Integrating Chemists Preferences for Shape-Similarity Clustering of Series

Laurent A. Baumes[*,1], Remi Gaudin[2], Pedro Serna[1], Nicolas Nicoloyannis[2] and Avelino Corma[1]

[1]*Instituto de Tecnologia Quimica, UPV-CSIC, Av. de los naranjos, s/n E-46022 Valencia, Spain*

[2]*Lab. ERIC, 5 Av. Pierre mendés France, Univ. Lumière - Lyon 2, 69676 Bron, France*

Abstract: This study shows how chemistry knowledge and reasoning are taken into account for building a new methodology that aims at automatically grouping data having a chronological structure. We consider combinatorial catalytic experiments where the evolution of a reaction (*e.g.*, conversion) over time is expected to be analyzed. The mathematical tool has been developed to compare and group curves taking into account their shape. The strategy, which consists on combining a hierarchical clustering with the k-means algorithm, is described and compared with both algorithms used separately. The hybridization is shown to be of great interest. Then, a second application mode of the proposed methodology is presented. Once meaningful clusters according to chemist's preferences and goals are successfully achieved, the induced model may be used in order to automatically classify new experimental results. The grouping of the new catalysts tested for the Heck coupling reaction between styrene and iodobenzene verified the set of criteria "defined" during the initial clustering step, and facilitated a quick identification of the catalytic behaviors following user's preferences.

Keywords: Combinatorial, high throughput, heterogeneous catalysis, heck, data mining, time series, clustering, hierarchical, *k*-means.

1. INTRODUCTION

The fast automated procedures for discovery of new materials, as well as the development of focused libraries for subsequent lead optimization, has turned into a new paradigm to boost material science [1-3]. However, the use of such strategies for materials and especially heterogeneous catalysts still remains controversial. We are convinced that the interplay of computing strategies from all domains (statistics, artificial intelligence, molecular modeling, design of experiments, data bases, etc.) will become, sooner or later, an integral part of any discovery/optimization program as was observed in the pharmaceutical field thirty years ago. Here it is shown how chemical knowledge and reasoning are taken into account for the creation of a new methodology. A reverse engineering is done starting from the raw data structure, the chemist/user wishes, the study constraints, and the different possible and available methodologies. Bearing in mind what should be a "good" algorithm for the task in hand, existing solutions are hybridized and adapted to give "improved" results in the sense that changes are done to better handle the result expected by the user. Such stepwise work underlines the benefit of an iterative and controlled algorithm conception, *i.e.* a task-guided construction.

In this manuscript, we stress automatic data treatment and, more precisely, clustering and classification, when experimental responses are a chronological series. The exploitation of such information is considered through a real case, where evolution of chemical reactions is obtained *via* high throughput (HT) experimentation. A relatively large number of catalysts has been iteratively synthesized and tested for the Heck coupling reaction. The catalysts composition has been intentionally designed in a very wide space in order to create diversity into output responses and to test the methodology. To the best of our knowledge, the proposed strategy, which is able to handle and group data following a shape-similarity criterion, represents the first work in the material science domain. A hybridization of two classical procedures, namely k-means and hierarchical clustering, is proposed to automatically capture and highlight the characteristics of catalytic curves through cluster information. A comparison with both the algorithms used separately is investigated. Since the HT approach usually simplifies the information of unsteady state experiments by using a discrete number of parameters, the whole behavior of the catalysts through time is forgotten. The hybridization is shown to be of great interest, allowing first the identification of such different behaviors between catalysts, and then the induced model is reused for the automatic classification of new HT experiments. It is shown that the results and the corresponding knowledge gain cannot be obtained using the other traditional clustering strategies.

We first focus on the general problem of methodologies selection. Then, the importance of data treatment aiming at automatically creating "interesting" groups of series is depicted through various examples dealing with material science. The different available solutions are quickly reviewed in section 4, and the new approach is proposed. The elaboration of the algorithm is examined and thoroughly detailed. Finally, the use of the technique for the chosen application in heterogeneous catalysis is presented and discussed.

2. THE SELECTION/CREATION OF AN ADEQUATE STRATEGY

HT experimentation results in large amounts of data to be handled, administered, stored [4-6], and analyzed, but also

*Address correspondence to this author at the Instituto de Tecnologia Quimica, UPV-CSIC, Av. de los naranjos, s/n E-46022 Valencia, Spain;
E-mail: baumesl@itq.upv.es

1386-2073/08 $55.00+.00

hardware/software to be integrated/interconnected and maintained [7, 8]. However, very few new or adapted algorithms have been developed aiming at minimizing the loss of information [9-14]. Although researchers are often able to propose strategies for a particular problem, *i.e.* to select a family of existing algorithms; knowing which one will perform the best remains a challenging task [15]. When particularities regarding a specific study must be considered, the performance of traditional methodologies is usually limited by the possibility of the algorithm to integrate such particular traits through the representation mode that is employed. The proof is given by the no free lunch (NFL) theorem [16]. Algorithms are developed to work in a pre-defined way and give favor to a concrete data representation. Facing a given problem, one should select or modify strategies according to its specific characteristics. However, data particularities may force the researcher to elaborate new techniques. Therefore, the knowledge about the problem may be used/integrated to design the strategy, depending on the specific interests. The new algorithm should improve the results compared to other "impersonal" tools regarding the defined goals or priorities.

3. MONO-CURVE CLUSTERING

'Mono-curves', as the evolution of one variable with regard to another one which is inherently (*i.e.* natural) ordered, differ from punctual results in the concept of the variation (magnitude and direction for example). In data mining research, and more precisely for biological patterns, meteorology, GPS tracking, and video surveillance domains, the treatment of series represents an actual challenge due to the unique structure of such data. This invokes a new field of research called temporal data mining [18], that includes association rules [19], indexing (query by content) [20], feature mining [21] discovery of recurrent or surprising patterns [22], classification [23], and clustering [24-26]. Here, the application of this field to material science is focused on clustering for which different approaches have been proposed such as the use of Hidden Markov Models [25], or the *k*-means algorithm [26]. Clustering can be considered as the most important unsupervised learning problem. It deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be, "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Next, we draw some examples from chemistry domain for which the application of clustering on ordered data structure may result in a significant knowledge gain.

3.1. A Broad Range of Applications for Chemistry

It is possible to find many situations where the shape of series provides information on the problem to be handled. For example, the evolution of species concentration during one reaction may be directly correlated with some aspects of the reaction pathway. Concretely, representations in the form of yield *vs* conversion lead to curves that allow identifying primary stable products, secondary stable products, secondary unstable products, etc., just attending to the shape of such curves. Taking into account that labeling each molecule makes reference to the shape of the curves, independently of the absolute values of conversion and yield, classical clustering algorithms usually fail for grouping curves from the chemistry point of view. Fig. **1** shows an illustrative example of the mentioned situation.

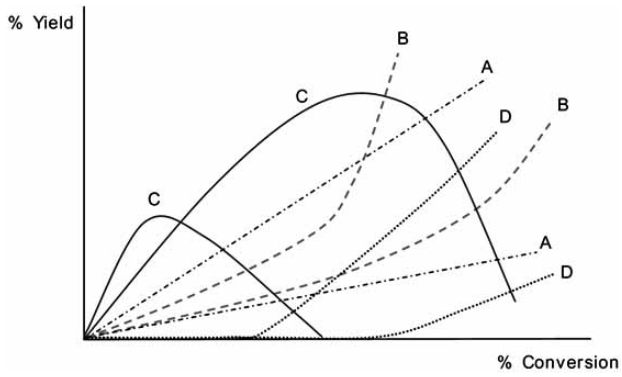


Fig. (1). Examples of *Yield vs Conversion* curves from kinetic experiments. The shape of the curves allows distinguishing primary stable products (**A**), primary and secondary stable products (**B**), primary unstable products, and secondary unstable products (**D**).

Another case of great importance deals with XRD data used for the identification/segregation of crystalline structures. Despite the reported use of traditional clustering algorithms to automatically group samples [3] an adapted methodology is still lacking if one considers that a specific structure can present differences in the XRD diffractogram, regarding both the intensity of peaks and the 2θ diffraction angles, depending on its level of crystallinity and its chemical composition. These limitations must be overcome for facing further challenging applications on XRD data.

More illustrative examples can be found dealing with the adsorption properties of microporous and mesoporous materials. Such materials are widely employed attending to the capacity to retain a gas inside their porous structure. The measurement of the adsorption of one gas regarding its partial pressure, under isothermal conditions, produces the so-called isotherm adsorption curve, which provides information about the material structure and the interaction between such material and one adsorbate. In addition, quantitative information about porous volume and porous size distribution can be inferred. In spite of the fact that HT adsorption equipment is still under development, powerful mathematical tools able to automatically group adsorption curve shapes, whatever the scale of partial pressures and adsorbed volumes, would be crucial for the treatment of massive adsorption data.

Lastly, it would be possible to find a broad variety of examples with regard to many spectroscopic techniques, *i.e.* IR, Raman, XPS, diffuse reflectance, etc., where bands constituting the curves can be related to different characteristics of the samples.

3.2. Study Case: Macroscopic Catalytic Behavior and Evolution of Material Properties

Catalytic results are usually summarized using only one parameter as a reference for the comparison between the behavior of different catalysts or catalytic conditions (turnover number, turnover frequency, or simply the level of ac-

tivity/selectivity at a fixed time). Due to time and cost considerations, catalytic tests are planned in order to obtain only few points of the evolution of the reaction with time. However, it is possible to increase the amount of the experiments using HT technologies: a greater number of assays are carried out in less time. Automated treatment of the data must be developed to avoid the loss of information and the loss of time through manual treatment of data, since otherwise, there is little justification to use expensive HT equipments.

Reaction curves express the variation of the reaction composition with time. Considering a batch reactor, where there is no input/output of material during the reaction, the conversion goes from 0 to 100% (exhaustion of one reactant). The reaction rate, *i.e.* the speed of such evolution, changes as reactants are consumed going asymptotically from a maximum value at the initial reaction rate to 0 at 100% conversion. A monotone increasing curve "conversion *vs* time" is obtained. Under fixed reaction conditions, the dependency between reaction rate and conversion is determined by the action of the catalyst, so that different curves can be generated. The shape of these curves involves aspects such as the kinetic behavior, adsorption processes, resistance to deactivation, or even some activation phenomena. In the present paper, a new algorithm is provided for improving the unreliable results from existing methodologies when shape-similarity of series must be highlighted. Moreover, once meaningful clusters are successfully achieved, the model is coupled with HT equipment in order to automatically classify new experimental results. The separation of the new catalysts verifies the set of criteria "defined" during the initial clustering step allowing a quick identification of the catalytic behaviors according to chemist's preferences and goals.

We have chosen to treat the evolution of the conversion with time for the Heck reaction between styrene and iodobenzene (Fig. **2**) considering a broad variety of catalysts. This reaction consists on a cross-coupling carbon-carbon reaction between an aryl halide or vinyl halide and activated alkenes in the presence of a catalyst and a base. Although some examples have been reported about using Au [27], Rh, Ir, Au, Ni, Co, and Cu catalysts for cross-coupling reactions [28], Pd catalysts have shown the best results. Due to the mechanism of the catalytic cycle, reaction curves present a sigmoidal shape. The induction period observed at the beginning of the reaction is associated with a change in the oxidation state of the Pd sites (Pd^{II} to Pd^{0}) to form the active species [29], but the length of this period depends both on the catalyst (support, Pd content, presence of co-catalysts, etc.) and the reaction conditions (temperature, solvent, concentration and nature of reactants, etc.). A diverse set of kinetic curves has been generated by testing iteratively different catalysts (a total of 105) under fixed reaction conditions. The synthesis of the catalysts was designed following a combinatorial method, covering a wide range of variables, such as type of support, the presence of simultaneous doping metals, or the intensity of the thermal post-treatment, but always keeping constant the percentage of Pd. Both the synthesis of catalysts and their catalytic tests were performed by means of HT equipments (Fig. **3**). The reactions were followed by analyzing aliquots at different times. Thus, each experiment characterizes a given catalyst by means of one curve corresponding to reaction rate *vs* time. Fig. **4** shows the final dataset of curves. The picture clearly shows that making groups or clusters from many curves is not trivial, especially because the induction period shifts them along the time axe. We will show how a new methodology can be useful for improving the automatic clustering/classification of the curves regarding a set of chemical criteria.

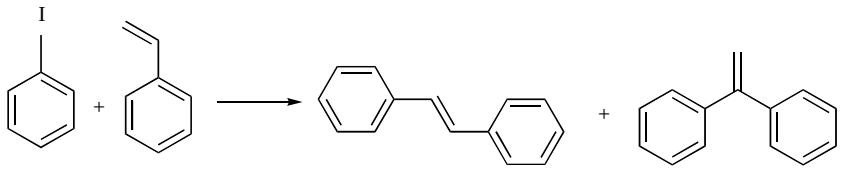


Fig. (2). Heck reaction scheme between styrene and iodobenzene.

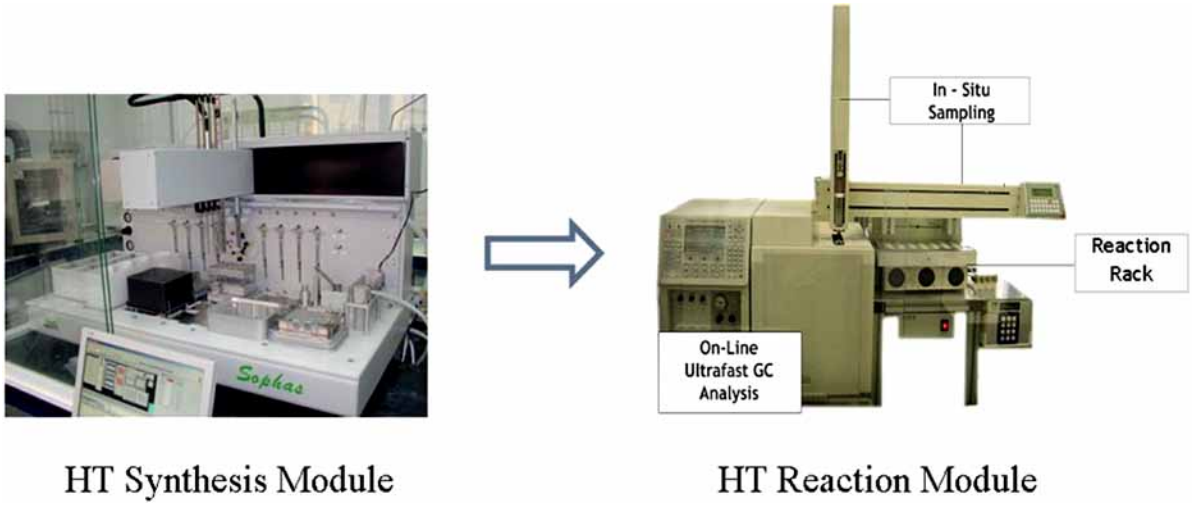


Fig. (3). High throughput modules for the synthesis of catalysts and the reactivity tests.

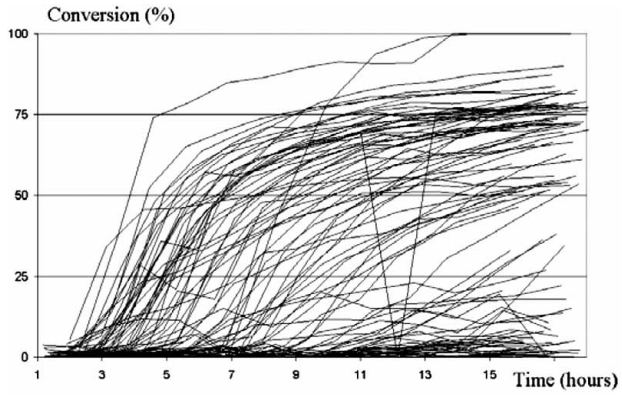


Fig. (4). Conversion *vs* time. 105 different catalysts tested for Heck coupling reaction in HT batch reactors.

4. TOWARD AN ADEQUATE STRATEGY

The creation of an efficient methodology follows a trial and error process starting from the precise definition of the problem and the use of available solutions. Sophistications are then applied through adequate modifications which take into account the interpretation of intermediary expectations, results, and errors. The user is usually concerned by the following algorithm properties: type of attributes being handled, scalability to large dataset, ability to work with high-dimensional data, ability to find cluster of irregular shape, handling outliers, time complexity, data order dependency, strict or fuzzy, reliance on *a priori* knowledge, user defined parameters, and interpretation of results.

4.1. First Examination

4.1.1. Problem Form

To fix the context and to clarify prolific terminology, O is a dataset consisting of curves noted c_i, with $i=1..n$. Each curve is composed of a set of points $\{(x_{i1},t_1),\ldots,(x_{ij},t_{ji})\}$ also noted x_{ij} the j^{th} quantitative value of the observed variable (here the conversion) as shown in Fig. **5** for the two curves c_1 and c_2 (respectively grey, and big black dots). The curves c_i and c_p are respectively composed of ji and jp measures, with ji and jp non necessary equal. Whatever l, the l^{th} values of a given t for two curves, t_{il} and t_{pl} may be different (see c_1 and c_2 measures which are not vertically aligned in Fig. **5**).

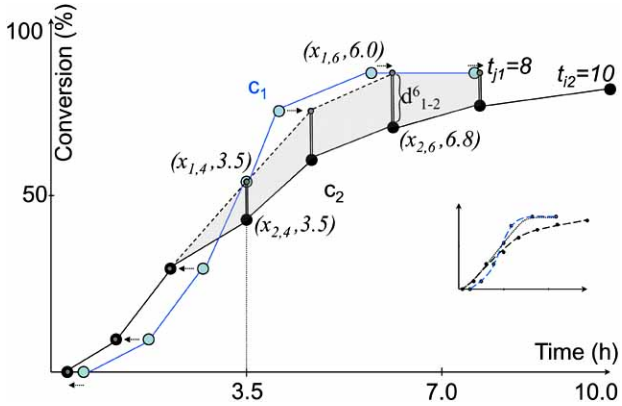


Fig. (5). Examination of the data and notation.

4.1.2. Preliminary Treatments

Data pre-treatment refers to a range of processing steps preceding the detailed analysis. In our case, some transformations of the original data appear necessary due to the following reasons: *i*) presence of outliers, *ii*) all curves do not have an identical length, and *iii*) the measures are not obtained at the same exact moment for all reactors due to the sequential order imposed by the GC analysis.

i) Detection and treatment of outliers. We define as outliers the data points that can be excluded from the curve because they are inconsistent, and will adversely affect the clustering results. Here, a data point is removed as an outlier when it creates either as an increase or a decrease out of a reasonable range. The value is replaced through Hermite curve interpolation [30]. Note that it has been decided not to use smoothing procedures for the whole curve, since the new methodology is expected to be robust facing the inherent experimental noise which differs from the notion of outliers previously defined.

ii) Curve length. If there is a difference of length between curves the most common solution is to rescale the whole set to equal length. However, this pre-process may perturb greatly the clustering as emphasized by Fig. **6**. Another strategy that results in a great loss of information consists on cutting all the series to the minimum period of time. In our implementation, curves are not rescaled and comparisons are restricted onto the shared range of time considering each pair of curves. This allows not hypothesizing on the behavior of unknown regions as it would be using extrapolation methodologies.

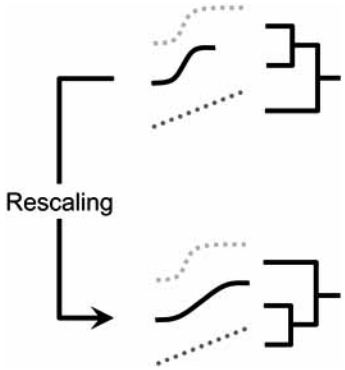


Fig. (6). Rescaling problem. The two time series on the top have very similar shapes, but one of them is early cut. In case of rescaling, the resulting shape is distorted and the corresponding clustering is greatly affected.

iii) Distribution of the points. As noted above, the l^{th} conversion measure does not occur at the same precise moment for all catalytic samples due to technical aspects. By only taking into account the order of the sequence instead of the real time, the shape of the curve may change. This is emphasized in Fig. **5** where real time (big grey dots) curve roughly shows a sigmoid shape, and the corresponding order-based curve (small dots) becomes nearly linear.

The use of the Euclidian distance for sizing how far one curve is from another one requires the points to be re-sampled through approximation (interpolation) in order to obtain vertically aligned points (Fig. **7**). The similarity/dissimilarity criterion that the clustering algorithms use to

define the groups is required to be global for the entire curve, such as the sum of the distances between curves' points. Consequently, the distribution of points influences both criteria. It may be observed that the presence of a relatively larger frequency in a given range of time influences the criterion, such as providing a higher weight to such period. And thus, one could take advantage of the necessity of re-sampling the points for distance calculations. We have chosen a uniform distribution of points not to make the methodology more complicated.

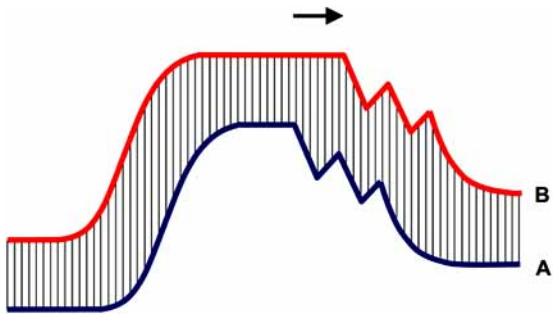


Fig. (7). Euclidian distance.

4.1.3. Expected Goal

An important step we consider belonging to the pre-treatment phase is the preliminary analysis regarding the expected outcome. Considering a set of artificial curves (Fig. **8**), the determination of what would be a good/bad solution allows to quickly establish if a given existing solution provides satisfactory results and, consequently, to point what are the characteristics to be improved. For group 1, *i.e.* G1 in Fig. **8**, the Euclidian-based criterion groups C_A with C_B, and C_C with C_D, since they are the closest to each other (see the areas between curves). From a chemistry point of view, these two clusters make sense with regard to the level of general activity of the catalysts. However, we would like to focus on the importance of the shape of the curves, taking into account that whereas catalysts producing C_A and C_D curves show an important influence of the conversion level on the reaction rate (sharp shape), C_B and C_C present a softer variation at high conversion values. Following such way of thinking c_3 and c_2 in G2 should be grouped together while c_1 would be alone. However, if we consider now G2 to be composed of c and c' curves, it seems possible that a given strategy may form the two following groups $\{c_1, c_{1'}, c_{2'}, c_2\}$ and $\{c_3\}$. Such grouping may be the consequence of the so-called "chaining" effect: because c_1 is very closed to $c_{1'}$, $c_{1'}$ to $c_{2'}$, and $c_{2'}$ to c_2, the c_3 curve appears as the odd element in a separated cluster. Such effect must be discarded due to the non-respect of the shape concept. Finally, G3 is considered with c_a, c_b, and c_c, besides c_3 from G2 (*i.e.* all the non-dotted lines). In this case we want to highlight the effect of the induction period on the clustering analysis, especially because two different expected goals could be chosen. On one hand, we would like $\{c_3, c_c\}$, and $\{c_a,c_b\}$ to be formed, attending to their sharp shape as previously explained. On the other hand, one could be interested on merging curves under a certain limit regarding the induction time, since this parameter provides information about the easiness of the Pd sites activation for each catalyst. Therefore, groups $\{c_3,c_b\}$, and $\{c_a,c_c\}$ could also be expected. Complementarily, large differences on induction period usually produce uncompleted curves, due to temporal limits on the experimental stage, so that the shape of the curves does not appear sufficiently consistent, and the level of uncertainty in the analysis increases. The clustering tool should be flexible enough for considering all the mentioned criteria.

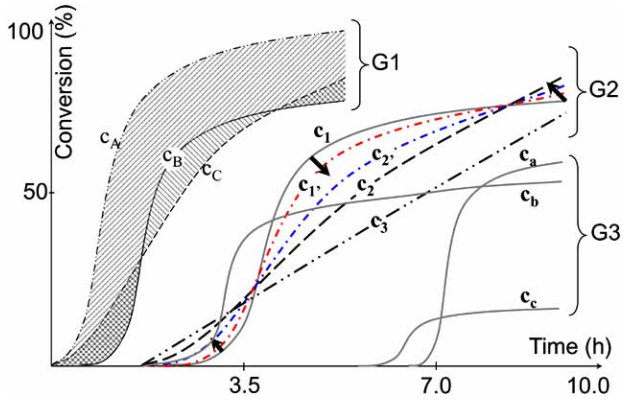


Fig. (8). Algorithm of expected outcomes.

4.2. Examination of Existing Methodologies: Advantages and Drawbacks

Cluster analysis can be divided into two major categories, namely hierarchical and partitional (*i.e.* non-hierarchical) clustering.

4.2.1. Hierarchical Clustering

In hierarchical clustering a series of partitions takes place, and can be subdivided into *agglomerative* methods (CAH), which proceed by series of fusions of the *n* objects into groups, and *divisive* methods (CDH), which separate *n* objects successively into finer groupings. Agglomerative techniques being more commonly employed, such technique is used for further comparisons with our strategy. Note that the proposed approach is divisive due to the hybridization. At each particular stage the method joins together the two clusters which are closest together (most similar).

4.2.2. Partitional Clustering

Another way to perform clustering is the use of a non-hierarchical approach. The number of clusters must be specified *a priori*. The procedure assigns each object to one of the *k* clusters so as to minimize a measure of dispersion within the clusters. Being time consuming, the computation of clusters makes often use of a fast heuristic method that generally produces good (but not necessarily optimal) solutions. The *k*-means [31] algorithm is one of them. *k*-means training starts with a single cluster with a centre as the mean of the data. This cluster is split into two, and the means of the new clusters are iteratively trained minimizing distances inside one cluster while maximizing them between clusters. The process continues until the specified number of clusters is obtained. In order to overcome the problem of starting values, the algorithm generates the *k* clusters' centers randomly, and goes ahead by fitting the data points in those clusters. This process is repeated for as many random starts as the user specified and the best starting value is kept.

The target algorithm must assign each curve into a set of *k* clusters. The use of curves makes the visual determination of the exact amount of clusters difficult (Fig. **4**). Moreover, such a criterion depends on the study, the number of elements to be clustered, the final use of clusters. However the user may have a rough idea or impose some constraints such as a minimum of curves per group. Consequently, the method should be flexible enough for proposing different solutions without a tremendous additional cost. In contrast to hierarchical clustering, *k*-means needs an *a priori* fixed number of clusters *k*. Considering the study, such assumption appear as a serious weakness. The advantage of hierarchical clustering is the flexible setting of the amount of clusters through the visualization of the obtained tree. Therefore the number of clusters *k* is *a posteriori* decided by the user based on the generated nested hierarchy. Taking into account temporal data, another important drawback of *k*-means is the computational cost due to the continuous re-calculation of the *k* clusters centers. On the other hand, the *k*-means approach generates kinds of "spherical" clusters due to the reference to a fictive mean in the implementation. Depending on the context such characteristic might be interesting if relatively homogeneous and proportional clusters are expected. In numerous cases, such clusters are less sensitive to outliers than hierarchical clustering.

4.3. Proposed Strategy

4.3.1. Distances

Whatever the method employed, the use of a similarity or dissimilarity measure is compulsory in order to first size the difference between objects and then proceed to cluster formation. Here, such criterion is defined through distances between objects. The simplest and most employed approach, due to the relative speed of computation to compare temporal sequences, consists in defining a similarity function based on the Euclidian distance. When considering chronological series, the operation consists in matching a given point from a first series with the point from the other one that occurs at the same moment (Fig. **7**). The Euclidean distance between series c_1 and c_2 noted $Eucl(c_1, c_2)$ is defined by $Eucl(c_1, c_2)^2 = \sum_{l=1}^{l=t_{\max}} (x_{1l} - x_{2l})^2$ with $t_{1l} = t_{2l}, \forall l$. However, depending on the context in hand, the use of the Euclidian distance may be unintuitive. It cannot deal with outliers, and is very sensitive to small distortions in the time axis. Therefore the resulting partition might not reflect correctly expected grouping [32]. Another distance, namely the Dynamic Time Warping (DTW), has been considered. The DTW distance allows stretching in time and comparing time-series of different lengths [3] (Fig. **9**), and thus it is merged into our strategy. DTW compares two time series together by allowing a given point to be matched with one or several points. Due to the importance of the meaning of the matching, a *warping window* noted δ is employed. A point that occurs at instant i can be matched with points from the other series that occur in $\left[(i-\delta)..(i+\delta)\right]$. δ is generally set as constant value. When series do not have the same length, the m points of the smaller series are compared with the $m + \delta$ points of the other. DTW is defined following the recursive Equation1 in Fig. **9** and needs a dynamic programming approach [34], see Equation 2 in Fig. **9**. An example of DTW calculation is given in Appendix.

$$DTW(Q,C) = \gamma(m,n) \text{ with } \gamma(i,j)$$
$$= \begin{cases} |q_i - c_j| & \text{if } i = j = 1 \\ |q_i - c_j| + \min\left[\gamma(i-1,j), \gamma(i,j-1), \gamma(i-1,j-1)\right], & \text{else} \end{cases} \quad \text{(Eq. 1)}$$

$$DTW(Q,C) = \begin{cases} \gamma(m,n) \;\; if \; |m-n| \le \delta \\ \gamma(m,m+\delta) \;\; if \; |m-n| > \delta \;\; and \; m < n \\ \gamma(n+\delta,n) \;\; if \; |m-n| > \delta \;\; and \; m > n \end{cases} \quad \text{(Eq. 2)}$$

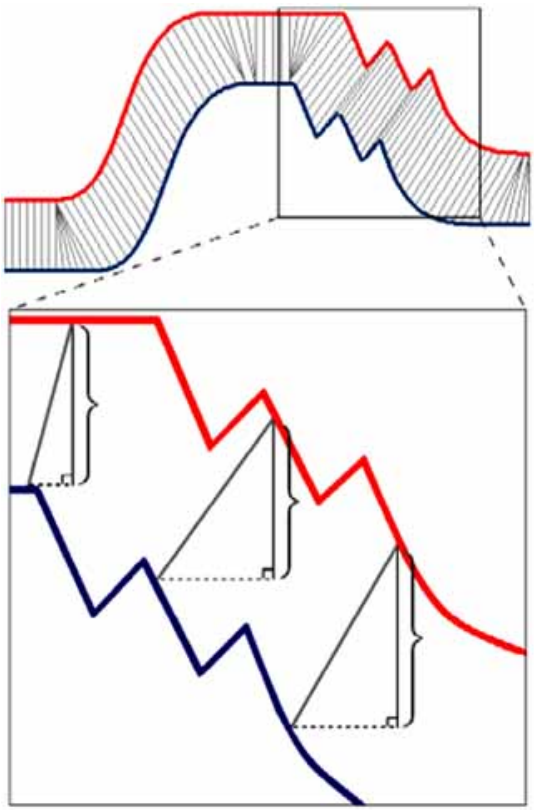


Fig. (9). Dynamic Time Wrapping distances between two series. Lines between the two series give the matching for each point. The distance between two matched points q_i and c_j is equal to $|q_i$-$c_j|$.

Table 1. Diday's *k*-Means Generalization with DTW Distance

Initialization	Define randomly k seeds $p = 0, S^p = \left\{ S_1^p \cdots S_k^p \right\}$
	Calculate the distance noted $L(x_i, S_j)$ between each the time series x_i and each seed S_j. $L(x_i, S_j), \forall i \in \left[1 \cdots n\right], \forall j \in \left[1 \cdots k\right]$ with $L(x_i, S_j) = \frac{1}{c} \sum_{y \in S_j} DTW(x_i, y)$
	Assign to each time series x_i its nearest seed, *i.e.* the seed S_j which minimizes $L(x_i, S_j)$ Calculate the intra-class variance $V^{p=0}(C)$ with $V^p(C) = \frac{1}{n} \sum_{j=1}^{k} \sum_{x_i \in C_j} Inert(x_i, C_j)$
Loop (p=p+1)	Redefine a new set of seeds S_j^p is made of the c time series x_i from C_j which verify: min $Inert(x_i, C_j) = \sum_{y \in C_j} DTW(x_i, y)$
	Assign to each time series x_i its nearest seed, *i.e.* the seed S_j which minimizes $L(x_i, S_j)$
	Calculate the intra-class variance $V^p(C)$
	If $V^p(C) - V^{p-1}(C) \leq \varepsilon$ then STOP

A set of n time series $X = \{x_1, \ldots, x_n\}$ must be split into k clusters according to a given similarity criterion. A so-called seed S_j is composed of c time series among the initial set X. One chronological series does not belong to more than one seed. The number of clusters is noted k, and each C_j cluster among $C = \{C_1, \ldots, C_k\}$ is composed of the time series for which S_j is the nearest seed.

4.3.2. Correction of k-Means Drawbacks

Our strategy which is based on the generalization of the Forgy algorithm [35, 36] permits to handle the following difficulties for the special case of time series: poor initial partitions and high computation cost of the mean curve for each cluster. Each cluster is characterized by a seed of c time series. Clusters are created based on the two following parameters: the number of clusters and the size of seeds, respectively noted c and k (Table **1**). Seeds are used to compute distances between time series and clusters as well as distances between each cluster. The c times series of a cluster are those that minimize the *Inertia* function. This function is also used to compute the intra-class variance between all clusters that evaluates the quality of the clustering. We proceed as indicated in Table **2**.

Table 2. The Optimal *c* Value

1	Given a maximal size of seeds *cmax* (*cmax* ≤ *RoundInf*(n/k)).
2	For i = 1 to *cmax*, do : Execute the algorithm with $c = i$. Save the intra-class variance final value.
3	The optimal value of c is the one that obtains the minimal variance.
4	Repeat step 2 and 3 until the decrease of the minimal variance is null or below a given threshold.

4.3.3. DTW Hierarchical k-Means

The process of hierarchical *k*-means consists in iteratively applying the *k*-means algorithm which produces a hierarchical tree. Such hybridization is of great interest because the user can stop the clustering independently in each sub tree, and can determine the best k value *a posteriori* according to the evolution of the clustering. Moreover, it allows modifying the δ value for each node before each new sub clustering and consequently may improve the quality of the result. Forgy algorithm allowing splits into n groups, arbitrary architecture of trees (*i.e.* not only binary trees) could be obtained. However, for the presentation of the algorithm n is restricted to 2. Hierarchical *k*-means algorithm got two parameters: *cmax* and δ. *cmax* is always set to 3 since its influence on the final clustering decreases tremendously when its value is superior or equal to such value. On the other hand, the setting of δ is very important as it is used to direct and refine the clustering toward user wishes. The hierarchical *k*-means algorithm preserves advantages of hierarchical clustering and classic *k*-means approaches (hierarchical tree and proportional clustering) without keeping their drawbacks (clustering disturbing by atypical curves and k parameter). A summary of algorithm properties is given in Table **3**.

5. RESULTS

In the case of Heck reaction catalyzed by solid catalysts, it is continuously under debate if the reaction takes place on

Table 3. Clustering Properties of DTW Hierarchical *k*-Means

Type of Attributes Being Handled	Quantitative
Scalability to large dataset	*Improved for time series through Diday's generalization adaptation*
Ability to work with high-dimensional data	*Mono-time series*
Ability to find cluster of irregular shape	*No (Not desired)*
Handling outliers	*Robust*
Time complexity	*Reasonable*
Data order dependency	*Yes (i.e. curves)*
Strict or fuzzy	*Strict*
Reliance on *a priori* knowledge	*Seed search, user can split and merge, delta is variable*
User defined parameters	*Integrated*
Interpretation of results	*Easy*

the surface of the catalyst, or by the contrary leached Pd species are the ones catalytically active, or the possibility that both the solid and the leached species contribute to the observed activity [29]. Furthermore, it is also claimed that in many cases the catalytic active species are generated during the exposure of the catalyst to the reactants and consequently an induction period in the kinetic process may occur. Because the shape of the curves is related to different aspects, such as the induction period length or the sharpness of the conversion evolution, clustering analysis becomes a difficult task, and classical algorithms hardly satisfy simultaneous user interests.

We show the different results obtained for CAH (*i.e.* Hierarchical and Agglomerative Clustering), *k*-means, and our strategy. Initially, the use of CAH and k-means is shown on the whole set of curves in order to better size the difference between the existing algorithms and the proposed one. The dataset is composed of 5 libraries of catalysts, 105 in total (5×21). Secondly, it is shown how the model is parameterized using only the first two generations (*i.e.* clustering), and used as a classification tool on the following libraries.

5.1. Clustering

Firstly, CAH methodology is used. Fig. **10** shows groups created with the Euclidean distance criteria for a different number of clusters. The clustering is largely imposed by the influence of the induction period, and the greatest differences are found between the zero conversion values (during the induction time) and non-zero values. For example, a group such as D (5 clusters) is divided into D and E (6 clusters) because even though they keep a similar shape, the former present lower induction time values. On the other hand, this methodology clearly suffer from the so-called *chaining effect*, which leads to clusters composed by just one curve, whereas curves apparently very dissimilar are included into the same cluster. Finally, after six clusters, it is still impossible to separate non active catalysts from active ones with a large induction period. Fig. **11** shows the final six clusters when Euclidean distance is substituted by DTW ($\delta = 10$). Thanks to the corrections that DTW introduces on the time axis, it can be seen that curves with a broad range of induction time are now found into the same clusters, being the resemblance between curves more focused on the shape. In spite of the fact that active and non active samples are distinguished, single curves are still composing clusters.

Next, the *k*-means algorithm is applied. Fig. **12** shows the results for one to six clusters when Euclidian distance is used. Comparing with the CAH methodology, both algorithms work under fixed conditions along the clusters formation (from initial parameters defined by the user). This is illustrated in Fig. **12**. The pyramidal structure is useful to check how the *k*-means algorithm (with Euclidean distances) is making narrower the differences on the induction period if more clusters are proposed. First division practically corresponds to curves with non null and nearly null conversion values before 8 hours. The result is logical considering that Euclidean distance do not introduce any correction on the data with regard to induction time. Increasing the number of clusters to 3 allows providing a narrower range of induction time, so that one group with nearly null conversion values until 11 hours, other with an induction time between 2 and 5 hours, and another one between 5 and 10 hours are obtained. However, one can observe that curves with different shapes are present into the same cluster. It can be checked that increasing the number of clusters produces new splits following the same trend, so that no other aspects can be highlighted.

As the main advantage, *k*-means offers homogeneous groups avoiding the existence of clusters composed by very few curves. However, the CAH tree-based architecture is lost. Finally, the proposed strategy "hierarchical *k*-means" is employed. This strategy allows combining the benefits of hierarchical strategies and *k*-means. On one hand, the algorithm is built under a tree-based architecture, and the user can select which cluster is split while controlling DTW parameters for each partition, so that it is possible to adapt the grouping criteria as the tree is being developed. Such flexibility is greatly important for rationalizing the clustering process under a semi-supervised methodology. It enables defining clustering parameters without necessarily expressing them mathematically. DTW merging would make such formalization of the set of criterion impossible. On the other hand, the *k*-means approach provides suitable clusters regarding the homogeneity.

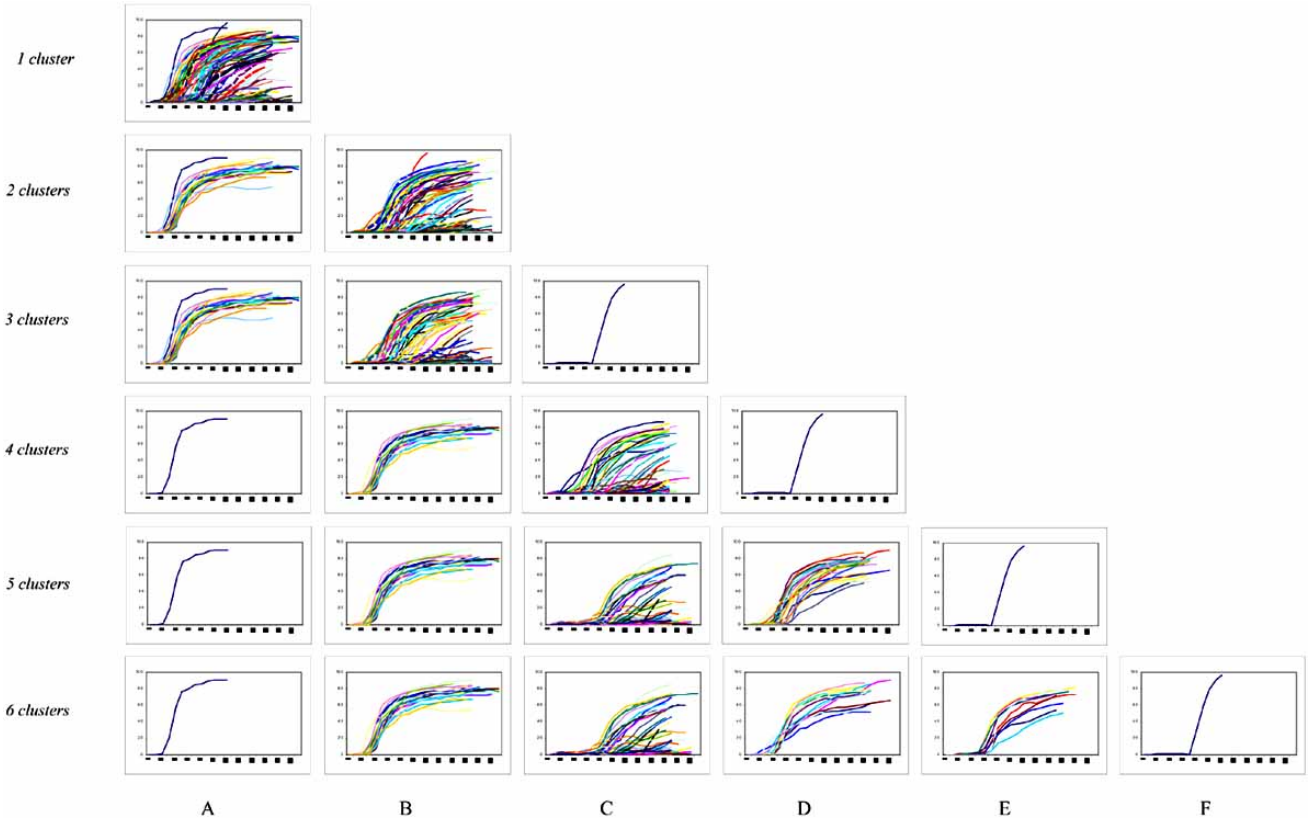


Fig. (10). CAH is used for automatically creating 1 to 6 clusters. The general criterion is the mean Euclidian distance.

Because different criteria can be selected to classify the curves, the hierarchy on the tree evolution can change from one users to others. For example, Fig. **13** shows a tree where the first partition if focused on splitting active and non active catalyst, which can be achieved minimizing the shift of the curves along the axis time ($\delta = 10$). Next, new groups from the active samples branch may be proposed attending to other criteria. Concretely, using a $\delta = 2$ value it is possible to separate sharpener curves from more linear shapes, obtaining at this level three perfectly defined clusters: non active, active with slow reaction rate at high conversion values, and active with little variation of reaction rate with the conversion. Finally, each group can be refined, *i.e.* different levels of deactivation (A and B clusters), different induction time values (C and D clusters), and different levels of activity (E and F clusters). Fig. **14** shows another example about the hierarchical *k*-means algorithm application. In this case, the first two groups have been created by using a low δ value, and thus the induction period plays the principal role on the division. New partitions on the clusters, using concrete δ values, allow distinguishing different levels of activity, deactivation or induction time length. Comparing with the result for final clusters on Fig. **13**, it can be observed that cluster B appears now refined on the induction time (B' and C' clus-

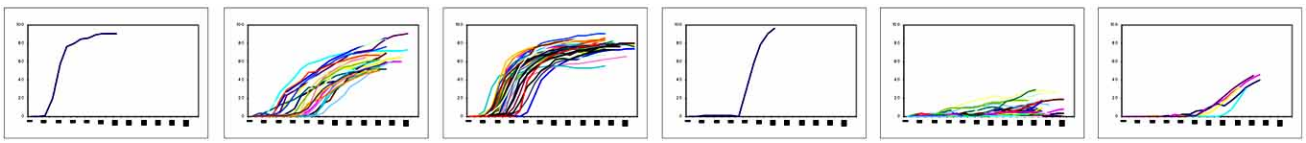


Fig. (11). CAH is used for automatically creating 1 to 6 clusters. The general criterion is the mean DTW distance, delta is fixed to 10.

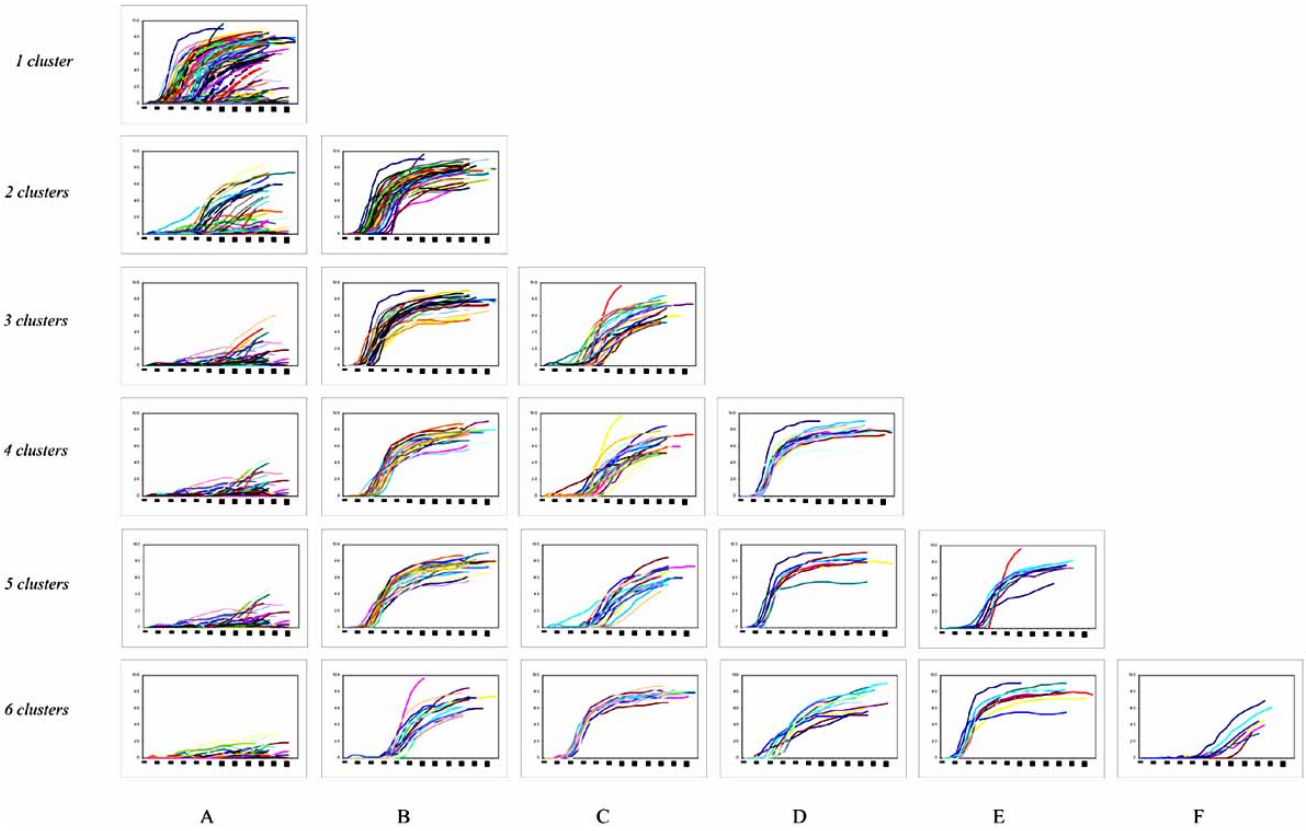


Fig. (12). *k*-means is used for automatically creating 1 to 6 clusters. The general criterion is the mean Euclidian distance.

ters), while no distinctions have been considered into the non active catalysts (F' cluster).

5.2. Clustering to Classification

The aim of data-mining treatment for combinatorial and HT studies is to make easier the analysis of large amounts of results in order to reduce time/cost and to acquire knowledge. In this sense, we show that the designed methodology, besides a robust but flexible behaviour regarding the clustering of series, can also be employed as a classification tool for new/unseen experiments. As illustration, we have organized the entire data set (reaction curves from the 105 synthesized catalysts) in 5 subsets of 21 catalysts, considering that this is the number of reactions we can simultaneously perform in the multibatch system. The final goal of using the proposed approach consists on training the algorithm with information coming from the first two sets of experiments (generations 1 and 2), and then to rapidly identify the behaviour of the following catalysts (generations 3 to 5). Therefore, the process for grouping all the data is organized in two successive steps: 1) identification of *prototype* groups through the previously exposed clustering approach, in which chemistry preferences are considered to establish the shape-similarities criteria; 2) automatic classification of new curves with the model induced by 1) allowing the retrieve of user preferences. Catalysts tested in the two firsts cycles of the reactor (42 samples) were randomly selected among the synthesized ones. The hierarchical *k*-means clustering on these samples shows, in essence, that same groups of curves can be identified in comparison with those generated from the whole population (Figs. **15**,**16**). Consequently, keeping an identical tree architecture during the training stage (same δ values for each partition) produces analogue final clusters. As classification of new cases is performed when using traditional induction trees or graphs, the new curves are dropped into the tree root and follow model rules, *i.e.* delta values. In our case, we use the delta value in order to define in which of the two following leafs the curve will be. Each one is associated into the group (*i.e.* leaf) which contains the curve that is the nearest to the current test curve using the previously defined delta value. Such calculation is successively applied until reaching

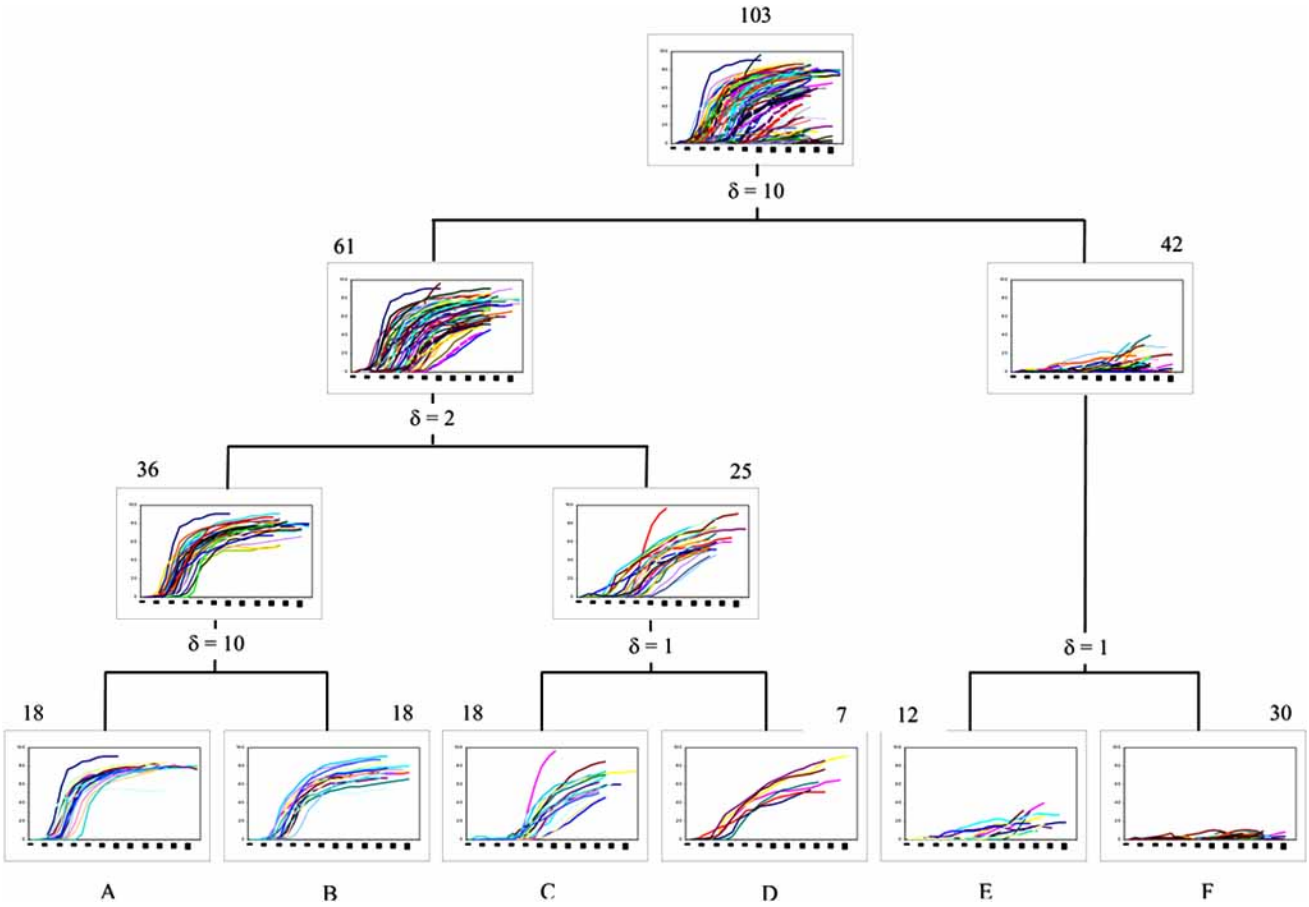


Fig. (13). Hierarchical *k*-means is used for automatically creating 6 clusters. The general criterion is the mean DTW (Version 1).

one final leaf. All new samples, coming from posterior cycles of reaction, can be automatically classified into the correct groups (Figs. **13**,**14** are obtained). This means that the shape-similarity preferences integrated through the tree conception is conserved/maintained. The flexibility of the methodology represents a key point in the selection of different criteria for grouping the curves, while a robust behaviour is shown when previously unseen data must be treated. On the contrary, other clustering algorithms suffer from important limitations to work with such data structure and their use for automatically classifying curves is clearly unsatisfactory. In the end, a powerful clustering/classification tool has been developed for the rapid treatment of large amount of data in the form of series. The hierarchical *k*-means option represents the most suitable alternative for the integration of a grouping strategy into a global HT scheme.

6. DISCUSSION

The main goal of clustering techniques is to determine "intrinsic" groupings in a set of unlabeled data. There is no absolute "best" criterion which would be independent of the final aim of the clustering. For instance, one could be interested in finding representatives for homogeneous groups (*data reduction*), in finding "natural clusters" and describe their unknown properties (*"natural" data types*), in finding useful and suitable groupings (*"useful" data classes*) or in finding unusual data objects (*outlier detection*). The proposed methodology allows the user to direct/supervise the clustering while taking into account various criteria. Heck reaction curves are characterized by a particular sigmoidal shape. According to this, our method may consider, in a single clustering, the three following principal aspects for defining the behavior of the catalysts: *i*) the length of the induction time, which delays the starting point of the catalytic cycle along the time axe, *ii*) the general level of conversion values once the catalytic cycle has started, *iii*) the sharpness of conversion evolution.

The rigidity around clusters formation when traditional algorithms are employed strongly limits the incorporation of knowledge during the process. The use of a DTW hierarchical *k*-means represents a semi-supervised way of clustering analysis. This methodology shows numerous advantages. Firstly, the hierarchical architecture allows obtaining a tree structure. The user directly decides which branches of the tree must be more developed, with regard to practical criteria. Secondly, δ can independently varies into the different

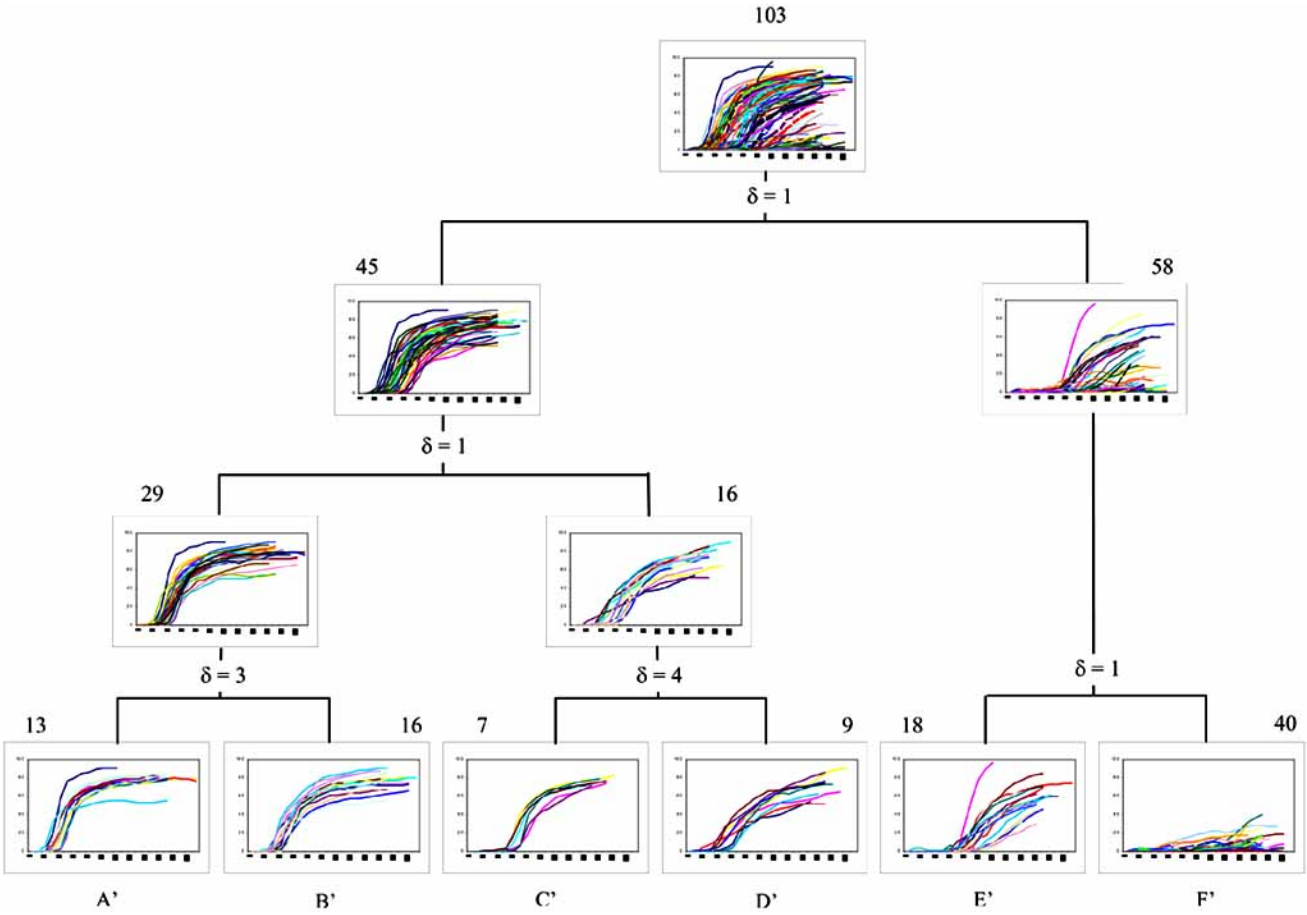


Fig. (14). Hierarchical *k*-means is used for automatically creating 6 clusters. The general criterion is the mean DTW (Version 2).

divisions/splits, in such a way that one may give a logical order to the clustering sequence. Thirdly, the user can propose different trees, depending on which aspects are considered. Two representative examples have been shown for two opposite situation with regard to the whole data set division.

Considering the tuning of the proposed clustering approach into a classification tool, perfect results have been obtained since the labelling of all curves is identical either for the clustering methodology on the entire dataset, or for the two-step strategy: clustering with a lower amount of data, and then using the induced model (delta parameters) on unseen cases (1-nearest neighbour classification). However, considering a machine learning approach, the role of the sampling is crucial. When employing such strategies, the user must always assume that the training set is a representative subset of the search space. For our application, it appears logical, due to the use of a similarity criterion, that trying to set the model (*i.e.* clustering step) with a dataset that do not contains any curve of a given group will fail classifying new experiments expected to belong to such group. Then, the user plays a central role checking if all the search space and expected groups are covered by examples. The integration of the chemist (knowledge) into such methodology is of great interest since it would be possible to create artificial curves (unsynthesized materials) for defining groups that are not represented by the training set.

CONCLUSIONS

We have proposed using shape-similarities analogies between data in the form of series to overcome the typical loss of information that certain high throughput applications suffer from. Concretely, we reviewed the way that combinatorial and data-mining methodologies make use of the information recovered from catalytic experiments (reaction curves), showing that there was a lack of a consistent methodology to manage the clustering/classification of such data structure. Then, a specific data treatment is proposed in order to handle the present problematic, demonstrating that it is

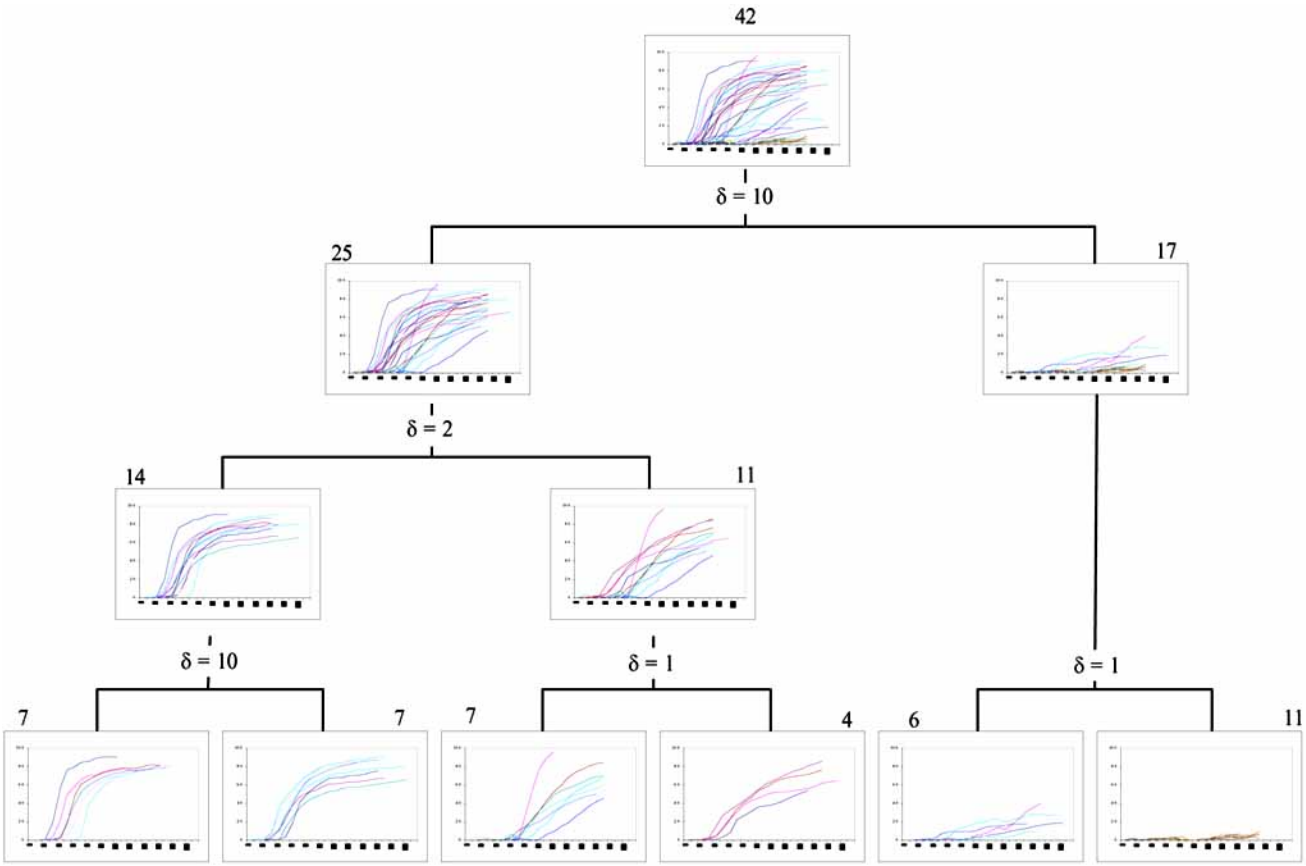


Fig. (15). Generation of clusters for a subset of only 42 curves from the whole data set (Version 1).

possible to adapt the data mining process to the user requirements and criteria. It is shown that chemistry preferences can be integrated in the clustering conception. The flexibility of the algorithm allows looking for groups under a semi-supervised approach, while the criteria can be kept for automatically classifying new data in an unsupervised form. The applicability of the method has been tested on a real case and compared against other common algorithms. It is shown that such a strategy is of great interest considering the special case of series. The possibility of using the proposed approach as a classification routine opens the way to the application of such algorithm into an iterative HT loop. Further experiments are currently done, and correlation between clusters formation and synthesis variables is investigated.

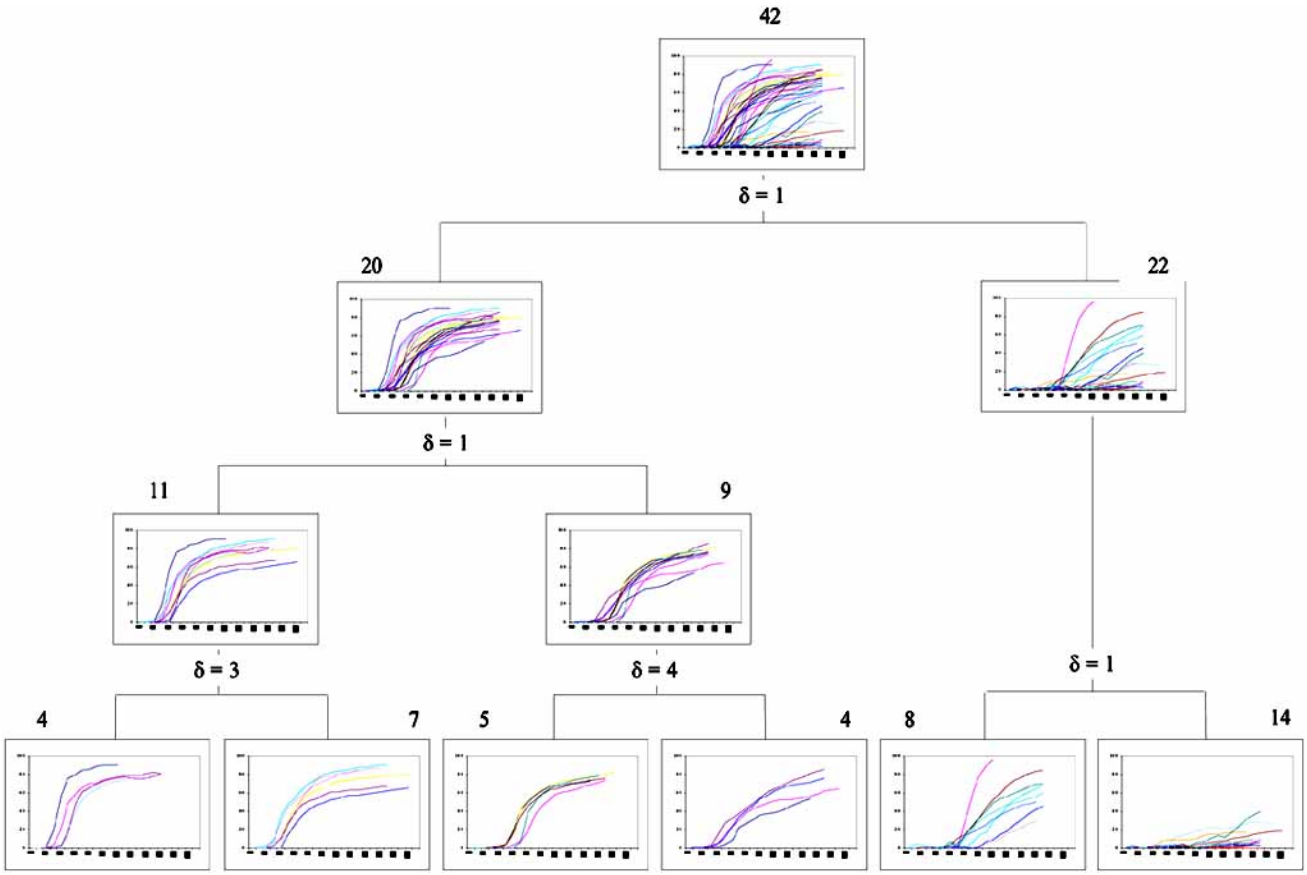


Fig. (16). Generation of clusters for a subset of only 42 curves from the whole data set (Version 2).

Appendix: DTW Calculation

Let us consider two curves Q and C of length m and n, here both equal to 8.

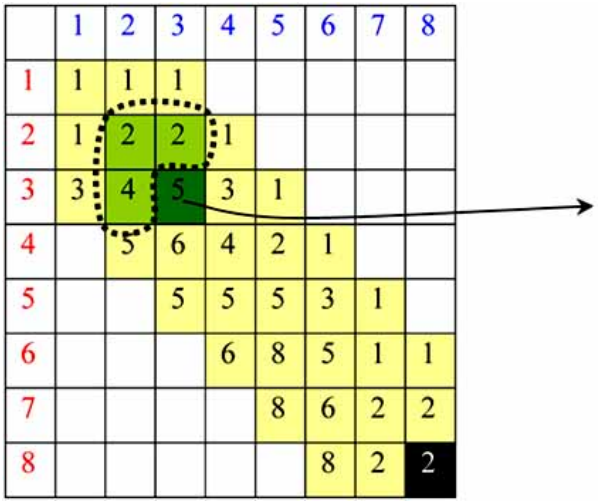
	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
Q	1	2	4	3	1	1	2	1
C	2	1	1	2	4	3	1	1

Inside a $m \times n$ matrix X, the value X(i,j) is stored in the cell (i, j).

$$X(i,j) = \begin{cases} |q_i - c_j| & \text{if } i = j = 1 \\ |q_i - c_j| + \min\left[X(i-1,j), X(i,j-1), X(i-1,j-1)\right], & \text{else} \end{cases}$$

The value contained in X(m,n) is equal to the DTW distance between Q and C, *i.e.* $DTW(Q,C) = X(m,n)$

With Delta = 2,



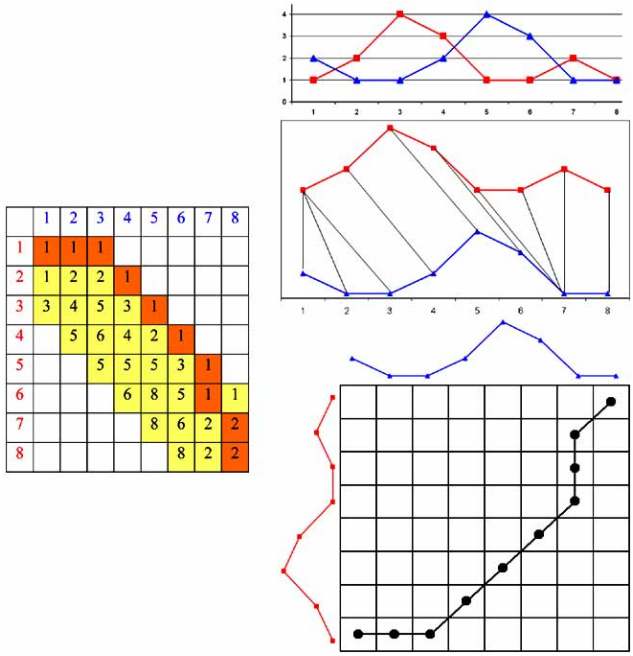
=X(3,3)

$=|Q_3-C_3|+ \text{Min}\{ X(2,3) ; X(3,2) ; X(2,2)\}$

=|4 -1|+ Min{2 ; 4 ; 2}

=3+ 2 = 5

The path beginning in position (1, 1) and finishing in (m, n) that minimizes the sum of the cells it goes through, indicates which points of Q and C are matched. Note that if the path goes through X(i,j), then the preceding cell is either X(i-1,j), X(i,j-1), or X(i-1,j-1). If there is a warping window delta, the warping path can goes through X(i,j) only if |i - j| ≤ delta. Such path is drawn with dark cell in the figure below.



ACKNOWLEDGEMENTS

Financial support from Spanish government (Project MAT 2003-07945-C02-01, and grants FPU AP2003-4635) and EU Commission (TOPCOMBI Project) is gratefully acknowledged. We thank Santiago Jiménez Serrano for the help concerning the construction of the ITQ platform called hIT$_e$Q we used for data treatment.

REFERENCES

[1] (a) Jandeleit, B.; Schaefer, D.J.; Powers, T.S.; Turner, H.W.; Weinberg, W.H. *Angew. Chem. Int. Ed.*, **1999**, *38*, 2494. (b) Senkan, S.M. *Angew. Chem. Int. Ed.*, **2001**, *40*, 312. (c) Reetz, M.T. *Angew. Chem. Int. Ed.*, **2001**, *40*, 284. (d) Newsam, J.M.; Schuth, F. *Biotechnol. Bioeng.*, **1999**, *61*, 203-216. (e) Gennari, F.; Seneci, P.; Miertus, S. *Catal. Rev.-Sci. Eng.*, **2000**, *42*, 385.

[2] Harmon, L.A.; Vayda, A.J.; Schlosser, S.G. *Abstr. Pap. - Am. Chem. Soc.*, **2001**, *221*, BTEC-067.

[3] (a) Corma, A.; Moliner, M.; Serra, J.M.; Serna, P.; Díaz-Cabañas, M.J.; Baumes, L.A. *Chem. Mater.*, **2006**, *18*, 3287. (b) Rajagopalan, A.; Suh, C.; Li, X.; Rajan, K. *Appl. Catal. A: General*, **2003**, *254*, 147. (c) Corma, A.; Díaz-Cabañas, M.J.; Moliner, M.; Martínez, C. *J. Catal.*, **2006**, *241*, 312.

[4] (a) Baumes, L.A. Combinatorial Stochastic Iterative Algorithms and High Throughput Approach: from Discovery to Optimization of Heterogeneous Catalysts (in English). Univ. Claude Bernard Lyon 1, Lyon, France, **2004**. (b) Farrusseng, D.; Baumes, L.A.; Mirodatos, C. In *High Throughput Analysis: A Tool For Combinatorial Materials Science*, Potyrailo, R.A.; Amis, E.J.; Eds. Kluwer Academic/Plenum Publishers: **2003**; pp. 551-579. (c) D. Farrusseng, L.A.; Baumes, C.; Hayaud, I.; Vauthey, P.; Denton, C. Mirodatos. Kluver Academic Publisher, NATO series, edited by E. Derouane. Proc. NATO Advanced Study Institute on Principles and Methods for Accelerated Catalyst Design, Preparation, Testing and Development, Vilamoura, Portugal, 15-28 July **2001**. eds. E. Derouane, V. Parmon, F Lemos, F. Ribeiro. Book Series: NATO SCIENCE SERIES: II: Mathematics, Physics and Chemistry. 69, 101-124, Kluwer Academic Publishers, Dordrecht. Hardbound, ISBN 1-4020-0720-5. July **2002** (d) http://www.fist.fr/article259.html website accessed the 10th March 2007.

[5] Adams, N.; Schubert, U.S. *Macromol. Rapid. Commun.*, **2005**, *25*, 48.

[6] (a) Scheidtmann, J.; Frantzen, A.; Frenzer, G.; Maier, W.F. *Meas. Sci. Technol.*, **2005**, *6*, 119. (b) Frantzen, A.; Sanders, D.; Scheidtmann, J.; Simon, U.; Maier, W.F. *QSAR Comb. Sci.*, **2005**, *24*, 22.

[7] (a) Adams, N.; Schubert, U.S. *QSAR Comb. Sci.*, **2005**, *24*, 58. (b) Ohrenberg, A.; von Torne, C.; Schuppet, A.; Knab, B. *QSAR Comb. Sci.*, **2005**, *24*, 29. (c) Saupe, M.; Fodisch, R.; Sunderrmann, A.; Schunk, S.A.; Finger, K.E. *QSAR Comb. Sci.*, **2005**, *24*, 66.

[8] Gilardoni, F.; Curcin, V.; Karunanayake, K.; Norgaard, J.; Guo, Y. *QSAR Comb. Sci.*, **2005**, *24*, 120.

[9] Nicolaides, D. *QSAR Comb. Sci.*, **2005**, *24*, 15.

[10] (a) Klanner, C.; Farrusseng, D.; Baumes, L.A.; Mirodatos, C.; Schuth, F. *QSAR Comb. Sci.*, **2003**, *22*, 729. (b) Klanner, C.; Farrusseng, D.; Baumes, L.A.; Mirodatos, C.; Schuth, F. *Angew. Chem. Int. Ed.*, **2004**, *43*, 5347. (c) Farrusseng, D.; Klanner, C., Baumes, L.A.; Lengliz, M.; Mirodatos, C.; Schüth, F. *QSAR Comb. Sci.*, **2005**, *24*, 78. (d) Schüth F.; Baumes, L.A.; Clerc, F.; Demuth, D.; Farrusseng, D.; Llamas-Galilea, J.; Klanner, C.; Klein, J.; Martinez-Joaristi, A.; Procelewska, J.; Saupe, M.; Schunk, S.; Schwickardi, M.; Strehlau, W.; Zech, T. *Catal. Today*, **2006**, *117*, 284. (e) Baumes L.A.; Farruseng D.; Lengliz, M.; Mirodatos, C. *QSAR Comb. Sci.*, **2004**, *29*, 767.

[11] (a) Baumes, L.A.; Jouve, P.; Farrusseng, D.; Lengliz, M.; Nicoloyannis, N.; Mirodatos, C. In *7th Int. Conf. on Knowledge-Based Intelligent Information & Engineering Systems* (KES'2003). Univ. of Oxford, UK. Springer-Verlag in Lecture Notes in AI (LNCS/LNAI series). Eds. Palade, V.; Howlett R.J.; Jain, L.C. Sept. 3rd-5th, **2003**. (b) Serra, J.M.; Corma, A.; Farrusseng, D.; Baumes, L.A.; Mirodatos, C.; Flego, C.; Perego, C. *Catal. Today*, **2003**, *82*, 67.

[12] *Experimental design for combinatorial and high throughput materials development*, Eds. Cawse, J.N. Wiley and Sons: Hoboken, NJ. **2003**.

[13] (a) Baumes, L.A. *J. Comb. Chem.* **2006**, *8*, 304. (b) Baumes, L.A.; Serra, J.M.; Serna, P.; Corma. A. *J. Comb. Chem.*, **2006**, *8*, 583. (c) Serra, J.M.; Baumes, L.A.; Moliner, M.; Serna, P.; Corma. A. *Comb. Chem. High Throughput Screen.*, **2007**, *10*(1), 13.

[14] Corma, A.; Serra, J.M.; Serna, P.; Valero, S.; Argente, E.; Botti, V. *J. Catal.*, **2005**, *229*, 513.

[15] Baumes, L.A.; Moliner, M.; Corma, A. *QSAR Comb. Sci.*, **2007**, *26*, 255.

[16] (a) Wolpert, D.H.; Macready, W.G. *IEEE Trans. Evol. Computat.*, **1997**, *1*, 67. (b) Wolpert, D.H. *Complex Sys.*, **1992**, *6*, 47.

[17] (a) Baumes, L.A.; Serra, J.M.; Serna, P.; Corma. A. *J. Comb. Chem.*, **2006**, *8*, 583. (b) Serra, J.M.; Baumes, L.A.; Moliner, M.; Serna, P.; Corma, A. *Comb. Chem. High Throughput Screen.*, **2007**, *10*, 13.

[18] (a) Antunes, C.M.; Oliveira, A.L. In *Proc. of the Workshop on Temporal Data Mining, at the 7th Int. Conf. on Knowledge Discovery and Data Mining (KDD'01)*, San Francisco, CA, 1-15, **2001**. (b) Lin, W.; Orgun, M.A.; Williams, G.J. Macquarie University and CSIRO Data Mining. The Australasian Data Mining Workshop. **2002**.

[19] (a) Das, G.; Gunopulos, D.; Mannila, H. In *Proc. of Principles of Data Mining and Knowledge Discovery, 1st European Symposium*. Trondheim, Norway, 88-10. **1997**. (b) Das, G.; Lin, K.; Mannila, H.; Renganathan, G.; Smyth, P. In *Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining*. New York, Aug. 27-31, 16-22, **1998**.

[20] (a) Keogh, E.; Ratanamahatana, C.A. In *Knowledge and Information Systems* (KAIS04). May **2004**. (b) Yi, B. K.; Jagadish, H.; Faloutsos, C., In *IEEE Int. Conf. Data Engineer*. **1998**, 201-208.

[21] (a) Kudenko, D.; Hirsh, H. In *Proc. of the 15th National Conf. Artificial Intelligence (AAAI'98)*, Menlo Park, California, 733-739, **1998**. (b) Lesh, N.; Zaki, M.J.; Ogihara, M. *IEEE Intell. Sys.*, **2000**, *15*, 48.

[22] (a) Buhler, J.; Tompa, M. *J. Comp. Biol.*, **2002**, *9*, 225. (b) Chiu, B.; Keogh, E.; Lonardi, S., In *9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. Aug. 24-27. Washington-DC, USA, 493-498, **2003**. (c) Keogh, E.; Lonardi, S.; Chiu, W. In *8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Edmonton*, Alberta, Canada, 550-556, **2002**. (d) Lin, J.; Keogh, E.; Patel, P.; Lonardi, S. In *Proc. of the 2nd Workshop on Temporal Data Mining, at the 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD'02)*, Edmonton, Alberta, Canada, 53-68, **2002**.

[23] (a) Chappelier, J.C.; Gori, M.; Grumbach, A., In R. Sun and G.L. Giles, editors. *Sequence Learning: Paradigms, Algorithms and Applications*, 105-134. Springer-Verlag, **2000**. (b) James, D.L.; Miikkulainen, R. *Adv. Neural Process. Systems*, **1995**, 7: 577-584. (c) Keogh, E.; Pazzani, M. In *Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining*. New York. 239-241. **1998**. (d) Somervuo, P.; Kohonen, T. *Neural Process. Lett.*, **1999**, *10*, 151.

[24] Gaudin, R.; Nicoloyannis, N. In *Proc. of the 5th Journées d'Extraction et de Gestion des Connaissances (EGC'05)*, Paris, France, 201-212, **2005**.

[25] (a) Lin, F.R.; Hsieh, L.S.; Pan, S.M. In *Proc. of the 38th Annual Hawaii Int. Conf. on System Sciences (HICSS'05)*, **2005**. (b) Oates, T.; Firoiu, L.; Cohen, P.R., In IJCAI-99 Workshop on Sequence Learning, 17-21, **1999**.

[26] (a) Lin, J.; Vlachos, M.; Keogh, E.; Gunopulos, D. In *Proc. of the 9th Conf. on Extending Database Technology (EDBT 2004)*. Crete, Greece, **2004**. (b) Vlachos, M.; Lin, J.; Keogh, E.; Gunopulos, D. Workshop on Clustering High-Dimensionality Data and its Applications, SIAM Datamining, San Francisco, **2003**.

[27] (a) Arellano, G.; Corma, A.; Iglesias, M.; Sanchez, F. *J. Catal.*, **2006**, *238*, 497. (b) Carrettin, S.; Guzman, J.; Corma, A. *Angew. Chem.*, **2005**, *44*, 2242.

[28] (a) Ishiyama, T.; Hartwig, J. *J. Am. Chem. Soc.*, **2000**, *122*, 12043. (b) Hirabayashi, T.; Sakaguchi, S.; Ishii, Y. *Adv. Synth. Catal.*, **2005**, *347*, 872. (c) Carrettin, S.; Guzman, J.; Corma, A. *Angew. Chem.*, **2005**, *44*, 2242. (c) Ikeda, S.-i.; Miyashita, H.; Taniguchi, M.; Kondo, H.; Okano, M.; Sato, Y.; Odashima, K. *J. Am. Chem. Soc.*, **2002**, *124*, 12060. (d) Korn, T.J.; Knochel, P. *Angew. Chem.*, **2005**, *44*, 2947. (e) Thathagar, M.B.; Beckers, J.; Rothenberg, G. *J. Am. Chem. Soc.*, **2002**, *124*, 11858.

[29] Phan, N.T.S.; Van Der Sluys, M.; Jones, C.W. Cover Picture: *Adv. Synth. Catal.* **2006**, *348*(6), 597.

[30] Xu, L.; Shi, *J. Comp. Aid. Geomet. Des.*, **2001**, *18*, 817.

[31] McQueen, J (Eds.) L. Le Cam. J. Neyman. In *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA. 1, 281-297. **1967**.

[32] Keogh, E.; Pazzani, M. In *Proc. of the 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Aug. 20-23. Boston, MA, USA. pp. 285-289. **2000**.

[33] (a) Keogh, E.; Pazzani, M. In *Proc. of the 21st Int. Conf. on Very Large Databases*, Boston, MA, **2000**, pp. 285-289. (b) Kruskall, J. B.; Liberman, M. *Time Warps, String Edits and Macromolecules*, Addison-Wesley, **1983**.

[34] (a) Bellman, R. *Dynamic Programming*, Princeton Univ. Press, New Jersey, **1957**. (b) Berndt, D.J.; Clifford, J. *Adv. Knowl. Discov. Data Min.*, **1996**, 229.

[35] Forgy, E. *Biometrics*, **1965**, *21*, 768.

[36] Diday, E. *Int. J. Comput. Sci.*, **1973**, *2*, 1.

Received: June 26, 2007 **Revised: November 5, 2007** **Accepted: November 5, 2007**