of the population in the ES. Indeed, the average of population fitness is significantly lower in the case of the ES which indicates a higher diversity and in turn a greater driving force to seek global optima.

Because the benchmark consists at several global maxima algorithms can converge preferentially to one or to the other. Statistics on the distribution of global maxima are calculated on 50 runs (Table 2).

Although SA shows better skills for converging towards higher fitness values, GA and ES approaches are preferred for use in practice because of the use of parallel tools for synthesis and testing of large library at a time. In addition, SA and TS converge preferentially towards large maxima whereas GA and ES enable to discover maxima with very low probability to be found on a random manner.

4. Conclusions

The implementation of four different optimization algorithms was performed in OptiCat and tested on a custom virtual function which is used as benchmark. The reliability of the algorithms was validated by measuring the confidence intervals on 50 runs. The effects of

Table 2
Probability to obtain at least an individual exhibiting fitness above 0.95 at the last iteration and statistical distribution among the three maxima

Method	Highest fitness >0.95 (%)	$PtPd/Al_2O_3$ (%)	Au/TiO_2 (%)	Cu/CeO_2 (%)
SA	50	52	32	16
TS	28	78	7	14
GA	38	42	21	36
ES	45	36	31	31

algorithm settings, such as population size, sharing and others, are under investigation on different benchmarks. We believe that OptiCat is a powerful tool for library and experimental design in the frame of high throughput experimentation.

5. Referee general comments

OptiCat was independently tested by Dr. Francois Gilardoni, Inforsense Ltd (London GB). "This software will allure the user by the modular type of design which is proposed: to obtain a program by a construction of various software "bricks" of bases, presented in the form of graph objects, available in an important library of algorithms. Its field of application of origin (combinatory catalysis) can be largely open to many other applications in the field of the analysis and the data processing. However it would be interesting to have "on line" examples with accompanying notes and detailed to facilitate the training."

6. Supporting data

Mathematical expression of the surface response used as benchmark and algorithm parameters used in the study are can be found in Supporting Data.

Acknowledgments

We thank the European Commission for supports in the frame of TopCombi (Integrated Project No. NMP2-CT2005-515792). We also thank the "Region Rhône-Alpes" in the frame of the program student mobility and training. We thank R. Rakotomalala and Prof. N Nicoloyannis for stimulating scientific discussions.

References

[1] B. Jandeleit, D.J. Schaefer, T.S. Powers, H.W. Turner, W.H. Weinberg, Angew. Chem., Int. Ed 38 (1999) 2494–2532.
[2] J. Cawse (Ed.), Experimental Design for Combinatorial and High Throughput Materials Development, Wiley-VCH, Weinheim, 2003.
[3] F. Glover, G. Kochenberger, Handbook of Metaheuristics, Springer, 2003.
[4] J. Holland, Adaptation In Natural and Artificial Systems, Ann Arbour, 1975.
[5] H.P. Schwefel, Evolution and Optimum Seeking, Wiley & Sons, New York, 1995.
[6] M. Mitchell, J. Holland, S. Forrest, Adv. Neural Inf. Process. Systems (1994) 51.
[7] v. Cerny, J. Optim. Theory Appl. 45 (1985) 41–51.
[8] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science 220 (1983) 671–680.
[9] F. Glover, M. Laguna, in: C Reeves (Ed.), Modern Heuristic Techniques for Combinatorial Problems, Blackwell Scientific Publishing, Oxford, England, 1993.
[10] D. Farrusseng, Report for The Catalyst Group Resources — Catalytic Advances Program, 2006.
[11] OptiCat can be downloaded at http://eric.univ-lyon2.fr/~fclerc/ and http://a.farrusseng.free.fr.
[12] Cecill license terms can be found at http://www.cecill.info/.
[13] K. Deb, S. Agrawal, Understanding Interactions among Genetic Algorithm Parameters, 1998.
[14] J.G. Digalakis, K.G. Margaritis, Intern. J. Comput. Math 79 (2002) 403–416.