Meta Modeling for Combinatorial Catalyst Optimization

Fréderic Clerc[†,††], David Farrusseng[†], Ricco Rakotomalala[††], Nicolas Nycoloyannis[††], Claude Mirodatos[†]

[†]Institut de la Recherche sur la Catalyse, CNRS-UPR 5403, F-69626 Villeurbanne, France
[††]Laboratoire ERIC, Université Lumière Lyon 2, F-69500 Bron, France

Summary
Our aim is to find the best catalyst, the best combination of compounds, in order to optimize a chemical reaction. The chemists use mainly a heuristic algorithm, especially an evolutionary algorithm, to achieve the best combination. In this paper, we outline a variant of evolutionary optimization algorithm, says meta modeling. Our idea is to combine a statistical learning algorithm with the optimization process. The goal is a better use of the past experience, the labelled individuals, in the guidance of the search exploration of the optimal solution. The approach is especially useful in the combinatorial catalysis optimization because the fitness function is unknown and the labelled individual is obtained by real chemical reaction. This is highly costly and takes time. We show on a well-known chemists' benchmark that our process slightly the average performance of the standard evolutionary algorithms. But numerous problems remain opened. We try to inventory them in order to define our future work to improve the approach.
Key words: *Optimization, Data Mining, Combinatorial Catalysis*

1. Introduction

Optimization algorithms receive great interest both in the academy and the industry. In particular, evolutionary algorithms (EA) have been demonstrated well fitted for solving discrete, discontinuous or noisy problems. Consequently, these types of methods are often used in real-world applications. Evolutionary strategies (ES) and genetic algorithms (GA) are becoming more and more used in combinatorial catalysis. For example, in 2000, efficient catalysts for oxidative dehydrogenation of propane were found with this methodology [1]. Libraries of catalysts have been synthesized and tested iteratively, according to the proposals made by an ES. After some generations, the targeted compound presenting the best performance, the *optimum*, was found.
But standard EAs algorithms are not so well suited to catalysis. Indeed, in the general situation of the use of EAs, a large number of evaluations i.e. labelling each individual, are often needed to find the best individual. It is not scarce to let dozens of individuals evolve through hundreds of generations. The label of an individual is the output value of the mathematical function to optimize. In the catalyst context, the problem to solve has no mathematical expression. In terms of optimization: the fitness function is not *a priori* known. Each evaluation is in fact an experiment where the chemical reaction is really synthetized and the output measured. Consequently, even if a few real experiments show promising results, they remain scarce because of their cost [2].
In order to improve the optimization process, especially to quicken the achievement of the best solutions, the main idea of this paper is the better use of the past experiences, the labelled individuals, using data mining methods. Data mining is the practice of automatically extracting important patterns from data using machine learning or statistical algorithms. Although it is usually associated with commercial sales and purchases [3], numerous applications exist in the industry [4], and solutions using data mining have been proposed in combinatorial catalysis [5]. The idea is to use accumulated labelled data about known catalysts for a given reaction for building a mathematical model. This model can be used for predicting the performances of new, unknown catalysts. But for now, the state of knowledge does not really enable to build a reliable model for designing formulations with targeted performance. This is because of the lack of reliable data: no open wide database exists and literature often presents small datasets that poorly maps the catalysts search space. Moreover, a catalyst is not universal: each chemical reaction requires very different compounds.
In this paper, we present a system which enable to save experiments by combining advantages of data mining and evolutionary optimization algorithm: a meta modeling algorithm. The basic idea was already reported in other studies [6]. The starting point consists in a real catalyst library which is synthesized and then tested. The corresponding information is stored in a database. We build a statistical model from this database using a knowledge discovery algorithm in order to label new virtual individuals proposed by genetic operators. The best individuals according to the predicted label from the statistical model are really synthetized i.e. we make the chemical reaction; we obtain the true label and the

Manuscript received October 9, 2006.
Manuscript revised October 25, 2006.

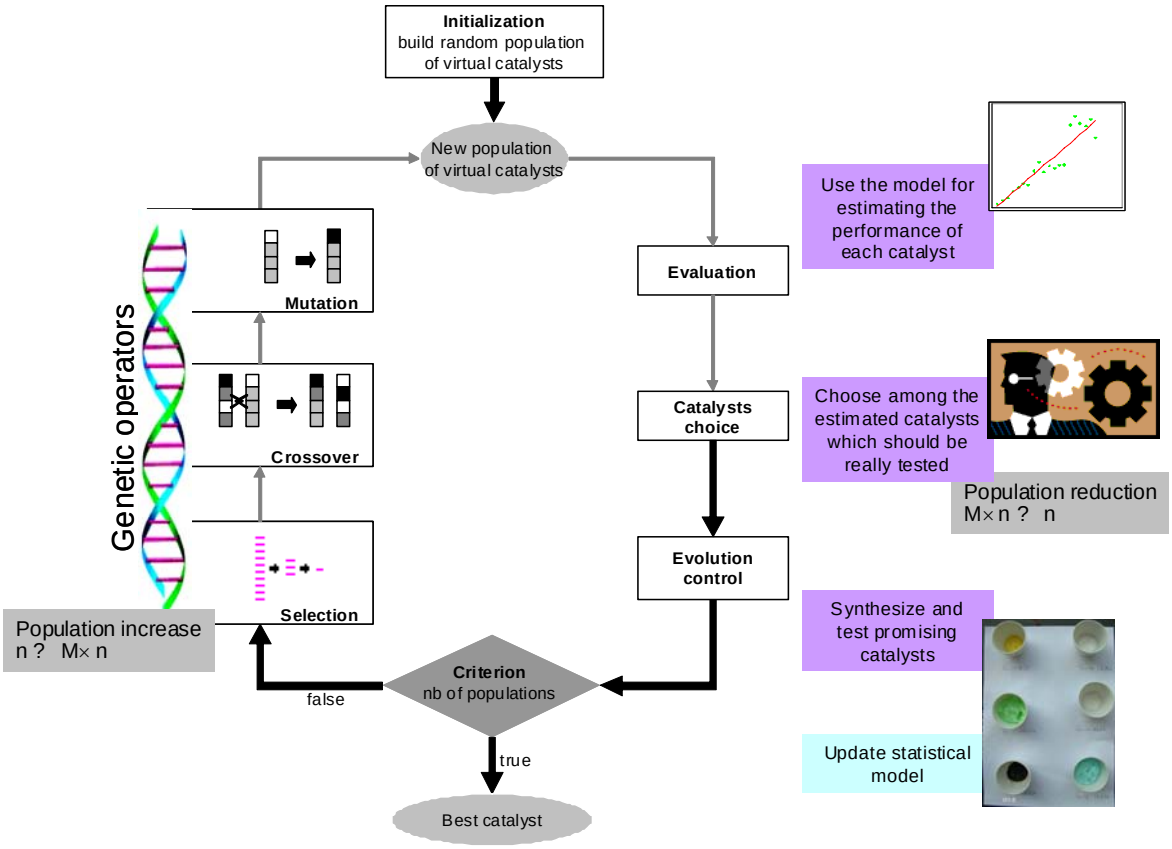

Fig. 1 A meta-modeling algorithm with model management in catalysis

resulting information are added to the database. At the next iteration, the knowledge algorithm will fine up the prediction. This process is repeated until the checking of a given criterion. The improvement of statistical model after each generation shall enable to direct the design of the libraries by a virtual pre-screening.

In the next section, we present the concept of meta modeling and we discuss its fitting with regard to catalysis. The aim of this paper is proving the validity of the concept before real experimentation. Consequently, for this computer experiment, we used a catalysts virtual response surface that we depict in the third section. This catalysis-suggested function to be optimized comes from literature. For demonstrating the efficiency of meta modeling, we compare its performance with standard evolutionary optimization algorithms. We will conclude by discussing why the meta modeling approach is a good choice for catalysts design.

2. Meta modeling in the combinatorial catalysis optimization context

2.1 Meta Modeling

Suggested by genetic algorithms [7], different meta modeling (MM) techniques exist: fitness inheritance [8], knowledge-relative genetic operators [9] or model management [10]. We retain this last option whose basic principles are illustrated on figure 1. This approach makes more sense because statistical models are a well-known methodology in catalysis. In the following, when we will refer to “meta modeling algorithm”, we assume the model management approach.

The steps of the MM algorithm are:

1. **Initialization**. Generate a random population of virtual catalysts (M x n individuals).
2. **Evaluation**. Estimate the performance of the catalysts through a statistical model, based on previously acquired real data. In the first iteration, because no dataset is available, we assign a random label.

3. **Catalysts choice**. Choose the most promising catalysts according to the statistical labelling. The population size is reduced to *n* individuals.
4. **Evolution control**. Synthesize and test promising catalysts pointed up by the model computation. This is the bottleneck of process. Each real chemical reaction experiment is very costly and takes a lot of time.

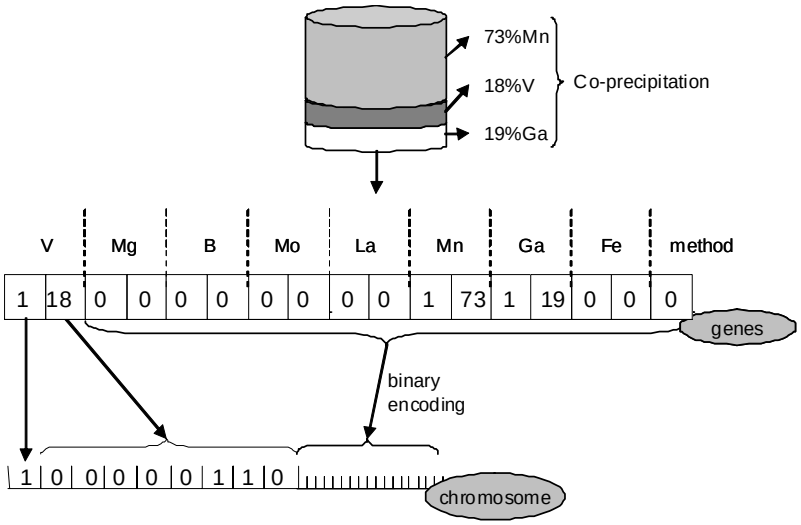

Fig. 2 Binary encoding, from the native description to a chromosomal description of a catalyst

5. **Update the statistical model**. New examples are available in the database; we can use them in order to update the statistical model that we use to assign a label to individuals in the step 2.
6. **Criterion**. If verified then the evolution stops. The main criterion is the number of iterations because the budget (time and money) is the real constraint of our optimization process.
7. **Genetic operators**. Produce a new population of virtual catalysts. The population size is increased to $M \times n$ individuals.
   a. Selection: pick an amount of catalysts in the population
   b. Crossover: mix the characteristics of the catalysts
   c. Mutation: include random changes in formulations
8. Back to step 2

Increasing the size of the population improves the efficiency of the EA algorithm. It is a well-known phenomenon. But in the catalyst optimization process, we cannot make use of this property because the real labelling of the individuals, the synthesis step (Step 4), is very costly. Our idea is to substitute to the real labelling, a statistical labelling, in order to better use the increasing diversity of the population in the step 7 (Genetic operators). The learning statistical model step, multiplying the individuals and applying the statistical model are costless operations, especially compared with the synthesis step.

Practically, a population multiplier M is applied at the selection step. If n catalysts compose the population, then $(M \times n)$ virtual individuals will be generated through the genetic operators. Meta modeling allows the choice of the individuals that will be synthesized and tested. The increase of population is motivated by the human choice: more opportunities are offered to the experimenter. Nevertheless, this is a delicate operation and should be done very carefully. The diversity proposed by the algorithm should be respected, and should not be guided by subjective matters such as the catalyst synthesis easiness. Once chosen, the catalysts for evolution control need weeks of human labor.

2.2 Meta modeling and standard genetic algorithm

We want to compare our MM approach to standard evolutionary algorithms. We especially study two algorithms: evolutionary strategy and genetic algorithm. The particularity of the GA is the binary encoding of the values (catalyst composition) during the genetic operator phase (step 7).

The main difference between the meta modeling approach and these standard EA are the increasing of the population (Step 7), the virtual labelling (step 2) and the selection (Step 3). For the standard EA, M=1 i.e. the size of the population does not vary during the optimization process. There is a lack of diversity of the population if n is small. In an optimization problem with a mathematical fitness function, we can increase n like we want. In a catalyst problem, n is often a constraint specified by the

characteristics of the robots used for the chemical reaction. Numerous of them can handle simultaneously a few number of reactions for an experiment. An experiment runs for a long time. Using the multiplier M (M > 1) improves the diversity of the population. But all the individuals cannot be synthesized. The selection process based on the predicted label from the statistical model allows leading the process on the best solutions, without increasing in any manner the cost of the experiments.

2.3 Data mining algorithms

The statistical model used for computing the predicted label of individuals plays a crucial role in the process. In the ideal, it should reproduce the fitness function to be optimized. But it is seldom possible. The question that should be raised then is: "which characteristic of the classifiers must be pointed up in this context?"
A rough model allows much freedom for the search space. Consequently, it avoids a premature convergence towards a local optimum. But on the other hand the guidance is weak, the convergence to the global optimum is slow. At the opposite, an accurate model will guide steadily the process toward an optimum, but this optimum is often in relation to the representation bias of the classifier that is not the fitness function to optimize. The ideal combination would be a model formulation capable of progress during the evolution, it must above all indicate us the good directions to be explored for optimization. The important issue in evolutionary computation is the browsing/exploitation ratio [11].
Among various data mining algorithms [4], we retained the classical linear regression. It consists in a mathematical function built according to known data; any new individual is estimated using this function. We chose this method because it is particularly simple and still efficient. Not reported here, other learning algorithms such as nearest neighbor or PLS (Partial Least Square) regression are also used but did not improve significantly the results.

3. A benchmark catalysis for the experiments

3.1 A synthetic chemical reaction

At this step of the advancement of our researches, it is not possible to evaluate the efficiency of our approach on a real chemical reaction. To show the efficiency of the meta modeling approach, we examine the optimization of virtual catalysts according to a theoretical response surface. Considering the evolution control step, this means that instead of synthesizing and testing the catalyst, a theoretical fitness value will be attributed. It is calculated according to a mathematical function, which reflects the behavior of the catalyst in real conditions.
We name Φ the function suggested by the oxidative dehydrogenation of ethane and propane [16]. Possible catalysts can be composed of eight elements: V, Mg, B, Mo, La, Mn, Fe and Ga. Moreover, an additional value is considered: the preparation method, either co precipitation or impregnation. Algebraically, we write a catalyst X as in equation 1

$$\begin{aligned} &X = (x_v, x_{Mg}, x_B, x_{Mo}, x_{La}, x_{Mn}, x_{Fe}, x_{Ga}, method) \\ &x \in [0,1] \\ &method \in \{0,1\} \end{aligned} \quad (1)$$

The performance of a catalyst, the response value, is defined in equation 2

$$\Phi = \begin{cases} X_1 \times S_1, \text{ if method} = 0 \\ X_2 \times S_2, \text{ if method} = 1 \\ 0, \text{ if } x_{La} > 0 \text{ or } x_B > 0 \end{cases} \quad (2)$$

Where

$$\begin{aligned} S_1 &= 66 x_v x_{Mg} (1 - x_v x_{Mg}) + 2 x_{Mo} - 0.1 x_{Mn} - 0.1 x_{Fe} \\ X_1 &= 66 x_v x_{Mg} (1 - x_v x_{Mg}) - 0.1 x_{Mo} + 1.5 x_{Mn} + 1.5 x_{Fe} \\ S_2 &= 60 x_v x_{Mg} (1 - 1.3 x_v - x_{Mg}) \\ X_2 &= 60 x_v x_{Mg} (1 - 1.3 x_v - x_{Mg}) \end{aligned}$$

The optimum is a compound containing 32% of Vanadium, 32% of Magnesium and 36% of Molybdenum, the method being co precipitation: $\Phi_{\max}$ (0.32, 0.32, 0, 0.36, 0, 0, 0, 0) = 7.55. Moreover, Φ presents a local optimum Φ_{local} (0.66, 0.33, 0, 0, 0, 0, 0, 1) = 7.1.
For genetic algorithm and meta modeling approach, we use a binary encoding during the genetic operators phase (Step 7). Seventeen *genes* represent the composition of the catalyst. Two genes represent each element: a Boolean for its presence, a real for its percentage. The seventeenth gene is also a Boolean; it represents the preparation method, 1 for co-precipitation, 0 for impregnation. An example of a chromosome is presented on figure 2.

3.2 Parameters of the experiments and evaluation criteria

To simulate the real conditions of an experimentation, we limited the total number of evaluations to 400 by organizing them in the following way: at each iteration, we consider that the robots can handle simultaneously 40 chemical reactions, the reactions that a robot can handle simultaneously is called a library; the number of iterations

is 10. At the end of the process, we pick the best proposal of each algorithm.
In order to obtain more reliable results, we repeat 50 times the whole process and we propose two evaluation criteria. The first one measures the average performance of the optimization algorithm over the 50 experiments. This indicates the performance of the approach, the expected value of the optimization when we use it. The second one measures the number of times where the approach reaches the optimal value (response value is equal to 7.5). This indicates the reliability of the approach, the capacity to find the real global optimum. These criteria are complementary. The good optimization approach must always reach the global optimum (reliability criterion), but we want that when it does not find the global optimum, it reaches nevertheless a good solution (performance criterion).

4. Results of experiments

4.1 Results and comments

Comparative results with other evolutionary algorithms are depicted in the table 1.

Table 1 Results over the 50 experiments. Comparison of meta modeling with other EA algorithms

Opt. Algorithm	Performance	Reliability
Meta modeling	7.275	10
Evolutionary Strategy	7.05	12
Genetic algorithm	6.675	8

The multiplier M is set to 100 in our experiments. The meta modeling obtains good results. The average performance over the 50 experiments is 7.275 while the global optimum is 7.5. We can be confident with the proposal of this approach during its use. We note that on 10 executions among 50, we reach the global optimum. These are an encouraging results for the utilization of this approach in a real chemical reaction. The approach often suggests solutions near the global optimum.
Compared to other EA optimization methods, MM seems slightly better. In fact, the real confrontation can be made between the MM and the GA that we can consider like a MM approach with a multiplier (M = 1) and without statistical prediction. The difference between ES and GA seems to show that there are some problems during the binary encoding. That could make some interference into our approach that is also based on a binary encoding of the data.
Numerous open discussions can be started from these results. We did not really evaluate the influence of the multiplier M. We can set this parameter as large as possible, but there is probably a value from which it is not necessary to increase it any more. Another open problem is the role of the learning algorithm. We use a simple linear regression in our experiments. It seems that it is not necessary to use very sophisticated classifiers in the meta modeling framework. But we suspect that the learning characteristics must be examined in relation to chemical reaction characteristics.
Precisely, at this stage of our advance, we use only simulated chemical reactions in our experiments. Our benchmark is well known in the chemists' community. It is a smooth landscape composed of only one global optimum that is quite easy to find using standard heuristic optimization algorithm, even if the surface is discontinuous. The main difficulty is the limited number of evaluations. The acceleration of the convergence is a crucial property of the optimization algorithm in this context. We cannot say on the other hand if an excessive acceleration will not be penalizing when surface to be explored is very wiggly with many local optima.

4.2 Other optimization algorithms

Another question is the behavior of other heuristic optimization such as simulated annealing or taboo approach. But with regard to the real high throughput experiment (HTE) conditions, we can discard these methods because they are difficult to use. They require synthesizing and testing the catalysts one by one. This does not fit the real experiments scheme.
We nevertheless wanted to evaluate them in our context by fixing the maximum number of evaluation at 400 to put them on the same baseline with the others. It appears that the simulated annealing approach presents a performance of 6.825 and a reliability of 12. It is comparable to the evolutionary approaches. Surprisingly, the taboo algorithm shows very bad results. The performance is 3.6 and the reliability is 7. We think that this is mainly due to the benchmark characteristics. The surface is discontinuous and the taboo search does not allow wide individual moves. So when a solution is located in a poor zone of the search space, the possibilities for reaching the best zones are reduced.

5. Conclusion

We outline in this paper a promising optimization method, says meta modeling, based on evolutionary approach. It is well suited in the combinatorial catalysis optimization because the fitness function is not known. The labelling of the individuals consists on a real chemical reaction that is very costly, in money and in time. Our main idea is a better use of the past experience, the individuals really labelled, with a statistical learning algorithm. This allows us to multiply the individuals during the genetic operators, and select the best ones according to the predicted label from the statistical model before the real chemical reaction.

It seems that when the number of iterations is very small, our approach allows to better guidance of the optimization process.
Our first results are mainly experimental. In a future work, we want to deeply examine the influence of the parameters of the approach: the statistical learning algorithm, the value M of the multiplier, the binary encoding technique. Furthermore, these parameters must be studied according to the characteristics of the chemical reaction and the catalysis components. We hope that when the fitness function is very hard to optimize, and the representation space is large, a better guidance will be more efficient in the context of very small number of iterations.
To conclude this article on an optimistic note, we are currently evaluating this methodology on a real chemical reaction.

References

[1] O.V. Buyevskaya, D. Wolf, and M. Baerns, Catal. Today **62** (1), 91 (2000).
[2] A. Corma, J.M. Serra, and A. Chica, in *Principles and methods for accelerated catalyst design and testing*, edited by E.G. Derouane, V. Parmon, F. Lemos et al. (Kluver Academic Publishers, Dordrecht - NL, 2002), pp. 153; G. Kirsten and W.F. Maier, Appl. Surf.Sci **223** (1-3), 87 (2004).
[3] Michael J. A. Berry and Gordon S. Linoff, *Mastering Data Mining: The Art and Science of Customer Relationship Management*. (Wiley, 2000).
[4] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. (Massachusetts Institute of Technology, 2001).
[5] A. Corma and J.M. Serra, Catal. Today **107-108**, 3 (2005); Y. Yamada, T. Kobayashi, and N. Mizuno, Catalysts & Catalysis **43**, 310 (2001).
[6] A. Corma, J.M. Serra, P. Serna et al., J. Catal. **229** (2), 513 (2005); U. Rodemerck, M. Baerns, M. Holena et al., App. Surf. Sci. **223**, 168 (2003); Y. Jin, Soft Computing **9** (1), 3 (2003).
[7] J. Holland, *Adaptation In Natural and Artificial Systems*. (Ann Arbour, 1975).
[8] K. Sastry, D.E. Goldberg, and M. Pelikan, in *Proceedings of the Genetic and Evolutionary Computation Conference* (2001), pp. 551—558.
[9] K. Rasheed and H. Hirsh, in *Proceedings of the Genetic and Evolutionary Computation Conference* (2000), pp. 628; K.S. Anderson and Y.H. Hsu, in *Congress on Evolutionary Computation*, edited by IEEE (Washington, 1999), pp. 527.
[10] A. Ratle, PPSN **V**, 87 (1998); M. El-Beltagy, P.B. Nair, and A.J. Keane, in *Genetic and Evolutionary Computation Conference* (1999), pp. 196.
[11] L. Baumes, P.E. Jouve, D. Farrusseng et al., in *KES*, edited by V. Palade, R.J. Howlett, and L.C. Jain (Springer, Oxford, 2003), Vol. 2773, pp. 265.
[12] J. Dréo, A. Pétrovski, P. Siarry et al., *Métaheuristiques pour l'optimisation difficile*. (Eyrolles, 2003).
[13] F. Glover and M. Laguna, in *Modern Heuristic Techniques for Combinatorial Problems*, edited by C. Reeves (Blackwell Scientific Publishing, Oxford, England, 1993).
[14] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. (Addison-Wesley, 1989).
[15] H.P. Schwefel, *Evolution and Optimum Seeking*. (Wiley & Sons, New York, 1995).
[16] D. Wolf, O. Buyevskaya, and M. Baerns, Appl. Catal., A, **200(1-2)**, 63 (2000).

Frederic Clerc holds a Ph.D. in computational chemistry from University Lyon 1 and a master in data mining and decision sciences from university Lyon 2. He is currently project leader in Bayesia S.A. His main research interests are data mining, combinatorial optimization and bayesian networks.

David Farrusseng holds a Ph.D. in materials sciences and a master of chemistry from University of Montpellier. He is project leader in the "Institut de Catalyse et d'Environnement de Lyon" (IRCELYON). He is in charge of innovating projects in Combinatorial Sciences in heterogeneous catalysis.

Ricco Rakotomalala is associate professor of computer science at the University Lumière (Lyon 2) since 1998. He is member of the ERIC Laboratory. His main research area is data mining, especially the utilization of the machine learning methods in a real application domain. He designed TANAGRA , a free data mining software which is widely distributed on the web.

Nicolas Nicoloyannis is professor of statistics and computer science at the University Lumière Lyon 2. He manages the ERIC Laboratory and supervises the projects of the laboratory since 2001. His main research area is feature selection, boosting and more generally speaking the statistical learning methods. He supervises several PhD Students at the ERIC Laboratory.

Claude Mirodatos is Research Director (Professor). He manages the Institut de la Recherche sur la Catalyse (IRC) which is one of the most important laboratory in France about the catalysis problem. He participates to several European projects in this domain.