Innovative Food Science and Emerging Technologies 25 (2014) 67–77

Contents lists available at ScienceDirect

Innovative Food Science and Emerging Technologies

journal homepage: www.elsevier.com/locate/ifset

Food model exploration through evolutionary optimisation coupled with visualisation: Application to the prediction of a milk gel structure

Evelyne Lutton [a,*], Alberto Tonda [a], Sébastien Gaucel [a], Alain Riaublanc [b], Nathalie Perrot [a]

[a] UMR 782 Génie et Microbiologie des Procédés Alimentaires, AgroParisTech, INRA, 78850 Thiverval-Grignon, France
[b] INRA, Rue de la Géraudière, BP 71627, 44316 Nantes Cedex 3, France

ARTICLE INFO

Article history:
Received 22 June 2013
Accepted 1 February 2014
Available online 11 February 2014

Keywords:
CMA-ES
Evolutionary algorithms
Milk gel model
NSGA-II
Optimisation
Visualisation

ABSTRACT

Obtaining reliable in-silico food models is fundamental for a better understanding of these systems. The complex phenomena involved in these real-world processes reflect in the intricate structure of models, so that thoroughly exploring their behaviour and, for example, finding meaningful correlations between variables, become a relevant challenge for the experts. In this paper, we present a methodology based on visualisation and evolutionary computation to assist experts during model exploration. The proposed approach is tested on an established model of milk gel structures, and we show how experts are eventually able to find a correlation between two parameters, previously considered independent. Reverse-engineering the final outcome, the emergence of such a pattern is proved by physical laws underlying the oil–water interface colonisation. It is interesting to notice that, while the present work is focused on milk gel modelling, the proposed methodology can be straightforwardly generalised to other complex physical phenomena.

Industrial relevance: Sustainability is nowadays at the heart of industrial requirements. The development of mathematical approaches should facilitate common approaches to risk/benefit assessment and nutritional quality in food research and industry. These models will enhance knowledge on process–structure–property relationships from the molecular to macroscopic level, and facilitate the creation of in-silico simulators with functional and nutritional properties. The stochastic optimisation techniques (evolutionary algorithms) employed in these works allow the users to thoroughly explore the systems: when coupled with visualisation, they make it possible to provide the experts with a restricted set of significant data, helping them to highlight eventual issues or possible improvements in the model. With regard to the complexity of the food systems and dynamics, the challenge of the mathematical approaches is to realise a complete dynamic description of food processing. In order to reach this objective, it is mandatory to use innovative strategies, exploiting the most recent advances in cognitive and complex system sciences.

Glossary

Name	Description	Unit
m_p	Total mass of proteins in the solution (constant)	g
m_{wp}	Mass of native whey proteins in the solution	g
m_{cas}	Mass of casein micelles in the solution	g
S_0	Initial lipid surface	m^2
S_{fall}	Lipid surface available for adsorption of both native whey proteins and casein micelles	m^2
S_{fres}	Lipid surface left by casein micelles due to steric effects for native whey proteins	m^2
k_{wp}	Adsorption rate of native whey proteins	s^{-1}
k_{cas}	Adsorption rate of casein micelles	s^{-1}

(*continued*)

Name	Description	Unit
s_{wp}	Surface area occupied by 1 g of native whey proteins	$m^2 \cdot g^{-1}$
s_{cas}	Surface area occupied by 1 g of casein micelles	$m^2 \cdot g^{-1}$
α	Fraction of the adsorbed surface of a casein micelle reserved for native whey proteins	Dimensionless
w_{wp} (0)	Initial mass percentage of native whey proteins in the solution, $w_{wp}\ (0) = m_{wp}\ (0)/m_p\ (0)$	%
w_{cas} (0)	Initial mass percentage of casein micelles in the solution, $w_{cas}\ (0) = m_{cas}\ (0)/m_p\ (0)$	%
$w_{wp_{int}}$	Final mass percentage of native whey proteins at lipid interface relative to the total mass of adsorbed proteins	%
$w_{cas_{int}}$	Final mass percentage of casein micelles at lipid interface relative to the total mass of adsorbed proteins	%
Γ	Final interfacial concentration which corresponds to the quantity of adsorbed proteins per 1 m^2 of lipid surface	$mg \cdot m^{-2}$
$d_{3.2}$	Average diameter of lipid droplet	m
ρ_l	Lipid density	$g \cdot m^{-3}$
m_l	Mass of lipid (constant)	g
μ	Population size parameter for the evolutionary algorithms used in the experience	Dimensionless

(*continued on next page*)

* Corresponding author.
E-mail addresses: evelyne.lutton@grignon.inra.fr (E. Lutton), alberto.tonda@grignon.inra.fr (A. Tonda), sebastien.gaucel@grignon.inra.fr (S. Gaucel), alain.riaublanc@nantes.inra.fr (A. Riaublanc), nathalie.perrot@grignon.inra.fr (N. Perrot).

1466-8564/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ifset.2014.02.003

(*continued*)

Name	Description	Unit
λ	Offspring size parameter for the evolutionary algorithms used in the experience	Dimensionless
η_{operator}	Distribution index for a genetic operator in the evolutionary algorithm NSGA-II	Dimensionless

1. Introduction

Building in-silico models for food processes is an important but difficult task, as there exist various known bottlenecks (Perrot, Trelea, Baudrit, Trystram, & Bourgine, 2011). The process of model design, for instance, often relies on computationally expensive optimisations to match a theoretical model with available data (parameter learning). Scarcity of data is a classical source of troubles for the optimisation process, resulting in badly conditioned problems. Solutions provided by optimisation cannot be exploited directly and must be revisited by experts, in order to disambiguate equivalent sets of solutions. Facilitating a high level expert analysis of computational results, or even more, interaction of expert knowledge with computational processes, is a challenging task: interactive optimisation is an active field of research (Takagi, 1998), and its potential applications in the domain of food process modelling are numerous.

Optimisation tools are actually often used in a "black box" manner, and computational optimal results may then yield imprecise, ambiguous or even wrong parameter setting. In this paper, we present a methodology based on evolutionary algorithms (also know as "genetic algorithms"). Their iterative, population-based, algorithmic structure, if appropriately exploited, allows the highlighting of various features of the search space, which correspond to possible pathologies of the considered model. Experts may of course have access to these pathological features via appropriate theoretical analysis, as soon as they know what to search for. As we will see below, the observation of the successive population distributions of the evolutionary algorithms allows as to get some intuitions about the possible degeneracies of the searched model, thus making the task of the expert easier. We exemplify this approach on a complex test case, the prediction of the structure of a milk gel.

Table 1
Milk gel data used for training (database 1, L_1 to L_7) and validation (database 2, V_1 to V_4).

Sample	w_{cas} (0) (%)	$d_{3.2}$ (μm)	m_p (0) (g)	$w_{cas_{int}}$ (%)	Γ (mg·m^{-2})	Database
L_1	13	0.6	2.47	5	5.6	1
L_2	19	0.7	2.44	9	4.8	1
L_3	21	0.6	2.42	16	4.0	1
L_4	26	0.65	2.40	43	4.9	1
L_5	32	0.55	2.40	65	5.6	1
L_6	49	0.56	2.39	71	4.2	1
L_7	80	0.9	2.37	84	9.3	1
V_1	13	0.59	8.79	0	4.66	2
V_2	22	0.74	8.47	33	4.4	2
V_3	31	0.87	8.64	46	6.88	2
V_4	80	0.75	9.18	91	6.93	2

Evolutionary algorithms (EAs) are the generic name for a large set of techniques that rely on the computer simulation of natural evolution mechanisms (Artificial Darwinism). Since pioneering works in the second half of 20th century (Bremermann, 1962; Fraser, 1957; Holland, 1962; Rechenberg, 1973), Artificial Darwinism techniques have progressively gained importance in stochastic optimisation and artificial intelligence domains for the resolution of difficult optimisation problems, and particularly for learning the optimal parameters of complex models (Bäck & Schwefel, 1993).

The main idea of EAs is to copy, in a very rough manner, the principles of natural evolution, that let a population be adapted to his environment. According to Darwin's theory (Darwin, 1859), adaptation is based on very simple mechanisms: random variations, inheritance, and survival/reproduction of the fittest individuals. Transposed into optimisation algorithms, this scheme has the major advantage of making few assumptions on the function to be optimised (there is no need to have a continuous or derivable function for instance). In short, evolutionary algorithms consider a population of potential solutions exactly as a population of individuals of a natural population that live, fight and reproduce. The environment pressure is replaced by an "optimisation" pressure: the function to be optimised is considered as a measurement of the adaptation of the individual to its environment (fitness). In this way, individuals that reproduce are the best ones with respect to the problem to be solved, and reproduction consists in generating new solutions via genetic operators (called crossover and mutation by analogy to nature).

Evolutionary optimisation techniques are particularly well suited to difficult problems, where classical methods fail. The major reason of this success is the tunable combination of oriented and random search mechanisms that allow injecting a priori, incomplete informations in the genetic operators, while letting some other more unpredictable components be randomly searched.

Considering evolutionary optimisation as a "black box", however, is not a good strategy in general. The first reason is that one may lose the opportunity to adapt the mechanisms to the specifics of the problem, which usually improves the efficiency of the algorithms and reduces its computation time. Another reason is related to the internal mechanisms of the algorithm that performs a sampling of the search

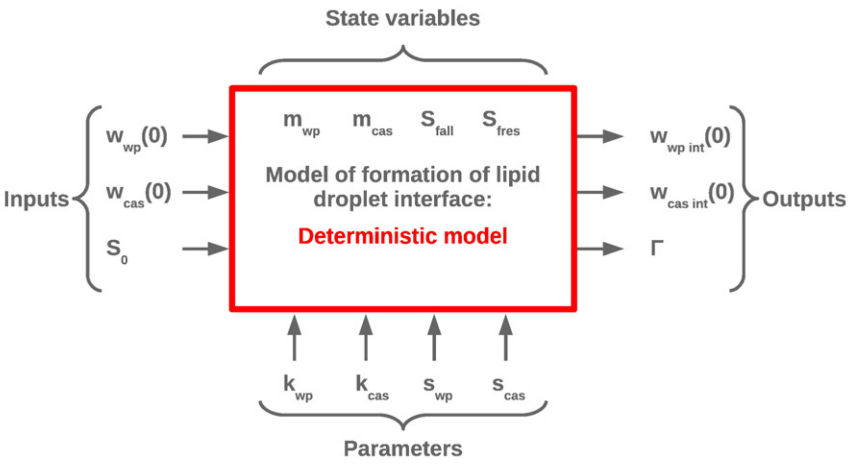


Fig. 1. Model of milk gel formation.

Table 2
Intervals of validity for each parameter in the considered optimisation problem. Parameters' values are obtained from literature and expertise on the subject.

Parameter	Minimum	Maximum
k_{wp}	0	100
k_{cas}	0	100
s_{wp}	0	1500
s_{cas}	0	300
α	0	1

Table 3
Parameters of the two EAs used during the experience. μ is the size of the population, λ is the size of the offspring generated at each iteration. While NSGA-II is terminated after 100 iterations (or generations), CMA-ES stops when a stagnation condition is reached: in particular, when the difference in fitness value between all solutions in the population is under a user-defined threshold. For CMA-ES, initial points in the middle of the search space are specified, for each dimension, and initial standard deviation to generate solutions is set; the algorithm will self-adapt the standard deviation during the run. For NSGA-II, P(operator) represents the probability of applying a specific genetic operator when a new solution is produced. η_{operator} is the distribution index of the genetic operator, regulating how much the children solutions will differ from the parents.

CMA-ES		NSGA-II	
Parameter	Value	Parameter	Value
Objective	Minimise ($f0 * f1$)	Objective 1	Minimise($f0$)
		Objective 2	Minimise($f1$)
Stop condition	Stagnation (10^{-12})	Stop condition	100 generations
μ	250	μ	500
λ	500	λ	500
Initial points	0.5	P(crossover)	0.9
Initial standard deviations	0.3	P(mutation)	$\frac{1}{\text{problem-dimension}}$
		$\eta_{\text{crossover}}$	20
		η_{mutation}	20

space via the evolution of its population. Observing how the population is distributed, then concentrated along generations; how diversity is lost or how it persists; and what is the appearance of the optimal set of solutions (a point or a significant subset): all these factors provide important information about the nature of the optimisation problem. In the case of model learning, this analysis makes it possible, for instance, to know if the learning set is large and discriminative enough.

Classical uses of EAs only consider the best individual of the last population as an estimation of the optimum, and do not exploit all useful informations provided by the algorithm, that may for instance help to assess if the optimum is correct and robust. A recent work points out the potential benefit of visualising data collected during the execution of an EA (Lutton & Fekete, 2011; Lutton, Foucquier, Perrot, Louchet, & Fekete, 2011), and shows how a multidimensional visualisation tool, GraphDice (Bezerianos, Chevalier, Dragicevic, Elmqvist, & Fekete, 2010), can help to efficiently navigate inside the data set collected during the execution of an EA. In this paper, we follow the previously described research line, and develop the proposed methodology for the specific case of a milk gel model.

Oil-in-water emulsions are dispersed systems stabilised by surface-active molecules, including proteins, polymers, and ionic and non-ionic surfactants (Dickinson, 2011). Proteins, as the main emulsifier in food systems, adsorb to the freshly formed interface of oil droplets created during homogenisation. They stabilise the emulsion thanks to their ability to generate repulsive interactions (steric and electrostatic) between oil droplets (McClements, 2004). Milk proteins have excellent emulsifying properties and are one of the most convenient ingredients used in food processing (Dickinson, 1999; Surel et al., 2014). Recent researches on milk gels (Dickinson, 2001, 2011; Gaygadzhiev, Hill, & Corredig, 2009; Knudsen, Ogendal, & Skibsted, 2008; Murray, 2002) highlight the major role of nanoscopic and microscopic dynamics during interface stabilisation on the qualitative characteristics of the gel, both macroscopic and nutritional. But if the structural characteristics of pure protein aggregates submitted to heat treatment are widely studied (Rabe, Verdes, & Seeger, 2011), research on aggregates of casein coupled to whey proteins (denatured or not) is still in the initial stages (Morand, Dekkari, Guyomarc'h, & Famelart, 2012): the interpretation of the surface composition in emulsions containing the full range of aggregated milk proteins (caseins and whey proteins) is quite complex and certainly not yet fully resolved. And if experimental data about the phenomena are collected, they are still rarely exploited in modelling approaches.

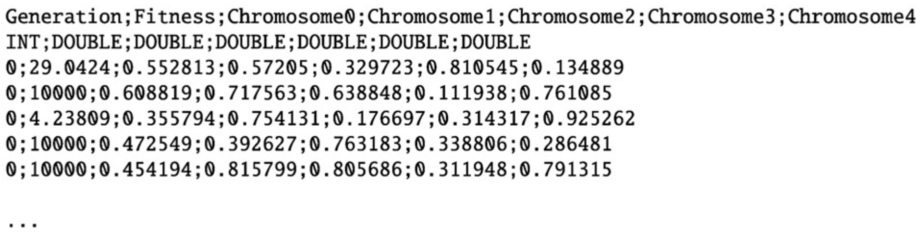
Among the main research lines on milk gel, an important part is devoted to the development of models with the ability to replicate the dynamics of gel formation at relevant scales, linking the structure to macroscopic properties. These models aim thus at including all levels and correctly predicting the complex interactions between elements at different scales (Erni, Windhab, & Fischer, 2011). For instance, the colonisation of the lipid droplet interface at the nanoscopic and microscopic scales (from the size of the surfactants like whey proteins with a diameter of around 3 nm, to micelles and aggregates with a diameter of around 100 nm) influences the formation of milk polymers: to predict the properties of the product at a macroscopic level, such as consistency, some knowledge on this process is directly needed. Although modelling approaches start to address this multi-scale reconstruction problem (Foucquier et al., 2012), there is considerable space for improvements. To cope with the scarcity of data, we propose an approach combining computational exploration, based on an evolutionary algorithm, with visualisations of the results.

The paper is organised as follows: Section 2 first describes the experimental dataset, the model and the evolutionary algorithm used for parameter learning. Section 3 details how the visualisation of the data collected along the execution of the EA allows to draw important conclusions on some model parameters. A feedback into the equations of the initial model allows a revision of the evolutionary algorithm structure to better address the parameter learning problem (Section 3.3). Finally, Section 4 draws some conclusions about the potential generalisations of the method.

2. Materials and methods

2.1. Experimental data

Data has been collected during two emulsification experiences, where the continuous phase of the emulsion is formed by dissolving



```
Generation;Fitness;Chromosome0;Chromosome1;Chromosome2;Chromosome3;Chromosome4
INT;DOUBLE;DOUBLE;DOUBLE;DOUBLE;DOUBLE;DOUBLE
0;29.0424;0.552813;0.57205;0.329723;0.810545;0.134889
0;10000;0.608819;0.717563;0.638848;0.111938;0.761085
0;4.23809;0.355794;0.754131;0.176697;0.314317;0.925262
0;10000;0.472549;0.392627;0.763183;0.338806;0.286481
0;10000;0.454194;0.815799;0.805686;0.311948;0.791315

...
```

Fig. 2. A simple .csv file collected during a run of the single-objective EA (CMA-ES). Chromosome0; Chromosome1; Chromosome2; Chromosome3; and Chromosome4 are respectively k_{wp}, k_{cas}, s_{wp}, s_{cas} and α.

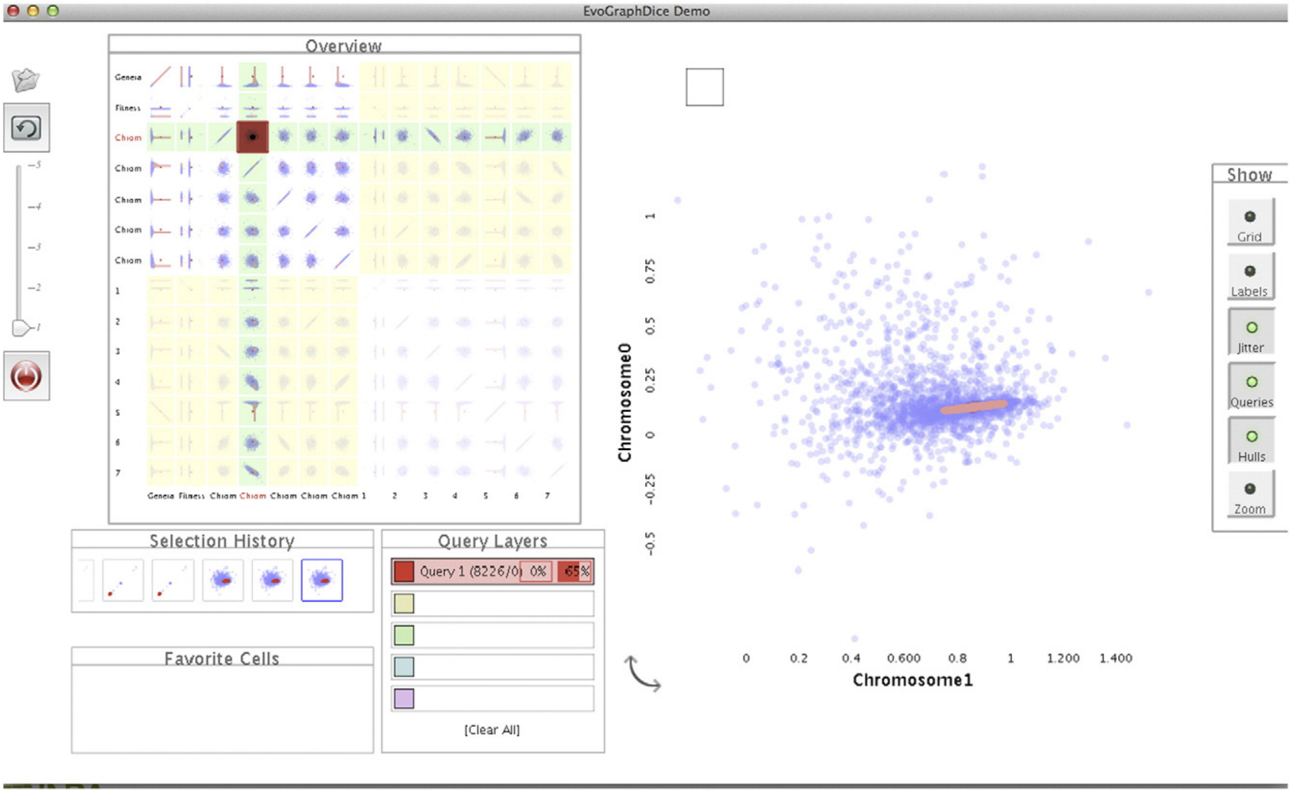


Fig. 3. Main window of EvoGraphDice (yellow column and lines numbered 1 to 7 are additional dimensions computed by the system for the purpose of analysis) for the visualisation of data collected during a run of CMA-ES searching the 5 parameter space. Red points correspond to best fitness values: they are obviously distributed along a line.

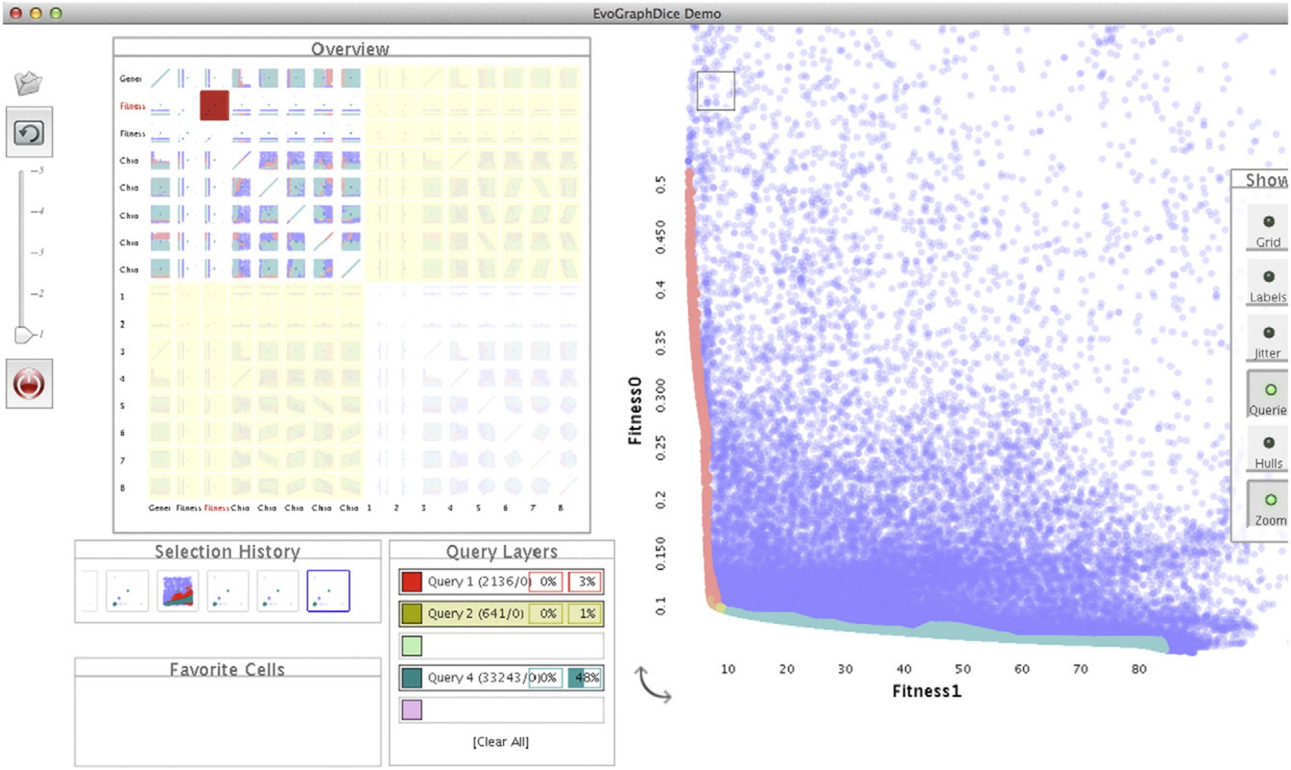


Fig. 4. Visualisation of a multiobjective run (NSGA-II) in the 5 parameter space. Pareto front (right): preference of fitness0 over fitness1 in green, and reversely in red, and equivalent compromise in yellow. Blue points are non-Pareto optimal.

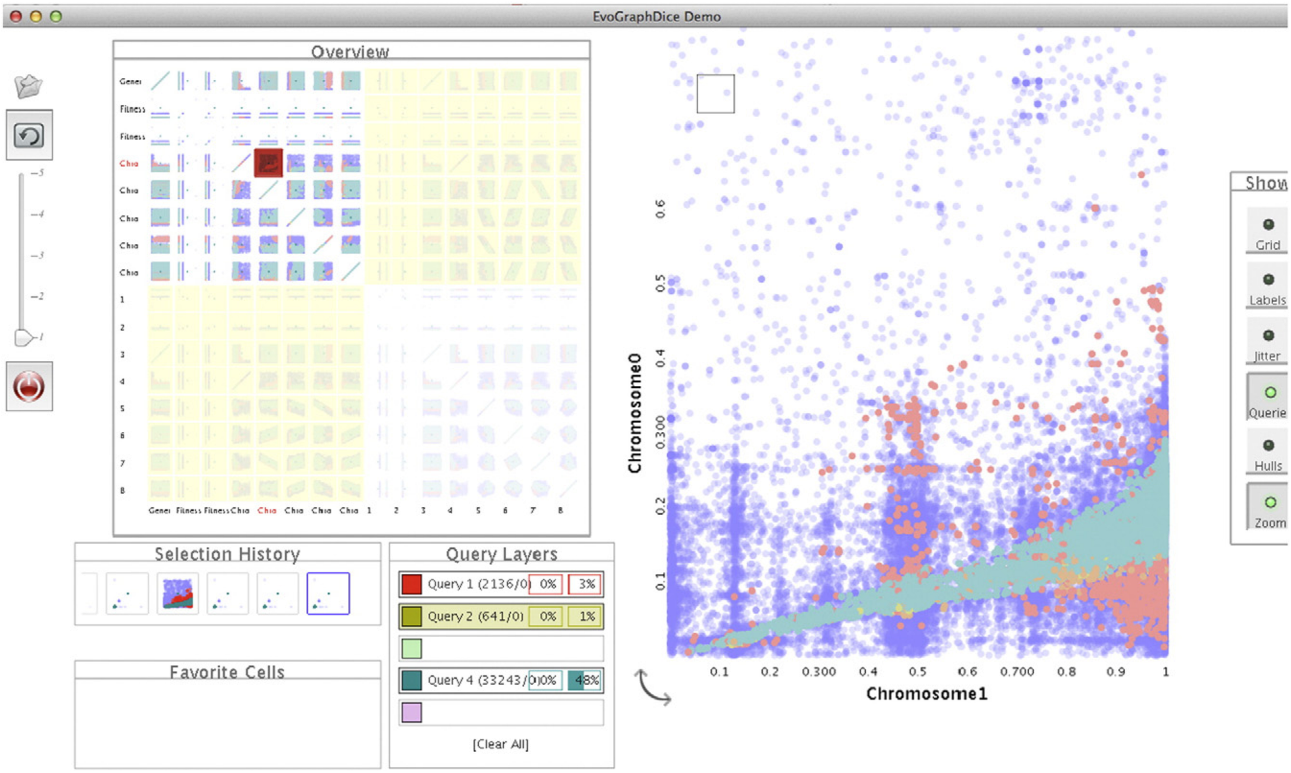


Fig. 5. Visualisation of a multiobjective run (NSGA-II) in the 5 parameter space. (k_{wp}, k_{cas}) projection (right), with the same colouring of the Pareto frontier. Yellow points are at the bottom line of the green cone.

milk proteins in permeate. The considered milk proteins are the following:

- a mixture of caseins (Promilk 852B, IDI company, France, with 5% moisture, 1.5% fat, 85.5% nitrogenous matter/dry matter, 8.5% mineral matter, 4% lactose, 81% nitrogenous matter in powder, 92% casein micelle, 2.6% Ca, 1.5% P, 0.3% K, 0.1% Na and 0.1% Mg); and
- native whey proteins (BiPro, DAVISCO company, Minnesota with 5% max moisture, 95% min protein, dry basis, 1% max fat, 3% max ash, 1% max lactose and pH between 6.7 and 7.5) with milk permeate powder (Armor Protines, France with pH of 6.0 min, 3% max moisture, 3% min proteins, 1% max fat, 82% lactose and 8% ashes).

The continuous phase has been prepared the day before the experience, waved at 4 °C and then heated at 60 °C before emulsification. The dispersed phase of the emulsion is a saturated liquid; anhydrous milk fat (AMF) which has been heated at 60 °C to become liquid.

Two sets of experiments were made as to characterise the emulsifications dynamics (so called database 1 and database 2). They differ in the homogenisation process and in the volume of emulsion produced. The process for database 1 consisted in mixing 49 g of the protein phase and 21 g of milk fat using a rotor stator homogenizer and a low pressure homogenizer (50 bar). The process for database 2 consisted in mixing 182 g of the protein phase and 78 g of milk fat using a blender and a high pressure homogenizer (300 bar). Although they produced different amounts of emulsion, respectively 70 g and 260 g for databases 1 and 2, the proportions of the different components were kept constant: 30% w/w of milk fat and 70% w/w of water phase leading to about 3.4% w/w of proteins. Several initial conditions have been tested, and the emulsions are characterised using measurements at the micro/nanoscale (Foucquier et al., 2011; Surel et al., 2014).

- Laser light scattering has been used to measure the diameter of the lipid droplets in the emulsion and evaluate the size distribution. For database 1, measurements have been performed using a Saturn DigiSizer 5200 (Micromeritics, Norcross, USA),while for database 2, they have been made with a LS 13320 laser diffraction particle size analyser (Beckman Coulter). The gathered information has been used to compute the initial free lipid surface S_0. The measurement error for the aforementioned devices is around 10%;
- For database 1, the composition of the interface of lipid droplets has been studied using the Patton and Huston technique (Patton & Huston, 1986) to separate droplets. The interfacial protein concentration has been then quantified through the Markwell method (Markwell, Haas, Bieber, & Tolbert, 1978). Subsequently, SDS-PAGE electrophoresis has been applied to determine the concentration of each protein at the interface. In database 2, SDS-PAGE electrophoresis has also been carried out and gels have been purchased from Invitrogen Ltd. (Paisley, UK). The measurements give the interfacial concentration (Γ) and the percentage of adsorbed caseins ($p_{cas_{ads}}$). The measurement errors of the Patton techniques and of the electrophoresis are each at around 10%. The resulting measurement error for this experimentation is thus at around 20%.

Database 1 is used for learning and database 2 for validation (see Table 1): the training set is made of 7 samples, L_1 to L_7, and 4 samples are used as the validation set, V_1 to V_4.

2.2. Modelling milk gel competition at the interface

Fig. 1 is an overview of the model that will be used to verify the efficacy of our approach. The model, developed in our previous work (Foucquier et al., 2011), has been chosen because of the availability of data and expert knowledge on the process' behaviour. It is important to notice that the focus of this paper is on the coupling of evolutionary computation and visualisation techniques, not on the model itself, or its efficiency: the proposed approach can be generalised to any kind of model.

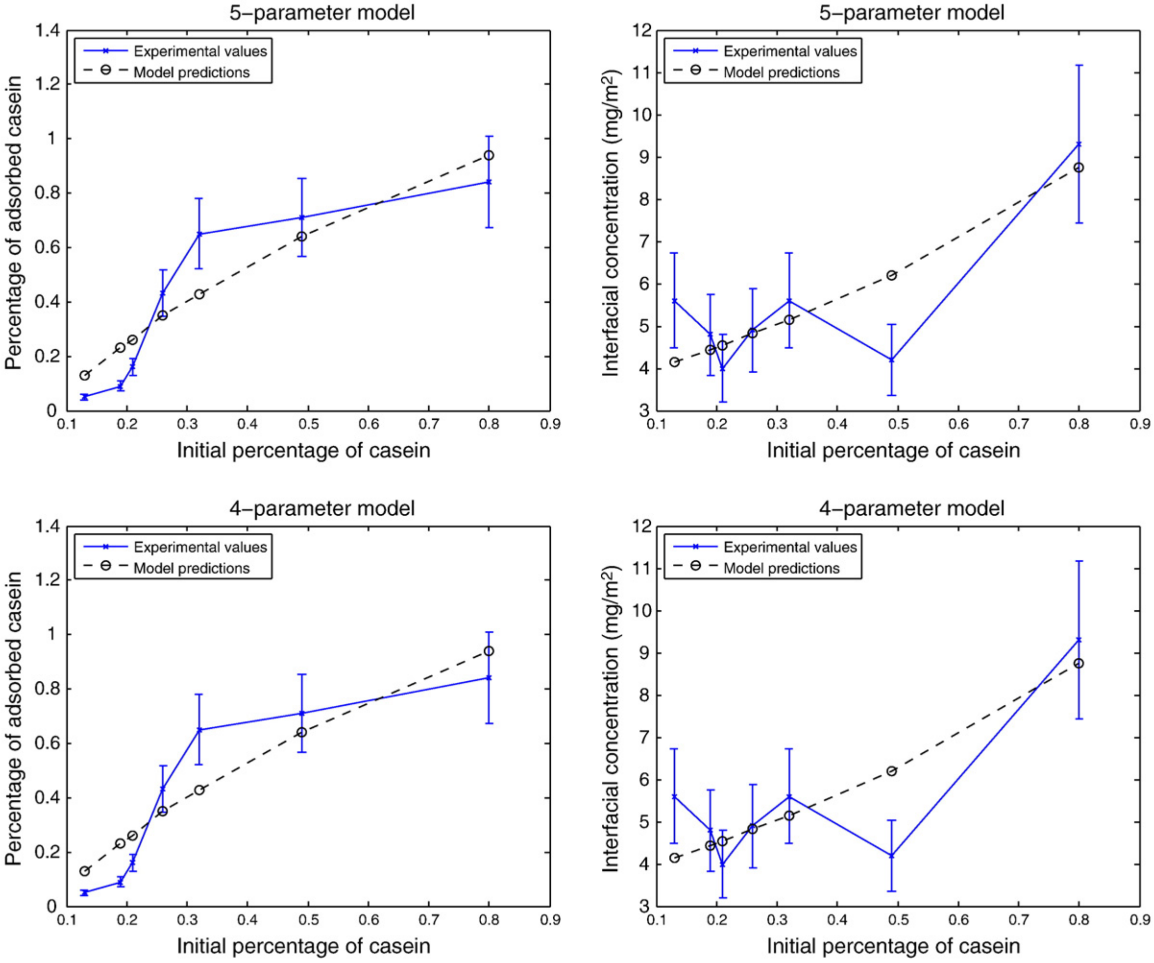


Fig. 6. Fitting for the two outputs of the casein model on the training set (database 1), with 5 parameters (top) and 4 parameters (bottom). The bars represent the measurement error in the experimental data.

The amount of native whey proteins in the solution, m_{wp}, and casein micelles, m_{cas}, as well as the surfaces S_{fall} and S_{fres} evolves with time. S_{fall} is the lipid surface area available for colonisation of native whey proteins and casein micelles, and S_{fres} is the lipid surface left by casein micelles, owing to steric effects, for native whey proteins.

The model predicts the structure characterised by the mass percentage of adsorbed casein micelles and native whey proteins relatively to the total amount of proteins adsorbed at the lipid interface, $w_{cas_{int}}$ and $w_{wp_{int}} = 1 - w_{cas_{int}}$ respectively, and the interfacial concentration, Γ, which corresponds to the quantity of adsorbed proteins for 1 m^2 of lipid surface. These outputs depend on:

- the structure of the model built from expert knowledge;
- the parameters of the model i.e., 5 parameters: adsorption rate of native whey proteins, k_{wp}, adsorption rate of casein micelles, k_{cas}, surface occupied by 1 g of native whey proteins, s_{wp}, surface occupied by 1 g of casein micelles, s_{cas} and fraction of the adsorbed surface of a casein micelle left for native whey proteins, α. k_{wp} and k_{cas} are associated with a mean representation of the reactions that take place at a local level according to the organisation of each protein at the nanoscale. It induces specific diffusion and protein–surface interactions that are taken into account at a higher scale through these overall parameters. s_{wp} and s_{cas} are associated with a mean representation of actual adsorption, conformational reorganisation and structural consolidation that take place at a lower scale level; and
- the inputs, i.e., the initial mass percentage of casein micelles in the water phase, w_{cas_0}, initial mass percentage of native whey proteins in the water phase, $w_{wp_0} = 1 - w_{cas_0}$, initial mass of proteins, $m_p(0)$, initial mass of lipid, $m_l(0)$, and initial lipid surface, $S_{0,}$ depending on $d_{3.2}$ (average diameter of lipid droplets):

$$S_0 = \frac{m_l}{\rho_l}\frac{6}{d_{3.2}} \tag{1}$$

where m_l is the constant mass of lipid in the solution and ρ_1 is the lipid density (0.920 10^6 g · m^{-3}).

The system described in Eq. (2) is an ordinary differential equation system (ODE), that can be solved with a Runge–Kutta fourth order method (RK4) (Butcher, 1987).

$$\begin{cases} \dfrac{dm_{wp}}{dt} = -k_{wp}\ m_{wp}\left(\dfrac{m_{wp}}{m_{wp}+m_{cas}}\right)\dfrac{S_{f_{all}}+S_{f_{res}}}{S_0} \\ \dfrac{dm_{cas}}{dt} = -k_{cas}\ m_{cas}\left(\dfrac{m_{cas}}{m_{wp}+m_{cas}}\right)\dfrac{S_{f_{all}}}{S_0}\left(1-\dfrac{\left(m_{cas_0}-m_{cas}\right)s_{cas}}{S_0}\right) \\ \dfrac{dS_{f_{all}}}{dt} = \dfrac{dm_{wp}}{dt}s_{wp}\left(\dfrac{S_{f_{all}}}{S_{f_{all}}+S_{f_{res}}}\right)+\dfrac{dm_{cas}}{dt}s_{cas} \\ \dfrac{dS_{f_{res}}}{dt} = \dfrac{dm_{wp}}{dt}s_{wp}\left(\dfrac{S_{f_{res}}}{S_{f_{all}}+S_{f_{res}}}\right)-\dfrac{dm_{cas}}{dt}\alpha s_{cas} \end{cases} \tag{2}$$

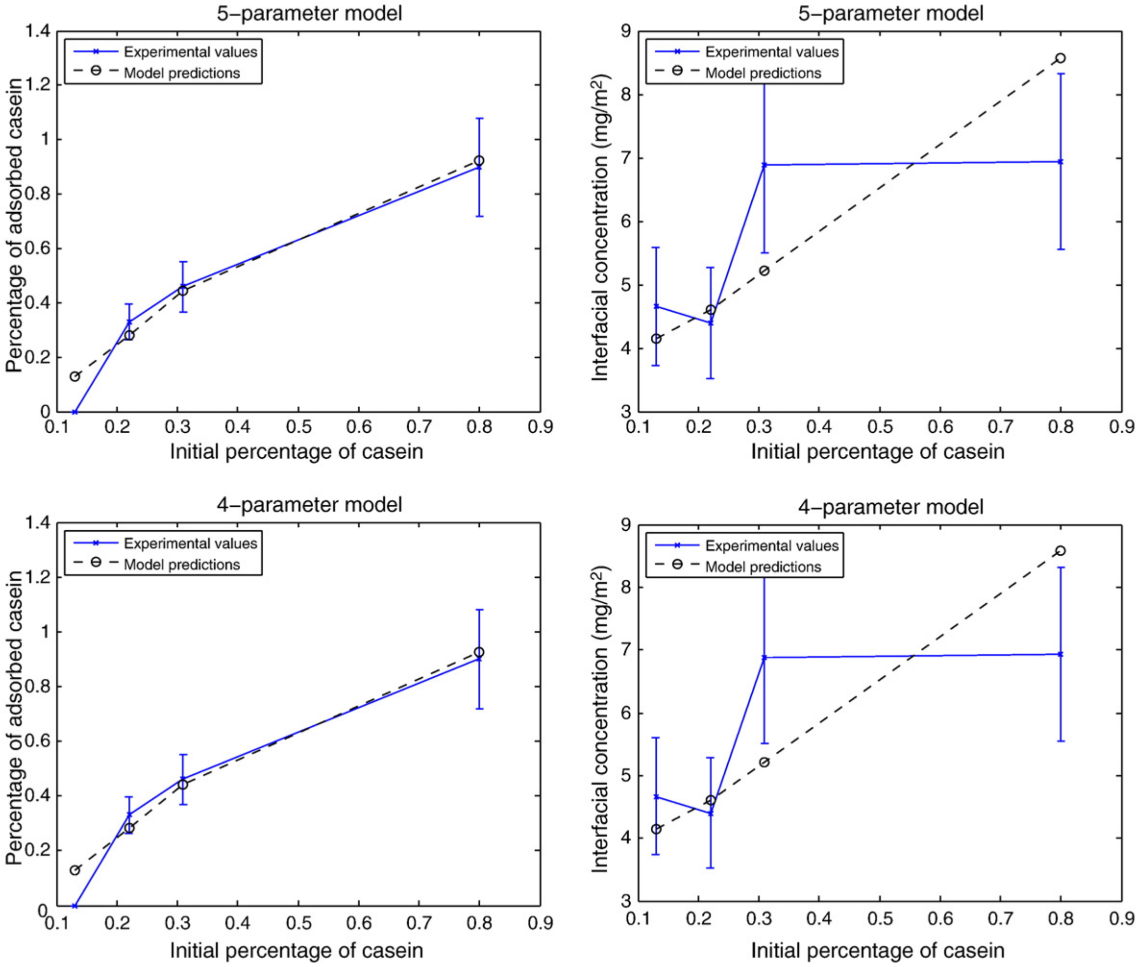


Fig. 7. Fitting for the two outputs of the casein model on the validation set (database 2), with 5 parameters (top) and 4 parameters (bottom). The bars represent the measurement error in the experimental data.

Initialisation of the state variables was set by using Table 1 and the following relations with S_0 computed according to Eq. (1):

$$
\begin{aligned}
&m_{wp}(0) = m_p(0)\ \ w_{wp}(0)\\
&m_{cas}(0) = m_p(0)\ \ w_{cas}(0)\\
&S_{fres}(0) = 0\\
&S_{fall}(0) = S_0.
\end{aligned}
$$

2.3. Learning the parameters of the model using an evolutionary approach

The learning task can be turned into the following optimisation problem: find the optimal parameter setting that best matches the training dataset. To run an EA on such a problem, the following features have been fixed:

- Search space, structure of an individual of the population: a candidate solution is a vector of real numbers, describing the values of model parameters k_{wp}, k_{cas}, s_{wp}, s_{cas} and α. An individual is thus a set of 5 real values in the interval (0, 1): each value is mapped in the appropriate interval of validity for the corresponding parameter, before evaluation. A complete list of intervals is reported in Table 2.
- Fitness function, how to assess the quality of a solution: the quality of a solution is given by the average squared error between values predicted by the system of Eq. (2) (solved using a Runge–Kutta fourth order method) and experimental data (learning set). It is important to note that the model has several outputs, and a candidate solution might perform very well for one output, and badly for the others. In this experience, we are focusing on two outputs, the interfacial concentration and the final casein percentage in the mixture. There are two ways to consider this: single-objective and bi-objective approaches, as EAs are also able to deal with multi-objective problems (Deb, 2001).
- Genetic engine, operators and reproduction strategies: given the problem characteristics, we chose two established evolutionary algorithms particularly suited to study the nature of the problem: covariance matrix adaptation evolution strategy (CMA-ES) (Hansen, Müller, & Koumoutsakos, 2003) and non-sorting genetic algorithm II (NSGA-II) (Deb, Pratap, Agarwal, & Meyarivan, 2002). CMA-ES is considered one of the best real-value optimizers for single-objective problems, delivering high-quality results in a very limited amount of time. NSGA-II is the de facto state-of-the-art for multi-objective optimisation.
Both algorithms are purposefully set with a large population, in order to obtain more insights on the nature of the problem from the distribution of candidate solutions in the search space. For CMA-ES, the fitness function to be minimised is the product of the average squared errors on each output; NSGA-II simply considers the average squared error on each output as an objective to minimise. Complete parameters for both algorithms are reported in Table 3.
For every algorithm, a single run is performed, until a stagnation condition is reached (for CMA-ES) or 100 iterations of the process have expired (NSGA-II).

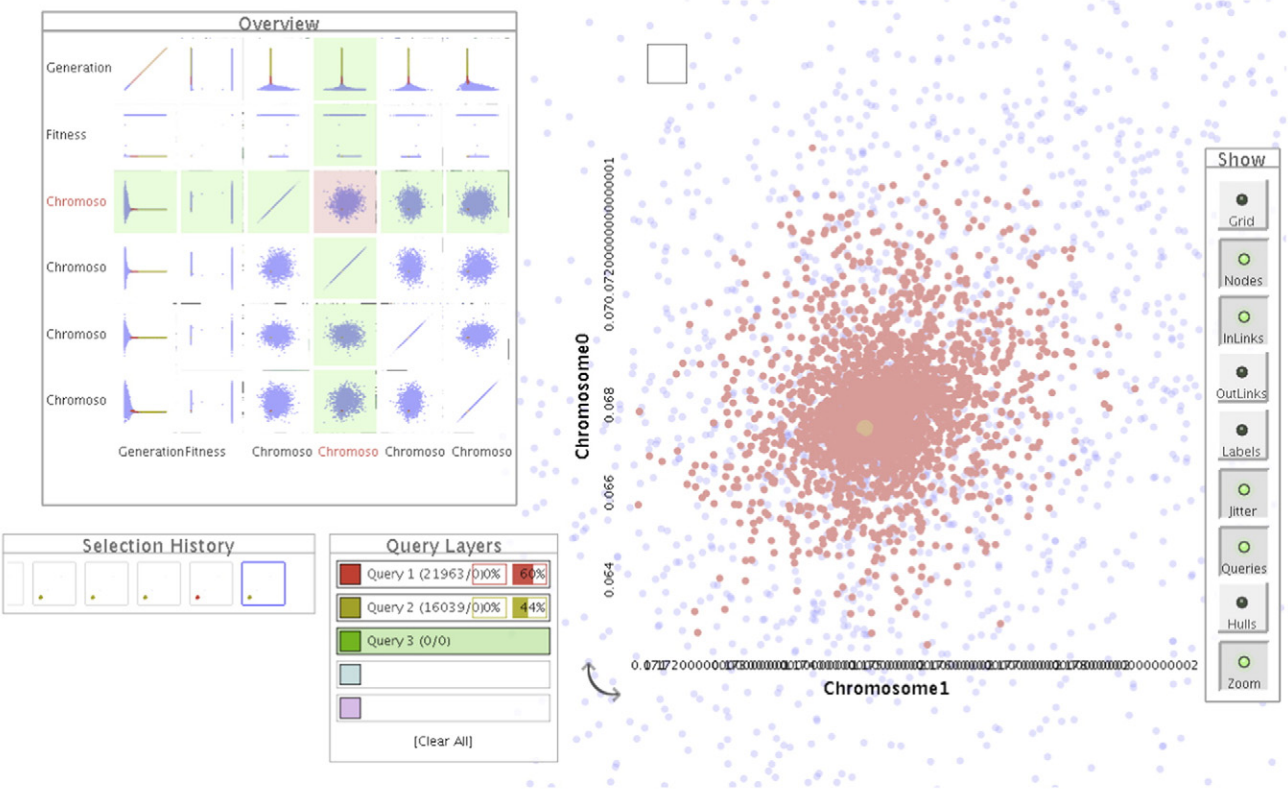


Fig. 8. Zoomed view, using GraphDice. Red points correspond to good fitness values and yellow points to best fitness values. The optima values are concentrated on a point.

2.4. Visualisation using the GraphDice environment

GraphDice (Bezerianos et al., 2010) is an evolution of ScatterDice (Elmqvist, Dragicevic, & Fekete, 2008), a multidimensional visual exploration tool, that enables the user to navigate in a multidimensional set via simple 2D projections, organised as scatterplot matrices. The visual coherence between various projections is based on animated 3D transitions. A scatterplot matrix presents an overview of the possible configurations, thumbnails of the scatterplots, and support for interactive navigation in the multidimensional space. Various queries can be built using bounding volumes in the dataset, sculpting the query from different viewpoints to become more and more refined. Furthermore, the dimensions in the navigation space can be reordered, manually or automatically, to highlight salient correlations and differences among them[1].

GraphDice (Bezerianos et al., 2010) uses the same principles but with many additional features, allows reading the same type of data (.csv files), and other more sophisticated formats, as it also embeds graph visualisation utilities[2]. GraphDice can be used to visualise data collected during the run of an EA (Lutton & Fekete, 2011). At each generation, the content of the current population can be written into a ".csv" file as shown in Fig. 2, creating what can be called as a "cloud" of successive populations made of multidimensional points. The figures presented in the next section have been generated using EvoGraphDice, another extension of GraphDice, specially devised to analyse dependencies between variables (an extension of a PCA analysis of multidimensional data, based on an interactive evolutionary algorithm, allowing to consider various linear or non-linear dependencies) (Cancino, Boukhelifa, & Lutton, 2012).

The visualisation system (Figs. 3 to 9) displays, at the same time:

- an overview scatterplot matrix (top left, entitled "Overview") showing the original data set of 7 dimensions, namely generation, fitness and 5 parameters, plus additional dimensions (1…7) for EvoGraphDice;
- a main plot view (top right), corresponding to a zoom of one of the cells of the overview scatterplot matrix. It corresponds to the red cell at the intersection of the green line and columns in the scatterplot matrix;
- a tool bar for main plot view giving access to zoom, convex hulls and other functionalities; and
- a selection query window, to manage various subsets of points that are interactively selected (lasso selection using the mouse).

3. Results and discussion

3.1. Sensitivity analysis

A global sensitivity analysis was performed in Foucquier et al. (2011), on the basis of the variance-based method described in Saltelli (2002). The results are summarised in Table 4. The relation $w_{cas_{int}} + w_{int} = 1$ induced identical sensitivity results for output variables $w_{cas_{int}}$ and $w_{wp_{int}}$. For the sake of clarity, Table 4 only presents the results for $w_{cas_{int}}$ and Γ.

This analysis recommends keeping the 5 parameters relative to the established structure. No matter what the initial conditions are, k_{wp} and k_{cas} seem to have an impact on the variance of the percentage of adsorbed caseins (Table 4). Their impact on the variance of the interfacial concentration, however, is verified only for high initial percentages of caseins. Nevertheless, it is important to find good values for all parameters: k_{wp} and k_{cas} have an impact for at least one given initial condition and one output; α and s_{cas} seem to have a major impact on the variance of the outputs of the model for high initial percentage of caseins; and, for all initial conditions, s_{wp} has a high impact on the variance of the outputs of the model.

[1] A demo of ScatterDice can be launched from http://www.aviz.fr/~fekete/scatterdice/, it accepts standard .csv files (although it may be necessary to add a second line after the header giving the data type for each column — INT, STR, REAL, etc).

[2] A demo of GraphDice is also accessible at http://www.aviz.fr/graphdice/.

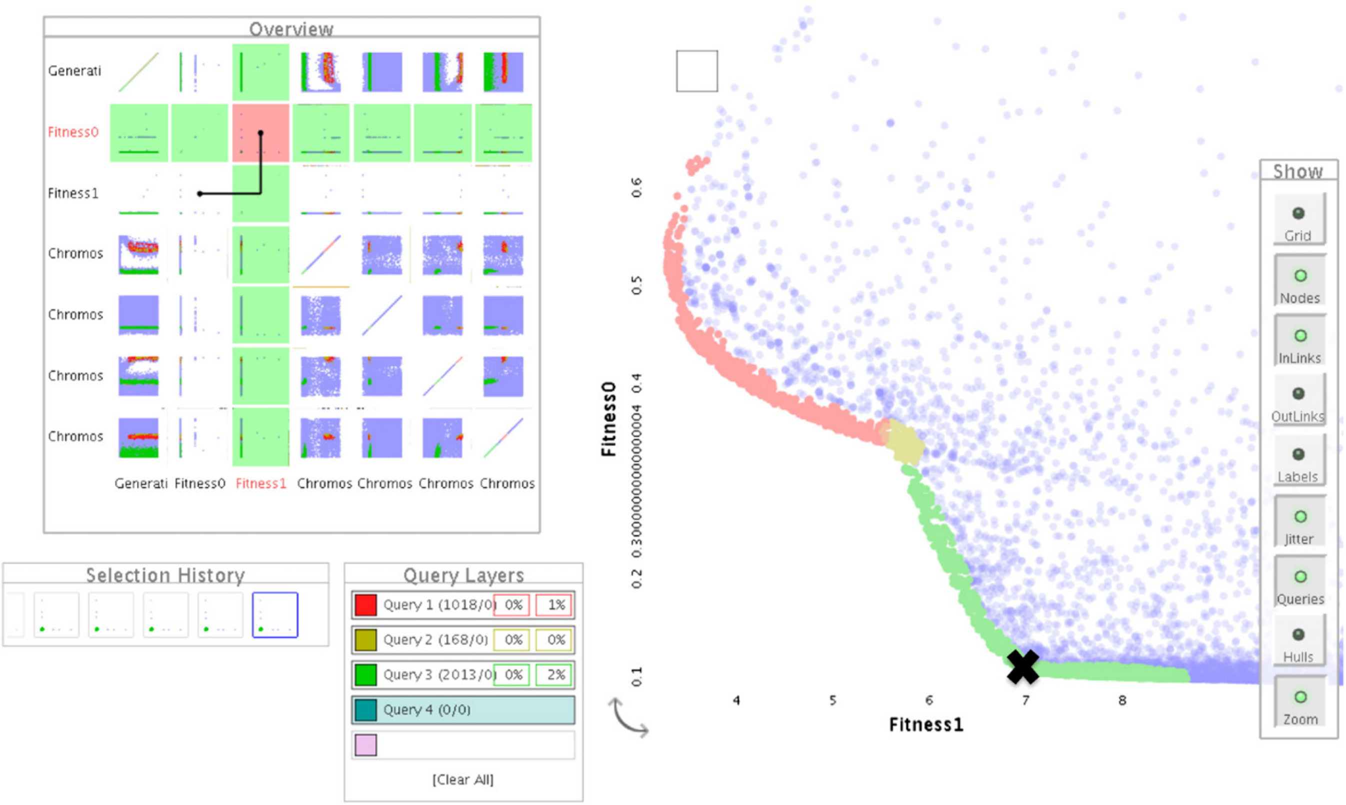


Fig. 9. The Pareto front has been highlighted in three colours: red for the points where fitness1 is dominating, green for the points where fitness0 is dominating, and yellow for balanced fitness area. The optimal point found with CMA-ES (4 parameters problem) is situated under the “X” mark.

Sensitivity analysis makes it possible to gain some insight on the model, and it is a starting point for the visual exploration. Is has been made evident that 5 parameters have to be kept for a complete description of the dynamics of the system. However, questions remain about the possible simplifications at the equilibrium state.

3.2. Visual exploration of the model

A mono-objective EA has been first run on the 5 parameter search space according to the settings presented in Section 2.3. It is important to notice that the available datasets only represent equilibrium states (when $m_{wp} + m_{cas}$ goes to zero): observation results are thus only valid for these specific conditions. A best fitting corresponds to the following values:

$$k_{wp} = 11.976 \ \text{s}^{-1}, \ \ k_{cas} = 80.783 \ \text{s}^{-1}, \ \ s_{wp} = 261.209 \ \text{m}^2 \cdot \text{g}^{-1}, \ \ s_{cas} = 104.876 \ \text{m}^2 \cdot \text{g}^{-1} \ \text{and} \ \alpha = 9.748 \cdot 10^{-12}.$$

However, a visual exploration of the EA data collected during the run (a sample file is given in Fig. 2) shows a convergence toward a rather large area of values for the couple of parameters k_{wp} and k_{cas}. The right part of Fig. 3 displays a projection onto the plane (k_{wp}, k_{cas}) of the distribution of the points visited by the EA along generations, the best fitness values are coloured in red. It is made evident that, even if the EA is converging in a satisfying manner, there exists a whole set of optimal values distributed along a line segment. This evidence is even more salient when examining the same type of data collected during the run of a multi-objective EA (NSGA-II). Fig. 4 shows in the main plot (right part), a zoom on the Pareto front, i.e. the projection defined by the two fitness values (fitness0, fitness1), corresponding to the two aims being optimised (i.e. respectively the average error for adsorbed casein, and the average error for interfacial concentration). A set of queries highlights some parts of the Pareto front: preference of fitness0 over fitness1 in green, and reversely in red, and equivalent compromise in yellow. The same colour encoding is used in Fig. 5 that displays the projection in the (k_{wp}, k_{cas}) plane of the same dataset. It is made evident that the green area corresponds to a large set of equivalent points (a cone). Yellow points are distributed along the bottom line of the green cone.

Visualisations of Figs. 3 to 5 show some evidence about a possible dependence between k_{wp} and k_{cas} for optimal values. Experiments run with 4 parameters only are presented in Figs. 8 and 9. The 4 parameters are k, s_{wp}, s_{cas} and α with $k_{wp} = k \cdot k_{cas}$.

Fig. 8 shows a convergence toward a point with respect to all projections in the 4-parameter space, and optimal values are:

$$k_{wp} = 1 \ \text{s}^{-1}, \ \ k_{cas} = 6.748 \ \text{s}^{-1}, \ \ s_{wp} = 261.264 \ \text{m}^2 \cdot \text{g}^{-1}, \ \ s_{cas} = 104.786 \ \text{m}^2 \cdot \text{g}^{-1} \ \text{and} \ \alpha = 1.352 \cdot 10^{-12}.$$

It is noticeable how the values found for s_{wp}, s_{cas} and α are extremely close to values previously found for the 5-parameter model. The ratio $\frac{k_{wp}}{k_{cas}}$ is 0.148 for both the 5-parameter and 4-parameter models. Small differences are due to the stochastic nature of the optimisation techniques applied to the problem.

Figs. 6 and 7 show the model fitting to experimental points of data, for the 5-parameter model and the 4-parameter model, with optimised parameter value. It is intuitive to notice how the shapes of the curves and the points are almost exactly the same for 5 and 4 parameters. Theoretical analysis of Section 3.3 confirms this visual evidence.

The observation of average fitting errors per sample for both solutions (Table 5 and Figs. 6 and 7), also shows that the prediction of adsorbed casein seems to be an easier task than the prediction of the interfacial concentration.

Table 4
Total effects of the parameters on the variances of the outputs of the model. Meaning of symbols: 0, no or very low impact ($S_{T_i} \leq 0.1$); +, low impact ($0.1 < S_{T_i} \leq 0.3$); ++, average impact ($0.3 < S_{T_i} \leq 0.6$); +++, high impact ($0.6 < S_{T_i} \leq 1$); and ++++, very high impact ($S_{T_i} > 1.0$).

	$w_{cas}(0) = 13\%$		$w_{cas}(0) = 21\%$		$w_{cas}(0) = 49\%$		$w_{cas}(0) = 80\%$	
S_{Ti}	$w_{cas_{int}}$	Γ	$w_{cas_{int}}$	Γ	$w_{cas_{int}}$	Γ	$w_{cas_{int}}$	Γ
k_{wp}	++	0	++	0	+	0	+++	+
k_{cas}	++	0	+++	0	++	0	+++	+
α	0	0	0	0	0	0	++++	+
s_{wp}	+	++++	+	+++	+	++	+++	+
s_{cas}	0	0	0	0	+	0	++++	+++

3.3. Theoretical discussion

A change of variables has been performed in order to circumvent the mathematical singularity of system 2, when $m_{wp} + m_{cas}$ goes to zero. Introducing the ratio $P = \frac{m_{wp}}{m_{cas}}$ allows rewriting the Eq. (2) as an equivalent regular system 3.

$$\begin{cases} \dfrac{dP}{dt} = \dfrac{P}{S_0(P+1)}\left[k_{wp}P\left(S_{f_{all}} + S_{f_{res}}\right) - k_{cas}S_{f_{all}}\left(1 - \dfrac{\left(m_{cas_0} - m_{cas}\right)s_{cas}}{S_0}\right)\right] \\ \dfrac{dm_{cas}}{dt} = -k_{cas}\dfrac{S_{f_{all}}\ m_{cas}}{S_0(P+1)}\left(1 - \dfrac{\left(m_{cas_0} - m_{cas}\right)s_{cas}}{S_0}\right) \\ \dfrac{dS_{f_{all}}}{dt} = -\dfrac{S_{f_{all}}m_{cas}}{S_0(P+1)}\left[k_{wp}\ s_{wp}\ P^2 + k_{cas}\ s_{cas}\left(1 - \dfrac{\left(m_{cas_0} - m_{cas}\right)s_{cas}}{S_0}\right)\right] \\ \dfrac{dS_{f_{res}}}{dt} = -\dfrac{m_{cas}}{S_0(P+1)}\left[k_{wp}\ s_{wp}\ S_{f_{res}}\ P^2 - k_{cas}\ \alpha s_{cas}\ S_{f_{all}}\left(1 - \dfrac{\left(m_{cas_0} - m_{cas}\right)s_{cas}}{S_0}\right)\right] \end{cases} \quad (3)$$

The available data only characterise the emulsion at initial and final states of the process; therefore, it is not possible for the model to catch the dynamic of the process, while it provides accurate results for the final state of the emulsion. Consequently, we only focus on the asymptotic values of the solutions of system 2 or 6. If we set a new timescale, $u = k_{cas}t$, and introduce the ratio between adsorption rates, $k = \frac{k_{wp}}{k_{cas}}$, system 6 rewrites as follows.

$$\begin{cases} \dfrac{dP}{du} = \dfrac{P}{S_0(P+1)}\left[kP\left(S_{f_{all}} + S_{f_{res}}\right) - S_{f_{all}}\left(1 - \dfrac{\left(m_{cas_0} - m_{cas}\right)s_{cas}}{S_0}\right)\right] \\ \dfrac{dm_{cas}}{du} = -\dfrac{S_{f_{all}}m_{cas}}{S_0(P+1)}\left(1 - \dfrac{\left(m_{cas_0} - m_{cas}\right)s_{cas}}{S_0}\right) \\ \dfrac{dS_{f_{all}}}{du} = -\dfrac{S_{f_{all}}m_{cas}}{S_0(P+1)}\left[ks_{wp}\ P^2 + s_{cas}\left(1 - \dfrac{\left(m_{cas_0} - m_{cas}\right)s_{cas}}{S_0}\right)\right] \\ \dfrac{dS_{f_{res}}}{du} = -\dfrac{m_{cas}}{S_0(P+1)}\left[ks_{wp}\ S_{f_{res}}P^2 - \alpha s_{cas}\ S_{f_{all}}\left(1 - \dfrac{\left(m_{cas_0} - m_{cas}\right)s_{cas}}{S_0}\right)\right] \end{cases} \quad (4)$$

Table 5
Average errors for 4- and 5-parameter solutions, for both training and validation sets. The relative error for each data point is computed as $\frac{abs\ (p_i - e_i)}{e_{\max} - e_{\min}}$, where p_i is the value predicted by the model, e_i is the experimental value, $e_{\max}$ and $e_{\min}$ are the maximum and minimum experimental values, respectively.

Model	Average error for adsorbed casein samples		Average error for interfacial concentration samples	
Training set				
	Absolute	Relative	Absolute	Relative
4 parameters	0.113	14.3%	0.773	20.9%
5 parameters	0.113	14.3%	0.773	20.89%
Validation set				
	Absolute	Relative (%)	Absolute	Relative (%)
4 parameters	0.055	0.06%	1.008	44.42%
5 parameters	0.055	0.06%	1.008	44.38%

The change of timescale does not impact on the stationary states, which means that systems 3 and 4 lead to the same asymptotic values. Notice also that only the ratio $k = \frac{k_{wp}}{k_{cas}}$ is involved in 4, which means that it is possible to express the asymptotic values only with respect to k. This result is in accordance with the visual analysis presented above.

4. Conclusions

The coupling of evolutionary algorithms with visualisation proves to be an efficient strategy for exploring a food model. For the milk gel model, it was possible to reduce the complexity of the model, due the nature of the available data. Experimental measurements did actually not give access to the temporal data: some parameters governing the dynamics were thus impossible to define in a unique way, and using a ratio was enough to fit the model to the data.

This approach can be used for exploring other models, including models that are not fully analytically defined. It has to be noticed, however, that both the structure of the model and the nature of available data have a strong impact on the models that can be explored. For instance, if there are no data that make it possible to access dynamic behaviours, there is no way to instantiate dynamic parameters in an unique way. A way to address this issue might be to exploit interactions with human experts as an additional source of information. Further studies will consider the integration of expert knowledge, by allowing the expert to interactively re-design the model, or the injection of uncertainty informations for data sample points. Specific methods for dynamic models will also be considered.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under the grant agreement FP7-222 654-DREAM.

The authors would like to thank Romain Reuillon and Mathieu Leclaire for their precious help with Scala programming; Claire Surel, Marc Anton from INRA BIA Nantes, and Alan Mackie, Marine Rouland from IFR Norwich, for their precious help in building the model and conducting the experiments.

References

Bäck, T., & Schwefel, H. -P. (1993). An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation*, *1*(1), 1–23.

Bezerianos, A., Chevalier, F., Dragicevic, P., Elmqvist, N., & Fekete, J. -D. (2010). Graphdice: A system for exploring multivariate social networks. *Computer Graphics Forum*, *29*(3), 863–872 (Proc. EuroVis 2010).

Bremermann, H. -J. (1962). Optimization through evolution and recombination. In M. C. Yovits, G. T. Jacobi, & G. D. Goldstein (Eds.), *Self-organizing systems* (pp. 93–106). Washington DC: Spartan Books.

Butcher, J. C. (1987). *The numerical analysis of ordinary differential equations: Runge–Kutta and general linear methods.* New York, NY, USA: Wiley-Interscience.

Cancino, W., Boukhelifa, N., & Lutton, E. (2012, June 10–15). EvoGraphDice: Interactive evolution for visual analytics. *IEEE Congress on Evolutionary Computation, June 10–15*. Brisbane, Australia: IEEE Computational Intelligence Society.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray.

Deb, K. (2001). *Multi-objective optimization.* Multi-objective optimization using evolutionary algorithms, 13–46.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197.

Dickinson, E. (1999). Caseins in emulsions: Interfacial properties and interactions. *International Dairy Journal*, *9*(3–6), 305–312.
1997 Hannah Symposium on caseins and caseinates — Structures, interactions, networks (pp. 305–312). Ayr, Scotland: Hannah Res Inst.

Dickinson, E. (2001). Milk protein interfacial layers and the relationship to emulsion stability and rheology. *Colloids and Surfaces B: Biointerfaces*, *20*(3), 197–210.

Dickinson, E. (2011). Mixed biopolymers at interfaces: Competitive adsorption and multilayer structures. *Food Hydrocolloids*, *25*(8), 1966–1983.

Elmqvist, N., Dragicevic, P., & Fekete, J. -D. (2008). Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1141–1148 (Proc. InfoVis 2008).

Erni, P., Windhab, E. J., & Fischer, P. (2011). Emulsion drops with complex interfaces: Globular versus flexible proteins. *Macromolecular Materials and Engineering*, *296*(3–4), 249–262.

Foucquier, J., Chantoiseau, E., Le Feunteun, S., Flick, D., Gaucel, S., & Perrot, N. (2012). Toward an integrated modeling of the dairy product transformations, a review of the existing mathematical models. *Food Hydrocolloids*, *27*(1), 1–13.

Foucquier, J., Gaucel, S., Surel, C., Riaublanc, A., Baudrit, C., & Perrot, N. (2011). Modelling the formation of fat droplet interface during homogenisation in order to describe the texture. In G. Saravacos, P. Taoukis, M. Krokida, V. Karathanos, H. Lazarides, N. Stoforos, C. Tzia, & S. Yanniotis (Eds.), *11th International Congress on Engineering and Food (ICEF11). Procedia Food Science, Vol. 1*. (pp. 706–712). Sara Burgerhartstraat 25, PO Box 211, 1000 AE Amsterdam, Netherlands: Elsevier Science BV (11th International Congress on Engineering and Food (ICEF), Athens, GREECE, MAY 22–26, 2011).

Fraser, A. S. (1957). Simulation of genetic systems by automatic digital computers. I. Introduction. *Australian Journal of Biological Sciences*, *10*, 484–491.

Gaygadzhiev, Z., Hill, A., & Corredig, M. (2009). Influence of the emulsion droplet type on the rheological characteristics and microstructure of rennet gels from reconstituted milk. *Journal of Dairy Research*, *76*(3), 349–355.

Hansen, N., Müller, S. D., & Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, *11*(1), 1–18.

Holland, J. H. (1962). Outline for a logical theory of adaptive systems. *Journal of the Association for the Computing Machinery*, *9*(3), 297–314.

Knudsen, J. C., Ogendal, L. H., & Skibsted, L. H. (2008). Droplet surface properties and rheology of concentrated oil in water emulsions stabilized by heat-modified beta-lactoglobulin B. *Langmuir*, *24*(6), 2603–2610.

Lutton, E., & Fekete, J. -D. (2011, July 12–16). Visual analytics of EA data. *Genetic and Evolutionary Computation Conference, GECCO 2011*. Dublin, Ireland: ACM.

Lutton, E., Foucquier, J., Perrot, N., Louchet, J., & Fekete, J. -D. (2011). Visual analysis of population scatterplots. *10th Biannual International Conference on Artificial Evolution (EA-2011), Angers, France*.

Markwell, A. K., Haas, S. M., Bieber, L. L., & Tolbert, N. E. (1978). Modification of Lowry procedure to simplify protein determination in membrane and lipoprotein samples. *Analytical Biochemistry*, *87*(1), 206–210.

McClements, D. J. (2004). Protein-stabilized emulsions. *Current Opinion in Colloid & Interface Science*, *9*(5), 305–313.

Morand, M., Dekkari, A., Guyomarc'h, F., & Famelart, M. H. (2012, AUG). Increasing the hydrophobicity of the heat-induced whey protein complexes improves the acid gelation of skim milk. *International Dairy Journal*, *25*(2), 103–111.

Murray, B. S. (2002). Interfacial rheology of food emulsifiers and proteins. *Current Opinion in Colloid & Interface Science*, *7*(5–6), 426–431.

Patton, S., & Huston, G. E. (1986). A method for isolation of milk-fat globules. *Lipids*, *21*(2), 170–174.

Perrot, N., Trelea, I. C., Baudrit, C., Trystram, G., & Bourgine, P. (2011). Modelling and analysis of complex food systems: State of the art and new trends. *Trends in Food Science & Technology*, *22*(6), 304–314.

Rabe, M., Verdes, D., & Seeger, S. (2011). Understanding protein adsorption phenomena at solid surfaces. *Advances in Colloid and Interface Science*, *162*(1–2), 87–106.

Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technicher system nach prinzipien der biologischen evolution*. Stuttgart: Fromman Holzboog.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297.

Surel, C., Foucquier, J., Perrot, N., Mackie, A., Garnier, C., Riaublanc, A., et al. (2014). Composition and structure of interface impacts texture of o/w emulsions. *Food Hydrocolloids*, *34*, 3–9.

Takagi, H. (1998, Sept 17–19). Interactive evolutionary computation: System optimisation based on human subjective evaluation. *IEEE Int. Conf. on Intelligent Engineering Systems (INES'98), Vienna, Austria*.

Update

Innovative Food Science and Emerging Technologies

Volume 32, Issue , December 2015, Page 208

DOI: https://doi.org/10.1016/j.ifset.2015.09.007

Innovative Food Science and Emerging Technologies 32 (2015) 208

Contents lists available at ScienceDirect

Innovative Food Science and Emerging Technologies

journal homepage: www.elsevier.com/locate/ifset

Corrigendum

Corrigendum to "Food model exploration through evolutionary optimization coupled with visualization: Application to the prediction of a milk gel structure" [INNFOO/25 (2014) 67–77]

Evelyne Lutton[a,*], Alberto Tonda[a], Sébastien Gaucel[a], Alain Riaublanc[b], Nathalie Perrot[a]

[a] UMR 782 Génie et Microbiologie des Procédés Alimentaires, AgroParisTech, INRA, 78850 Thiverval-Grignon, France
[b] INRA, Rue de la Géraudière, BP 71627, 44316 Nantes Cedex 3, France

Please consider an additional author in fourth position in the list of authors of the paper IFSET-D-13-00261R, as follows: Evelyne Lutton, Alberto Tonda, Sébastien Gaucel, Julie Foucquier, Alain Riaublanc, Nathalie Perrot.

The author would like to apologize for any inconvenience caused.

DOI of original article: http://dx.doi.org/10.1016/j.ifset.2014.02.003.

* Corresponding author at: INRA, Rue de la Géraudière, BP 71627, 44316 Nantes Cedex 3, France. Tel.: +33 1 1 30 81 53 79.

E-mail address: evelyne.lutton@grignon.inra.fr (E. Lutton).

http://dx.doi.org/10.1016/j.ifset.2015.09.007