Primordial sex facilitates the emergence of evolvable protocells

Sam Sinai[a,b,1], **Jason Olejarz**[a,1], **Iulia A. Neagu**[a,d], **and Martin A. Nowak**[a,b,c]

[a]Program for Evolutionary Dynamics, Harvard University, 1 Brattle square, suite 6, 02138; [b]Department of Organismic and Evolutionary Biology; [c]Department of Mathematics; [d]Department of Physics, Harvard University

This manuscript was compiled on April 4, 2016

Membranes, forming protocells, are widely considered beneficial or even essential to the maintenance of cooperation in early evolution [1–5]. Moreover, there are strong arguments from chemistry to suggest that membranes played a critical role in pre-evolutionary dynamics [6–9]. In this study we propose a novel reason why membranes are beneficial even before the presence of replication or selection. We argue that the ability of lipid membranes to fuse and share their contents, "primordial sex", improves the efficiency of finding minimal evolvable protocells. We analyze and quantify a model of merging membranes that resembles a sexual repair mechanism known as *multiplicity reactivation* in modern viruses [10]. We then argue that this mechanism could shorten the timescale and increase the probability of finding evolvable combinations of simple functional elements significantly. This in turn suggests that assembling complicated sets of functions at random may not be as probabilistically implausible as it first appears. Hence, in the presence of sex, large assemblies and functional networks can form without requiring evolution. Finally, we establish a quantitative framework to analyze how parasites, thought to be a serious impediment in early life, affect the accumulation of functions. We show that while parasites may hurt the accumulation process, under most circumstances, the benefits of sex massively outweigh the risks of exposure to parasitic elements.

Origin of Life | Protocells | Origin of Sex | Multiplicity Reactivation

Membranes are ubiquitous across all domains of modern life, yet their importance stretches far back to the origins of the very first cells [6–8, 11]. Prebiotic chemists [6, 7, 9], as well as origin of life theorists [2–4, 8, 12], have been interested in understanding the specific roles that membranes, in self-organized lipid vesicles (also referred to as protocells), could have played in early evolution.

The "RNA world hypothesis" concerns itself with how RNA or similar bio-polymers gave rise to information-coding and enzymatic activities that eventually lead to their central role in living organisms [13–15]. However, well-mixed populations of such molecules often suffer from well-known pitfalls, including the error catastrophe for replicases [16, 17] and parasitism for cooperative enzymes [1, 2, 5, 12]. Further, despite decades of effort in prebiotic chemistry, and some exciting progress (e.g. [18, 19]), building efficient, stable, and prebiotically plausible replicases (sometimes called the "holy grail" of the RNA world) in lab has remained a challenge [20]. Population assortment through dividing membranes seems to alleviate the parasitism problem [2–5, 12]. Apart from mathematical reasons, chemists also argue that membranes play a crucial role (e.g. producing an electro-chemical gradient) in maintaining a metabolism in early cells [6, 7, 9, 21].

While early presence of membranes has many potential benefits, it is prudent to consider whether they could have been present in abiotic earth. There is good evidence in support. It has been shown that amphiphilic molecules, like simple fatty acids that are building blocks for the lipid-membrane, could be produced in a prebiotically plausible manner [22]. These molecules are able to spontaneously assemble into vesicles in aqueous conditions [23–25]. Alternatively, lipids could have been imported to earth by chondrite meteorites[26–28]. Hence, it is commonly assumed that such molecules were present in sufficient abundances [2–4, 6–9, 12, 24] and could have produced lipid vesicles.

A "lipid world" may have preceded or coexisted with the RNA world [6–9, 29–31]. In a lipid world, protocells can contain and protect catalytic and information-bearing molecules. After the onset of replication (on a molecular or cellular level), a key step in the RNA world, protocells help selection for cooperative polymers, in particular replicases [3, 4]. Because of the potential benefits of protocells, a multitude of successful experiments in the past decade have focused on the dynamics of simple co-polymerization inside lipid vesicles [32–36].

There are several abstract properties of protocells that are of interest. First and most obviously, the contents of protocells are held near each other (are "co-localized"), and share the same fate. This results in higher concentrations, increased interactions within the protocell, and decreased interactions with outside environment. It also means that the protocell can house a "compositional genome", i.e. the information within the protocell need not be stored in one (or few) contiguous polymer [11, 37]. It may also dampen the effects of side

Significance Statement

Protocells are thought to play important roles in the origins of life. Meanwhile, some propose that sex —sharing informational content among protocells— provides benefits in early evolution. We use mathematical modeling to suggest that even before the emergence of replication (and evolution), sex could have been enormously beneficial. In particular, we show that while assembling protocells with a desired set of components is nearly impossible if the number of components is large, sex would improve the efficiency of making such cells by orders of magnitude. We quantify how much this primitive sex (which also appears in viruses) "speeds up" the inception of evolvable protocells, and once evolution begins can further increase the speed at which complexity arises.

S.S., J.O., M.A.N. designed research, S.S., J.O., I.A.N, M.A.N. performed research, S.S. J.O. I.A.N., M.A.N analyzed data, S.S., J.O., I.A.N, M.A.N. wrote the paper.

The authors declare no conflict of interest

[1]S.S. and J.O. contributed equally to this work.

[2]Martin Nowak E-mail:martin_nowak@harvard.edu

DRAFT

EFTA01183812

reactions for auto-catalytic cycles that may be required to start and maintain a metabolism [1]. Second, protocells can divide into daughter cells that inherit parts of their contents [38]. This property is at the heart of many group selection models, like the stochastic corrector [12], that alleviate problems arising from parasitism[3, 4, 39–41]. Third, protocells are able to merge and share their contents under certain conditions. While this property of protocells has been scrutinized before [1, 42, 43], it has received far less attention relative to division mechanisms. Nonetheless, there has been recent experimental success in protocell fusion models, suggesting that fusion may play a role in the development of early cells [21]. Note that some of these properties are also exhibited by other non-organic boundaries. For instance, bubbles [44] and porous materials [45] (where fluids flow through small holes and pipes) can increase local interactions, divide material, and merge them. In this study, we primarily focus on merging and its role in constructing evolvable protocells, keeping in mind that our results are general and are applicable to many processes that exhibit such properties.

We assume that in order to be evolvable, a protocell needs to contain a certain number of functions (molecules of various complexity). In early life, these could be molecules as simple as ions, co-factors, and nutrients, or more complicated polymers, like oligo-peptides, and even elementary ribozymes and simple unlinked genes [1, 19, 32–34, 46–48]. Similar models of functional assemblies have been employed successfully in the past [2, 11, 12, 37, 49–51] and simple examples have been experimentally observed [21, 52]. We call the smallest set of functions from which an evolvable protocell can be made a *minimal evolvable protocell*. More precisely, the target set should result in an auto-catalytic network that results in an evolvable cell with non-negligible probability.

In the absence of evolution through replication, a protocell will need to collect all of those functions through some random process. If the number of necessary functions that have to co-occur in a protocell is large, this process is very inefficient in a landscape where there is no evolution and replication. The absolute worst case scenario would be that out of pure luck, a membrane is formed around all the required functions at once, and results in an evolvable protocell. As this is incredibly unlikely, it is used as a criticism against approaches that require many components (or in some scenarios "genes") at inception [1]. However, as we show in the following section, alternative random mechanisms of accumulation are made possible by protocells. These mechanisms may reduce the probabilistic burden significantly enough, that even under no evolution, the target set of functions may be achievable for large number of functions.

Model and Results

The goal of our study is to compare the efficiency of mechanisms that lead to construction of a minimal evolvable protocell, in terms of the information it contains. While we take an algorithmic perspective (see [53] for a related discussion), the results can be interpreted biologically. Our target set would entail a lipid membrane that encloses all the necessary functions for starting a simple metabolism (e.g. an auto-catalytic cycle) and eventually a replication process.

We study the average-case trajectory of single cell in the population of protocells until it accumulates all the necessary

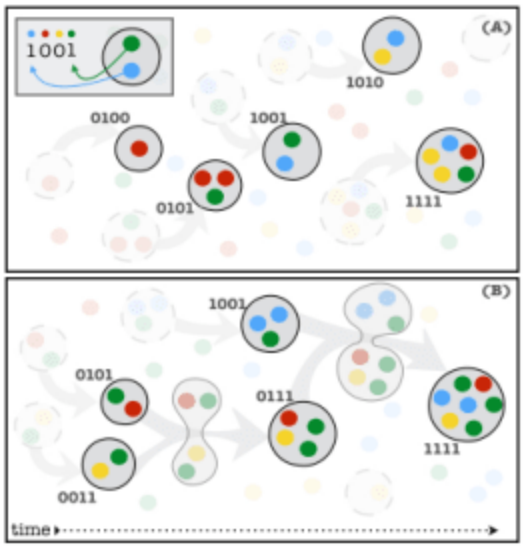

Fig. 1. Merging occurs between randomly assembled cells. A protocell consisting of all of the necessary functionalities could be constructed by (A) random assembly or (B) merging. Each merging event can be seen as a bitwise OR operation between strings of length n that represent protocell contents. Here each color represents a functionality (and so does each position on the representative strings). A cell with all the functions would be represented by a string of all 1s.

functionalities. We use the number of operations for each process, relative to the number of functions we need to reach, as our measure of complexity. A process that takes fewer steps to finish is considered more efficient. This approach allows us to analyze the processes in the same framework, and compare their efficiency. It is possible to map this measure to physical time or energy cost in a continuous chemical process, depending on the problem of interest. Here we concern ourselves with a protocell's abstract properties.

In order to mathematically measure the number of operations, we represent the functional (or genetic) content of each protocell as a binary string of length n (a boolean conjunction used similarly in [54]). For simplicity, and without loss of generality, we ignore the redundancy (or dose) of each function in the protocell, and are only concerned with their presence. If a protocell contains a particular function i, then the string will have a value of 1 at the ith position and 0 otherwise. We assume that the probability for the presence of each component (component frequency) in a sample is p (and define $q \equiv 1 - p$). We also assume that p is identical across components, as this simplifies our model but is not a key issue for our conclusions.

A process begins with an empty protocell that we track, called the accumulator. The accumulator then collects functions by "sampling" protocells and updating its own contents. Sampling is defined as picking random strings from the environment. We can measure the average-case trajectory of the accumulator protocell by calculating the expected number of independent samples required so that the accumulator acquires all n functions. We can now model an accumulation algorithm within this framework. The methods used here are commonplace in analysis of random algorithms and have also been applied to evolutionary processes [55]. The detailed calculations for the following results are provided in the SI.

DRAFT

EFTA01183813

Random assembly. In a world in which protocells cannot fuse, and material transport across the membrane is insignificant, protocells are constructed by spontaneous formation of a membrane around a random set of functions. In this world, protocells keep the set of functions they already contain. A full set is only generated when all the required functions happen to be enclosed within a membrane upon formation. Under the random assembly process, for each step, the accumulator takes the value of the latest sample. Hence, each new sample represents a protocell forming around a random set of molecules. A protocell produced in this way will contain X functions, where X is a random variable, distributed binomially with parameters n, p. Therefore, on average np functions are contained inside the protocell when it is assembled at random. However, if we need all n functions to co-occur in the same cell, then $T_R(n)$, the expected hitting time, grows exponentially in n:

$$T_R(n) = \left(\frac{1}{p}\right)^n \quad [1]$$

This calculation lays the foundation for our model as it provides a point of reference for the performance under the worst-case scenario.

Merging. As we can see from above, most of the cells generated by random assembly will contain a subset of n possible functions. However, if merging is allowed, we can intersect their contents to produce the desired set. In order to model this process we merge the accumulator protocell with random samples taken from the environment while counting the number of merges.

Table 1. Merging protocells efficiently produce cells with n components.

Number of Functions (n)	Random Assembly (Eqn 1.)	Merging (Eqn 2.)	Merging (Simulation)
10	10^{20}	291.93	291.16 ± 5.24
25	10^{50}	380.18	381.96 ± 5.47
50	10^{100}	448.17	448.48 ± 5.55
100	10^{200}	516.64	516.22 ± 5.53
250	10^{500}	607.51	608.32 ± 5.41

As an example, fixing the concentration parameter at $p = 0.01$, we compare the number of steps it takes to accumulate n functions through random assembly, and that of the merging process. While finding sets of n functions by random assembly grows exponentially fast in n, if those same randomly assembled protocells were able to merge, target compositions can be found with very few merges. We also verify the model numerically. The Monte Carlo simulation results are the mean hitting times over 2000 trials (with corresponding 95% confidence interval).

If protocells are able to merge with each other, and generate a new protocell that encloses all the functions from the two original cells, then their contents are the union of the parental cells that are generated by random assembly. Like before, merging occurs with samples that contain X functions, where X is binomially distributed. Hence, a sampled string will on average contain np functions, not all of which are necessarily new additions to the accumulator. Note that when two protocells merge, the value of the resulting string at every position i is simply determined by a bitwise OR operation (an OR operation on the ith bit of the original protocells taken together). Now, the problem can be seen as the probability of finding a string that has a 1 at every position, by sampling many strings and merging them with the current accumulator. The hitting time of this process is[1]:

$$T_M(n) = \sum_{i=1}^{n} \binom{n}{i} \frac{(-1)^{i+1}}{1-q^i} \quad [2]$$

This function is $O(\log n)$ (intuitively, grows no faster than $k \log n$ as $n \to \infty$ for some constant k). This captures the idea that you start from many protocells and merge them together to arrive at all n functions. Further, it is possible to show that the distribution of the number of merges is reasonably tight around the mean (see SI).

Note that in the limit where $p = 1/n$, each new merging protocell contains a single function on average. We call this limit the "membrane transport" process as each operation entails absorption of a single function from the outside environment into the protocell. The reader may notice that this characterization is identical to the coupon collector's problem [58]. In this limit, the hitting time is:

$$T_S(n) = nH_n \quad [3]$$

Here H_n is the n-th harmonic number. More intuitively, this function is $O(n \log n)$. As a nice check, we can see that the formula provided for the merging model (setting $p = 1/n$) is a good approximation to the hitting time predicted by the coupon collector process. This is indeed the case, and a verification is provided in the SI.

The membrane transport process is a special case of the merging process. Both processes are far more efficient that the prohibitively slow random assembly. As an aside, note that our analysis of merging membranes easily maps to other membrane transport phenomena such as heat-cycles, where protocells become more permeable to surrounding material in a periodic manner [59]. In such a case, $T(n)$ would capture the number of cycles that the protocell undergoes to capture all n functions (assuming there is net inflow).

Loss of functions. We observed above that in an ideal setting, where all samples can be incorporated into the accumulator without interruptions, merging significantly reduces the number of steps it takes to reach the target set. Obviously, while functions are being accumulated, membrane integrity may be lost, the protocell may get infected by a parasite, or the protocell may simply divide. Hence, the key test of the performance of the merging process is to understand if it can accumulate functions efficiently even in cases where it is regularly set back by events like division or death.

To address this question, which is the main contribution of our study, we consider the possibility of a restart in the accumulator. A restart can be total or partial. A total restart is equivalent to protocell death; i.e., all functions are lost, and the accumulation of functions starts anew. A partial restart occurs if a protocell divides and loses some—but not all—of its functions. Obviously, division performs better than death in the merging process, as the accumulater gains a head start on the number of functions. Therefore, calculating the hitting time by assuming a total restart (death) provides us with an upper bound on the performance of the merging process with protocell division.

[1] A reader with interest in algorithms may recognize this result as the expected height of a probabilistic skip list with n elements [56, 57].

DRAFT

EFTA01183814

We introduce a new parameter, δ, which denotes the probability of death at any given step. The accumulator makes a step (samples another protocell) with probability $1-\delta$. Given δ, we can revisit the mechanisms introduced above and incorporate the death parameter into them. For random assembly, $\delta = 1$, i.e. either all functions are accumulated on the first step, or the process restarts. We now turn our attention to the cases where $0 < \delta < 1$ (with arbitrary $0 < p < 1$) and extend the results that we obtained in the previous sections for merging.

To calculate the hitting time, we define a sequence as a series of merging events starting from a randomly assembled protocell that terminates either by accumulating all n functionalities or by being reset due to death. After each death event, a new sequence begins. Denote by F the probability that a given sequence results in all n functionalities being accumulated without being reset. We have:

$$F = \sum_{z=1}^{\infty} (1-\delta)^{z-1} \left[(1-q^z)^n - \left(1-q^{z-1}\right)^n \right]$$

Denote by $P(z)$ the probability mass function for the number of samples, z, needed to accumulate all n functionalities when starting with a randomly assembled protocell given that all n functionalities are accumulated before death. We have:

$$P(z) = \frac{(1-\delta)^{z-1} \left[(1-q^z)^n - \left(1-q^{z-1}\right)^n \right]}{F}$$

Similarly, denote by $A(z)$ the probability mass function for the number of samples, z, taken before the protocell is reset to having no functionalities when starting with a randomly assembled protocell given that the protocell dies. We have:

$$A(z) = \frac{\delta(1-\delta)^{z-1}\left[1-(1-q^z)^n\right]}{1-F}$$

Hence, the expected number of samples needed to accumulate all n functionalities is given by the exact result:

$$T(n) = \frac{\sum_{z=1}^{\infty} z[FP(z) + (1-F)A(z)]}{F} \quad [4]$$

We show a numerical verification for Eq. (4) in Figure 2. We can calculate the expected number of steps exactly through Eq. (4). However to understand the trade-off between component frequency p, death δ, and the number of functions n better we provide the following approximations. For arbitrary values $p, \delta \in (0,1)$ and large n, the expected number of steps has the following asymptotic behavior:

$$T(n) \approx \frac{-(1-\delta)\log(1-p)}{\delta^2 \Gamma(k)} n^k, \text{ where } k = \frac{\log(1-\delta)}{\log(1-p)} \quad [5]$$

There are many possible cases to consider. For example, we can simplifyEq. (5) further for small p, δ (hence $k \approx \delta/p$), and if δ is not too large relative to p. In this case, we can approximate the growth of $T(n)$ by:

$$T(n) \approx \frac{n^k}{\delta} \quad [6]$$

This equation is also plotted and verified via simulation in Figure 2. Using Eq. (5) and Eq. (6) we can see that the ratio k is the primary factor in determining the number of operations

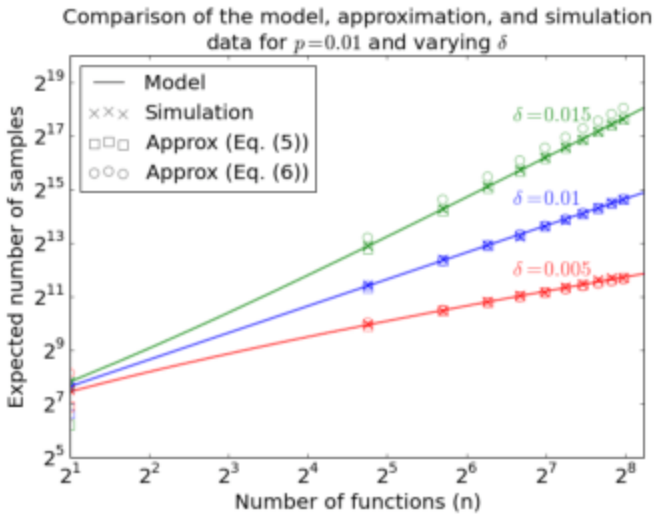

Fig. 2. Numerical verification of the merging process with death. For different values of the death parameter δ, we show the number of samples required to reach a minimal evolvable protocell. Simulation results and the approximations in Eq. (5) and Eq. (6) are provided for comparison.

required to reach the target set of functions. Remarkably, the merging process achieves a complete set of functions in low-order polynomial time for a sizable segment of the parameter space. For example, for $\delta < p$, the upper bound on the growth of $T(n)$ is $O(n)$. As another example, for $1-\delta > (1-p)^2$, the upper bound on the growth of $T(n)$ is $O(n^2)$. As long as $\delta < 1$—i.e., while there is merging of protocells—the merging process accumulates a complete set of functions in polynomial time. To clarify this further we provide a visualization for the growth of $T(n)$ with respect to p and δ in Figure 3.

Discussion

Membrane merging, and sharing of informational content, could be seen as a primitive form of sex. The idea that sex (or a similar fusion and genetic sharing mechanism) may have existed since the RNA world has been discussed for decades [1, 30, 42, 43, 60], but, to our knowledge, the time complexity of this process has not yet been quantified. We offer a simple model in the previous and use it to quantify the time complexity of the accumulation processes that result in functional or genetic assemblies (akin to compositional genomes [11, 37] or auto-catalytic sets [50]). These results establish the quantitative scale of improvement that is possible through merging (and transport across a membrane), in terms of number of operations. The time $T(n)$ required to assemble a protocell with all the necessary functionalities is reduced from an exponential number of attempts to a low-order polynomial by the merging process. Critically, our observations remain relevant even if the protocells undergo division or death during this process or if they are infected with parasites. This is the key result of this analysis. If merging is possible, the idea of cells with many co-occurring functions is no longer a probabilistic miracle, but sometimes even inevitable.

There are conjectures that in the primordial world, genomes may have been segmented and a lot of mixing and reassortment may have taken place [1, 11, 61, 62]. Our results take the benefits of sex to even before evolution (self-replication of components or protocells) started. In other words, the population

DRAFT

EFTA01183815

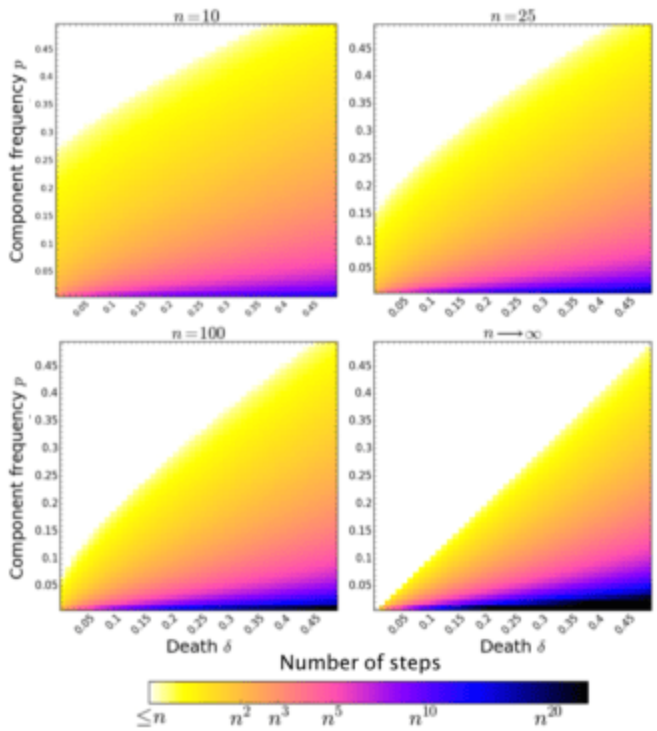

Fig. 3. Target protocells are found within polynomial number of steps through merging with death. The four panels provide a general overview of the interplay between the probability of component per sample p , probability of death δ, and multiple examples of n (number of functions). The colors represent the number of steps $T(n)$ as a function of n.

structure, and operations proposed here, improve the efficiency of finding an evolvable cell, without the need for any selection or explicit replication. The process only requires protocells, or a similar compartmentalization agent that is capable of fusion. Of course, the same results could equally apply to protocell interactions after the emergence of replication.

In that case, these results may suggest that among other benefits that sex could provide for protocells, like allowing good combinations to form, select for good "mixers" [63], and repairing lost functionalities [42], it could have also played a role in finding them quickly. Under such scenario, primordial sex through merging and content sharing preceded primordial replication (or "prelife" [64]) and lasted throughout early evolution. Obviously, these results are applicable if membranes, or similar compartments, appear early and in sufficient abundance, and the number of functions is not trivially small.

A well-recognized pitfall of early life dynamics is the problem of parasitism [2, 4]. In particular, sex increases the possibility of exposure to parasitic elements [1, 43]. However, our results show that while parasites do harm the efficiency of the process significantly, even in their presence the merging process remains tractable and reasonably efficient for a large set of parameters. Notably, here we assume that merging with a single parasite is sufficient to kill the cell, which is a strictest possible bound. In other words, barring relatively high probability of encountering a parasite at each merging, in most regimes, it is beneficial for the protocell to fuse with others (to gather functions). Hence, while the issue of parasites cannot be ignored, we address the issue at its heart by establishing an analytical framework to quantify the trade-off between abundance of parasites and performance gain provided by merging.

The merging model developed here also has similarities to well-known biological phenomena in modern viruses. The first, *Multiplicity Reactivation (MR)* [10, 65], is captured by our model. It is a process to generate an infectious particle by combining multiple non-functional mutant viruses of the same strain. In experiments, the viral particles would be subject to intense radiation such that they accumulate too many deleterious mutations and would not be able to replicate in their host. However, if several of these mutants were introduced into the same host cell, the mutant particles would "cover" each other's loss-of-function mutations, and ultimately result in a functioning virus. Our calculations complement the early models proposed by Luria and Baricelli [10, 65]. They can also be used to calculate the expected multiplicity of infection required for sexual repair in viruses, given any level of genetic damage (or mutation). Second, in *multi-compartment viruses* multiple distinct components need to co-infect the same host in order to produce a new virion. In many plant viruses, such as the genus *Tymovirus*, the infection occurs when two or more functionally distinct virions infect the same host [66, 67]. Similarly, some viral satellites and virophages need to co-infect a host in the presence of their target organism in order to reproduce [68]. These satellites are thought to transfer genetic and functional material between their hosts. These processes could serve as modern examples of similar mechanisms in early life. The fact that this type of combinatorial reproduction is present in many RNA viruses, which are thought to be ancient [62], is consistent with the suggestion that such mechanisms could have been present for a long time. If one assumes a virus-early point of view [61, 69, 70], we can readily see how this process could have contributed to the increase in complexity of cellular life. There are in fact several suggestions that RNA viruses with segmented genomes may be very ancient, and in fact may have undergone some form of mating [61, 62].

Excitingly, recent experiments have invoked fusion successfully in vitro in order to produce "self-sustaining" protocells for three generations [21]. Kurihara et al. used "conveyer protocells" (which correspond to our *samples*) to restore the chemical composition of their "giant vesicle" (*accumulator*), and thereby produced a recursive mechanism by which protocells can grow and divide for multiple generations. Our results indicate that not only can such an approach be used to construct the basic accumulator from scratch, and further provide it with metabolic nutrients, but also it can be used to efficiently increase the genetic and functional information content of a complex vesicle.

The insights we gain through this analysis could prove useful in progress towards synthetic genomes. A "minimal bacterial cell" may require a few hundred genes in order to self-sustain [71–73]. Current attempts at making such cells use a reductionist approach, where non-essential genes are pruned by trial and error to the point that all remaining genes are required for a cell to grow in a stress-free environment at a reasonable rate. Recall that in our model the ratio between frequency of fatal outcomes δ and proportion of components p (in this case genes) that carry essential functions in a given context is the key factor that determines the efficiency of the process. If this ratio is small ($\delta/p \sim 1$) in this case, it means

DRAFT

EFTA01183816

that it is possible to construct genomes by random samples of genes from pools of simple genomes within a feasible number of trials.

We hope that in light of our results, the role of protocell fusion in pre-life and early life is revisited and further considered both by theoreticians and experimentalists. In this study, we have shown that merging significantly improves the plausibility of producing protocells with a high number of components through a random process. Our results are applicable to assemblies of molecules before and after the onset of evolution. We have also provided a clear quantitative model that captures the effects of parasites (or other fatal causes) on the efficiency of the merging process. Finally, we hope that these results can be useful in analyzing viral mechanisms such as multiplicity reactivation, reassortment, and their evolutionary backgrounds.

Materials and Methods

Models were verified using Monte Carlo simulations written in Mathematica and Python. Each simulated mean is generated from 2000 independent trials, and confidence intervals are calculated using the t-Procedure. Simulation code can be provided by request from the authors.

ACKNOWLEDGMENTS. We thank Krishnendu Chatterjee for technical comments on the manuscript. We also thank Leslie Valiant and Scott Linderman for helpful comments in the initial phases of this project. We thank Robert Israel for pointing us to related literature. We thank Jeffrey M. Gerold, Carl Veller, Michael Nicholson, Ben Adlam, Nicolas Fraiman, and Tibor Antal for helpful discussions. This research was conducted using the resources provided by the Program for Evolutionary Dynamics at Harvard University. PED is supported by the John Templeton Foundation and in part by a grant from B Wu and Eric Larson.

1. Szathmáry E, Santos M, Fernando C (2005) Evolutionary potential and requirements for minimal protocells in *Prebiotic Chemistry*. (Springer), pp. 167–211.
2. Szathmáry E, Demeter L (1987) Group selection of early replicators and the origin of life. *Journal of theoretical biology* 128(4):463–486.
3. Bianconi G, Zhao K, Chen IA, Nowak MA (2013) Selection for replicases in protocells. *PLoS Comput Biol* 9(5):e1003051.
4. Markvoort AJ, Sinai S, Nowak MA (2014) Computer simulations of cellular group selection reveal mechanism for sustaining cooperation. *Journal of theoretical biology* 357:123–133.
5. Hogeweg P, Takeuchi N (2003) Multilevel selection in models of prebiotic evolution: compartments and spatial self-organization. *Origins of Life and Evolution of the Biosphere* 33(4-5):375–403.
6. Lane N, Martin WF (2012) The origin of membrane bioenergetics. *Cell* 151(7):1406–1416.
7. Luisi PL, Walde P, Oberholzer T (1999) Lipid vesicles as possible intermediates in the origin of life. *Current opinion in colloid & interface science* 4(1):33–39.
8. Segré D, Ben-Eli D, Deamer DW, Lancet D (2001) The lipid world. *Origins of Life and Evolution of the Biosphere* 31(1-2):119–145.
9. Paleos C (2015) A decisive step toward the origin of life. *Trends in Biochemical Sciences* 40(9):487–488.
10. Luria SE, Dulbecco R (1949) Genetic recombinations leading to production of active bacteriophage from ultraviolet inactivated bacteriophage particles. *Genetics* 34(2):93.
11. Segré D, Ben-Eli D, Lancet D (2000) Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies. *Proceedings of the National Academy of Sciences* 97(8):4112–4117.
12. Fontanari JF, Santos M, Szathmáry E (2006) Coexistence and error propagation in pre-biotic vesicle models: a group selection approach. *Journal of theoretical biology* 239(2):247–256.
13. Crick FH, et al. (1968) The origin of the genetic code. *Journal of molecular biology* 38(3):367–379.
14. Orgel LE (1968) Evolution of the genetic apparatus. *Journal of molecular biology* 38(3):381–393.
15. Woese C (1967) The genetic code.
16. Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58(10):465–523.
17. Nowak MA (2006) *Evolutionary dynamics*. (Harvard University Press).
18. Attwater J, Wochner A, Holliger P (2013) In-ice evolution of rna polymerase ribozyme activity. *Nature chemistry* 5(12):1011–1018.
19. Lincoln TA, Joyce GF (2009) Self-sustained replication of an rna enzyme. *Science* 323(5918):1229–1232.
20. Higgs PG, Lehman N (2014) The rna world: molecular cooperation at the origins of life. *Nature Reviews Genetics*.
21. Kurihara K et al. (2015) A recursive vesicle-based model protocell with a primitive model cell cycle. *Nature communications* 6.
22. McCollom TM, Ritter G, Simoneit BR (1999) Lipid synthesis under hydrothermal conditions by fischer-tropsch-type reactions. *Origins of Life and Evolution of the Biosphere* 29(2):153–166.
23. Yamamoto S, Maruyama Y, Hyodo Sa (2002) Dissipative particle dynamics study of spontaneous vesicle formation of amphiphilic molecules. *The Journal of chemical physics* 116(13):5842–5849.
24. Deamer DW (1986) Role of amphiphilic compounds in the evolution of membrane structure on the early earth. *Origins of Life and Evolution of the Biosphere* 17(1):3–25.
25. Mavelli F, Luisi PL (1996) Autopoietic self-reproducing vesicles: a simplified kinetic model. *The Journal of Physical Chemistry* 100(41):16600–16607.
26. Yuen G, Blair N, Des Marais DJ, Chang S (1984) Carbon isotope composition of low molecular weight hydrocarbons and monocarboxylic acids from murchison meteorite.
27. Lawless JG, Yuen GU (1979) Quantification of monocarboxylic acids in the murchison carbonaceous meteorite.
28. Deamer DW (1985) Boundary structures are formed by organic components of the murchison carbonaceous chondrite.
29. Russell MJ et al. (2014) The drive to life on wet and icy worlds. *Astrobiology* 14(4):308–343.
30. Haldane JBS (1929) The origin of life. *Rationalist Annual* 148:3–10.
31. Oparin AI, et al. (1957) The origin of life on the earth. *The origin of life on the Earth*. (3rd Ed).
32. Mansy SS et al. (2008) Template-directed synthesis of a genetic polymer in a model protocell. *Nature* 454(7200):122–125.
33. Zhu TF, Szostak JW (2009) Coupled growth and division of model protocell membranes. *Journal of the American Chemical Society* 131(15):5705–5713.
34. Budin I, Debnath A, Szostak JW (2012) Concentration-driven growth of model protocell membranes. *Journal of the American Chemical Society* 134(51):20812–20819.
35. Adamala K, Szostak JW (2013) Nonenzymatic template-directed rna synthesis inside model protocells. *Science* 342(6162):1098–1100.
36. Hanczyc MM, Fujikawa SM, Szostak JW (2003) Experimental models of primitive cellular compartments: encapsulation, growth, and division. *Science* 302(5645):618–622.
37. Segré D, Lancet D, Kedem O, Pilpel Y (1998) Graded autocatalysis replication domain (gard): kinetic analysis of self-replication in mutually catalytic sets. *Origins of Life and Evolution of the Biosphere* 28(4-6):501–514.
38. Markvoort A, Smeijers A, Pieterse K, van Santen R, Hilbers P (2007) Lipid-based mechanisms for vesicle fission. *The Journal of Physical Chemistry B* 111(20):5719–5725.
39. Simon B, Fletcher JA, Doebeli M (2013) Towards a general theory of group selection. *Evolution* 67(6):1561–1572.
40. Traulsen A, Shoresh N, Nowak MA (2008) Analytical results for individual and group selection of any intensity. *Bulletin of mathematical biology* 70(5):1410–1424.
41. Traulsen A, Nowak MA (2006) Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences* 103(29):10952–10955.
42. Bernstein H, Byerly HC, Hopf FA, Michod RE (1984) Origin of sex. *Journal of Theoretical Biology* 110(3):323–351.
43. Santos M, Zintzaras E, Szathmáry E (2003) Origin of sex revisited. *Origins of Life and Evolution of the Biosphere* 33(4-5):405–432.
44. Russell MJ, Hall A (1997) The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and ph front. *Journal of the Geological Society* 154(3):377–402.
45. Baaske P et al. (2007) Extreme accumulation of nucleotides in simulated hydrothermal pore systems. *Proceedings of the National Academy of Sciences* 104(22):9346–9351.
46. Fishkis M (2007) Steps towards the formation of a protocell: the possible role of short peptides. *Origins of Life and Evolution of Biospheres* 37(6):537–553.
47. Chen IA, Nowak MA (2012) From prelife to life: how chemical kinetics become evolutionary dynamics. *Accounts of chemical research* 45(12):2088–2096.
48. Black RA et al. (2013) Nucleobases bind to and stabilize aggregates of a prebiotic amphiphile, providing a viable mechanism for the emergence of protocells. *Proceedings of the National Academy of Sciences* 110(33):13272–13276.
49. Gánti T (1975) Organization of chemical reactions into dividing and metabolizing units: the chemotons. *BioSystems* 7(1):15–21.
50. Kauffman SA (1986) Autocatalytic sets of proteins. *Journal of theoretical biology* 119(1):1–24.
51. Ruiz-Mirazo K, Mavelli F (2008) On the way towards 'basic autonomous agents': Stochastic simulations of minimal lipid–peptide cells. *Biosystems* 91(2):374–387.
52. Vaidya N et al. (2012) Spontaneous network formation among cooperative rna replicators. *Nature* 491(7422):72–77.
53. Walker SI, Davies PC (2013) The algorithmic origins of life. *Journal of the Royal Society Interface* 10(79):20120869.
54. Valiant LG (2009) Evolvability. *Journal of the ACM (JACM)* 56(1):3.
55. Chatterjee K, Pavlogiannis A, Adlam B, Nowak MA (2014) The time scale of evolutionary innovation. *PLoS Comput Biol* 10(9):e1003818.
56. Pugh W (1990) Skip lists: a probabilistic alternative to balanced trees. *Communications of the ACM* 33(6):668–676.
57. Kirschenhofer P, Martínez C, Prodinger H (1995) Analysis of an optimized search algorithm for skip lists. *Theoretical Computer Science* 144(1):199–220.
58. Baum LE, Billingsley P (1965) Asymptotic distributions for the coupon collector's problem. *The Annals of Mathematical Statistics* pp. 1835–1839.
59. Mansy SS, Szostak JW (2008) Thermostability of model protocell membranes. *Proceedings of the National Academy of Sciences* 105(36):13351–13355.
60. Norris V, Raine D (1998) A fission-fusion origin for life. *Origins of Life and Evolution of the Biosphere* 28(4-6):523–537.
61. Koonin EV, Senkevich TG, Dolja VV (2006) The ancient virus world and evolution of cells. *Biol Direct* 1(1):29.
62. Nasir A, Caetano-Anollés G (2015) A phylogenomic data-driven exploration of viral origins and evolution. *Science advances* 1(8):e1500527.
63. Livnat A, Papadimitriou C, Dushoff J, Feldman MW (2008) A mixability theory for the role of sex in evolution. *Proceedings of the National Academy of Sciences* 105(50):19803–19808.

EFTA01183817

64. Nowak MA, Ohtsuki H (2008) Prevolutionary dynamics and the origin of evolution. *Proceedings of the National Academy of Sciences* 105(39):14924–14927.
65. Barricelli NA (1960) An analytical approach to the problems of phage recombination and reproduction: I. multiplicity reactivation and the nature of radiation damages. *Virology* 11(1):99–135.
66. Smith JM, et al. (1989) *Evolutionary genetics*. (Oxford University Press).
67. Rao A (2006) Genome packaging by spherical plant rna viruses. *Annu. Rev. Phytopathol.* 44:61–87.
68. La Scola B et al. (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* 455(7209):100–104.
69. Forterre P (2010) Defining life: the virus viewpoint. *Origins of Life and Evolution of Biospheres* 40(2):151–160.
70. Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. *Virus research* 117(1):5–16.
71. Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences* 93(19):10268–10273.
72. Gil R, Silva FJ, Peretó J, Moya A (2004) Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews* 68(3):518–537.
73. Hutchison CA et al. (2016) Design and synthesis of a minimal bacterial genome. *Science* 351(6280):aad6253.
74. Eisenberg B (2008) On the expectation of the maximum of iid geometric random variables. *Statistics & Probability Letters* 78(2):135–143.

DRAFT

EFTA01183818