Article:
Dykeman, Eric Charles (2017) A model for viral assembly around an explicit RNA sequence generates an Implicit fitness landscape. Biophysical Journal. pp. 506-516. ISSN 0006-3495

https://doi.org/10.1016/j.bpj.2017.06.037

A model for viral assembly around an explicit RNA sequence with applications to viral evolution and fitness landscapes

Eric Charles Dykeman[1]

21 June 2017

[1]Corresponding author. Address: Department of Mathematics, University of York, Deramore Lane, York, North Yorkshire YO10 5GE, United Kingdom

Abstract

Previously, a stochastic model of ssRNA virus assembly was created to model the cooperative effects between capsid proteins and genomic RNA that would occur in a packaging signal-mediated assembly process. In such a assembly scenario, multiple secondary structural elements from within the RNA, termed packaging signals (PS), contact coat proteins and facilitate efficient capsid assembly. In this work, the assembly model is extended to incorporate explicit nucleotide sequence information as well as simple aspects of RNA folding which would be occurring during the RNA/capsid co-assembly process. Applying this new paradigm to a dodecahedral viral capsid, a computer derived nucleotide sequence is evolved *de novo* that is optimal for packaging the RNA into capsids, while also containing capacity for coding for a viral protein. Analysis of the effects of mutations on the ability of the RNA sequence to successfully package into a viral capsid reveals a complex fitness landscape where the majority of mutations are neutral with respect to packaging efficiency with a small number of mutations resulting in a near complete loss of RNA packaging. Moreover, the model shows how attempts to ablate PSs in the viral RNA sequence may result in redundant PSs already present in the genome fulfilling their packaging role. This explains why recent experiments that attempt to ablate putative PSs may not see an effect on packaging. This modelling framework presents an example of how an implicit mapping can be made from genotype to a fitness

parameter important for viral biology, i.e. viral capsid yield, with potential applications to theoretical models of viral evolution.

Key words: RNA; Virus; Viral Assembly; Viral Evolution; Fitness Landscape; Packaging Signal

Introduction

Self-assembly of proteins into large biomolecular structures is ubiquitous throughout protein biochemistry. One well known and well studied example is the self-assembly of viral capsids, the protein containers which surround and protect a virus' genetic material. Although viruses employ several different mechanisms of capsid assembly and genome packaging (1–3), this paper focuses on the co-assembly mechanism that is present in plus sense single-stranded RNA (ssRNA) viruses, one of the largest class of viruses infecting a variety of hosts including humans, plants and animals. In the co-assembly process, nucleic acid and coat proteins interact to spontaneously assemble the capsid shell around the viral genome. Recent experimental and theoretical modeling work has demonstrated for a number of ssRNA viruses that specific interactions between sites within the nucleic acid (termed packaging signals - PSs) and coat proteins, facilitate the co-assembly process and are important for efficient assembly of the virion (4–6). Additional experiments with the plant satellite tobacco necrosis virus (7) have also shown that fragments of the wild-type viral genomic RNA sequence are better able to promote assembly than mutated versions, suggesting that the overall RNA sequence is important for fitness contributions related to assembly and packaging. Although theoretical descriptions of the co-assembly process exist for ssRNA viruses (8–12), none of these models are able to take the specific sequence effects on the assembly process into account, or by implication, the effects that sequence mutations would have on the packaging capacity of a viral sequence.

The observed link between a viral RNA sequence and its capacity to package presents an opportunity to construct an implicit genotype-phenotype-fitness landscape for a ssRNA virus, where fitness is measured by the yield of correctly assembled virus particles. Single-stranded RNA viruses are under a unique set of constraints to ensure that their genomes have high assembly and packaging fitness due to the fact that their genomes must perform multiple functions in the host cell. First, they must act as messenger RNAs (mRNA) by providing a template for host ribosomes to translate viral proteins. Moreover, since they do not enter a DNA stage in which DNA is integrated into the cellular genome, they must also regulate the synthesis of the different viral proteins to ensure that each is present at the concentration required for optimal replication. In addition, they must also be packaged into their protective protein shells, but only late in the infection cycle, since premature packaging of the viral RNA would result in low titres of progeny virus due to a lack of mRNA templates. Finally, the presence of other cellular mRNA competitors presents the virus with an additional challenge: how to distinguish viral RNA from host RNA.

The temporal coordination of viral protein translation and RNA packaging events that occur *in vivo* is critical for the efficient assembly of viral progeny and is believed to be controlled in part by RNA dynamics. Specifically, as the RNA genome folds into different secondary structures, it will present different structural elements with competing regulatory roles. Such regulatory elements have been identified in bacteriophage MS2 where the translational repressor (TR) serves as both a regulatory element to shut off synthesis of the viral replicase gene and as an assembly initiation site (13).

A further example is in the plant Turnip crinkle Virus (TCV), where RNA structures at the 3' end control minus strand synthesis (14). Packaging signals themselves also serve a regulatory role in facilitating efficient packaging of RNA into viral capsids. The ability of these PSs (or any regulatory elements in general) to be presented in the RNA genome will potentially impact on viral processes such as translation of viral proteins and/or packaging of viral genomes into capsids.

In order to develop a modelling paradigm which can explore the role of RNA dynamics in virus capsid assembly, as well as more broadly the regulation of the overall viral life-cycle, this paper focuses on the development of a new stochastic framework which incorporates RNA dynamics into models for packaging signal-mediated virus assembly. The resulting model fully incorporates the primary sequence of the RNA (i.e. an RNA sequence of A,U,G, and Cs) and uses RNA folding rules based on the BARRIERS method (15) and the Turner rules (16) to generate simplified dynamics. The model is then applied to a dodecamer capsid model where the capsid shell is built from 12 pentameric units, similar to capsid assembly in picornaviruses, to generate a small RNA sequence which has enhanced assembly efficiency when compared with other random RNA sequences. An interesting consequence of this model is that, by coupling RNA sequence to assembly yield, an implicit fitness landscape can be constructed where single or multiple mutations to a sequence and their effects on fitness can be assessed. This presents the opportunity of modeling recent experimental observations of PS mutations and their effects on assembly (4, 7). The need for biologically realistic fitness functions for studying evolutionary processes has been previously discussed

by Stadler (17) and this model presents a potential framework in which viral evolution could be explored using a more realistic fitness function. Some of the features of the implicit fitness landscape arising from the model, such as large regions of neutrality and examples of epistasis, are discussed.

Theory and Methods

A number of different theoretical modelling techniques exist for the prediction of viral capsid assembly kinetics. These include stochastic methods such as the Gillepsie method (8, 12), theoretical models based on energy minimization(11, 18), ODE based methods (19), and Brownian dynamics models (10). Several of these methods have been adapted to study the specific problem of RNA-coassembly that occurs in the class of ssRNA viruses (8, 10, 12). However, when incorporating the explicit RNA sequence into the model along with features of RNA dynamics such as hairpin folding, issues of computational complexity must be considered. Any assembly model which includes the RNA and its sequence specific effects, as well as mutational effects, on capsid assembly must be able to compute RNA folding kinetics fast enough such that multiple assembly simulations can be completed in minutes to hours of computational time. Thus, more advanced Brownian dynamics models will be likely too computationally expensive to accomplish this. Instead, the Gillespie method developed for the assembly of viral capsids around a static RNA genome (8) is adapted for simulating RNA hairpin folding using a variant of the Gillespie algorithm, i.e. the next reaction method (20). The advantage of the next reaction method is that

it will allow for the construction of a binary queuing system with a computational cost per reaction fired of $O(Log_2(N))$ in contrast to the $O(N)$ cost in the traditional Gillespie algorithm, allowing for effects such as mutations to the RNA sequence and its impact on assembly to be explored in less computational time.

RNA Folding Kinetics Model for ssRNA Virus Assembly

The RNA kinetics model that is used in combination with the assembly model for ssRNA viruses is based on the BARRIERS method (15). The BARRIERS method uses the Wuchty sub-optimal RNA folding algorithm (21) to identify a set of RNA folds that are within a specified energy difference of the minimum free energy fold. Given this set of RNA folds representing a subspace of the complete RNA folding space, BARRIERS identifies a set of local minima and saddle points. Once all local minima and saddles have been identified, BARRIERS constructs transitions between pairs of local minima that connect through a saddle point and calculates the transition rate based on the height of the energy barrier between the two local minima. The BARRIERS method can then perform RNA kinetics using the set of local minima as the different RNA folding states and a stochastic based method (such as Gillespie (22)), or a numerical method for solving the set of coupled differential equations that results, can be used to simulate the kinetics of RNA folding.

The general BARRIERS method will produce RNA structures which have long-distance interactions, hairpins, and multi-loops. However, for a variety of viruses including bacteriophage MS2 (6), satellite tobacco necrosis

virus (5), and Human Parecho virus (4), experiments have shown that the RNA structures which are involved in the binding to coat proteins are short simple hairpins (of about 20 nt) with specific sequence or structural features. Moreover, many ssRNA viruses, such as those from the picornavirus family, are also believed to assemble from the 5' ends of RNAs during synthesis of the plus strand. This suggests that local hairpin structures are likely to be more important than long range interactions for some ssRNA viruses such as those from the picornavirus family. Given this information, the RNA folding model for virus assembly is simplified to only include local hairpins spanning a user-specified number of nucleotides which will allow for a substantial increase in the computational efficiency of the RNA folding part of the assembly algorithm. This results in the RNA sequence being modelled as a linear chain of RNA hairpins with different sequence, bulge, and apical loop configurations which can fold and melt at different rates depending on their base-pairing and stacking interactions. To simplify the transitions between RNA states, individual hairpins can either form or melt, but cannot transition to another hairpin state (c.f. Figure 1c). Finally, based on the sequence and structure of the RNA hairpin, the on and off binding rates to coat protein can then be assigned based on user specified rules which will depend on the specific virus of interest. It may be possible in the future to incorporate long distance interactions in RNA kinetics, however current coarse grained folding algorithms can take up to 24 hours to complete 100 seconds of folding (23), which is not currently fast enough to simulate multiple thousands of assembly reactions.

The RNA folding states are computed in three steps. First, using a

variant of the Waterman-Byers algorithm (24), all RNA hairpins that are on a given RNA sequence which contain a user-specified maximum number of nucleotides are computed. Second, hairpins which are local minima are identified by testing if the removal or addition of a base-pair results in a lower energy structure. Finally, the energy barrier is calculated between the folded and unfolded states of the hairpin and the folding and melting rates computed.The construction of the RNA folding states only needs to be pre-computed once prior to simulating the assembly of the virus. Once the RNA folding states are computed (i.e. the set of all possible hairpins that can form), the assembly simulation chooses from a set of reactions to fire which include CP binding/unbinding to the folded RNA hairpins and CP-CP association dissociation (Figure 1b), or RNA hairpin folding/melting (Figure 1c).

Next Reaction Method for Virus Assembly with RNA Folding

Before presenting the next reaction method for virus assembly and RNA folding, the Gillespie model for virus assembly around a static RNA (8) is first briefly discussed as its algorithmic procedure serves as the basis for the assembly model which incorporates RNA folding. The Gillespie method for virus assembly around a static RNA containing a number of binding sites (as depicted in Figure 1a), stores information on the PS binding sites, their affinities, as well as the capsid proteins which these PSs are in contact with for each RNA in the simulation. From this configuration information, one can quickly calculate the total number reactions that are possible for any RNA in the system which includes CP-CP association/dissociation and PS-

CP binding/unbinding as depicted in Fig. 1b. The reaction flux $a_0(\alpha)$ for a single RNA α in the simulation is then calculated by summing over all $i = \{1, M_\alpha\}$ reaction rates, $a_i(\alpha)$, that are possible for this RNA/capsid complex

$$a_0(\alpha) = \sum_{i=1}^{M_\alpha} a_i(\alpha). \tag{1}$$

From this, the total reaction flux Φ is computed for the entire system of $\alpha = \{1, N_r\}$ RNAs by summing over the reaction flux for each individual RNA, i.e.

$$\Phi = \sum_{\alpha=1}^{N_r} a_0(\alpha). \tag{2}$$

To choose a reaction to "fire", the assembly algorithm first chooses a random number r between zero and one, $r = [0, 1]$, then computes $\bar{\Phi} = r\Phi$. Following the traditional Gillespie stochastic algorithm, the RNA μ is identified which satisfies the partial sum inequality

$$\sum_{\alpha=1}^{\mu} a_0(\alpha) \geq \bar{\Phi}. \tag{3}$$

After choosing the RNA μ based on the value $\bar{\Phi}$, a specific reaction to fire out of the M_μ possible reactions for this RNA is identified by finding the reaction j such that

$$\sum_{\alpha=1}^{\mu-1} a_0(\alpha) + \sum_{i=1}^{j} a_i(\mu) \geq \bar{\Phi}. \tag{4}$$

Now that the specific reaction to fire in RNA μ has been chosen, the system can be updated according to this reaction and the time incremented by τ

according to

$$\tau = \frac{-ln(r)}{\Phi}, \tag{5}$$

where r is a random number between zero and one. Since only one RNA and its list of reactions and reaction rates change after each Gillespie step, a binary tree containing partial sums of the reaction fluxes for each RNA, $a_0(\alpha)$, can be used to quickly identify in $Log_2(N_r)$ time the RNA μ containing the reaction to fire next as well as re-sum the total flux Φ, greatly speeding up the reaction selection and total flux summation tasks. This procedure follows a similar binary tree method which was used to speed up a Gillepie model of RNA kinetics at single base-pair resolution (25).

To incorporate RNA folding into this algorithm, a variant of the traditional Gillespie method called the next reaction method is used (20). The next reaction method differs to the traditional Gillespie method in two key ways. First, instead of calculating the total flux Φ in Eq. 2, picking a reaction to fire, and then calculating the time that reaction occurs using Eq. 5, the next reaction method samples the wait time for each possible reaction $i = [1, M_\alpha]$ in RNA α using

$$\tau_i(\alpha) = \frac{-ln(r_i(\alpha))}{a_i(\alpha)}, \tag{6}$$

where the $r_i(\alpha)$ are random numbers between zero and one. Second, wait times are only sampled once, i.e. the random numbers are re-used, until the reaction actually fires. The next reaction method simulates a series of chemical reactions by selecting the smallest wait time to be the next reaction that is fired. Once the reaction has been fired, a new time is sampled for

the reaction and the process repeats. Although the procedure of selecting the reaction to fire in the next reaction method differs from the traditional Gillespie algorithm, one can show that they are mathematically equivalent and sample the same chemical kinetics (20).

In the case of virus assembly with RNA folding, where the RNA is a simple linear chain of N_h hairpins that are either present or not, a queue table in the form of a binary tree can be employed to allow for the selection of the next reaction with minimum wait time in $O(1)$ time with update of the queue table in $O(Log_2(N_h))$ time. Consider the case of $N_h = 2^n$ hairpins which can have states of either folded (and thus present in the RNA strand) or un-folded. Assign to each hairpin a wait time to fold or unfold (depending on its current state) according to Eq. 6 and construct a queue table in the form of a binary tree on an array of length N_h. Once the sorted binary tree has been constructed, the $m = N_h/2$ element of array points to the hairpin number which has the minimum wait time. After either folding/unfolding the hairpin with the minimum wait time, a new wait time for the hairpin to unfold/fold is sampled using Eq. 6 and the queue table can be updated in $O(Log_2(N_h))$ steps. Supplementary Figure 1 illustrates the queue table and the selection and update process. When an unfolded hairpin (hairpin A) which has minimum wait time in the queue overlaps with another hairpin in the RNA sequence (hairpin B) which is already present, a new folding time for hairpin A is queued, corresponding to the time to wait for hairpin B to unfold plus the time to wait for hairpin A to fold. In this fashion, reactions that are forbidden due to changes in the RNA fold are re-queued to occur at an appropriate time in the future. The advantage of the next

reaction method over the traditional Gillespie algorithm is that it removes the time consuming step of looping over all possible hairpin reactions (which may be in the thousands even for a small RNA sequence), checking if each can occur, and then adding the appropriate reaction flux for that hairpin folding/melting reaction into the total flux which would require $O(N_h)$ time whenever a reaction is fired. A similar queue system can be constructed for the capsid assembly reactions for consistency. Benchmarking of the assembly model has shown that the simulation of capsid assembly with around 2000 copies of a 360 nt RNA sequence can be completed in around 10-20 min on a single processor, making it feasible to explore RNA sequence space and its impact on assembly via a genetic algorithm.

Results

To illustrate the features of the assembly model and the resulting implicit fitness landscape, the assembly of a small capsid comprising 12 pentameric units which assemble into a dodecahedral capsid around a small 360 nt RNA is examined using the next reaction method. As before, we consider the assembly model depicted in Figure 1 which requires 12 PS sites to be present in the genome, one for each pentameric unit, with appropriate affinities to successfully promote virus assembly. In this way, this model follows previous work (8), but with the added complexity that the explicit RNA sequence is present and must fold PSs to present to the coat protein for binding. As a result, not all PSs may be able to fold due to competing hairpins which may block their folding and hence binding to CPs. As discussed above, al-

though long-distance interactions are neglected to simplify the computation, this assembly scenario is related to the picornavirus genus of viruses which includes the virus families Parechoviruses (HPEV), Apithoviruses (FMDV), and Enteroviruses (Polio). In these viruses, assembly is believed to take place during synthesis of the plus strand via interaction with local hairpin structures (4). After choosing model parameters in the next section, an RNA sequence with high assembly efficiency is evolved using the assembly model and details of the sequence's fitness are examined.

Choice of Model Parameters

There are only three types of model parameters that are required to be chosen in order to model capsid assembly; (1) the CP-CP association/dissociation rates, (2) the CP-RNA binding/unbinding rates for each hairpin, and (3) the folding/melting rates for each hairpin. The CP-CP association/dissociation rates, κ_a and κ_d, can be determined from the relation

$$\frac{\kappa_a}{\kappa_d} = e^{-\beta \Delta G_p}, \tag{7}$$

where $\beta = 1/k_b T$, and ΔG_p is the change in free energy due to coat protein association with a partially formed capsid. Assuming the CP-CP association rate κ_a has a rate of $\kappa_a = 10^6 s^{-1}$, consistent with previous RNA-CP assembly models (8), dissociation rates κ_d can then be computed using Eq. 7 and the number of contacts n_c made between the incoming CP and the partially formed capsid, i.e. $\Delta G_p = n_c C$. In this model, a free energy change per contact of $C = -2.5 \text{kcal} M^{-1}$ is used, which is approximately

the value needed for free pentamers to assemble into complete capsids in the RNA-free situation (19).

The remaining constants are attributable to the N_h individual hairpins that have been identified via the BARRIERS method. Each hairpin will have a total of four rates associated with it; a rate of binding to CP, a rate of unbinding from CP, a rate of folding, and a rate of melting. The folding and melting rates for hairpin i, $\kappa_f(i)$ and $\kappa_m(i)$, are calculated from the minimum free energy barrier between the folded and single stranded states, $\Delta G_f(i)$, using the formula

$$\kappa_f(i) = Ae^{-\beta \Delta G_f(i)} \tag{8}$$

$$\kappa_m(i) = Ae^{-\beta \Delta G_m(i)} \tag{9}$$

where $\Delta G_m(i) = \Delta G_f(i) - G(i)$ is the free energy barrier for melting and $G(i)$ is the free energy of the fully folded hairpin i, calculated using the Turner rules (16). The constant A is related to the attempt frequency in the Arrhenius equation and is set to $A = 10^7$ which yields folding rates on the order of $10 - 100$ns, consistent with estimates for small hairpins (26, 27).

For the binding and unbinding rates for hairpin i, $\kappa_b(i)$ and $\kappa_u(i)$, affinities are assigned based on both the secondary structure of the hairpin as well specific features of its sequence. First, the secondary structure of the hairpin is checked against the three possible structures that are allowed to bind to CP (c.f. Figure 1d). If the hairpin matches one of these, it is assigned the generic dissociation constant for that structure as shown in Figure 1d. Next, the dissociation constant is adjusted by a multiplicative factor $F = F_1F_2F_3$

depending on the specific nucleotide sequence of the hairpin at the three specific positions (numbered in Figure 1d). For each position $i = [1 \dots 3]$, the multiplicative factor F_i is obtained from Table 1. Once F has been calculated, the total dissociation constant for hairpin i, $K_D(i)$, is obtained. The following relation can then be used to calculate the binding and unbinding rates $\kappa_b(i)$ and $\kappa_u(i)$

$$K_D = \frac{k_u(i)}{k_b(i)} = e^{\beta \Delta G_b}, \tag{10}$$

where $k_u(i) = \kappa_u(i)$ and $\kappa_b(i) = \frac{k_b(i)}{V}$. Stopped flow kinetic binding experiments have estimated the binding rate for these hairpins to CP are diffusion limited and are roughly on the order $k_b = 1.1 \times 10^7\ M^{-1}s^{-1}$ (28). The same value is used here for all hairpins with one of the three structural types shown in Figure 1d. Using a volume of $V = 0.7\mu m^3$, a typical volume of a small cell, the generic binding rate for binding competent hairpins is $\kappa_b(i) = 0.0261s^{-1}$. The unbinding rate is then computed from Eq. 10 using the $K_D(i)$ value calculated for the hairpin based on its sequence. All other hairpin structures not shown in Figure 1d, binding and unbinding rates are set to a value consistent with very weak binding, e.g $\Delta G_b > -3.0kcal/M$.

An RNA capable of efficient assembly

In order to identify an RNA sequence with high assembly efficiency, a genetic algorithm which searches the ensemble space of all RNA sequences of length 360 nucleotides is employed. Using a starting population of 2000 random RNAs which have been seeded with 12 hairpins that can bind CP according to the rules in the previous section, the RNA sequences are optimized with

respect to assembly efficiency (i.e. fast assembly) and yield of correctly assembled particles containing one packaged RNA. Each of the RNA sequences is subjected to assembly in the presence of 24000 CP pentamers, enough to package all 2000 RNAs into complete capsids. Assembly is stopped after 200 seconds has elapsed, which provides the selective pressure for fast assembly. This time is reduced in later rounds to further increase the selective pressure. After each of the 2000 RNAs has been tested for its assembly yield after 200 seconds (or smaller time), the top 25% of sequences which have the most virus capsids assembled are selected to move on to the next round and provide the sequence diversity for the subsequent generation. Single nucleotide mutations at random positions in the genome and recombination events between pairs of RNA sequences which swapped up to 30 nt between a pair of RNAs are used to construct a new population of 2000 RNAs. Despite the enormous search space of $4^{360} \approx 10^{216}$ different sequences, the genetic algorithm converges rapidly (within 30-40 generations) to sequences which can package $> 90\%$ of the 2000 RNAs into completed capsids in less then 200 seconds. This sequence optimization process was repeated 3 more times with different starting populations. Each of these additional optimization runs also converged in a similar number of generations, but to different sequences with similar yields (see supplementary Figure 2), suggesting that the sequence space has many equivalent solutions. Although the optimized sequences have different genotypes, they do have a distribution of PSs with similar affinities indicating that they have similar phenotypes. The fast convergence (within 30-40 generations) combined with the fact that different random starting populations can converge to a solution suggests that there

are a huge number of RNA sequences capable of efficient packaging in the sequence space.

Since ssRNA viral sequences must also have the ability to code for viral gene products as well as assemble, the sequences were checked for their ability to code for a gene product with a single AUG start codon followed by a series of codons coding for amino acids and terminating in a single stop codon (UAA, UAG, UGA). Figure 2 illustrates the assembly kinetics for one such RNA sequence which will be referred to as the "wild type" sequence. Figures 2a and 2b show the assembly kinetics for intermediates containing 2-11 coat proteins with the dashed line labelled capsid indicating the fully formed correctly assembled viruses. As can be seen from the figure, 95% of the RNAs are able to be packaged into capsids. Figure 2c shows the RNA sequence with the amino acid sequence of the gene product and PSs used during packaging below and above the nucleotide sequence, respectively. Packaging signals with square brackets indicate PSs that are used 50-90% of the time while those with round brackets are used over 90% of the time during assembly, while Figure 2d illustrates the PSs secondary structure (numbered 1-12, c.f. Figure 1b).

It should be noted that an alternative way of searching for a WT sequence with both high assembly efficiency and the ability to code for an amino acid sequence would be to first fix an amino acid sequence, then perform synonymous mutations on the RNA until a sequence with high assembly efficiency is identified. Although this method may actually be closer to the type of mutational pressures that a virus might be subject to since the structure and function of the viral protein must be preserved, it is quite

difficult to implement in practice since it is not clear that every amino acid sequence is able to be optimized for assembly under synonymous mutations. Furthermore, for a given amino acid sequence, it may be possible to make non-synonymous mutations to a few of the amino acids so that the sequence is then able to be optimized for assembly while the structure and function of the protein is preserved. These issues make a full study of such a search scenario difficult.

Exploration of the fitness landscape and effects of mutations on assembly

The assembly model developed here links an explicit nucleotide sequence to the number of capsids assembled, a measure of viral fitness, presenting an opportunity to explore the implicit fitness landscape which is formed as a result, as well as the effect of mutations. Although the size of the search space makes a complete exploration of the fitness landscape impossible, the local neighbourhood of the fitness landscape around the WT sequence can be explored. The local neighbourhood of sequences comprise of the set of sequences which are a mutational distance of 1 away from the WT sequence. Mutations can be either a single nucleotide insertion, deletion, or polymorphism. For the 360 nt WT sequence, there are a total of 2880 sequences in this local neighbourhood. Note that a subset of these 2880 sequences will be synonymous mutations which preserve the protein coding sequence. Each of these sequences were generated and then tested for their viral yield after 200 seconds of simulated assembly. Figure 3 shows the distribution of fitness yields, i.e. the percentage of viral capsids assembled, for the local

neighbourhood of mutated sequences. Interestingly, most of the sequences in the local neighbourhood (approximately 75%) have very little difference in assembly yield when compared to the WT sequence. However, a small fraction of the sequences in the local neighbourhood of the WT sequence (approximately 6%) have essentially no capsid assembled after 200 seconds. Similar behaviour is seen in another sequence solution where roughly 7% of mutations are deleterious (see supplementary Figure 3). Thus the mutations associated with these sequences represent the critical areas of the WT sequence which are very sensitive to mutation and would be expected to be highly conserved in a real viral sequence. The sensitivity of some areas of the genome to mutation leads to the natural question of whether these sites are associated with areas of the genome which are critical for forming the secondary structures of the PSs needed for interaction with the CP during assembly. The surprising answer is most but not all; 12 out of the 244 mutations which reduce the yield of capsid to $<10\%$ are outside of PSs and in the regions between them.

To further investigate this, mutant sequences in the local neighbourhood of the WT sequence were identified which ablate the ability of either PS 11 or PS 6 to bind to CP while preserving the WT protein coding sequence (i.e. synonymous mutations). Three synonymous mutations (G277U, G277A, and G277C) were identified which ablate WT PS 11 binding to CP. Figure 4a shows their effect on the secondary structure of PS 11 while Figures 4b-c illustrate their impact on virus assembly. Although the mutations disrupt binding of CP to PS 11, their effect on assembly is mostly small (G277A or G277C) or non-existent (G277U). This is because although the G277

mutation ablates PS 11 binding to CP, an alternative PS 11 (PS 11 mut in Figure 4a) can fulfil the role of the WT PS 11 and complete assembly. The small reduction in viral capsid yield from the G277A and G277C mutations is due to the formation of a hairpin competitor in these cases which is unable to bind CP. This results in a temporary kinetic trap where the partially formed capsid must wait for the hairpin competitor to melt and the PS 11 mutant to fold in its place.

In contrast to PS 11, ablating PS 6 binding to CP via the mutation G163A dramatically reduces the viral assembly yield from over 95% to less then 1.0%, roughly a 2 log reduction in yield (Figure 5). The G163A mutation is unable to produce a stem-loop which is capable of binding CP in the place of WT PS 6 and, as a result, assembly is permanently stalled in the capsid intermediate containing 5 CPs (c.f. Figure 5b and c). Secondary mutations which both restore assembly and preserve the ability of the RNA to code for protein can be identified by examining the local neighbourhood of sequences that are one mutation away from the G163A sequence and calculating their assembly fitness. There are 6 possible secondary mutations which restore assembly fitness (listed in Figure 5d). Interestingly, all 6 will result in a change to the primary protein sequence. Moreover 4 of the mutations (marked with a black dot in Figure 5d), when assessed individually in the absence of the G163A mutation, do not significantly alter the assembly yield with respect to the WT sequence. This suggests that several mutational pathways in the fitness landscape exist such that the G163A mutation can be incorporated into the RNA sequence without impacting on capsid yield. Figure 6 illustrates these pathways by showing a

simple two-dimensional slice of the fitness landscape. To illustrate the two-dimensional landscape, a pair of mutations are selected and their effect on the assembly yield are shown using bar graphs. Figure 6a shows the effects of the G163A and C167A mutations on the WT sequence in isolation and together. The WT sequence and its yield are shown in the bottom left corner of the square. Each edge of the square represents the presence of either the G163A or C167A mutation. The empty box at the top of the square represents the yield of capsid (less then 1%) after the G163A mutation is applied to the WT sequence. One can see that application of either the G163A or C167A mutations results in a reduction in viral yield, with the combined mutation restoring assembly yield to WT levels, suggesting an epistatic effect. In contrast, for the pair of mutations G163A and G166C shown in Figure 6b, the G166C mutation alone is able to maintain WT assembly yields. These results suggest that the fitness landscape which results from this simple RNA folding and assembly model is highly complex with areas of neutrality as well as complex epistatic effects.

In addition to monitoring assembly yield, the assembly model can also monitor the effect of mutations on the choice of the 12 PSs used by the RNA sequence during assembly. Figure 6c illustrates how the structure of PS 6 is altered to the mutant PS 6 under the two mutations C167A followed by G163A. After the first mutation, an alternative structure (PS 6 option 2) is created, allowing CP to choose between option 1 and option 2. However, the binding rules make PS 6 option 2 unable to bind to CP due to the larger bulge. This presents a competing structure which affects the yield of capsid. After the mutation G163A is applied, it ablates PS 6 option 1 and closes

the bulge of option 2, making PS 6 mutant which is CP binding competent and restoring capsid yield. In contrast, Figure 6d illustrates how PS 6 usage and structure evolves during the G166C and G163A mutations. Mutating G166C creates a second binding competent PS 6 (option 2). Both are able to bind equally well and the effect on assembly is minimal. Mutating G163A ablates PS 6 option 1 and strengthens the stem of option 2, maintaining capsid yield at WT levels. Figure 6d illustrates how a virus may undergo random drift in the protein coding space (since G166C alters the primary protein sequence) while capsid assembly efficiency is unaltered. However, it is clear from the mutational study of G163A that not all areas of the sequence space may be directly assessable without specific compensatory mutations elsewhere in the sequence.

Discussion

RNA viruses present an example of cooperative co-assembly where RNA-CP contacts, termed packaging signals, mediate packaging of the genomic RNA into the capsid container. Previously it has been demonstrated using a stochastic assembly model with fixed PSs that the position and affinity of the PSs in the linear genomic sequence enables the virus to efficiently assembly. Extension of this model to incorporate both genomic sequence information as well as RNA folding effects has allowed for further exploration of the roles that PSs may play during capsid assembly.

An additional feature of this assembly model is that specific RNA sequences can be mapped to a fitness value based on the number of infectious

virus particles that are completely assembled within a given time frame. Admittedly, viral fitness is an abstract concept since in a biological context "fitness" is likely a complex interplay between replication speed, replication efficiency, the ability to evade host immune responses, and success at producing infectious virions in the host cell. While only one aspect of viral fitness is examined here, number of infectious viral copies produced due to successful genome packaging, the resulting implicit fitness function does reveal some interesting features which may be relevant to virus evolution. Figure 7 illustrates the implicit fitness mapping which arises as a consequence of the assembly model. Given an RNA sequence, kinetic folding of the RNA using the folding rates allows the hairpin structures of the RNA to fold and melt in competition with each other. The kinetics of the folding process gives rise to a dominant phenotype for that RNA sequence which may contain small hairpins that are capable of binding to CP with varying affinity. In the presence of CPs that bind to specific RNA sequences and secondary structures, this phenotype can then be assigned a fitness value, i.e. the number of viral capsids that are successfully assembled, using the RNA virus assembly rules and CP-RNA binding rules. The genotype to phenotype to fitness mapping introduced here extends previous work which introduced an implicit phenotype to fitness mapping for virus assembly based on a simple 12-dimensional space of possible RNA phenotypes, with each dimension representing the affinity of a packaging signal and fitness measured by viral yield (31). Additionally, genotype to phenotype mappings have also been used previously in evolutionary models of RNA folding (32). In this example, Schuster and Fontana use the Turner energy rules to fold various RNA sequences and

assign the sequence a phenotype based on the sequences ability to form into a tRNA structure. The resulting implicit mapping links a sequence to a biologically relevant phenotype, in this case the tRNA structure. Here, the implicit fitness function performs both mappings simultaneously, linking a genotype directly to a fitness parameter relevant to ssRNA viruses, i.e. the ability to package a viral genome into the capsid container efficiently.

The recent interest in the experimental identification of packaging signals in a number of RNA viruses using SELEX (4, 5) has led to a number of experiments which attempt to ablate these putative PS sites in order to prove function. The ability of the assembly model to map sequences to biologically relevant fitness measures allows for mutational effects on viral capsid assembly to be explored and hence theoretically predict the effects of mutations which ablate PSs on assembly fitness. Paradoxically, the assembly models show that mutating the RNA sequence such that the binding of a PS to CP is ablated may not have any effect at all on assembly yield, leading an experimentalist to conclude (incorrectly) that this PS is not truly a packaging signal and has no biological function. However, the model shows how a PS could have a proper biological function (promotion of assembly) but have no effect on assembly if it was mutated in such a way that it cannot bind to CP. This is due to alternative options for PSs existing at some sites within the genome and ablating the main PS causes an alternative PS structure to be utilized in its place. This highlights the potential difficulties in designing PS knockout experiments via reverse genetics and may explain why recent PS knockout experiments in Human Parechovirus (4) were "hit and miss" where some PS knockouts resulted in essentially no viral titre while others

resulted in viral titres that were similar to WT. The model here suggests that this could be due to some PSs being sensitive to mutational change (e.g. PS 6 in the simulation above) while others are more mutationally robust since they have mutually exclusive PSs to replace the ablated one (e.g. PS 11).

It is the hope that the model developed here will enable more sophisticated theoretical modelling of RNA virus assembly as well as lead to the incorporation of additional features of the viral life cycle. Such models will present opportunities to develop implicit fitness landscapes that are biologically relevant and allow for the exploration of the evolutionary landscape of RNA viruses. Although the framework developed here has a very simplified RNA folding model with room for improvement, the model discussed here has provided one example of how an implicit genotype to phenotype to fitness mapping can be made for ssRNA viruses.

Author Contributions

ECD designed the research, performed the research, contributed analytic tools, analyzed data and wrote the paper.

Acknowledgments

ECD would like to thank the Leverhulme Trust for support via an Early Career Fellowship (ECF-2013-019) and for constructive conversations with Prof. Reidun Twarock, Dr. Richard Bingham, and Dr. German Leonov. The assembly code is available for download from GitHub at:

https://github.com/edykeman/RNAassembly.

References

1. Rao, A., 2006. Genome packaging by spherical plant RNA viruses. *Annu. Rev. Phytopathol.* 44:61–87.

2. Sun, S., V. B. Rao, and M. G. Rossmann, 2010. Genome packaging in viruses. *Current opinion in structural biology* 20:114–120.

3. Hutchinson, E. C., J. C. von Kirchbach, J. R. Gog, and P. Digard, 2010. Genome packaging in influenza A virus. *Journal of general virology* 91:313–328.

4. Shakeel, S., E. C. Dykeman, S. White, A. Ora, J. J. Cockburn, S. Butcher, P. Stockley, and R. Twarock, 2017. Genomic RNA folding mediates assembly of human parechovirus. *Nature communications* In Press.

5. Bunka, D. H., S. W. Lane, C. L. Lane, E. C. Dykeman, R. J. Ford, A. M. Barker, R. Twarock, S. E. Phillips, and P. G. Stockley, 2011. Degenerate RNA packaging signals in the genome of satellite tobacco necrosis virus: implications for the assembly of a T= 1 capsid. *Journal of molecular biology* 413:51–65.

6. Dykeman, E. C., P. G. Stockley, and R. Twarock, 2013. Packaging signals in two single-stranded RNA viruses imply a conserved assembly

mechanism and geometry of the packaged genome. *Journal of molecular biology* 425:3235–3249.

7. Patel, N., E. C. Dykeman, R. H. Coutts, G. P. Lomonossoff, D. J. Rowlands, S. E. Phillips, N. Ranson, R. Twarock, R. Tuma, and P. G. Stockley, 2015. Revealing the density of encoded functions in a viral RNA. *Proceedings of the National Academy of Sciences* 112:2227–2232.

8. Dykeman, E. C., P. G. Stockley, and R. Twarock, 2013. Building a viral capsid in the presence of genomic RNA. *Physical Review E* 87:022717.

9. Dykeman, E. C., P. G. Stockley, and R. Twarock, 2014. Solving a Levinthal's paradox for virus assembly identifies a unique antiviral strategy. *Proceedings of the National Academy of Sciences* 111:5361–5366.

10. Perlmutter, J. D., and M. F. Hagan, 2015. The role of packaging sites in efficient and specific virus assembly. *Journal of molecular biology* 427:2451–2467.

11. Hagan, M. F., and R. Zandi, 2016. Recent advances in coarse-grained modeling of virus assembly. *Current opinion in virology* 18:36–43.

12. Smith, G. R., L. Xie, and R. Schwartz, 2016. Modeling Effects of RNA on Capsid Assembly Pathways via Coarse-Grained Stochastic Simulation. *PloS one* 11:e0156547.

13. Peabody, D. S., 1990. Translational repression by bacteriophage MS2 coat protein expressed from a plasmid. A system for genetic analysis of

a protein-RNA interaction. *Journal of Biological Chemistry* 265:5684–5689.

14. Simon, A. E., 2015. 3 UTRs of carmoviruses. *Virus research* 206:27–36.

15. Flamm, C., I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger, 2002. Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie International journal of research in physical chemistry and chemical physics* 216:155.

16. Mathews, D. H., J. Sabina, M. Zuker, and D. H. Turner, 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology* 288:911–940.

17. Stadler, P. F., 2002. Fitness landscapes. *In* Biological evolution and statistical physics, Springer, 183–204.

18. Yu, Z., M. J. Dobro, C. L. Woodward, A. Levandovsky, C. M. Danielson, V. Sandrin, J. Shi, C. Aiken, R. Zandi, T. J. Hope, et al., 2013. Unclosed HIV-1 capsids suggest a curled sheet model of assembly. *Journal of molecular biology* 425:112–123.

19. Zlotnick, A., 1994. To build a virus capsid: an equilibrium model of the self assembly of polyhedral protein complexes. *Journal of molecular biology* 241:59–67.

20. Anderson, D. F., 2007. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *The Journal of chemical physics* 127:214107.

21. Wuchty, S., W. Fontana, I. L. Hofacker, P. Schuster, et al., 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49:145–165.

22. Gillespie, D. T., 1977. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* 81:2340–2361.

23. Huang, J., and B. Voß, 2014. Analysing RNA-kinetics based on folding space abstraction. *BMC bioinformatics* 15:60.

24. Waterman, M. S., 1995. Introduction to computational biology: maps, sequences and genomes. CRC Press.

25. Dykeman, E. C., 2015. An implementation of the Gillespie algorithm for RNA kinetics with logarithmic time update. *Nucleic acids research* 43:5708–5715.

26. Zhang, W., and S.-J. Chen, 2006. Exploring the complex folding kinetics of RNA hairpins: I. General folding kinetics analysis. *Biophysical journal* 90:765–777.

27. Proctor, D. J., H. Ma, E. Kierzek, R. Kierzek, M. Gruebele, and P. C. Bevilacqua, 2004. Folding thermodynamics and kinetics of YNMG RNA hairpins: specific incorporation of 8-bromoguanosine leads to stabilization by enhancement of the folding rate. *Biochemistry* 43:14004–14014.

28. Lago, H., A. M. Parrott, T. Moss, N. J. Stonehouse, and P. G. Stockley, 2001. Probing the kinetics of formation of the bacteriophage MS2 trans-

lational operator complex: identification of a protein conformer unable to bind RNA. *Journal of molecular biology* 305:1131–1144.

29. Beckett, D., H.-N. Wu, and O. C. Uhlenbeck, 1988. Roles of operator and non-operator RNA sequences in bacteriophage R17 capsid assembly. *Journal of molecular biology* 204:939–947.

30. Dykeman, E. C., P. G. Stockley, and R. Twarock, 2013. Packaging signals in two single-stranded RNA viruses imply a conserved assembly mechanism and geometry of the packaged genome. *Journal of molecular biology* 425:3235–3249.

31. Bingham, R., E. C. Dykeman, P. G. Stockley, and R. Twarock, 2017. A Quasi-species theory based model of RNA virus evolution reveals a drug target with a high barrier to resistance. *Proceedings of the National Academy of Sciences* Submitted.

32. Fontana, W., and P. Schuster, 1998. Continuity in evolution: on the nature of transitions. *Science* 280:1451–1455.

Table 1: Effects of mutations on coat protein RNA binding. NT labels the nucleotide number in Figure 1d which, when altered according to the table, changes the dissociation constant K_D by the factor listed. Mutations to multiple nucleotide positions affect K_D in a multiplicative manner. Estimates for the dissociation constant are based on experimental data for bacteriophage MS2 (see ref. (29) and (30)).

NT	A	G	U	C
1	1	2	2	2
2	100	100	1	0.18
3	1	1000	1000	1000

Figure Legends

Figure 1.

A stochastic assembly model for ssRNA viruses with RNA folding reactions. **(a)** In the packaging signal mediated model of viral assembly, viral coat proteins (pentamer shapes) co-assembly with viral ssRNA to form a viral capsid containing the viral RNA genome. The coat proteins interact via RNA secondary structures (here RNA hairpins) present in the RNA genome. **(b)** Two types of reactions are used to model capsid assembly, an RNA/CP reaction where CP can bind at rate κ_b or unbind at rate κ_u to any hairpin present in the RNA genome and a CP/CP reaction where two coat proteins which neighbour on the RNA strand associate or dissociate from each other with rates κ_a and κ_d, respectively. Binding rates for the RNA/CP reactions vary depending on the sequence and secondary structure features of each RNA hairpin. **(c)** During assembly of the capsid, CP-free RNA hairpins can melt and ssRNA areas can fold altering the local secondary structure of the RNA. Rates of folding and melting (κ_f and κ_m) are sequence and structure dependant and are estimated using the BARRIERS methodology and the Turner 99 rules. **(d)** Structures of hairpins with high affinity for MS2 CP ($K_D = 1$nM) which represent the three structures which can bind to CP in the model. The three critical sequence elements effecting the binding of RNA to CP (labeled 1 to 3) where probed experimentally by Ulenbeck et al. (29). Table 1 lists the effects of mutations at these sites on the binding affinity. Multiple mutations are modelled as being multiplicative in their effect on the binding affinity.

Figure 2.

A genotype with high fitness and protein coding capacity evolved from a random population of RNAs. Assembly kinetics for 2000 copies of the WT sequence for RNA/Capsid intermediates containing **(a)** two to six CPs and **(b)** seven to twelve CPs. **(c)** Nucleotide sequence of the identified high fitness RNA. Packaging signals (PSs) that are used during assembly more then 90 percent of the time are shown above the RNA sequence with parentheses while PSs that are used less then 90 precent of the time are shown with square brackets. The protein sequence is shown below the RNA sequence with the stop codon labelled by a star. **(d)** Secondary structures and nucleotide positions of the PSs interacting with CP during capsid assembly in the majority of capsids.

Figure 3.

Effect of single nucleotide mutations (insertion, deletion, or polymorphism) on viral assembly fitness of the wild type sequence. The majority of mutations ($> 75\%$) result in at least 85% of the RNAs being packaged into complete capsids within a 200 second time frame. A small minority of mutations ($\approx 6\%$) are deleterious and result in essentially no capsid being assembled.

Figure 4.

Effects of G277N, a synonymous mutation ablating PS11, on viral assembly fitness. **(a)** The valine (GUG) codon in the WT RNA sequence

(boxed) is mutated to the three alternative codons (GUC,GUU,GUA) which alters the secondary structure of PS11 and ablates its ability to bind to CP according to the RNA-CP binding rules used in the assembly model. The result of all three mutations is the creation of a mutant PS11 capable of binding CP. G277A and G277C also create a stable hairpin unable to bind CP which competes with PS10 and PS11. **(b)** Assembly kinetics of the G277U mutant showing essentially unaltered assembly compared with WT. **(c)** Assembly kinetics for G277A mutant showing a temporary kinetic trap formed at RNAs bound to nine CPs. The trap is formed due to a delayed folding of PS10 and PS11 from the hairpin competitor shown in (a). **(d)** Assembly kinetics for G277C mutant. A less severe temporary trap is formed due to the folding of the non-CP binding RNA competitor shown in (a).

Figure 5.

Effects of G163A, a synonymous mutation ablating PS6, on viral assembly fitness. **(a)** Mutation of the arginine (AGG) in the WT RNA sequence (underlined in the sequence and boxed in the secondary structure) to the alternative codon AGA results in a mutant PS6 which is unable to bind to CP according to the binding rules of the assembly model. Assembly kinetics of RNAs containing **(b)** two to six CPs show that the majority of RNAs are permanently trapped in intermediates containing five CPs in complex with RNA. **(c)** Only six single nucleotide mutations are able to restore assembly fitness and keep protein coding capacity. Each of these mutations require an amino acid change in the protein sequence. Four of these mutations (labelled with a black dot) do not alter viral assembly fitness

with respect to the WT sequence (i.e. in absence of the G163A mutation).

Figure 6.

Two dimensional assembly fitness landscapes and mutational pathways. (a) The assembly fitness of WT, the two single mutants G163A and C167A, and the double mutant containing both mutations are arranged in a two dimensional fitness plot. Assembly fitness (in terms of percentage of capsid assembled) are illustrated as shaded boxes. Edges in the two dimensional landscape indicate either the G163A or C167A mutation to the genome, with the corner opposite to WT containing the double mutation. **(b)** Same as in (a) but for the mutations G163A and G166C. **(c)** The mutation C167A (lower case a) alters the WT PS6 and creates an alternative PS6 (option 2). Option 2 is unable to bind to CP under the binding rules due to a larger bulge and thus acts as a competitor to the option 1 PS6 and results in a kinetic trap and reduced viral capsid yield as shown in (a). The mutation G163A alters PS6 option 2 to become binding competent, ablates PS6 option 1 and restores assembly efficiency. **(d)** The mutation G166C (lower case c) alters WT PS6 and also creates an alternative, binding competent PS6 (option 2) which results in the RNA retaining assembly efficiency. The subsequent mutation G163A ablates PS6 option 1 and further stabilizes option 2.

Figure 7.

An implicit genotype to phenotype to fitness mapping based on a ssRNA viral assembly model. A specific genotype is mapped to an

RNA secondary structure phenotype via RNA folding and melting reactions in the model. The kinetics and competition between different hairpins in the sequence determines the ability of the RNA sequence to present secondary structures capable of promoting CP binding and assembly. The CP-RNA binding rules and assembly reactions allow the phenotype to be mapped to a value of assembly yield, i.e. the number of correctly assembled capsids within a specific time, which gives a measure of viral fitness.