Optimal Placement of Origins for DNA Replication

Jens Karschau,[1,2] J. Julian Blow,[2] and Alessandro P. S. de Moura[1]

[1]*Institute for Complex Systems and Mathematical Biology, SUPA,*
King's College, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom
[2]*Wellcome Trust Centre for Gene Regulation and Expression,*
College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom[*]
(Dated: February 3, 2012)

DNA replication is an essential process in biology and its timing must be robust so that cells can divide properly. Random fluctuations in the formation of replication starting points, called *origins*, and the subsequent activation of proteins lead to variations in the replication time. We analyse these stochastic properties of DNA and derive the positions of origins corresponding to the minimum replication time. We show that under some conditions the minimization of replication time leads to the grouping of origins, and relate this to experimental data in a number of species showing origin grouping.

The replication of the DNA content of a cell is one of the most important processes in living organisms. It ensures that the information needed to synthesize proteins and cellular components is passed on to daughter cells in a robust and timely fashion. Replication takes place during the S-phase of the cell cycle, and it starts from specific locations in the chromosome called *origins*. In order to function in a particular round of the cell cycle, possible origin locations (loci) must undergo a sequence of binding events before S-phase starts. This culminates in the clamping of one or more pairs of ring-shaped Mcm2-7 molecules around the DNA; this is known as *licensing*. Below we denote a pair of Mcm molecules as pMcm. Features of human replication have been studied with the help of the yeast *S. cerevisiae* and *X. laevis* frog embryos. In *S. cerevisiae* licensing is only possible at a set of specific points in each chromosome, characterized by the presence of specific DNA sequences, whereas in *X. laevis* embryos the licensing proteins can bind at virtually any location in the genome [1]. When a licensed locus activates in S-phase, two replication forks are created at the origin, and they move in opposite directions with approximately constant speed, duplicating the DNA as they travel through the chromosome [Fig. 1(a)]. Both origin licensing and origin activation time are stochastic events, since they result from molecular processes involving low-abundance species. In *X. laevis*, both the loci that are licensed and licensed loci selected for activation vary randomly from cell to cell, whereas in *S. cerevisiae* each of the fixed loci has a certain probability of being activated in any given cell — the *competence* — which reflects the fraction of cells in a population in which that locus has had time to be licensed before the S-phase starts [2].

The total time it takes to replicate a cell's DNA — the *replication time* — is a quantity of crucial importance for biology, since it is clearly an evolutionary advantage for replication to be rapid as it affects the minimum time required for cells to duplicate. The location of the origins is one of the crucial factors determining the replication time of cells, and it is reasonable to expect that the loci have been selected by evolution such that the replication time is minimized. There are a number of recent theoretical and modeling works on the dynamics of DNA replication (reviewed in [3]). Previous theoretical works on *S. cerevisiae* have used the experimentally determined loci as given parameters, without attempting to understand why the origins are located where they are [2, 4–6]. Inspection of the loci on a *S. cerevisiae* genome map shows groups of two or three very close origins which are very prominent in most chromosomes [7]. There is also experimental evidence for grouping in *X. laevis*, where origins seem to be distributed with groups of 5 to 10 pMcms [8–10].Most of the existing models of replication in *X. laevis* [9, 11–15] – an exception is [16] – assume the origins to be random and independent of each other, and so they cannot explain pMcm grouping or the observed maximum-spacing of 25 kilobases (kb) between adjacent origins [9].

In this Letter, we use a simplified mathematical model of the DNA replication process to determine the optimal origin location in a chromosome which leads to the shortest average replication time, and how this optimal placement depends on parameters such as the origin competences and the width of the activation time probability distribution. We show that contrary to what one might expect, in many cases the replication time is minimized by placing origins close together in groups like those observed in real chromosomes. This suggests that origin locations have been selected to minimize the replication time. Analysis of our model reveals that grouping is favored for low-competence origins and for origins with large stochastic fluctuations in their activation time. The reason for this is that if origins have an appreciable likelihood of either failing to activate (low competence) or of taking a very long time to activate, grouping origins together helps reduce the risk of large regions of the chro-

mosome not being replicated on time. If the origins are highly competent and have a well-defined activation time, it is optimal to have maximal coverage and distribute the origins evenly on the chromosome. We further show that there is an abrupt transition in the optimal configuration of origins, from isolated to grouped, as the locus' competence decreases (in *S. cerevisiae*), and also as the width of the activation time distribution increases. We give an intuitive explanation of this phenomenon, and argue that it is robust, and independent of any particular details of the model. These results are derived analytically, and tested through numerical simulations. We also compare quantitatively the predictions of our theory with the available experimental data for both *S. cerevisiae* and *X. laevis*, and find that they match well.

We start by analyzing the case of stochasticity in licensing of fixed origin loci, as in *S. cerevisiae*. DNA is modeled as a one-dimensional segment of unit length, and we for simplicity consider only two loci in the chromosome. The two origin loci have competences p_1 and p_2 — these are the probabilities that origins have been licensed and can therefore start replication forks. We initially make the assumption that origins activate at a well-defined time (which we set to $t = 0$). All replication forks travel at the same unit speed across the DNA. We consider the geometry depicted in Fig. 1(a); d_1 (d_2) is the distance from the left (right) end of the chromosome to the left (right) most locus. If both loci fail to be licensed we postulate that replication will eventually take place anyway, with a replication time T_0 — for example, we can imagine that stretch of DNA will be replicated by forks originating from origins outside of the region we are considering. Our results do not depend on T_0, as will be clear shortly; this is just a mathematical device to prevent us dealing with infinite replication times.

If only one of the loci fails to become licensed, the replication time depends on the time it takes for the fork to reach the furthest end of the segment, so $T_{d_1} = 1 - d_1$ for locus 1 and $T_{d_2} = 1 - d_2$ for locus 2. If both loci have been licensed the replication time $T_{d_1,d_2} = \max\{d_1, d_2, (1 - d_1 - d_2)/2\}$ is defined by the longest time for a fork to reach the end of the segment or for two forks to collide. It can be shown that the replication time of an asymmetric placement of loci is never less than a corresponding symmetric configuration (that is, with $d_1 = d_2$). Therefore we consider only symmetrical locus placements, and use $d_1 = d_2 = d$ with $0 \leq d \leq 1/2$. The average replication time is then given by $T_{\mathrm{rep}}(d) = (1 - p_1)(1 - p_2)T_0 + (p_1 + p_2 - 2p_1p_2)(1 - d) + p_1p_2 \max\{d, (1 - 2d)/2\}$.

This is a piecewise-linear function with discontinuities at $d = 1/4$ and $1/2$. Hence, T_{rep} can only have a minimum found at $d = 0$, $d = 1/2$, or at $1/4$. Placing loci at the end of a segment ($d = 0$) is obviously not a minimum of T_{rep}. The replication times for $d = 1/4$ and $1/2$ are $T_{\mathrm{rep}}(d = 1/2) = (1 - p_1)(1 - p_2)T_0 +$ $(p_1 + p_2 - p_1p_2)/2$ and $T_{\mathrm{rep}}(d = 1/4) = (1 - p_1)(1 - p_2)T_0 + (3p_1 + 3p_2 - 5p_1p_2)/4$. We conclude that the two loci group together ($d = 1/2$) to achieve minimum replication time if $T_{\mathrm{rep}}(d = 1/2) < T_{\mathrm{rep}}(d = 1/4)$, which leads to the condition

$$p_2 < \frac{p_1}{3p_1 - 1}. \tag{1}$$

Notice here that T_0 drops out. The inequality Eq. (1) defines two regions on the p_1–p_2 plane, corresponding to grouped or isolated loci being optimum. This is shown in Fig. 1(b), where this analytical result is confirmed by stochastic simulations. The region above the curve corresponds to competences for which T_{rep} is minimized by loci being apart ($d = 1/4$) and below the curve for organising these in a group ($d = 1/2$). In general, if one of the loci has low competence grouping gives the minimum T_{rep}. In fact, it can be shown that if one of the loci has a competence lower than 50%, grouping is the optimal situation regardless of the competence of the other — even if the other is close to 100% competent.

For the case of equal competences, $p_1 = p_2 = p$, the grouped configuration is optimal if $p < 2/3$. We ran a numerical optimization algorithm (using genetic algorithms [17]) to find the loci corresponding to the least T_{rep} for a range of p; these results are shown in Fig. 1(c). The same transition also takes place for non-identical values of p_1 and p_2 — whenever one crosses from the dark to the light regions of Fig. 1(b).

The above results may seem at first quite counter-intuitive; one might expect that the configuration with the least replication time would correspond to isolated loci ($d = 1/4$). However, if the origins have a significant chance of failing to activate, this configuration would mean that often one side of the chromosome would have to wait for a fork which originated at the origin on the other site to replicate it, therefore increasing T_{rep}. So in the case of low competences, it becomes advantageous to have both loci centered, which is near any point in the chromosome. This explains the condition for grouping if $p < 2/3$.

In reality eukaryotic chromosomes have more than two loci [3], so next we investigate the case of a chromosome on which there are many loci and examine the conditions under which it becomes favorable to have isolated origin loci compared to groups. In this analysis we will assume for simplicity that the loci all have identical competence. We consider a group of loci as one single locus with an effective competence p_{eff}. For a group consisting of m loci p_{eff} is the competence that at least one locus will be licensed there, and is given by $p_{\mathrm{eff}} = 1 - (1 - p)^m$. We assume that one large group of n identical loci breaks up into two groups of equal size, each consisting of $n/2$ loci. A locus organized with others in a group of size $m = n/2$ rather than with n loci will give minimum T_{rep}, as long as the locus' competence is less than its
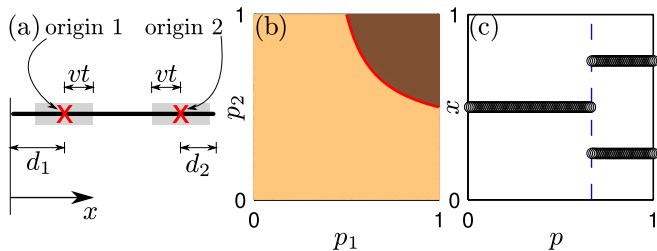
FIG. 1. (color online). (a) Coordinate system for origin loci with d_1, d_2 being the distance from the left- or right-end of the chromosome, respectively. x is the position coordinate along the chromosome. Replication forks travel at a speed v away from the origins. The grey regions show the replicated DNA at time t. (b) Simulation results, showing optimal loci to achieve minimal T_{rep} for 2 loci with different competences, are shown for p_1-p_2 combinations on a lattice grid. Color indicates $d_1, d_2 = 1/2$ (beige) or $d_1, d_2 = 1/4$ (brown). The two regimes are separated by a coexistence line matched by the condition Eq. (1) in red. (c) Optimal position of 2 loci with respect to their competence p to minimize the replication time T_{rep} (circles) and $p = 2/3$ (dashed line).

FIG. 2. (color online). (a) Probability at which groups separate p_c vs. loci/group n. Shown are simulations (circles) and analytical prediction for $p_c = 1 - 1/\sqrt[n]{9}$ (line). (b) Distribution of origin loci on yeast chromosome VI with known (grey) and unknown competences [2, 18]. The distribution results from our simulation in search for minimum T_{rep} (only grey origins considered). The group in the middle of the chromosome with a low and highly competent locus was recovered.

critical probability p_c, given by $p_{\text{eff}} = 2/3$, which yields $p_c = 1 - 1/\sqrt[n]{9}$. Figure 2(a) confirms our analytical result showing the value of p_c for increasing group sizes in our simulations. These results clearly show that large groups of many highly competent loci are unfavorable, but that groups tend to form for low-competence loci.

Our hypothesis is that selective pressure has influenced the position of origin loci through the minimization of the replication time. We show in Fig. 2(b) locus competence and location data for yeast chromosome VI, which has been studied extensively [2, 18]. Competences cannot be measured for all loci (in white), because either they are too close to the end of the chromosome or to an adjacent locus. We performed a search for the optimal position for the loci in the region with known competences using a genetic algorithm [17]. We remark that there is an identifiability problem as all strong loci have $p \sim 90\%$ and we therefore constrained the ordering during the optimization. Although in this result we do not consider inter-origin variations in the origin activation time, the predicted locus distribution from these simulations bears a good resemblance to the actual spacing with a score of $F = 0.11$ [19]; in particular we recover the group in the middle, in which an origin locus with 58% competence is placed next to one with 88% competence. Even multiple repeats of the optimisation algorithm produce minimum T_{rep} solutions which have on average $F = 0.12$.

The above discussion focused on the case of pre-defined loci in yeast, and ignored additional noise such as the variation in origin activation time. We show in Fig. 3(d) that the previous pattern of origin grouping is preserved in the two-origin model with stochastic variation in origin activation time. Grouping is important for swift replica-
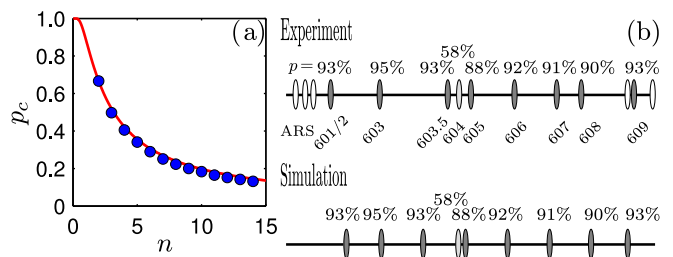
tion under conditions of low competence and large noise which we will explain in the remainder of this letter.

We now examine the case of stochastic activation time for X. laevis embryos as a model organism. Unlike loci in yeast, any DNA locus in an X. laevis embryo is capable of binding with pMcm to become an origin. Surprisingly, biologists find roughly equally-spaced groups of 5–10 pMcms separated by approximately 10 kb [8–10]. We will use the same approach as above, but now with respect to stochasticity in the replication time. In this case, an "origin" is defined as a locus where at least one pMcm has bound to it, and so it corresponds to 100% competent locus in the notation we have used so far. It is well accepted by biologists, however, that origin activation time is stochastic. For simplicity we assume that the pMcms at an origin can activate with uniform probability at any time within a window which has a lower boundary at $t_0 = 0$ min and an upper at t_b, which is at maximum the length of an S-phase (20 min). In addition, pMcms are assumed to be all identical with the same activation probability distribution (standard deviation $\sigma = t_b/\sqrt{12}$).

The expectation is that we will again see a transition of the optimal configuration from isolated pMcms to groups as σ increases; this is akin to varying competence in our previous scenario. We test this prediction using the two-origin model with one pMcm bound to one origin; we find numerically the optimal (minimum T_{rep}) positions for the origins as a function of σ. The results are presented in Fig. 3(a). We again use a segment of unit length and forks progress at unit speed of $v = 1$. We observe a sharp transition at $\sigma \approx 0.25$, above which it is best to place both origins in the middle of the segment, as observed in the case with varying competence. A minor difference between this case and the previous one is that for $\sigma < 0.25$, the optimal location of the origins is not constant.

We now apply this model for more origins and pMcms, using realistic parameters so that we can relate the results to what is experimentally known about X. lae-
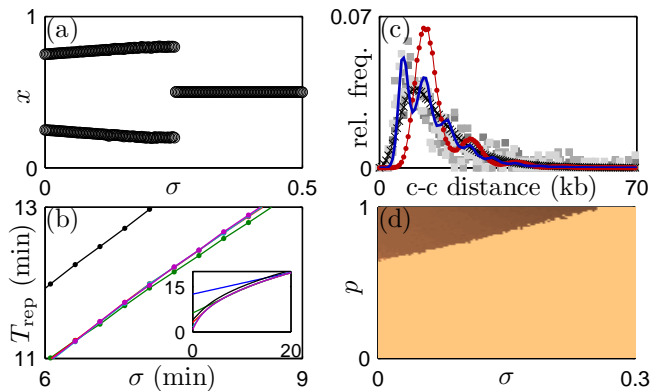
FIG. 3. (color online). (a) Origin position x so that T_{rep} for 2 pMcms is minimal on a segment of unit length, when the standard deviation σ for their activation time increases. (b) Inset: T_{rep} as a function of σ for realistic parameters as given in the text. Origins are distributed in 4 equally-spaced groups of 16 pMcms (blue); 8 groups of 8 pMcms (green); 16 groups of 4 pMcms (red); 32 groups of 2 pMcms (cyan); 64 single pMcms (magenta); 64 pMcms placed randomly (black). Main: zoom around realistic $\sigma \sim 8$ min. For $6 < \sigma < 20$ min minimal T_{rep} is achieved for groups of 8 pMcms. (c) Center-center distribution from 3 experiments [9] (squares) and from simulations 5 min after replication started. The simulation is positioning groups of 4 pMcms every 6.3 kb (solid line), groups of 8 pMcms every 12.5 kb (circles), or all randomly (crosses). A small random amount was added to the group location of fixed distances which was picked from a Gaussian distribution with $\sigma \sim 16\%$ of group distances. The pMcm/length ratio was fixed [cf. (b)]. (d) Phase diagram of the two-origin model to minimize replication time with changing competence and increasing the σ. Color indicates origin position relative to chromosome ends $d_1, d_2 = 1/2$ (beige) or $d_1, d_2 = 1/4$ (brown).

vis' pMcm distribution. We model a stretch of DNA of size 100 kb and $v = 1$ kb/min [20]. To determine whether the minimum-replication-time configuration requires pMcm grouping, we distributed 64 pMcms in total, i.e. that there is on average $1/1.5$ pMcm/kb as found in nature [8]. The pMcms are then placed in $64/n$ groups of $n \in \{1, 2, 4, 8, 16\}$ origins, so that the origins are uniformly distributed through the 100 kb chromosome, or completely random. Other authors have identified σ to be 6–10 min in *X. laevis* [13, 21] as well as in *S. cerevisiae* [2, 5, 20, 22]. Our results [Fig. 3(b)] indicate that grouping with an equal spacing of up to 12.5 kb achieves precise and fast DNA synthesis before the end of S-phase (20 min) for σ within these limits. We also find that 8 groups of 8 pMcms gives the advantage of a 1.1 min quicker T_{rep} than using random loci; even when the number of pMcms at these 8 groups varies, a quicker T_{rep} is achieved (data not shown). Grouping pMcms also protects the overall replication process against fluctuations from one round of the cell cycle to another; a similar problem is discussed in [14]. This is because one initiation event at an origin is sufficient to activate replication forks.

One might expect that in a natural environment there would not be strict equal spacing of groups. We now relax our previous assumption by taking evenly-spaced groups and perturb the location of each group by a small random amount drawn from a Gaussian distribution. The introduction of such variation allows us to compare our simulation with available experimental data of replicated genomic regions, which were captured as center-center distances at around 5 min after the onset of replication (cf. [9]). Figure 3(c) shows that our result is in agreement with the current understanding of the biological community, i.e. groups of 5-10 pMcms about every 10 kb. This may be achieved by a regulation of pMcm-loading proteins, whose affinity to bind decreases around existing origins [23, 24]. Although a random placement represents the data similarly well, T_{rep} remains smaller in this case where the origin groups are not equally-spaced as seen before [cf. Figure 3(b)]. This shows that grouping of origins remains favorable even in a more general setting [Fig. 3(d)].

[*] jens.karschau@abdn.ac.uk
[1] T. J. Kelly and G. W. Brown, Annu. Rev. Biochem. **69**, 829 (2000).
[2] A. P. S. de Moura *et al.*, Nucl. Acids Res. **38**, 5623 (2010).
[3] O. Hyrien and A. Goldar, Chromosome Res. **18**, 147 (2009).
[4] T. W. Spiesser, E. Klipp, and M. Barberis, Mol. Genet. Genomics **282**, 25 (2009).
[5] A. Brümmer *et al.*, PLoS Comput. Biol. **6**, e1000783 (2010).
[6] S. C.-H. Yang, N. Rhind, and J. Bechhoefer, Mol. Syst. Biol. **6**, 404 (2010).
[7] C. A. Nieduszynski *et al.*, Nucl. Acids Res. **35**, D40 (2007).
[8] H. M. Mahbubani *et al.*, J. Cell. Biol. **136**, 125 (1997).
[9] J. J. Blow *et al.*, J. Cell. Biol. **152**, 15 (2001).
[10] M. C. Edwards *et al.*, J. Biol. Chem. **277**, 33049 (2002).
[11] S. Jun and J. Bechhoefer, Physical Review E **71**, 011909 (2005).
[12] H. Zhang and J. Bechhoefer, Phys. Rev. E. **73**, 051903 (2006).
[13] A. Goldar *et al.*, PLoS ONE **3**, e2919 (2008).
[14] S. C.-H. Yang and J. Bechhoefer, Phys. Rev. E. **78**, 41917 (2008).
[15] J. J. Blow and X. Q. Ge, EMBO reports **10**, 406 (2009).
[16] S. Jun *et al.*, Cell Cycle **3**, 211 (2004).
[17] PGAPack, http://ftp.mcs.anl.gov/pub/pgapack/ .

[18] K. Shirahige *et al.*, Mol. Cell Biol. **13**, 5043 (1993).

[19] $F = \frac{1}{9} \sum_{i=1}^{n=9} d_i^o / d_i^r$ is a measure of the difference between the gap distribution of the optimized and random cases. A gap is defined as the separation between the ith experimental locus position p_i^e and that of the optimization p_i^o: $d_i^o = |p_i^e - p_i^o|$. d_i^r is akin; the average separation that arises from placing a locus uniformly and randomly and p_i^e. $F = 0$ means that the optimization fits the experimental loci positions perfectly; $F \sim 1$ indicates no difference to that of a random placement.

[20] M. K. Raghuraman *et al.*, Science **294**, 115 (2001).

[21] J. Herrick, J. Mol. Biol. **320**, 741 (2002).

[22] M. D. Sekedat *et al.*, Mol. Syst. Biol. **6**, 353 (2010).

[23] A. Rowles, S. Tada, and J. J. Blow, Journal of cell science **112**, 2011 (1999).

[24] M. Oehlmann, A. J. Score, and J. J. Blow, The Journal of cell biology **165**, 181 (2004).