Frustration and Fidelity in Influenza Genome Packaging

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Nida Farheen

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

May, 2019

Supervisor: Dr. Mukund Thattai

Certificate

This is to certify that this dissertation entitled Frustration and Fidelity in Influenza Genome Packaging towards the partial fulfillment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Nida Farheen at National Centre for Biological Sciences under the supervision of Dr. Mukund Thattai, National Centre for Biological Sciences during the academic year 2018-2019.

Dr. Mukund Thattai

Committee:

Dr. Mukund Thattai

Dr. M. S. Madhusudhan

This thesis is dedicated to the four years at IISER

Declaration

I hereby declare that the matter embodied in the report entitled Frustration and Fidelity in Influenza Genome Packaging, are the results of the work carried out by me at the National Centre for Biological Sciences, Bangalore under the supervision of Dr. Mukund Thattai, and the same has not been submitted elsewhere for any other degree.

Nida Farheen

Nida Farheen

Acknowledgments

I am immensely thankful to Dr. Mukund Thattai for giving me an opportunity to work on this project and for being a wonderful mentor. It was great fun to learn and discuss science with him. I would also like to acknowledge IISER Pune, INSPIRE-SHE and NCBS for funding and support.

I want to thank my TAC member Dr. M. S Madhusudhan for reviewing this work. I would be always grateful to Madhu for shaping my scientific abilities and interests during my academic trajectory at IISER Pune. I have been fortunate to have mentors like Mukund and Madhu who have a contagious excitement for science.

I would like to thank my lab members Sachit, Pragati, Ashish, Rahul, Anshuman, and Sachin and friends Amit, Alkesh, Dilawar, Ashwini, Saptarishi, Abhishek, and Alex without whom I would not have enjoyed my everyday life at NCBS. I wish to express gratitude to my friends from IISER- Vrushali, Rahul, Poornima, Sanyukta, Sravya, Kaushik, Kaustubh, and Neeladri who comforted me from miles away.

Finally, I would like to thank my family for their support.

Abstract

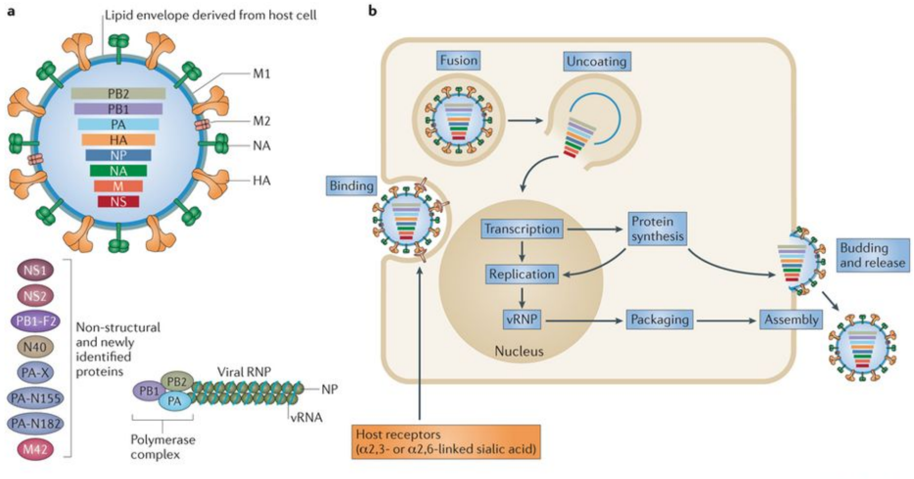
Influenza genome is organized as eight distinct RNA segments, each coding for proteins essential for the virus life cycle. Post-infection, segments are replicated in the host's nucleus and packaged into the budding progenies. Previous experimental results show that most viral progenies contain the complete genome, i.e., one copy of each of the eight segments. It is unclear how the virus efficiently assembles its genome from a pool of replicated segments. There is strong evidence suggesting that segments form specific RNA-RNA interactions and these inter-segment interactions lead to the genome assembly. However, the precise interaction network remains unresolved. Here, we investigated the nature of the interaction network by asking which network topologies would be most efficient in assembling the genome. It was shown that out of many possible network topologies, only a few of them would guarantee genome packaging. Two hypothetical models were constructed to predict the topologies that are most likely to evolve. Furthermore, the segment interaction network for Influenza virus was inferred from three published experimental datasets. This study makes testable predictions on the interactions that underlie the Influenza genome assembly, with the hope that it would provide insights into the mechanism of genome packaging and viral evolution.

Contents

Introduction

Influenza is one of the major epidemic diseases caused by virus, affecting more than three million people every year. The genome of Influenza virus is organized as eight disjointed segments composed of single-stranded RNA (vRNAs). These eight segments code for eleven proteins namely hemagglutinin (HA), neuraminidase (NA), matrix 1 (M1), matrix 2 (M2), nucleoprotein (NP), non-structural protein 1 (NSP1), non- structural protein 2 (NS2), polymerase acidic protein (PA), polymerase basic protein 1 (PB1), polymerase basic protein 2 (PB2) and polymerase basic protein 1- F2 (PB1-F2) ([1]). vRNAs are generally present in a complex with NP and polymerases, which together is called as viral ribonucleoproteins (vRNPs).The segmented form of the genome provides evolutionary advantage for the virus to shuffle its segments with other viral strains. This process, known as viral recombination/reassortment, has previously led to the emergence of novel Influenza viruses such as H1N1 (Spanish flu, 1918), H2N2 (Asian flu, 1957) and more recently H1N1 (Swine flu 2009) [2].

While on the one hand, recombination is a crucial force in driving viral evolution, the segmented nature also complicates the process of genome assembly inside the host cell. Influenza virus gains an entry in the cell through endocytosis, which is triggered upon the interaction of HA with sialic acid present on plasma membrane. After entering the cell, the viral membrane fuses with endosomal membrane, releasing its genome into the cytoplasm. The vRNPs are trafficked to the nucleus, wherein genome replication and transcription take places. A quantitative study on viral replication dynamics showed that vRNA level increases up-to 10000 molecules per segment within the first four hours of infection before leveling out [3]. The replicated vRNPs are exported out of the nucleus and actively transported to plasma membrane. It is believed that genome assembly, i.e., bringing together eight distinct segments, takes place on the route to plasma membrane. The production of an infectious viral progeny depends not only on the genome assembly but also on the localization of viral proteins at the plasma membrane [4]. Viral progenies start to bud off from the plasma membrane four hours post infection, at the rate of 1000 virions per hour [3]. The Influenza virus's life cycle is summarized in Figure 1.

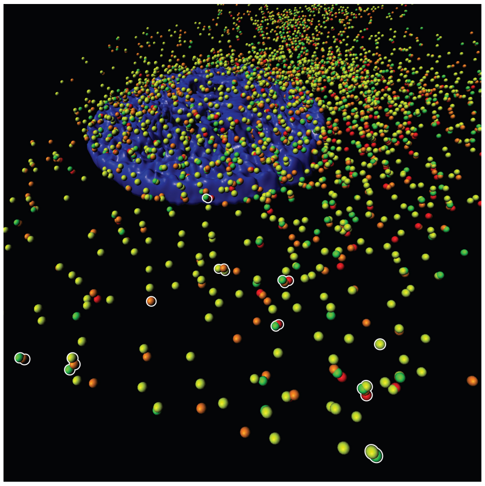


Nature Reviews | Microbiology

Figure 1: (a) Eight genomic segments of Influenza are labeled as PB2 (1), PB1 (2), PA (3), HA (4), NP (5), NA (6), M (7) and NS (8); numeric labeling denoted in brackers. Segments are present in complex with nucleoproteins (NP) and polymerases. (b) The life cycle of virus from entry into the host to release of progeny (Figure adapted from [5])

From a fitness point of view, the goal of the virus is to maximize production of infectious progenies. Since each segment codes for an essential protein, packaging of the complete genome is a necessary condition for generating infectious viral particles [4]. Interestingly, Influenza viruses show remarkable ability in assembling the eight segments. Electron tomography of the released viral particles shows that at least 80% of the virions contain the full genome [6, 7, 8]. This result is further supported by fluorescent in-situ hybridization study of vRNPs at single virus resolution which demonstrated that precisely one copy of each distinct segment is packaged in the virions [9].

How the virus manages to package its genome with such a high efficiency is an unresolved question. A 'selective packaging model' has been proposed in the literature. According to this model, vRNPs bind to each other through specific RNA-RNA interactions, leading to the assembly of all segments. This model has now gained several lines of evidence. Firstly, electromobility shift assay on co-incubated vRNAs show that nine vRNA pairs can interact in-vitro [10, 8, 11]. Secondly, terminal regions of vRNAs are conserved and mutations in these regions are associated with decreased genome packaging [12, 13, 14, 15, 16]. Packaging signals have been mapped on the genome using inferences from mutational analysis [17]. Thirdly, base-paired inter-segment interactions have been identified using cross-linking techniques and predicted to be energetically favourable [18]. Secondary structures of vRNA 7 and vRNA 8 suggest an interaction between segment 4 and 7, and between segment 2 and 8 [19, 20]. Recent studies also indicate towards the role of nucleoproteins in mediating genome packaging [21, 22, 23]. Electron tomography of genome inside the virions shows that segments are interconnected on a platform like surface [8]. Further, Figure 2 shows segment co-localization inside the cytoplasm, thereby capturing the assembly process in real time[24]. Taken together, these results provide convincing evidence in support of selective packaging model [25, 26, 27].



Lakdawala SS, et al. 2016.
Annu. Rev. Virol. 3:411–27

Figure 2: Co-localization of segments observed in the cytoplasm through four-color fluorescence in-situ hybridization. Segments are labelled as yellow, green, orange and red. (Figure adapted from [24])

Despite the molecular information gathered about the packaging signals, the precise segment interaction network remains unknown. Knowledge of this interaction network would not only extend our understanding of genome assembly but would also provide insights into the mechanisms of segment re-assortment and broaden the vaccine strategies, which currently, require annual renewal. In this study, we have attempted to delineate the segment interaction network underlying genome packaging from three independent directions. Firstly, we studied the dependence of network topology on its efficiency to assemble the genome. Out of many possible topologies, it was found that only a subset of them have maximal packaging efficiency. We then constructed two models to investigate which topologies are more likely to evolve. Finally, the interaction network was inferred using three different experimental data-sets and was compared to the predictions.

Chapter 1

Definitions

i) **vRNP**: Single stranded viral RNA in complex with nucleoproteins and polymerases. Note that the term 'vRNP' is used interchangeably with 'segment'
ii) **vRNA**: Single stranded viral RNA
iii) **Genome Assembly**: The process of bringing together eight distinct vRNPs of Influenza.
iv) **Interaction Network**: Set of interactions between segments
v) **Connected Network**: In a connected network, each segment interacts with at-least one other segment.
vi) **Disconnected Network**: In a disconnected network. there is atleast one segment which does not interact with any of the seven segments
vii) **Tree Topology**: Connected network with only seven interactions (for eight nodes)/Connected network with only two interactions (for three nodes).
viii) **Cycle Topology**: Connected network with number of interactions ranging from eight to twenty-eight (for eight nodes)/Connected network with three interactions (for three nodes).
ix) **Cluster**: Refers to a group of segments bound together through interactions
x) **Assembly Reaction Efficiency**: The proportion of completely assembled clusters out of total clusters formed at steady-state
xi) **Virion**: Viral progenies released from the host cell
xii) **Nearest nighbour segments**: For a given segment, the segment present at its right, left and center inside the virion are its nearest neighbours.

Chapter 2

Assembly Efficiency of Networks

The goal of genome assembly is to bundle the newly synthesized vRNPs into clusters (Definitions ix), such that each cluster contains only one copy of each of the eight distinct segments. One can imagine that a minimum of seven interactions between eight vRNPs is necessary to assemble the genome. Since there are only twenty-eight vRNPs pairs ($\binom{8}{2}$) and each vRNP pair can either interact or not, the total number of possible interaction networks (Definitions iv) is equal to 2^{28}. In this chapter, we explore if the 2^{28} networks differ in their efficiency of genome packaging. Efficiency is defined as the number of completely assembled clusters out of the total clusters formed at steady state. Specifically, we ask how many interactions are required to package the genome efficiently. All networks can be classified into two topologies: tree or cycle (Definitions vii, viii & Figure 2.1) depending on the number of interactions. In the figures below, segments are represented as nodes and segment interactions are drawn as edges between the nodes.

The results of this chapter are based on two assumptions. First, all interactions of a given network are strong and irreversible in cellular timescale. That is not to deny the possibility of weak reversible interactions. However, we are only considering strong and stable interactions which would be necessary for a successful assembly. The alternative hypothesis of having transient-cooperative interactions is discussed in the Conclusions. This assumption is supported by an in-vivo experiment in which the formation of a stable assembly of three vRNPs was observed [28]. Second, since released viral particles contain only one copy of the eight distinct segments, it was assumed that each vRNP could bind to only one copy per interaction partner. Had this not been the case, the cluster would continue to bind and grow beyond eight segments.

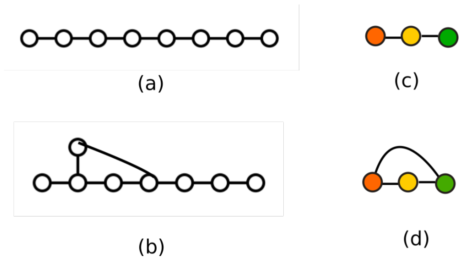


Figure 2.1: Examples of networks with tree and cycle topology - (a) tree with eight nodes (b) cycle with eight nodes containing a loop between 4 nodes (c) Minimal tree network (d) Minimal cycle network.

2.1 Efficiency of Cycle Networks

The loop size, i.e the number of segments that are part of the loop in cyclic networks (of n nodes) can range anywhere between n+1 to $\binom{n}{2}$. Here, we computed the assembly efficiency (η) for a loop size of 3 or the minimal cycle network (Figure 2.1 (d)). In this network, three distinct monomers A, B, and C interact with each other to assemble a cluster ABC. The interacting bonds can take two possible conformations- bonds pointing outwards shown in Figure 2.2 (a) and bonds pointing inwards shown in Figure 2.2 (b). The bond direction will dictate the accessibility for further interactions.

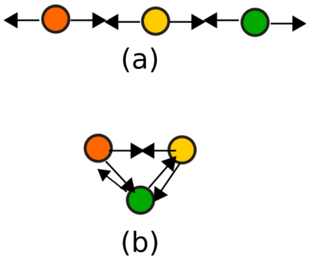


Figure 2.2: Possible orientation of interacting bonds: (a) Bonds point outwards, allowing for additional interactions (b) Bonds are oriented inwards, leading to steric hindrance for further interactions.

Case1: No Steric Hindrance in Interactions

In cyclic networks, each segment interacts with at least two other segments. For this particular example of minimal cycle network, if after the formation of a three-segment cluster, the interaction sites remain available for more segments to come and bind, then the cluster would continue to polymerize (Figure 2.3). Polymerization would trap free monomeric segments and generate undesirable clusters, thereby not allowing the reaction to reach its maximum efficiency.

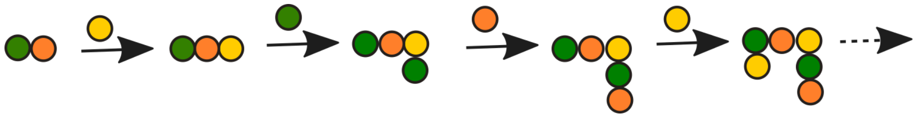


Figure 2.3: In the case of no steric hindrance, segments would continue to interact and polymerize beyond the desired three node cluster.

Case2: Steric Hindrance in Interactions

If we impose steric hindrance and exclude the possibility of polymerization (Figure 2.2 (b)), the mass conservation equations and simulations show that the reaction efficiency would be less than one at steady state (Methods I). Figure 2.4 depicts the assembly simulation with time and shows that at steady state, segments are trapped in incomplete dimer assemblies (here AB, BC, and AC), which are left with no free A, B and C to form the desired trimer (ABC). Note that ABC/BCA/ACB... are all equivalent. The state in which dimers are unable to react further and proceed to assembly completion is referred here as the 'frustrated state'.

The equations at best point towards a condition for which system would not reach maximum efficiency. To further quantify, five hundred stochastic simulations of assembly reaction following minimal cycle network was carried out (Methods II) and the average η was found to be 53% (Figure 2.5). Similar simulations were performed to study the dependence of efficiency on the number and size of loops. Figure 2.6 shows that the efficiency decreases on increasing either the number of loops or the loop size. Interestingly, the decrease in efficiency is much more on increasing loops, with η changing from 60% to less than 20%, as compared to change on increasing loop size. While we investigated the frustration in genome packaging independently, the same effect has also been recently shown in a general study of self-assembly [29].

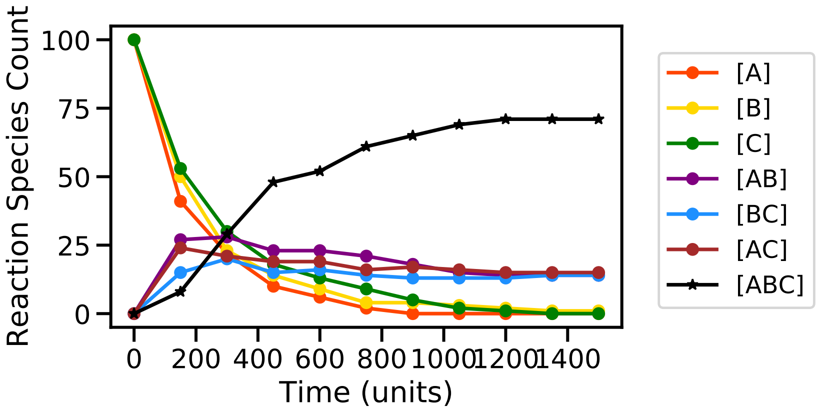


Figure 2.4: Evolution of a system undergoing assembly reaction between three monomers A, B, and C which interact according to a minimal cycle network. At time = 0, the system contains 100 copies of A, B, and C. At steady state, monomers reach zero and dimers stay non-zero.

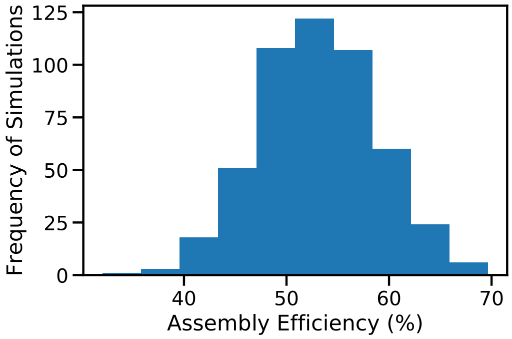


Figure 2.5: Histogram of reaction efficiency obtained for 500 stochastic simulations of assembly reactions following minimal cycle network.

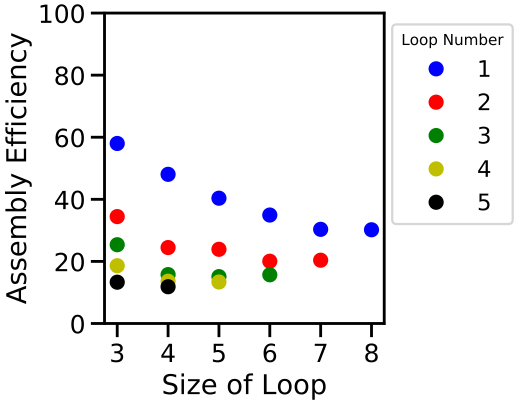


Figure 2.6: Dependence of assembly reaction efficiency (average obtained over 500 simulations) on the number & size of loops in interaction network.

2.2 Efficiency of Tree Networks

For an assembly reaction in which segments interact according to a tree network, proofs and simulation show that efficiency would always reach one at steady state (Method I and Figure 2.7). Unlike cycles, this reaction would not undergo frustration or polymerization.

We, therefore, believe that the segment interaction network in Influenza would be a tree,

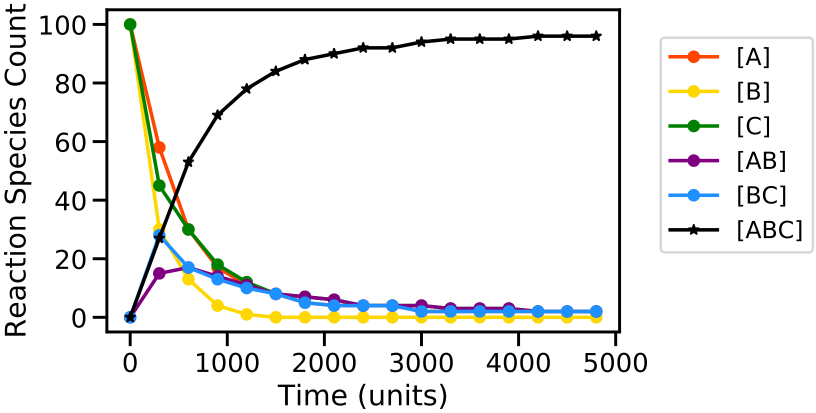


Figure 2.7: Evolution of a system undergoing assembly reaction between three monomers A, B, and C which interact according to a minimal linear network (Figure 2.1 (c)). At time = 0, the system contains 100 copies of A, B and C. At steady state, all monomers and dimers have interacted to form the product ABC.

to maximize its chances of genome packaging. For eight segments, trees can further be classified into twenty-three distinct topologies (Figure 2.9). We next investigated if the rate of assembly differs between these topologies. Stochastic simulations do not indicate any significant difference between the trees in the time taken for assembly Figure 2.8).

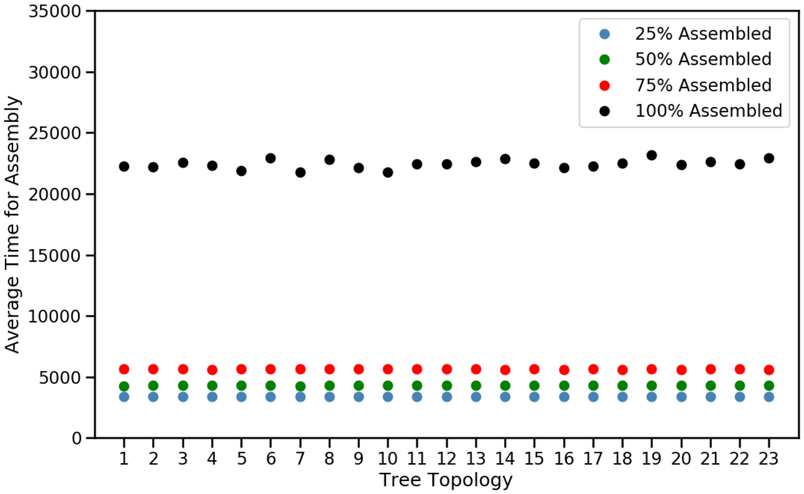


Figure 2.8: Time taken to assemble 25 %, 50%, 75% and 100% of the genome with interaction network of different tree topologies

To conclude, in this chapter, we discussed the effect of interaction network topology on the efficiency of assembly reactions. It was shown that reactions following a cycle network can get struck in frustration or polymerization and as a consequence of that, would not reach maximum efficiency. On the other hand, a network of tree topology is guaranteed to assemble the genome, given the system to allowed to reach steady state. Further, no significant difference in the rate of the assembly was observed between the twenty-three tree topologies.

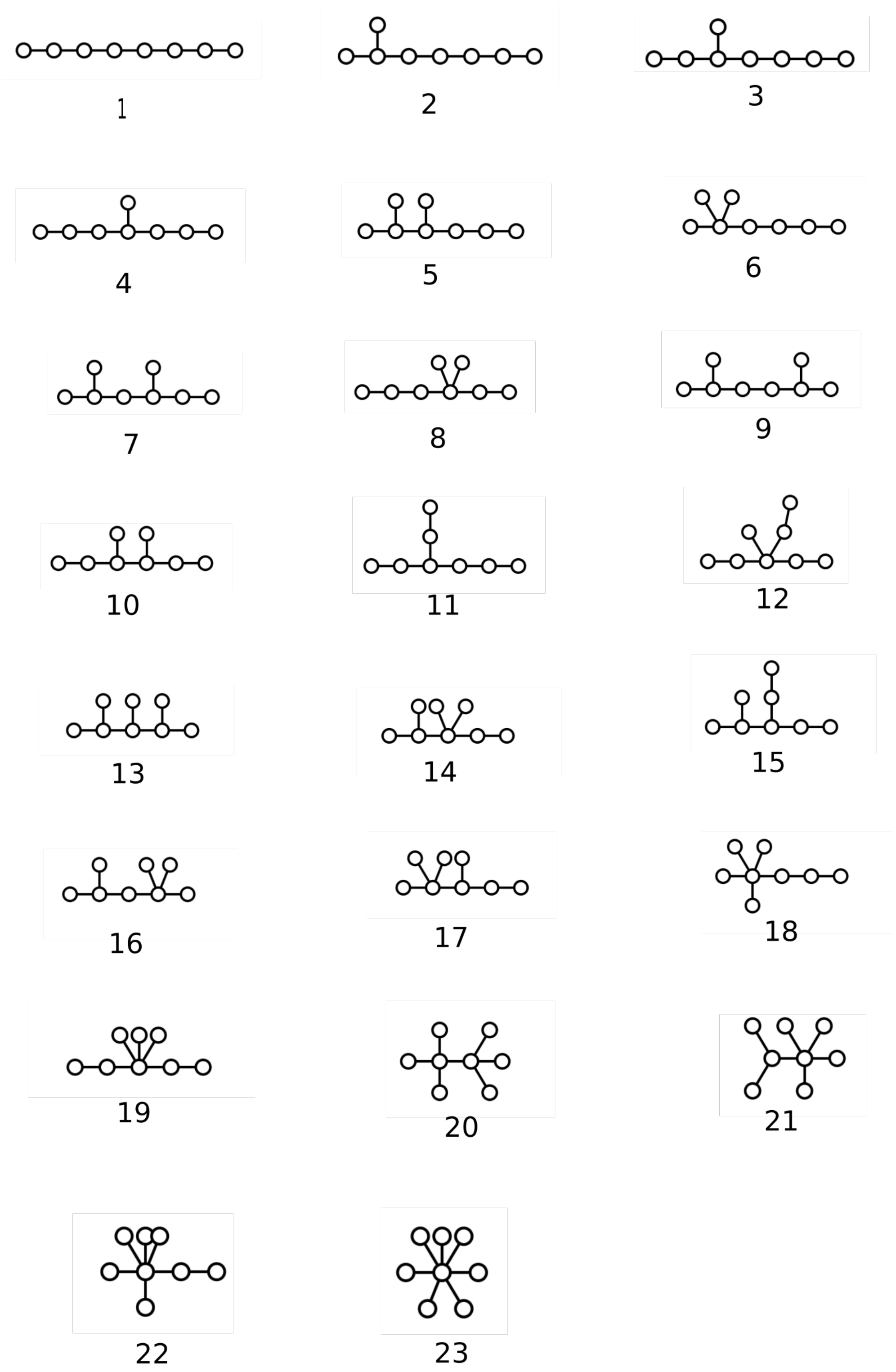


Figure 2.9: Twenty-three tree topologies for eight nodes. The tree labels are used for referencing the tree topologies in this report.

Chapter 3

Evolution of Trees

In the previous chapter, we concluded that any tree network would have 100% efficiency in assembling the genome. Within the broad category of trees, for eight segments, there are twenty-three distinct topologies. Stocastic simulations showed that there is no significant difference in the rate of assembly between reactions following different tree topologies. This result leaves the following question open: Given that all trees are equally competent to assemble the genome, why would one tree topology evolve over the other? Here, we address this question by constructing two hypothetical models on the evolution of trees - Gain Model and Gain-Loss Model.

3.1 Gain Model

In this model, we hypothesize that, at the very beginning, the viral genome had only one segment. Over the evolutionary time, the virus gained foreign genomic segments, which could form interactions with the existing ones. Therefore, the evolution of segmented genome and the corresponding interaction network started from a single segment and the rest of the seven segments were added subsequently to the growing network. Depending on the order in which segments were added, one can back-trace different paths through which a given tree could have evolved. The main idea, here, is that topologies that can be constructed in maximum pathways are most likely to evolve.

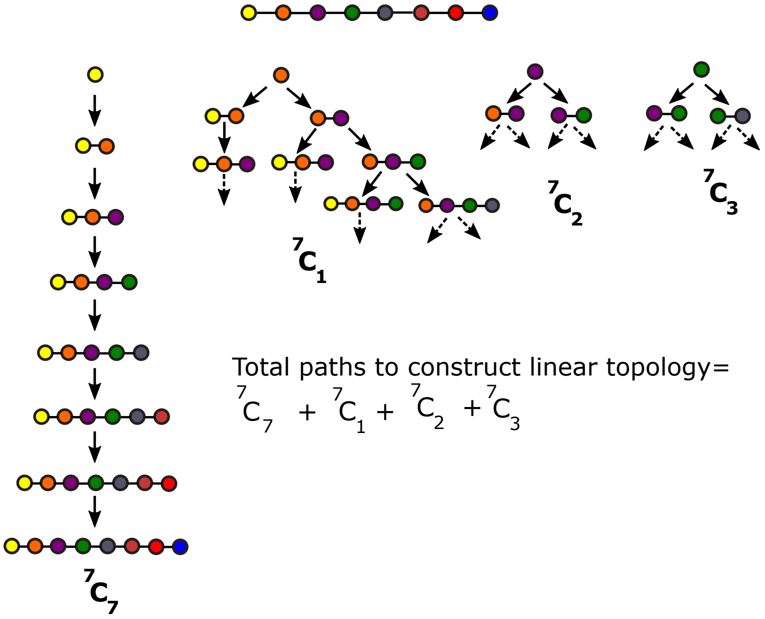

$^7C_1$
$^7C_2$
$^7C_3$
Total paths to construct linear topology=
$^7C_7 + {}^7C_1 + {}^7C_2 + {}^7C_3$
$^7C_7$

Figure 3.1: Example of evolution of a linear topology through gain model. The segments marked * were considered as the most anscestral segments. At each step, a new segment was added and the choice of that segment depended on the existing segments in network.

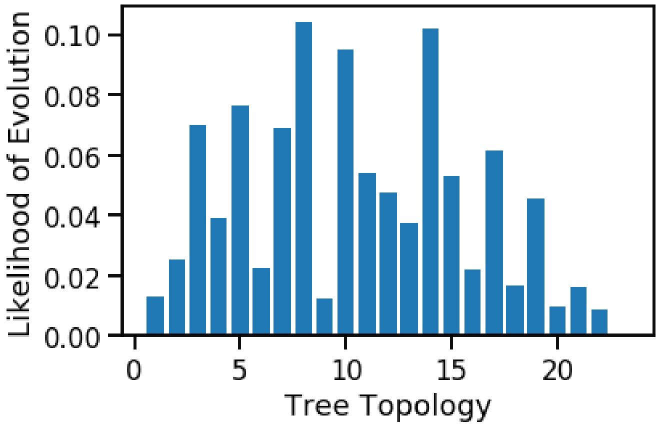


Figure 3.2: Evolutionary likelihood of tree topologies as predicted from Gain Model

Figure 3.1 demonstrates this hypothesis and shows possible evolutionary paths for the formation of a linear topology network. The total pathways possible to construct each of the twenty-three topologies was counted and is described in Methods III. The proportion of total paths available for given topology out of all ways to construct any eight-segment tree was inferred as its evolutionary likelihood (Figure 3.2). This model predicts that linear topology (type 1), as well as trees with high out-degree (type 20, 21, 22, 23), have a very low chance of evolution. Tree types 8, 10 and 14 are predicted most likely to evolve (probability ~ 0.1).

3.2 Gain-Loss Model

The previous model only considered gain of interactions, and did not include interaction loss. To account for the latter, we modeled the evolution of tree networks as a continuous process of gain and loss of interactions. According to this model, at the very beginning, the viral genome was present as eight segments which interacted via a tree network. Occasionally, mutations would cause segments to lose an existing interaction or gain an additional interaction. Unlike gain model, here, there is no further addition of foreign segments to the genome. Of the seven interactions between eight segments, loss of any interaction would cause the network to disconnect. An assembly reaction guided by a disconnected network would not be able to bring together the genome and result in zero fitness. Alternatively, gain of interaction would change the network topology from tree to cycle. Our previous calculations show that cycles have less than 100% packaging efficiency and hence, lower fitness compared to trees. Since the efficiency is not completely zero, an assembly reaction following a cycle topology could still produce progenies with complete genomes. If this progeny, with its eight segments interacting through a cycle network, undergoes a further interaction loss, its topology can change to either the same/different tree type (Figure 3.4). This mechanism of gain followed by a loss could potentially allow for the co-existence of different tree topologies within the same population. The question is after many such successive gain and loss of interactions, what is the abundance of different tree topologies at steady state? Are certain topologies more abundant than rest? To answer this, we modeled it as Markov process (Method IV). Using steady state abundances of topologies as a proxy for their likelihood to evolve, this model predicts that tree topology 3 has the maximum chance of evolution. For verifying the calculations, the steady-state abundances were also obtained from simulations and were found similar to the results from Markov Matrix (Pearson correlation coefficient = 0.996). Note that, all gain of interactions are assumed equally likely to happen and similarly, all interactions are considered equally likely to break. This might not be true in reality, wherein the probability of gaining/losing an interaction could depend on the underlying nature of the interaction. Since we do not know about the interactions during evolution, a simple model of gain followed by loss without any bias in which interaction would be gained/lost was used.

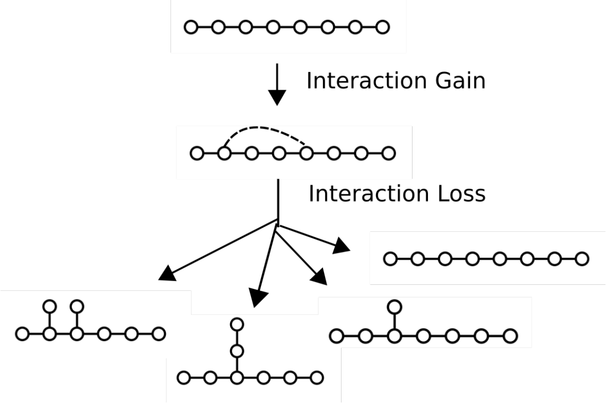


Figure 3.3: Schematic of evolution of trees from the Gain-Loss Model. The starting tree can gain an interaction, changing its topology to cycle. This cycle network can undergo a further loss of interaction changing its topology back to tree of the original type or a different type.

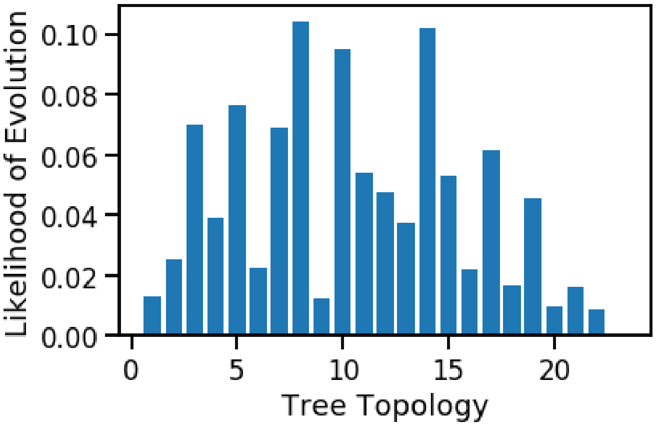


Figure 3.4: Evolutionary likelihood of tree topologies as predicted from Gain-Loss Model

3.3 Comparison of Evolutionary Models

The two models described above are based on different hypotheses for how interaction networks would have evolved. Figure 3.5 shows a correlation plot of the predicted likelihood of topologies from Gain and Gain-Loss model. Trees with high out-degree (Type 20, 21, 22, 23) are predicted least likely to evolve from both models. Trees types 5, 6, 7, 8, and 9 also have similar likelihoods predicted from the two models. On the other hand tree types 1, 2, 3 and 4 are predicted twice more likely to evolve from the gain-loss model that the gain model. Tree types 12, 14, 17 and 19 are predicted to have twice more abundance from the gain model.

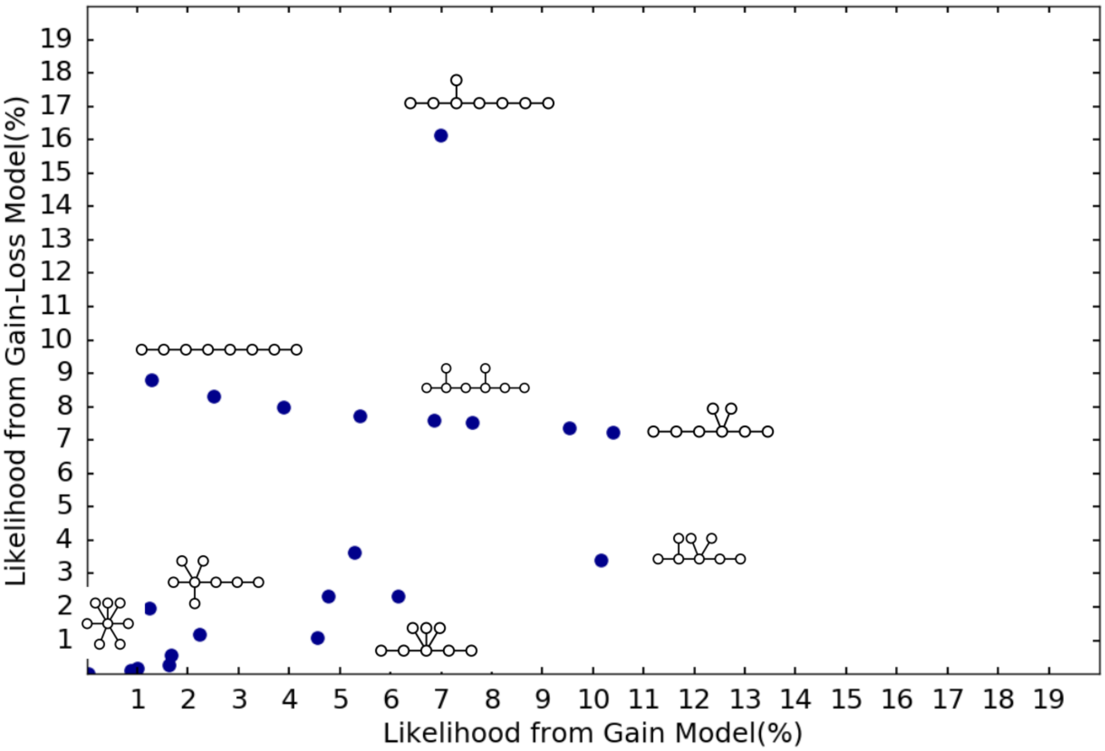


Figure 3.5: Results from Gain and Gain-Loss Model

In conclusion, we formulated two different models for the evolution of the interaction network topologies. For the gain model, the number of ways to construct a topology was used as a proxy for its evolutionary likelihood. In the gain-loss model, topologies were subjected to repeated gain followed by loss of interaction and the steady-state abundances were used as an indicator of the likelihood to evolve. Both models make a strong negative prediction about the hub-spoke topology (tree type 23) as unlikely to evolve. However, it should be noted that the evolutionary likelihoods only differed by a maximum factor of 10 and therefore, it is difficult to make any strong prediction about the topology that is most likely to evolve.

Chapter 4

Inferring Network from Experimental Data

In the previous chapters, we concluded that networks with tree topology are maximally efficient to assemble the genome and attempted to predict the topologies that are more likely to evolve. Independent of this, here, the segment interaction network is inferred from three published experimental datasets. These experiments have studied various aspects of Influenza genome assembly: segment arrangement inside the virions, segment interactions and co-localization of segments in the cytoplasm. Since all three experiments were done on Influenza A/WSN/33 (H1N1) strain and MDCK (MadinDarby canine kidney) cell line, we could integrate results and pinpoint the interactions that were consistently observed.

4.1 Electron Tomography of Virions

Inside a virion, the eight genomic segments are arranged in a characteristic 7+1 pattern, with seven vRNPs on the periphery surrounding a central vRNP (Figure 4.1). Takeshi Noda et al. further elucidated the identity of individual segments within the 7+1 arrangement for thirty virions by using segment length as a proxy for segment identity (Figure 4.2 & [7]). The length differences between segments were significant to distinguish five out of eight; however, the first three vRNPs (PB2, PB1 and PA) could not be resolved from each other. Among the thirty viral particles, segment 4 was observed at the center position for twelve virions. For the rest eighteen, segment 1/segment2/segment3 occupied the center. There was no single consistent segment order in the periphery as well, but, certain vRNP pairs occurred more often next to each other than others. For example, the heat map of nearest neighbor shows that segment 6 and segment 7 were found positioned next to each other in only 2 virions, whereas segment 6 and 8 are were nearest neighbor in 10 virions (Figure 4.10). Interestingly, the authors observed many 'thread' like structures between the RNPs, which were speculated to be RNA-RNA interactions (Appendix Figure 6.2). However, the exact nature of these 'threads' remains inconclusive.

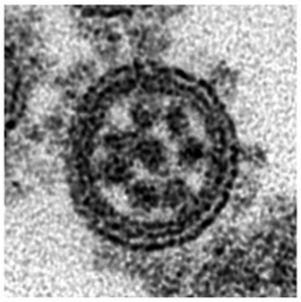


Figure 4.1: Electron microscopy of virions shows that segments are positioned in a 7+1 configuration with seven segments on the periphery and one in the center (Figure adapted [30])

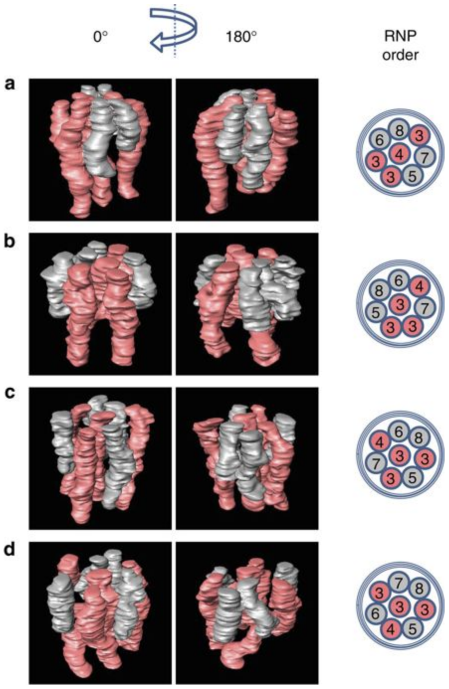


Figure 4.2: Electron tomography reveals the segment arrangement inside virions. Out of the eight segments, the first three (PB2, PB1 and PA) could not be resolved and are marked 3 in this figure. The data on the other twenty-six virions is shown in Apppendix Figure 6.1 (Figure adapted from [7])

The selective packaging model predicts that there is a set of RNA interactions underlying genome assembly. The experimental data on segment arrangement provides information on the positioning of segments inside virions. If one assumes that only nearest neighbor segments (Definitions xii) can interact, then the segment arrangement data essentially indicates the set of plausible and non-plausible interactions. We asked if one can derive the interaction network from this data, assuming that interactions take place only between nearest neighbor segments. This assumption is based on the fact that no electron density was observed between segments which are not immediately positioned next to each other. Further such interactions would face steric hindrances. If we were to include interactions with the second nearest neighbours as well, then every segment can potentially interact with all others, and this would reduce our ability to predict specific interactions.

Eight vRNPs can form a total of twenty-eight interactions ($\binom{8}{2}$), and each interaction is either present or not in actuality. Therefore, there are only 2^{28} networks, forming the entire space of all plausible networks between eight segments. To find the network that best represents the segment arrangement data, all connected networks ($\sim 60,500,000$) were scored against the experimental data. Note that within cycle networks, interactions between 6 and 7, and 6 and 8 were not included to reduce the computational load. These two interactions were observed as nearest neighbours in only 2 and 3 virions respectively. Networks were assigned penalty scores based on the number of network interactions that cannot form inside the virions (Methods V). Hence, the higher the network score, the less representative it is of the experimental data. An ideal network would have a score of zero, indicating that all interactions of that network are between segments that are positioned as nearest neighbours in virions.

In the brute force based analysis, none of the networks scored zero. The minimum score obtained was 34, corresponding to the two networks shown in Figure 4.3. Since we cannot distinguish between segment 1/2/3, all six permutations (swapping 1/2/3 positions) of both networks would also have the same score. Interestingly, 131 networks (distinct permutations) were of tree topology with lowest scores ranging from 34 to 43. This analysis points out that over the entire space of trees and cycles, the networks which best fit the experimental data are of the tree topology, thus supporting our prediction of tree being the likely interaction network.

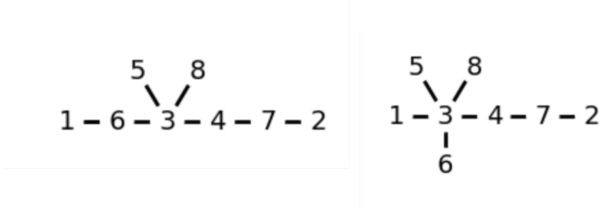


Figure 4.3: Networks that best fit the segment arrangement data (Lowest score of 34 over all tree/cycle networks)

The non-zero score of the network indicates that all seven interactions cannot form inside virions. In fact, for the lowest scoring network (Figure 4.3), thirteen virions had two interactions missing, eight virions had one missing, and nine virions contained all seven interactions (Figure 4.4). Since a minimum of seven interactions is necessary for the assembly of eight segments, our analysis would predict that few virions contain the genome as disconnected sub-assemblies. It is unlikely that segments assemble that way because then the virus would need a non-interaction based mechanism to ensure that the right sub-assemblies come together to package the complete genome. It is possible that our assumption is not entirely correct and certain non-neighbour interactions do take place. The second possibility is that after the budding, the connected assembly breaks into sub-assemblies which freely change their relative arrangement. This could cause segments that were previously interacting to break apart and position as non-nearest neighbours.

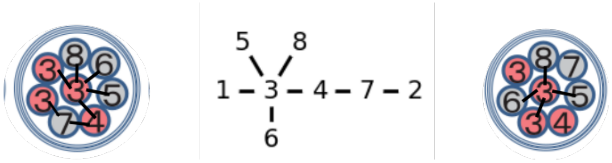


Figure 4.4: Non-zero scores of networks imply that all interactions of the network cannot form inside virions. Example of a network (Figure 4.3) which can assemble the segments in a virion but is unable to form 4-7 and 7-2 interaction in another virion.

Alternatively, one can ask what is the lowest scoring network that can package the genome as a single connected assembly in all thirty virions. Such a network is of cycle topology with 11 interactions and has a score of 96 (Figure 4.5). This network contains two master nodes 3 and 4, forming five and four interactions respectively. This is in somewhat a trivial case because master node 3 forms interaction with all segments except 1 and 4 and similarly master node 4 connects with all except 1, 8 and 2.

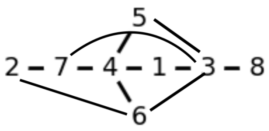


Figure 4.5: Minimum scoring network that can assemble eight segments in all thirty virions (Score = 96). Note that six permutations (swapping 1/2/3 positions) of this network would also have same score.

We next checked if the low scoring networks actually capture the electron tomography data or if they would score similar on a random configuration of virions. For this, 1000 synthetic datasets were generated, each containing 30 virions, with random order of segments on the periphery. The central segment was chosen to be 4 in 12 virions and 3 in 18 virions, as observed in the experiments. The brute force search was repeated over the entire space of trees to find the best fitting/least scoring network on each dataset. Cycles were excluded because it was computationally expensive to search over all networks. The score of the best fit network for 78% of the datasets was 48, with 42 being the least score overall, obtained only on three datasets. Majority of the best-fit networks had hub-spoke topology (type 23) (Figure 4.6). Since hub-spoke topology has a single master segment (here segment 3) which interacts the rest seven, this network would completely connect the genome in virions where segment 3 is the center. For the rest twelve, four interactions would be missing per virion, hence the score of 48.

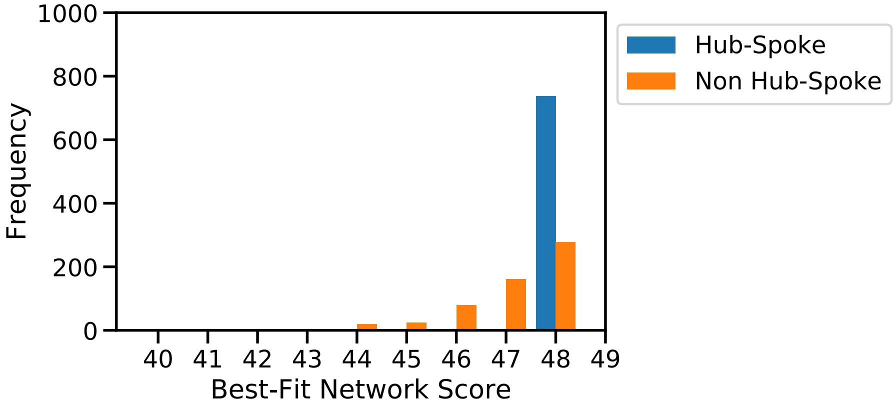


Figure 4.6: Histogram of scores/topologies of best-fit networks obtained on 1000 random datasets

The fact that for the random datasets, none of the best fit networks were non hub-spoke and had scores as low as obtained for the experimental dataset, this suggests that the low scoring networks inferred from experimental data do capture the highly non-random configuration of real virions. Since the least score on synthetic datasets was 42, further analysis was continued only on networks inferred from experimental data which scored less than 42.

4.2 Interaction Map from SHAPE-MaP and SPLASH

Electron tomography on virions cannot distinguish between the first three segments and hence the six permutations (swapping positions of 1/2/3) of any network also had the same scores. To further rank among the networks with score < 42 and their permutations, we used data from another experiment which generated an interaction map between the vRNPs (Appendix Figure 6.3). Here, the authors used a technique SHAPE-MaP (Selective 2-Hydroxyl Acylation Analysed by Primer Extension and Mutational Profiling) to probe the conformation of nucleotides and search for secondary structures [18]. The SHAPE profiles of vRNA (i.e. without nucleoproteins) was compared to vRNPs and it was observed that the profiles were different suggesting that nucleoproteins play a role in limiting the accessibility of nucleotides. To characterize the interactions, SPLASH was done on purified virions (Sequencing of Psoralen Crosslinked, Ligated, and Selected Hybrids) which cross-links base-paired nucleotides. The experiment reported the frequency at which a given interaction (i.e. base pairing between two segments) was observed. Two replicates on SPLASH (here referred as WSN2 and WSN3) were obtained which correlated with R^2 of 0.87. Since purified virions were analyzed, it is likely that the observed interactions are composed of main/primary interactions required for assembly, interactions happening post packaging and incidental interactions due to experimental protocols.

Apart from scoring potential networks against segment arrangement data, we assigned a SPLASH score (Method VI). The SPLASH score was indicative of how often interactions of a given network were observed experimentally, with higher scores corresponding to network being more close to the experimental data. To combine the two scoring schemes - electron tomography of segment arrangement scores (ET score) and SPLASH scores, a Pareto front was constructed over all tree networks. Pareto front represents the set of networks which have the most optimal scores on both scoring axes (maximum SPLASH score and minimum ET score). Figure 4.7 shows the scatter plot of scores of all tree networks ($\sim$ 260000) and the Pareto front. The networks corresponding to the Pareto front with ET scores $<$ 42 are shown in Figure 4.8.

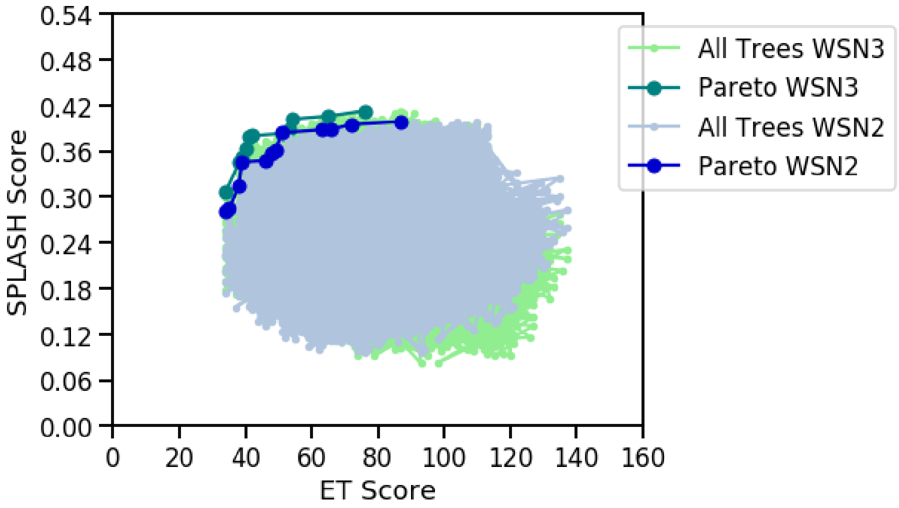
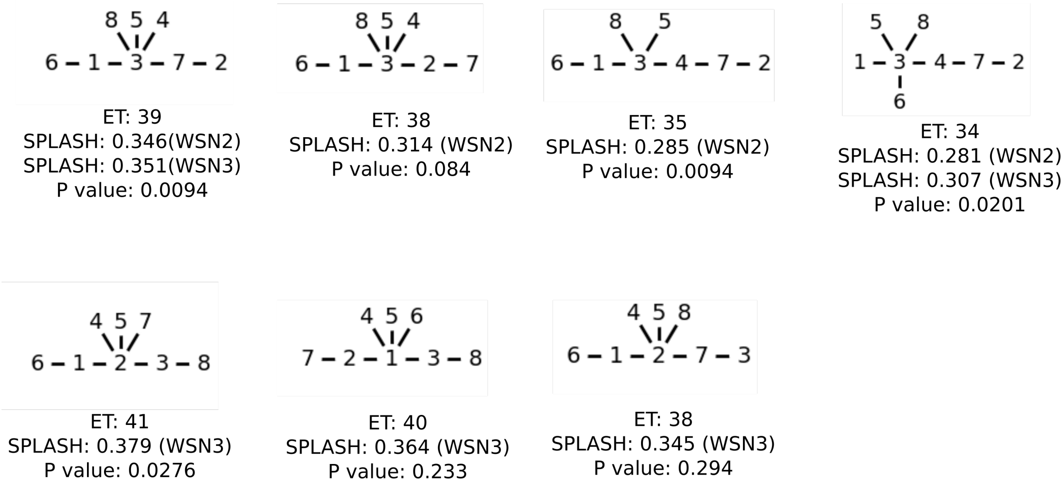


Figure 4.7: Score of all tree networks against electron tomography data of segment arrangement inside virions (ET score) and SPLASH data of interaction maps (SPLASH score). Note that lower ET score and higher SPLASH score is indicative of the network being closer to the experimental data.



ET: 39
SPLASH: 0.346(WSN2)
SPLASH: 0.351(WSN3)
P value: 0.0094

ET: 38
SPLASH: 0.314 (WSN2)
P value: 0.084

ET: 35
SPLASH: 0.285 (WSN2)
P value: 0.0094

ET: 34
SPLASH: 0.281 (WSN2)
SPLASH: 0.307 (WSN3)
P value: 0.0201

ET: 41
SPLASH: 0.379 (WSN3)
P value: 0.0276

ET: 40
SPLASH: 0.364 (WSN3)
P value: 0.233

ET: 38
SPLASH: 0.345 (WSN3)
P value: 0.294

Figure 4.8: Pareto trees

Another experiment quantified the colocalization coefficients (on a scale of 0 to 1) of the twenty-eight segment pairs by using fluorescent tags inside the cytoplasm [28]. These coefficients were obtained from a snapshot of the assembly process (8 hours post infection) for multiple single cells (Appendix Figure 6.4). Since only four segments were visualized at a time and the coefficients showed high variation among the cells, with 20 pairs having standard deviation more than/equal to 0.2, we did not infer the network independently here. This data was used only as a confirmatory test to check if the segments forming interactions in Pareto networks co-localize significantly more than the ones that do not interact (Method VII). Five Pareto trees were found to have significant colocalization between its interacting segments ($\alpha = 0.05$), with the lowest p-value being 0.0094 (Figure 4.8).

If one ignores the segment arrangement data and uses only high-frequency interactions from SPLASH to construct a tree, then the p-value of that tree is 0.17 (WSN2) and 0.058 (WSN3), both being higher than the values obtained for Pareto trees. Similarly, if one ignores the SPLASH data and uses only the segment arrangement data, the best fit network obtained is one of the Pareto trees. However, the heat map of interactions present in the Pareto trees (score < 42) is different from the heat map of nearest neighbours in segment arrangement data (Figure 4.9 & Figure 4.10). For example, relying solely on the assumption that nearest neighbour interact, EM data would predict interactions between segment 5 & 6, 5 & 7, 5 & 8 and, 4 & 7. None of these interactions come up if one overlays this with SPLASH data. We believe that combining the results from the two experiments have improved our ability to make predictions on non-incidental and consistently observed interactions.

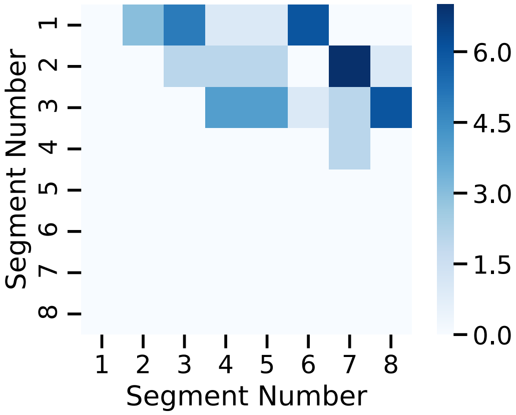


Figure 4.9: Heat map of frequency at which segment interactions are observed within Pareto trees.

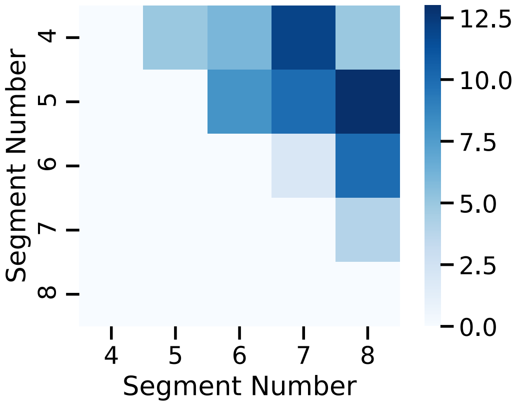


Figure 4.10: Heat map of frequency at which segments are observed as nearest neighbours inside thirty virions.

It is tempting to predict that the topmost Pareto tree in Figure 4.8 is the most likely interaction network because it has the lowest p-value of 0.0094, relatively good ET score of 39 (four more than lowest ET score obtained over all networks), SPLASH score of 0.351 (0.05 less than the maximum splash score), and is present in both replicates. However, the heat map of interactions observed in Pareto trees (Figure 4.9) generates a more modest prediction on the set of interactions that are statistically significant and consistently observed.

Chapter 5

Conclusion

This study was centered on understanding the mechanism of influenza genome packaging. The genome of Influenza virus is present as eight distinct RNA segments, each in complex with nucleoproteins and polymerases. These segments code for proteins essential for the production of infectious progenies from host cells. Interestingly, the virus can assemble its replicated segments and package the complete genome in majority of its progenies. It has been proposed that segments interact with each other through specific RNA-RNA interactions, leading to the assembly of eight segments. However, the underlying mechanism/interactions that give rise to a robust and accurate assembly remains unclear.

Influenza genome packaging falls under the category of self-assembly processes often observed in nature. Self-assembled structures can emerge either from strong interactions or transient but cooperative interactions. It is not known yet which of these two mechanisms operate for the genome assembly in Influenza. In chapter 2, we started with the assumption that the segments form strong specific interactions among each other, these interactions being irreversible in cellular timescale. The chapter discussed the following question; how many strong interactions are required for an efficient assembly? Using stochastic simulations of assembly reaction and numerical calculations, it was shown that seven interactions (i.e a network of tree topology) would guarantee genome packaging, whereas more than seven interactions (i.e network of cycle topology) would decrease the efficiency of assembly reactions. The absolute value of decrease would depend on the number of interactions and size of loops, with the efficiency decreasing as one increases both these parameters. To the best of our knowledge, there is no strong evidence for the presence of transient cooperative interactions. Therefore, we can only conclude that an assembly following a tree network comprising of seven strong interactions would package the genome with high efficiency, as observed in experiments. An assembly guided by cycle network would not lead to maximum efficiency unless some interactions are weak and there is an order/cooperativity in the assembly process. This chapter was approached from a reaction efficiency point of view, and we did not account for the robustness of networks during evolution. Since interactions are formed by base-pairing, having more interactions in a network would cause the network to be more robust against mutations. It would be interesting to explore if there is a trade-off between robustness during evolutionary timescale and accuracy during assembly in cellular timescale.

In chapter 3, we continued with the hypothesis of tree being the likely topology of the interaction network and studied the evolutionary likelihoods of different tree topologies by constructing two models - Gain Model and Gain-Loss Model. The gain model was based on the idea that the segmented genome and its interaction network evolved through continuous gain of segments. The order of segment gain would dictate the trajectory for the evolution of a topology. The hypothesis was that the more the number of possible trajectories, the more the likelihood of evolution. Tree topology 8 were predicted most likely to evolve, and topology with high out-degrees (type 21, 22, 23) was predicted unlikely to evolve. The second model conceptualized the evolution of trees as a Markov process of gain and loss of interactions. The abundance of different topologies at steady state were inferred as their evolutionary likelihoods. Tree topology 3 was predicted most likely to evolve and consistent with the gain model, trees with high out degrees had a low chance of evolution. It should be pointed out that the evolutionary likelihoods only differed by a factor of 10 and therefore, it is difficult to conclude that topology 8 or topology 3 is highly likely to evolve.

Independent of these results, in chapter 4, we attempted to infer the interaction network from three published experimental results. These experiments provided insights into various aspects of influenza genome assembly; namely, segment arrangement inside the virions, segment interactions map and the co-localization of segments during assembly in cytoplasm. To find the network that best fits the experimental results, all possible networks between eight segments were scored against the segment arrangement data, and it was found that the best fitting networks were of tree topology. Further, by combining results from the other two experiments, we constructed a heat map to pinpoint the interactions that were consistently observed and statistically significant. It should be noted that only nearest neighbor segments were assumed to interact inside virions. As a consequence of this assumption, the predicted best fit networks do not cluster together eight segments in all virions. This can be explained in several ways. First, perhaps a few non-neighbor interactions do happen, which would allow the predicted networks to cluster all segments. Second, since purified virions were being analyzed, it is probable that the interactions that happened at the time of assembly were no longer present and segments in the sub-assemblies had changed relative positioning. Thirdly, we cannot falsify the existence of many weak cooperative interactions underlying genome assembly. Such a mechanism would predict that influenza has a highly connected interaction network and different subset of interactions allow genome assembly in different virions.

Previously, vRNAs have been shown to interact in-vitro using EMSA [10]. Since vRNA is present in a complex with nucleoproteins in cells and nucleoproteins can dictate the availability of interacting region, it might be erroneous to test predictions against the EMSA study of vRNAs. An ideal test would be to check if the vRNPs predicted to interact can indeed form interactions in-vitro. One can further see if the presence of other segments alters the interactions. This experiment would inform about the interactions taking place during the assembly and if the interactions between vRNPs are strong/stable or transient/cooperative.

This study and it's experimental test would hopefully provide insights into the precise set of interactions underlying influenza genome packaging, mechanism of viral reassortment and genome assembly in other viruses with segmented genomes.

Chapter 6

Methods

I. Calculations on Assembly Efficiency of Networks

In this section, the reaction efficiency (η) is computed for two cases i) assembly reaction following tree network and ii) assembly reaction following a minimal cyclic network. The efficiency of a reaction is defined as the proportion of completely assembled clusters out of all clusters formed at steady state. The calculations/proofs are valid for a closed system, which at its initial state contains N copies of each of the M distinct monomers. The central assumption is that all interactions of a network are equally probable and irreversible. Square brackets are used to denote the count of a given reaction species at a particular time instant. For M=3, monomers are labeled as A, B and C, and the fully assembled cluster is denoted by ABC.

Packaging Efficiency of Trees

Theorem 1: The efficiency of an assembly reaction following a tree interaction network is 1.

Proof Using Mass Conservation (M=3)

At initial state ($Time = 0$):

$$[\mathrm{A}] = [\mathrm{B}] = [\mathrm{C}] = \mathrm{N}$$

At steady state ($Time \rightarrow \infty$):

$$[\text{ABC}] = \text{P}$$
$$[\text{AB}] = \text{X},\ [\text{BC}] = \text{Y}$$
$$[\text{A}] = [\text{B}] = [\text{C}] = 0$$

where values of P, X and Y are determined by solving Equation 6.1-6.3.

Since it is a closed system, the amount of monomers A, B and C should be conserved at all time points.

$$[AB] + [ABC] = N \rightarrow X + P = N \tag{6.1}$$

$$[AB] + [BC] + [ABC] = N \rightarrow X + Y + P = N \tag{6.2}$$

$$[BC] + [ABC] = N \rightarrow Y + P = N \tag{6.3}$$

By solving the equations above for steady state, we obtain,

$$[\text{AB}] = [\text{BC}] = 0$$
$$[\text{ABC}] = \text{N}$$

Therefore, $\eta = \dfrac{[ABC]}{[ABC] + [AB] + [BC] + [A] + [B] + [C]} = 1$

Proof By Contradiction (Generalized)

Proposition: The assembly efficiency of a reaction following a interaction network of tree topology is 1.

Proof: Suppose there is a reaction in which segments interact and assemble according to a tree network and the efficiency of this reaction is < 1.0

$\Longrightarrow$ At steady state, there is at least one incomplete cluster C, having less than M monomers.

$\Longrightarrow$ C lacks at-least one monomer. Let us label the missing monomer as M_1.

$\Longrightarrow$ Let us say M_1 has an interacting partner M_2 in C.

$\Longrightarrow$ There are no free unbound M_1 monomers that can bind to M_2 and merge with C.

$\implies$ All N copies of M_1 are bound to all N copies of M_2.
$\implies$ But, there is one copy of M_2 monomer in C.

The total count of M_2 monomers cannot exceed N. Hence, the case described above is in contradiction with system properties and therefore for a tree interaction network, system will always reach 100% packaging efficiency.

Packaging Efficiency of Cycles

Theorem 2: The efficiency of an assembly reaction wherein monomers interact based on a minimal cycle network can be less than 1.

To prove this, it is suffice to show that there exists at-least one steady state for which $\eta < 1$.

At initial state ($Time = 0$):

$$[\mathrm{A}] = [\mathrm{B}] = [\mathrm{C}] = \mathrm{N}$$

At steady state ($Time \to \infty$):

$$[\mathrm{ABC}] = \mathrm{P}$$
$$[\mathrm{AB}] = \mathrm{X},\ [\mathrm{BC}] = \mathrm{Y},\ [\mathrm{AC}] = \mathrm{Z}$$
$$[\mathrm{A}] = [\mathrm{B}] =[\mathrm{C}] =0$$

where values of P, X and Y are determined by solving Equation 6.4-6.6.

Since the system is closed, the amount of each monomer should be conserved at all time points.

$$[AB] + [AC] + [ABC] = N \to X + Z + P = N \tag{6.4}$$

$$[AB] + [BC] + [ABC] = N \to X + Y + P = N \tag{6.5}$$

$$[BC] + [AC] + [ABC] = N \to Y + Z + P = N \tag{6.6}$$

By solving the equations above for steady state, we obtain,

$$[\text{AB}] = [\text{BC}] = [\text{AC}] = \text{X}$$
$$[\text{ABC}] = \text{N - 2* [AB]}$$

Therefore, $\eta = \dfrac{[ABC]}{[ABC] + [AB] + [BC] + [AC] + [A] + [B] + [C]} <= 1$

(If [AB] = [BC] = [AC] = 0, $\eta = 1$, otherwise $\eta < 1$)

II. Stochastic Simulation for Genome Assembly

Stochastic simulations of assembly reaction following a given interaction network were carried out to compute the reaction efficiency of that network. All interactions of the network were considered irreversible and equally probable. Each simulation was started with 100 copies of each of the distinct monomers and was continued till steady state (no further reaction possible). The system state was stored as a list of all reaction species at any time instant. In every iteration, two reaction species were picked at random from the list and checked for the possibility of interaction based on the network defined. If the two reactant species had no common monomers and could interact, they were put back into the system as a fused cluster. If no interaction was possible, both reactants were put back as it is. At the end of the simulation, efficiency was computed as the proportion of completely assembled clusters out of total clusters present.

III. Evolutionary Likelihood Calculation: Gain Model

This section illustrates the method for computing evolutionary likelihood of tree topologies from Gain Model through an example for the linear topology. The calculation is centered around counting the total number of paths through which a given topology can evolve, the idea being that topologies which have more pathways are more likely to evolve. In other words, given an interaction network, one has to back-trace all paths through which that network could have evolved.

Figure 3.1 shows an interaction network consisting of eight non-identical segments (represented with different colors), arranged in a linear topology. As per the model, the evolution of a network began from a single segment and the rest seven segments were added subsequently to the growing network. Any one of these eight could have been the starting/most ancestral segment. Since ancestral segments which are topologically symmetric would lead to equal number of pathways to form the final topology, we only considered the set of non-symmetric segments as potential ancestors, just to avoid over-counting (in the example: yellow, orange, violet and green).

The number of evolutionary pathways is dependent on the number of already present segments in the evolving network to which a new segment can form an interaction with. For example, if the most ancestral segment is considered to be yellow, the only segment that can be next added to the network is orange because yellow does not interact with any other segment. The segment that can be further added is violet for the same reason and so on. Hence, there is only one way to construct the given tree network starting from the yellow. Instead, if the starting segment is red, the second segment can be either yellow or violet because red interacts with both. The options for the third segment would depend on the choice of the second segment. The order of addition of the yellow segment can be between second to eighth. Therefore, there are seven choices and once its order is fixed, the other six segments can be added in only way. Therefore, starting from red, there are 7C1 pathways to construct linear topology. Similarly, if the starting segment is taken as violet, yellow and red segments can be added in 7C2 orders and once the order of addition of these two segments is fixed, there is only one way to construct the rest of the network. The total number of ways of ways in which a linear topology can be constructed is equal to the sum of number of paths from each ancestral segment: 1 (starting: yellow), 7C1 (starting: red), 7C2 (starting: violet) and 7C3 (starting: green). While this example is specifically for the linear case, the same methodology was followed for likelihood calculation for other topologies.

IV. Markov Matrix for Gain-Loss model

A 23 X 23 markov matrix was computed to capture the probabilities of transitioning from one tree topology to another through a single gain followed by a loss of interaction. If a tree network gains an interaction, it's topology would change to cycle. A further loss of interaction in a cycle network (with eight interactions) would shift the network back to tree, either of the original tree topology or a different one. To compute the probabilities of transitioning from i^{th} tree topology to j^{th}, the i^{th} topology was subjected to all possible combinations of gain-loss of interactions and the proportion of j^{th} topologies obtained was calculated and used as the probability of transitioning from i^{th} to j^{th} topology.

V. Scoring Network against Segment Arrangement Data

To judge how well a network fits the experimental data on segment arrangement inside virions, the network was scored against each of the thirty virions and was summed over to obtain an overall score. We assumed that only nearest neighbour segments inside the virion can interact. The scoring was based on the number of network interactions that cannot form inside a given virion, owing to the segments being non nearest neighbors. Since segment 1, 2 and 3 are indistinguishable in the data, six permutations were generated for each virion by swapping the positions of these three segments. The network was scored against the six permutations and the minimum score over the six was taken as the score of that virion (Eq. 6.7).

$$\text{Score of Network N} = \sum_{V=1}^{30} min(P_{V1}, P_{V2}, P_{V3}, P_{V4}, P_{V5}, P_{V6}) \tag{6.7}$$

where P_{Vi} : number of interactions of network N are cannot form in i^th permutation of virion V because interacting segment do not occur as nearest neighbours.

VI. Scoring Network against SPLASH Data

The frequency at which a given interaction is observed in SPLASH data was obtained from the authors for all pairwise interactions ([18]). The normalized sum of the frequencies of seven interactions in a given network was used as the score for that network (Eq 6.8).

$$\text{SPLASH Score of Network N} = \frac{\sum_{Ni=1}^{7} F_{Ni}}{\sum_{j=1}^{28} F_j} \tag{6.8}$$

where F_{Ni} is the interaction frequency reported from SPLASH for the i^{th} interaction of network N. F_j refers to the observed frequency for j^{th} interaction, where j is ranged over all twenty-eight interactions.

VII. Assigning P-value to Network using Colocalization Data

The average co-localization coefficients for each of the twenty-eight RNP pairs were derived from the plots (Appendix Figure 6.4) by using 1.7 cm on the scale as equivalent to a coefficient of 1. A p-value (Mann Whitney U test one-tailed) was assigned to networks to capture if the segments forming interactions in that network (n=7 pairs) have significantly higher co-localization coefficients as compared to their co-localization with other vRNPs (n=21 pairs).

Appendix

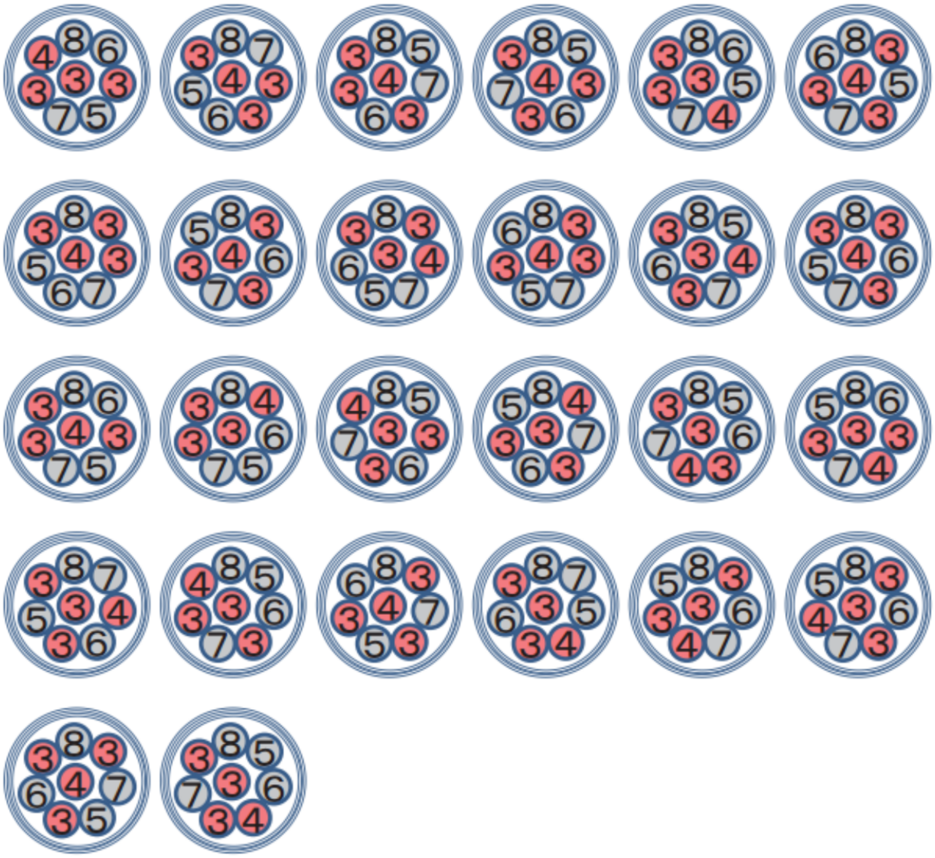


Figure 6.1: Segment arrangement inside twenty-six virions obtained using electron tomography. Figure adapted from [7]

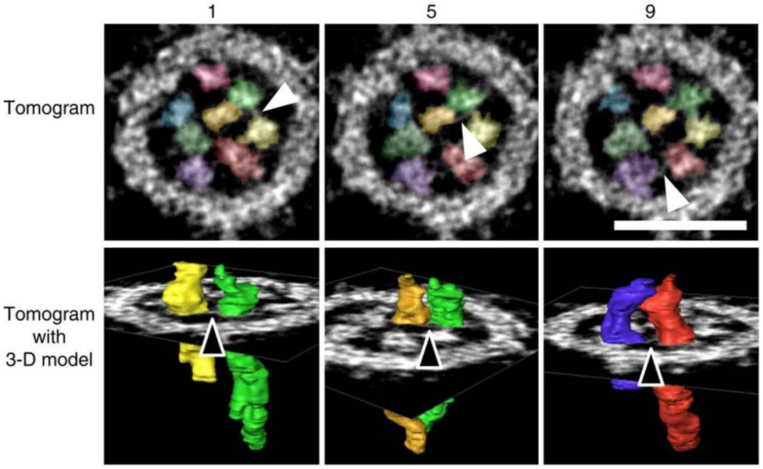


Figure 6.2: Thread like structures observed between genomic segments inside virions using electron tomography. Figure adapted from [7]

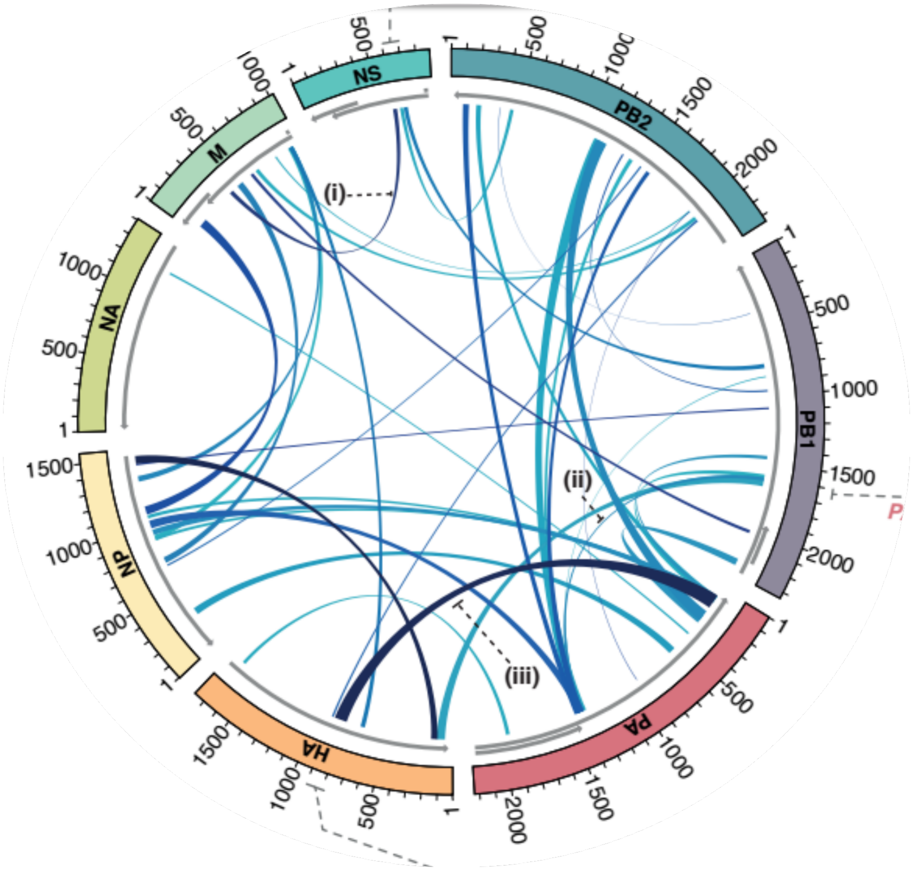


Figure 6.3: Interaction map between genomic segments inferred from SPLASH. Line thickness is correlated to the observed interaction frequency. Figure adapted from [18].

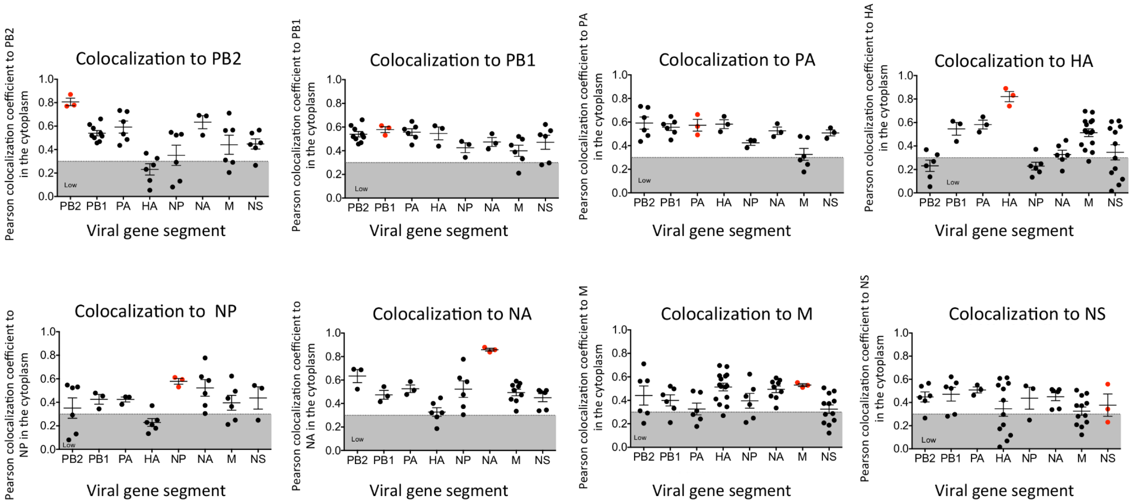


Figure 6.4: Colocalization coefficients of segments in cytoplasm obtained using fluorescence in-situ hybridization. Figure adapted from [28]

Bibliography

[1] Tasleem Samji. Influenza A: Understanding the viral life cycle, 2009.

[2] John Steel and Anice C. Lowen. Influenza a virus reassortment. *Current Topics in Microbiology and Immunology*, 2014.

[3] Timo Frensing, Sascha Y. Kupke, Mandy Bachmann, Susanne Fritzsche, Lili E. Gallo-Ramirez, and Udo Reichl. Influenza virus intracellular replication dynamics, release kinetics, and particle morphology during propagation in MDCK cells. *Applied Microbiology and Biotechnology*, 2016.

[4] Nicole M. Bouvier and Peter Palese. The biology of influenza viruses. *Vaccine*, 2008.

[5] Yi Shi, Ying Wu, Wei Zhang, Jianxun Qi, and George F. Gao. Enabling the 'host jump': Structural determinants of receptor-binding specificity in influenza A viruses, 2014.

[6] Sumiho Nakatsu, Hiroshi Sagara, Yuko Sakai-Tagawa, Norio Sugaya, Takeshi Noda, and Yoshihiro Kawaoka. Complete and incomplete genome packaging of influenza A and B viruses. *mBio*, 2016.

[7] Takeshi Noda, Yukihiko Sugita, Kazuhiro Aoyama, Ai Hirase, Eiryo Kawakami, Atsuo Miyazawa, Hiroshi Sagara, and Yoshihiro Kawaoka. Three-dimensional analysis of ribonucleoprotein complexes in influenza A virus. *Nature Communications*, 2012.

[8] Emilie Fournier, Vincent Moules, Boris Essere, Jean Christophe Paillart, Jean Daniel Sirbat, Catherine Isel, Annie Cavalier, Jean Paul Rolland, Daniel Thomas, Bruno Lina, and Roland Marquet. A supramolecular assembly formed by influenza A virus genomic RNA segments. *Nucleic Acids Research*, 2012.

[9] Y.-y. Chou, R. Vafabakhsh, S. Doganay, Q. Gao, T. Ha, and P. Palese. One influenza virus particle packages eight unique viral RNAs as shown by FISH analysis. *Proceedings of the National Academy of Sciences*, 2012.

[10] Cyrille Gavazzi, Catherine Isel, Emilie Fournier, Vincent Moules, Annie Cavalier, Daniel Thomas, Bruno Lina, and Roland Marquet. An in vitro network of intermolecular interactions between viral RNA segments of an avian H5N2 influenza A virus: Comparison with a human H3N2 virus. *Nucleic Acids Research*, 2013.

[11] C. Gavazzi, M. Yver, C. Isel, R. P. Smyth, M. Rosa-Calatrava, B. Lina, V. Moules, and R. Marquet. A functional sequence-specific interaction between influenza A virus genomic RNA segments. *Proceedings of the National Academy of Sciences*, 2013.

[12] Ulrich Desselberger, Vincent R. Racaniello, James J. Zazra, and Peter Palese. The 3' and 5'-terminal sequences of influenza A, B and C virus RNA segments are highly conserved and show partial inverted complementarity. *Gene*, 1980.

[13] G. A. Marsh, R. Rabadan, A. J. Levine, and P. Palese. Highly Conserved Regions of Influenza A Virus Polymerase Gene Segments Are Critical for Efficient Viral RNA Packaging. *Journal of Virology*, 2008.

[14] Yuki Kobayashi, Bernadeta Dadonaite, Neeltje van Doremalen, Yoshiyuki Suzuki, Wendy S. Barclay, and Oliver G. Pybus. Computational and molecular analysis of conserved influenza A virus RNA secondary structures involved in infectious virion production. *RNA Biology*, 2016.

[15] E. C. Hutchinson, M. D. Curran, E. K. Read, J. R. Gog, and P. Digard. Mutational Analysis of cis-Acting RNA Signals in Segment 7 of Influenza A Virus. *Journal of Virology*, 2008.

[16] Edward C. Hutchinson, Helen M. Wise, Katerine Kudryavtseva, Martin D. Curran, and Paul Digard. Characterisation of influenza A viruses with mutations in segment 5 packaging signals. *Vaccine*, 2009.

[17] Marie Gerber, Catherine Isel, Vincent Moules, and Roland Marquet. Selective packaging of the influenza A genome and consequences for genetic reassortment, 2014.

[18] Bernadeta Dadonaite, Egle Barilaite, Ervin Fodor, Alain Laederach, and David LV Bauer. The structure of the influenza a virus genome. *bioRxiv*, 2017.

[19] Elzbieta Lenartowicz, Julita Kesy, Agnieszka Ruszkowska, Marta Soszynska-Jozwiak, Paula Michalak, Walter N. Moss, Douglas H. Turner, Ryszard Kierzek, and Elzbieta Kierzek. Self-folding of naked segment 8 genomic RNA of influenza a virus. *PLoS ONE*, 2016.

[20] Agnieszka Ruszkowska, Elzbieta Lenartowicz, Walter N. Moss, Ryszard Kierzek, and Elzbieta Kierzek. Secondary structure model of the naked segment 7 influenza A virus genomic RNA. *The Biochemical journal*, 2016.

[21] Étori Aguiar Moreira, Anna Weber, Hardin Bolte, Larissa Kolesnikova, Sebastian Giese, Seema Lakdawala, Martin Beer, Gert Zimmer, Adolfo García-Sastre, Martin Schwemmle, and Mindaugas Juozapaitis. A conserved influenza A virus nucleoprotein code controls specific viral genome packaging. *Nature Communications*, 2016.

[22] Z. Li, T. Watanabe, M. Hatta, S. Watanabe, A. Nanbo, M. Ozawa, S. Kakugawa, M. Shimojima, S. Yamada, G. Neumann, and Y. Kawaoka. Mutational Analysis of Conserved Amino Acids in the Influenza A Virus Nucleoprotein. *Journal of Virology*, 2009.

[23] Hardin Bolte, Miruna E. Rosu, Elena Hagelauer, Adolfo García-Sastre, and Martin Schwemmle. Packaging of the influenza A virus genome is governed by a plastic network of RNA/protein interactions. *Journal of Virology*, 2018.

[24] Seema S. Lakdawala, Ervin Fodor, and Kanta Subbarao. Moving On Out: Transport and Packaging of Influenza Viral RNA into Virions. *Annual Review of Virology*, 2016.

[25] Edward C. Hutchinson, Johann C. von Kirchbach, Julia R. Gog, and Paul Digard. Genome packaging in influenza A virus, 2010.

[26] S. D. Duhaut and J. W. McCauley. Defective RNAs inhibit the assembly of influenza virus genome segments in a segment-specific manner. *Virology*, 1996.

[27] Y. Fujii, H. Goto, T. Watanabe, T. Yoshida, and Y. Kawaoka. Selective incorporation of influenza virus RNA segments into virions. *Proceedings of the National Academy of Sciences*, 2003.

[28] Seema S. Lakdawala, Yicong Wu, Peter Wawrzusin, Juraj Kabat, Andrew J. Broadbent, Elaine W. Lamirande, Ervin Fodor, Nihal Altan-Bonnet, Hari Shroff, and Kanta Subbarao. Influenza A Virus Assembly Intermediates Fuse in the Cytoplasm. *PLoS Pathogens*, 2014.

[29] Jim Madge, David Bourne, and Mark A. Miller. Controlling Fragment Competition on Pathways to Addressable Self-Assembly. *Journal of Physical Chemistry B*, 2018.

[30] Takeshi Noda, Hiroshi Sagara, Albert Yen, Ayato Takada, Hiroshi Kida, R. Holland Cheng, and Yoshihiro Kawaoka. Architecture of ribonucleoprotein complexes in influenza A virus particles. *Nature*, 2006.