

Chapter 2

Optimal Origin Placement for Minimal Replication Time

Eukaryotic genomes vary in their size and are much larger than their bacterial counterpart, e.g. that of *Saccharomyces cerevisiae* is $\sim 10^7$ bp long, those of *Xenopus laevis* or humans are $\sim 10^9$ bp in length, whereas *Escherichia coli* is $\sim 10^6$ bp. Bacteria also only have one single origin locus of replication from which they start replication, with each fork propagating at ~ 4 kbp/min [1, 2]; this allows for replication completion in under 40 min. Eukaryotic replication forks however exhibit a much slower characteristic speed, and experimental data shows that the speed of synthesis is ~ 1.5 kbp/(min·fork) [3, 4] in *Saccharomyces cerevisiae* and at ~ 0.6 kbp/(min·fork) in early *Xenopus laevis* frog embryos [5]. Let us then consider the time required for *Saccharomyces cerevisiae* DNA replication here if there were only one single origin of replication: it would take *Saccharomyces cerevisiae* almost three days to complete its genome replication.¹ In a laboratory environment yeast completes replication of its entire genome in less than about 30 min [6, 7], more than 100 times faster than what we calculated—it is clearly not the case that there is only one of replication.

The time until replication completion is accelerated by partitioning the chromosome into smaller replication domains; each of these requires an origin of replication that has formed at an origin locus. Origin loci therefore need to be placed in a manner such that replication time is minimal, i.e. replication completes by the end of S-phase. An initial guess is to space origin loci at regular intervals across a chromosome (Fig. 2.1a), if we assume origins always become licensed and activate at the same time. Such a scenario is the optimal case to result in quickest replication as compared to having the same number of origins sparsely spaced but instead groups (Fig. 2.1). Within a group only one origin is able to become active which then means that replication forks must travel farther prolonging the overall replication process (Fig. 2.1b). Therefore grouping seems to be a waste of origin resources. However we do show in this chapter that grouping is necessary to achieve minimum replication time. This is if there is uncertainty for a locus to become licensed. To compensate for not activated origins, replication forks need to travel farther than in the ideal

¹ The replication time of the yeast genome for the case of replication starting from one single origin of replication is $1.2 \cdot 10^7 \text{ bp} / (2 \cdot 1.5 \cdot 10^3 \text{ bp/min}) = 2.9$ days.

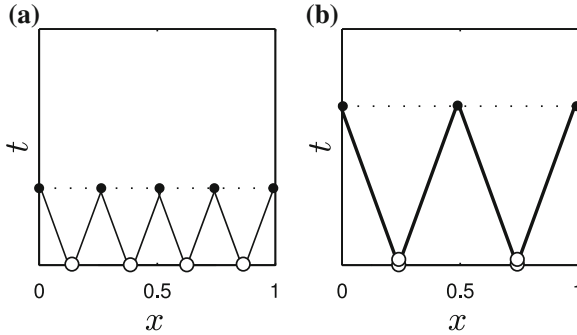


Fig. 2.1 Space-time diagram of four origins of replication (*hollow circles*). Schematic representation of origin loci distributed along the x -coordinate of a unit-sized chromosome. They have all been licensed and activate at the same time $t = 0$ min. The replication fork movement (*black line*) along the x coordinate at a given point in time is shown, and forks terminate when they coalesce or arrive at an end of the chromosome (*black filled circles*). **a** The resulting replication time is minimal if all four origins are regularly spaced. **b** If origins are grouped (shown on top of each other) only one origin of a group is able to activate. The forks must travel a longer distance at the same speeds as in (**a**); the replication time is hence longer

scenario (Fig. 2.1a). It becomes a balancing act of either spreading out origins but risking failure and longer fork travelling times, or grouping origins to compensate for the likelihood of failure and initially have longer gaps between groups. Using mathematical modelling we show that there exist certain regimes between grouped and separated origin loci positions depending on the likelihood of activation.

We relate our modelling to budding yeast *Saccharomyces cerevisiae*, which has origin loci at specific genomic positions on a chromosome—some origins in groups and some separated. For the *Saccharomyces cerevisiae* origin distribution we investigate through our model what the optimal origin distribution must be, and find that grouping of origin loci is present within *Saccharomyces cerevisiae* origin distribution to minimise replication time. This is done through an evolutionary model which searches for loci positions to give minimum replication time, and our simulations results of optimal origin positions compare well to the experimental origin distribution. We also extend our model of specific genomic positions to apply it to the case of a circular chromosome. Finally, we also introduce uncertainty in origin activation time. An origin might never have the chance to activate if it has a high chance of activating later than other origins so that it always becomes replicated by forks that originated elsewhere. We show that in such a scenario origin grouping is also a means to minimise replication time. We use the example of *Xenopus laevis* where origins appear to take random positions. In experiments, groups of origins however appear to be regularly spaced [8] which we show gives indeed minimum replication time in our model.

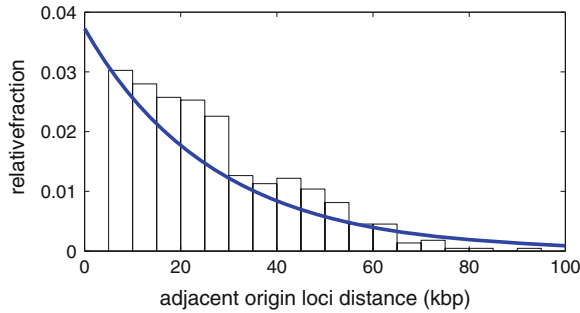


Fig. 2.2 Histogram of *Saccharomyces cerevisiae* inter-origin distances. The separation from one origin to its nearest neighbour is determined and then binned at intervals of 5 kbp (black bars). The origin position data was kindly provided by Hawkins et al. [9]. The mean of this data is 26 kbp, which is used to plot an exponential distribution with the same mean value (blue solid line)

2.1 Properties of Origins of Replication in *Saccharomyces cerevisiae*

If the origin loci in *Saccharomyces cerevisiae* would take their position within the genome randomly, then their nearest neighbour distances should be exponentially distributed. A histogram of inter-origin loci distances *Saccharomyces cerevisiae* however shows that this is not case. We show this in Fig. 2.2 where we plot a histogram of a recent study by Hawkins et al. [9]. The mean distance of the experimental data is 26 kbp which does not fit an exponential distribution with the same mean value. Also the inspection of a map for loci on the *Saccharomyces cerevisiae* genome reveals that there are groups of two or three very closely spaced origin loci which are prominent in most chromosomes [10]. We show such a map of *Saccharomyces cerevisiae* origin loci in Fig. 2.3 from the origin location data that was used in the study by Hawkins et al. [9]. Furthermore a similar map of origin loci of the fission yeast *Schizosaccharomyces pombe* gives a similar predominant grouping behaviour of origin loci (Fig. 2.4). It is to note that *Schizosaccharomyces pombe* has fewer but longer chromosomes than *Saccharomyces cerevisiae* which still require a large cohort of possible origin sites that have to be spaced with minimal gaps between to allow replication within the time allowed by the cell cycle. The data was taken from the oriDB database [10], and origin loci are shown for those classified as ‘confirmed’ or as ‘likely’.

Previous theoretical works on *Saccharomyces cerevisiae* have used the experimentally determined loci as given parameters, without attempting to understand why the origins are located where they are [11–14]. Here, we will first show an analysis of *Saccharomyces cerevisiae* origin data addressing this, and then use mathematical modelling to explain the origin loci distribution for a specific chromosome.

As discussed in Sect. 1.4, DNA replication is divided into two distinct phases; the licensing phase and the synthesis phase (S-phase). Origins in budding yeast carry a



Fig. 2.3 Map of *Saccharomyces cerevisiae* origin positions. The location of origins (*red bars*) is shown along each individual chromosome as numbered (*blue horizontal line*). For reference, the length of the smallest chromosome, chromosome 1, is 230 kbp. The origin position data was kindly provided by Hawkins et al. [9]

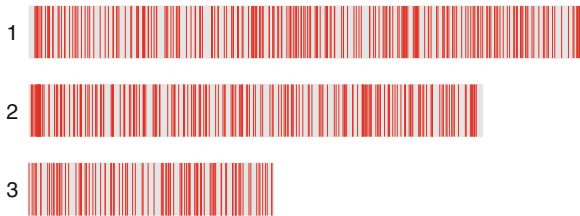


Fig. 2.4 Map of *Schizosaccharomyces pombe* origin positions. The location of origins (*red bars*) is shown along each individual chromosome as numbered (*blue horizontal line*). For reference, the length of the smallest chromosome, chromosome 3, is 2450 kbp. The origin position data was taken from the oriDB data base [10], and only those classified as either ‘confirmed’ or ‘likely’ have been considered here

certain sequence motif which allows ORC to specifically bind to a target location during licensing. This means that *Saccharomyces cerevisiae* proteins take fixed origin positions along chromosomes, and we term these positions *origin loci* to distinguish them later from licensed positions to which we refer to as *origins*. Although origin loci are at specific sites on the *Saccharomyces cerevisiae* genome this does not mean that every origin locus is going to become an active origin during each and every round of the cell cycle; i.e. not all origin loci become licensed every time. This is

because there are stochastic factors involved that hinder ORC from finding its DNA binding motif; also even once licensed an origin might not become activated during S-phase.

For the analysis, we assign to every origin locus certain (simplifying) properties. The first one is what we term *competence*, and it describes the likelihood of an origin locus to actually become licensed which will give it the ability to become activated. It is a value between zero and one and for example, a 50 % competent origin becomes on average licensed in every other round of the cell cycle, a 25 % competent one is licensed in every fourth—the larger the value, the higher the likelihood of licensing. The second property defines the time when a licensed origin activates; it is the origin activation time distribution assuming that the origin is not passively replicated. We characterise this probability density to activate in S-phase by a distribution, which in case of a Gaussian distribution has mean time of activation μ and standard deviation σ . Previous analysis of the origin activation time distribution suggests a bell-shape-like function [15], and thus a Gaussian distribution is a good first approximation. In a previous mathematical model of DNA replication which incorporates these origin properties Hawkins et al. [9] determined parameters of the entire origin population in budding yeast using a model developed by Retkute et al. [16]. They fitted their model to experimental replication timing curves of *Saccharomyces cerevisiae* to determine the competence, mean and spread of an origin activation time distribution. For their study, Hawkins et al. and Retkute et al. chose a Hill-type function to represent their origin activation time distribution which depends on two parameters t_{12} and t_w which are similar to mean and standard deviation of a Gaussian distribution. Their choice of a variant function manifests in the possibility of having origin activation prior to the begin of S-phase, which is biologically unphysical. A Hill-type function however gives origin activation times well defined between zero and later times although any other choice of function can display replication time data equally well (*personal communication with Renata Retkute*). Hawkins et al. study uncovers valuable information on the spatial distribution of origins along chromosomes, and the parameters of origin loci.

We here analyse their data which we will discuss for the remainder of this section. Of particular interest is whether specific genomic regions for origin loci are random or whether their spacing depends on the competence value of their neighbours. We calculate the sum of the competence values for adjacent origin pairs, and look at a plot of this against their genomic separation. Figure 2.5 shows that this separates groups with a low value from those with a high value. We expect that most points would be roughly near the diagonal, and the two off-diagonal corners to be empty.

Plotting the distribution of origin data shows a somewhat linear trend between the competence of neighbouring loci pairs and their separation (Fig. 2.6a). We emphasise on the left-hand tail of the distribution which shows that low competent origins *per se* are closely located for a certain parameter regime up to about 2/3. Highly competent pairs tend to be further separated from their nearest neighbour whereas low competent pairs have a tendency to be very close to each other; although there are also close nearby pairs for the case of highly competent origins. This tendency is also reflected in the correlation coefficient of 0.331 (p-value $\sim 10^{-13}$) for this data. As we show

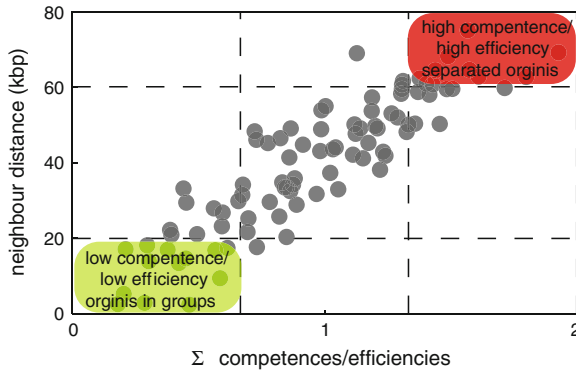


Fig. 2.5 Scheme for plotting origin neighbour distances. We plot the distance of adjacent origins of certain group size versus the sum of this competence or efficiency value of such a group. We expect that group with low competent/efficiency values have low distance to their neighbours and will be found in the *bottom left corner* (green region). As for highly competent/efficient origins, we expect these to be far away from their nearest neighbour and will be shown in the *top right hand corner* (red region)

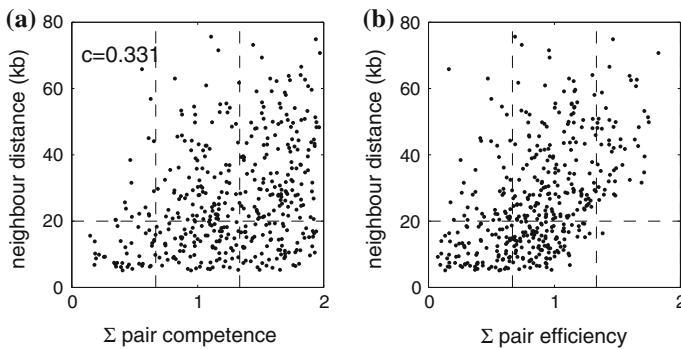


Fig. 2.6 Competence, efficiency and pair-wise neighbour distance in *Saccharomyces cerevisiae*. **a** Pairwise origin nearest-neighbour distance is plotted against their pairwise sum (Σ) of competences are. Highly competent pairs are found on the *right-hand side* of the vertical line at $4/3$, and low competent ones at the *left-hand side* of the vertical line at $2/3$. **b** Pairwise origin nearest-neighbour distances plotted here against the sum of their efficiency, i.e. the probability to become activated per round of the cell cycle

in Fig. 2.6b, a stronger trend for separation of highly competent origins holds for our analysis of efficiency—the probability of an origin being competent and also becoming activated in a particular round of the cell cycle. We emphasise that for the case of efficiency that there are no close and highly efficient origin pairs (bottom right corner) Fig. 2.6b.

This trend also persists if one considers sets of three nearest neighbouring origins. In Fig. 2.7a, b we compare the sum of competences with the maximal or minimal distance between direct origin neighbours out of a group of three adjacent origins.

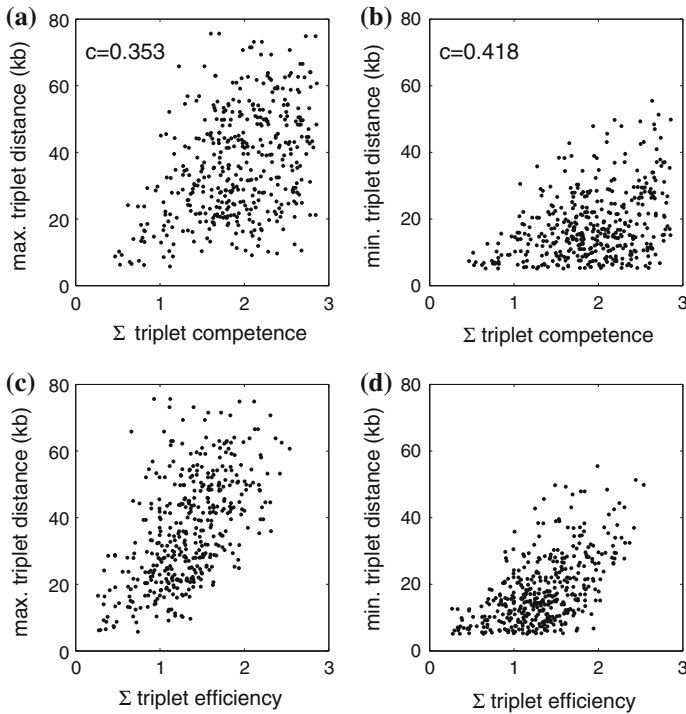


Fig. 2.7 Sets of three adjacent origins (*triplets*) are taken and either their maximum (a, c) or minimum (b, d) distance from one another within their triplet are plotted against either the (Σ) sum of their competences (a, b) or the sum of their efficiencies (c, d)

There is a striking difference in maximal, or minimal separation when considering low competent and highly competent groups of three, consistent with data for a group of two. The linear correlation between origin separation and their ability to eventually activate is even clearer when we also consider efficiency (Fig. 2.7c, d). Figure 2.7c also shows that as the efficiency of a group of three origins increases at least one origin becomes further and further separated from the other two origins. This also applies to the minimum distance of a group (Fig. 2.7d). The data gives reason to speculate that origin positions have thus been chosen preferably to compensate for origins that have little likelihood to activate by others in their surroundings.

So this data in Figs. 2.6 and 2.7 show that the proximity of origin loci correlates with their competence. These properties are therefore not independent. The remaining question is however under what conditions do origins group and whether the positions of origin loci have been favourably selected to minimise the average replication time.

2.2 A Mathematical Model for Optimal Origin Positions

The data showed that the separation of origin loci correlates with origin competence and efficiency of their neighbour(s). Yet it is unclear whether those position found in experiments are actually optimal loci positions—i.e. those giving the minimum replication time for an average of a cell population. To re-phrase the question, we can ask whether evolution has driven origin loci to their positions on the chromosome where they are found today.

2.2.1 A Simplified Two Origin Model

In a first attempt to establish a many origin model we consider the case of having only two origin loci that are positioned on a stretch of DNA. We also simplify further that origins only have a probability to activate (or fail). In other words, we only consider competence p_i for the i th origin locus. The DNA is modelled as a one-dimensional line of unit length, and we denote competences of two loci p_1 and p_2 . We initially make the assumption that origins activate at a well-defined time, $t = 0$. All replication forks travel at the same unit speed across the DNA. Specifically, we consider the geometry depicted in Fig. 2.8a where 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 this stretch of DNA will be replicated by forks originating from origins outside of the region we are considering. It will be clear shortly that our results do not depend on T_0 ; 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. Figure 2.8b illustrates 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

$$T_{\text{rep}}(d) = (1 - p_1)(1 - p_2)T_0 + (p_1 + p_2 - 2p_1p_2)(1 - d) + p_1p_2 \max\{d, (1 - 2d)/2\}. \quad (2.1)$$

This is a piecewise-linear function with discontinuity in its first derivative at $d = 1/4$, and with domain $[0, 1/2]$. Hence, T_{rep} can only have a minimum 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} . Placing both loci in the middle ($d = 1/2$) we assume that

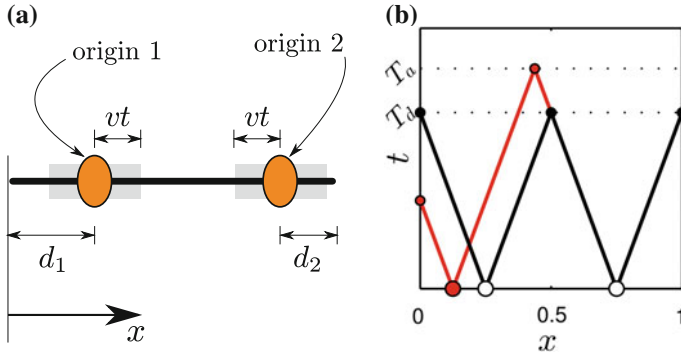


Fig. 2.8 Two origin model of DNA replication. **a** Coordinate system for origin loci with d_1, d_2 being the distance from the *left-* or *right-end* of the chromosome, respectively. x is the position coordinate along the chromosome. Replication forks travel at a speed v away from the origins. The *grey regions* show the replicated DNA at time t . **b** Space–time diagram of replication fork movement for the case of both origins starting replication at the same time $t = 0$ min. Forks move from each origin position and replication is completed once a fork reached the end of the chromosome or the last pair of forks coalesced. A symmetric placement of origins gives minimal replication time whereas an asymmetric one requires more time, i.e. $T_d < T_a$

both can activate at the same time, however the replication time is then $1/2$ for the last term in Eq. (2.1) as well as for the second term when only one activates. The replication times for $d = 1/4$ and $1/2$ are

$$T_{\text{rep}}(d = 1/2) = (1 - p_1)(1 - p_2)T_0 + (p_1 + p_2 - p_1 p_2)/2$$

and

$$T_{\text{rep}}(d = 1/4) = (1 - p_1)(1 - p_2)T_0 + (3p_1 + 3p_2 - 5p_1 p_2)/4.$$

We conclude that the two loci group together ($d = 1/2$) to achieve minimum replication time if $T_{\text{rep}}(d = 1/2) < T_{\text{rep}}(d = 1/4)$, which leads to the condition

$$p_2 < \frac{p_1}{3p_1 - 1}. \quad (2.2)$$

Notice here that T_0 drops out. The inequality Eq. (2.2) defines two regions on the p_1 – p_2 plane, corresponding to grouped or isolated loci being optimum. This is shown in Fig. 2.9a, where this analytical result is confirmed by stochastic simulations. These simulations are done employing a minimisation algorithm (using genetic algorithms [17]) which searches for the minimal replication time. The principal ingredients to the algorithms are as follows. First, origin loci are selected. Each origin locus is checked whether it will activate given its competence value, i.e. checking a random number against this probability. Finally the replication time is calculated, and this procedure repeats for several times to establish the average replication time. The positions of the origin loci are then changed, and the average replication time is

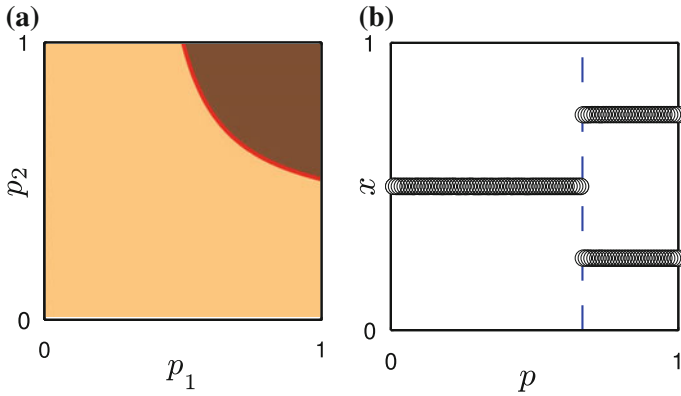


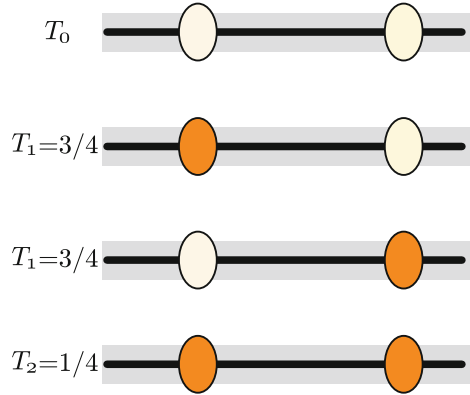
Fig. 2.9 Optimal locations in a two origin loci model. **a** Simulation results, showing optimal loci to achieve minimal T_{rep} for 2 loci with different competences, are shown for p_1 – p_2 combinations on a lattice grid. Colour indicates $d_1, d_2 = 1/2$ (beige) or $d_1, d_2 = 1/4$ (brown). The two regimes are separated by a coexistence line matched by the condition Eq.(2.2) in red. **b** Optimal position of 2 identical loci with respect to their competence p to minimize the replication time T_{rep} (circles) and $p = 2/3$ (dashed line)

calculated for this new configuration. It is then compared to other randomly selected loci positions to whether or not it results in minimum replication time. In Fig. 2.9a, 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 replication time. 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. This becomes clear with if one imagines that once a replication fork from an origin has to cover a distance more than $1/2$, such a grouped configuration becomes favourable. Figure 2.10 shows how the individual replication time (T_{rep}) terms change depending on how many origins become activated.

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 again to find the loci corresponding to the least replication time for a range of p ; these results are shown in Fig. 2.9b. The same transition also takes place for non-identical values of p_1 and p_2 —whenever one crosses from the dark to the beige region of Fig. 2.9a.

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

Fig. 2.10 The time it takes to replicate a given piece of DNA T_{rep} depends on the number of origins that activate (orange filled ovals) or not activating. This contributes to the different terms as for instance in Eq.(2.1)



both loci centered, which is near any point in the chromosome. This explains the condition for grouping if $p < 2/3$.

2.2.2 Many Origin Loci

In reality eukaryotic chromosomes have more than two loci [18], so next we investigate the case of a chromosome on which there are many loci and examine the conditions under which it becomes favourable 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

$$p_{\text{eff}} = 1 - (1 - p)^m. \quad (2.3)$$

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 larger than its critical probability p_c , given by $p_{\text{eff}} = 2/3$, which yields

$$p_c = 1 - 1/\sqrt[n]{9}. \quad (2.4)$$

Figure 2.11 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

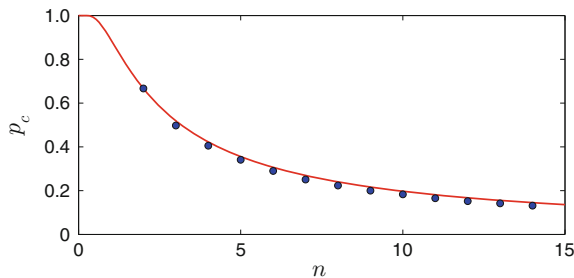


Fig. 2.11 Many origins with variable competence. **a** Probability at which groups separate p_c versus loci/group n . Shown are simulations (circles) and analytical prediction for $p_c = 1 - 1/\sqrt{n}$ (line)

low-competence loci. Our formula is also a good approximation to predict the probability at which a transition occurs for an odd number of origins in a group.

So we would expect for example a group of four origins, to break up at $p_c \approx 0.42$. Our simulations do show this to be the case (Fig. 2.12a). However the groups of four origins does not break up symmetrically into two groups of two origins, but rather into three groups. As p_{eff} increases through p , we first see two origins move out to positions $x = 1/4$ and $x = 3/4$ leaving two at $x = 1/2$. Only at a slightly larger p do we get two clusters of two. This is due to the fact that we assumed the simple case of two origins can be directly applied to the more complicated case of more origins. Figure 2.12b shows this as well where we plot the replication time for the individual configurations. This also illustrates that at first only two loci break out of the four origin group which is the crossover of the black with the blue line in Fig. 2.12b; before the blue line crosses with the red one.

2.2.3 Evolutionary Pressure Drives Yeast Origin Loci to Optimal Positions

Our hypothesis from this modelling is that selective pressure has influenced the position of origin loci through the minimization of the replication time. The theoretical result—low competence loci group, high competence loci are spread out—is also in line with our data analysis presented in Sect. 2.1. The competence data used for the analysis there however resulted in silico by model fitting to experimental data. So it required a proxy that could be potentially biased, and as a further example we now use a *Saccharomyces cerevisiae* chromosome for which origin positions and competence values are experimentally known. We then apply a search algorithm for it to find the optimal loci positions to achieve minimal replication time. This will show that in silico optimisation matches a known set of locations.

We show in Fig. 2.14 locus competence and location data for *Saccharomyces cerevisiae* chromosome VI, which has been studied extensively [12, 19]. Competences

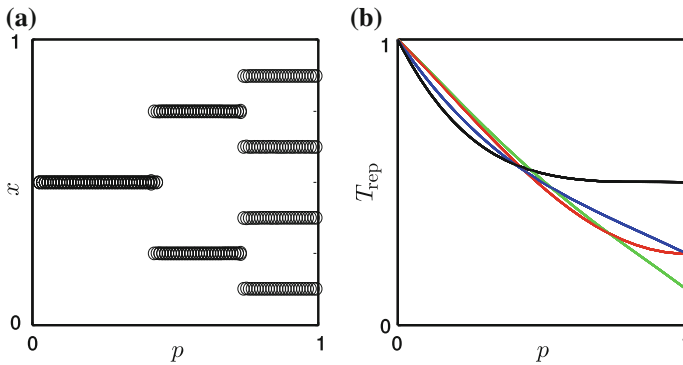


Fig. 2.12 Four origins with variable competence. **a** Simulation results for positions x of four identical origin loci with probability p . As p increases spreading loci along the chromosome of unit length results in minimal replication time. **b** The average replication time T_{rep} of arranging four origin loci positions with the same probability [corresponding to configuration shown in (a)]. The different colours of the curves correspond to: all 4 loci clustered at the middle position (*black*); 2 loci at $x = 1/2$, and 2 loci at either $x = 1/4$ or $x = 3/4$ (*blue*); groups of 2 individual loci at either $x = 1/4$ or $x = 3/4$ (*red*); individual loci $x = 0.2, 0.4, 0.6, 0.8$ (*green*)

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]. The algorithm mimicks an evolutionary process by first selecting sets of random origin locations for a parent generation of 50 individuals. The parent generation is then tested for its individual set location to give minimum replication time. The most optimal of the minimal sets are selected for the next round of iteration. They then become reshuffled amongst each other to yield a new collection of origin loci positions on this chromosome. The sets of locations are in tournament. A pair of randomly selected individuals is set to tournament, meaning the one with lower replication time succeeds. Ten new sets of location are drawn randomly and replace the ten worst (maximal replication time) location sets out of the tournament. The remaining sets produce children. They result from crossing over 85 % of the parents which are selected randomly, i.e. 15 % of the best part of a population remains unchanged to the next generation. The selection of new locations from parents results from crossover of the two parental sets of locations, i.e. either picking location 1 from parent 1 or parent 2 and so forth. They produce two children sets so that each child inherits a particular location from a particular parent to 50 %; termed crossover. Note that the number of origins always stays fixed. We then determine the replication time for the individual position sets just as before. The genetic algorithm was run with a population of 100 chromosomes of the parent generation and optimised over 2,000 iterations meaning the genomes evolve over 2,000 generations. The procedure repeats for 18,000 times with different seeds of the random number generator. Figure 2.13 summarises the algorithm detailed above. The details of parameters here lead to a local minimum set of origin location in a reasonable amount of computation time.

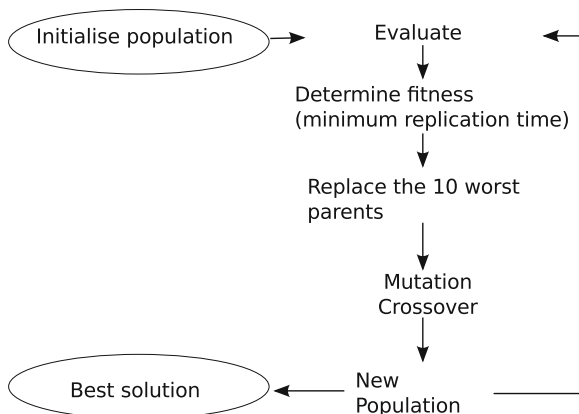


Fig. 2.13 Genetic algorithm. Summary of the steps of the genetic algorithm to determine the set of origin location to give minimum replication time

Although an appropriate choice of biological evolutionary-like parameters can be used to mimic evolution to occur over millions of years however this requires a substantial amount of extra computation time as most of the runs end in a local minimum similar to the one we find below (Figs. 2.14 and 2.15).

We remark that in our example of chromosome VI is an identifiability problem as all strong loci have $p \sim 90\%$, and we therefore constrained the ordering during the optimisation. 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 $F = 0.11^2$; 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 replication time solutions which have on average $F = 0.12$ (Fig. 2.15). This indicates that evolution has generated a near optimal solution for the proper placement of origin loci over many generations. Our study here shows a possible means to minimise replication time by choosing optimal origin loci positions. Mutations such as the translocation of genetic sequences occur frequently in unicellular, eukaryotic organisms such as yeast [20]. The rearrangement of genetic sequences—origin loci in our model—over many generations is therefore also a legitimate device in an evolutionary context to achieve minimal replication timing.

² $F = \frac{1}{9} \sum_{i=1}^{n=9} d_i^o / d_i^r$ is a measure of the difference between the gap distribution of the optimised and random cases. A gap is defined as the separation between the i^{th} 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 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.

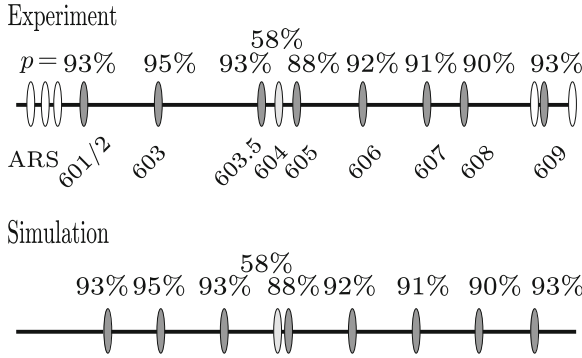


Fig. 2.14 Distribution of origin loci on yeast chromosome VI with known (grey) and unknown competences [12, 19]. The distribution results from our simulation in search for minimum T_{rep} (only grey origins considered). The group in the middle of the chromosome with a low and highly competent locus was recovered

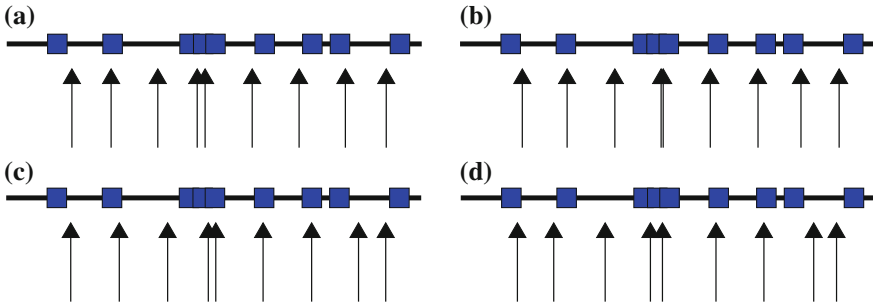


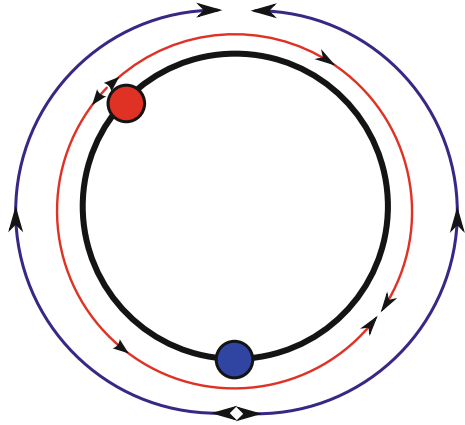
Fig. 2.15 Optimisation results for finding the minimum T_{rep} by varying the origin loci positions given their competences. The blue boxes are the experimental origin loci positions, the arrows show the positions found in simulations of individual runs. **a** Origin loci distribution that has the overall minimum replication time corresponds to Fig. 2.14. **b–d** Some distributions that give minimum replication time close to the overall minimum solution

2.2.4 Loci Competence and Circular Chromosomes

Most prokaryotes, for example the bacterium *Escherichia coli*, carry their genomic information on a single, circular chromosome. They have no compartmentalisation, meaning DNA is contained within the cytoplasm and not within a nucleus. Therefore there is no separation of licensing and origin activation as is in eukaryotes, and prokaryotes can start replication as soon as their origin locus becomes replicated. So here we can have re-replication since there is no separation of licensing from synthesis. This way they can produce concurrent copies of their DNA during exponential, unlimited growth conditions.

Their organisation of DNA replication on circular chromosome also has the advantage of only one replication fork being able to replicate its entire genome. For instance,

Fig. 2.16 One origin model of circular chromosome. The replication time for one origin (blue or red) is independent of its location due to the symmetry of a *circle*. Replication forks will always meet after travelling half circumference



the fork can start from any position on a chromosome, from which it takes the same time until to complete synthesis; it completes a full circle. This is different from the previous case of having a linear chromosome. There fork movement is more constrained because a fork cannot go around and one requires at least two forks travelling from either direction of an origin to complete DNA or have a fork starting from an edge of the chromosome. This edge effect was shown in Sect. 2.2.1 to result in preferred origin locations; only two locations that are symmetrically around the centre of a DNA segment result in the minimal replication time.

A circular chromosome also has advantage over failing origins or stalling replication forks to be easily recovered by a fork travelling towards them from elsewhere on the circle as illustrated in Fig. 2.16. So we note that all positions on a circle with a circumference we set to unit length result in the same replication time of $T_{\text{rep1}} = 1/2$ (2 forks, each replicating half of the circle); and therefore any position serves equally well to act as an origin locus. In principle, we will always observe the same T_{rep1} for a population of cells no matter where each individual cell starts its replicating from. The remaining question is whether there also exist similar origin placement conditions as we observed previously—grouped or separated; and if, so how many origins are required along with their competence value to achieve minimal replication time. We consider growth to be limiting, so that there are at maximum two copies of a chromosome and not multiple ones as during exponential growth, and again ask the question which loci positions give minimum replication time.

2.2.4.1 Two Origins

The case of two origins, shown in Fig. 2.17, results in a shorter replication time of $T_{\text{rep2}} = 1/4$, if both origins origins are maximally apart as is the case for a symmetric placement in Fig. 2.17a. An asymmetric placement however results in a replication time less optimal, depending on the maximum distance between the two origins it

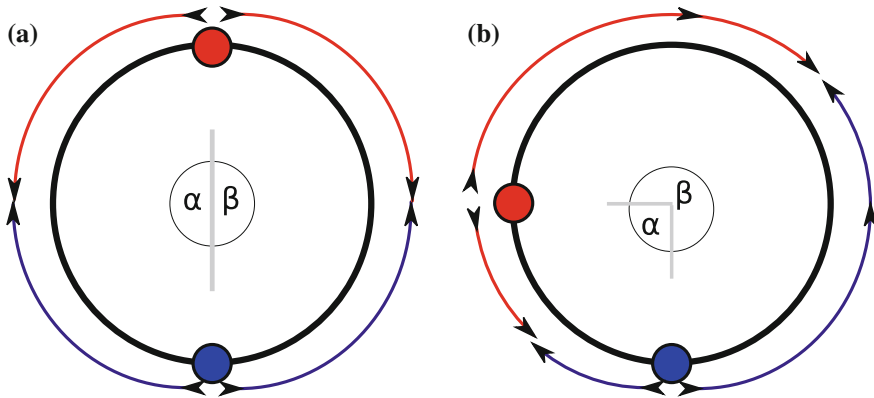


Fig. 2.17 Two-origin loci model with the angle α and β between them indicated by the grey bar. **a** The minimum replication time of $T_{\text{rep}2} = 1/4$ is achieved by placing the origin loci furthest apart from each other with distances clockwise and anticlockwise to the other origin being equal. **b** An asymmetric placement results in a longer replication time, because it takes longer for two forks to coalesce

will be $1/4 \leq T_{\text{rep}2} \leq 1/2$ (Fig. 2.17b). The other extreme is placing both origins on top of each other, for which we recover the same result as in the one-origin case. There exists only one optimal configuration, which is placing origins furthest apart which we show analytically. We define the angle between adjacent origins α and β . We note that the time of the replicated piece of the chromosome by two forks is defined by $T = \alpha / (2 \cdot 360^\circ)$, which then gives the mean replication time

$$T_{\text{rep}2} = (1 - p_1)(1 - p_2)T_0 + p_1(1 - p_2)\frac{1}{2} + p_2(1 - p_1)\frac{1}{2} + p_1p_2 \max \left\{ \frac{\alpha}{2 \cdot 360^\circ}, \frac{\beta}{2 \cdot 360^\circ} \right\}. \quad (2.5)$$

The first term accounts for neither of the origins activating, the second and third terms account for only either origin to activate and the last term if both do. The angles are constrained by one full round around the circle $360^\circ = \alpha + \beta$ which gives $\beta = 360^\circ - \alpha$. We can only find the minimum of Eq.(2.5) at either $\alpha = 0^\circ$, $\alpha = 360^\circ$ or the discontinuity of the maximum function $\alpha = 360^\circ - \alpha$ which is for $\alpha = 180^\circ$. $\alpha = 0^\circ$ and $\alpha = 360^\circ$ mean that both origins would sit on top of each other; $\max\{\alpha, \beta\} = 360^\circ$. This only leaves the configuration shown in Fig. 2.17a with both origins maximally apart to give minimum replication time.

We now write Eq. (2.5) in terms of different competence values p_1 and p_2 and include our knowledge that the minimum replication time can only be found for either $\alpha = 180^\circ$ or $\alpha = 0^\circ$, i.e. if both origins activate $T = 1/4$ or $T = 1/2$, respectively (cf. Fig. 2.17). We set $T_0 = 1$. The average replication time of both cases is then given by

$$T_{\text{rep}2}^b(p_1, p_2) = \frac{1}{4}p_1p_2 - \frac{1}{2}p_1 - \frac{1}{2}p_2 + 1, \text{ and} \quad (2.6)$$

$$T_{\text{rep}2}^b(p_1, p_2) = \frac{1}{2}p_1p_2 - \frac{1}{2}p_1 - \frac{1}{2}p_2 + 1. \quad (2.7)$$

The minimum is found using the configuration for $\alpha = 180^\circ$ [Eq. (2.6)], because $T_{\text{rep}2}^a(p_1, p_2) < T_{\text{rep}2}^b(p_1, p_2)$ for $p_1, p_2 \in (0, 1]$. So even for origins with different competence it is always best to be farthest apart from each other. This result differs from our analysis of a linear chromosomes in Sect. 2.2.1. We showed that there exists a sharp transition from finding origins together or apart depending on the parameter p_1 and p_2 for a linear chromosome.

2.2.4.2 Three Origin Loci Break Circular Symmetry: And Group Together

We now examine an odd number of origin loci and continue our analysis in terms of the time a fork travels. We take the example of three origins and place them as depicted in Fig. 2.18. The case which results in minimum $T_{\text{rep}3} = 1/6$ is again placing all origins maximally apart from each other (Fig. 2.18a). Maximum replication time is achieved by placing all three origin loci on top of each other, which is obviously not the preferred configuration to achieve an optimal replication time. This leaves two possible scenarios to arrange the origins. We place two of them maximally apart and the third one on top of any of the two (Fig. 2.18b), or the third origin is placed somewhere in the remaining halves (Fig. 2.18c). We note that: if all origin loci are always competent to activate ($p = 100\%$) then the resulting $T_{\text{rep}2} = 1/2$ which is independent of the arrangement of the third origin locus. In a more general approach, we write an expression for the average replication time

$$T_{\text{rep}3}(p) = (1 - p)^3T_0 + 3p(1 - p)^2T_1 + 3p^2(1 - p)T_2 + p^3T_3, \quad (2.8)$$

with which we show analytically that placing origin loci maximally apart is the only optimal configuration. The four different terms in Eq. (2.8) account for the possible number of origins activating during a round of the cell cycle. T_0 is the time resulting of all origins failing, but we note that it can be chosen arbitrarily as it will not influence our analysis. We choose $T_0 = 1$, as this is the longest time it takes for one single fork to complete replication. T_1 accounts for the time, if only one of the three origins activates is always independent of the placement of the failing origins. We know from the case of one origin locus that $T_1 = T_{\text{rep}1} = 1/2$. T_2 and T_3 both depend on the chosen configuration for the origin loci, and are defined by when the last coalescence event happens, so by the maximum distance a fork must travel.

The average replication times $T_{\text{rep}3}^a$, $T_{\text{rep}3}^b$ and $T_{\text{rep}3}^c$ for a circle of circumference $c = 1$ and origin loci at the positions shown in Fig. 2.18a–c are given by

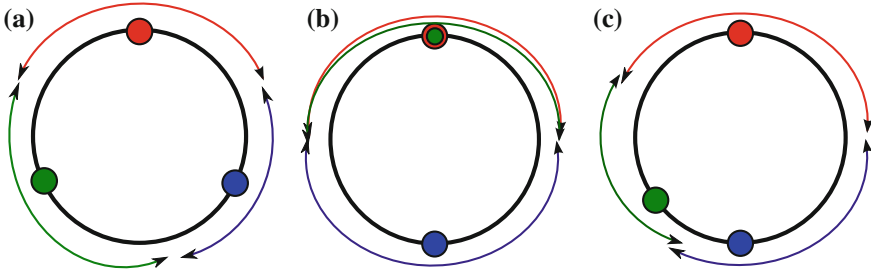


Fig. 2.18 Three origin loci on a circular chromosome. Three origin loci model with origins in either *green*, *blue* or *red*, and their corresponding forks shown as *lines*; origin loci 1, 2 and 3 respectively. Forks coalesce at the positions where two *arrowheads* meets. There exist three possible configurations to achieve minimum replication time. **a** All origins are spaced maximally apart. **b** Two origins are at either side along the diameter of the *circle* and the third origin at the same location of one of the two other. **c** The third origin can be placed in either half of the chromosome. However it does not contribute to the minimum replication time since it will always take longer to replicate the *right-hand side* of the *circle*

$$T_{\text{rep}3}^a = -1/3 p^3 + p^2 - \frac{3}{2} + 1, \quad (2.9)$$

$$T_{\text{rep}3}^b = -1/4 p^3 + p^2 - \frac{3}{2} + 1, \quad (2.10)$$

$$T_{\text{rep}3}^c = -1/4 p^3 + p^2 - \frac{3}{2} + 1. \quad (2.11)$$

We note that $T_{\text{rep}3}^a < T_{\text{rep}3}^b = T_{\text{rep}3}^c$ as well as $T_{\text{rep}3}^b = T_{\text{rep}3}^c$ is for all origin sites with the same p ; a group of two origin loci can either be situated at the top half or the bottom of the circle [configurations (2) and (3) in Fig. 2.19a]. We conclude that for all identical origin loci $T_{\text{rep}3}^a$ is the only optimal configuration, i.e. three origin loci are best placed maximally apart from each other. The cases for $T_{\text{rep}3}^b$ and $T_{\text{rep}3}^c$ both result in the same average replication time; the open boundary allows replication forks to travel around the circle. Those cases however are relevant for origin loci that differ in their competence as we show below.

We fix two loci with competence equal to 1, say $p_3 = p_2 = 1$ (red and blue loci respectively). Using a general expression for the average replication time [Eq. (2.8) for individual p_i values] one can show that the positioning of the third origin locus with variable competence has no contribution to the average replication. This is for as long as its competence value is below 0.5. We give the analytic expression of the average replication time for the configuration shown in Fig. 2.18a, c ($p_2 = p_3 = 1$), which we call $T_{\text{rep}3}^{a*}$ and $T_{\text{rep}3}^{c*}$ respectively:

$$T_{\text{rep}3}^{a*} = 1/3 - 1/6 p_1, \quad (2.12)$$

$$T_{\text{rep}3}^{c*} = 1/4. \quad (2.13)$$

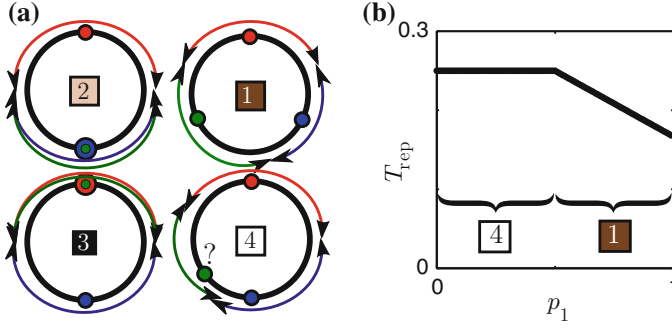


Fig. 2.19 Distribution of three loci on a circular chromosome. Origin 1, 2 and 3 have colours *green*, *blue* and *red*, respectively. **a** Three loci can be distributed in four different ways on a circular chromosome. In configuration (1) all origins are equally spread out, in configurations (2) and (3) one locus pairs with another one, and in configuration (4) the third locus can be positioned anywhere. **b** Average replication time T_{rep} of a 3 origin system with two loci of competence 100% and one origin having varying competence p_1 . T_{rep} is independent of p_1 for $p_1 < 0.5$ [configuration (4) in (a)] and for values $p_1 > 0.5$ it contributes [configuration (1) in (a)]

We see that Eq.(2.13) is independent of p_1 , the green origin, which is confirmed through stochastic simulations shown in Fig. 2.19a. There are only two possible configurations for this setting which are depicted as configuration (1) and (4) in Fig. 2.19b. Origin loci are either best placed far apart from each other, or only two origins contribute to the replication time. Minimum average replication time is achieved for the condition $T_{\text{rep}3}^{a*} < T_{\text{rep}3}^{c*}$ for $p_1 > 0.5$. Therefore a less competent origin will not influence the average replication time if combined with two highly competent origins.

Now we vary the competence of two origins, say the red origin that has $p_3 = 1$ here. We will see that there are four different configurations for this case. These are shown in Fig. 2.19a. Again using the general expression Eq.(2.8), we find that the other two green and blue origins cluster together; the red origin, origin 3, stays isolated as in Fig. 2.19a configuration (2). This is if the following condition is justified

$$p_1 < \frac{p_2}{3p_2 - 1}, \quad (2.14)$$

which corresponds to the beige region in Fig. 2.20a. The relative position of origin 1 and 2 (green and blue loci) to origin 3 (red locus of Fig. 2.19) is plotted in this figure; beige indicates locus 1 and 2 group together [configuration (2) in Fig. 2.19a], black they are 1/2 apart from each other [configuration (3) in Fig. 2.19a], brown all loci are maximally apart from each other [configuration (1) in Fig. 2.19a].

We now lower p_3 as in for example Fig. 2.20b–d where $p_3 = 0.75, 0.50, 0.25$, respectively. This makes the above mentioned four regions more visible; each corresponds to an optimal configuration. If two origin loci have same competence, the location of the weaker third origin locus can be chosen freely as it will not affect

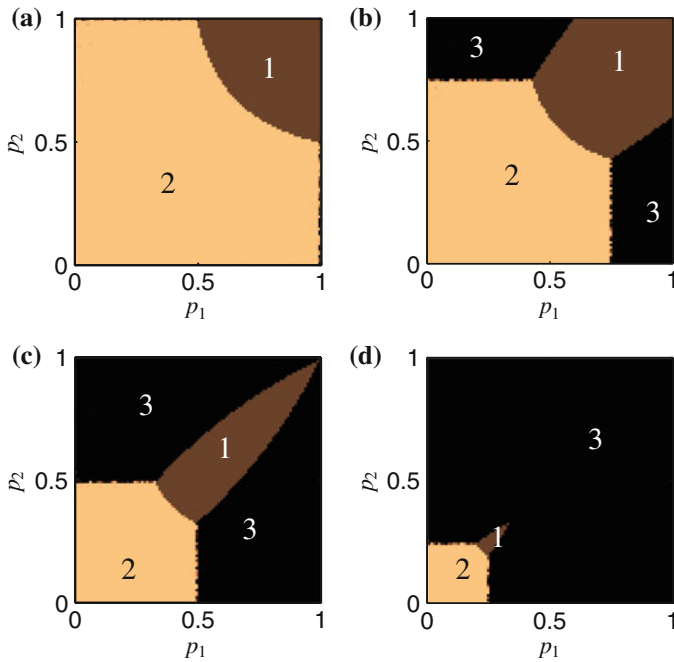
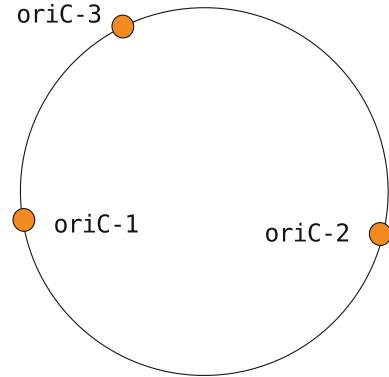


Fig. 2.20 The configurations giving minimal replication are shown for the case of three origin loci. The competence of locus 3 is fixed to either the value of $p_3 = 1.00$ in (a), $p_3 = 0.75$ in (b), $p_3 = 0.50$ in (c), or $p_3 = 0.25$ in (d). Competences p_1 and p_2 of loci 1 and 2 are varied. The colour code and numbering correspond to the configurations as shown in Fig. 2.19a. The brown (1) colour indicates complete separation of all loci. Beige (2) and black (3) colours correspond to grouping of two loci, i.e. configurations (2) and (3) of Fig. 2.19a. There is a fourth regime along $p_1 = p_2 = 1.00$ in (a), $p_1 = p_2 = 0.75$ in (b), $p_1 = p_2 = 0.50$ in (c), and $p_1 = p_2 = 0.25$ in (d) where the position of the fourth locus positions can be chosen arbitrarily. As the competence p_3 decreases the number of possible configurations of (1), where we find maximal separation, decrease; as do those for configuration (2)

the result of the average replication time, [cf. Fig. 2.19a configuration (4)]. This is the case for the randomly coloured shades as one crosses from the beige to the black region at $p_2 = p_3 = 0.50$; the configurations change from (2)→(4)→(3) (Fig. 2.19a) in Fig. 2.20b. As p_3 decreases even further the region of configuration (1) shrinks even further. Once p_3 drops below 0.50, i.e. going from Fig. 2.20c, d, the regime of configuration (3) increases. This is in agreement with Fig. 2.19b where we showed that grouping or not requires a minimum value to contribute to the average replication time.

This analysis shows that origin loci grouping is also a means of minimising replication time for a circular chromosome. If all origins are sufficiently competent they will be furthest apart from each other. A transition from where it becomes best to group two origins if they are weaker compared to a third. Then the individual loci and the group of two take a configuration similar to a two origin model; they are

Fig. 2.21 An archaeal chromosome with 3 origin loci. Schematic representation of the arrangement of the origin loci (oriC-1, oriC-2, oriC-3) of the archaea *Sulfolobus solfataricus*



1/2 apart from each other. There are only a few examples in nature where there are three origins on a circular chromosome. Most of the organisms with circular chromosomes and multiple origins are part of the kingdom of archaea [21, 22]. In Fig. 2.21, we show an example of a *Sulfolobus solfataricus* chromosome with three origin loci [23]. The arrangement of its origin loci bears resemblance with what we have shown here to be the optimal positions for loci with high competence (see also Figs. 2.19a and 2.18), and there are also several further examples as for instance in *Haloferax volcanii* [24] or *Sulfolobus islandicus* [25] with similar loci arrangements.

2.3 Optimal Origin Loci and Stochasticity in Origin Activation Time

The above discussions focused on the case of pre-defined loci in yeast and archaea, and ignored additional noise such as the variation in origin activation time. Stochastic origin activation is also well accepted by biologists and we now examine the case of stochastic activation time for *Xenopus laevis* embryos as a model organism. We remind that unlike loci in *Saccharomyces cerevisiae*, any DNA locus in a *Xenopus 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 kbp [26–28]. However do these give minimal replication time for biological relevant parameters with such an activation time distribution?

We first turn to the case where origin loci have been licensed, and there is a delay during their activation given by some activation time distribution. 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). The probability for an origin to activate at some time t is distributed according to

$$f_T(t) = \begin{cases} 1/t_b & 0 \leq t \leq t_b \\ 0, & \text{otherwise} \end{cases} \quad (2.15)$$

This distribution has mean $\mu = t_b/2$ and standard deviation $\sigma = t_b/\sqrt{12}$ and represents, for example the grey area in Fig. 2.22a. As an origin activates later than its neighbour the overall replication is delayed as well. In this scenario replication completes when all forks have either coalesced or reached the end of the DNA segment. If an origin does not activate by the time its locus is replicated from a replication fork which originated elsewhere, it then cannot become activated anymore. The replicating fork then has to continue synthesis until it reaches the end of the chromosome; which prolongs overall replication time.

We will use the same approach as for a linear chromosome in Sect. 2.1 now incorporating the uniform activation time distribution. 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. In addition, pMcms are assumed to be all identical with the same activation probability distribution (standard deviation $\sigma = t_b/\sqrt{12}$). We apply this probability distribution to the two-origin model depicted in Fig. 2.22a, and we also use the genetic search algorithm [17] to find the positions resulting in minimum replication time as σ increases. 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. If for most cases an origin activates too late it becomes replicated and cannot activate anymore. The active replication fork then has to travel a much farther to complete replication at a much later time as if all origins had activated. We test this prediction using the two-origin model with one pMcm bound to one origin; we find numerically the optimal (minimum average replication time) positions for the origins as a function of σ which are shown in Fig. 2.22b. These results show that origin grouping is also preserved in the two-origin model with stochastic variation in origin activation time. Grouping is important for swift replication under conditions of low competence and large noise which we will explain in the remainder of this chapter. We again use a segment of unit length and forks progress at unit speed of $v = 1$ kbp/min. We observe a sharp transition at $\sigma \approx 0.25$ min, above which it is best to place both origins in the middle of the segment, as observed in the case with varying competence. This is consistent with Fig. 2.22c which shows the average replication time. A minor difference between this case and the previous one in Sect. 2.1 is that for $\sigma < 0.25$, the optimal location of the origins is not constant. Origins move by a small amount further towards the edges of the chromosome. Using a Gaussian activation time distribution, as suggested by for example Goldar et al. [15] or Herrick et al. [29], also gives the transition from separated to grouped origin for a similar σ -value of around 0.25 if we fix the mean at zero (cf. also Fig. 2.23b). A uniform distribution is thus a good approximation and further has the advantage that replication can only occur after a set time $t = 0$. A Gaussian distribution however has the complication that by its definition an activation prior to the begin of S-phase is possible. The transition for the uniform distribution we use here is also

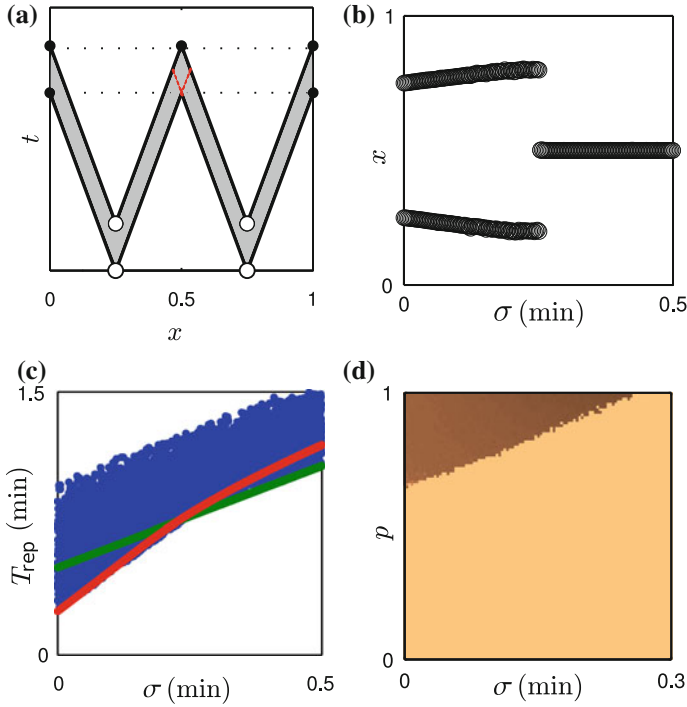


Fig. 2.22 **a** Schematic representation of space-time diagram for a two-origin system. Origins (hollow circles) can activate randomly within a time window (grey area). This will change the replication completion time (dotted lines). Forks arrive later at an edge (filled circle), and also forks from an early origin have to travel further until they coalesce with those of a late origin (red dotted line). **b** Origin position x so that the average replication T_{rep} for 2 pMcs is minimal on a segment of unit length, when the standard deviation σ for their activation time increases. **c** T_{rep} curves for two-origin systems of (b) at fixed positions (green $x = 1/2$; red: $x = 1/4$ and $x = 3/4$); or both at random sites (blue). **d** Phase diagram of the two-origin model to minimize replication time with changing competence and increasing the σ . Colour indicates origin position relative to chromosome ends $d_1, d_2 = 1/2$ (beige) or $d_1, d_2 = 1/4$ (brown)

reflected in Fig. 2.22c where the fixed positions at $x = 1/4$ and $x = 3/4$ (red solid line) result in a slightly higher replication time than compared to a random sampling of all possible configurations (blue area). Intuitively speaking, the origins group if the fork travelling towards the middle position needs to travel beyond the position of the other, i.e. it travels a distance longer than 0.5 and then has to continue until it reaches the end (see also 2.22a). Figure 2.22d shows that the transition between the group and ungrouped regimes also holds if we vary competence as well as varying σ of the uniform activation time distribution.

We also remark that our result is independent from the particulars of origin activation time distribution. Figure 2.23 depicts examples for the case of two origins and using either a distribution where an origin can activate with probability 1/2 at

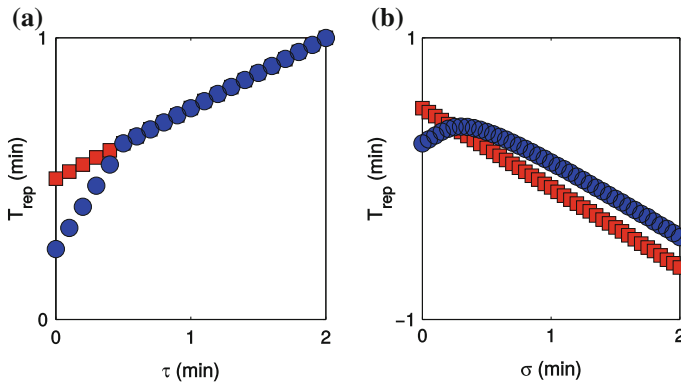
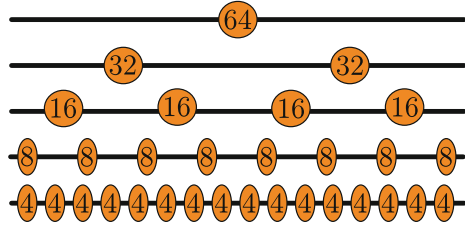


Fig. 2.23 Spread of replication times using different activation time distributions and varying the time between activations. Two origins are placed on a line of unit length either at positions $x = 1/4$ and $x = 3/4$ (blue circles), or both at position $x = 1/2$ (red squares). In (a) an origin can either activate at time $t = 0$ min or later with equal probability, thus τ denotes the difference between those times. In (b) the origin activation time is given by a Gaussian activation time distribution with zero mean. We increase its standard deviation σ . This allows for activation at times earlier than zero, hence the decreasing (and ‘negative’) average replication times

$t=0$ min or with equal probability at some later time. The difference between these times is shown as τ in Fig. 2.23a. It is clear that once the difference in origin activation time is larger than $1/2$ the configuration of having origins positioned at $1/4$ and $3/4$ does not display any advantage compared to the case of having both origins at the middle position. Similarly as the standard deviation σ of a Gaussian activation time distribution passes over a threshold value the grouped configuration gives minimum replication time (Fig. 2.23b). We note here that once the Gaussian activation time distribution becomes very wide we achieve minimum replication times as the mean is fixed at zero, however left hand tail of the distribution stretches towards negative value allowing (at least one of the) origins to start at some ‘negative’ time.

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 pMcm distribution of *Xenopus laevis*. We model a stretch of DNA of size 100 kbp and $v = 1$ kbp/min [3]. 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/kbp as found in nature [26]. The pMcms are then placed in $64/n$ groups of $n \in \{1, 2, 4, 8, 16, 32, 64\}$ origins, so that the origins are uniformly distributed through the 100 kbp chromosome, or completely random. As the group size decreases the spacing between origins becomes closer as for instance shown in Fig. 2.24. Other authors have identified σ to be 6–10 min and $\mu \sim 15$ min (Gaussian-like) in *X. laevis* [29, 30] as well as in *S. cerevisiae* [3, 4, 12, 14]. As works by Herrick et al. and Goldar et al. [29, 30] have identified the activation distribution at a fixed mean in *Xenopus laevis*, using a uniform distribution and varying σ is a good approximation for our analysis here. Our results (Fig. 2.25a) indicate

Fig. 2.24 Cartoon illustration of distributing a total of 64 pMcms at origins in groups of varying sizes to simulate the pMcm distribution in *Xenopus laevis*. As the groups of pMcms at an origin decrease the separation between individual origins decreases as well



that grouping with an equal spacing of up to 12.5 kbp 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 [31]. This is because one initiation event at an origin is sufficient to activate replication forks and result in a shorter mean time for an activation event at an origin, as we show below.

The probability of the i th pMcm activating by the time t^* given our uniform activation time distribution is

$$P(X_i = t^*) = \frac{t^*}{t_b}, \quad (2.16)$$

and the probability that a pMcm activates later than t^* is

$$\begin{aligned} P(X_i > t^*) &= \int_{t^*}^{t_b} \frac{t'}{t_b} dt' \\ &= \frac{t_b - t^*}{t_b}, \end{aligned} \quad (2.17)$$

We consider there to be a group of n identical pMcms at an origin. The probability of at least one of those activating by t^* then follows as

$$\begin{aligned} P(\min(X_i) = t^*) &= \sum_{i=1}^n \left\{ P(X_i = t^*) \prod_{j=1, j \neq i}^n P(X_j > t^*) \right\}, \\ &= n P(X_i = t^*) P(X_j > t^*)^{n-1}, \end{aligned} \quad (2.18)$$

$$= n \frac{1}{t_b} \left(\frac{t_b - t^*}{t_b} \right)^{n-1} \quad (2.19)$$

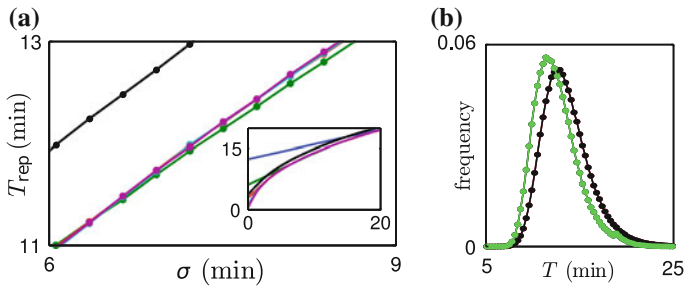


Fig. 2.25 Replication timing in *Xenopus laevis*. **a** Inset: T_{rep} as a function of σ for realistic parameters as given in the text. Origins are distributed in 4 equally-spaced groups of 16 pMcms (blue); 8 groups of 8 pMcms (green); 16 groups of 4 pMcms (red); 32 groups of 2 pMcms (cyan); 64 single 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. **b** Distribution of replication times T for a 100 kbp chromosome under the condition that $\sigma = 8$ min. Shown is the distribution of 64 pMcms in 8 equally-spaced groups of 8 pMcms (green) and placed randomly (black)

According to this origin activation time distribution the mean activation time is $t_b/(n+1)$. This shows that activation is earlier for a certain group of pMcm compared to an individual that has mean activation time $t_b/2$. So as the average activation increases through t_b it becomes a balancing act to be able to activate before a replication fork has moved across from another origin elsewhere. The origin must also not be too sparsely placed to leave small enough gaps between groups to replicate on time. Grouping is therefore a useful tactic to achieve this by lowering the overall activation time of a group of pMcm.

In a natural environment, one might expect that there would not be strict equal spacing of groups as we show it here. 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 centre-centre distances at around 5 min after the onset of replication (for instance in Blow et al. [27]). Figure 2.26 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 kbp. This may be achieved by a regulation of pMcm—loading proteins, whose affinity to bind decreases around existing origins [32, 33]. 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. Fig. 2.25b).

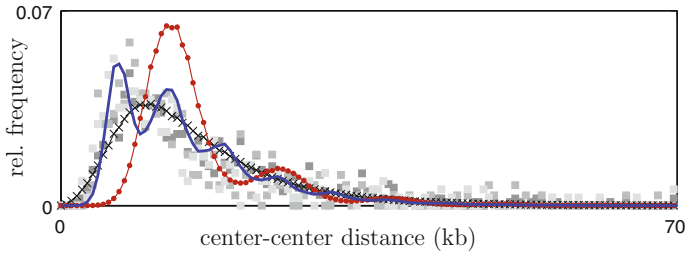


Fig. 2.26 Centre–centre distribution from three experiments [27] (*squares*) and from simulations 5 min after replication started. The simulation is positioning groups of 4 pMcms every 6.3 kbp (*solid line*), groups of 8 pMcms every 12.5 kbp (*circles*), or all randomly (*crosses*). A small random amount 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 with a total of 64 origins distributed per 100 kbp of DNA (cf. Fig. 2.25a)

2.4 Summary

Grouping is a means by which replication time is minimised and it strongly depends on the parameters of an origin. Some of the previous models of DNA replication neglected this fundamental question of where origins should be placed to minimise replication time.

We have shown that random fluctuations in the formation of origins, and the subsequent activation of proteins lead to variations in the replication time. We analysed these stochastic properties of DNA and derive the positions of origins corresponding to the minimum replication time. This was done calculating the relation between the competence of the origin to activate and the replication time; low-competence replication origins tend to group in order to minimise replication, and so do origins with long delay in their activation time. This delay is independent of the shape of the activation time distribution of the origins. Moreover we intuitively showed that origin grouping occurs to compensate for the origin failure. It thus only depends on whether or not an origin had become activated before it becomes passively replicated by a replication fork that originated elsewhere.

We have related this to experimental data in a number of species. All of those organisms show that origin grouping on linear as well as circular chromosomes is a means for minimising replication time. We finally showed arguments to prove our hypothesis that evolution has driven origins to the locations where they are found today. For this we used *Saccharomyces cerevisiae* as an example, however we propose that our results also applies to other yeast species such as *Schizosaccharomyces pombe*.

References

1. R. Reyes-Lamothe, C. Possoz, O. Danilova, D.J. Sherratt, Independent positioning and action of *Escherichia coli* replisomes in live cells. *Cell* **133**(1), 90–102 (2008). doi:[10.1016/j.cell.2008.01.044](https://doi.org/10.1016/j.cell.2008.01.044)
2. T.M. Pham, K.W. Tan, Y. Sakumura, K. Okumura, H. Maki, M.T. Akiyama, A single-molecule approach to DNA replication in *Escherichia coli* cells demonstrated that DNA polymerase III is a major determinant of fork speed. *Mol. Microbiol.* (2013). doi:[10.1111/mmi.12386](https://doi.org/10.1111/mmi.12386)
3. M.K. Raghuraman et al., Replication dynamics of the yeast genome. *Science* **294**(5540), 115–21 (2001). doi:[10.1126/science.294.5540.115](https://doi.org/10.1126/science.294.5540.115)
4. M.D. Sekedat, D. Fenyő, R.S. Rogers, A.J. Tackett, J.D. Aitchison, B.T. Chait, GINS motion reveals replication fork progression is remarkably uniform throughout the yeast genome. *Mol. Syst. Biol.* **6**, 353 (2010). doi:[10.1038/msb.2010.8](https://doi.org/10.1038/msb.2010.8)
5. H.M. Mahbubani, T. Paull, J.K. Elder, J.J. Blow, DNA replication initiates at multiple sites on plasmid DNA in *Xenopus* egg extracts. *Nucleic Acids Res.* **20**(7), 1457–1462 (1992)
6. A. Lengronne, P. Pasero, A. Bensimon, E. Schwob, Monitoring S phase progression globally and locally using BrdU incorporation in TK+ yeast strains. *Nucleic Acids Res.* **29**(7), 1433–1442 (2001)
7. C.A. Müller et al., The dynamics of genome replication using deep sequencing. *Nucleic Acids Res.* **42**(1), e3 (2013). doi:[10.1093/nar/gkt878](https://doi.org/10.1093/nar/gkt878)
8. J.J. Blow, Control of chromosomal DNA replication in the early *Xenopus* embryo. *EMBO J* **20**(13), 3293–3297 (2001). doi:[10.1093/emboj/20.13.3293](https://doi.org/10.1093/emboj/20.13.3293)
9. M. Hawkins, R. Retkute, C.A. Müller, N. Saner, T.U. Tanaka, A.P. de Moura, C.A. Nieduszynski, High-Resolution Replication Profiles Define the Stochastic Nature of Genome Replication Initiation and Termination. *Cell Rep.* **5**(4), 1132–1141 (2013). doi:[10.1016/j.celrep.2013.10.014](https://doi.org/10.1016/j.celrep.2013.10.014)
10. C.A. Nieduszynski et al., OriDB: a DNA replication origin database. *Nucl. Acids Res.* **35**, 40–46 (2007). doi:[10.1093/nar/gkl758](https://doi.org/10.1093/nar/gkl758)
11. T.W. Spiesser, E. Klipp, M. Barberis., A model for the spatiotemporal organization of DNA replication in *Saccharomyces cerevisiae*. *Mol. Genet. Genomics* **282**(1), 25–35 (2009). doi:[10.1007/s00438-009-0443-9](https://doi.org/10.1007/s00438-009-0443-9)
12. A.P.S. de Moura, R. Retkute, M. Hawkins, C.A. Nieduszynski, Mathematical modelling of whole chromosome replication. *Nucleic Acids Res.* **38**(17), 5623–5633 (2010). doi:[10.1093/nar/gkq343](https://doi.org/10.1093/nar/gkq343)
13. S.C.-H. Yang, N. Rhind, J. Bechhoefer, Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Mol. Syst. Biol.* **6**, 404 (2010). doi:[10.1038/msb.2010.61](https://doi.org/10.1038/msb.2010.61)
14. A. Brümmer, C. Salazar, V. Zinzalla, L. Alberghina, T. Höfer, Mathematical modelling of DNA replication reveals a trade-off between coherence of origin activation and robustness against rereplication. *PLoS Comput. Biol.* **6**(5), e1000783 (2010). doi:[10.1371/journal.pcbi.1000783](https://doi.org/10.1371/journal.pcbi.1000783)
15. A. Goldar, M.-C. Marsolier-Kergoat, O. Hyrien, Universal temporal profile of replication origin activation in eukaryotes. *PLoS One* **4**(6), e5899 (2009). doi:[10.1371/journal.pone.0005899](https://doi.org/10.1371/journal.pone.0005899)
16. R. Retkute, C.A. Nieduszynski, A. de Moura, Mathematical modeling of genome replication. *Phys. Rev. E* **86**(3), 031916 (2012). doi:[10.1103/PhysRevE.86.031916](https://doi.org/10.1103/PhysRevE.86.031916)
17. D. Levine, Users guide to the PGAPack parallel genetic algorithm library. (1996), <http://ftp.mcs.anl.gov/pub/pgapack/>, doi: 10.12172/366458
18. O. Hyrien, A. Goldar, Mathematical modelling of eukaryotic DNA replication. *Chromosome Res.* **18**(1), 147–161 (2010). doi:[10.1007/s10577-009-9092-4](https://doi.org/10.1007/s10577-009-9092-4)
19. K. Shirahige, T. Iwasaki, M.B. Rashid, N. Ogasawara, H. Yoshikawa, Location and characterization of autonomously replicating sequences from chromosome VI of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **13**(8), 5043–5056 (1993). doi:[10.1128/aANMCB.13.8.5043](https://doi.org/10.1128/aANMCB.13.8.5043)
20. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, *Molecular Biology of the Cell* (Garland Publishing, New York, 1994)

21. L.M. Kelman, Z. Kelman, Multiple origins of replication in archaea. *Trends Microbiol.* **12**(9), 399–401 (2004). doi:[10.1016/j.tim.2004.07.001](https://doi.org/10.1016/j.tim.2004.07.001)
22. O. Hyrien et al., From simple bacterial and archaeal replicons to replication N/U-domains. *J. Mol. Biol.* **425**(23), 4673–4689 (2013). doi:[10.1016/j.jmb.2013.09.021](https://doi.org/10.1016/j.jmb.2013.09.021)
23. I.G. Duggin, N. Dubarry, S.D. Bell, Replication termination and chromosome dimer resolution in the archaeon *Sulfolobus solfataricus*. *EMBO J.* **30**(1), 145–153 (2011). doi:[10.1038/emboj.2010.301](https://doi.org/10.1038/emboj.2010.301)
24. C. Norais, M. Hawkins, A.L. Hartman, J.A. Eisen, H. Myllykallio, T. Allers, Genetic and physical mapping of DNA replication origins in *Haloferax volcanii*. *PLoS Genet.* **3**(5), e77 (2007). doi:[10.1371/journal.pgen.0030077](https://doi.org/10.1371/journal.pgen.0030077)
25. R.Y. Samson et al., Specificity and function of archaeal DNA replication initiator proteins. *Cell Rep.* **3**(2), 485–96 (2013). doi:[10.1016/j.celrep.2013.01.002](https://doi.org/10.1016/j.celrep.2013.01.002)
26. H.M. Mahbubani, Cell Cycle Regulation of the Replication Licensing System: Involvement of a Cdk-dependent Inhibitor. *J. Cell Biol.* **136**(1), 125–135 (1997). doi:[10.1083/jcb.136.1.125](https://doi.org/10.1083/jcb.136.1.125)
27. J.J. Blow, P.J. Gillespie, D. Francis, D.A. Jackson, Replication origins in *Xenopus* egg extract Are 5–15 kilobases apart and are activated in clusters that fire at different times. *J. Cell Biol.* **152**(1), 15–25 (2001)
28. M.C. Edwards, A.V. Tutter, C. Cvetic, C.H. Gilbert, T.A. Prokhorova, J.C. Walter, MCM2-7 complexes bind chromatin in a distributed pattern surrounding the origin recognition complex in *Xenopus* egg extracts. *J. Biol. Chem.* **277**(36), 33049–33057 (2002). doi:[10.1074/jbc.M204438200](https://doi.org/10.1074/jbc.M204438200)
29. J. Herrick, S. Jun, J. Bechhoefer, A. Bensimon, Kinetic Model of DNA Replication in Eukaryotic Organisms. *J. Mol. Biol.* **320**(4), 741–750 (2002). doi:[10.1016/S0022-2836\(02\)00522-3](https://doi.org/10.1016/S0022-2836(02)00522-3)
30. A. Goldar et al., A dynamic stochastic model for DNA replication initiation in early embryos. *PLoS One* **3**(8), e2919 (2008). doi:[10.1371/journal.pone.0002919](https://doi.org/10.1371/journal.pone.0002919)
31. S.C.-H. Yang, J. Bechhoefer, How *Xenopus laevis* embryos replicate reliably: investigating the random-completion problem. *Phys. Rev. E: Stat. Nonlin. Soft Matter Phys.* **78**(4), 41917 (2008)
32. A. Rowles, S. Tada, J.J. Blow, Changes in association of the *Xenopus* origin recognition complex with chromatin on licensing of replication origins. *J. Cell Sci.* **112**, 2011–2018 (1999)
33. M. Oehlmann, A.J. Score, J.J. Blow, The role of Cdc6 in ensuring complete genome licensing and S phase checkpoint activation. *J. Cell Biol.* **165**(2), 181–90 (2004). doi:[10.1083/jcb.200311044](https://doi.org/10.1083/jcb.200311044)

Mathematical Modelling of Chromosome Replication and
Replicative Stress

Karschau, J.

2015, XIII, 76 p. 57 illus., 9 illus. in color., Hardcover

ISBN: 978-3-319-08860-0