A SEMI-AUTOMATIC MODELLING APPROACH FOR THE PRODUCTION AND FREEZE-DRYING OF LACTIC ACID BACTERIA

Thomas Chabin, Marc Barnabé, Alberto Tonda, Nadia Boukhelifa, Fernanda Fonseca
Eric Dugat-Bony, Hélène Velly, Evelyne Lutton, Nathalie Méjean Perrot
UMR 782 GMPA, INRA
1 av. Lucien Brétignières, 78850, Thiverval-Grignon, France
email: {firstname.lastname}@inra.fr

KEYWORDS
Complex Systems, Multiscale Modelling, Interactive Modelling, Optimisation, Expert Knowledge, Visualisation, Freeze-Drying of Lactic Acid Bacteria.

ABSTRACT

The production system of freeze-dried lactic acid bacteria involves several processes, but its impact on bacteria resistance is still not well understood. This system can be defined as a complex one since it depends on multiple scales: the Genomic, the Cellular and the Population scale. The scarcity of data available for building models leads us to propose an approach that makes use of expert knowledge. In this paper we present a semi-automatic modelling tool, LIDEOGRAM and discuss how it contributes to insight formulation and rapid hypothesis testing. New results show that LIDEOGRAM is able to produce more robust modelling hypotheses when experts can interact and revisit the genomic data preprocessing.

INTRODUCTION

Complex systems are found in many features of nature and science, ranging from the economic and social structure of a city to the global climate, or from the behaviour of a single cell to the behaviour of the intricate interactions of the human brain. Complex systems involve numerous components linked by non-trivial relationships, and are challenging to study. Modelling such complex systems, by summarising available knowledge into a mathematical or computational representation, is not a trivial task.

Concentrates of Lactic Acid Bacteria (LAB) are widely used in the food industry for manufacturing products such as yoghurt, cheese, fermented meat, vegetables and fruit beverages. The production of freeze-dried LAB is a complex food system due to its multi-scale and multi-step properties. One of the main challenges that need to be tackled ahead is to understand the origin of the LABs resistance and/or their sensitivity to the whole production system. Models describing parts of the system for a specific strain of bacteria are found in the literature (Passot et al. (2011)). However, to the best of our knowledge, no models have been proposed to represent the whole process.

Automatic modelling approaches have already been proposed for complex systems such as metabolic networks (Schmidt et al. (2011)), or for various multi-scale processes (Hasenauer et al. (2015)). These approaches often require a significant amount of data, however, gathering data on the freeze-drying process of LAB is expensive and time-consuming. Little amount of data is thus generally available, and this is a major issue for automatic modelling. To compensate for the lack of data, expert knowledge on the process can be exploited. We show in this paper how such knowledge can be integrated within a modelling process, based on a semi-automatic scheme.

The paper is organised as follows: First we present some background on complex systems, on semi-automatic modelling and on expert knowledge integration. The target system and the dataset are then described. Next we detail our semi-automatic modelling software and show some experimental results. Finally, results are discussed and conclusions are drawn.

BACKGROUND

Complex Systems in Biology and Food Systems: Expert Knowledge Integration Methods

A complex system is defined as a system made of multiple processes, entities, and nested subsystems. Global properties emerge through a series of phenomena occurring at different scales (Ladyman et al. (2013)). Appropriate descriptions with high expressiveness and little uncertainty of the underlying mechanisms is needed to elucidate such systems. Building models of complex systems is crucial, but highly difficult. It usually requires a robust framework, with strong iterative interaction combining computational intensive methods, formal reasoning and experts from different fields. In such context, optimisation plays an important role (Lutton et al. (2016)). Properties of food systems (such as uncertainty and variability, heterogeneity of data, coexistence of qualitative and quantitative information, conjunction of different perspectives) raise the focus on another es-

sential issue, that can be called the *human factor*. In order to gain a better understanding of food systems, human expertise and decision making are of major importance, and should thus be integrated into automatic modelling approaches (Lutton and Perrot (2015)). Numerous papers propose to take advantage of a structured prior knowledge of a system to improve machine learning methods. Among those, prior knowledge was used to improve the predictions of neural networks model of chemical systems (Thompson and Kramer (1994)), or to build a genetic network using Bayesian methods (Le et al. (2004)). Expert knowledge was also used by Baudrit et al. (2010) to improve the quality of a cheese ripening process model using Bayesian networks. In other studies, structured knowledge about the topic of interest is not well defined and exist more in the form of insights. In such cases, approaches relying on visual exploration of the data, and interactions through software have been proposed (Turkay et al. (2017), Cancino et al. (2012) and Krause et al. (2014). In this way, it is possible to confirm and elucidate new hypotheses.
In this context, this work aims at presenting in the following sections a new approach for modelling multi-scale systems, interactively and iteratively, through visual exploration, machine learning and knowledge integration.

PRODUCTION AND FREEZE-DRYING OF LACTIC ACID BACTERIA

Concentrates of LAB, also called starters, are food ingredients widely used for producing fermented meat, vegetables, fruit and dairy products. The commercialisation of these starters requires the application of successive operations: fermentation, concentration and preservation by freezing or freeze-drying (or lyophilisation). The viability and acidification activity of the cells are the two main quality attributes of the starters. They depend on many control parameters of the multi-steps process: Fermentation, Concentration, Formulation, Freeze-Drying and Storage, See Carvalho et al. (2004) for a detailed description.
The bacteria's levels of resistance to the processes is also dependent on the biochemical and biophysical properties and organisation of their membrane, which in turn is determined by the expression of the bacterial genome. This case study is based on the work of Velly et al. (2015) about the resistance of *Lactococcus lactis* subsp. *lactis TOMSC161* to freeze-drying. This strain, very sensitive to freeze-drying, is used for the manufacture of *Tomme de Savoie*, a French cheese, for its texturing and acidification characteristics. Several bacterial properties were measured for two fermentation temperatures (22 C and 30 C), and two cell growth phases (at the beginning of the stationary growth phase and 6 hours later). The dataset features 12 data points, corresponding to previous detailed four fermentation conditions and three biological repetitions of each experimental condition. Different scales were considered:

- Genomic: Transcriptomic data obtained on 2744 genes by RNA-seq,
- Cellular: Relative composition of main fatty acids present in the bacterial membrane determined by Gas Chromatography-Mass Spectrometry after extraction and the Anisotropy of the membrane (rigidity) assessed by flow cytometry,
- Population: Viability by numbering on agar plates and acidification activity in milk quantified using the CINAC system at the end of the following steps: Concentration, Freezing, Drying, and three months of Storage.

LIDEOGRAM

Experts of the domain seek answers about how a given bacterial strain becomes resistant to the process. Mathematical tools, including mathematical formulas are generally used to help them address these questions. But finding reliable formulas linking the different variables of such a system is indeed challenging. In biological data, repetitions of a given experimental condition are often highly variable. Moreover, experiments are usually time-consuming and expensive, resulting in few data being obtained, which makes the task of characterising the existing variability difficult.
LIDEOGRAM (***L****ife-based* ***I****nteractive* ***DE****velopment* ***O****f* ***GRA****phical* ***M****odels*) tries to address these challenges with an original approach based on semi-automatic modelling (See Chabin et al. (2017a;b) for details). The goal of LIDEOGRAM is to provide experts with a design tool for modelling their complex process. Each non-input variable is modelled by a mathematical formula involving other variables of the system. It is then possible to create a multi-scale model where each scale of the process is defined with variables of a lower scale and with experimental conditions. A global model is therefore a concatenation of mathematical equations that figure relationships between different variables at different scales. This is similar to other successful multi-scale modelling approaches such as for grape berries ripening by Dai et al. (2007) or for cow's milk production by Cros et al. (2003): the global model is made of stacked sub-models.
However, it is difficult to find the "right" equation in a context of high variability in the dataset. It is for instance frequent to come up with over-fitted equations that perfectly represent a dataset including its noise. In order to rule out over-fitted equations, our strategy is to involve experts in the course of the modelling process rather than splitting the (small) available dataset into training and test subsets. The expectation is that experts are able, thanks to their knowledge of the pro-

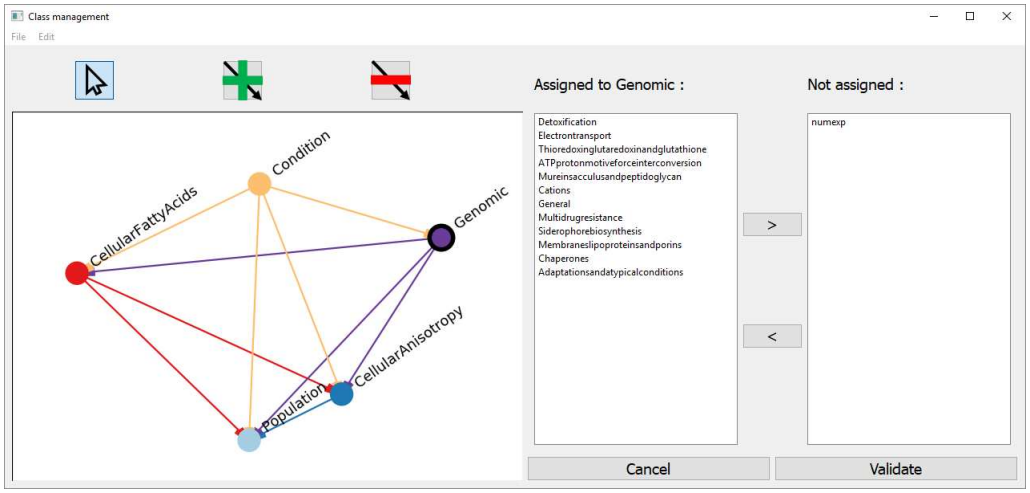

Figure 1: Screenshot of the interface where users choose the authorised links between the defined classes. A link between two classes means that all variables associated to the parent class can be involved in the equations for all variables of the child class. The displayed graph represents the selected constraints chosen for this experiment. The selected class here is the Genomic class (circled in black). The variables assigned to this class can be seen on the right side.

cess, to identify over-fitted, under-fitted or inappropriate equations.

Therefore, as a first optimisation step, LIDEOGRAM runs Orthogonal Matching Pursuit (OMP) on each variable. This technique introduced by Pati et al. (1993) is a linear regression that makes it possible to choose the number of predictive variables used in an equation. Using this approach, a set of candidate equations with different levels of complexity and fitting error is associated to each variable. Constraints can be defined beforehand by the user, using the interface presented in Figure 1. It makes it possible to attribute each variable to a given class of variables, and to authorise of forbid links between them, in the sense that only the variables from a parent class can be used for calculating the variables of the child class.

A qualitative view of the proposed equations is then presented to the user as a graphical network (See Figure 2).

The purpose of this view is to help the user identify the critical variables, where expert feedback is most needed. Nodes of the graph represent variables. Colours of nodes correspond to their attributed class. A link between two variables means that the parent node is used at least once in the set of equations of the child node. Colours of links correspond to a numerical value computed using all mathematical equations featuring the parent node in the child node. A green link corresponds to a good mean fitting of the data for the corresponding equations. Conversely, a red link represents a poor fitting. The network may be difficult to read, since the displayed graphical network can have a considerable amount of links. A slider filters the links based on a level of importance. This level of importance is defined for a link by the number of equations in the child node that use the parent node, divided by the total number of equations of the child node.

When a node is selected, the equations found by OMP are displayed on the top-right side (See Figure 2). Similarly, a click on an equation triggers the corresponding plot of experimental versus predicted data. The user can then interact with the system by deleting an equation, deleting a link between a parent node and a child node (i.e. all equations using the parent node in the child node are deleted), or deleting a variable (in this case all equations using the deleted variable are deleted). After this, few or no equations may remain for some nodes, the user can choose to restart the OMP on any node, with new constraints. After this, an evolutionary optimisation builds a global model by selecting one equation for each variable, thus taking into account the coherence between the scales.

USING EXPERT KNOWLEDGE VIA A SEMI-AUTOMATIC MODELLING SCHEME

The first step when starting LIDEOGRAM is to organise the set of variables into classes and to define define the possible links between classes. Five classes of variables, linked to the different scales of the process, were designed: Condition, Genomic, CellularFattyAcids, CellularAnisotropy and Population. The authorised links between the classes are presented in Figure 1.

A preprocessing of the variables at the Genomic scale was needed before running LIDEOGRAM on them. With 2744 genes measured by transcriptomics and with a high variance in the measurements, it is hard to explore and make sense of the function of each individual gene in a model. For this purpose, two solutions

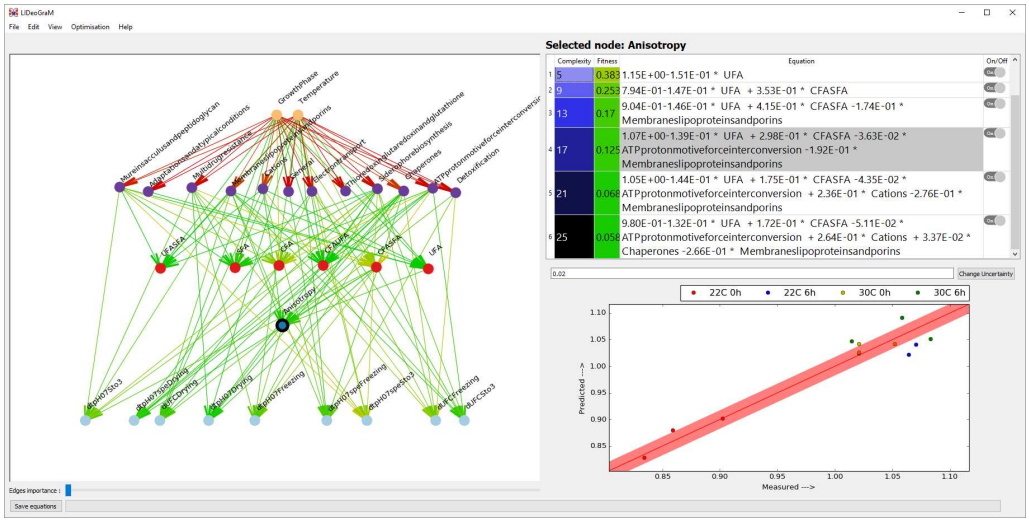

Figure 2: Screenshot of LIDEOGRAM. Left: graphical model representing the mean fitness for all local models. Top-right: list of equations proposed for the selected node (Anisotropy); Bottom-right: plot of the measured versus predicted data corresponding to the selected equation.

were explored using human expertise. The first pre-processing (PP1) is a classification proposed by Bolotin et al. (2001). The 82 functional classes of genes were reviewed by the experts in order to keep the most relevant ones. Twelve functional classes were selected at the genomic level. For each functional class, the sum of the genes expressions corresponding to this functional class is computed. Following experts advices, a second pre-processing (PP2) was performed to select only the genes showing a strong differential expression with respect to the conditions of fermentation. The selection criterion, computed for each gene, is the variance calculated on the mean expression for each condition, divided by the maximum of the variances calculated for each fermentation condition. Only the genes with a criterion larger than 2 were kept, which yielded 26 genes.
LIDEOGRAM was used to rapidly access the best way to pre-process the genomic data for the modelling step. Results are reported in Table 1, where 0 represents the best possible prediction and 1 the worst one.

Table 1: Experimental results on the local and global models using PP1 and PP2.

	PP1	**PP2**
Mean of the Local models at the Genomic Scale	All variables > 0.7	All variables < 0.5
Global Fitness on 10 runs	0.702 var. 1.4×10^{-4}	0.283 var. 5.0×10^{-5}

From these results, it seems that the variance criterion approach (PP2) for the Genomic scale performs better, both in terms of local and global model. These two hypotheses were explored in less than an hour, giving users a convenient and versatile way to test various modelling hypotheses. The equations proposed for the Anisotropy were also explored, using PP1 (shown in top-right side of Figure 2). Each equation was evaluated by the experts, who stated that the first three equations were compatible with their knowledge. Indeed, it was shown by Velly et al. (2015) that the anisotropy associated to membrane rigidity is anti-correlated with the Unsatturated Fatty Acids (UFA), but correlated to the ratio between Cyclic Fatty Acids and Saturated Fatty Acids (CFASFA). In the third equation, it is proposed that the anisotropy is also anti-correlated with the expression of genes that are associated to the membrane's lipoproteins and porins. According to our experts, this could be explained by the localisation of such proteins anchored (lipoproteins) to the membrane and crossing it (porins), respectively. One can thus hypothesise that their interactions with membrane components such as fatty acids could possibly modify membrane physical properties with a lowering effect on anisotropy (rigidity). Finally, some of the proposed equations were hard to validate. Our hypothesis is that these are over-fitted equations which may not represent the underlying reality. Such equations can then be discarded by the expert.

DISCUSSION AND CONCLUSION

We have presented above a semi-automatic approach for multi-scale modelling, that relies on both expert knowledge and automatic optimisation. LIDEOGRAM lets domain experts easily test various modelling hypotheses for multi-scale systems. Tests have been made on a production system for freeze-dried Lactic Acid Bacteria. LIDEOGRAM is also used for other applications, for ex-

ample, for modelling a cheese ripening ecosystem, and for modelling a grape berry maturity prediction system. Future improvements will be focused on the graphical interface and the various views: new interaction techniques will be proposed for analysing and modifying the proposed models. Indeed, it has been suggested that providing a variety of views and interactions (including change of focus of interest) is an excellent manner to engage users, reduce their fatigue and boost their creativity (Lutton et al. (2003), Boukhelifa et al. (2016)). Non-linear local models will be implemented in a future version of the software. We actually expect a better accuracy with the proposed models. Previous attempts using non-linear models provided promising results, however technical difficulties make the full use of a non-linear modelling approach more complex. Pareto front exploration of the local models is another important issue that will be developed in a further work.

REFERENCES

Baudrit C.; Sicard M.; Wuillemin P.H.; and Perrot N., 2010. *Towards a global modelling of the Camembert-type cheese ripening process by coupling heterogeneous knowledge with dynamic Bayesian networks. Journal of Food Engineering*, 98, no. 3, 283–293.

Bolotin A.; Wincker P.; Mauger S.; Jaillon O.; Malarme K.; Weissenbach J.; Ehrlich S.D.; and Sorokin A., 2001. *The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403. Genome research*, 11, no. 5, 731–753.

Boukhelifa N.; Bezerianos A.; Tonda A.; and Lutton E., 2016. *Research prospects in the design and evaluation of interactive evolutionary systems for art and science.* In *CHI workshop on Human Centred Machine Learning*. np.

Cancino W.; Boukhelifa N.; and Lutton E., 2012. *Evographdice: Interactive evolution for visual analytics.* In *Evolutionary Computation (CEC), 2012 IEEE Congress on*. IEEE, 1–8.

Carvalho A.S.; Silva J.; Ho P.; Teixeira P.; Malcata F.X.; and Gibbs P., 2004. *Relevant factors for the preparation of freeze-dried lactic acid bacteria. International Dairy Journal*, 14, no. 10, 835–847.

Chabin T.; Barnabe M.; Boukhelifa N.; Fonseca F.; Tonda A.; Velly H.; Lemaitre B.; Perrot N.; and Lutton E., 2017a. *LIDeOGraM: an interactive evolutionary modelling tool.* In *Biennial International Conference on Artificial Evolution (EA-2017)*. np.

Chabin T.; Barnabe M.; Boukhelifa N.; Fonseca F.; Tonda A.; Velly H.; Perrot N.; and Lutton E., 2017b. *Interactive evolutionary modelling of living complex food systems: freeze-drying of lactic acid bacteria.* In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. ACM, 267–268.

Cros M.J.; Duru M.; Garcia F.; and Martin-Clouaire R., 2003. *A biophysical dairy farm model to evaluate rotational grazing management strategies. Agronomie*, 23, no. 2, 105–122.

Dai Z.W.; Vivin P.; and Génard M., 2007. *Modelling the effects of leaf-to-fruit ratio on dry and fresh mass accumulation in ripening grape berries.* In *VIII International Symposium on Modelling in Fruit Research and Orchard Management 803*. 283–292.

Hasenauer J.; Jagiella N.; Hross S.; and Theis F.J., 2015. *Data-driven modelling of biological multi-scale processes. Journal of Coupled Systems and Multiscale Dynamics*, 3, no. 2, 101–121.

Krause J.; Perer A.; and Bertini E., 2014. *INFUSE: interactive feature selection for predictive modeling of high dimensional data. IEEE transactions on visualization and computer graphics*, 20, no. 12, 1614–1623.

Ladyman J.; Lambert J.; and Wiesner K., 2013. *What is a complex system? European Journal for Philosophy of Science*, 3, no. 1, 33–67.

Le P.P.; Bahl A.; and Ungar L.H., 2004. *Using prior knowledge to improve genetic network reconstruction from microarray data. In silico biology*, 4, no. 3, 335–353.

Lutton E.; Cayla E.; and Chapuis J., 2003. *ArtiE-fract: The artists viewpoint.* In *Workshops on Applications of Evolutionary Computation*. Springer, 510–521.

Lutton E. and Perrot N., 2015. *Complex systems in food science: Human factor issues. 6th International Symposium on Delivery of Functionality in Complex Food Systems Physically-Inspired Approaches from the Nanoscale to the Microscale.*

Lutton E.; Perrot N.; and Tonda A., 2016. *Evolutionary Algorithms for Food Science and Technology.* John Wiley & Sons.

Passot S.; Fonseca F.; Cenard S.; Douania I.; and Trelea I.C., 2011. *Quality degradation of lactic acid bacteria during the freeze drying process: Experimental study and mathematical modelling.*

Pati Y.C.; Rezaiifar R.; and Krishnaprasad P.S., 1993. *Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition.* In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 40–44.

Schmidt M.D.; Vallabhajosyula R.R.; Jenkins J.W.; Hood J.E.; Soni A.S.; Wikswo J.P.; and Lipson H., 2011. *Automated refinement and inference of analytical models for metabolic networks. Physical biology*, 8, no. 5, 055011.

Thompson M.L. and Kramer M.A., 1994. *Modeling chemical processes using prior knowledge and neural networks. AIChE Journal*, 40, no. 8, 1328–1340.

Turkay C.; Slingsby A.; Lahtinen K.; Butt S.; and Dykes J., 2017. *Supporting theoretically-grounded model building in the social sciences through interactive visualisation. Neurocomputing.*

Velly H.; Bouix M.; Passot S.; Penicaud C.; Beinsteiner H.; Ghorbal S.; Lieben P.; and Fonseca F., 2015. *Cyclopropanation of unsaturated fatty acids and membrane rigidification improve the freeze-drying resistance of Lactococcus lactis subsp. lactis TOMSC161. Applied microbiology and biotechnology*, 99, no. 2, 907–918.