Validation of a Damage Accumulation Model of Replicative Ageing in *S. cerevisiae*

Master's thesis in Complex Adaptive Systems

AMANDA OLMIN

MASTER'S THESIS 2019

Validation of a Damage Accumulation Model of
Replicative Ageing in *S. cerevisiae*

AMANDA OLMIN

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
GOTHENBURG UNIVERSITY
Gothenburg, Sweden 2019

Validation of a Damage Accumulation Model of Replicative Ageing in *S. cerevisiae*
AMANDA OLMIN

Cover: Fit of simulated output from the damage accumulation model of replicative ageing to experimental data obtained for wild-type yeast.

Validation of a Damage Accumulation Model of Replicative Ageing in *S. cerevisiae*
AMANDA OLMIN
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Age-related diseases and conditions give rise to societal challenges and pose a threat to healthy ageing. At the same time, the more recent evolutionary theories of ageing hypothesise that the process of ageing is a consequence of living rather than an evolutionary strategy. Consequently, it is implied that ageing is not as inevitable as many might believe and, as a consequence, it is of interest to study this biological process and its underlying mechanisms.

On a cellular level, accumulation of damage is often regarded as the main cause of ageing. Since the basic properties of ageing between unicellular and multicellular organisms are similar on this level, it is common to use the unicellular yeast *Saccharomyces cerevisiae* as a model organism in the field of ageing research.

The aim of this project is to validate a mathematical damage accumulation model of replicative ageing in yeast. The model represents a cell by intact protein and damage and describes how these quantities change as the cell grows. In addition to cell growth, the model takes asymmetric division, retention and cell death into account.

For the purpose of validating the model of replicative ageing, structural and numerical identifiability methods are applied and continuous optimisation is performed using single-cell area data. The model is fit to experimental data obtained for wild-type yeast and the two deletion strains *sir2Δ* and *fob1Δ*. Moreover, replicative lifespan data of 4,698 single-gene deletion strains is analysed and, in conjunction to this, it is investigated how the model parameters affect the replicative lifespan of the simulations.

The results show that the parameters in the model of replicative ageing that describes the rate of change of intact protein and damage in the cell, are structurally identifiable. In spite of this, they are not numerically identifiable based on the experimental data available; the parameter estimates obtained have high variances and are moderately or highly correlated with each other. Likewise, it is possible to generate parameter sets that make the mathematical model reproduce the replicative lifespans of the investigated strains, if a replicative lifespan constraint is inferred on the optimisation.

For future work, it is suggested that new experimental data is generated as to fit the model of replicative ageing to growth curves belonging to cells of later life stages. Ultimately, the data should be sufficient enough for the optimisation to generate parameter sets that make the model adapt to the characteristics of the investigated strains, without having additional constraints added to the objective function.

Keywords: systems biology, ageing, yeast, damage accumulation, optimisation, parameter estimation, parameter identifiability

Acknowledgements

Contents

Contents

List of Figures

List of Figures

List of Tables

Chapter 1

Introduction

Demographic changes in terms of an increasing percentage of elderly people in the population and an increased expected lifespan give rise to several societal challenges [1]. Among those challenges are age-related diseases and conditions, posing a threat to healthy ageing. As a consequence of this, it is of interest to study ageing and to understand the underlying mechanisms of this biological process.

A commonly used model organism in the study of biological processes is the baker's yeast *Saccharomyces cerevisiae*. It has the advantage of being well-studied, having a short generation time and of using cell mechanisms that are similar to those used by human cells [2]. In addition to this, it is relatively easy to introduce gene alterations inside of a unicellular microorganism like *S. cerevisiae* [2], making it convenient to obtain experimental data from it. *S. cerevisiae* is also a good organism in which to study ageing, since the basic properties of ageing on a cellular level are similar between unicellular and multicellular organisms [3].

The development of new measurement techniques in the field of biology has led to the generation of a great amount of experimental data. However, to interpret the data and to understand how the sub-components of a biological system give rise to emergent properties, other tools than the human intellect need to be employed [3]. The field of systems biology integrates mathematical modelling and computer simulations with biology in order to get a better insight into biological processes [4]. Previous knowledge based on experimental data is used to construct mathematical models of biological systems. The models can then be used as tools to gain further insight into the systems and their underlying mechanisms as well as for making predictions regarding the biological processes in question.

In the context of systems biology, a mathematical damage accumulation model of replicative ageing in yeast was developed with the purpose to study the ageing process [5]. The model represents a cell by intact protein and damage and describes how the amounts of these variables change with time as the cell grows. Apart from cell growth, the model takes asymmetric cell division, retention and cell death into account.

This project aims to validate the model of replicative ageing. The investigation includes structural and numerical identifiability analyses, to determine if the parameters of the model are identifiable, as well as parameter estimation. The parameter estimation is performed using single-cell area data generated for wild-type *S. cerevisiae* and the two deletion strains *sir2*Δ and *fob1*Δ. The *sir2*Δ deletion strain has been shown to have a shortened replicative lifespan compared to wild-type *S. cerevisiae* while the *fob1*Δ strain has a prolonged replicative lifespan [6]. In addition

to the parameter estimation, replicative lifespan data of 4,698 single-gene deletion strains is analysed and, in conjunction to this, it is investigated how the parameters of the model affect the replicative lifespan of model simulations.

Chapter 2

Theory

This chapter gives a biological and mathematical background to the project. The biological background consists of an introduction to evolutionary theories of ageing and the ageing of unicellular organisms. For a mathematical background, the damage accumulation model of replicative ageing in yeast is explained in detail. In addition, the parameter estimation and identifiability analysis methods applied in this project are presented.

2.1 Evolutionary theories of ageing

The more recent evolutionary theories of ageing has cast aside the belief that ageing is a programmed process, selected for during the course of evolution since it was beneficial for the fitness of a population. These theories indicate that ageing is a side effect of living rather than a strategy used to adapt to the environment. It can be hypothesised that deleterious genes that give an effect only at later stages in life have been surpassed by natural selection as a consequence of a decreasing selection pressure with age [7]. In another perspective, these deleterious genes might in fact have increased the fitness of an individual at an early life stage and therefore have propagated in the population [7]. The basis for *the disposable soma theory* is that an organism has limited resources and needs to allocate them between maintenance and reproduction [8]. Ageing is then a result of a trade-off where translation accuracy in germ cells are favoured over accuracy in somatic cells, why the latter group of cells will accumulate errors over time.

2.2 Ageing of unicellular organisms

The yeast *S. cerevisiae* is commonly used as a unicellular model organism in the field of ageing research. One standard measurement used to access the age of a yeast cell is the *replicative lifespan* (RLS). The replicative lifespan of a cell is equal to the number of times that the cell divides during the course of its life [9]. Typically, the RLS of wild-type *S. cerevisiae* is 20-30 [10].

On a cellular level, damage accumulation is a generally accepted fundamental cause of ageing [3]. In yeast, oxidatively damaged proteins and extrachromsomal ribosomal DNA circles (ERCs) are part of the damage that have a higher abundance in ageing cells. These are grouped as *ageing factors*, partly since they accelerate

ageing [3].

There are mainly two mechanisms by which yeast handles damage accumulation. Firstly, it divides asymmetrically, giving rise to a mother cell that is larger than the daughter cell [5]. Secondly, it has been observed that some unicellular organisms, including yeast, have a retention mechanism that enables them to retain a larger amount of damaging factors in the mother cell during cell division [5]. In this way, the age of a newly formed daughter cell is reset.

The retention of some ageing factors, such as ERCs, seems to be dependent upon the protein sir2p [11]. It has been shown that *sir2Δ S. cerevisiae* mutants have a reduced replicative lifespan compared to wild-type yeast, with one explanation being the accumulation of ERCs [12]. The mean RLS of the *sir2Δ* strain has been found to be 14.0 with a standard deviation of 4.80 [6]. In contrast to this, a deletion of the gene *FOB1* in wild-type *S. cerevisiae* increases the yeast replicative lifespan, resulting in an RLS of 31.8 ± 11.8 [6], in part due to the reduction of ERC formation [13].

2.3 Model of replicative ageing

A mathematical model of replicative ageing in *S. cerevisiae* based on the accumulation of damage has been developed [5]. The model describes the rate of change of intact or functioning proteins, denoted P, and damage, D, inside of a cell, taking cell growth, damage formation and repair into account [5]. Included in the model are asymmetrical cell division and retention together with cell death. Cell division and death occurs if the amount of intact protein or damage in the cell, respectively, reaches a threshold. The system is described by the following equations.

$$
\begin{cases}
\dfrac{\mathrm{d}P}{\mathrm{d}t} = \mu(S)P\left(\dfrac{g \cdot D_{death} - D}{D_{death}}\right) - k_1 P + k_2 D \\
\dfrac{\mathrm{d}D}{\mathrm{d}t} = k_1 P - k_2 D \\
P(0) = f_P(s, re),\ D \in [0, D_{Death}],\ P \in [0, P_{div}] \\
D(0) = f_D(s, re),\ t \in \mathbb{R}_+,\ g \in (1, \infty),\ k_1, k_2 \in \mathbb{R}_+
\end{cases}
\tag{2.1}
$$

In System (2.1), k_1 is the damage formation rate, k_2 the repair rate, s the relative size of the mother cell after cell division and re a retention factor [5]. P_{div} and D_{death} are the thresholds for cell division and cell death. The term $\mu(S)P\left(\dfrac{g \cdot D_{death} - D}{D_{death}}\right)$ in System (2.1) models the cell growth, where the growth rate $\mu(S)$ is described by the Monod equation.

$$
\mu(S) = \mu_{max}\frac{S}{K_S + S}
\tag{2.2}
$$

In Equation (2.2), μ_{max} is the maximum specific growth rate, K_S the Monod constant and S is the concentration of substrate [5]. The maximum specific growth

rate of *S. cerevisiae* has been found to be $\mu_{max} = 0.5h^{-1}$ [14]. The second part of the growth term, $\left(\dfrac{g \cdot D_{death} - D}{D_{death}}\right)$, has been included in the model to account for a declining growth rate with age, where the parameter g is larger than 1 [5].

The model of replicative ageing has a continuous and a discrete part. The continuous part consists of the two differential equations describing the rate of change of intact protein, P, and damage, D. The discrete part of the model takes into account cell division and cell death, that are both described by instantaneous events. When the amount of damage in the cell reaches the threshold D_{death}, the cell dies and the simulation is terminated [5]. Similarly, when the amount of intact protein in the cell reaches the threshold P_{div}, the protein and damage content of the cell is reset based on the functions $f_P(s, re)$ and $f_D(s, re)$.

$$P(0) = f_P(s, re) = \begin{cases} sP_{div} - re(1 - s)D & [Mother] \\ (1 - s)P_{div} + re(1 - s)D & [Daughter] \end{cases} \qquad (2.3)$$

$$D(0) = f_D(s, re) = \begin{cases} (s + (1 - s)re)D & [Mother] \\ (1 - s)(1 - re)D & [Daughter] \end{cases} \qquad (2.4)$$

In Equations (2.3) and (2.4), D is the amount of damage in the cell prior to division [5].

The non-dimensionalised version of the model of replicative ageing is shown below.

$$\begin{cases} \dfrac{d\tilde{P}}{dt} = \tilde{\mu}(\tilde{S})\tilde{P}(g - \tilde{D}) - \tilde{k}_1\tilde{P} + \tilde{k}_2Q\tilde{D} \\[2mm] \dfrac{d\tilde{D}}{dt} = \dfrac{\tilde{k}_1}{Q}\tilde{P} - \tilde{k}_2\tilde{D} \\[2mm] \tilde{P}(0) = f_P(s, re), \ \tilde{D} \in [0, 1], \ \tilde{P} \in [0, 1] \\[1mm] \tilde{D}(0) = f_D(s, re), \ \tau \in \mathbb{R}_+, \ \tilde{k}_1, \tilde{k}_2 \in \mathbb{R}_+, Q \in (0, 1] \end{cases} \qquad (2.5)$$

$$f_P(s, re) = \begin{cases} s - re(1 - s)Q\tilde{D} & [Mother] \\ (1 - s) + re(1 - s)Q\tilde{D} & [Daughter] \end{cases} \qquad (2.6)$$

$$f_D(s, re) = \begin{cases} (s + (1 - s)re)\tilde{D} & [Mother] \\ (1 - s)(1 - re)\tilde{D} & [Daughter] \end{cases} \qquad (2.7)$$

In System (2.5), the following non-dimensional states and parameters has been introduced: $\tilde{P} = \dfrac{P}{P_{div}}$, $\tilde{D} = \dfrac{D}{D_{death}}$, $\tau = \mu_{max} \cdot t$, $Q = \dfrac{D_{death}}{P_{div}}$, $\tilde{k}_1 = \dfrac{k_1}{\mu_{max}}$, $\tilde{k}_2 = \dfrac{k_2}{\mu_{max}}$, $\tilde{S} = \dfrac{S}{S_{in}}$, $\tilde{K}_S = \dfrac{K_S}{S_{in}}$, $\tilde{\mu}(\tilde{S}) = \dfrac{\tilde{S}}{\tilde{K}_S + \tilde{S}}$. The parameter S_{in} is the substrate concentration in the inflow [5].

The model of replicative ageing was used to formulate three parameter constraints that need to be fulfilled in order for the cell to age. The first constraint is the starvation constraint that relates ageing to substrate uptake [5].

$$\tilde{k}_1 + \tilde{k}_2 < \tilde{\mu}(\tilde{S}) \cdot g \tag{2.8}$$

The two other constraints, the clonal senescence constraint and the immortality constraint, are based on the *lethal initial damage threshold, D_T*. If a cell has a damage proportion that is smaller than D_T after cell division, it should divide at least once more [5]. Based on a daughter cell with maximal fitness, the clonal senescence constraint is formulated as follows.

$$D_T(k_1, k_2)|_{P=(1-s)} > 0 \tag{2.9}$$

The immortality constraint ensures that the cell has a finite replicative life span. A cell that has a larger damage proportion than D_T after cell division should undergo cell death [5]. The immortality constraint is based on a mother cell of minimal fitness.

$$D_T(k_1, k_2)|_{P=(s-re(1-s)Q)} < s + (1-s)re \tag{2.10}$$

The above mentioned constraints form a triangle in the $(k_1,\ k_2)$ parameter space, *the triangle of ageing*, within which the cell ages [5].

2.4   Parameter estimation

In order for the model of replicative ageing to be biologically relevant, values of the model parameters need to be determined. One way to find values of model parameters is by literature search. However, a higher biological relevance is obtained if the model is fit to experimental data with the help of optimisation. For this purpose, experimental data of single-cell area of growing *S. cerevisiae* has been generated. In addition to this, data of a large-scale analysis of the replicative lifespan of 4,698 *S. cerevisiae* single-gene deletion strains [15] is at hand.

2.4.1   Optimisation

Optimisation is the process of finding, or getting close to, an optimum of an objective function that depends on the parameter values that are to be determined [16]. In continuous optimisation, the objective function and the constraints are continuous and the ground set is closed and convex [16].

When the purpose of an optimisation problem is to minimise the objective function, one approach is to take successive steps in a descending direction in the solution space until a stop condition is reached [16]. In this case, the parameter values are initialised at a chosen point in the solution space and updated according to the equation below.

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \cdot \mathbf{p} \tag{2.11}$$

In the equation, $\mathbf{x}_k$ is the set of parameter values at step k and α is the step size [16]. The step size is used to ensure that $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$, where one of the simpler update rules is to start with $\alpha = 1$ and then continuously bisect α until the condition is satisfied. The last parameter in Equation (2.11), $\mathbf{p}$, is the descent direction [16].

In Newton's method, the descent direction is based on that the objective function, f, is twice differentiable and is defined as $\mathbf{p} = -\dfrac{\nabla f(\mathbf{x}_k)}{\nabla^2 f(\mathbf{x}_k)}$, where $\nabla^2 f(\mathbf{x}_k)$ is the Hessian of f [16].

In cases where the Hessian is computationally costly, it can be approximated by a matrix $\mathbf{B}_k$ in a quasi-Newton method [16]. Based on a first-order Taylor expansion of the Hessian, $\mathbf{B}_k$ can be calculated from the gradient of f [16].

$$\mathbf{B}_k(\mathbf{x}_k - \mathbf{x}_{k-1}) = \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1}) \tag{2.12}$$

In the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, $\mathbf{B}_k$ is updated as follows.

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \frac{(\mathbf{B}_k\mathbf{s}_k)(\mathbf{B}_k\mathbf{s}_k)^T}{\mathbf{s}_k^T\mathbf{B}_k\mathbf{s}_k} + \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k} \tag{2.13}$$

In Equation (2.13), $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$ [16].

A number of different termination criteria can be applied to an optimisation algorithm in order to decide when to terminate the search. It is common to observe the changes in the gradient, the function value and the parameter values from one iteration to the next and stop if one or several of them are small or close to zero [16].

In the case of fitting a mathematical model to experimental data, the objective function depends on the values of the model parameters. If it is assumed that the residuals between the experimental data points and the estimated data points obtained from the model, $e_i = z_i - y_i$, are normally distributed, the objective function can be based on a likelihood function.

$$f(\mathbf{e}|\mathbf{p}) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2\sigma_i^2}((z_i - y_i) - \mu_i)^2} \tag{2.14}$$

Here, $\mathbf{e}$ is the vector of residuals e_i, while σ_i and μ_i are the variance and mean, respectively, of residual e_i [17]. The objective is to find the set of parameters $\mathbf{p}$ that maximises the likelihood function. Observe that if the means, μ_i, of the normal distributions are assumed to be zero and $w_i = 1/\sigma_i^2$, this is the same as to minimise the weighted least squares criteria.

$$LS(\mathbf{p}) = \sum_{k=0}^{N} w_k(z_k - y_k(\mathbf{p}))^2 \tag{2.15}$$

In Equation (2.15), $\mathbf{p}$ is the set of parameters, N is the number of data points, z_k is the kth measured data point and y_k is the kth estimated data point [17].

When the purpose of the optimisation is to find parameter values within a defined region, constrained optimisation can be applied. Interior point algorithms visually adds a barrier in the parameter space to keep the solution within the feasible set. This is achieved by the addition of a penalty term to the objective function.

$$\chi s(\mathbf{x}) \approx \nu \hat{\chi} s(\mathbf{x}) := \begin{cases} \nu \sum_{i=1}^{m} \phi[g_i(\mathbf{x})], & \text{if } g_i(\mathbf{x}) < 0, \, i = 1, ..., m \\ +\infty, & \text{otherwise} \end{cases} \qquad (2.16)$$

The function χs is called the *indicator function* of the set S where S is the feasible set [16]. In Equation (2.16), $\mathbf{x}$ is the parameter vector, ν is a penalty parameter, $g_i(\mathbf{x})$ is constraint i and m is equal to the number of constraints [16]. The function ϕ is characteristic for the type of interior point method applied.

2.5 Parameter identifiability

When it comes to the estimation of parameters in the modelling of a biological system, an identifiability analysis is performed to determine if the parameters can be uniquely estimated based on the available input-output data [18]. An identifiability analysis can be either a priori or data-based. In an a priori approach, it is determined whether a parameter is *structurally* identifiable or not, meaning that its value can be estimated based on an ideal set of noise-free input-output data [17]. When the identifiability analysis is data-based, it is instead investigated if a parameter is *numerically* identifiable. A numerically identifiable parameter is a structurally identifiable parameter that can be estimated given real data with noise [17].

2.5.1 Structural identifiability

There are several approaches available that can be used to determine if a parameter is structurally identifiable. Approaches used for this purpose are either global, meaning that the identifiability holds for all possible values of the parameter, or local [18]. A package was developed for `Mathematica` that can be used to determine the structural identifiability of model parameters given a measured model output [19]. In this package, the method used is local and is based on the rank-test for structural identifiability. The basic idea is that a set of parameters are structurally identifiable if they can be uniquely determined based on the time derivatives of the inputs and outputs of the system.

2.5.2 Numerical identifiability

To determine if a parameter is numerically identifiable, a sensitivity analysis can be performed. The sensitivity analysis is used to decide which parameter values affect the model output, and thereby give a high relative parameter sensitivity, and which do not [17]. If the model output is not significantly affected by the variation of a

parameter, the sensitivity of that parameter is low and its value hard to identify [17]. To check this, it is common to observe the variance of the parameter estimate and to investigate how it correlates with other model parameters. A parameter estimate with a small variance and with low correlation with other parameter estimates is likely to be a good estimate [17]. The variances and correlations of parameters can be obtained from the covariance matrix where the non-diagonal elements can be used to calculate the correlations and the diagonal elements are the variances [17]. In order to get a relative measure of the variance of a parameter estimate, the *coefficient of variation* can be calculated. The coefficient of variation for a parameter, p_i, is equal to the quotient between the standard deviation, σ_i, and the parameter value.

$$cv(p_i) = \frac{\sigma_i}{p_i} \tag{2.17}$$

A lower bound on the covariance matrix can be found with the Fisher Information Matrix.

$$\mathrm{COV}(\hat{\mathbf{p}}) \geq F^{-1}(\hat{\mathbf{p}}) \tag{2.18}$$

In Equation (2.18), $\hat{\mathbf{p}}$ is the vector of parameter estimates and F is the Fisher Information matrix [17].

$$F(\mathbf{p}) = \left(\frac{\partial \mathbf{Y}}{\partial \mathbf{p}}\right)^T \boldsymbol{W} \frac{\partial \mathbf{Y}}{\partial \mathbf{p}} \tag{2.19}$$

The Fisher Information matrix is calculated with the Jacobian, $\dfrac{\partial \mathbf{Y}}{\partial \mathbf{p}}$, and the data covariance matrix, $\boldsymbol{W}$. The data covariance matrix have diagonal entries $w_{ii} = \dfrac{1}{\sigma_i^2}$ where σ_i^2 is the variance of data point z_i [17]. The non-diagonal elements of the matrix are the covariances of the data points. It can be noted that the Hessian of the objective function can be approximated using the same calculation as for the Fisher Information matrix [17].

$$\hat{\mathbf{H}} \simeq \left(\frac{\partial \mathbf{Y}}{\partial \mathbf{p}}\right)^T \boldsymbol{W} \frac{\partial \mathbf{Y}}{\partial \mathbf{p}} \tag{2.20}$$

The Jacobian is calculated as shown below.

$$\frac{\partial \mathbf{Y}}{\partial \mathbf{p}} = \begin{bmatrix} \frac{\partial y_1}{\partial p_1} & \frac{\partial y_1}{\partial p_2} & \cdots & \frac{\partial y_1}{\partial p_m} \\ \frac{\partial y_2}{\partial p_1} & \frac{\partial y_2}{\partial p_2} & \cdots & \frac{\partial y_2}{\partial p_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial p_1} & \frac{\partial y_n}{\partial p_2} & \cdots & \frac{\partial y_n}{\partial p_m} \end{bmatrix} \tag{2.21}$$

Here, y_i is the output obtained at time point t_i, p_j is parameter j, n is the number of data points and m the number of parameters [17].

The above-mentioned approach to a numerical sensitivity analysis is local since the Jacobian is evaluated at a given set of parameter values. This means that the conclusion regarding whether the model parameters are numerically identifiable or not can only be drawn around a given point in the parameter space. In contrast to this, there exist *global* identifiability approaches that allow the investigator to determine if a parameter is numerically identifiable on the entire set of feasible parameter values. Variance-based sensitivity analysis approaches aims at resolving which uncertain parameters that explain the most of the variance in the model output. One such approach is the calculation of Sobol' indices. These indices are based on a decomposition of the output variance into sums of contributions of first-order, second-order, third-order, and so on, parameter effects [20]. For a set of parameters $\mathbf{u} \subset 1, ..., d$, where d is the number of parameters, the Sobol' index is defined by the following equation.

$$S_{\mathbf{u}} = \frac{Var[G_{\mathbf{u}}(\mathbf{X_u})]}{Var[G(\mathbf{X})]} \tag{2.22}$$

In Equation (2.22), $G(\mathbf{X})$ is the model output as a function of $\mathbf{X}$ consisting of random vectors X_j with $j = 1, ..., d$ and $G_{\mathbf{u}}(\mathbf{X_u})$ is the model output of $\mathbf{X_u}$ consisting of random vectors X_j for all $j \in \mathbf{u}$ [20]. This means that the term $Var[G(\mathbf{X})]$ is the total variance of the model output while $Var[G_{\mathbf{u}}(\mathbf{X_u})]$ is the variance explained by the parameters in $\mathbf{u}$. Monte Carlo-based estimators can be used to calculate these quantities. The first-mentioned quantity can be calculated as follows.

$$Var[G_{\mathbf{u}}(\mathbf{X_u})] = \frac{1}{2N} \sum_{i=1}^{N} [(G(\mathbf{A})^{(i)} - \mu(\mathbf{A}))^2 + (G(\mathbf{B})^{(i)} - \mu(\mathbf{B}))^2]$$

$$\mu(\mathbf{A}) = \frac{1}{N} \sum_{i=1}^{N} G(\mathbf{A})^{(i)} \tag{2.23}$$

$$\mu(\mathbf{B}) = \frac{1}{N} \sum_{i=1}^{N} G(\mathbf{B})^{(i)}$$

In Equation (2.23), $\mathbf{A}$ and $\mathbf{B}$ are sampling matrices of size $N \times d$ where each row is a sampled point in the parameter space [20].

The variance explained by a single parameter j can be estimated by the following equation.

$$S_j \cdot Var[G(\mathbf{X})] = \frac{1}{N} \sum_{i=1}^{N} G(\mathbf{B})^{(i)} (G(\mathbf{A_{B_{(\{j\})}}})^{(i)} - G(\mathbf{A})^{(i)}) \tag{2.24}$$

In the equation, $\mathbf{A_{B_{(\{j\})}}}$ is the sampling matrix $\mathbf{A}$ where column j has been replaced with the corresponding column from matrix $\mathbf{B}$ [20].

The rows of matrices $\mathbf{A}$ and $\mathbf{B}$ in Equations (2.23) and (2.24) are samples from the parameter space $S \in \mathbb{R}^d$, with d the number of uncertain parameters. For the

construction of sampling matrices, one commonly used method is latin hypercube sampling. The columns of a Latin Hypercube Sample (LHS) is initialised with d random permutations of the integers 1,...,n where n is the number of sampled points [21]. Each element $d_j^{(i)}$ in row i of column j is then transformed to get the distribution, F, of parameter j.

$$x_j^{(i)} = F^{-1}\left\{\frac{d_j^{(i)} - 1}{n - 1}\right\}$$

Here, $x_j^{(i)}$ is the final element in row i, column j of the LHS [21].

Chapter 3

Methods

Three data sets were used in the parameter estimation process in order to fit the model of replicative ageing to the following *S. cerevisiae* strains: wild-type, *sir2Δ* and *fob1Δ*. Continuous and constrained optimisation was applied with a quasi-Newton method in order to fit the model to experimental data. A structural identifiability analysis was performed as well as numerical sensitivity analysis on a local and global scale. In addition to this, replicative lifespan data of yeast deletion strains was analysed and categorised. To connect the RLS data analysis to the model of replicative ageing, it was investigated how the parameters of the model affect the RLS of the model simulations.

3.1 Experimental data

The cell area data generated with the purpose of parameter estimation consists of single-cell growth curves with the area given in μm^2 and the time given in minutes. The data was obtained with holographic microscopy for wild-type yeast and the *sir2Δ* and *fob1Δ* deletion strains. The number of curves and the total number of data points for each of the data sets are shown in Table 3.1.

Table 3.1: An overview of the data sets available for wild-type yeast and the *sir2Δ* and *fob1Δ* deletion strains. The table shows the number of daughter cell and mother cell curves and the total number of data points in each data set.

	Daughter cells	Mother cells	Data points
Wild-type	6	9	11,188
sir2Δ	12	10	30,221
fob1Δ	27	23	89,391

Examples of a daughter cell curve and a mother cell curve from the wild-type yeast data set are shown in the figure below.
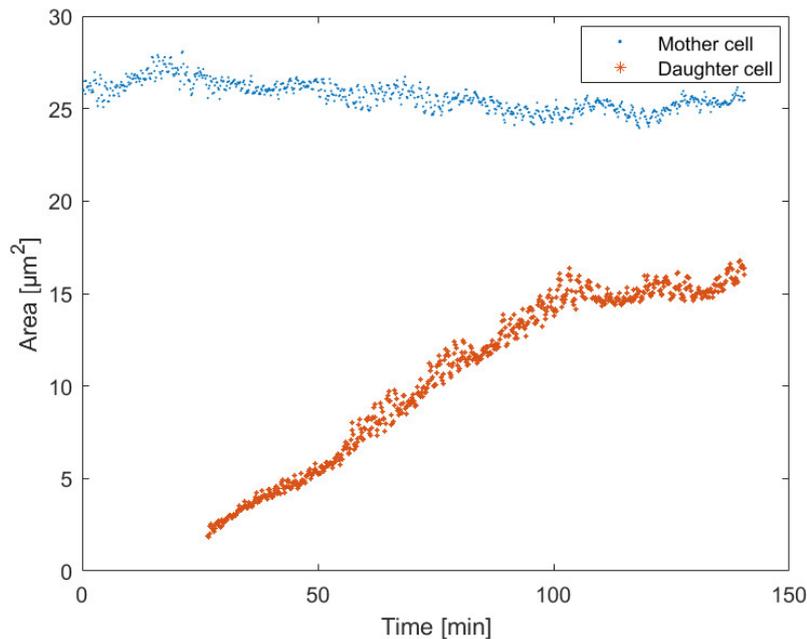
Figure 3.1: An example of a daughter cell curve and a mother cell curve from the wild-type yeast single-cell area data set.

3.2 Optimisation

The built-in `MATLAB` function `fmincon` was used to perform constrained continuous optimisation on the non-dimensionalised version of the model of replicative ageing, System (2.5). For simplification, all of the variables and parameters in the non-dimensionalised model will from hereon be denoted as in the original model. An interior point method was utilised to infer the starvation constraint, Equation (2.8), on the optimisation. Upper and lower bounds where set on the parameters as follows: $Q \in [0,1]$, $g \in [1,10]$ $k_1 \in [0,1]$ and $k_2 \in [0,1]$. The Hessian of the objective function was updated with the BFGS method. To solve the system of ODE:s, the ODE-solver `ode45` in `MATLAB`, that applies a Runge-Kutta method, was used.

For the purpose of the optimisation, three data sets were created; one for wild-type yeast and one for each of the *sir2*Δ and *fob1*Δ deletion strains. Of the available experimental data over single-cell area, only those curves that were identified as belonging to daughter cells were used for the optimisation. It was assumed that the cell area is proportional to the mass of the cell and that the daughter cells have $D \sim 0$. As a result of these assumptions, the cell area was approximated with the total amount of intact protein, P, in the cell.

The experimental data was non-dimensionalised in order to match the output of the model. The area of all of the data curves was scaled with the generation time. It was assumed that the growth curves are approximately linear and that the cell area at the generation time should be equal to 1, the maximum amount of intact protein in the non-dimensionalised model. Scaling was performed for each data curve as follows.

$$\boldsymbol{y} = \frac{t_{end}}{A_{end} \cdot g_t} \cdot \boldsymbol{A} \qquad (3.1)$$

In the equation, $\boldsymbol{y}$ is the vector of scaled data points, t_{end} and A_{end} are the time and the area of the last data point of the curve, respectively, and $\boldsymbol{A}$ is the vector of non-scaled data points. Based on a daughter cell curve from the *sir2*Δ data set for which a division was observed, the generation time, g_t, of the daughter cells was estimated to 120 minutes. Since the cells are of an early generation, it was judged valid to assume that the generation time is the same for all of the three investigated strains. The time was scaled as $\tau = t \cdot \mu_{max}$ with $\mu_{max} = 0.5h^{-1}$. The Monod constant was set to $K_S = 2/3$ and the substrate concentration to $S = 1$.

The approach was to fit the model to all of the growth curves in a data set as follows. For each growth curve, a simulation was run with initial conditions $(P_0, D_0) = (y_0, 0)$ where y_0 is the scaled area of the first data point in the curve. The simulation was terminated when P reached the y-value of the last data point. Linear interpolation with extrapolation was applied to the model output, P, with the `MATLAB` function `interp1` to obtain model data point estimates at the time points of the experimental data. The least-squares equation, Equation (2.15), with weights $w_k = 1$ was used for the objective function. The objective function was calculated as a sum over the least squares, averaged over the number of data points of each individual curve.

3.2.1 Optimisation without RLS constraint

The optimisation approach described above was run for each of the three data sets. To avoid ending up in a local optimum, the optimisation was run using 150 sets of initial conditions with Q ranging from 0.2 to 0.4 with step size 0.2, g ranging from 1.5 to 2.5 with step size 0.5 and k_1, k_2 ranging from 0.05 to 0.85 with a step of 0.2. The minima generating the smallest objective function values were selected. For these, simulations were run to obtain the RLS of the model given the selected sets of parameters. The RLS was found as follows; initial conditions were set to $(P_0, D_0) = (1 - s, 0)$ with $s = 0.75$ and simulations were run until $D = 1$. Each time $P = 1$ was reached, the variables were reset according to Equations (2.6) and (2.7). The RLS was calculated based on the number of resets made, with the maximum set to 1,000. The retention constant was set to $re = 0.875$ for the wild-type and the *fob1*Δ strains. For the *sir2*Δ strain it was assumed that the retention mechanism is non-functional and $re = 0$ was used.

3.2.2 Optimisation with RLS constraint

In order to select for parameter values that generate replicative lifespans close to those of the investigated strains, a penalty term was added to the objective function. The penalty term was formulated as $\dfrac{|RLS_{mean} - RLS(\boldsymbol{p})|}{RLS_{mean}}$, where RLS_{mean} is the mean experimental RLS of the strain and $RLS(\boldsymbol{p})$ is the replicative lifespan obtained from simulation as explained in the previous subsection. The optimisation was

performed as in the preceding case. However, the parameter Q was set to a fixed value of 0.33, and only the values of g, k_1 and k_2 were optimised for. A total of 75 initial conditions were tested.

3.3 Parameter identifiability

3.3.1 Structural identifiability analysis

The structural identifiability package developed in `Mathematica` [19] was used to determine which parameters in the continuous part of the model of replicative ageing that are structurally identifiable given that $y = P$ is the measured output. The initial conditions were set as $(P_0, D_0) = (1 - s, 0)$ with $s = 0.75$.

3.3.2 Numerical identifiability analysis

For the minima found with optimisation, a numerical identifiability analysis was performed. A lower bound on the covariance matrix of the parameter estimates was found with Equation (2.18). Based on the experimental data, the variance of the data points was estimated to $\sigma_i^2 = 0.04$ for the wild-type yeast data set and $\sigma_i^2 = 0.01$ for the other two data sets. It was assumed that the data points are non-correlated. The rate of change of the sensitivities were found by derivation of the model using $\frac{\partial}{\partial t}(\frac{\partial \tilde{P}}{\partial p_i}) = \frac{\partial}{\partial p_i}(\frac{\partial \tilde{P}}{\partial t})$. For the full derivations and ODE:s, see Appendix 1. The ODE:s were solved with the built-in `MATLAB` function `ODE45` with initial conditions $(P_0, D_0, (\frac{dQ}{dt})_0, (\frac{dg}{dt})_0, (\frac{dP}{dt})_0, (\frac{dD}{dt})_0) = (y_0, 0, 0, 0, 0, 0)$ where y_0 denotes the value of the scaled area of the first experimental data point in a data curve. The sensitivities were interpolated to obtain values at the time points of the experimental data. The entries of the lower bound on the final covariance matrix were taken as averages over the values obtained for each of the curves in a data set. Correlation matrices were calculated from the lower bounds on the covariance matrices. The coefficient of variation for each model parameter were calculated from the diagonal of the covariance matrix according to Equation (2.17).

A global sensitivity analysis was performed in order to investigate how insecurities in the values of the model parameters affect the value of the objective function. Individual Sobol' indices were estimated for the investigated parameters with Equations (2.23) and (2.24). LHS sampling was used to construct sampling matrices of 1,000 observations. The parameters were assumed to have uniform distributions with $Q \in [1 \cdot 10^{-5}, 1]$, $g \in [1, 10]$ and $k_1, k_2 \in [0, 1]$. A total of 10 indices for each of the model parameters were estimated for all of the three data sets investigated.

3.4 RLS data analysis

A summary of RLS experiments performed on *S. cerevisiae* gene-deletion strains was available [15]. Each experiment consisted of a set of RLS measurements obtained on a given gene deletion strain as well as a set of measurements obtained on wild-type

yeast. The experiments used for analysis were selected as follows. Only the experiments on strains with single gene-deletions were investigated. The experiments with a sample size of five cells or smaller and those where the reference strain had a mean RLS outside of the normal range of wild-type yeast (20-30) were removed. Experiments obtained on the same gene-deletions strains regardless of mating type were merged and new mean replicative lifespans were calculated for the gene-deletion strains as well as the reference measurements. P-values were calculated using the Wilcoxon signed rank test and experiments yielding a p-value smaller than 0.05 were kept for further analysis. Out of the gene-deletions strains that were left, strains were grouped as short-lived if they had a decrease in RLS of 30 % or more compared to the reference strain. Strains that had an increase in RLS of 30 % or more were grouped as long-lived. For the purpose of investigating the gene products of the deleted genes of the two groups, a cell map tool [22] was used to group the gene products based on cellular function.

In conjunction to the analysis of the RLS data, it was investigated how the parameters Q, g, k_1, k_2, s and re influence the variance of the RLS of model simulations. Sobol' indices were estimated for the model parameters over 10 samples with 1,000 observations. The sampling matrices were constructed with latin hypercube sampling and the distributions of the parameters were taken as uniform distributions with Q, k_1, k_2, $re \in [0, 1]$, $g \in [1, 3]$ and $s \in [0.5, 1]$. Simulations were performed with the non-dimensionalised model of replicative ageing to find the RLS, using the sampling points of the constructed matrices, given that the initial amount of damage in a cell is equal to zero. The maximum RLS of the simulations was set to 1,000, assuming that a cell is immortal if the RLS is equal to or larger than that.

Chapter 4

Results

This chapter presents the outcome of the optimisation, including parameter estimates for all of the three investigated strains. The results of the identifiability and sensitivity analyses are used to evaluate the certainty of the estimates and the two optimisation approaches applied are compared. Lastly, the results from the RLS data analysis are presented.

4.1 Optimisation

4.1.1 Optimisation without RLS constraint

The minima and objective function values obtained with optimisation for each investigated strain are shown in Table 4.1. Shown in the table are also the replicative lifespans found with model simulations using the given sets of parameters.

Table 4.1: Minima found with optimisation. Shown in the table are the objective function and parameter values at the minima and the RLS obtained with model simulations using the given sets of parameters.

	LS value	Q	g	k_1	k_2	RLS
Wild-type	0.06826	0.04946	7.285	0.9999	0.9999	0
***sir2*Δ**	0.03942	0.04097	7.977	0.9683	0.9920	0
***fob1*Δ**	0.09692	0.04493	7.353	0.9999	0.9999	0

The differences between the values of the parameter estimates obtained for the three strains are minor, Table 4.1. For all three strains, the estimated values for the parameter Q are small, which would mean that the cells barely have any resilience to damage accumulation. In addition, the values of g are higher than expected; the prediction would be that the parameter should have a value above, but still close to, 1. The reasoning behind this is that an old cell with a damage amount of $D \sim 1$ should have a slow growth, such that the growth term $\mu(S)P(g - D)$ is close to zero. Moreover, the values of k_1 and k_2 are close to their upper limits and almost equal for all strains. As a consequence of this, in combination with the assumption that $D \sim 0$, more damage will form than will be repaired since the damage formation is proportional to P while the repair of damage is proportional to D. It is indeed expected that damage will accumulate with age. However, a

higher damage formation in combination with a low damage resilience, Q, leads to the violation of the clonal senescence constraint, equation (2.9).

In order to visualise how the model fits to the experimental data given the obtained minima, Table 4.1, model simulations were run and the output, P, was plotted together with the data curves, Figure 4.1. The simulated curves were initialised at $(P_0, D_0) = (y_0^i, 0)$ and terminated at $P = y_{end}^i$ where y_0^i and y_{end}^i are the values of the scaled cell area of the first and the last data point, respectively, of experimental curve i.

Figure 4.1: Plots of the experimental data together with simulated output P. Simulations were run with initial conditions $(P_0, D_0) = (y_0^i, 0)$ and terminated at $P = y_{end}^i$ where y_0^i and y_{end}^i are the values of the scaled area of the first and the last data point, respectively, of experimental curve i. The colours represent individual curves, where curve i has the same colour in both the experimental and the simulated case. Parameter values were set according to the minima obtained with optimisation. Top panel: wild-type. Middle panel: *sir2Δ*. Bottom panel: *fob1Δ*. To the left in each panel: all curves in the data set and their simulated correspondences. To the right: a representative example.

The cells in the model simulations die before they have time to divide, Figure 4.1. The reason for this is that the damage content becomes too large. This, in addition, renders the assumption of zero damage invalid. The objective function used in the optimisation does not take into account what happens in the simulations beyond the last experimental time point, hence the bad fits beyond this point are overlooked during the parameter estimation process.

4.1.2 Optimisation with RLS constraint

The results from the optimisation with the RLS constraint are shown in Table 4.2.

Table 4.2: Minima found by optimisation with the addition of an RLS constraint to the objective function. Shown in the table are the objective function and parameter values at the minima and the RLS obtained with model simulations using the given sets of parameters. The parameter Q was set to 0.33.

	LS value	g	k_1	k_2	RLS
Wild-type	0.1975	4.230	0.3569	0.7471	23
$sir2\Delta$	0.4832	6.056	0.8596	0.2043	9
$fob1\Delta$	0.2596	4.185	0.2356	0.4336	32

Next, the values of the parameter estimates obtained with the two optimisation approaches, Tables 4.2 and 4.1, are compared. The estimates found with the RLS constrained optimisation approach are judged as more reasonable. The estimates for the parameter g are still higher than expected, but at least smaller in comparison. The most difference is seen in the estimates for k_1 and k_2, which are no longer close to their upper limits. For wild-type yeast and the $fob1\Delta$ strain, the value of k_2 is approximately twice as large as that of k_1. This means that the repair rate is large enough to compensate for the damage formation. In addition to this, with Q set to 0.33, the cells have some damage resilience, meaning that they can proliferate even in the presence of damage inside of the cell.

In the case of the *sir2Delta* strain, the estimated value of the damage formation rate, k_1, is larger than the estimate for the repair rate, k_2. This implies that the $sir2\Delta$ strain does not repair damage to the same extent as the other two investigated strains do. However, note that it was assumed that the $sir2\Delta$ strain have a malfunctioning retention mechanism, with $re = 0$. As a consequence, less damage accumulation occur inside of the mother cell and the need for damage repair is seemingly smaller. On the other hand, with a non-functional retention mechanism, daughter cells are formed with more damage and therefore get a possibly shorter replicative lifespan than their mothers. It would as a result have been of interest to evaluate the RLS of the model as an average over a whole population where not all daughter cells were born with zero damage. In parallel to this, the assumption of zero damage might not be valid for the $sir2\Delta$ strain. For further investigations, the assumption that $D \sim 0$ should be removed so that the model could be fit to mother cell data or to data generated on cells of older generations.

The replicative lifespans found by simulations using the minima obtained with optimisation, Table 4.2, shows that it is possible to fit the model to experimental

data and at the same time generate parameter estimates that give a model RLS close to that of the investigated strain. For the wild-type yeast and the *fob1*Δ strain, the replicative lifespans found by model simulations are close to the experimental mean values. However, the RLS obtained for the *sir2*Δ strain is slightly outside of the range 14.0 ± 4.80 found by experiments.

The fit of the model output P, using the found minima, to experimental data was visualised, Figure 4.2.
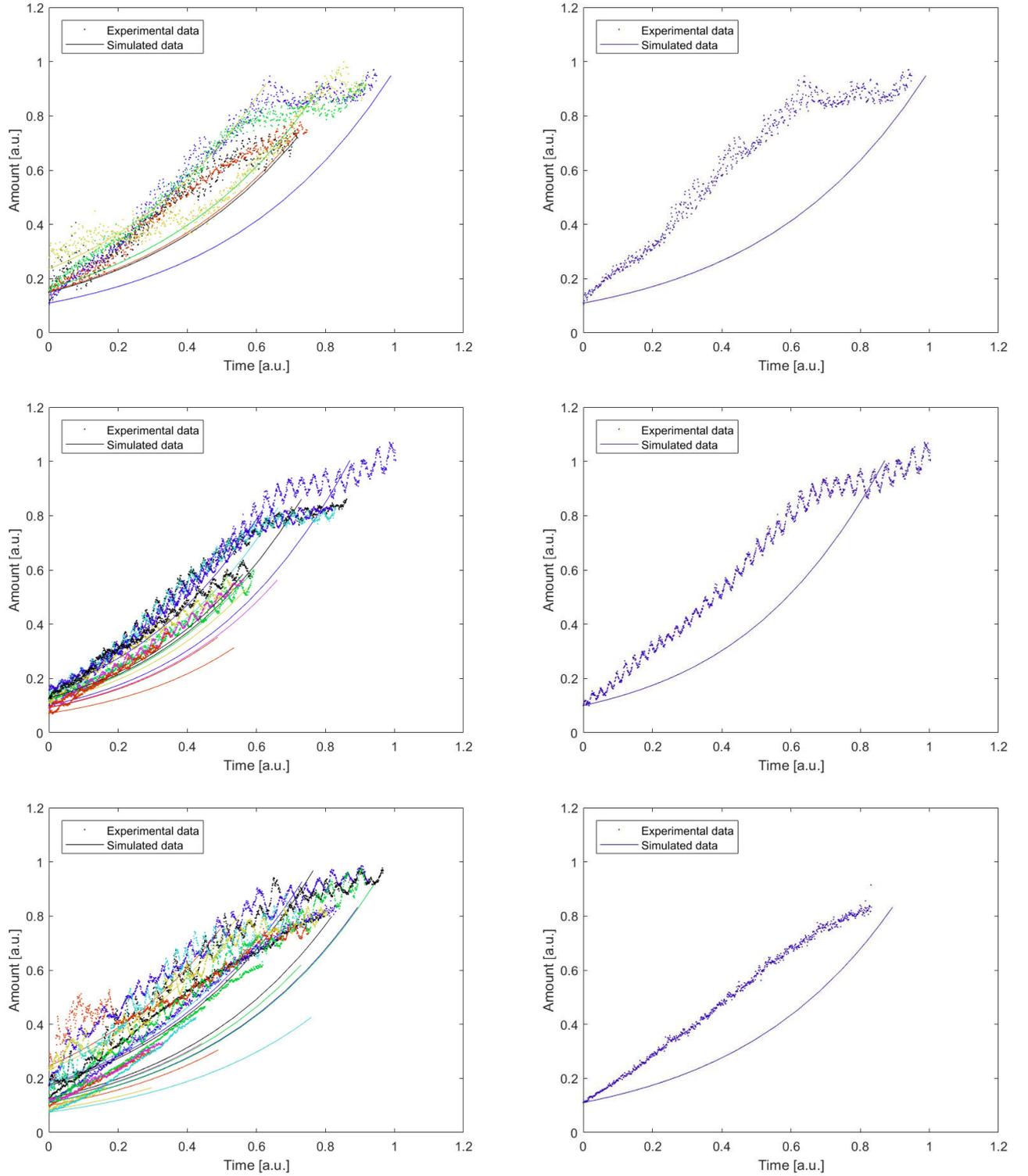
Figure 4.2: Plots of the experimental data together with simulated output P. Simulations were run with initial conditions $(P_0, D_0) = (y_0^i, 0)$ and terminated at $P = y_{end}^i$ where y_0^i and y_{end}^i are the values of the scaled cell area of the first and the last data point, respectively, of experimental curve i. The colours represent individual curves, where curve i has the same colour in both the experimental and the simulated case. Parameter values were set according to the minima obtained by optimisation with an RLS constraint. Top panel: wild-type. Middle panel: $sir2\Delta$. Bottom panel: $fob1\Delta$. To the left in each panel: all curves in the data set and their simulated correspondences. To the right: a representative example.

4. Results

The fits obtained with the RLS constrained optimisation procedure are not as good as the ones obtained with the first approach, considering only the time span of the experimental data. This is reflected in the fit plots, Figures 4.1 and 4.2, as well as in the objective function values at the minima. Specifically, the experimental growth curves seem to have a logarithmic shape while the simulated growth curves are more close to being exponential.

4.2 Parameter identifiability

4.2.1 Structural identifiability

The local evaluation of the structural identifiability was performed for the parameters in the continuous part of the model of replicative ageing. It showed that all of the parameters Q, g, k_1 and k_2 are structurally identifiable given that $y = P$ is the measured output.

4.2.2 Numerical identifiability

4.2.2.1 Optimisation without RLS constraint

The coefficients of variation for Q, g, k_1 and k_2, Table 4.3, were calculated as the quotients between the estimated lower bounds on the standard deviations and the values of the parameter estimates.

Table 4.3: Coefficients of variation for the parameter estimates obtained by optimisation without RLS constraint.

	Q	g	k_1	k_2
Wild-type	5.440	1.123	5.426	2.476
sir2Δ	2.941	0.5515	2.999	0.8451
fob1Δ	2.131	0.4358	2.114	0.6819

All of the coefficients of variation calculated for the estimates of Q and k_1 are larger than 1, Table 4.3, indicating that these parameters are locally non-identifiable based on the experimental data. However, for the two deletions strains, the coefficients of variation of g and k_2 are smaller than 1. This implies that g and k_2 are numerically identifiable for the deletion strain data sets.

The correlation matrices for the parameter estimates, Tables 4.4 to 4.6, were found from the lower bounds on the covariance matrices.

Table 4.4: Correlation matrix for the parameter estimates obtained for wild-type yeast by optimisation without RLS constraint.

$CORR^{wild-type}(\hat{\mathbf{p}}) =$

	Q	g	k_1	k_2
Q	1			
g	0.9939	1		
k_1	0.9933	0.9998	1	
k_2	-0.6468	-0.5621	-0.5551	1

Table 4.5: Correlation matrix for the parameter estimates obtained for *sir2*Δ by optimisation without RLS constraint.

$CORR^{sir2\Delta}(\hat{\mathbf{p}}) =$

	Q	g	k_1	k_2
Q	1			
g	0.9973	1		
k_1	0.9976	0.9999	1	
k_2	-0.5782	-0.5194	-0.5205	1

Table 4.6: Correlation matrix for the parameter estimates obtained for *fob1*Δ by optimisation without RLS constraint.

$CORR^{fob1\Delta}(\hat{\mathbf{p}}) =$

	Q	g	k_1	k_2
Q	1			
g	0.9977	1		
k_1	0.9984	0.9997	1	
k_2	-0.8530	-0.8202	-0.8232	1

The correlation matrices indicate a high correlation between the parameter estimates for all of the data sets. An exception can be observed for the damage repair rate, k_2, in the cases of the wild-type yeast and the *sir2*Δ strain. The estimates of this parameter are in these cases only moderately correlated with the other parameter estimates. Likewise, for the wild-type yeast, the estimate for k_2 has a large coefficient of variation, Table 4.3.

The Sobol' indices for parameters Q, g, k_1 and k_2 were estimated for each of the three strains investigated, using the objective function as the output. The values of the indices are visualised in a boxplot, Figure 4.3.
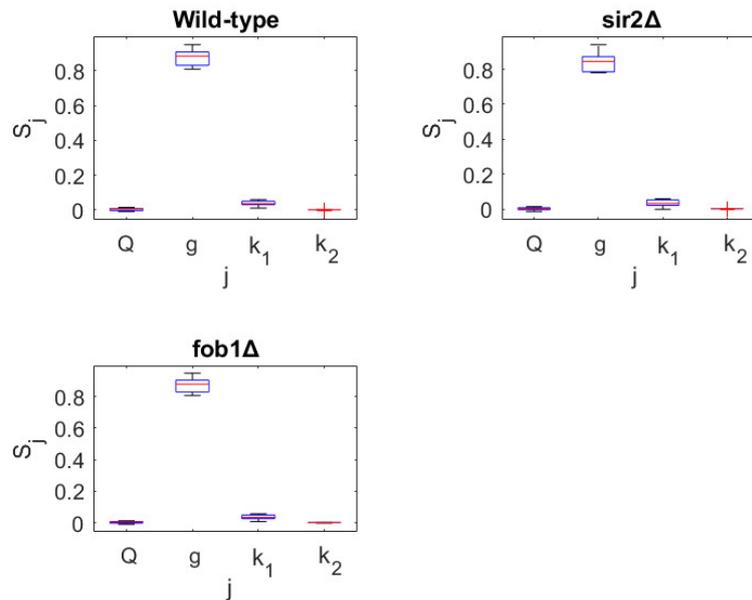
Figure 4.3: Estimated individual Sobol' indices, S_j, for the model parameters with the objective function as the output. The values were obtained for 10 samples of size 1,000, generated with LHS sampling.

The estimated Sobol' indices, Figure 4.3, show that the parameter g explains most of the variance in the objective function. All of the other parameters explain only a small fraction of the variance observed. This holds especially for parameters Q and k_2, that explain less than 1% of the observed variance for all of the three data sets. As a result, g is supposedly the parameter that is the easiest to identify on a global scale in the optimisation problem. Note that the results might be affected by the broader span of investigated values for g. In addition, there might be significant second-order, or higher, effects not investigated here.

4.2.2.2   Optimisation with RLS constraint

The coefficients of variation for the estimates of g, k_1 and k_2 are shown in Table 4.7.

Table 4.7: Coefficients of variation for the parameter estimates obtained with optimisation inferring a constraint on the RLS.

	g	k_1	k_2
Wild-type	0.2177	1.803	4.248
sir2Δ	0.3967	1.905	11.20
fob1Δ	0.2725	3.286	9.719

Although the coefficients of variation for the parameter g are smaller than 1, Table 4.7, the coefficients for k_1 and especially k_2 are large. This indicates that only g is numerically identifiable.

The correlation matrices for the three sets of parameter estimates, Tables 4.8 to 4.10, were obtained from the lower bounds on the covariance matrices.

Table 4.8: Correlation matrix for the parameter estimates obtained for wild-type yeast by optimisation with an RLS constraint.

$CORR^{wild-type}(\hat{\mathbf{p}}) =$

	g	k_1	k_2
g	1		
k_1	0.9417	1	
k_2	0.3324	0.6290	1

Table 4.9: Correlation matrix for the parameter estimates obtained for *sir2*Δ by optimisation with an RLS constraint.

$CORR^{sir2\Delta}(\hat{\mathbf{p}}) =$

	g	k_1	k_2
g	1		
k_1	0.9944	1	
k_2	0.9046	0.9442	1

Table 4.10: Correlation matrix for the parameter estimates obtained for *fob1*Δ by optimisation with an RLS constraint.

$CORR^{fob1\Delta}(\hat{\mathbf{p}}) =$

	g	k_1	k_2
g	1		
k_1	0.9757	1	
k_2	0.6216	0.7770	1

The correlations between the parameter estimates are moderate to high, Tables 4.8 to 4.10. As in the previous case, less correlation is observed for the estimates of the damage repair rate, k_2, but only for the wild-type yeast and the *fob1*Δ deletion strain. Likewise, the coefficients of variation obtained for the estimates of this parameter are large, Table 4.7.

The estimated individual Sobol' indices for the objective function with an RLS constraint are displayed in a boxplot in Figure 4.4.
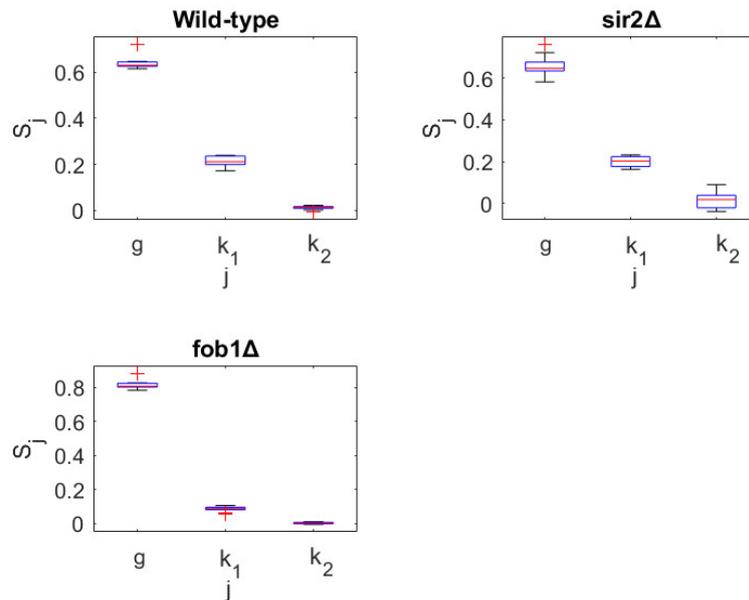
Figure 4.4: Estimated individual Sobol' indices, S_j, for the model parameters. The investigated output is the objective function with an inferred RLS constraint. The values were obtained for 10 samples of size 1,000, generated with LHS sampling.

If the estimated Sobol' indices obtained for the objective function with and without an RLS constraint are compared, it can be deduced that the parameter g has a smaller relative effect on the objective function when an RLS constraint is added to it. On the other hand, the damage formation rate, k_1, explains a higher percentage of the variance observed in the output with the constraint inferred. However, the variance explained by the parameter k_2 is still small in comparison. This implies that the damage repair rate, k_2, is hard to identify with the given objective function and experimental data. The individual Sobol' indices does not sum to one for any of the data set, indicating that there are second-order, or higher, effects present.

4.3   RLS data analysis

Out of the replicative lifespan data, a total of 668 single-gene deletion strains were used for analysis [15]. Of these strains, 136 were grouped as short-lived (a decrease in RLS of 30 % or more compared to wild-type yeast) whereof one was the *sir2Δ* strain. A total of 40 deletion strains were grouped as long-lived (an increase in RLS of 30 % or more compared to wild-type yeast). The *fob1Δ* strain had a mean increase of 25 % in RLS and was therefore not classified as long-lived in this analysis. Cell maps over the gene products of the deleted genes are presented in figure 4.5, where the cell map on the top shows all of the gene products (yellow) and the one in the bottom displays the gene products whose deletions resulted in short-lived (red) and long-lived (green) strains.
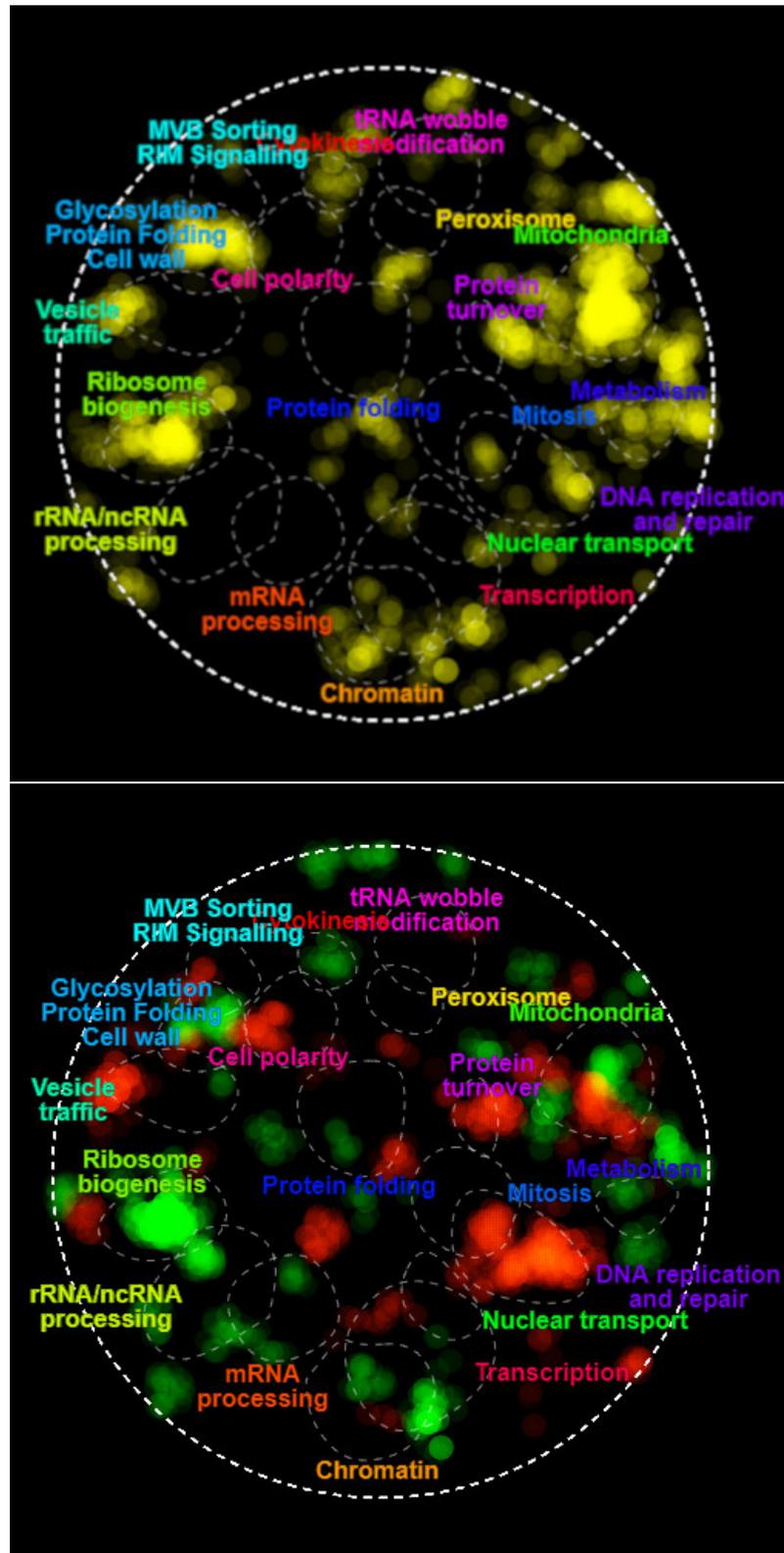
Figure 4.5: Cell maps over the products of the deleted genes of investigated single-gene deletions strains. Top: cell map over the products of all the deleted genes. Bottom: cell map over the gene products whose removal led to a significant decrease (red) or increase (green) in RLS.

It can be concluded that not all of the areas of the cell map are represented by the gene deletions investigated, top cell map in Figure 4.5, meaning that no general conclusions can be drawn about the gene product distribution. However, a distinction can be observed between the long-lived and the short-lived groups, bottom cell map in Figure 4.5, which is not likely to be a result of insufficient representation. Two main areas are revealed where either the short-lived or the long-lived group dominates; *DNA replication and repair* and *Ribosome biogenesis*.

There is one intuitive explanation to why the deletion of genes involved in DNA replication and repair leads to a decrease in the RLS. If the replication or the DNA repair process is faulty, errors will be surpassed. This will lead to the production of damaged proteins and, ultimately, to the accumulation of damage. In the perspective of the model of replicative ageing, this would be the same as to have a high damage formation rate, k_1. Since a high damage formation rate will lead to a faster increase in the amount of damage D inside of the cell, the damage threshold will be reached at an earlier stage in the yeast lifespan and the cell will undergo premature death. In the case of the *sir2*Δ strain, the sir2 protein (sir2p) is believed to stabilise ribosomal DNA, leading to a decrease in the formation of ERCs. The removal of this protein would therefore result in an increase in damage formation [23]. Note that this is apart from sir2p being involved in the retention mechanism.

That the deletion of genes involved in ribosome biogenesis extend the RLS of yeast cells has been found previously [15]. It is proposed that the cells have evolved to have a high fitness rather than a long lifespan [15]. This is connected to the evolutionary theory of ageing where a decreasing selection pressure with age lead to the occurrence of deleterious genes that have an effect only at later life stages.

The conclusion that can be drawn from the RLS data analysis is that there exist cellular processes that are fundamental in order for a cell to have a normal lifespan and that partly seem to be connected to the damage control of the cell. At the same time, there are gene products that, on the contrary, seem to reduce the replicative lifespan when present in the cell, indicating that the cells have not evolved to live as long as possible. Instead, it is argued that, during the course of evolution, the fitness of the population has been of greater importance than the lifespan of the individual.

In order to connect the replicative lifespan data analysis to the model of replicative ageing, it was investigated how the model parameters affect the RLS of the model. Individual Sobol' indices were estimated for the model parameters with the replicative lifespan of the model as output, Figure 4.6.
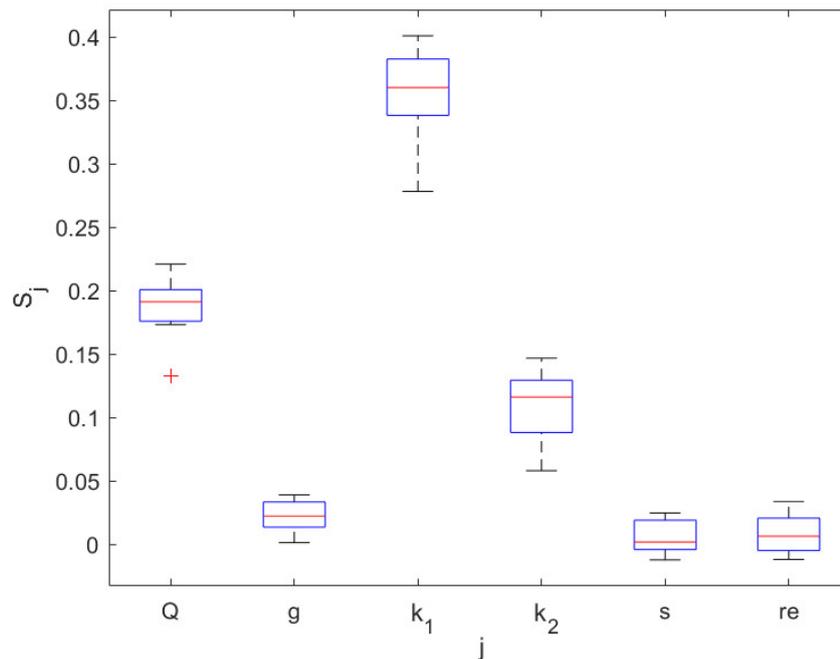
4. Results

Figure 4.6: Estimated Sobol' indices, S_j, for the model parameters, with the RLS of the model as the output. The values were obtained for 10 samples of size 1,000, generated with LHS sampling.

The estimated Sobol' indices, Figure 4.6, indicate that the damage formation rate, k_1, explains most of the variance in the replicative lifespan of the model, followed by the damage resilience parameter, Q, and the damage repair rate, k_2. What these three parameters have in common is that they are all connected to how the individual cell handles damage. This is in accordance with the three strategies for an increased individual RLS established previously; to decrease the damage formation rate (k_2) and to increase the repair rate (k_1) or the damage resilience (Q) [5]. In comparison to this, parameters s and re, determine how much damage the daughter cell inherits from the mother cell and thus might explain the variance in the RLS on a population basis rather than on an individual basis.

Chapter 5

Discussion and conclusion

In the field of systems biology, the level on which an analysis can be performed is in part determined by the experimental data. On the one hand, there is the large-scale population data; like the replicative lifespan data analysed in this project. This kind of experimental data allows for the observation of general trends and patterns, and conclusions can be drawn regarding, for example, a whole population or genome. On the other hand, it is possible to zoom in and record data on a single-cell level, in order to get a more detailed understanding. This was exploited in the parameter estimation, where microfluidics data was used to fit the damage accumulation model of replicative ageing to the behaviour of individual cells. To take the research further, the next step would be to merge the inferences made from investigations on several levels to create a bigger picture; the details of the small-scale analysis can be used to better understand the general trends observed in the large-scale analysis and vice versa. The following discussion and conclusion will mainly concern the parameter estimation.

The analysis performed on the damage accumulation model of replicative ageing indicate that the continuous model parameters Q, g, k_1 and k_2 are structurally identifiable given that the output is $y = P$. This means that it is theoretically possible to determine the values of these parameters with optimisation, assuming zero or negligible damage inside of the cells.

The parameter estimates obtained by optimisation without the RLS constraint are judged as unreasonable given the structure of the model of replicative ageing. Specifically, the estimates for the damage resilience parameter Q are small, generating an environment where the simulated cells die before they have time to divide. In other words, the clonal senescence constraint is violated. Moreover, the numerical identifiability analysis indicate that the validity of some of the parameter estimates is uncertain. The global sensitivity analysis showed that most of the variance in the objective function used is explained by the parameter g and that only a small fraction of the observed variance is explained by the three other parameters. This implies that all of the investigated parameters but g are hard to identify on a global scale with the implemented objective function. That the optimisation was performed with daughter cell curves, for which growth is dominant and damage handling secondary, is likely a reason.

In the second optimisation approach, a replicative lifespan constraint was inferred on the optimisation and the parameter Q was set to a fixed value. The minima found with this approach do not violate the clonal senescence constraint. However, the coefficients of variance obtained for the estimates of k_1 and k_2 are large

and the correlations between the estimates for all of the investigated parameters are moderate to high. This means that, even though the minima found generate simulations with reasonable replicative lifespans, the parameters are numerically non-identifiable. The global sensitivity analysis that was performed indicate that the addition of an RLS constraint induces a shift in the explained variance of the objective function, where k_1 explains more of the variance observed. However, a majority of the variance is still explained by the parameter g. In addition, only a small fraction of the variance is explained by k_2, indicating that this parameter is globally hard to identify with the given experimental data and objective function.

In the optimisation without the RLS constraint added to the objective function, not much difference could be observed between the minima obtained for the three investigated strains. The results from the second optimisation approach demonstrate that it is possible to find parameter sets for which the model simulations generate the desired replicative lifespans, if a replicative lifespan constraint is added to the objective function. However, optimally, the experimental data should have been sufficient enough for the model to be adapted to the differences between the investigated strains and, as a result, also reproduce their correct replicative lifespans. With the aim to achieve this, the next step should be to fit the model of replicative ageing to data recorded for both mother and daughter cells and to remove the assumption of zero damage.

The recommendation is to record single-cell mass or area data following one or several cells from the time when a daughter cell is newly formed and at least for one or two divisions. If cell divisions can be observed in the data, the constant s can be estimated based on the differences in cell mass, or area, prior to and after division. Assuming that a newly formed cell has zero damage, the threshold P_{div} can be estimated from the first data point y_0 of the curve as $P_{div} = y_0/(1-s)$ and used for scaling of the experimental data in place of the generation time. The assumption that daughter cells are formed with zero damage is less valid for the *sir2*Δ strain, due to the loss of the retention mechanism. However, the assumption could still be judged valid in this context if the cells are of an early generation.

If the constant P_{div} can be estimated from the experimental data, there is no need to assume that the cells stay damage free. In this way, the cell mass, or area, can be approximated by $y = P + QD$, the total amount of intact protein and damage in the cell. The drawback of this approach is that the parameter Q is not structurally identifiable. On the advantageous side, if the assumption of zero damage is removed, the optimisation can be performed including mother cell data, making the resulting parameter estimates more reliable. This will be valuable, since the differences between the investigated strains are likely to be more pronounced when comparing cells of an older age, that have had time to accumulate damage. As a result, the model could be better adapted to strain characteristics, rendering it even more useful in the study of the mechanisms behind ageing.

Bibliography

[1]   European Commisson. *Population ageing in Europe*. Brussel, 2014.

[2]   Chang Mell J. and Burgess S.M. "Yeast as a Model Genetic Organism". In: *Encyclopedia of Life Sciences* (2003), pp. 1–8. DOI: `https://doi.org/10.1038/npg.els.0000821`.

[3]   Borgqvist J., Dainese R., and Cvijovic M. "Systems Biology of Ageing". In: *Systems biology*. Ed. by Hohmann S. Nielsen J. New Jersey: John Wiley & Sons, Incorporated, 2017.

[4]   Österlund T., Cvijovic M., and E. Kristianssion. "Integrative Analysis of Omics Data". In: *Systems biology*. Ed. by Hohmann S. Nielsen J. New Jersey: John Wiley & Sons, Incorporated, 2017.

[5]   Borgqvist J., Welkenhuysen N., and Cvijovic M. "The triangle of ageing: a model of how repair, resilience to damage and asymmetric segregation are responses to replicative ageing". In: *BioRxiv* (2018). DOI: `https://doi.org/10.1101/446302`.

[6]   Liu P. and Acar M. "The generational scalability of single-cell replicative aging". In: *Science Advances* 4 (1 2018), eaao4666. DOI: `https://doi.org/10.1126/sciadv.aao4666`.

[7]   Partridge L. and Barton N. H. "Optimality, mutation and the evolution of ageing". In: *Nature* 362 (1993), pp. 305–311. DOI: `http://dx.doi.org//10.1038/362305a0`.

[8]   Kirkwood T.B.L. "Evolution of ageing". In: *Nature* 270 (1977), pp. 301–304. DOI: `http://dx.doi.org//10.1016/S0047-6374(01)00419-5`.

[9]   Polymenis M. and Kennedy B.K. "Chronological and replicative lifespan in yeast: do they meet in the middle?" In: *Encyclopedia of Life Sciences* 11 (19 2012), p. 3531. DOI: `https://doi.org/10.4161/cc.22041`.

[10]   Steffen K.K., Kennedy B.K., and Kaeberlein M. "Measuring replicative lifespan in budding yeast". In: *Journal of Visualised Experiments* (28 2009), p. 1209. DOI: `https://doi.org/10.3791/1209`.

[11]   Aguilaniu H. et al. "Asymmetric Inheritance of Oxidatively Damaged Proteins During Cytokinesis". In: *Science* 299 (5613 2003), pp. 1751–1753. DOI: `https://10.1126/science.1080418`.

[12] Kaeberlein M., McVey M., and Guarenete L. "The SIR2/3/4 complex and SIR2 alone promote longevity in Saccharomyces cerevisiae by two different mechanisms". In: *Genes and development* 13 (19 1999), pp. 2570–2580. DOI: https://10.1101/gad.13.19.2570.

[13] Defossez P.A. et al. "Elimination of Replication Block Protein Fob1 Extends the Life Span of Yeast Mother Cells". In: *Molecular cell* 3 (4 1999), pp. 447–455. DOI: https://doi.org/10.1016/S1097-2765(00)80472-4.

[14] Snoep J.L. et al. "Control of specific growth rate in saccharomyces cerevisiae". In: *Microbiology* 155 (5 2009), pp. 1699–1707. DOI: https://doi.org/10.1038/nature07212.

[15] McCormick M.A. et al. "A comprehensive analysis of replicative lifespan in 4,698 single-gene deletion strains uncovers conserved mechanisms of aging". In: *Cell metabolism* 22 (5 2015), pp. 895–906. DOI: https://doi.org/10.1038/npg.els.0000821.

[16] Andréasson N., Evgrafov A., and Patriksson M. *An Introduction to Continuous Optimization*. 2nd ed. Lund: Studentlitteratur AB, 2013.

[17] DiStefano III J. *Dynamic systems biology modeling and simulation*. 1st ed. Cambridge: Academic Press, 2014.

[18] Raue A. et al. "Comparison of approaches for parameter identifiability analysis of biological systems". In: *Bioinformatics* 30 (10 2014), pp. 1440–1448. DOI: http://dx.doi.org/10.1093/bioinformatics/btu006.

[19] Karlsson J., Anguelova M., and Jirstrand M. "An Efficient Method for Structural Identifiability Analysis of Large Dynamic Systems". In: *IFACS Proceedings Volumes* 45 (16 2012), pp. 941–946. DOI: http://dx.doi.org/10.3182/20120711-3-BE-2027.00381.

[20] Prieur C. and Tarantola S. "Variance-Based Sensitivity Analysis: Theory and Estimation Algorithms". In: *Handbook of Uncertainty Quantification*. Ed. by Owhadi H. Ghanem R. Higdon D. Cham: Springer, 2015.

[21] Woods D.C. and Lewis S.M. "Design of Experiments for Screening". In: *Handbook of Uncertainty Quantification*. Ed. by Owhadi H. Ghanem R. Higdon D. Cham: Springer, 2015.

[22] Usaj M. et al. "TheCellMap.org: A Web-Accessible Database for Visualizing and Mining the Global Yeast Genetic Interaction Network". In: *G3: Genes, Genomes, Genetics* 7 (5 2017), pp. 1539–1549. DOI: https://doi.org/10.1534/g3.117.040220.

[23] Sinclair D.A. and Guarente L. "Extrachromosomal rDNA Circles - A Cause of Ageing in Yeast". In: *Cell Press* 91 (7 1997), pp. 1033–1042. DOI: https://doi.org/10.1016/S0092-8674(00)80493-6.

Chapter A

Appendix 1

A.1 Sensitivities

In the general case, the ODE:s of the sensitivities for P and D are defined as follows.

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{P}}{\partial p_i}\right) = \frac{\partial}{\partial p_i}\left(\frac{\partial \tilde{P}}{\partial t}\right) = \frac{\partial h_1}{\partial \tilde{P}}\frac{\partial \tilde{P}}{\partial p_i} + \frac{\partial h_1}{\partial \tilde{D}}\frac{\partial \tilde{P}}{\partial p_i} + \frac{\partial h_1}{\partial p_i} \tag{A.1}$$

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{D}}{\partial p_i}\right) = \frac{\partial}{\partial p_i}\left(\frac{\partial \tilde{D}}{\partial t}\right) = \frac{\partial h_2}{\partial \tilde{P}}\frac{\partial \tilde{P}}{\partial p_i} + \frac{\partial h_2}{\partial \tilde{D}}\frac{\partial \tilde{P}}{\partial p_i} + \frac{\partial h_2}{\partial p_i} \tag{A.2}$$

Where p_i is a parameter, $h_1 = \dfrac{dP}{dt}$ and $h_2 = \dfrac{dD}{dt}$.
The rate of changes of the sensitivities for Q are shown below.

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{P}}{\partial Q}\right) = (\tilde{\mu}(\tilde{S})(g - \tilde{D}) - \tilde{k}_1)\frac{\partial \tilde{P}}{\partial Q} + (\tilde{k}_2 Q - \tilde{\mu}(\tilde{S})P)\frac{\partial \tilde{D}}{\partial g} + \tilde{k}_2 D \tag{A.3}$$

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{D}}{\partial Q}\right) = \frac{\tilde{k}_1}{Q}\frac{\partial \tilde{P}}{\partial Q} - \tilde{k}_2\frac{\partial \tilde{D}}{\partial Q} - \frac{\tilde{k}_1}{Q^2}\tilde{P} \tag{A.4}$$

The rate of changes for the sensitivities for g were derived as follows.

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{P}}{\partial g}\right) = (\tilde{\mu}(\tilde{S})(g - \tilde{D}) - \tilde{k}_1)\frac{\partial \tilde{P}}{\partial g} + (\tilde{k}_2 Q - \tilde{\mu}(\tilde{S})P\frac{\partial \tilde{D}}{\partial g} + \tilde{\mu}(\tilde{S})\tilde{P} \tag{A.5}$$

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{D}}{\partial g}\right) = \frac{\tilde{k}_1}{Q}\frac{\partial \tilde{P}}{\partial g} - \tilde{k}_2\frac{\partial \tilde{D}}{\partial g} \tag{A.6}$$

The rate of changes of the sensitivities for k_1 follows below.

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{P}}{\partial \tilde{k}_1}\right) = (\tilde{\mu}(\tilde{S})(g - \tilde{D}) - \tilde{k}_1)\frac{\partial \tilde{P}}{\partial \tilde{k}_1} + (\tilde{k}_2 Q - \tilde{\mu}(\tilde{S})P)\frac{\partial \tilde{D}}{\partial \tilde{k}_1} - P \tag{A.7}$$

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{D}}{\partial \tilde{k}_1}\right) = \frac{\tilde{k}_1}{Q}\frac{\partial \tilde{P}}{\partial \tilde{k}_1} - \tilde{k}_2\frac{\partial \tilde{D}}{\partial \tilde{k}_1} + \frac{P}{Q} \tag{A.8}$$

For k_2, the ODE:s for the sensitivities were derived as shown below.

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{P}}{\partial \tilde{k}_2}\right) = (\tilde{\mu}(\tilde{S})(g - \tilde{D}) - \tilde{k}_1)\frac{\partial \tilde{P}}{\partial \tilde{k}_2} + (\tilde{k}_2 Q - \tilde{\mu}(\tilde{S})P)\frac{\partial \tilde{D}}{\partial \tilde{k}_2} + Q\tilde{D} \tag{A.9}$$

$$\frac{\partial}{\partial t}\left(\frac{\partial \tilde{D}}{\partial \tilde{k}_2}\right) = \frac{\tilde{k}_1}{Q}\frac{\partial \tilde{P}}{\partial \tilde{k}_2} - \tilde{k}_2\frac{\partial \tilde{D}}{\partial \tilde{k}_2} - \tilde{D} \tag{A.10}$$