19

Physiological Analysis of Yeast Cell by Intelligent Signal Processing

Andrei Doncescu[1], Sebastien Regis[1], Katsumi Inoue[2] and Nathalie Goma[3]
[1]*University of Toulouse, LAAS CNRS UPR 8001*
[2]*National Institute of Informatics*
[3]*IPBS CNRS*
[1,3]*France*
[2]*Japan*

1. Introduction

Before getting into the details of this chapter, let us make some light upon the abiotic/biotic debate. The issue could be summarized as two main differences between biotic and abiotic: the first one is the internal regulation of biotic systems and what is generally called by some researchers "cellular intelligence" related to the possibility of communication between cells. More precisely, in both types of systems it is possible to identify mass and energy exchanges (thermodynamic laws control), but within the biotic systems only one can find a certain project capability, using an ascending informational system (genome towards metabolic networks and environment adaptation systems) and a descending one (inverse).

The living organisms science gave birth to two main research areas: biomimetics and systemic biology. Biomimetics is a new discipline based not on what could be extracted from nature but on what could be learned from it, through the biologic systems or biosystems. Many attempts on defining systems in general exist. The classic one is the definition given by Bertalanffy: the system is an "organized complex", delimited by the existence of "strong interactions or non trivial interactions", *i.e.* non linear interactions. A biologic system is supposed to replicate itself and develop a system of reactions to exterior perturbations. This replication is done on the basis of a non-systemic, not organized or less organized exploitation of the surrounding environment. As the biologic system "works" as an algorithm, it is quite normal that since the first days of the molecular biology (1959) the engineers intensified and diversified their references to biology, even before the general acceptance in the scientific community of the term nanotechnology. The convergence between biotechnology and nanotechnology is due to the conceptual statement "bio is nano". Many examples of biomimetics achievements may be recalled: the artificial pancreas, the artificial retina, biomaterials. The latter is very interesting because biomaterials are not homogeneous. Biosystems are also used to develop innovating methods allowing measuring physiologic parameters, to find diagnostics for diseases, and to evaluate the effectiveness of new therapeutic compounds. Encouraging activities take place in the biomedical field: the cancer, brain/heart vascular diseases, infectious diseases.

The biologic systems, as a result of billions of years of evolution, are complex and degenerated systems and this double characteristic makes their understanding extremely difficult. In general, for a biologic system, the cyclic issues are related to the measure redundancy, to variables pertinence and to the significant correlation between the parameters and the type of the model that has to be used. The difficulty of these models is the intrinsic nature of some of the constituting elements and their phenomenological reductionism. By giving a few examples their limited character will better understood: the phenomenological modeling does not take into account the metabolic capabilities of the system, the stoechiometrical modeling does not take into account the dynamics of the system, thus a "time-space" analysis is not possible, etc. Such observations drive us to envisage a new approach of biologic problems: passing from the analytic paradigm to the complexity paradigm, via the so called approach of biology of systems or biosystemics or systemic biology.

The systemic biology is, as a simple definition, the integration of mathematics, biology, physics and computer science for creating new models for the biologic systems. Kitano, one of the fathers of Biology of Systems defines the biosystemic strategy (13) as:

1. Defining and analysing the structure of these systems;
2. Studying their behaviour and characteristics under different conditions;
3. Studying the regulation processes through which the systems control their equilibrium states and manage their variations;
4. Identifying the processes that allow building systems that are adapted to a given function

From a purely biologic standpoint, there exist complex groups of interacting proteins, performing:

1. the metabolism;
2. the synthesis of DNA;
3. information processing;

These interactions are network-like organized. The purpose of this biochemical diagram where each node represents a specific protein, which regulates the biochemical conversions, is to explain the cellular physiology starting from the dynamics of the net, in a cells population context. We can now understand that, in this context, the biologic systems are uncertain and subordered to viability constraints. From a qualitative dynamics standpoint, the viability constraints induct two states: the homeostasis and the chaos. The homeostasis defined by C. Bernard and W. Cannon is the capability of a living system to preserve its vital functions, by maintaining, in a certain structural stability range, its parameters and, in a certain viability range, its internal variables. The longevity of this living is a function of this stability. The chaos, in its common definition, is more opposed the concept of order than to those of stability and viability. It is possible to continuously pass from order to chaos, by increasing the complexity of the studied system, and this can be done by simple variations of some of its bifurcation parameters, variations that might provoke a change in the nature of the systemÕs dynamics attractors by overriding critical bifurcation values. The two notions are derived from the notion of attractor of a dynamic system in the context of a wide regulation network. We can give as an example the nature of biologic regulation: direct, indirect, or causal, governing the response of these systems. As the evolution of these systems is still

rather unknown due to the continuous or discrete way of evolution we can suppose the existence of response thresholds. Biology inspires an analysis of the non-linear dynamics in terms of feedback loops (positive or negative). A positive feedback loop within the interaction graph of a differential system is a necessary condition of multistationarity which becomes, in biology, the cellular differentiation. A negative feedback loop within the interaction graph is a necessary condition for a periodic stable behaviour, very important in biology. It is quite possible that the oscillators coupling and multiple feedback loops interaction could lead to better oscillations coherence. The biologic regulation networks allow, among other things, to model genes interactions within a cell. Consequence to genome sequency, including the humain genome sequence, the postgenomics field has the goal to characterize the genes, the functions and the intercations beetween genes. The network of regulation takes an important and difficult role to assume the analysis at different scales, which means: "setup of new models using the biologic data and predicts the behavior of biologic systems from information extracted from genome sequence". But how Monod had enhanced: "everything which exists in universe is the result of hazard and necessity which it is not in contradiction with Occam principle to do not introduce supplementary clauses when it is not necessary. It is important to mention that the hazard presented in biology is not stochastic which means "to drive" the hazard but tychastic which describes some phenomenaÕs which escape to all statistic regularity. Therefore it is impossible to speak about the predictive capacity of the model which is the final goal of theoretical biology. From a realist view point we have made the hypothesis that the DNA could not include the finest maps of the organism but only some information about the iterative process of bifurcation and growth. If the approaches proposed before are in-silico, in the case of the in vitro the extraction of correct information is primordial. Basically, the information obtained is realized by direct observation and also by the interpretation of the response of the system. To be more accurate in this explication this approach is based on the perturbation of the analyzed system, supposed in a steady state, by a pulse of metabolite which belongs to the non linear differential model which described the biological system. Using the response of the system it is possible to deduce the relations between this variable and some variables of the model. For the complex models the analysis of the picks and the slope of the response directly link to the concentration of the metabolite it is not possible. The extraction of the information from responses supposes the cooperation of two parties :

1. the development of a model able to observe the dynamic of the system with a lot of precision
2. using of the mathematical tools allowing to tune the model to experimental observations

which means to solve a regression problem. If the model is linear the regression is linear therefore without any theoretical interest. Comparing with linear models the non-linear models offer infinity of possibility without the difficulties enclosed in the resolution of this kind of approach. One strategy is the Lotka-Voltera modelling. This method has been applied in ecology and focuses on the interaction between two species of type predator Ð prey. The inconvenient of this method in study of metabolic pathway or fermentation is that he metabolite depends of many compounds. The deduction of a non linear model from experimental data is an inverse problem which could be solved by a regression method or genetic algorithms which minimize the error between the model and the data. In practice the local minima which stop the convergence of the algorithms. One smart solution is the utilisation of Bayesian Methods or Simulated Annealing when the systems are small

size, no noise and we use a PC cluster. Another approach is the non linear estimation by NARMAX but these methods did not work with strong non linearties. The challenge in the case of the modeling of biological systems is the elucidation of the optimal evolution of the equations system when initial conditions are incomplete or missing. In this case, the analysis of the trajectories of the system is the central point to make the difference between regular trajectories (periodical or quasi periodic) and chaotic paths in the space of phases. The change of trajectories is directly related to the parameters of the model. Periodic trajectories will be identified by Poincaré sections and by using the Harmonic Wavelet Transform we will be able to make the difference between quasi-periodical theoretical and the chaos. The interest of the analysis by wavelet is the linearization of the model to find the steady states stationary of the biological system excited by a non-stationary process. The original system will be decomposed into linear subsystems, each having the response in a frequency band of well defined. The final response is calculated by the addition of each subsystem response. This type of analysis allows the studying of the influence of modes of regulation on the time of relaxation of the cell and to find out the stationary states of the cellular cycle.

Today, the pace of progress in fermentation is fast and furious, particularly since the advent of genetic engineering and the recent advances in computer sciences and process control. The high cost associated with many fermentation processes makes optimization of bioreactor performance trough command control very desirable. Clearly, control of fermentation is recognized as a vital component in the operation and successful production of many industries. Despite the complexity of biotechnological processes, biotechnologists are capable of identifying normal and abnormal situations, undertaking suitable actions accordingly. Process experts are able to draw such conclusions by analysing a set of measured signals collected from the plant. The inexistence of satisfactory mathematical models impedes model-based approaches to be used in supervisory tasks, thus involving other strategies to be applied. That suggests that, despite the lack of process models, measured data can be used instead in the supervisory system development. The advances in measurement, data acquisition and handling technologies provide a wealth of new data which can be used to improve existing models.

In general, for a biologic system, the cyclic issues are related to the measure redundancy, to variables pertinence and to the significant correlation between the parameters and the type of the model that has to be used. The difficulty of these models is the intrinsic nature of some of the constituting elements and their phenomenological reductionism. Moreover, the dynamic nature and the inherent non-linearity of bio-processes make system identification difficult. The majority of kinetic models in biology are described by coupled differential equations and simulators implement the appropriate methods to solve these systems. Particularly, analysis of states occuring during experiences is a key point for optimization and control of these bioprocesses. Thus, model-based approaches using differential equations (26), expert system (31), fuzzy sets and systems (23), (9), neural networks (8) have been developed. However, although model-based approaches give more and more accurate results close to real outputs (10), these methods using simulation techniques can lead to wrong conclusions, because of lack of description parameters or during an unexpected situation. Non-model-based methods have an increasing success and are based on the analysis of the process biochemical signals. The detection and the characterization of the physiological states of the bioprocess are based on signal processing and statistical analysis of signals. For

example, methods based on covariance (21) and moving averages (4) have been proposed, but they do not take account of the changes occurring in the signals. Wavelet transform is a powerful tool for non-stationary signal analysis due to its good localization in time and frequency domains. Wavelets are thus sensitive to changes in signals. Bakshi and Stephanopoulos (3), then Jiang et al. (12) have successfully used wavelets to analyze and detect states during bioprocesses.

One of the main contribution of Artificial Intelligence to biological or chemical processes turns out to be the classification of an increasing amount of data. Can we do more than that and can an AI program contribute to help in discovery of hidden rules in some such complex process. In fact, even if we can predict, for instance, mutagenicity of a given molecule or the secondary structure of proteins, with high degree of accuracy, this is not sufficient to give a deep insight of the observed behavior. In this paper we present a method using Maximum of Modulus of Wavelets Transform, Hölder exponent evaluation and correlation product for the detection and the characterization of physiological states during a fermentation fed-batch bioprocess. Therefore, we consider the estimation of nonoscillating and isolated Lipschitz singularities of a signal.

2. Yeast biotechnology

The main process we are concerned is a bio-reaction, namely the dynamical behavior of yeast during chemostat cultivation. Starting from the observation of a set of evolutive parameters, our final aim is to extract logical rules to infer the physiological state of the yeast. Doing so, we obtain not only a better understanding of the system's evolution but also the possibility to integrate the inferred rules in a full on-line control process. The first thing we have to do is to capture and analyze the parameters given by the sensors. These signals must be treated to be finally given to the logic machine. Thus, two things have to be done : first, to denoise the signals, secondly to compute the local maximum values of the given curves. In fact, we are more interested in the variations of the signals than in their pure instantaneous values. We use a method issued from wavelets theory (1) and which tends to replace classical Fourier analysis. At the end of this purely analytic treatment, we dispose of a set of clean values for each critical parameter. Now, our idea is to apply Inductive Logic Programming to exhibit, starting from a finite sample set of numerical observations, a number of logical formulae which organize the knowledge using causal relationships. Inductive logic programming is a sub-field of machine learning based upon a first-order logic framework. So instead of giving a mathematical formula (for instance a differential equation) or a statistical prediction involving the different parameters, we provide a set of implicative logical formulae. A part of these formulae can generally be inferred by a human expert, so it a a way to partially validate the mechanism. But its remains some new formulae which express an unknown causality relation : in that sense, this is a kind of knowledge discovery. As far as we know, one of the novelties of our work is the introduction of a time dimension to simulate the dynamic process. In logic, this time variable is in general not considered except with some specific modal logics. So, we modelize the time with an integer-valued variable.

The methodology has been applied to a biotechnological process. Saccharomyces Cerevisiae is studied under oxidative regime (i.e., no ethanol production) to produce yeast under a laboratory environment in a bioreactor. Two different procedures are applied: a batch

procedure that is followed by a continuous procedure. The batch procedure is composed by a sequence of biological stages. This phase can be thought as a start-up procedure. Biotechnologists state that the behaviour in the batch procedure influences later in induced phenomena in the continuous phase. So complete knowledge of the batch phase is of great importance for the biotechnologist. The traditional way to get acquainted of such knowledge is at present carried out through offline measurements and analysis which most of the time produce results when the batch procedure has ended, thus lacking of real time performance. Instead, the proposed methodology allows for real time implementation. This example deals with the batch procedure. Among the set of available on-line signals the expert chooses the subset of signals which, according to the expert knowledge contain the most relevant information to determine the physiological state:

1. DOT : partial oxygen pressure in the medium.
2. O2 : oxygen percent in the output gas
3. CO2 : carbon dioxide percent in the output gas
4. pH.
5. OH- ion consumption : derived from control action of the pH regulator and the index of reflectivity.

The consumption of negative OH ions is evaluated from the control signal of the pH regulator. The actuator is a pump, switched by an hysteresis relay, that inoculates a basic solution (NaOH). The reflectivity, which is measured by the luminance, seems to follow the biomass density. Nevertheless its calibration is not constant and depends on the run. Yeasts are a very well-studied micro-organisms and today, such micro-organism like *Saccharomyces cerevisiae* which make the object of this study, are largely used in various sectors of the biomedical and biotechnology industrial processes. So, this is a critical point to control such processes. Two directions have been explored:

1. the *on-line* analysis : it does not allow to identify in an instantaneous manner and with certainty the physiological state of the yeast.
2. the *off-line* analysis : it allows to soundly characterize the current state, but generally too late to take into account this information and to adjust the process on the fly by actions of regulators allowing to adjust some critical parameters such that pH, temperature (addition of basis, heats, cooling).

To remedy these drawbacks, computer scientists in collaboration with micro-biologists develop tools for supervised control of the bioprocess. They use the totality of informations provided by the sensors during a set of sample processes to infer some general rules to which the biological process obeys. These rules can be used to control the next processes. This is exactly the problem we tackle in this paper. To sum up, our application focus on the evolutive behavior of a *bio-reactor* (namely yeast fermentation) that is to say an evolutive biological system whose interaction with physical world, described with pH, pressure, temperature, etc..., generates an observable reaction. This reaction is studied by the way of a set of sensors providing a large amount of (generally) numerical data, but, thanks to the logical framework, symbolic data could also be integrated in the future. For an approach based upon classification

and fuzzy logic, one can see (24) : this work is devoted to discover the different states of the bio-reactor but not to predict its behavior.

In a yeast culture, measures result of biology phenomena and physical mechanisms. That is why to bring the culture, it is always decisional between biology and physico-chemical. The biological reaction is function of the environment and an environmental modification will improve two types of biological responses. The first one is a quasi steady-state response, the micro-organism is in equilibrium with the environment. The biological translation of this state is kinetics of consummation, production and this phenomenon is immediate. The second biological response is a metabolic one, which can be an oxidative or fermentative mode, or a secondary metabolism. The characteristic of this response is that the time constants are relatively long. For cultures, in term of production, the essential parameters are metabolism control and performance (productivity and substrate conversion in biomass yield). With this goal, the process must be conducted by a permanent intervention in order to bring the culture to an initial point to a final point. This control can be done from acquired measures on process, which are generally gases. Indirect measures show the environmental dynamic, which is shown by gas balance, with respiratory quotient (RQ) and pH corrector liquid (see figure 1).

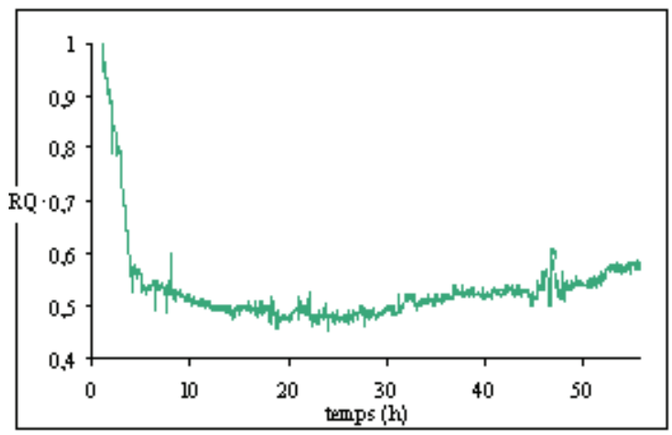


Fig. 1. An example of respiratory quotient evolution during a culture. x-axis is the time of the experience, y-axis is the amplitude of the signal.

Then, there are physical phenomenon, which are associated to real reactors. These mechanisms can be decomposed in many categories : transfer phenomenon (mass, thermal and movement quantity), regulation (realised by an operator), introduction of products, and mixing. These mechanisms interfere with biology and it is significant to notice that relaxation times of these phenomena are of size order of response time of biological response. With all these phenomena, a variable can be described by the following equation (see (26)) :

$$\frac{dV}{dt} = \Delta.\left(\frac{V_{equilibrium} - V(t)}{\tau_{physical}}\right) + r_{V(t)} + \Phi_{V(t)} \quad (1)$$

where:

- $\frac{dV}{dt}$ corresponds to the dynamic of the system.

- $\Delta.\left(\frac{V_{equilibrium}-V(t)}{\tau_{physical}}\right)$ is variable variation between biological and physical parameters. $\tau_{physical}$ is the time constant of physical phenomena; this constant can not be characterised because it depends on reaction progress.
- $r_{V(t)}$ is the volumic density of reaction of the variable V, it is a biological term.
- $\Phi_{V(t)}$ corresponds to an external intervention which results of a voluntary action.

Moreover, it is essential to observe that there is a regulation loop between biology and physic (see figure 2). The problematic is, from measures, to isolate or eliminate perturbations. These responses depend on physical phenomena or human interventions (process regulation). It is to quantify biological kinetics and by this way to optimise biological kinetics and control that is to say identify modifications of the biological behaviour. For example, in the case of yeast production, it is important to maintain an oxidative metabolism by the control of glucose residual concentration, fermentative metabolism is prejudicial to the yield. The aim is to maintain an optimal production to avoid the diminution of substrate conversion yield, that is to say to remark the biological change between oxidative and fermentative metabolism.

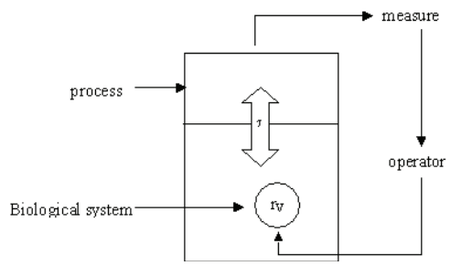


Fig. 2. Interactions between the biological system, the process and the operator.

3. Knowledge based methodology

In learning there is a constant interaction between the creation and the recognition of concepts. The goal of the methodology is to obtain a model of the process, which can be used in a supervisory system for condition monitoring. The complexity of this model imposes the co-operation of data mining techniques along with the expert knowledge. When only expert knowledge is used to identify process situations or states, any of these situations can arise: Ø the expert can express only a partial knowledge from process, Ø he does know the existence of several states but he ignores how to recognise them from on-line data, or/and Ø he doesn't have a clear idea on which states to recognise. For example, in the yeast production batch phase, biotechnologists apply expert rules when recognising some of the physiological states from on-line data. Nevertheless those rules usually don't take into account other phenomena that can change the evolution of signals without any influence in the physiological state. This leads to wrong conclusions. It is mainly due to the fact that the expert is not able to draw conclusions from the analysis of multiple signals between which there exist true relationships. Nevertheless, a classification tool copes well with this drawback. This proves the need of an iterative methodology to identify the biological states, which refines the expert knowledge with the analysis of past data sets.

Initialize : $E' = E$ (initial set of examples)
$H = \emptyset$ (initial hypothesis)
While $E' \neq \emptyset$ do
Choose $e \in E'$
Compute a covering clause C for e
$H = H \cup \{C\}$
Compute $Cov = \{e' \mid e' \in E, B \cup H \models e'\}$ $E' = E' \setminus Cov$ End while

Fig. 3. General Progol scheme

3.1 Standard ILP task

We stay within the pure setting i.e. where programs do not involve negation. In that case, the meaning of a logic program is just its least Herbrand model, which is a subset of the Herbrand universe i.e. the full set of ground atoms. In that setting, a concept C is just a subset of the Herbrand base. As shortly explained in our introduction, an ILP machine takes as input :

- a finite proper subset $E =< E^+, E^- >$ (the training set in Instance Based Learning terminology) where E^+ can be considered as the positive examples i.e. the things known as being true and is a subset of C, E^- as the negative examples and is a subset of $\overline{C}$.
- a logic program usually denoted B (as background knowledge) representing a basic knowledge we have concerning the concept to approximate. This knowledge satisfies two natural conditions : it does not explained the positive examples : $B \not\models E^+$ and it does not contradict the negative ones : $B \cup E^- \not\models \bot$

So the ILP task consists in finding a program H such that $H \cup B \models C$. One of the most popular method is to find H such that $H \cup B \models E^+$ and $H \cup B \cup E^- \not\models \bot$. In the field of classification, it is known that this approach, minimizing the error rate over the sample set (here we have zero default on the sample set) does not always guaranty the best result for the whole concept C.

Nevertheless, as far as we know, no alternative induction principle is used for ILP. Of course, as explained in the previous section, an ILP machine could behave as a classifier. Back to the introduction, the sample set $S = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}$ is represented as a finite set of Prolog facts $class(x_i, y_i)$ constituting the set E^+. The ILP machine will provide an hypothesis H. Consulting H with a Prolog interpreter, for a given element x, we get the class y of x by giving the query $class(x, Y)$? to the interpreter.

3.2 Progol machinery

Back to the standard ILP process, instead of searching for consequences, we search for premises : it is thus rather natural to reverse standard deductive inference mechanisms. That is te case for Progol which uses the so-called inverse entailment mechanism ((20)). Progol is a rather complex machine and we only try to give a simplified algorithm schematizing its behavior in figure 3. One can read the tutorial introduction of CProgol4.4[1] from which we take our inspiration.

The main point we want to update is the choice of the relevant clause C for a given training example e. Let us precise here how this clause is chosen.

[1] available on **http://www.cs.york.ac.uk/mlg/progol.html** where a full and clear description is given.

3.3 The choice of the covering clause

It is clear that there is an infinite number of clauses covering e, and so Progol need to restrict the search in this set. The idea is thus to compute a clause C_e such that if C covers e, then necessarily $C \models C_e$. Since, in theory, C_e could have an infinite cardinality, Progol restricts the construction of C_e using mode declarations and some other settings (like number of resolution inferences allowed, etc...). Mode declarations imply that some variables are considered as input variables and other ones as output variables : this is a standard way to restrict the search tree for a Prolog interpreter.

At last, when we have a suitable C_e, it suffices to search for clauses C which θ-subsume C_e since this is a particular case which validates $C \models C_e$. Thus, Progol begins to build a finite set of θ-subsuming clauses, $C_1, \ldots, C_n$. For each of these clauses, Progol computes a natural number $f(C_i)$ which expresses the *quality* of C_i : this number measures in some sense how well the clause explains the examples and is combined with some compression requirement. Given a clause C_i extracted to cover e, we have :

$$f(C_i) = p(C_i) - (c(C_i) + h(C_i) + n(C_i))$$

where :

- $p(C_i) = \#(\{e \mid e \in E, B \cup \{C_i\} \models e\})$ i.e. the number of covered examples
- $n(C_i) = \#(\{e \mid e \in E, B \cup \{C_i\} \cup \{e\} \models \bot\})$ i.e. the number of incorrectly covered examples
- $c(C_i)$ is the length of the body of the clause C_i
- $h(C_i)$ is the minimal number of atoms of the body of C_e we have to add to the body of C_i to insure output variables have been instantiated.

The evaluation of $h(C_i)$ is done by static analysis of C_e. Then, Progol chooses a clause $C = C_{i_0} \equiv \arg\max_{C_i} f(C_i)$ (i.e. such that $f(C_{i_0}) = max\{f(C_j) \mid j \in [1, n]\}$). We may notice that, in the formula computing the number $f(C_i)$ for a given clause C_i covering e, there is no distinction between the covered positive examples. So $p(C_i)$ is just the number of covered positive examples. The same computation is valuable for the computation of $n(C_i)$ and so success and failure could be considered as equally weighted.

To abbreviate, we shall denote $Progol(B, E, f)$ the output program P currently given by the Progol machine with input B as background knowledge, E as sample set and using function f to chose the relevant clauses. In the next section, we shall explain how we introduce weights to distinguish between examples.

4. A boosting-like mechanism for Progol

As explained in our introduction, a Progol machine is a consistent learner i.e. it renders only hypothesis with no error on the training set : so the sample error at the end of a learning loop, $\epsilon^t = \Sigma_{\{i|x_i \text{ misclassified}\}} w_t(i)$, is 0 since each example is necessarily correctly classified. So we cannot base our solution over the computation of such an error since the nullity of this error is a halting condition for a standard boosting algorithm. So we introduce a new way to adjust the weights. Given an example e_i, since it is covered we have $B \cup H_t \vdash e$. Given

an other problem instance e, we claim that the longer the proof for $B \cup H_t \vdash e$, the riskier the prediction for e.

5. Inductive logic programming : basic concepts

Mathematical logic has always been a powerful representation tool for declarative knowledge and Logic Programming is a way to consider mathematical logic as a programming language. A set of first order formulae restricted to a clausal form, constitutes a logic program and as such, becomes executable by using standard mechanisms of theorem proving field, namely unification and resolution. Prolog is the most widely distributed language of this class. In this context, the data and their properties, i.e. the observations, are represented as a finite set of logical facts E. E could generally been discomposed into the positive examples E^+ and the negative ones E^-. In case of background knowledge, it is described as a set of Horn clauses B. This background knowledge is supposed to be insufficient to explain the positive observations and the logical translation of this fact is : $B \not\models E^+$ but there is no contradiction with the negative knowledge: $B \cup E^- \not\models \bot$. So an ILP machinery ((20)), with input E and B, will output a program H such that $B \cup H \models E$. So H constitutes a kind of explanation of our observations E. Expressed as a set of logical implications (Horn clauses) $c \rightarrow o$, c becomes a possible cause for the observation $o \in E$. We give here a simple scheme giving a functional view of an ILP machine.

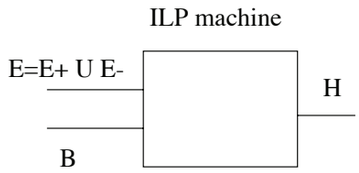


Fig. 4. ILP machine functional scheme

It is important to note that a logic program is inherently non deterministic since a predicate is generally defined with a set of distinct clauses. To come back to our introductive notation, we can have two clauses of the form (using Prolog syntax) $o \leftarrow c$ and $o \leftarrow c'$: this means that c and c' are potential explanations for o. The dual situation $o \leftarrow c$ and $o' \leftarrow c$ where the same cause produces distinct effects is also logically consistent. So this is a way to deal with uncertainty and to combine some features of fuzzy logic. The main difference is that we have a sound and complete operational inference system and this is not the case for fuzzy logic.

6. Formalization of our problem

We have 4 potential states for the bio-reactor : we shall denote e_1, e_2, e_3 and e_4 these states to avoid unusefull technical words. e_4 will be considered as a terminal state where the bio-reactor is stable because of the complete combustion of the available ethanol. We add a specific state e_5 corresponding to a stationary situation where the process is going on without perturbation. The transition between two states is a critical section where the chosen observable parameters (bio-mass, pH and O_2 rate) give rise to great variations.

The predicate to learn with our ILP machine is :

$$\textbf{to-state}(E_i, E_t, P_1, P_2, P_3, T)$$

meaning that the bio-reactor is going into state E_t knowing that at step T, the current bio-reactor parameters are the P_i's and the current state E_i. It is thus clear that we can deal with as many parameters as we want, but we restrict to 3 in this experiment. As explained in our introduction, we introduce the variable T to simulate the dynamic behavior of our process. As far as we know, previous experiments using inductive logic programming generally compute causal relationship between parameters which do not involve time. So we have to learn a transition system where we do not know what are the basic actions activating a transition. Our informations about the system are given by sensors providing numerical signals for p_1 (bio-mass), p_2 (pH value) and p_3 (O_2 rate),. These signals are analyzed using a wawelet-based system, we visualize the curve of the different functions and we extract the values of the differential for each given function. These values constitutes the input of our learning system. So, we want to obtain a causal relationship between the transitions of the system and the values of the differentials of the curve describing the evolution of our parameters.

So, we add a predicate **derive**$(P, T, P1)$ which expresses the fact that, for the curve of the parameter P,

at time T, the value of the differential is $P1$. It is thus easy to describe what is a pike for the curve describing P : this is included in our background knowledge. These pikes correspond to

local minima/maxima for the given parameter. So, we are also interested in the sign of

the derivative and we include specific predicates (**positive/2**, **negative/2**)

to compute and test this sign.

As background knowledge (corresponding to the B input of our scheme 4), we have the definitions of predicates **derive/3**,

positive/2, **negative/2**, **pike/2**. Here is an overview of the mode declaration

to describe the potential causes of a state transition.

```
:- modeb(*,pike(+parameter,-float))?
:- modeb(3,pike(-parameter,+float))?
:- modeb(1,positive(+parameter,+float))?
:- modeb(1,negative(+parameter,+float))?
:- modeb(1,between(+float,+float,+float))?
:- modeb(1,between(-float,+float,+float))?
% a (very little) part of

% our background knowledge.
pike(P,T) :- derive(P,T, P1),P2 is P1, between(P2,-0.001,0.001).
```

The results in this paper are obtained using the last implementation of Progol, namely **CProgol4.4** freely available on the following site **http://www.cs.york.ac.uk/mlg/progol.html**. Of course, we get a lot of rules, depending on the quantity of introduced examples. Some of them are not really interesting : they only generalize one or two examples. But we get, for instance, the next one (among the simplest ones to explain) :

```
to_state(E,E,A,B,C,T) :- derive(p1,A,T),
         derive(p2,B,T), derive(p3,C,T),
         positive(p1,T), positive(p2,T),
         positive(p3,T).
```

This rule indicates that there is no evolution of the metabolism state (the bio-reactor remains in the same state) when the parameters have an increasing slope but that we do not encounter maxima or minima. In general, the obtained rules are long except those ones generalizing only one or two examples. Nevertheless, there are some observations where this rule could be overcame : this means that we need (at least) an other parameter $p4$ to better understand the behaviour of the machinery.

7. Detection and characterization of physiological states

In microbiology, a physiological state (or more simply, a state) is, qualitatively, the set of potential functionalities of a micro-organism, and, quantitatively, the level of expression of theses functionalities. The environment has a strong influence on the activity of the micro-organism due, on one hand, to its chemical composition (nature of substrate, pH...) and, on the other hand to its physical properties (temperature, pression...). Yeast can react on the availability of substrates such as carbon and nitrogen sources, or oxygen, by a flexible choice of different metabolic pathway. It is possible to analyze the global metabolism by genetical analysis, biochemical or biophysical analysis but the complexity of the biological system requires a simplification of the characterization by the analysis of some functionalities of some known mechanisms. The quantification of materials and energy interactions flows between the micro-organism and the environment enables to have a macroscopic characterization of several intrinsic metabolism of yeast population which, by correlation, enables to differentiate several physiological states even if the biological characterization is unknown. Thus the detection, as far as we know, is based on the analysis of biochemical signals measured during the bioprocess. A bioprocess is the set up of the fermentors protocol. Fermentors are composed of a number of different components which can be grouped by their functions, i.e. temperature control, speed control, continuous culture accessories. In this context the ultimate aim of bioprocess analysis therefore is a detailed monitoring of biological system, the chemical and physical environment and how these interact. However, no reliable technique exist to carry out real-time measurement of non-volatile substrates and metabolites in the fermentor. Several works using various approaches, lead to the conclusion that the limits of a state are linked to the singularities of biochemical signals: Steyer et al. (31) (using expert system and fuzzy logic), Bakshi and Stephanopoulos (3) (using expert system and wavelets) and Doncescu et al. (6) (using inductive logic) show that the beginning and the end of a state correspond to singularities of the biochemical signals measured during the process. In a fed-batch bioprocess, a physiological state can occur several times during the experience. After the detection of states, it is then necessary to characterize these states. The characterization is often based on the statistical properties of the biochemical signals. Experts in microbiology characterize the states by analysing and comparing the variations and the values of different biochemical signals and by a deductive reasoning using " if-then" rules. These approaches can be linked to mathematical methods based on correlation. Classification methods based on Principal Components Analysis (PCA) (27), adaptive PCA (15), and kernel

PCA (14) enable to distinguish and characterize the different states. However, these methods (except the adaptive PCA) do not take into account the temporal variation of the signals. The adaptive PCA is a PCA applied directly on wavelet coefficients in order to take into account the variations of the biological system. It has been shown that it can characterize the Lipschitz singularities of a signal by following the propagation across scales of the modulus maxima of its continuous wavelet transform. For identifying the boundaries of states, we propose to use the Maximum of Modulus of Wavelets Transform (17)(16) to detect the signals singularities. The singularities are selected according to their Hölder exponent evaluation between -1 and 1. The characterization of the states is based on the correlation product between the signals on intervals whose boundaries are the selected singularities.

8. Detection and selection of singularities by wavelets and Hölder exponent

The singularities of the biochemical signal correspond to the boundaries of the states. These signals are non-stationary and non-symmetrical; they are not chirps and have no infinite oscillations (see figure 5).
Several authors have proposed to use wavelets to detect the singularities of the signals for the detection of states: Bakshi and Stephanopoulos (3) and more recently Jiang et al. (12). Besides singularities correspond to maxima of modulus of wavelets coefficients. Bakshi and Stephanopoulos (3) propose to detect the maxima by analysing the variation of the wavelet coefficients through a multi-scale analysis but they don't explicitly characterize the nature of detected singularities. Jiang et al. (12) propose to select meaningful singularities by using a threshold on the finest scale, but the determination of the threshold remains empirical. After the detection of singularities by the Maxima of Modulus of Wavelet Transform, we propose to use the evaluation of Hölder exponent to characterize the type of singularities and eventually select meaningful singularities.

The wavelets are a powerful mathematical tool of non-stationary signal analysis, signals whose frequencies change with time. Contrarily to the Fourier Transform, Wavelet Transform can provide the time-scale localization. The performance of the Wavelet Transform is better than of the windowed Fourier Transform. Because of these characteristics, Wavelet Transform can be used for analyzing the non-stationary signals such as transient signals. Wavelets Transformation (WT) is a rather simple mechanism used to decompose a function into a set of coefficients depending on scale and location. The definition of the Wavelets Transform is:

$$W_{s,u}f(x) = (f \star \psi_{s,u})(x) = \int f(x)\psi(\frac{x-u}{s})dx \tag{2}$$

where ψ is the wavelet, f is the signal, $s \in R^{+*}$ is the scale (or resolution) parameter, and $u \in R$ is the translation parameter. The scale plays the role of frequency. The choice of the wavelet ψ is often a complicated task. We assume that we are working with an admissible real-valued wavelet ψ with r vanishing moments ($r \in N^*$).

The wavelet is translated and dilated as in the next relation :

$$\psi_{u,s} = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}) \tag{3}$$

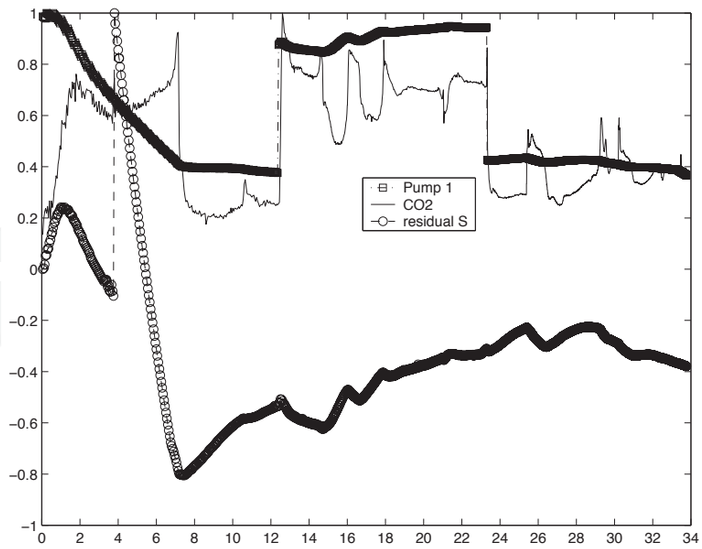


Fig. 5. Example of biochemical signals measured during the bioprocess. Pump 1 is a pump providing substance in the process, CO2 is the measured carbon dioxid and residual S is the residual substrate of the micro-organisms of the bioprocess. The signals have been normalized.

The dilation allows the convolution of the analyzed signal which different sizes of "window" wavelet function. For the detection of the singularities and of the inflexion points of the biochemical signal, we use the Maxima of Modulus of Wavelets Transform (16). The idea is to follow the local maxima at different scales and to propagate from low frequencies to high frequencies. These maxima correspond to singularities, particularly when the wavelet is the derivative of a smooth function:

$$\psi(x) = \frac{d\theta(x)}{dx}$$

$$W_{s,u}f(x) = f * \psi_{s,u} = f(x) * \frac{d\theta(x/s)}{dx}$$

Yuille and Poggio (35) have shown that if the wavelet is derivative of the Gaussian, then the maxima belong to connected curves which are continuous from a scale to another. The detection of the singularities of the signal is thus possible by using the wavelets (see for example figure 6).

The discretization form of Continuous Wavelet Transform is based on the next form of the Mother Wavelet :

$$\psi^{m,n}(t) = a_0^{-m/2}\psi(\frac{t - nb_0a_0^m}{a_0^m}) \tag{4}$$

By selecting a_0 and b_0 properly, the dilated mother wavelet constitutes an orthonormal basis of $L^2(R)$. For example, the selection of $a_0 = 2$ and $b_0 = 1$ provides a dyadic-orthonormal Wavelet Transform (DWT). The decomposed signals by DWT will have no redundant information thanks to the orthonormal basis.

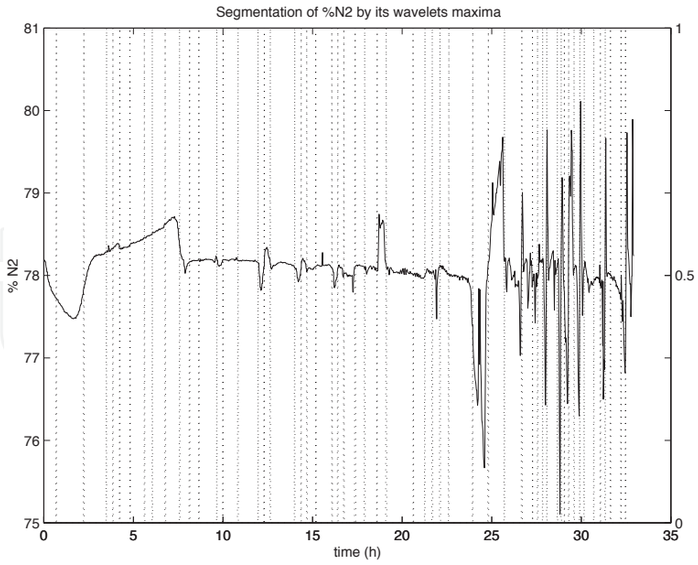


Fig. 6. Segmentation of N2 (nitrogen). Each vertical dotted line correspond to a singularity of the signal detected by wavelets. The wavelet is a DOG (first derivative of Gaussian) and the scales go from 2^0 to 2^3.

Jiang et al. (12) have proposed to select the maxima by using thresholding. Besides, all the singularities are not relevant only some of them are meaningful. However, as stated above, the thresholds proposed by Jiang et al. are chosen empirically. To select the meaningful singularities, we proposed using the Hölder exponent. The Hölder exponent is a mathematical value allowing characterization singularities. The fractal dimension could also be used but only the Hölder exponent can characterize locally each singularity. A singularity in a point x_0 is characterized by the Hölder exponent (also called Hölder coefficient or Lipschitz exponent). This exponent is defined like the most important exponent α allowing to verify the next inequality:

$$|f(x) - P_n(x - x_0)| \leq C|x - x_0|^{\alpha(x_0)} \tag{5}$$

We must remark that $P_n(x - x_0)$ is the Taylor Development and basically $n \leq \alpha(x_0) < n + 1$. Hölder exponent measures the remainder of a Taylor expansion and more of this measures the local differentiability:

1. $\alpha \geq 1$, $f(t)$ is continuous and differentiable.
2. $0 < \alpha < 1$, $f(t)$ is continuous but non-differentiable.
3. $-1 < \alpha \leq 0$, $f(t)$ is discontinuous and non-differentiable.
4. $\alpha \leq -1$, $f(t)$ is not longer locally integrable.

Therefore Hölder exponent could be extended to the distribution. For example the Hölder exponent of a Dirac is equal to -1. A simple computation leads to a very interesting result of the Wavelets Transform (11):

$$|W_{s,u}f(x)| \simeq s^{\alpha(x_0)} \tag{6}$$

This relation is remarkable because it allows to measure the Hölder exponent using the behavior of the Wavelets Transform. Therefore, at a given scale $a = 2^N$ the $W_{a,b}f(x)$ will be maximum in the neighborhood of the signal singularities. The detection of the Hölder coefficient is linked to the vanishing moment of the wavelet: if n is the vanishing moment of the wavelet, then it can detect Hölder coefficients less than n (16). We use a *DOG* wavelet (DOG: first derivative of Gaussian) with a vanishing moment equal to 1; consequently we can only detect Hölder coefficients smaller than 1. This is not a real problem because we are interested (*in this application*[2] by the singularities as step or dirac and the Hölder coefficient of these singularities are smaller than 1. Moreover, the meaningful singularities of the fed-batch bioprocess have Hölder exponents smaller than 1 which correspond to sharp singularities. This type of variations are meaningful for the fed-batch bioprocess fermentation because of many external regulations of the process. Moreover, for Hölder coefficients greater than 1 particularly for integer values, there are difficulties to interpret the Hölder coefficient (see (19) cited in (17)). To evaluate the Hölder coefficient using the wavelets, there are two main ways:

(1) the graphical method which consists in finding the maximum line i.e. the maximum which propagates through the scales, and computes the slopes of this maximum line (often using a log-log representation). The computed slope corresponds to the Hölder coefficient (16).
(2) the minimization method which consists in minimizing a function which has one of the parameters the Hölder coefficient (17). The function is the following:

$$\sum_j \left(ln_2(|s_j|) - ln_2(C) - j - \frac{\alpha(x_0) - 1}{2} ln_2(\sigma^2 + 2^{2j}) \right)^2 \tag{7}$$

where s_j represents the maximum at scale j, C is a constant depending on the singularity localized in x_0, σ is the standard deviation of a Gaussian approximation (see (17)), and $\alpha(x_0)$ the Hölder exponent.

In (17), a gradient descent algorithm is proposed to solve the minimization, but this technique is very sensitive to local minima. Recently, a minimization using Genetical Algorithms has been proposed (18) and used in bioprocess. More precisely it uses Differential Evolutionary (DE) algorithms. The DE algorithms was introduced by Rainer Storn and Kenneth Price (33).

9. Differential evolution

Differential Evolution (DE)(33) is one of Evolutionary Algorithms (EA) which are a class of stochastic search and optimization methods including Genetic Algorithms (GA), evolutionary programming, evolution strategies, genetic programming and all methods based on genetics and evolution. Through its fast convergence and robustness properties, it seems to be a promising method for optimizing real-valued multi-modal objective functions. Compared to traditional search and optimization methods, the EAs are more robust and straightforward to use in complex problems : they are able to work with minimum assumptions about the objective functions. These methods are slower because due to the generation of the population

[2] However it is always possible to use other wavelets with greater vanishing moment for others applications in bioprocesses

and the selection of individuals for crossing. The goal is to obtain the trade-off between accuracy and computing time.

The generation of the vectors containing the parameters of the model is made by applying an independent procedure :

$$X_{i,G} = X_{1,i}.....X_{D,i} \tag{8}$$

with $i = 1...NP$, is the index of one individual of the population; D is the number of parameters which have to be estimated; NP is the number of individuals in one population; G is the index of the current population; i is one individual of the population and $X_{j,i}$ is the parameter j of the individual i in the population G.

As Genetic Algorithms are stochastic processes, the initial population has been chosen randomly, but the initialization of the parameters is based on experts knowledge. Trial parameter vectors are evaluated by the objective function. Several objective functions are tested to produce results on Hölder coefficient detection. For simple GA algorithms the new vectors are the result of the difference between two population vectors and the result is added to a new one. It's a simple crossing operation. The objective function determines if the new vector is more efficient than a candidate population member and replace it if this simple relation is true. In the case of the DE the generation of the new vectors are realized by the difference between the "old vectors" given an weight to each one.

We have tested and compared different schemes of individual generations :

- $DE/rand/1$: For each vector $X_{i,G}$ a perturbed vector $V_{i,G+1}$ is generated according to :

  $$V_{i,G+1} = X_{R1,G} + F * (X_{R2,G} - X_{R3,G})$$

  $R1, R2, R3 \in [1, NP]$: individuals of population, chosen randomly, $F \in [0, 1]$: controls the amplification $(X_{R2,G} - X_{R3,G})$
  $X_{R1,G}$: the perturbed vector. There is no relation between $V_{i,G+1}$ and $X_{i,G}$. The objective function must evaluate the quality of this new trial parameter with respect to the old member. If $V_{i,G+1}$ yields a lower objective function value, $V_{i,G+1}$ is set to $X_{i,G+1}$ in the next generation or there is no effect.

- $DE/best/1$: It is like $DE/rand/1$ but is generating $V_{i,G+1}$ by integrating the most performante vector :

  $$V_{i,G+1} = X_{best,G} + F * (X_{R1,G} - X_{R2,G})$$

  $X_{best,G}$: best vector of population G, $R1, R2 \in [1, NP]$: individuals of population, chosen randomly. As $DE/rand/1$, the objective function compares the quality of $V_{i,G+1}$ and $X_{i,G}$; the smallest of the two is kept in the next population.

- Hydrid Differential evolution algorithms: As DE algorithms, a perturbed vector is generated, but the weight F is a stochastic parameter.

To increase the diversity potential of the population, a crossover operation is introduced. $X_{i,G+1} = (X_{1i,G+1}, X_{2i,G+1}.....X_{Di,G+1})$ becomes :

$$V_{ji,G+1} \begin{cases} j = (n)_D, (n+1)_D, (n+L-1)_D \\ X_{ij,G} \text{ otherwise} \end{cases}$$

$n \in [1, D]$: starting index, chosen randomly
$(n)_D$ = n mod D
$L \in [1, D]$: number of parameters which are going to be exchanged

10. Use of GA for Hölder's coefficients detection

10.1 Implementation

The cost function we have to minimize is the following (17):

$$\sum_j \left(log_2(|a_j|) - log_2(C) - j - \frac{h(x_0) - 1}{2} log_2(\sigma^2 + 2^{2j}) \right)^2 \quad (9)$$

In the Holder objective function three parameters have to be estimated : $h(x_0)$, C and σ. Thus, one individual X in GA's population is represented by the vector X_h, X_C, X_σ. In our case, the size of population equals 30.

Using the graphical method and the DE, the Hölder coefficient found is quite close to -1, whereas the value computed by gradient descent is not correct. Moreover, if we consider the Hölder coefficient of the Step function, only DE provides quite good results while the graphical method and the values of the gradient descent are too far from the theoretical value. The last median square is not so accurate as DE's. The results obtained indicate that the DE can be used for the analyzed data .

For this simulation, the results are summarized in the following table :

Singularity	Dirac	Step 1	Step 2
Theoritical Hölder Coef.	-1	0	0
Hölder Coef. by Graph. Method	-0.5	0.51	0.51
Hölder Coef. by Grad. Descent	0.26	0.89	0.89
Least Median Square	-0.5	0.802301	0.802301
Hölder Coef. by AG	-0.5	-0.03	-0.04

We note that the graphical method is the fastest and used method, but the evaluation of the Hölder coefficient is sometimes imprecise as noted in (34), (22).

On a simple signal (see figure 7), this new method using DE provides better results than those of existing methods as shown in table 1.

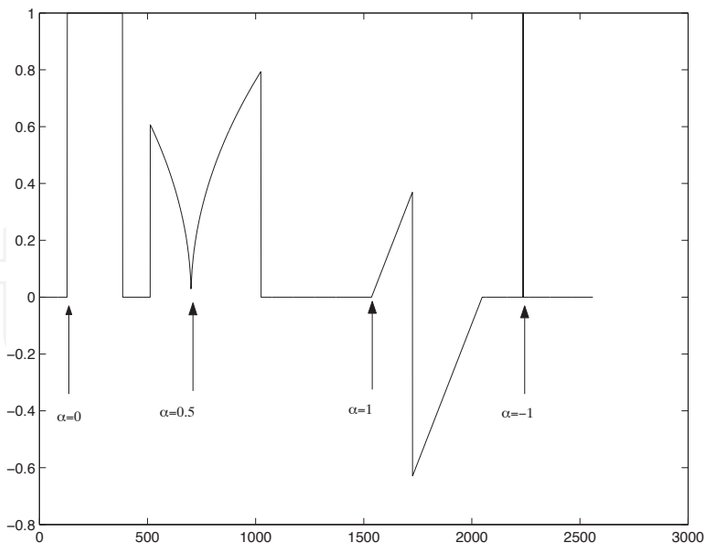


Fig. 7. Simple signal with singularity (step, cups, ramp and dirac) whose Hölder exponents α are known .

Singularity	Dirac	Step	cups	Ramp
theoritical Hölder coef.	-1	0	0,5	1
Hölder exponent by graphical method	-1,13	0,16	0,61	0,84
Hölder exponent by gradient descent	-0,24	0,39	0,74	1,20
Hölder exponent by DE	-1,02	0,02	0,52	1,0007

Table 1. Results of Hölder exponent evaluation by several methods. The wavelet used here is a LOG (second derivative of Gaussian).

11. Characterization by correlation product and classification

Once the states are bounded by the detected and selected singularities using the wavelets, they are characterized by the analysis of the correlations between the biochemical signals. On each interval defined by the singularities, a product of correlation is computed between all pairs of signals. The correlation coefficient (also called Bravais-Pearson coefficient, see (30)) is given by the equation:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sigma_x\sigma_y} \tag{10}$$

where x_i represent the values of one parameter (in a given interval), y_i the values of the second parameter (in the same interval), n the number of elements, $\overline{x}$ the average of the elements x_i (of the first biochemical signal), $\overline{y}$ the average of the elements y_i (of the second biochemical signal), et σ_x et σ_y the standard deviation of each of the two signals.
The correlation coefficient is equivalent to the cosine of the scalar product between two

biochemical signals projected in the correlation circle of a PCA realized between the two biochemical signals. On each interval, the sign of each correlation coefficient between two signals is kept. Each interval is thus characterized by a set of positive of negative signs. The intervals with the same set of signs are put in the same class as illustrated in the figure 8.

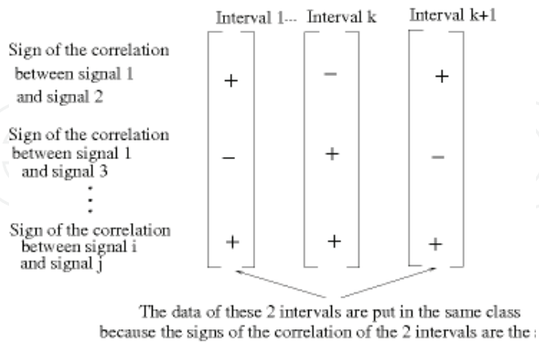


Fig. 8. Principle of the classification method based on wavelets, Hölder exponent and correlation coefficient

Ruiz et al. (27) propose a classification method based on PCA for a neighboring application (wastewater treatment): the data are projected in the space generated by the two first principal components. The method enables to reduce the size of the data space and to take account of the correlation of the signals. However the PCA doesn't take account of the time: the temporal evolution of the process is not taken into account. Ruiz et al. propose to use time analysis window of fixed size. But as the window has a fixed size, it doesn't really take account of the changes occurring during the bioprocess. So the method proposed in this article seems to be more adapted if it is necessary to take account of the variation of the process.

12. Experimental results

Tests have been done on two fed-batch fermentation bioprocesses and the first results have been presented in (25). The two bioprocesses are biotechnological processes using yeast called *Saccharomyces Cerevisiae*. In the first bioprocess we have applied the method to differentiate intrinsic biological phenomena from reactions of micro-organism to extern actions (changes in the environment). In the second bioprocess we directly use the method to detect and classify the states of the bioprocess. For the two fed-batch, the maximum scale is chosen empirically. Mallat and Zhong (17) propose to use as maximal scale $log_2(N) + 1$ where N is the number of measured samples of the signals. However if we use this maximal scale, several singularities would be removed. The empirical value which has been found it is 12. Concerning the Hölder exponent we are interested by the singularities between -1 and 1. For the evaluation of Hölder exponent using Genetical Algorithms, tests have shown that 100 iterations are sufficient for an acurate evaluation (18).

12.1 Differentiation between biophysical and biological phenomena

The first bioprocess is a bioprocess lasting about 25 hours. 12 biochemical signals have been measured during the bioprocess.
In a fed-batch bioprocess, there are two kinds of signals: the signals given by parameters

regulated by an extern action (expert in microbiology or control system) and the signals given by non regulated parameters. An example of regulated parameter is the agitation which is the speed of the rotor of the bioreactor and an example of non regulated parameter is the N2 (nitrogen). The actions on regulated parameters induce modifications of the physiology of the micro-organisms and physical changes in the bioreactor: there are *biophysical* phenomena. On the other hand, during the bioprocess, the micro-organisms have intrinsic physiological behavior: there are *biological* phenomena.
Is it possible to distinguish biophysical phenomena and biological phenomena?
To answer this question, we propose the following steps:

1. search the variations of the regulated signals. These variations are sharp variations which correspond to singularities as Dirac or step.
2. compare the sign of correlation product between regulated signals and non regulated signals before and after each detected singularity of the regulated signals. If the sign is the same before and after, there is no influence: it is a biological phenomenon. If the sign changes before and after, there is an influence: it is a biophysical phenomena.

We must note that:

- only the singularities of the regulated signal are detected and selected,
- to compare the sign of correlation product before and after each singularity, me must choose a reference temporal interval. Besides, the first temporal interval (delimited by the detected singularities) is considered as a biological interval as the bioprocess begins and the initial conditions are considered as biological.

An example of comparison between the agitation and the nitrogen is given in figures 9 and **??**.

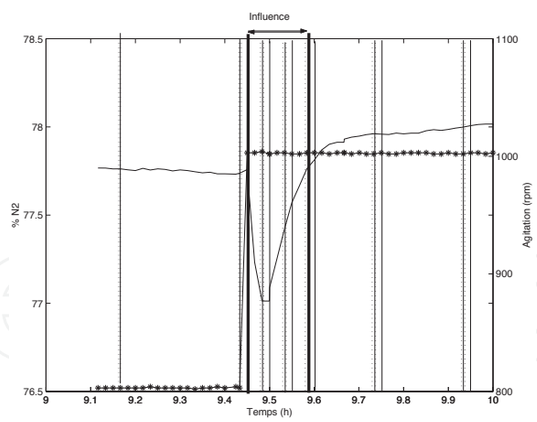


Fig. 9. An example of intervals (horizontal lines are singularities and correspond to boundaries of the temporal interval) with the segmentation given by the detection of regulated signals. This example has a duration of one hour (from 9 hours to 10 hours) taken from the first fed-batch. There are 14 intervals. Signals are agitation (stars) expressed in rotation per minute (rpm) and the percentage of nitrogen N2 (solid line).

Results confirm the observations of the expert. All the intervals considered as biological by the proposed method are considered as biological by the expert. Particularly, the last interval is considered as a biological phenomenon, which is well known by experts, as at

the end of a bioprocess, regulated signals are not modified. Another example is given by biological intervals located in the middle of the bioprocess which correspond to spontaneous oscillations.

12.2 Detection and classification of states

We have studied Saccharomyces cerevisiae dynamical behaviour during fed-batch aerated cultivation in oxidative metabolism. The maximal growth rate of this yeast was calculated to 0,45 h-1. The aim of our work was to determine by on line analysis, different physiological states of the yeast behavior only with the available sensors (pH, temperature, oxygen Ě). Off-line metabolites and intracellular carbohydrate reserve analysis help in a first approach to identify the physiological states. State recognition is performed by signal processing technics. The second bioprocess is a bioprocess lasting about 34 hours. 11 biochemical signals have been measued during the bioprocess.
We recall the used method for the characterization of intervals for the classification is given in section 4 and summarised in figure 8. The classification provided by the method gives interesting results shown in the figure 10. Once again, results obtained correspond to the experts observations. Particularly, the most interesting result concerns the detection and the characterization of a state resulting of an external action. Besides, the class number 8 corresponds to the addition of an acid[3] (the acid is not a regulated parameter as in the first example, but is directly introduced by the expert during the experience) in the bioprocess. All apparition of class 8 correspond exactly to an acid addition. These results were confirmed and validated. As far as we know, it is the first time that this kind of non-model-based approach can find and characterize automatically the addition of acid in a fed-bacth process. The results are promising and further analysis of the classification is necessary.

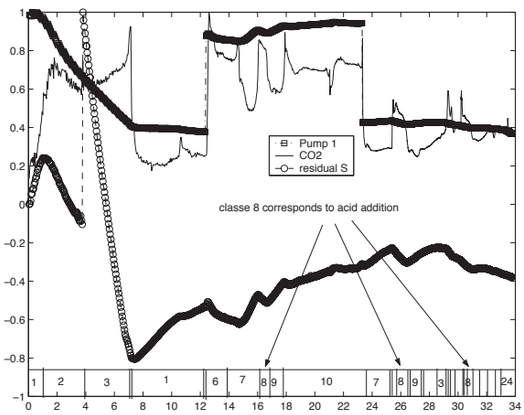


Fig. 10. Classification provided by the method. The wavelet is a DOG and the scales go from 2^0 to 2^{10}.

13. Discussion and conclusion

We apply logical tools to get explanation rules concerning the behavior of a bio-reactor. The ability to incorporate background knowledge and re-use past experiences marks out ILP as a

[3] because of industrial confidentiality, we are not allowed to gives more information

very effective solution for our problem. Instead of simply giving classification results, we get some logical rules establishing a causality relationship between different parameters of the bio-machinery. Among these rules, some ones are validated by expert knowledge, but some new ones have been provided. It yet appears that some previous rules have to be removed or modified to fit with new observations.

One of the main interest of this kind of approach is the fact that the resulting theory is easy to understand, even for a non specialist : the first order logic is, from a syntactic viewpoint, close to the natural language.

Intelligibility of resulting explanations is an other argument in favor of the ILP tools. A drawback of standard logic is the difficulty to deal with the time dimension : in some sense, standard logic is static and thus, not well suited to described dynamic process. One could hope that modal logic would be of some help, but it remains to design an inductive machine dealing with the temporal modalities, i.e. a way to reverse temporal logic inference system.

14. References

[1] Arneodo A. and al. Ondelettes, multifractales et turbulence de l'ADN aux croissances cristallines. *DIDEROT EDITEUR*, Paris, 1995.

[2] J. Aguilar-Martin, J. Waissman-Vilanova, R. Sarrate-Estruch, and B. Dahou. Knowledge based measurement fusion in bio-reactors. In *IEEE EMTECH*, May 1999.

[3] Bakshi, B. and Stephanopoulos, G. (1994). Representation of process trends-III. multiscale extraction of trends from process data. *Computer and Chemical Engineering*, 18(4):267–302.

[4] Cao, S. and Rhinehart, R. (1995). An efficient method for on-line identification of steady state. *Journal of Process Control*, 5(6):363–374.

[5] Domingo P. and Pazzani M.. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29,103-130, 1998.

[6] Doncescu, A., Waissman, J., Richard, G., and Roux, G. (2002). Characterization of bio-chemical signals by inductive logic programming. *Knowledge-Based Systems*, 15(1-2):129–137.

[7] Edelman, G. M. and Gally, J. A. (2001). Degeneracy and complexity in biological systems. In *Proc Nat Acad Science USA*.

[8] Gadkar, K., Mehra, S., and Gomes, J. (2005). On-line adaptation of neural networks for bioprocess control. *Computer and Chemical Engineering*, 29:1047–1057.

[9] Guillaume, S. and Charnomordic, B. (2004). Generating an interpretable family of fuzzy partitions from data. *IEEE Transactions on Fuzzy Systems*.

[10] Hvala, N., Strmcnik, S., Sel, D., Milanic, S., and Banko, B. (2005). Influence of model validation on proper selection of process models-an industrial case study. *Computer and Chemical Engineering*.

[11] Jaffard, S. (1997). Multifractal formalism for functions part 1 and 2. *SIAM J. of Math. Analysis*, 28(4):944–998.

[12] Jiang, T., Chen, B., He, X., and Stuart, P. (2003). Application of steady-state detection method based on wavelet transform. *Computer and Chemical Engineering*, 27(4):569–578.

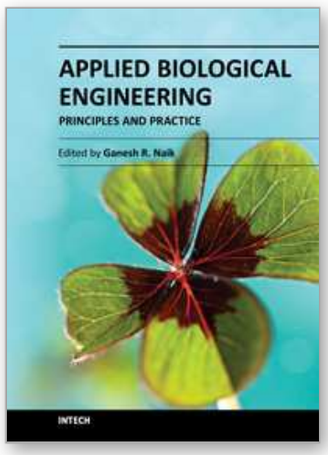
[13] Kitano, H. (2002). Computational systems biology. In *Nature 6912, 206*.

[14] Lee, J.-M., Yoo, C., Lee, I.-B., and Vanrolleghem, P. (2004). Multivariate statistical monitoring of nonlinear biological processes using kernel PCA. In *IFAC CAB'9*, Nancy, France.
[15] Lennox, J. and Rosen, C. (2002). Adaptative multiscale principal components analysis for online monitoring of wastewater treatment. *Water Science and Technology*, 45(4-5):227–235.
[16] Mallat, S. and Hwang, W.-L. (1992). Singularity detection and processing with wavelets. *IEEE Trans. on Information Theory*, 38(2):617–643.
[17] Mallat, S. and Zhong, S. (1992). Characterization of signals from multiscale edges. *IEEE Trans. on PAMI*, 14(7):710–732.
[18] Manyri, L., Regis, S., Doncescu, A., Desachy, J., and Urribelarea, J. (2003). Holder coefficient estimation by differential evolutionary algorithms for saccharomyces cerivisiae physiological states characterisation. In *ICPP-HPSECA*, Kaohsiung, Taiwan.
[19] Meyer, Y. (1990). *Ondelettes et Opérateurs*, volume I. Hermann.
[20] S. Muggleton. Inverse entailement and Progol. New Gen. Comput.,13:245-2,1998.
[21] Narasimhan, S., Mah, R., Tamhane, A., Woodward, J., and Hale, J. (1986). A composite statistical test for detecting changes in steady state. *American Institute of Chemical Engineering Journal*, 32(9):1409–1418.
[22] Nugraha, H. B. and Langi, A. Z. R. (2002). A wavelet-based measurement of fractal dimensions of a 1-d signal. In *IEEE APCCAS*, Bali, Indonesia.
[23] Polit, M., Estaben, M., and Labat, P. (2002). A fuzzy model for an anaerobic digester, comparison with experimental results. *Engineering Applications of Artificial Intelligence*, 15(5):385–390.
[24] Rocca J.. Technical report of Laboratoire d'automatisme et architecture des sytemes. 1998.
[25] Régis, S., Doncescu, A., Faure, L., and Urribelarea, J.-L. (2005). Détection et caractérisation d'un bioprocédé par l'analyse de l'exposant de hölder. In *GRETSI 2005*, Louvain-la-Neuve, Belgium.
[26] Roels, J. (1983). *Energetics and kinetics in biotechnology*. Elsevier Biomedical Press.
[27] Ruiz, G., Castellano, M., Gonzàlez, W., Roca, E., and Lema, J. (2004). Algorithm for steady states detection of multivariate process: application to wastewater anaerobic digestion process. In *AutMoNet 2004*, pages 181–188.
[28] S. Régis. *Segmentation, classification, et fusion d'informations de séries temporelles multi-sources: application à des signaux dans un bioprocédé*. Thèse de Doctorat (PhD thesis), Université des Antilles et de la Guyane, Novembre 2004.
[29] S. Régis, L. Faure, A. Doncescu, J.-L. Uribelarrea, L. Manyri, and J. Aguilar-Martin. Adaptive physiological states classification in fed-batch fermentation process. In *IFAC CAB'9*, Nancy, France, March 2004.
[30] Saporta, G. (1990). *Probabilités, et Analyse des données et Statistique*. Technip.
[31] Steyer, J., Pourciel, J., Simoes, D., and Uribelarrea, J. (1991). Qualitative knowledge modeling used in a real time expert system for biotechnological process control. In *IMACS International Workshop "Decision Support Systems and Qualitative Reasoning"*.
[32] J.P. Steyer. *Sur une approche qualitative des systèmes physiques : aide en temps rèel à la conduite des procédés fermentaires*. Thèse de Doctorat, Université Paul Sabatier, Toulouse France, Décembre 1991.

[33] Storn, R. and Price, K. (1996). Minimizing the real functions of the icec'96 contest by differential evolution. In *Proc. of the 1996 IEEE International Conference on Evolutionary Computation*.

[34] Struzik, Z. R. (1999). *Fractals: Theory and Application in Engineering*, pages 93–112. Springer Verlag.

[35] Yuille, A. and Poggio, T. (1986). Scaling theorems for zero-crossing. *IEEE Transaction for zero-crossing*, 8(1):15–25.



Applied Biological Engineering - Principles and Practice

Edited by Dr. Ganesh R. Naik

ISBN 978-953-51-0412-4
Hard cover, 662 pages
Publisher InTech
Published online 23, March, 2012
Published in print edition March, 2012

Biological engineering is a field of engineering in which the emphasis is on life and life-sustaining systems. Biological engineering is an emerging discipline that encompasses engineering theory and practice connected to and derived from the science of biology. The most important trend in biological engineering is the dynamic range of scales at which biotechnology is now able to integrate with biological processes. An explosion in micro/nanoscale technology is allowing the manufacture of nanoparticles for drug delivery into cells, miniaturized implantable microsensors for medical diagnostics, and micro-engineered robots for on-board tissue repairs. This book aims to provide an updated overview of the recent developments in biological engineering from diverse aspects and various applications in clinical and experimental research.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen