Interactive evolutionary modelling of living complex food systems: freeze-drying of lactic acid bacteria

T. Chabin, M. Barnabé, N. Boukhelifa, F. Fonseca, A. Tonda, H. Velly, N. Perrot, E. Lutton
UMR 782 GMPA - Agroparistech, INRA, Université Paris-Saclay
Thiverval-Grignon F-78850, France

ABSTRACT

Modelling the production and stabilisation process of lactic acid starters has several practical applications, ranging from assessing the efficacy of new industrial methods, to proposing alternative sustainable systems of food production. In order to reach this objective, however, it is necessary to overcome several obstacles, tied to the complex nature and interactions of the target processes. In this paper, we present a novel complex system modelling approach, exploiting both stand-alone evolutionary search and visual interaction with the user. The presented framework is then tested on a real-world case study, for which it shows promising results.

KEYWORDS

Complex systems, Lactic acid bacteria, Interactive modelling, Symbolic regression, Living food system

ACM Reference format:
T. Chabin, M. Barnabé, N. Boukhelifa, F. Fonseca, A. Tonda, H. Velly, N. Perrot, E. Lutton . 2017. Interactive evolutionary modelling of living complex food systems: freeze-drying of lactic acid bacteria. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3075992

1 INTRODUCTION

Agri-food processes can be regarded as complex systems, as they are characterised by uncertain and intricate interaction effects between physical, chemical, and biological components.[1] A first difficulty to address when modelling such processes is the availability of experimental data at the different scales of interest. Data may be sparse and uncertain as well as high dimensional. A second feature of the domain concerns the importance of expert knowledge in the modelling process.[3] Building a model in these conditions is a complex optimisation where experts knowledge can drastically modify the shape of the search space, the relative impact of data, or even the optimisation aims. We propose here an interactive modelling approach based on a two-level evolutionary optimisation scheme, which correspond to what we call our *local* and *global* models. Users can interact with the constructed models via a graphical user interface (GUI), run various optimisation steps, revisit optimisation results, restart the process, add constraints, and take decisions.

GECCO '17 Companion, Berlin, Germany

DOI: http://dx.doi.org/10.1145/3067695.3075992

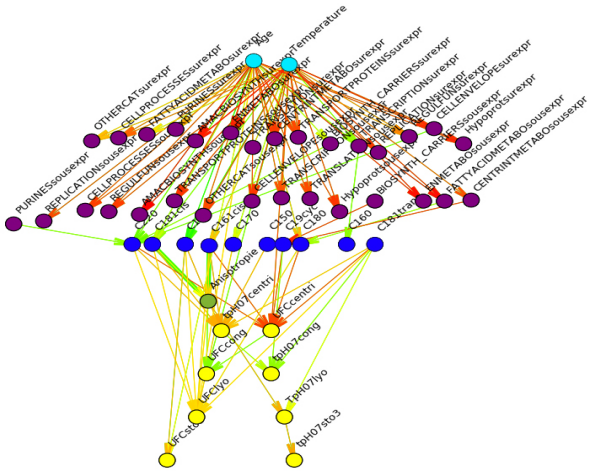

Figure 1: Graphical model representing the mean fitness of the local models obtained by symbolic regression. Node color correspond to classes.

2 PROPOSED APPROACH

LIDeOGraM (***L**ife-based **I**nteractive **De**velopment **O**f **Gra**phical **M**odels*) implements an original approach of semi-automatic modelling of biological complex systems. Its goal is to help domain experts (biologists) build a global model by characterising each non-input variable by a mathematical formula that involves other variables of the system. Finding the right equation in a context with high variability in the dataset is an ambitious task since it is easy to come up with over-fitted equations. A solution to rule out these equations is to involve experts in the modelling process. Symbolic regression using a Pareto-like approach such as the one implemented in Eureqa[2], constitutes a compelling approach to take advantage of the expert's insight. Indeed, by providing a set of formulas according to different compromises between fitness and complexity, the approach allows the experts to filter out incoherent equations or even designate the most suitable one. Therefore, as a first optimisation step, LIDeOGraM runs Eureqa on each variable using user-predefined constraints in the search : Each variable is attributed to a given class, so that dependencies will be searched only with variables of other classes (no intra-class dependencies). A qualitative view of these results is presented to the user in the form of a graphical network (See Figure 1).

The goal of this display, is to help the user focus on the variables that need the most the expert's feedback. In this prospect, variables are represented as nodes in the graph. The colour of a link represents the mean value of either the fitness or the complexity of the equations involving the parent node in the child

node. By clicking on a node, the equations found by Eureqa are displayed to the user and in order to get a better idea about the quality of an equation, the user can also click on it, and have a plot of the experimental measures versus what is predicted by the equation. The user can then act on these results 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), deleting a variable (i.e. all the equations using the deleted variable are deleted), forcing a specific link to be present in the global model, deleting all the equations in the child node that do not contain the parent node. After adding these constraints, few or no equations might still be available for some nodes. In order to get more equations, the user can choose to restart a symbolic regression on any node. The user can iterate the process for as long as desired, redefine classes, add constraints, and restart symbolic regression on nodes. When he is satisfied with the local models, a computation of a global model can be triggered. A global model is derived from the set of local models by selecting one equation only at each node. Finding a global model is a complex problem, because the value predicted by a an equation depends on the value predicted by its parent variables. The global model is thus built using an evolutionary optimisation process. The fitness function, to be minimised for the global model is the mean value of the fitness calculated for all non-input nodes. The fitness function of a single node computes a value based on the Pearson correlation coefficient using the measured and the predicted data considered in a 2D-space. After evolution, remaining incoherent choice of equation can be edited by the user. The user can go back to the local view, change local models and restart a new global optimisation. A global model is thus iteratively built via user interaction, based on chained local and global optimisations.

3 EXPERIMENT AND CONCLUSIONS

The case study is based on the work of H. Velly et al. [4], on the resistance of *Lactococcus lactis* subsp. *lactis TOMSC161* to freeze-drying. The resistance of the bacteria is studied for 4 different conditions of fermentation: 22°C and 30°C, evaluated at the beginning of the stationary growth phase and 6 hours later. The dataset featured 12 data points, with 3 biological repetitions of each experimental condition. The dataset is made of 2 input variables, the temperature of fermentation and the time at which the fermentation is stopped and 49 variables measured at 4 different steps (fermentation, concentration, freeze-drying and storage) for 3 biological scales (Genomic, Cellular and Population).

A (μ +λ)-evolutionary algorithm is used to optimise the global model. The genome of a candidate global model is a string of integers, of size equal to the number of variables in the process. Each gene is associated to a variable, and can assume a value between 1 and the number of equations available to describe that variable, thus representing an index for a candidate equation in that node. The parameters of the evolutionary optimization algorithm used for the global model are as follows : μ : 100, λ: 80, Number of generations: 100, Probability of crossover: 0.8, Probability of mutation: 0.2, Selection: Tournament of size 2, Crossover function: Uniform, Mutation function: With a probability 0.05 for each gene, change the selected equation to the previous or the next complex one. Feedback on the proposed local models was given by an expert researcher with 20 years of experience in the field of the bacteria freeze-drying process. The local models, were explored by the expert during 20 minutes. The expert chose to remove 5 equations out of a total of 232. and 2 nodes. The deletion of those two variables removed 14 more equations. With such major deletions, some variables were left with only a few equations, therefore, the expert chose to restart a symbolic regression on 3 nodes, obtaining 12 new equations in total. To reveal the contribution of the expert, the global model optimisation was performed 10 times using expertise, and 10 times without. The fitness evolution of these runs are shown in Figure 2. To obtain an accurate comparison of the models, the fitness computed for optimisation without the expertise did not take into account the two removed nodes. The global models obtained using expertise have a median fitness of 0.787 with a standard deviation of 0.010 whereas the global models obtained without expertise have a median fitness of 0.801 with a standard deviation of 0.013. The expert was asked to provide

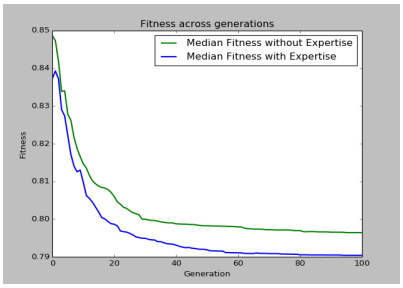

Figure 2: Comparison of the evolution of the best fitness across generations for 10 runs

feedback for the last step of the modelling process in which one of the global model obtained was submitted to his expertise. The results were explored during 10 minutes, and the equations for three node were changed. The fitness of the final global model was slightly degraded, changing from a fitness of 0.789 to a fitness of 0.801, but the produced model is able to better reflect the underlying reality of the process.

We proposed a time-saving tool of modelling for the experts, allowing them to design a better global model of their process by a semi-interactive approach. Figure 2 shows that the resulting models are "better", not only according to the expert requirements, but also with respect to the numerical data (faster and better convergence).

REFERENCES

[1] Nathalie Perrot, Ioan-Cristian Trelea, Cédric Baudrit, Gilles Trystram, and P Bourgine. 2011. Modelling and analysis of complex food systems: state of the art and new trends. *Trends in Food Science & Technology* 22, 6 (2011), 304–314.

[2] Michael Schmidt and Hod Lipson. 2009. Distilling free-form natural laws from experimental data. *Science* 324, 5923 (2009), 81–85.

[3] Mariette Sicard, Cédric Baudrit, MN Leclerc-Perlat, Pierre-Henri Wuillemin, and Nathalie Perrot. 2011. Expert knowledge integration to model complex food processes. Application on the camembert cheese ripening process. *Expert Systems with Applications* 38, 9 (2011), 11804–11812.

[4] H Velly, M Bouix, S Passot, C Penicaud, H Beinsteiner, S Ghorbal, P Lieben, and F Fonseca. 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, 2 (2015), 907–918.