PLOS ONE

RESEARCH ARTICLE

Positive Selection or Free to Vary? Assessing the Functional Significance of Sequence Change Using Molecular Dynamics

Jane R. Allison[1,2,3]*, Marcus Lechner[4], Marc P. Hoeppner[5], Anthony M. Poole[2,6]*

1 Centre for Theoretical Chemistry and Physics & Institute of Natural and Mathematical Sciences, Massey University Albany, Auckland, New Zealand, **2** Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand, **3** Maurice Wilkins Centre for Molecular Biodiscovery, Massey University Albany, Auckland, New Zealand, **4** Department of Pharmaceutical Chemistry, Philipps-University Marburg, Marburg, Germany, **5** Christian-Albrechts-University of Kiel, Institute of Clinical Molecular Biology, Kiel, Germany, **6** School of Biological Sciences, University of Canterbury, Christchurch, New Zealand

* j.allison@massey.ac.nz (JA); anthony.poole@canterbury.ac.nz (AP)

OPEN ACCESS

Citation: Allison JR, Lechner M, Hoeppner MP, Poole AM (2016) Positive Selection or Free to Vary? Assessing the Functional Significance of Sequence Change Using Molecular Dynamics. PLoS ONE 11 (2): e0147619. doi:10.1371/journal.pone.0147619

Editor: Narayanaswamy Srinivasan, Indian Institute of Science, INDIA

Received: July 17, 2015

Accepted: January 6, 2016

Published: February 12, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received funding from the following sources: Royal Society of New Zealand Marsden Fund Fast Start award (13-MAU-039) to JRA; Royal Society of New Zealand Rutherford Discovery Fellowship (RDF-11-UOC-013) to AMP; Biomolecular Interaction Centre, University of Canterbury (to AMP, JRA); Massey University Research Fund Early Career Grant (to JRA); and Swedish Research Council (to AMP). The funders had no role in study design, data collection and

Abstract

Evolutionary arms races between pathogens and their hosts may be manifested as selection for rapid evolutionary change of key genes, and are sometimes detectable through sequence-level analyses. In the case of protein-coding genes, such analyses frequently predict that specific codons are under positive selection. However, detecting positive selection can be non-trivial, and false positive predictions are a common concern in such analyses. It is therefore helpful to place such predictions within a structural and functional context. Here, we focus on the p19 protein from tombusviruses. P19 is a homodimer that sequesters siRNAs, thereby preventing the host RNAi machinery from shutting down viral infection. Sequence analysis of the p19 gene is complicated by the fact that it is constrained at the sequence level by overprinting of a viral movement protein gene. Using homology modeling, in silico mutation and molecular dynamics simulations, we assess how non-synonymous changes to two residues involved in forming the dimer interface—one invariant, and one predicted to be under positive selection—impact molecular function. Interestingly, we find that both observed variation and potential variation (where a non-synonymous change to p19 would be synonymous for the overprinted movement protein) does not significantly impact protein structure or RNA binding. Consequently, while several methods identify residues at the dimer interface as being under positive selection, MD results suggest they are functionally indistinguishable from a site that is free to vary. Our analyses serve as a caveat to using sequence-level analyses in isolation to detect and assess positive selection, and emphasize the importance of also accounting for how non-synonymous changes impact structure and function.

analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Evolutionary arms races between host and pathogens can lead to rapid change in key genes associated with immune response or pathogenicity. In eukaryotes, RNA interference (RNAi) is a broadly conserved [1–6] mechanism that provides effective defense against a range of genomic pathogens, including viruses. Numerous viruses in turn encode viral suppressors of RNA silencing (VSR) that act via diverse mechanisms to suppress RNAi-mediated antiviral defense [7–9]. The distribution and variety of VSRs shows that viral suppressors have evolved numerous times independently, confirming the ongoing arms race between viruses and their hosts.

Evolutionary arms races may be detectable at the sequence level as positive selection. Positive selection is an excess of non-synonymous (amino-acid changing) nucleotide substitutions relative to synonymous substitutions (that do not affect protein sequence). Key genes involved in the RNAi response have been shown to be under positive selection [10], and it therefore seems reasonable to expect that VSR genes in viruses might also show evidence of positive selection.

Several factors complicate sequence-level analyses of positive selection. False positive predictions may arise when some synonymous sites are under greater constraint [11–13]. This may be a particular concern in viruses, where some protein-coding genes are known to be overprinted on other genes [11, 14, 15]. Overprinting refers to situations where two protein products, encoded in different frames or orientations, arise from the same nucleic acid sequence [15]. The de novo emergence of genes overprinted on existing genes is well documented in viruses [15–18]. Another complication, and the main focus of this study, is the impact that non-synonymous changes can have on function; sequence-level changes may not have equivalent effects on protein structure or function, and variation may be functionally significant even where a sequence is not under positive selection.

Given concerns with false-positive predictions of positive selection [19–21], we have examined the intriguing case of p19, a viral suppressor of RNAi from the tombusvirus family of plant viruses [22]. P19 has significant constraints on sequence change due to the p19 gene being overprinted on the tombusvirus movement protein (MP) gene (**Fig 1A**). P19 is directly involved in viral suppression of the host RNAi machinery. It suppresses plant RNAi silencing by binding and sequestering siRNAs produced in response to viral infection [9, 22]. Structural studies indicate that p19 forms a homodimer that binds dsRNA in a sequence-independent but size-selective manner [23, 24], and this in turn prevents systemic spread of siRNA [25, 26].

P19 is therefore a clear candidate for participation in a host-pathogen arms race [7, 9]. Moreover, the p19 gene is overprinted on the MP gene in all known tombusvirus genomes [27] (**Fig 1A and Figure A in S1 File**). The structure of p19 is inherently linked to its function—it is an obligate dimer with respect to siRNA size selection and binding [23, 24]—and p19 appears to exhibit greater sequence diversity than MP. We therefore sought to establish whether p19 shows evidence of being subject to positive selection, as has been shown for other overprinted genes [14] and whether this signal appears genuine or could be an artifact of sequence level constraints attributable to its unusual genic organization. Following sequence level analyses to identify residues possibly under positive selection, we considered the structural and functional aspects of the identified sequence changes to give a more complete picture of their biological implications.

Sequence-based analyses yielded results consistent with p19 being under positive selection, despite the fact that MP overprinting constrains variation at p19 residues. Among sites putatively under positive selection, we chose to investigate those specifically involved in dimer formation. To establish how sequence changes impact function, we used molecular dynamics (MD) simulations to probe the effect on the p19 homodimer of all the observed natural

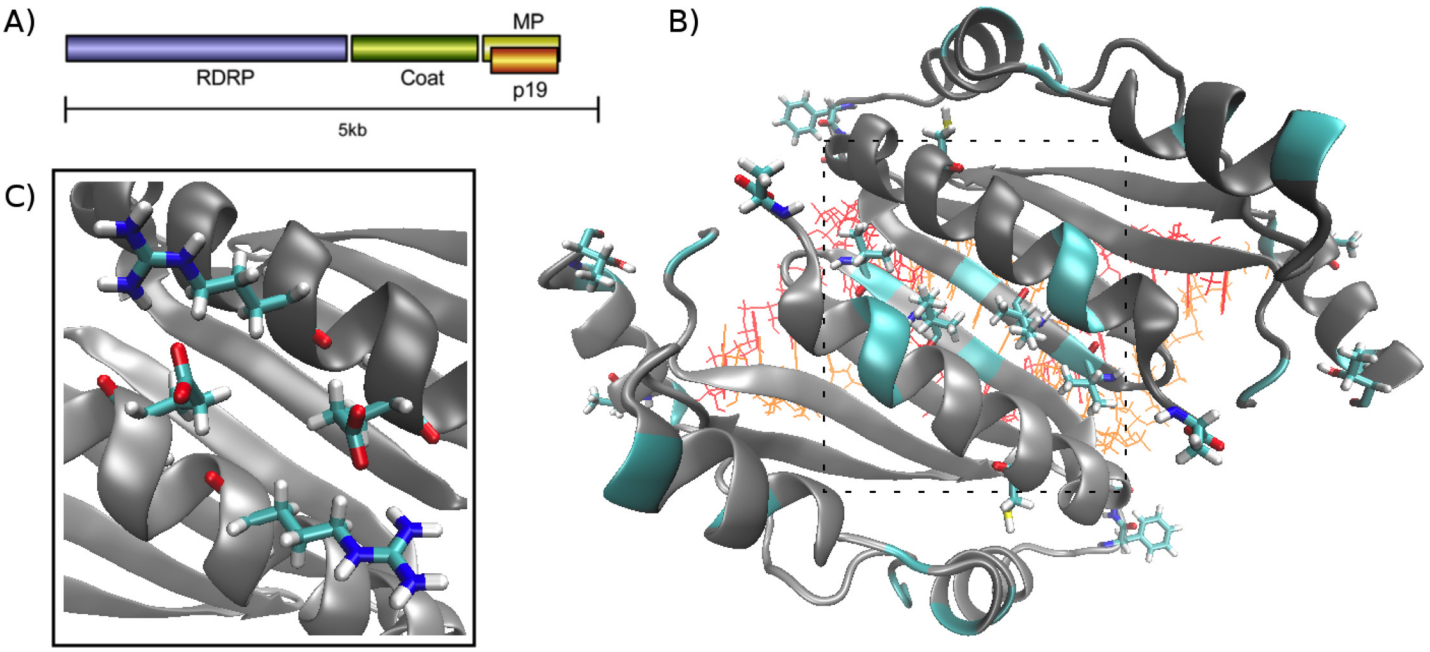

Fig 1. Tombusvirus genome structure and p19 crystal structure. (A) Schematic diagram showing the arrangement of the tombusvirus genome, including the overprinting of p19 on the movement protein (MP). (B) Crystal structure of the tomato bushy stunt (TBS) virus p19 dimer (PDB ID 1R9F) [24] bound to a 19 bp RNA fragment. The protein subunits are drawn in cartoon style and coloured dark and light grey; the two RNA strands are drawn as van der Waals spheres and coloured red and orange. Residues found to be under positive selection in at least two analyses are coloured cyan (cartoon representation), and residues identified as being under positive selection by all four analyses are drawn explicitly and coloured according to atom type (cyan: carbon; red: oxygen; blue: nitrogen; white: hydrogen). The dashed square indicates the region shown in panel C. (C) Close up view of the dimer interface with residues Arg139 and Glu143 of each subunit drawn explicitly.

doi:10.1371/journal.pone.0147619.g001

variants at key interface residues (hereafter termed *observed*) plus any additional permissible mutations of these residues (where where nonsynonymous changes in p19 result in synonymous changes in the overprinted MP protein—hereafter *permissible*). Surprisingly, the majority of *observed* and *permissible* mutations of these sites do not impact dimer formation, even where the *permissible* mutations are expected to be disruptive, and have not been observed in nature. The robustness of the p19 dimeric structure to mutations suggests that while molecular evolutionary analyses support the inference of positive selection, the identification of positively selected sites provides only partial insight into structure and function. It may therefore be non-trivial to disentangle positive selection from sites that are simply robust to non-synonymous substitutions.

Results and Discussion

Sequence analysis suggests p19 may be under positive selection

To identify individual sites putatively under positive selection, we first analyzed available p19 sequences from tombusvirus whole genomes (**Fig 2 and Tables A-H and Figures A-C in S1 File**) using two Maximum Likelihood-based methods: Codeml, from the PAML package [28], and Fixed Effects Likelihood (FEL), implemented in the HyPhy package [20, 29, 30], a likelihood-based analogue to traditional, more conservative site-by-site counting methods. As p19 is overprinted, we also ran the Kaki package, which aims specifically to deal with more complex constraints on sequence evolution [12]. Notably, PAML and HyPhy both indicate that p19 is under positive selection, while Kaki results indicate that, when variable baseline substitution is taken into account, p19 is not under positive selection (**Tables I and J in S1 File**). That said,

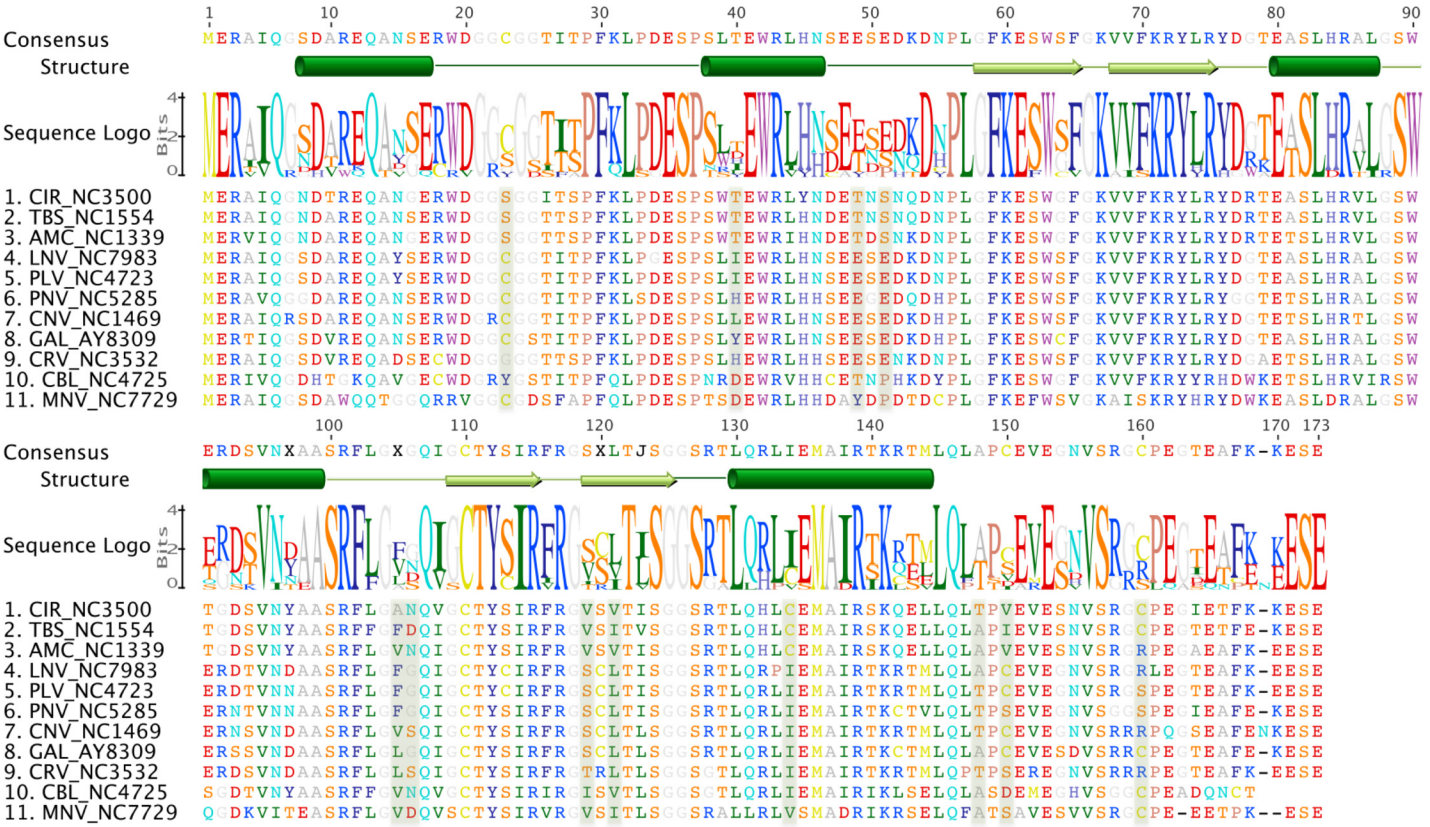

Fig 2. Alignment of amino acid sequences of p19 from 11 different tombusvirus species. The single-letter residue codes are coloured according to the nature of the amino acid side chain: (red) negatively-charged (D,E); (dark blue) aromatic (F,Y); (blue) positively-charged (K, R); (cyan) large polar amide-containing (Q,N); (orange) small polar hydroxyl-containing (S,T); (yellow) sulfur-containing (C,M); (grey) small aliphatic (A,G); (green) medium aliphatic (I,V, L); (purple-blue) imidazole (H); (violet) indole (W); (pink-brown) cyclised secondary amine (P). Sites identified as being under positive selection by all four analyses are highlighted in pale green. The secondary structure elements are indicated above the sequences: (barrels) α-helices; (arrows) β-strands. Viral species and NCBI accession numbers are as follows: CIR (Carnation italian ringspot virus, NC003500), TBS (Tomato bushy stunt virus, NC001554), AMC (Artichoke mottle crinkle virus, NC001339), LNV (Lisianthus necrosis virus, NC007983), PLV (Pear latent virus, NC004723), PNV (Pelagornium necrotic streak virus, NC005285), CNV (Cucumber necrosis virus, NC001469), GAL (Grapevine algerian latent virus, AY830918), CRV (Cymbidium ringspot virus, NC003532), CBL (Cucumber bulgarian latent virus, NC004725), MNV (Maize necrotic streak virus, NC007729).

doi:10.1371/journal.pone.0147619.g002

results from all three methods identified individual sites that are putatively under selection (Fig 3).

Fig 3. Residues identified as being under positive selection. See text for details of each of the four analyses. Light green indicates residues that only some of the analyses identified as being under positive selection. Dark green indicates residues found to be under positive selection in all four analyses. Note that some residues predicted to be under positive selection by HyPhy FEL2 (36, 43, 67, 107, 145) actually show no variation (see Fig 2).

doi:10.1371/journal.pone.0147619.g003

While there is a clear overlap in sites predicted to be under positive selection (Fig 3), each analysis yielded a number of potentially selected sites that were not picked up by one or more of the other methods. In contrast, PAML and HyPhy yielded no sites under positive selection in the MP or polymerase genes, which in fact appear to be under purifying selection (N.B. Kaki was not used here, as the polymerase gene is not overprinted). Our sequence-level analyses therefore suggest that p19 is under positive selection, whereas Kaki would predict this to be a false positive resulting from variable substitution rates. This latter result is expected given that p19 is overprinted on the MP gene.

Location and functional relevance of residues under positive selection

As dN/dS has in some instances been shown to be a poor predictor of functionally important changes [31], we sought to further investigate the functional significance of residues under positive selection. We first examined their location in the tomato bushy stunt (TBS) virus p19 dimer (PDB ID 1R9F), for which a crystal structure has been solved [24] (**Fig 1B**). We found that a significant fraction of sites displaying positive selection are located in coil and loop regions or on the exterior of the protein. While this could suggest that ongoing residue substitutions at these sites serve to reduce detection of p19 by the host, it is not clear whether or how such host detection might occur. It is equally likely that the high variance observed for these residues is simply a result of their location in regions where a broad range of residue types are tolerated.

Only one of the residues identified as being under positive selection (Fig 3) is involved directly in siRNA binding. This residue (position 40) lies at the end of the caliper helix and forms hydrogen bonds to the terminal phosphate of the siRNA. Residue 39, a key component of the RNA "caliper", was also found to be under positive selection in three of our analyses (Fig 3). However, the role of residue 39 is in siRNA size selection [23], and the natural variation at this site may in part be a reflection of the fact that multiple types of amino acids can perform this role [23].

While it is not known whether dimer formation occurs before, after, or concurrently with RNA binding, p19 is certainly an obligate dimer with respect to recognition of siRNAs of the correct length (~19–21 bp). Thus disruption or prevention of dimer formation represents a potential host defense mechanism, so the dimer interface is a likely location for positive selection to occur. Host defence might for example occur via a factor that directly blocks binding of the siRNA to the p19 dimer, or by binding at the dimer interface, thus preventing dimer formation and subsequent siRNA binding. This interface comprises the fourth β-strand (β4), which also lies at the center of the RNA-binding surface, and the fifth α-helix (α5) [24]. Both of these secondary structure elements contain several residues putatively under positive selection (**Fig 1B and Fig 3**).

Mutation of residues in β4 identified as being under positive selection (positions 119, 121 and 123) is unlikely to directly affect dimer formation, as the edge-to-edge linkage of β4 of each subunit to form the β-sheet that spans the dimer interface involves only the polypeptide backbone and is therefore sequence independent. However, highly disruptive mutations, such as to proline or amino acids with bulky or highly charged side chains, could have an indirect effect. Such changes are not observed in known tombusvirus sequences, however (**Figure C in S1 File**). Thus it is unclear whether positive selection of these dimer interface residues is symptomatic of an arms race in which the host is attempting to disrupt the interaction between β-strands across the dimer interface.

In contrast, the side chains of residues in α5 of each subunit interact across the dimer interface. Of particular note are residues 143 and 139, which form a pair of symmetrical salt bridges

or multiple hydrogen bonds, depending on the nature of residue 143, across the dimer interface (Fig 1C). Interestingly, although residue 143 was identified as being under positive selection in three of our four analyses (Fig 3), residue 139 (Arg) is conserved in all known tombusvirus sequences (Fig 2). While two changes in Arg139 (to Trp or Gly) are *permissible* (**Figure C in S1 File**), these are expected to be highly disruptive. The *observed* variation in residue 143 is limited to amino acids that are capable of maintaining the contacts across the dimer interface (Ser/Thr/Glu). This is noteworthy given there is a wider range of *permissible* changes to residue 143 (**Figure C in S1 File**), some of which are predicted to disrupt the interaction with residue 139. This suggests a preference for amino acids capable of forming hydrogen bonds with the invariant Arg139.

From our sequence and structural analyses, it remained unclear whether selection restricts amino acid changes to some optimum subset of arrangements conducive to dimer formation, while negotiating other evolutionary pressures necessitated by an arms race and/or overprinting. We therefore created structures and carried out MD simulations of all *permissible* residue changes at sites 139 and 143, as well as all *observed* tombusvirus p19 sequences, with the aim of establishing whether the changes observed across residue 143 are indeed indicative of positive selection and whether, given that Arg139 is invariant, changes at this site impact p19 function.

Evaluating the impact of dimer interface mutations

To test the effect of all *permissible* changes to residues 139 and 143, we created mutations within the context of the TBS tombusvirus sequence by directly mutating the crystal structure [24] in silico. Each sequence variant, including the wild-type TBS sequence, was simulated as a dimer both with and without a 19 bp siRNA bound. As a control, we simulated all *observed* tombusvirus p19 sequences (Fig 2) homology-modelled onto the TBS p19 crystal structure. This allowed us to test the assumption that other p19 sequences will form a structure similar to that of TBS p19. Importantly, it also allowed the effect of naturally occurring variants of residue 143 to be examined in context, where any potential disruption might be offset by compensatory mutations.

The homology modelling results confirm that the TBS p19 structure is likely to be shared by other known p19 proteins, with the most likely hit being the TBS p19 structure in all cases. Energy-minimisation and MD simulations of these *observed* variants showed them all to be structurally stable in the first instance, with no major loss of secondary, tertiary or quaternary structure in 50 ns of MD simulation with or without a siRNA bound (**Figures D-P in S1 File**). The caliper region shows the greatest flexibility (highest RMSF values), especially without the siRNA bound, as expected. Additionally, the structural stability of the dimer is nearly identical whether or not the siRNA is bound, other than the CBL, MNV and PLV proteins, where there was limited separation of the two subunits without RNA bound, resulting in large RMSD values, indicating that in most cases, RNA binding is not necessary for maintenance of p19 structure. Our assumption that other p19 sequences adopt a similar structure to that of TBS p19 is therefore valid.

To assess whether the *observed* variation in residue 143 requires buffering by compensatory mutations in order to maintain the overall structure, and therefore function, of p19, we examined the *permissible* single mutants of this residue and of the perfectly conserved residue Arg139 in the context of the TBS sequence (designated as wild-type hereafter). In this sequence, the combination of Glu143 and Arg139 maximises hydrogen-bonding capacity and allows for salt bridge formation.

We first assessed the effect of changes to the invariant residue Arg139 to determine whether MD simulations are capable of detecting undesirable substitutions, in this case the two

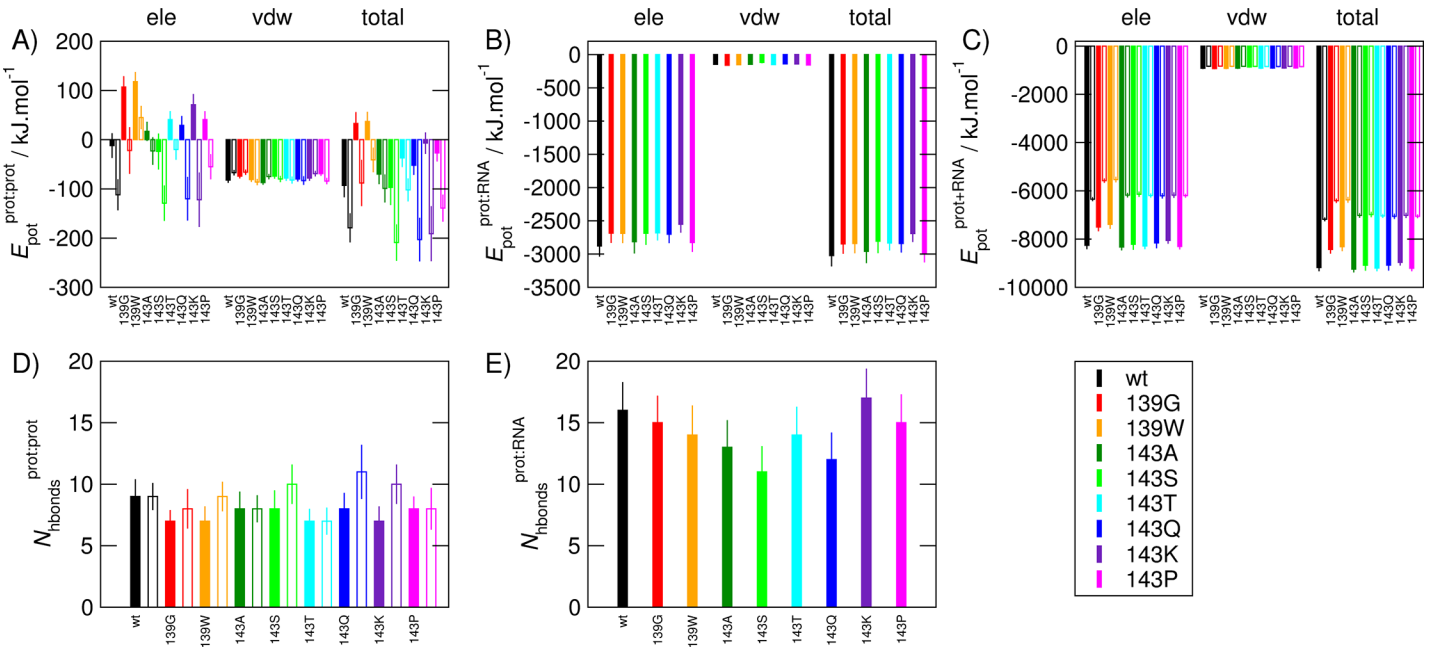

Fig 4. Potential energies and number of hydrogen bonds between different components of simulated p19 systems. Systems comprise wild-type and all *permissible* mutations of p19 alone or in complex with the siRNA. (A) Average potential energies (E_{pot}) of the interaction between the two protein subunits making up the p19 dimer (prot:prot). (B) Average potential energies (E_{pot}) of the interaction between the protein dimer and the RNA (prot;RNA). (C) Average potential energies (E_{pot}) of the complete protein/RNA complex (prot+RNA). The electrostatic and van der Waals contributions are shown separately alongside the total potential energy, as indicated above the graphs. (D) Average number of hydrogen bonds during the entire simulation between the two protein subunits making up the p19 dimer. (E) Average number of hydrogen bonds during the entire simulation between the protein dimer and the RNA. Solid bars correspond to simulations with RNA included, and empty bars to simulations without RNA present. Averages were calculated from the first 50 ns of simulation, after 10 ns equilibration, so that the averages calculated from the simulations of p19 with RNA bound (200 ns) are comparable to those of the simulations without RNA bound (50 ns). The error bars correspond to the standard deviation in all cases.

doi:10.1371/journal.pone.0147619.g004

permissible changes to this residue, both of which are expected to be highly disruptive. Although structurally stable, the Arg139 mutants lose the key salt bridge between residues 139 and 143 across the dimer interface (**Table K and Figure Z in** S1 File) and are the least energetically favourable (Fig 4A–4C), confirming that our simulations were sufficiently sensitive to detect the effects of unfavourable single amino acid changes. Additionally, the energetic effects of both *permissible* changes to Arg139 are much greater than those of even the most extreme changes to residue 143, suggesting that the complete conservation of Arg at position 139 is a result of selection rather than just chance.

Our sequence analyses suggested that residue 143 might be under positive selection (Fig 3), yet of the number of *permissible* substitutions at this position (Glu in the wild-type sequence), only three (Glu, Ser and Arg) are *observed*. Despite this, the overall structure of p19 was exceptionally resistant to changes to residue 143, with the structural properties of the mutants almost indistinguishable from those of the wild-type during the MD simulations (**Figures Q-Z in** S1 File). In almost all cases, the secondary, tertiary and quaternary structures were maintained throughout the simulations, with the caliper region exhibiting the greatest flexibility. One exception was the proline substitution, which, as expected, disrupted α5, providing further evidence that MD simulation is able to detect the effects of point mutations.

The attraction between the negatively charged RNA backbone and the positively-charged residues lining the RNA binding surface of the protein resulted in a large favourable protein-RNA interaction energy in all cases (Fig 4B). However, the protein-protein interaction energy (i.e. between subunits) of all *permissible* mutants is less favourable than that of the wild-type,

and is in fact positive (i.e. unfavourable) when the siRNA is bound (Fig 4A). When the siRNA is not bound, the protein-protein interaction energy improves due to rearrangement of the dimer interface that increases the number and fractional occupancy of hydrogen bonds (Fig 4D) and, in some cases, salt bridges between monomers, whereas with the siRNA bound, only limited formation of alternative interactions is possible (**Table K and Figure Z in S1 File**). This is exemplified by the Glu143Ser mutant, which lost secondary structure both with and without the siRNA bound, but retained a protein-protein interaction energy similar to that of the wild-type due to forming a greater variety of hydrogen bonds (**Fig 4D, 4E and Figures V and Z and Table K in S1 File**). In comparison, in Glu143Gln, the similarity in the size and chemical nature of the side chain of the glutamine to that of the wild-type glutamate results in formation of a number of different interactions with residue 139 that mimic those formed by the wild-type Glu, but are more transient (**Table K and Figure Z in S1 File**). Thus while the *permissible* variants at position 143 are still able to maintain the dimer interface, for the more dramatic changes in the nature of residue 143, the structural rearrangements required to optimise the interactions spanning the interface are less feasible when the siRNA is bound. The extremely favourable electrostatic interactions associated with binding of the siRNA mean that the p19 dimer interface mutants are still functional, however.

Overall, our MD simulations showed p19 to be robust to changes to two key residues at the dimer interface, Arg139, which is highly conserved, and Glu143, suggested by sequence-level analyses and structural examination to be under positive selection. Ideally, the thermodynamic effect of the mutations should be quantified by calculation of binding free energies across the protein-protein interface. Such calculations are infeasible for systems of the size and number investigated here, but the interaction potential energies supported the structural analysis of the simulations. Together, these showed that what may seem to be positive selection at a sequence level and appear to be functionally relevant at a structural level is not necessarily functionally significant once protein dynamics are taken into account. In this particular case, rearrangement of the dimer interface is able to compensate for even highly disruptive mutations.

Concluding Remarks

Detection of evolutionary arms races is of importance for understanding and managing host-pathogen relationships such as the competition between viruses and their plant or animal hosts. Arms races may be manifested as positive selection, which may be detected at a molecular level by the identification of fast-changing nucleotide sequences. Here, we sought to test whether sequence-level analyses are an appropriate means of detecting what is ultimately a phenotype-level effect. For this purpose, we chose a system where positive selection is expected to be occurring, namely the tombusvirus p19 protein, which is involved in suppression of the host plant RNAi response to viral infection. Importantly, the function of this protein depends largely on its structure, thus deconvoluting what is otherwise a complex relationship. Our sequence-level analyses did detect evidence of positive selection at a number of sites, and the structural context of these sites pointed to the dimer interface as being a key region in which positive selection resulting from an arms race might be taking place. We then evaluated the effect of mutations at the dimer interface by carrying out MD simulations of a comprehensive set of sequence variants. We observed a difference in the impact of mutations at a highly conserved residue compared to at a site putatively under positive selection, validating the use of MD to assess the effect of amino acid variation. The simulations of all *permissible* variations (given the constraints of overprinting) at a dimer interface site suggested to be under positive selection by sequence- and structural-level analyses revealed the p19 dimer to be robust to even highly disruptive changes to the dimer interface, calling into question the evolutionary impact

of single mutations in this particular example. We therefore conclude that identifying and assessing the validity of positive selection can be greatly aided by taking protein structure and function into account. Generating mutants and assessing the functional impact on infection under containment would provide an independent means of establishing whether effects not detectable by MD simulations account for the lack of distinction we see between *observed* and *permissible* variants. However, as such experiments are often difficult to undertake, our MD simulations highlight the need for caution in interpreting positive selection detected at sequence level, and permit a more nuanced way of interpreting positive selection signals. In the specific case of p19, we conclude it is difficult to separate tolerance to variation from positive selection.

Methods

Phylogenetic analyses & tests of positive selection

Guide trees for the PAML and FEL analyses were constructed from tombusvirus polymerase sequences (RdRP) using Neighbor-Joining with p-distances (MEGA [32], 500 bootstrap replicates) and ML-based algorithms (PhyML [33], 500 bootstrap replicates) (**Figure A in S1 File**). A posterior probability of 95% under each guide tree was used as criterion for a site to be considered as being under positive selection. The optimal substitution model for the ML tree (GTR +G) was calculated with Modeltest [34]. For the codeml analysis, we compared NSsites models M1 and M7 to M2 and M8, respectively–as recommended by the authors [28]. A likelihood ratio test was conducted to decide which of the two respective models (neutral evolution vs. positive selection) fitted our data best (**Tables C-F in S1 File**). Both guide trees produced equivalent results in codeml with the exception of codon 64, which was hence excluded from further analyses. The FEL analysis was carried out assuming both a 1-rate (dS held constant) and 2-rate model (dS adjusted for each site), respectively [20, 29, 30]. Similar to the codeml analysis, sites were only considered as being under positive selection if they were retrieved by both guide trees. An equivalent analysis was performed for MP genes (**Tables A and B and Figure A in S1 File**). Kaki was run according to the authors' instructions [12] as a means to address the complex constraints on the evolution of the overprinted p19 gene. Briefly, we tested for variability in the baseline substitution rate or whether substitution rate is homogeneous (M8-ρH vs M8-ρV and M8a-ρH vs M8a-ρV). We then tested for positive selection under both the assumption of sequence homogeneity (M8a-ρH vs M8-ρH) and where baseline substitution rate is variable (M8a-ρV vs M8-ρV) to assess whether there is selection at the DNA/RNA level. Results are presented in **Tables I and J in S1 File and Table M in S2 File**.

Molecular dynamics simulations

The list of all *permissible* changes to residues Arg139 and Glu143 of the Tomato bushy stunt (TBS) tombusvirus p19 that do not change the amino acid sequence of the overprinted MP was created taking codon degeneracy into account (**Figure C in S1 File**). In all *permissible* sequence variants, two additional mutations, Leu144Met and Leu147Met, which were in the TBS tombusvirus p19 crystallised by Ye et al. [24] were also present. Initial coordinates for each new sequence were derived from this X-ray structure of p19 with a 19-bp siRNA fragment bound (1R9F) by in silico site-directed mutagenesis using VMD [35]. Initial coordinates for the *observed* tombusvirus p19 sequences were generated by homology modelling using the PHYRE2 Protein Fold Recognition Server in intensive modelling mode [36].

All simulations were carried out using the CHARMM27 all-atom force field [37, 38] and the NAMD software [39]. The lengths of all bonds involving hydrogen atoms were constrained using ShakeH [40] with a tolerance of 1.0×10^{-9} nm, allowing for an integration time step of 2

fs. Van der Waals interactions were smoothed to zero at a cut-off distance of 1.2 nm using a switching function initiated at 1.0 nm. Electrostatic interactions between pairs of atoms separated by one or two bonds were excluded, and those between atoms separated by three bonds were scaled. Long-range electrostatic interactions outside a cut-off distance of 1.2 nm were treated using particle mesh Ewald (PME) [41, 42] with a direct space tolerance of 10^{-6}, interpolation order of 4 and grid spacing of 0.1 nm. The temperature was maintained at 293 K using the Langevin thermostat [43] with a damping coefficient γ of 1 ps^{-1}. A constant pressure of 1.01325 bar was maintained using the Berendsen algorithm [44] with an isothermal compressibility of 2.755×10^{-5} $(kJ\ mol^{-1}\ nm^{-3})^{-1}$, corresponding to a protein in water, and a relaxation time τ_p of 0.5 ps. Periodic boundary conditions were used.

The 1R9F X-ray structure and the initial structure predicted for each new sequence as detailed above, either with or without RNA bound, was energy-minimised in vacuum for 10,000 steps, then solvated in a cubic box of TIP3P water [45] with a minimum distance of 1.4 nm from the protein to the box edge using the VMD solvate package and neutralised by addition of sodium ions using the VMD autoionize package before a second 10,000 steps of energy-minimisation. The numbers of atoms protein, RNA, ion and solvent atoms and the simulation box dimensions are given in Table L in S1 File. The system was heated from 0 K to 293 K at a rate of 1 K every 2 ps (586 ps in total), followed by equilibration for 1.414 ns and finally a data collection phase of 50 ns (200 ns for all *permissible* mutations with RNA bound) during which coordinates were saved every 5 ps (100ps for 200ns simulations).

Analysis of the simulations was carried out using existing VMD packages (namdenergy, saltbr, hbonds) or analysis functions (rmsd, rmsf, sasa, secstruct) coupled to self-made tcl scripts (available upon request).

Supporting Information

S1 File. Supporting results for analyses presented in the main text. p19 dN/dS ratios (**Table A**). Movement Protein (MP) dN/dS ratios (**Table B**). PAML tests for positive selection, p19 (**Tables C and D**). PAML tests for positive selection, MP (**Tables E and F**). Codeml results (**Table G**). HyPhy results (**Table H**). Kaki results (**Tables I and J**). Formation of hydrogen bond and salt bridge interactions across the p19 dimer interface (**Table K**). Numbers of atoms and dimensions of simulation boxes (**Table L**). ML guide tree used for the PAML analysis, based on RDRP sequences from Tombusviruses (**Figure A**). dN/dS scatterplots (**Figure B**). Observed and permissible sequence variation (**Figure C**). Structural stability of all p19 variants studied with and without a 19 bp siRNA bound by molecular dynamics simulations (**Figures D-Y**). Potential energies and dimer interface stability of all *permissible* variants of the wild-type tomato bushy stunt virus p19 sequence with a 19 bp siRNA bound during 200 ns molecular dynamics simulation (**Figure Z**).
(DOC)

S2 File. Results in excel format from analyses using Kaki. Raw results for all four models described in the text are presented **(Table M).**
(XLS)

Acknowledgments

The authors wish to acknowledge the high performance computing facility at the University of Canterbury (UC HPC) for provision of supercomputing facilities, and financial support from the Biomolecular Interaction Centre at the University of Canterbury (JRA, AMP). JRA acknowledges support from a Massey University Research Fund Early Career Grant and a

Marsden Fund Fast Start award (13-MAU-039). AMP acknowledges current support via the Rutherford Discovery Fellowships programme of the Royal Society of New Zealand, and past support from the Swedish Research Council.

Author Contributions

Conceived and designed the experiments: AMP JRA MPH ML. Performed the experiments: JRA MPH ML. Analyzed the data: AMP JRA MPH ML. Wrote the paper: AMP JRA MPH.

References

1. Galiana-Arnoux D, Dostert C, Schneemann A, Hoffmann JA, Imler JL. Essential function in vivo for Dicer-2 in host defense against RNA viruses in drosophila. Nat Immunol. 2006; 7(6):590–7. PMID: 16554838.
2. Hamilton AJ, Baulcombe DC. A species of small antisense RNA in posttranscriptional gene silencing in plants. Science. 1999; 286(5441):950–2. PMID: 10542148.
3. Lu R, Maduro M, Li F, Li HW, Broitman-Maduro G, Li WX, et al. Animal virus replication and RNAi-mediated antiviral silencing in Caenorhabditis elegans. Nature. 2005; 436(7053):1040–3. PMID: 16107851.
4. Wang XH, Aliyari R, Li WX, Li HW, Kim K, Carthew R, et al. RNA interference directs innate immunity against viruses in adult Drosophila. Science. 2006; 312(5772):452–4. PMID: 16556799.
5. Zambon RA, Vakharia VN, Wu LP. RNAi is an antiviral immune response against a dsRNA virus in Drosophila melanogaster. Cell Microbiol. 2006; 8(5):880–9. PMID: 16611236.
6. Shabalina SA, Koonin EV. Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol. 2008; 23(10):578–87. PMID: 18715673. doi: 10.1016/j.tree.2008.06.005
7. Li F, Ding SW. Virus counterdefense: diverse strategies for evading the RNA-silencing immunity. Annu Rev Microbiol. 2006; 60:503–31. PMID: 16768647.
8. Ruiz-Ferrer V, Voinnet O. Viral suppression of RNA silencing: 2b wins the Golden Fleece by defeating Argonaute. Bioessays. 2007; 29(4):319–23. PMID: 17373696.
9. Incarbone M, Dunoyer P. RNA silencing and its suppression: novel insights from in planta analyses. Trends in plant science. 2013; 18(7):382–92. Epub 2013/05/21. doi: 10.1016/j.tplants.2013.04.001 PMID: 23684690.
10. Obbard DJ, Jiggins FM, Halligan DL, Little TJ. Natural selection drives extremely rapid evolution in antiviral RNAi genes. Curr Biol. 2006; 16(6):580–5. PMID: 16546082.
11. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol. 2012; 29(12):3767–80. Epub 2012/07/24. doi: 10.1093/molbev/mss179 PMID: 22821011; PubMed Central PMCID: PMC3494269.
12. Rubinstein ND, Doron-Faigenboim A, Mayrose I, Pupko T. Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. Mol Biol Evol. 2011; 28(12):3297–308. Epub 2011/06/22. doi: 10.1093/molbev/msr162 PMID: 21690564.
13. Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature reviews Genetics. 2006; 7(2):98–108. Epub 2006/01/19. doi: 10.1038/nrg1770 PMID: 16418745.
14. Hughes AL, Westover K, da Silva J, O'Connor DH, Watkins DI. Simultaneous positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. J Virol. 2001; 75(17):7966–72. PMID: 11483741.
15. Keese PK, Gibbs A. Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci U S A. 1992; 89(20):9489–93. PMID: 1329098.
16. Keese P, Gibbs A. Plant viruses: master explorers of evolutionary space. Curr Opin Genet Dev. 1993; 3 (6):873–7. PMID: 8118212.
17. Firth AE, Brown CM. Detecting overlapping coding sequences in virus genomes. BMC bioinformatics. 2006; 7:75. Epub 2006/02/18. doi: 10.1186/1471-2105-7-75 PMID: 16483358; PubMed Central PMCID: PMC1395342.
18. Belshaw R, Pybus OG, Rambaut A. The evolution of genome compression and genomic novelty in RNA viruses. Genome research. 2007; 17(10):1496–504. Epub 2007/09/06. doi: 10.1101/gr.6305707 PMID: 17785537; PubMed Central PMCID: PMC1987338.
19. Suzuki Y, Nei M. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom Thalassiosira weissflogii and the tax gene of a human T-cell lymphotropic virus. Mol Biol Evol. 2004; 21(5):914–21. Epub 2004/03/12. doi: 10.1093/molbev/msh098 PMID: 15014169.

20. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005; 22(5):1208–22. PMID: 15703242.

21. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005; 22(12):2472–9. PMID: 16107592.

22. Scholthof HB. The Tombusvirus-encoded P19: from irrelevance to elegance. Nat Rev Microbiol. 2006; 4(5):405–11. PMID: 16518419.

23. Vargason JM, Szittya G, Burgyan J, Tanaka Hall TM. Size selective recognition of siRNA by an RNA silencing suppressor. Cell. 2003; 115(7):799–811. PMID: 14697199.

24. Ye K, Malinina L, Patel DJ. Recognition of small interfering RNA by a viral suppressor of RNA silencing. Nature. 2003; 426(6968):874–8. PMID: 14661029.

25. Dunoyer P, Schott G, Himber C, Meyer D, Takeda A, Carrington JC, et al. Small RNA duplexes function as mobile silencing signals between plant cells. Science. 2010; 328(5980):912–6. Epub 2010/04/24. doi: 10.1126/science.1185880 PMID: 20413458.

26. Schott G, Mari-Ordonez A, Himber C, Alioua A, Voinnet O, Dunoyer P. Differential effects of viral silencing suppressors on siRNA and miRNA loading support the existence of two distinct cellular pools of ARGONAUTE1. The EMBO journal. 2012; 31(11):2553–65. Epub 2012/04/26. doi: 10.1038/emboj.2012.92 PMID: 22531783; PubMed Central PMCID: PMC3365429.

27. Merai Z, Kerenyi Z, Molnar A, Barta E, Valoczi A, Bisztray G, et al. Aureusvirus P14 is an efficient RNA silencing suppressor that binds double-stranded RNAs without size specificity. J Virol. 2005; 79 (11):7217–26. PMID: 15890960.

28. Yang ZH. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997; 13:555–6. PMID: 9367129

29. Pond SK, Muse SV. Site-to-site variation of synonymous substitution rates. Mol Biol Evol. 2005; 22 (12):2375–85. Epub 2005/08/19. doi: 10.1093/molbev/msi232 PMID: 16107593.

30. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005; 21 (5):676–9. Epub 2004/10/29. doi: 10.1093/bioinformatics/bti079 PMID: 15509596.

31. Nozawa M, Suzuki Y, Nei M. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci U S A. 2009; 106(16):6700–5. Epub 2009/04/03. doi: 10.1073/pnas.0901855106 PMID: 19339501; PubMed Central PMCID: PMC2672471.

32. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 2007; 24(8):1596–9. Epub 2007/05/10. doi: 10.1093/molbev/msm092 PMID: 17488738.

33. Guindon S, Delsuc F, Dufayard JF, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 2009; 537:113–37. Epub 2009/04/21. doi: 10.1007/978-1-59745-251-9_6 PMID: 19378142.

34. Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics. 1998; 14 (9):817–8. Epub 1999/01/27. PMID: 9918953.

35. Humphrey W, Dalke A, Schulten K. VMD—Visual Molecular Dynamics. J Molec Graphics. 1996; 14:33–8.

36. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nature protocols. 2015; 10(6):845–58. Epub 2015/05/08. doi: 10.1038/nprot.2015.053 PMID: 25950237.

37. Klauda JB, Brooks BR, MacKerell ADJ, Venable RM, Pastor RW. An ab initio study on the torsional surface of alkanes and its effect on molecular simulations of alkanes and a DPPC bilayer. J Phys Chem B. 2005; 109(11):5300–11. PMID: 16863197

38. Feller SE, MacKerell ADJ. An improved empirical potential energy function for molecular simulations of phospholipids. J Phys Chem B. 2000; 104(31):7510–5.

39. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. Journal of Computational Chemistry. 2005; 26:1781–802. PMID: 16222654

40. Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys. 1977; 23:327–41.

41. Darden TA, York DM, Pedersen LG. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. The Journal of Chemical Physics. 1993; 98:10089.

42. Essman E, Perera L, Berkowitz ML, Darden TA, Lee H, Pedersen LG. A smooth particle mesh Ewald method. The Journal of Chemical Physics. 1995; 103:8577–93.

43. Izaguirre JA, Catarello DP, Wozniak JM, Skeel RD. Langevin stabilization of molecular dynamics. The Journal of Chemical Physics. 2001; 114:2090–8.

44. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics. 1984; 81(8):3684–90.

45. Jorgensen WL, Chandrasekhar J, Madura JD. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics. 1983; 79:926.