Parameter Identification for a DNA replication model

K. Koutroumpas[1], Z. Lygerou[2] and J. Lygeros[1]

1. Automatic Control Laboratory, ETH Zurich, Switzerland
2. Laboratory of General Biology, School of Medicine, University of Patras, Rio, Patras, Greece

Abstract— **DNA replication is one of the most fundamental processes in the life of every cell. In earlier work a model to capture the mechanics of the DNA replication process was developed. The model allowed us to make novel predictions regarding the mechanisms behind DNA replication based on experimental data for the fission yeast. One of the difficulties we had to overcome in the process was tuning of the model parameters based on experimental data, which, for lack of better methods had to be done manually. Here we propose a methodology for systematizing this process, inspired by techniques for multi-objective optimization.**

I. Introduction

The life of any eukaryotic cell evolves through an ordered sequence of events, known as the cell cycle. The outcome of a successful cycle is the division of a single cell into two distinct cells, the "daughter cells" (mitosis). The cell cycle comprises four phases: G_1, a cell growth (gap) phase; *S*, the DNA synthesis phase; G_2, a second period of growth; and *M* (mitosis) phase, in which the cell is divided into two genetically identical cells [1] (Figure 1). It is important for the well-being of the cell that the four phases are precisely coordinated and follow one another in the correct order. Improper execution of the cell cycle (for example, uncontrolled division, or DNA re-replication) may lead to genomic instability, a characteristic of cancer cells [1]. Protein complexes called Cyclin Dependent Kinases (CDK) are supervising the cell cycle. Cell cycle events, such as entry into S phase and into M phase, are regulated by the periodic fluctuations in the activity of CDKs [2] (Figure 2).

DNA replication, the process of duplication of the cell's genetic material during the *S* phase, needs to be executed in a way that ensures that both daughter cells will have the same genetic information. DNA synthesis must always be carried out prior to cell division and within a specified amount of time. Failure to replicate even a small part of the genome would disrupt proper segregation of the genetic material to the two daughter cells during mitosis, leading to genomic instability.

In earlier work the authors and co-workers proposed a model to capture the "mechanics" of the DNA replication process [3]. The model was instantiated based on experimental data for the fission yeast (*Schizosaccharomyces pombe*). Fission yeast is an attractive model organism for studying DNA replication. As in all eukaryotes, DNA replication in the fission yeast initiates from multiple points along the genome, the so called origins of replication. Moreover, experimental evidence suggests that these origins are randomly selected from a pool of potential origins, and the time at which they initiate replication is randomly selected during the S phase [4], [5], [6], [7], [8]. This randomness appears to be a common theme for the DNA replication of higher eukaryotes, including humans. In our work [3] *in silico* analysis of a fission yeast DNA replication model instantiated based on the experimental data of [7] led to interesting conclusions and conjectures about the mechanisms that govern the DNA replication process.

One difficulty that had to be overcome in the process was tuning of the model parameters based on experimental data. The model we proposed contains several parameters, related to the efficiency with which origins fire during the S phase, the speed with which the replication forks move along the genome, the potential presence of low efficiency origins etc. To ensure that the model is realistic, values for all these parameters have to be selected to reflect experimental evidence. This is a rather painstaking process involving manual tuning and several trial-and-error *in silico* experiments. In this paper we propose a methodology for simplifying this process. The idea is to use tools from multi-objective optimization to systematically search for parameter values that "optimally" match experimental data. The key challenge for doing this is that the model is inherently stochastic, since it reflects the uncertainty in origin selection and initiation timing characteristic of eukaryotic cells. Therefore defining an appropriate notion of optimality is far from straight forward. In this paper we use empirical averages from multiple *in silico* experiments and attempt to match these to statistical data collected experimentally. Since the number of unknown parameters is relatively small, the search for optimal parameter values can then be conducted by "brute-force" gridding of the parameter space, followed by exhaustive search. We believe, however, that this approach can be extended to high dimensional parameter spaces, since it is straight forward to couple to randomized optimization methods such as Markov Chain Monte Carlo; for an example of the use of such methods for parameter identification in an unrelated biological model see [9].

The work is organized in 4 sections. Section II outlines the stochastic hybrid model for the DNA replication process. The parameter identification problem is formulated in Section III followed by the identification results. Finally, conclusions

This work was supported by European Commission under the project HYGEIA (FP6-NEST-004995).

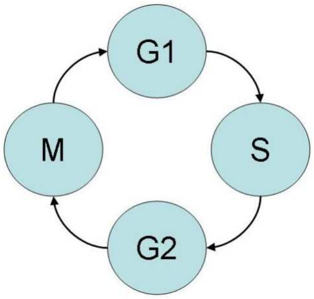

Fig. 1. The phases of Cell Cycle

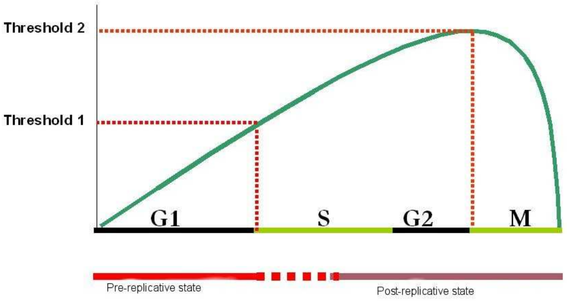

Fig. 2. Quantitative model of cell cycle regulation

based on the results and future objectives are reported in Section IV.

II. DNA REPLICATION MODEL

In this section a short description of the model for DNA replication is provided. The reader is referred to [3], [10] for a detailed description of the model.

A. Biological Background

DNA replication in eukaryotes initiates from multiple points called origins of replication. In a genome there is a number of regions that can serve as origins. However, only some of these potential origins are activated in each cell cycle. According to recent work [8], active origin selection is not deterministic. A specific origin will fire in some but not all cell cycles. Moreover, firing time of an active origin differs from cell to cell.

Initially, all origins are in the pre-replicative state. When CDK activity increases over a threshold (Threshold 1 in Figure 2), origins can initiate DNA replication. When an origin fires, two replication forks are created and start moving in opposite directions along the genome. To ensure that all bases are replicated once and only once a cell should be able to distinguish replicated from un-replicated regions. Origins locations that have been replicated (either because the origin fired or because they were passively replicated by a replication fork from a neighboring origin) automatically switch to the post-replicating state and can no longer support initiation of replication. Only after the end of M phase, when CDK activity resets to zero, origins can re-acquire the pre-replicative state. This supervising mechanism inhibits DNA re-replication.

B. Stochastic Hybrid Model

The DNA replication procedure, as described in the previous section, is a stochastic hybrid process comprising discrete transitions between origin states (pre-replicative and post-replicating), continuous evolution of replication forks and stochastic origin selection and firing. In earlier work by the authors and co-workers, a stochastic hybrid model of DNA replication has been developed to capture all these diverse elements [3] and has been instantiated using experimental data for the fission yeast [7]; see also [10] for the mathematical foundations on which the model is based.

The model of [3] contains several parameters, whose values play an important role in the predictive power of the model. Here we concentrate on three of these parameters.

- The firing propensity of each origin.
- The speed with which the replication forks move over different parts of the genome.
- The number and location of low efficiency origins, that are potentially present but undetectable due to the limitations of experimental methods.

For the first parameter, experimental evidence suggests that active origin selection and origin firing time are both random. To capture this, we assume that the firing time, T_i, of origin i is randomly extracted according to an exponential distribution

$$P[T_i \geq t] = e^{-\lambda_i t}. \tag{1}$$

The parameter λ_i reflects the intrinsic propensity of origin i to fire per unit of time. This parameter can be estimated using the experimentally observed firing probability of the origin (denoted by FP_i), i.e. the fraction of cell cycles in which the origin is observed to fire [7]. Assuming the exponential distribution (1) and setting the probability that origin i fires by a given time T_f equal to FP_i we obtain

$$\lambda_i = -\frac{1}{T_f}\ln(1-FP_i). \tag{2}$$

It is clear from (2) that origins with high experimental firing probability will have higher values of λ and be more likely to fire early in the S phase (Figure 3). Firing time is also related to whether an origin will be activated in a cell cycle or not. In a given cell cycle, origins due to fire later than the time when they are passively replicated by a fork emanating from a neighboring origin will not be active.

Note that the value of T_f is arbitrary in this process. Roughly speaking, the experimental data allows us to determine the relative propensity of origins to fire, but not their absolute propensity. A reasonable starting point is to set $T_f = 20$ minutes, the expected duration of S phase in *S.*

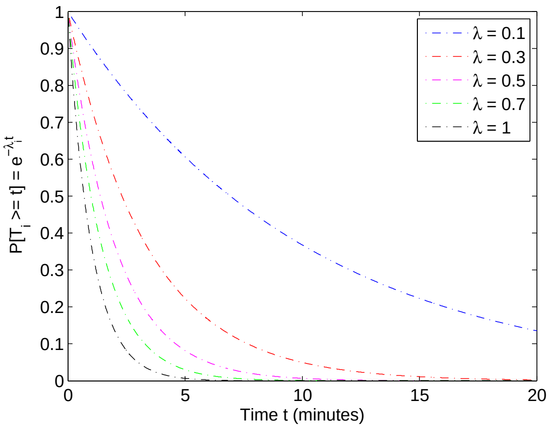

Fig. 3. Exponential distribution for different values of λ.

pombe. The first objective of our parameter tuning method will be to systematically determine a value for T_f.

The second parameter is the speed with which replication forks move. This speed will generally depend on the location of the genome that is currently being replicated. For simplicity, however, we will assume that this speed (denoted by v) is constant for the entire genome. Experimental evidence allows us to determine an approximate value for v at different places along the genome; the work of [7] suggests $v \approx 3000$ bases per minute. The second objective of the identification methodology will be to determine a more precise value of v. Even though this value will still be constant, it will hopefully help us account better for fluctuations along the genome.

Finally, it has been observed that in addition to "strong" origins that fire in many cell cycles and are easy to observe experimentally, there can be many more "weak" origins that fire in only a small fraction of the cell cycles (less than 10%). The location and firing propensity of each of these origins are difficult to determine experimentally, since they are observed very rarely. If there are many of them however, their cumulative effect may be significant. The last objective of the methodology outlined in the next section will be to estimate the number of such low-efficiency origins, denoted by N below.

III. Parameter Identification

A. Problem Statement

Let $\mathbf{M}(\theta)$ denote our stochastic model of DNA replication, where θ is a set of input parameters to be determined by the identification experiments; based on the discussion above we will consider a 3 dimensional parameter vector $\theta = (T_f, v, N)$. The objective is to determine the value of θ that best explains the experimental observations, denoted by $\mathbf{D}$. In subsequent discussion a 2 dimensional observation vector $\mathbf{D} = (Y_1^{obs}, Y_2^{obs})$ will be considered; Y_1^{obs} denotes the experimentally observed average S phase duration (generally accepted to be around 20 minutes) and Y_2^{obs} denotes the experimentally observed average number of active origins per cell cycle (160 according to the data of [7]).

In order to quantify the meaning of "best" in the above statement we have to determine an evaluation criterion to rank parameter values. Let $\mathbf{J}(\theta|\mathbf{D})$ be some cost function that quantifies how well the prediction of model $\mathbf{M}(\theta)$ fit experimental data $\mathbf{D}$; we assume that the lower the value of $\mathbf{J}(\theta|\mathbf{D})$ the better the fit. In this context, we want to find the parameter values $\hat{\theta}$ that minimize the cost function $\mathbf{J}(\theta|\mathbf{D})$:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \mathbf{J}(\theta|\mathbf{D}) \tag{3}$$

where Θ is our search space of possible parameter values.

The model for DNA replication comprises randomness that leads each execution to differ from every other execution, even for the same parameter values. This reflects the randomness that is also observed in experimental data. As a consequence a sole execution is not adequate to fully characterize a given set of parameter values; instead the aim of the optimization should be to somehow match the statistics observed in experiments to those predicted by the model. A simple way to do this is to compute empirical averages of the experimentally observed quantities by running multiple simulations of the model for fixed parameter values. We can then approximate the optimal parameter values θ^* as the ones whose empirical statistics best match the experimentally observed average values.

For a fixed value of θ let $Y_1^i(\theta)$ and $Y_2^i(\theta)$ for $i = 1, \dots M$ denote the DNA replication completion time and the number of firing origins predicted by each of M independent simulations of the model $\mathbf{M}(\theta)$. The empirical averages of these M experiments can then be computed as

$$Y_1(\theta) = \frac{1}{M}\sum_{i=1}^{M} Y_1^i(\theta) \text{ and } Y_2(\theta) = \frac{1}{M}\sum_{i=1}^{M} Y_2^i(\theta)$$

We can define now the score function $\mathbf{J}(\theta|\mathbf{D})$ as:

$$\mathbf{J}(\theta|\mathbf{D}) = \ln\left(\frac{1}{2}\left(\frac{|Y_1(\theta) - Y_1^{obs}|}{Y_1^{obs}} + \frac{|Y_2(\theta) - Y_2^{obs}|}{Y_2^{obs}}\right)\right).$$

Notice that the score function penalizes normalized deviations from the two experimental observations; the fractions in the expression can be thought of % error in the model predictions for each of the two experimentally observed quantities. The average of the two errors is then taken, to balance matching one observation against matching the other. Finally, the logarithm ensures that the score function is "sharp" around the optimal values, since a good and a bad match may differ by several orders of magnitude before the logarithm is taken.

B. Identification Results

Software was developed to solve the optimization problem outlined in the previous section. The software first creates a finite optimization problem by imposing a grid on the parameter space. It then explores the resulting finite space exhaustively. For each value of the parameters, θ, in this finite space several simulations (M) of the model $\mathbf{M}(\theta)$

TABLE I

OUTPUT RESULTS FOR PARAMETER SETS THAT MINIMIZE THE COST FUNCTION.

Parameters			Model Output		Cost Function	Area of -2 level set (Kb)
Potential Origins	Repl. Speed (Kb/min)	T_f (min)	Active Origins	Completion Time (min)		
863	9	6	166	21.2	-3.02	159.6
1000	9	6	170	20.5	-3.10	141.5
1250	9	8	157	21.4	-3.15	120.6
1500	9	8	167	19.4	-3.30	94.7
2000	9	10	166	19.5	-3.55	68.9
2500	9	12	164	19.9	-4.23	36.6

are executed and the empirical averages $Y_1(\theta)$ and $Y_2(\theta)$ are computed. This gives rise to a score $\mathbf{J}(\theta|\mathbf{D})$; parameter values are then ranked according to their score.

The results of the parameter identification are summarized in Table I. Replication speed v was selected from the set $\{1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 12, 15, 18\}$ (values Kb/min, where Kb stands for "thousands of base pairs"), time T_f from the set $\{1, 2, 3, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$ (values in minutes) and the number of potential origins from the set $\{863, 1000, 1250, 1500, 2000, 2500\}$. For the last parameter, the origin set for each number includes all the origins observed in [7] (863 in number), plus additional origins introduced at genomic regions according to a bioinformatic analysis of the properties of known fission yeast origins. Based on this study (and consistent with earlier studies in the literature) origins were assumed to be concentrated in inter-genic regions and the maximum AT content of 500 base windows was assumed to be a good predictor of origin activity. A moving AT content threshold was therefore used to determine inter-genic regions that were included in the simulations as weak origins. The firing probabilities of these additional origins were set to $FP_i = 8\%$, below the experimental threshold of approximately 10%.

The scores in the optimization procedure were computed based on empirical averages from $M = 1000$ simulations of the model $\mathbf{M}(\theta)$ for each possible triplet of the parameters θ. Table I summarizes the best parameter sets for each case of additional origins and the corresponding outputs of the model. Figure 5 demonstrates the score of the different parameter values. Recall that the score function is logarithmic, so a value of 3 indicates a match of the order of 10^{-3}.

Table I suggests that a better match of the experimentally expected values for the average S phase duration and the average number of active origins is obtained for larger numbers of additional, low efficiency origins. The original origin set of [7] can match the proposed cost function to within $1/1000$ for appropriate choices of replication speed and the T_f parameter. With a total number of 2500 origins, on the other hand, the match can be made better than $1/10000$. Superficially this would seem to suggest the presence of many low efficiency origins, too weak to be detected by the genome wide methods of [7]. Matching the proposed cost function is not, however, the end of the story. An additional consideration one has to take into account is the robustness of the process (see, for example [11] and the references therein). Robustness is difficult to quantify in this setting. Loosely speaking, one would expect small variations in the parameter values to lead to small variations in the cost and hence qualitatively similar behavior. In an attempt to quantify this notion we have computed the area of the level set of value -2 in the cost function for each of the cases A-F in Figure 5. That is the area[1] of the set

$$\Theta_N = \{\theta \mid \theta_3 = N \text{ and } \mathbf{J}(\theta|\mathbf{D}) \leq -2\}$$

for $N \in \{863, 1000, 1250, 1500, 2000, 2500\}$ (see Figure 4 for the case $N = 2000$). This area reflects the range of parameter values that match the cost function to within 1%, hence a larger area would suggest a more robust solution. The results are reported in the last column of Table I. Note that the trend is opposite from that of the cost: Fewer additional origins lead to more robust solutions, since larger variations in the replication speed and T_f can be tolerated with only a small impact on the average S phase duration and the average number of active origins.

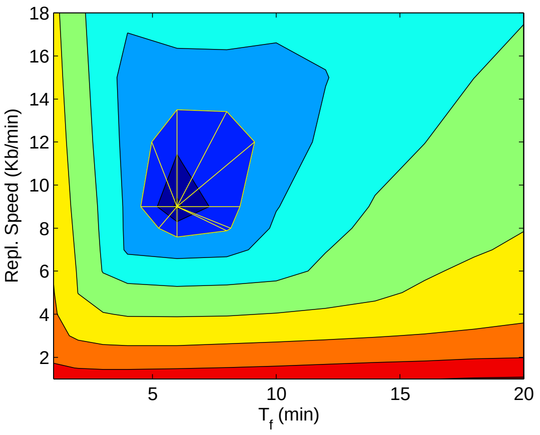

Fig. 4. Level set of value -2 for the case $N = 1000$.

A curious observation is that in all identified optimal parameter sets replication speed has the same value (9000 bases/minute). Although this value is far from experimental estimates of 3000 bases/minute [7], this intriguing invariance should be investigated further experimentally. It may, for example, reflect the maximum replication speed observed in

[1] "Area" in this case is measured in units of thousands of base-pairs.

selected parts of the genome characterized by large distances between potential origins. If this is the case, this could offer an alternative hypothesis for the correlation between inter-origin distance and replication speed reported in [12]: Rather than the replication speed being adjusted to match inter-origin distance, fewer (or weaker) origins may be placed in genomic regions where the replication speed tends to be high, due for example to a simpler chromatin structure.

IV. Discussion

In earlier work, a stochastic hybrid model for DNA replication has been developed [3]. Our aim here was to develop a mechanism for tuning the model parameters to "optimally" match experimental observations. The task is complicated by the fact that the model is stochastic and therefore matching experimental data needs to be interpreted in a statistical sense. We formulated an optimization problem based on empirical averages collected by simulating the model. The problem was solved by discretizing all parameter values on a finite grid. The results show that model estimates for the resulting optimal parameters are fairly close to the observed data. It still remains to be seen, however, if these optimal parameter values are biologically meaningful.

The optimization method proposed here was rather rudimentary, based on gridding the parameter space. This was adequate for our purposes, since the number of parameters was rather small. For higher dimensional spaces randomized optimization methods can be considered for this task. In this light, the use of empirical averages to reflect statistical match has an extra advantage, since empirical averages are readily amenable to estimation by Monte-Carlo simulation methods; different alternatives along these lines are discussed in [9].

In terms of the DNA replication problem itself, current work concentrates on investigating how more complex patterns of origin firing, such as coordination of replication speed with inter-origin distance [12] can be incorporated in the parameter identification framework presented here.

References

[1] D. O. Morgan, The Cell Cycle: Principles of Control. *New Science Press: London*, 2007.

[2] B. Stern and P. Nurse, A quantitative model for the cdc2 control of S phase and mitosis in fission yeast, *Trends in Genetics*, 12(9), 1996.

[3] J. Lygeros , K. Koutroumpas , S. Dimopoulos , I. Legouras , P. Kouretas , C. Heichinger , P. Nurse and Z. Lygerou, Stochastic hybrid modeling of DNA replication across a complete genome, *Proceedings of the National Academy of Sciences*, 105(34), 2008 , pp. 12295 - 12300.

[4] M. Segurado, A. de Luis and F. Antequera, Genome-wide distribution of DNA replication origins at A+T-rich islands in *Schizosaccharomyces pombe*, *EMBO Rep*, 4(11), pp. 1048 - 1053.

[5] W. Feng, D. Collingwood, M.E. Boeck , L.A. Fox, G.M. Alvino, W.L. Fangman, M.K. Raghuraman and B.J. Brewer, Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication, *Nat Cell Biol*, 2006;8:1488.

[6] M. Hayashi, Y. Katou, T. Itoh, M. Tazumi, Y. Yamada, T. Takahashi, T. Nakagawa, K. Shirahige and H. Masukata, Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast, *The EMBO Journal*, (2007) 26, 13271339.

[7] C. Heichinger, C.J. Penkett, J. Bahler and P. Nurse, Genome-wide characterization of fission yeast DNA replication origins, *The EMBO Journal*, 25(21), 2006, pp 5171-5179.

[8] P.K. Patel, B. Arcangioli, S.P. Baker, A. Bensimon and N. Rhind, DNA replication origins fire stochastically in fission yeast, *Mol Biol Cell*, 2006, 17(1), pp 308-316.

[9] K. Koutroumpas, E. Cinquemani, P. Kouretas, and J. Lygeros, "Parameter identification for stochastic hybrid systems using randomized optimization: A case study on subtilin production by *Bacillus Subtilis*," *Nonlinear Analysis: Hybrid Systems*, 2008, 2(3), pp 786-802.

[10] P. Kouretas, K. Koutroumpas, J. Lygeros and Z. Lygerou, Stochastic hybrid modelling of biochemical processes,In C.G. Cassandras and J.Lygeros, editors, *Stochastic Hybrid Systems*, number 9083, in CRC press, 2006.

[11] H. Kitano, Towards a theory of biological robustness, *Molecular Systems Biology*, 2007, 3:137.

[12] C. Conti, B. Sacca, J. Herrick, C. Lalou, Y. Pommier and A. Bensimon, Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells, *Mol Biol Cell*, 2007, 18(8), pp 3059-3067.

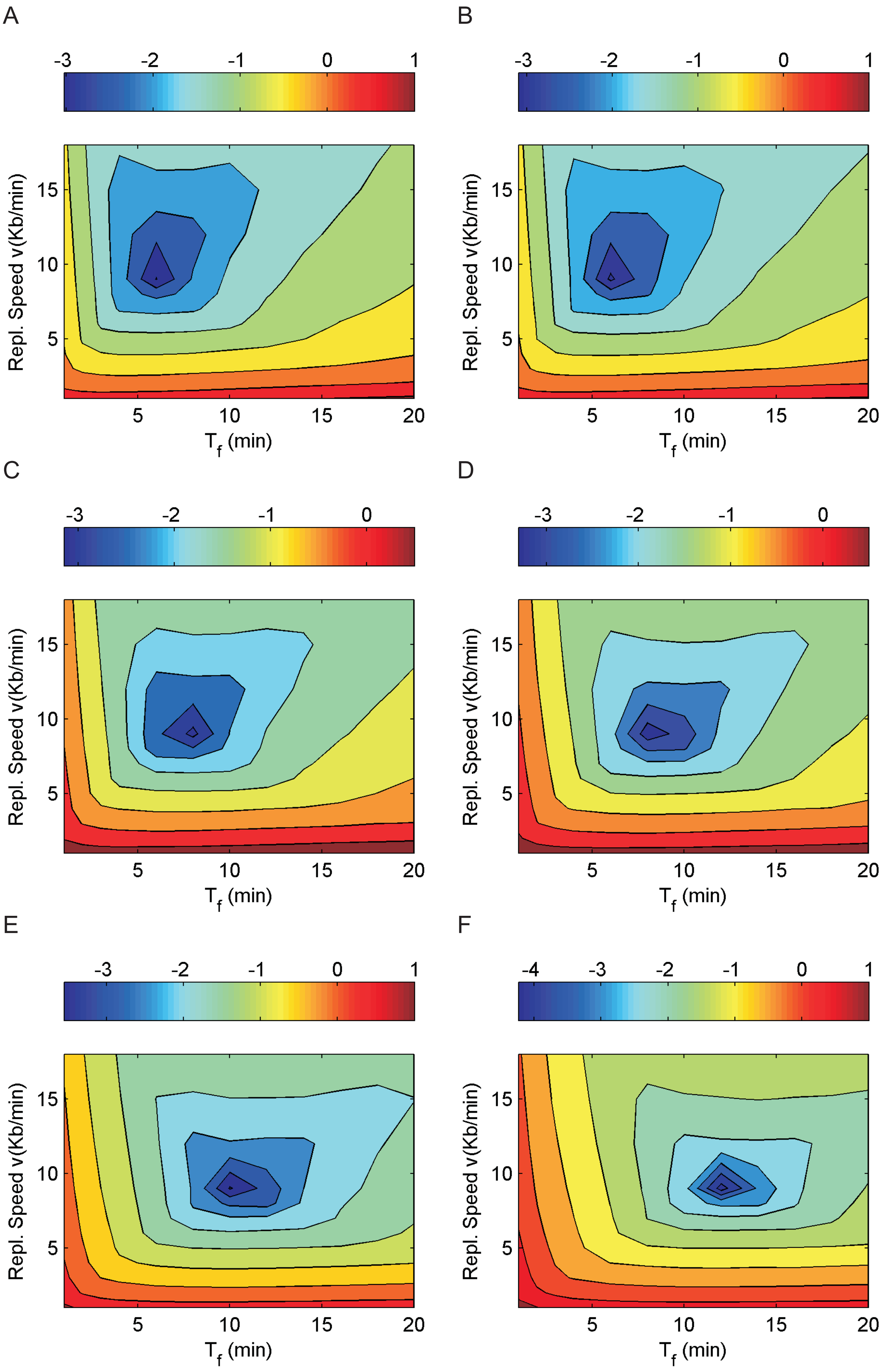

Fig. 5. Parameter identification results for 863 (A), 1000 (B), 1250 (C), 1500 (D), 2000 (E) and 2500 (F) potential origins.