

Identification of evolutionarily conserved nucleic acid secondary structures within viral genomes

by

MICHAEL THOMAS GOLDEN
STUDENT NUMBER: GLDMIC011

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN
In fulfilment of the requirements for the degree

MSc (Med) Bioinformatics

Faculty of Health Sciences
UNIVERSITY OF CAPE TOWN

Date of Submission

23 December 2013

Supervisor

Dr. Darren Martin

Computational Biology Group, Department of Clinical Laboratory Sciences, University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

DECLARATION

I, Michael Golden, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature:

Signed by candidate

Date: 03 March 2014 (Date of correction)

University of Cape Town

Abstract

Large single-stranded nucleotide sequences such as certain DNA and RNA virus genomes form complex secondary structures through Watson-Crick base-pairing interactions between bases on the same DNA or RNA strand. Both computational and experimental methods are available for determining the secondary structure conformations of nucleotide sequences and have been used to identify many thousands of structural elements within these sequences. However, to date only a small subset of these elements have been shown to have any biological functionality.

In this study, a software tool enabling the graphical visualisation of structural elements overlaid with additional biological information was developed in order to assist researchers in identifying elements with likely biological functionality within larger secondary structures. Additionally, a tool was developed for identifying conserved structural elements amongst a set of related sequences. Using these tools a number of conserved elements were identified within computationally predicted secondary structures of HIV and five flavivirus species. Further evidence of selection having acted to preserve the identified elements, in support of their likely biological importance, was provided by evolutionary analysis of related sequences.

In addition to the identification of individual structural elements, a large-scale analysis of HIV-1 and seven flavivirus species revealed statistically significant associations between base-paired nucleotide positions and various patterns of mutation. While the results of the various analyses carried out were consistent with the presence of between one and five highly-conserved functionally important secondary structures within the various genomes analysed, they were also consistent with the occurrence of many less-conserved (or only transiently conserved) structures that potentially have no specific biological functions. Although many of these less-conserved structures may, as has previously been suggested, defend these genomes against intra-cellular anti-viral siRNAs, a strong association was found between sites that are base-paired and codon sites that are apparently evolving under purifying selection: a finding which suggests that an additional or alternative role of these structures may simply be to reduce basal mutation frequencies, and hence amino acid substitution rates at codon sites that encode functionally important amino acids.

An investigation of HIV-1 natural inter- and intra-subtype recombinant sequences revealed that recombination breakpoints had a significant tendency to occur more frequently at base-paired nucleotide positions within the computationally predicted secondary structures of these sequences. Furthermore, it was found that these natural HIV-1 recombinants were less disruptive of protein tertiary structure and RNA secondary structure conformations than expected by chance: suggesting that natural selection favouring the maintenance of both protein and nucleic acid folding has likely had a major influence on the patterns of recombination that are detectable within HIV genomes.

Finally, a general method for predicting the secondary structures of a set of related sequences sharing similar secondary structure conformations was developed. By taking alignment uncertainty into account through the analysis of large numbers of similarly plausible multiple sequence alignments that are sampled using a statistical alignment procedure, the method was found to have better predictive accuracy than existing related RNA secondary structure prediction methods.

Acknowledgments

I would like to thank the many people who have helped me through the completion of this Master's thesis. The first is my supervisor, Dr. Darren Martin, for the patient guidance and mentorship he has provided to me from the start of my Honours year all the way through to the completion of my Masters degree. Darren has not only given me invaluable assistance with my research, but also encouraged me to have a broader understanding of the scientific endeavor. As a result of Darren's kind and generous nature I have had numerous opportunities made available to me and I consider myself extremely fortunate to have had him as a supervisor.

Many thanks go to the members of the Computational Biology group at the University Cape Town. In particular, I would like to thank Brejnev Muhire, Yves Semegni, Emil Tanov and Ben Murrell who I worked closely with on many joint investigations and provided great entertainment and many useful discussions over the years. A special thank you goes to the head of the Computational Biology group, Associate Professor Nicola Mulder, whose kind and quiet encouragement helped me choose to do my postgraduate studies in bioinformatics.

It is important that I acknowledge the work presented in chapter five which was carried out at the Oxford Summer School in Computational Biology 2012 (OSSCB12). This work was part of a six week project which I worked as a student alongside four hard-working and dedicated students and supervisors, namely: Preeti Arunapuram, Ingólfur Eðvarðsson, Ádam Novák, and James Anderson.

Computations for this thesis were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team (<http://hpc.uct.ac.za>). The High Performance Computing team was very helpful in dealing with all of my queries and I found their systems to be exceptionally reliable.

I would like to thank the National Research Foundation, administered by UCT's postgraduate funding office, and the Poliomyelitis Research Foundation for funding my studies. This funding has not only allowed me carry out my work, but also has given me the opportunity to attend a number of international conferences and courses which has added immense value to my Master's studies.

Finally, I would like to thank my family and friends whose support and encouragement over the past few years has made the completion of this Master's thesis possible.

Table of Contents

1	Introduction.....	1
	Identification of nucleic acid secondary structures.....	2
	The impact of genomic secondary structure on nucleotide substitution dynamics.....	5
	The impacts of genomic secondary structures on recombination.....	7
	Aims	9
	Candidate's contributions.....	10
2	Tool development.....	11
	Introduction	11
	Methods.....	15
	Nucleic Acid Visualisation and Analysis.....	15
	Substructure conservation analysis.....	29
	Results.....	36
	Nucleic Acid Visualisation and Analysis application.....	36
	Detection of conserved substructures in HIV-1	43
	Conclusions.....	45
3	Biologically relevant secondary structures in single-stranded RNA viruses.....	47
	Introduction	47
	Methods.....	48
	Dataset construction	48
	Prediction of nucleic acid secondary structures	49
	Inference of nucleotide substitution rates	49
	Inference of synonymous substitution rates	49
	Inference of non-synonymous substitution rates	51
	Inference of nucleotide coevolution	52
	Paired sites association test.....	54
	Tests of synonymous and non-synonymous substitution rates at paired and unpaired codon positions	55
	Paired sites correlation test.....	56
	Testing for coevolution of base-paired sites.....	58
	Detection of conserved nucleic acid substructures	59
	Large-scale investigation of nucleic acid secondary structures	59
	Detailed investigation of nucleic acid secondary structures.....	60

Results and Discussion	61
Large-scale analysis of secondary structures in various RNA viruses.....	61
Detailed analysis of conserved substructures in various RNA viruses	72
Conclusions.....	82
4 Protein and RNA structure is associated with recombination in HIV-1.....	84
Introduction	84
Methods.....	86
Dataset construction	86
Recombination Analysis	87
Simulating recombinants.....	87
Protein folding disruption tests.....	88
Nucleic acid fold-disruption tests	89
Tests for associations between recombination breakpoints and structured genome regions	90
Results and Discussion	91
Avoidance of protein fold disruption in HIV-1 recombinants	91
Co-localisation of recombination breakpoint positions and RNA secondary structures.....	94
Some evidence of RNA structure disruption in HIV-1M recombinants.....	96
Conclusions.....	99
5 Incorporating alignment uncertainty into RNA secondary structure prediction.....	100
Introduction	100
Methods.....	102
Structure prediction from large numbers of multiple sequence alignments	102
Reliability measure of secondary structure prediction accuracy	106
Results and Discussion	106
Predicting secondary structure from large numbers of multiple sequence alignments.....	106
A novel reliability measure of secondary structure prediction accuracy.....	109
Conclusions.....	110
6 Summary.....	111
7 References.....	115
8 Supplementary material.....	124

List of Figures

Figure 1.1: RNA secondary and tertiary structures	2
Figure 1.2: Pseudoknotted RNA secondary structure.....	3
Figure 2.1: Layout produced by VARNA for a ~9,600nt long Hepatitis C Virus genome secondary structure.....	12
Figure 2.2: Screenshot of the RRE substructure in DOOSS.....	13
Figure 2.3: NAVA data input and visualisation tabs.....	15
Figure 2.4: Flowchart describing NAVA workflow	17
Figure 2.5: NAVA project navigation panel.....	18
Figure 2.6: The "Add a structure overlay" dialog.....	22
Figure 2.7: The "Create substructure list" dialog.....	23
Figure 2.8: The "Add a 1-dimensional data overlay" dialog.....	24
Figure 2.9: The "Add a 2-dimensional data overlay" dialog.....	25
Figure 2.10: A graphical depiction of the NL4-3 HIV-1 genome annotations in the NAVA annotations panel	26
Figure 2.11: Examples of NAVA overlays demonstrated for a HIV-1 substructure	27
Figure 2.12: Dot bracket representation for nucleic acid secondary structures	29
Figure 2.13: Mountain representation for secondary structures	31
Figure 2.14: Mountain similarity calculation	31
Figure 2.15: Importing data files	37
Figure 2.16: Automatic data type detection.....	37
Figure 2.17: NAVA applications	38
Figure 2.18: Substructure visualisation.....	40
Figure 2.19: Paired sites comparison ranking.....	42
Figure 2.20: Substructure ranking.....	42
Figure 2.21: HIV-1 and HIV-2 genomic RNA substructure conservation analysis.....	44
Figure 3.1: Constraints imposed by the selective maintenance of a secondary structure on synonymous codon substitutions.....	51
Figure 3.2: Muse 95 paired character model.....	52
Figure 3.3: Example demonstrating the "Paired sites association test".....	55
Figure 3.4: Detecting correlations between base-paired sites.....	56
Figure 3.5: Summary of conserved substructures detected across HIV subtypes	78
Figure 3.6: Summary of conserved substructures detected amongst several HCV genotypes.....	79
Figure 3.7: Summary of conserved substructures detected amongst Dengue serotypes 1 - 4	80
Figure 3.8: Summary of conserved substructures detected between West Nile Virus, Japanese encephalitis virus and tick-borne encephalitis virus.....	81
Figure 4.1: Diagrammatic representation of generating simulated recombinants.....	88
Figure 4.2: The predicted sensitivity of HIV-1M proteins to recombinational disruption	92
Figure 5.1: Sampling alignments with StatAlign.....	103

Figure 5.2: Correlation of two reliability scores	109
Supplementary Figure 1: Example of how a data overlay mapping alignment and a structure overlay mapping alignment are used to create a mapping between a data overlay and a structure overlay	124
Supplementary Figure 2: Distributions of distances between base-paired nucleotides in different viruses.....	125
Supplementary Figure 3: Substructure visualisations for conserved substructures in Human Immunodeficiency Virus page 1.....	126
Supplementary Figure 4: Substructure visualisations for conserved substructures in Human Immunodeficiency Virus page 2.....	127
Supplementary Figure 5: Substructure visualisations for conserved substructures in Human Immunodeficiency Virus page 3.....	128
Supplementary Figure 6: Substructure visualisations for conserved substructures Hepatitis C Virus page 1	129
Supplementary Figure 7: Substructure visualisations for conserved substructures Hepatitis C Virus page 2	130
Supplementary Figure 8: Substructure visualisations for conserved substructures in Dengue virus page 1.....	131
Supplementary Figure 9: Substructure visualisations for conserved substructures in Dengue virus page 2.....	132
Supplementary Figure 10: Substructure visualisations for conserved substructures in West Nile Virus, Japanese Encephalitis Virus and Tick-borne Encephalitis Virus page 1	133
Supplementary Figure 11: Substructure visualisations for conserved substructures in West Nile Virus, Japanese Encephalitis Virus and Tick-borne Encephalitis Virus page 2	134

List of Tables

Table 2.1: Input and output file formats supported by NAVA for different data types	19
Table 2.2: Currently available NAVA applications	20
Table 3.1: Analysis of nucleotide substitution rates at paired and unpaired nucleotide positions in various viral coding sequences.....	62
Table 3.2: Analysis of both synonymous substitution rates and non- synonymous substitution rates at paired and unpaired codon positions in various viral coding sequences.....	65
Table 3.3: Table of different viruses and their synonymous substitution rate correlations.....	70
Table 3.4: Contingency tables and corresponding chi-squared association tests for genome-wide paired and coevolving nucleotide pairs in four viruses.	71
Table 4.1: Degrees of protein fold disruption in natural and simulated HIV-1 recombinants.....	93
Table 4.2: Breakpoint position tests performed on the full-genome recombination dataset	95
Table 4.3: Secondary structure fold-disruption tests performed on the full-genome recombination dataset.....	97
Table 4.4: Secondary structure fold-disruption tests performed on the coding-region recombination dataset.....	98
Table 5.1: Average F-scores for RNA secondary structure prediction benchmarks on 44 re-aligned RFam alignments	107
Supplementary Table 1: Top 20 out of 422 HIV-1 SHAPE structures with lower than expected SHAPE reactivities..	124
Supplementary Table 2: List of PDB structure files used in the recombination protein-fold disruption test.....	136
Supplementary Table 3: Degrees of protein fold disruption in natural and simulated HIV-1 recombinants.....	136
Supplementary Table 4: Breakpoint position tests for the coding-region recombination dataset	136

1 Introduction

As with their cellular counterparts, the primary sequences of virus genomes encode all the genetic information required for their survival. Unlike their cellular counterparts, however, virus genomes can consist either of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Although DNA is ordinarily thought of as a double-stranded molecule forming the now iconic double helical configuration (Watson and Crick, 1953), many viral genomes consist entirely of single-stranded DNA or RNA molecules. Such single-stranded nucleic acid molecules are far less constrained than double-stranded nucleic acid molecules in the variety of structures that they can form.

The structures that nucleic acid molecules form are commonly referred to as their secondary or tertiary structures. Whereas secondary structure is defined as the set of hydrogen bonding interactions between the constituent bases of a nucleic acid molecule, tertiary structure is defined as the arrangement of the nucleic acid molecule's constituent atoms in three-dimensional space (Figure 1.1). The most trivial example of a secondary structure is the base-pairing interactions between complementary bases in DNA double helices. The more interesting cases of secondary structure are those that form in single-stranded nucleic acid molecules due to the increased complexity of structures that can form in these molecules in the absence of reverse complementary sequences that constrain them to form double-stranded helices. As a result of their increased complexity the secondary structures of single-stranded nucleic acid molecules can have functions other than just information storage or transfer. For example, the Rev response element (RRE) which occurs in the genomes of all known HIVs plays a crucial role in the regulation of HIV virion expression by binding the HIV Rev protein which allows transfer of HIV genomes from the nucleus to the cytoplasm where translation and virion packaging occur (Heaphy *et al.*, 1990; Daugherty *et al.*, 2010).

Examples of tertiary structures that form in double-stranded DNA are the A-, B- and Z-conformations of DNA double helices (Watson and Crick, 1953; Mitsui *et al.*, 1970; Wang, 1979; Wang *et al.*, 1979). Ideally, when studying the function of a particular nucleic acid molecule one would want to determine its tertiary structure since this would contain three-dimensional coordinates of each atom and therefore include complete information about its secondary structure. Unfortunately, accurate tertiary structure models of nucleic acid molecules are not yet widely available for nucleic acid molecules as large as those of virus genomes as these are both difficult to infer both computationally and difficult to determine experimentally. For this reason, my thesis will solely focus on a far more tractable problem: the computational inference and analysis of virus genome-scale nucleic acid secondary structures.

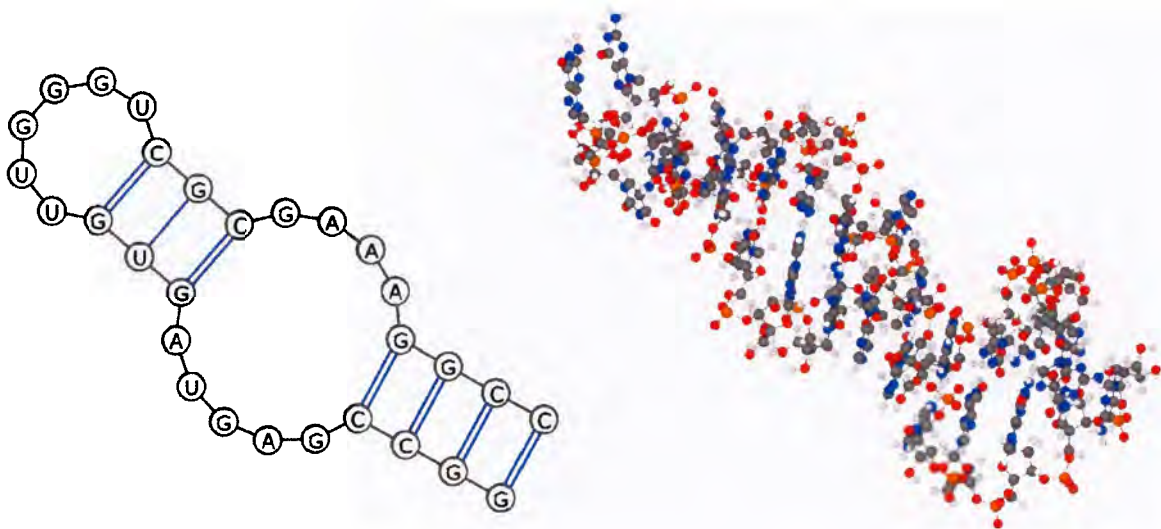


Figure 1.1: RNA secondary and tertiary structures. On the left is a graphical representation of the HCV-1B IRES domain IIID minimum free energy secondary structure predicted using UNAFold (Markham and Zuker, 2008) and visualised using VARNA (Darty *et al.*, 2009). Hydrogen bonding interactions are indicated using blue lines joining pairs of nucleotides. On the right is a tertiary structure representation of the same RNA molecule obtained from the Protein Data Bank (PDB ID 1F84; Lukavsky *et al.*, 2000) and visualised using Jmol (The Jmol Team, 2007), providing a graphical representation of the three-dimensional co-ordinates of each atom in the molecule.

Identification of nucleic acid secondary structures

Computational prediction of nucleic acid secondary structures is an important problem. With the advent of next generation sequencing and RNA-seq technologies, many RNA structural elements are being found to play important roles in the regulation of gene expression (Heaphy *et al.*, 1990; Mandal and Breaker, 2004). While high-throughput methodologies currently do not exist for the large-scale experimental determination of the nucleic acid secondary structures, an alternative is the use of computational secondary structure prediction methods (Markham and Zuker, 2008; Sükösd *et al.*, 2011; Bernhart *et al.*, 2008; Semegni *et al.*, 2011). These methods typically involve free energy minimization of experimentally-derived hydrogen bonding thermodynamic parameters between base-pairs within a nucleotide sequence (Mathews and Turner, 2006). The secondary structure with the lowest free energy is considered to be the most probable structure at equilibrium. For a single sequence, however, there may be multiple structures with free energies close to the minimum and even different free energy minima sharing similar free energies as the global minimum. For this reason it is possible that there exist multiple highly likely structural conformations a nucleic acid molecule can form at equilibrium. Despite this, however, it has been recognized that, for a given sequence, the structures within the set of most probable conformations typically closely resemble one another in that they share a significant proportion of base-pairings (Mathews, 2006).

Another major limitation of most free energy minimization techniques is that they are unable to predict a class of RNA secondary structures termed pseudoknots. Pseudoknots are any combination of non-nested base-pairings in a secondary structure (Figure 1.2, Rivas and Eddy, 1999, 2000; Staple and Butcher, 2005). Even without considering pseudoknots there is still much room for the improvement of computational approaches to secondary structure inference. For this reason approaches such as co-transcriptional RNA folding (Xayaphoummine *et al.*, 2005; Proctor and Meyer, 2013), the use of evolutionary information (Knudsen and Hein, 2003; Do *et al.*, 2006; Sükösd *et al.*, 2011) and co-estimation of secondary structures and alignments (Mathews and Turner, 2002; Siebert and Backofen, 2005) have been developed in order to improve the accuracy of structure prediction. Primary among these are comparative structure prediction approaches, such as that implemented in the computer programs PPfold (Sükösd *et al.*, 2011), NASP (Semegni *et al.*, 2010) and RNAalifold (Bernhart *et al.*, 2008), which utilise the information present in multiple homologous sequences to predict a consensus structure. To further increase the accuracy of secondary structure predictions for individual sequences, approaches such as selective 2'-hydroxyl acylation analyzed by primer extension (Wilkinson *et al.*, 2006) use experimental base-pairing information to constrain thermodynamics-based computational folding algorithms.

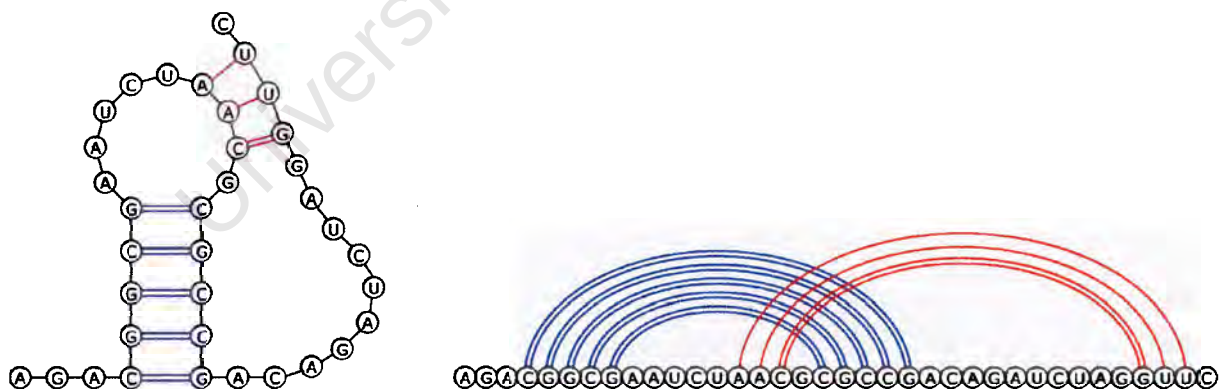


Figure 1.2: Pseudoknotted RNA secondary structure. These two graphic structure representations both represent the same pseudoknotted RNA secondary structure. Single and double lines joining pairs of nucleotides indicate base-pairing interactions. The fact that this structure is pseudoknotted is not entirely obvious from the graphical representation on the left, however, you will notice that in the representation on the right that the base-pairing interactions coloured in blue overlap with the base-pairing interactions coloured in red, this demonstrates that this secondary structure contains non-nested base-pairings. The fact that this secondary structure contains non-nested base-pairings implies that it is a pseudoknotted structure.

Initial investigations into the secondary structures of virus genomes have focused on the computational identification, experimental verification and functional analysis of individual structural elements that typically occur in the non-coding regions of viruses with single-stranded RNA genomes. Numerous such structures have been discovered and functionally characterized in this way and many more continue to be discovered each year using this same approach. The secondary structures thus discovered have a number of different biological functions including for example, internal ribosomal entry sites (IRESs) in a number viruses (Tsukiyama-Kohara *et al.*, 1992; Mokrejs *et al.*, 2010) and secondary structures involved in the promotion of genome dimerization (Darlix *et al.*, 1990; Bieth *et al.*, 1990; Moore and Hu, 2009).

More recent efforts have focused on the identification of the broad range of structures within entire viral genomes and have identified both conserved and non-conserved structures that are biologically relevant. A recent study (Davis *et al.*, 2008) used free energy minimization techniques to fold RNA sequences and detected an association between the amount of structure within a viral genome and viral persistence in various positive-strand RNA viruses. Viruses which contained higher degrees of ordered structure, referred to as genome-scale ordered RNA (GORS), were found to have increased viral persistence in comparison to those which did not. The mechanism by which GORS affects viral persistence is not yet well understood, but one possibility is that GORS sequesters nucleotides within genomic RNA preventing these from interacting with host small interfering RNAs (siRNAs; Simmonds *et al.*, 2004). siRNAs operate by binding to complementary RNA sequences within the viral genome, targeting the sequences for cleavage by host factors (Haasnoot *et al.*, 2007; Nielsen *et al.*, 2005). The role of genomic secondary structures in the avoidance of siRNA binding may be clinically relevant, because siRNAs have potential applications as antiviral therapeutics (Ding *et al.*, 2004; Tafer *et al.*, 2008). There is, however, evidence both of siRNA binding being affected by secondary structures at target sites (Westerhout *et al.*, 2005), and of viruses developing resistance to siRNAs through the evolution of alternative genomic secondary structures that sequester siRNA target sites (Tafer *et al.*, 2008). For these and other reasons accurately determining the secondary structures of virus genomes is crucial in efforts to both understand the functions of secondary structure and develop siRNA-based anti-viral therapies.

A noteworthy example of where the entire genomic secondary structure of a virus has been considered is the HIV-1M subtype B NL4-3 isolate (Watts *et al.*, 2009). Rather than rely solely on computational techniques for the determination of secondary structural elements within the viral genome, the SHAPE approach (Selective 2'-hydroxyl acylation analyzed by primer extension; Wilkinson *et al.*, 2006) was used. This hybrid computational-experimental approach used a method analogous to DNA sequencing to differentiate chemically modified unpaired

nucleotides from paired nucleotides (Wilkinson *et al.*, 2008). This so-called “SHAPE chemistry” was applied to gently purified genomic RNA thus maintaining the structural configuration of the molecule. After chemical modification of unpaired nucleotides, reverse transcription was used to detect modifications to the RNA allowing paired and unpaired nucleotides to be mapped with a large degree of accuracy. Nucleotides with a high SHAPE reactivity are considered more likely to be unpaired, whereas nucleotides with a low SHAPE reactivity are considered more likely to be paired. The SHAPE reactivity information was then used to constrain a thermodynamic folding algorithm, which allowed the generation of a secondary structure model which was reflective of the experimental data. Using the HIV-1M NL4-3 secondary structure model it was found that nucleotide sequences encoding both inter- and intra-protein domain boundaries were more structured than randomly selected regions within the genome. It was then experimentally demonstrated that these structures can cause ribosomal pausing which enables separate protein domains to be co-translationally folded independently of one another (Watts *et al.*, 2009).

The impact of genomic secondary structure on nucleotide substitution dynamics

Given that many of the secondary structures which are present in the genomes of single-stranded RNA and DNA viruses are known to have specific biologically relevant functions, one would expect that these structures might detectably impact patterns of evolution in these viruses. For example, a number of functionally important secondary structures occurring in the coding regions of HIV-1 group M sequences (such as the s14 stem-loop required for viral encapsidation and the Rev Response Element) correspond to regions of RNA which have very low synonymous substitution rates (Ngandu *et al.*, 2008; Sanjuán and Bordería, 2011). A plausible explanation for this is that nucleotide sequences within coding regions that are also within functionally important genomic secondary structural elements are constrained to evolve under two competing selection pressures: One favouring the maintenance of the amino acid sequence of the encoded protein(s) and the other favouring the maintenance of the structural element. At many such sites the only substitutions that will have no adverse effects on viral fitness would be those that did not affect either the encoded amino acid sequence or the secondary structural element. Besides the occurrence of biologically functional secondary structural elements within coding regions, a number of other factors might also result in the suppression in particular genome regions of synonymous substitution rates. Most important amongst these are overlapping open reading frames and conserved protein binding motifs (Plotkin and Kudla, 2011). Despite the fact that for any particular codon there are various possible causes of synonymous substitution rates being lower than those expected under

neutral evolution, analysing these rates remains a useful means of identifying and evaluating the potential biological significance of inferred genomic secondary structures (Tuplin *et al.*, 2004; Simmonds *et al.*, 2008; Budkowska *et al.*, 2011; Sanjuán and Bordería, 2011).

Another evolutionary signal that can be analysed so as to identify selectively maintained genomic secondary structures is the coevolution that is expected to occur between the base-paired nucleotides within these structures (Eddy and Durbin, 1994; Tuplin *et al.*, 2002, 2004; Cheng *et al.*, 2012). Many pairs of nucleotides within RNA molecules exhibit evidence of complementary coevolution such that whenever a substitution occurs in one partner of the pair, a complementary substitution occurs in the other partner in a manner that is consistent with the selective maintenance of Watson-Crick base-pairing between the nucleotides (Cheng *et al.*, 2012). Methods for detecting such complementarily coevolving positions, such as the mutual information approach of (Eddy and Durbin, 1994), can even be used to aid the computational inference of secondary structures. Accordingly, RNA comparative secondary structure prediction approaches such as those implemented in Pfold (Knudsen and Hein, 1999), use information about coevolving nucleotides inferred from sequence alignments to more accurately predict secondary structures. Evidence that base-paired nucleotides within a particular secondary structural element are coevolving is possibly the best indicator of that structural element having an important biological function. Unlike synonymous substitution rates, signals of complementary coevolution between nucleotide pairs will not be strongly influenced by either variations in basal mutation rates across different genome regions, or constraints on synonymous substitution rates that are unrelated to secondary structure (such as overlapping open reading frames or protein binding sites). Unfortunately, however, due to the dual evolutionary constraints in coding regions to maintain both secondary structural configurations and amino acid sequences, it is expected that there will be so little nucleotide variation at base-paired sites that there will often be insufficient information available to infer coevolution between such sites. In such cases the only evidence that an inferred secondary structure has some biological relevance might be that its constituent base-paired nucleotides occur within codons that display lower than expected synonymous substitution rates.

It is clear that biologically important nucleic acid structural elements can affect patterns of evolution due to selective processes that act to maintain structural interactions such as hydrogen bonding between base-paired nucleotides and other molecular interactions that determine the tertiary structural conformations. It is also possible, however, that the structural conformations of nucleic acid molecules can impact evolution prior to the action of selection by directly affecting the basal rates at which mutations arise within nucleotide sequences. The single-stranded regions of nucleic acid molecules are expected to be more prone

to chemical damage (Lindahl and Nyberg, 1974) and strand-cleavage (Parthasarathi *et al.*, 1995), and it has been hypothesised that there is a direct link between RNA secondary structure and the fidelity of RNA polymerases (Ji *et al.*, 1994; Yoshida *et al.*, 1997; Contreras *et al.*, 2002). Irrespective of the biological relevance of secondary structures it is therefore plausible firstly that unpaired nucleotide sites might experience globally higher basal mutation rates than base-paired nucleotide sites and, secondly, that unpaired nucleotide sites might also experience higher substitution rates (including both synonymous and non-synonymous substitution rates in coding regions) than base-paired nucleotide sites.

The impacts of genomic secondary structures on recombination

Besides influencing mutation and substitution processes, the structure of nucleic acid molecules can also potentially have a measurable impact on genetic recombination: another important process that occurs during the evolution of many virus species. Recombination results in either the reciprocal or non-reciprocal transfer of genetic information between different nucleotide sequences (Lai, 1992) and can result in the replacement (called homologous recombination), addition or deletion (called non-homologous recombination) of genetic material within the recipient sequence.

In many viruses recombination occurs during replication and does not involve the breakage and physical relegation of nucleic acid molecules. In HIV, for example, two genome copies are co-packaged within each virion and when upon entering a cell and these copies are reverse transcribed, The reverse transcriptase can disengage its current RNA template and sometimes drag the nascent DNA with it and re-engage with a new template producing a hybrid DNA molecule that carries genetic information derived from both templates (Billeter *et al.*, 1974; Beemon *et al.*, 1974; Onafuwa *et al.*, 2003; Chin *et al.*, 2005; Song *et al.*, 2007). This type of recombination, called copy-choice recombination, is common in many RNA viruses and is known in some cases to be strongly influenced by genomic secondary structure (Roda *et al.*, 2002; Moumen *et al.*, 2003; Galetto *et al.*, 2004). In HIV-1M, for example, it has been shown that secondary structure influences both the rate of recombination and the position of recombination breakpoints (Simon-Loriere *et al.*, 2010). Specifically, genomic secondary structures can cause polymerases and reverse-transcriptases to pause during the replication of nucleic acids. Such pausing is assumed to increase the possibility of the polymerases or reverse transcriptases disengaging from a template strand at genomic sites where there is a lot of RNA structure.

Another effect which secondary structure might have on the recombination patterns that occur in viruses is the influence that it might have on the viability of recombinant genomes. When a recombinant genome is generated from two sufficiently divergent parental viruses there will be a high probability that it will not have precisely the same secondary structure as its parental genomes. Specifically, if some of these structural differences are in biologically important secondary structures, then they could have a substantial negative impact on the recombinant's fitness. It is therefore plausible that the patterns of recombination that are observed in within virus genomes that are sampled from nature might display evidence of selective processes that have purged recombinant genomes containing disrupted secondary structures.

Although no evidence currently exists of recombination-induced secondary structure disruption having a major impact on recombination patterns in viruses sampled from nature, it is noteworthy that, for some groups of viruses at least, natural recombinants tend to display far lower degrees of protein folding disruption than would be expected in the absence of selection acting to purge recombinants that expressed misfolded chimaeric proteins (Voigt *et al.*, 2002; Lefeuvre *et al.*, 2007). This suggests that, given suitable tools to measure secondary structural disruptions, it may be possible to detect evidence of selection acting against recombinants with disrupted genomic secondary structures. In fact, such evidence has been detected, albeit only in an experimental setting, using computationally inferred genome-scale secondary structures of recombinants arising during experimental mixed infections (D. P. Martin *et al.*, 2011). It is unclear, however, how the approach used in this study would translate to the study of recombinants arising in nature. Unlike in natural settings where the actual parental genomes are never available for analysis and recombinants have usually accumulated multiple nucleotide substitutions subsequent to their generation, these recombination experiments of Martin *et al.* (2010) involved the analysis of over 50 recombinant genomes all of which were derived from the same two parental viruses. Nevertheless, if recombination-induced secondary structure disruption does indeed have a large impact on the survival of recombinants that arise in nature, then it might be possible to detect such a signal using a similar approach to experimental studies that make use of recombinants that are generated *in vitro*. Using such an approach it is relatively easier to accurately determine the position of breakpoints and the identity of parental sequences that contribute to recombinant sequence, which would aid our ability to detect such a signal.

Aims

This study investigates the overall biological relevance of nucleic acid secondary structure in the genomes of HIV and a number of flaviviruses by examining the relationships between secondary structure and patterns of genetic variation generated through mutation and recombination. A large portion of this study relied on the development of new computer software. Specifically, computational tools were developed both for identifying regions of conserved nucleic acid secondary structure evident amongst aligned sets of virus full-genome sequences, and for quantifying the biological relevance of the identified secondary structures based on evidence of natural selection acting to preserve these structures. Although not used as part of another investigation in this study, software was also developed for more accurately predicting RNA secondary structures from a set of related sequences by accounting for alignment uncertainty. I start in Chapter 2 by describing the development of a computer program for analysing secondary structures within nucleic acid sequences. This tool integrates various existing pieces of software into a graphical user interface that allows a user to both analyse nucleic acid sequences for the presence of secondary structures with likely biological relevance, and visualise these secondary structures in the context of relevant additional data such as synonymous substitution rates and degrees of nucleotide conservation. In Chapter 2 I also describe the development of a computational method and associated graphical interface for identifying regions of structural conservation within an aligned set of related secondary structures. Both these tools are used in Chapter 3 to analyse various virus genomes for the presence of conserved biologically relevant secondary structural elements. Furthermore, the effect of secondary structures on the evolution of various viruses is investigated and the relationships between secondary structures and underlying mutational processes demonstrated. Chapter 4 explores the relationship between genetic recombination and both RNA secondary structure and protein tertiary structure in HIV-1. In particular, naturally occurring recombinant sequences are analysed for evidence of them being less disruptive of nucleic acid secondary structure and protein tertiary structure interactions than would be expected if recombination occurred randomly and recombination was neutral. I additionally determine whether evidence exists to suggest that recombination breakpoint locations are predisposed to occur at genomic sites within nucleic acid secondary structural elements. Finally, in Chapter 5 a general method for predicting RNA secondary structures from set of related sequences is developed that accounts for alignment uncertainty and has a degree of predictive accuracy that is better than related structure prediction methods, in addition to this an improved metric for evaluating the quality of particular secondary structure prediction is presented.

Candidate's contributions

The computational tools developed in Chapter 2 were developed and programmed solely by the candidate, Michael Golden, with design recommendations thoughtfully suggested by Darren Martin, Brejnev Muhire and Emil Tanov. The analysis of various viral genomes carried out in Chapter 3 were developed and performed by Michael Golden, with interpretation of the results provided by both Michael Golden and supervisor, Darren Martin. The tests of protein fold-disruption in Chapter 4 were performed by Michael Golden using pre-existing tools available in the software RDP (Martin *et al.*, 2010). The secondary structure fold-disruption tests in Chapter 4 were conceived by Yves Semegni and Brejnev Muhire and were modified and re-implemented by Michael Golden such that it was computationally feasible to perform on longer RNA secondary structure (for example: the genomic secondary structures of entire ~9,200 nucleotide long HIV-1 genomes). Interpretation of the results for the tests in Chapter 4, were provided by both Michael Golden and Darren Martin. The work completed in Chapter 5, was carried out as part of a project at the Oxford Summer School in Computational Biology 2012. James Anderson, Ádam Novák and Jotun Hein conceived the project. Further ideas were contributed to the project by Michael Golden, and project members: Preeti Arunapuram, Ingólfur Eðvarðsson and project supervisors: Ádam Novák and James Anderson. One third of the software development was carried out by the candidate, Michael Golden (the other two thirds was carried out by Preeti Arunapuram and Ingólfur Eðvarðsson). Benchmarking of the software was performed by Michael Golden and drafting of the corresponding manuscript (Arunapuram *et al.*, 2013) was carried out by Michael Golden along with James Anderson.

2 Tool development

Introduction

Nucleic acid secondary structures can be predicted using either computational (Markham and Zuker, 2008; Bernhart *et al.*, 2008; Sükösd *et al.*, 2011; Semegni *et al.*, 2011) or experimental (Wilkinson *et al.*, 2006) approaches. Whilst, this is a challenging algorithmic problem in itself, an additional algorithmic problem subsequent to the prediction of a secondary structure is the graphical visualisation of the structure. Accordingly, there are a number of software tools available for the visualisation of nucleic acid secondary structures. Among the most popular of these are VARNA (Darty *et al.*, 2009) and RNAplot in the RNAstructure package (Reuter and Mathews, 2010).

The challenge of visualisation involves taking a raw secondary structure data file, which typically contains a list of paired and unpaired nucleotide positions, and producing an aesthetically pleasing 2-dimensional graphical layout consisting of the stems and loops that represent the secondary structure. Laying out the secondary structure such that stems and loops in different parts of the structure do not overlap with one another is one of the main challenges of structure visualisation (Brucoleri and Heinrich, 1988). Most visualisation tools such as VARNA (Darty *et al.*, 2009) have little trouble in laying out structures that are smaller than 300 nucleotides with minimal overlap. As secondary structures become larger, however, overlapping structural features become increasingly problematic. Many programs, including VARNA, allow the user to rotate different parts of the structure which can assist in reducing overlaps. However, for very large secondary structures, such as the ~9600 nt long Hepatitis C Virus genome, this user-guided process can be extremely tedious and there do not yet exist automated ways in which such large structures can be efficiently splayed-out in an unobstructed and visually appealing manner (see Figure 2.1). These problems, amongst others, make it difficult for researchers to easily interpret the functional and biological relevance of inferred secondary structures.

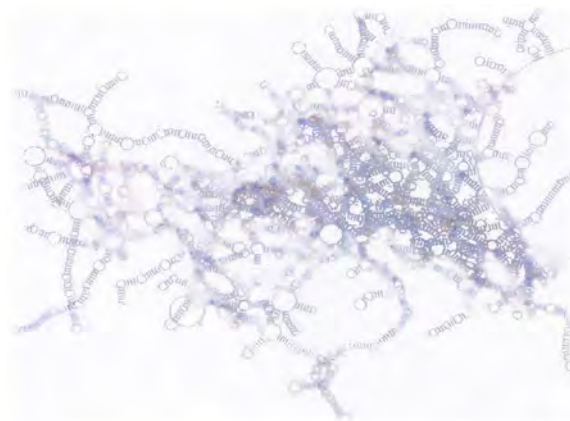


Figure 2.1: Layout produced by VARNA for a ~9,600nt long Hepatitis C Virus genome secondary structure. The large size of this secondary structure and the fact that much of the visualised structure is occluded due to overlap makes it difficult to give any useful biological interpretation.

Even if the problem of overlap were solved, the size alone of these larger secondary structures makes interpretation difficult for a researcher. In order to address the problem of visualising large secondary structures (typically structures which are >1000 nucleotides in length) we previously developed a tool called DOOSS (Data Overlaid On Secondary Structures; Golden and Martin, 2013). The key idea that alleviates the problem of visualisation is that large secondary structures are generally composed of discrete *substructures*. For large sequences, rather than visualising the entire secondary structure of a sequence, DOOSS takes the full-length secondary structure of that sequence and divides it into a number of individual, more easily displayed, substructures of between 10 and 350 nucleotides in length. We define a substructure as any sub-region of a secondary structure, for example: positions 1050 to 1250 in a particular secondary structure would be considered a substructure. The main purpose of DOOSS is to assist researchers in identifying substructures within a secondary structure that are likely to have biological functionality.

There are two primary reasons for considering substructures: smaller regions are more manageable for an investigator to deal with and functional components of secondary structures tend to correspond to sub-regions of a secondary structure (i.e. substructures). An example of a highly functional substructure in a virus genome is the HIV Rev Response Element (RRE). The RRE occurs in the protein-coding envelope gene (*env*) and is approximately 350 nucleotides in length. By binding to the HIV Rev protein, the RRE allows HIV genomic RNA to be exported from the nucleus to the cytoplasm where translation and packaging of virions can take place (Heaphy *et al.*, 1990; Daugherty *et al.*, 2010). Figure 2.2 shows a portion of the SHAPE-inferred HIV-1

RRE visualised in DOOSS (Watts *et al.*, 2009). There are numerous examples of similar biologically relevant substructures which have been identified in a variety of virus and animal species (Mason *et al.*, 2002; Tuplin *et al.*, 2004; Pedersen *et al.*, 2006).

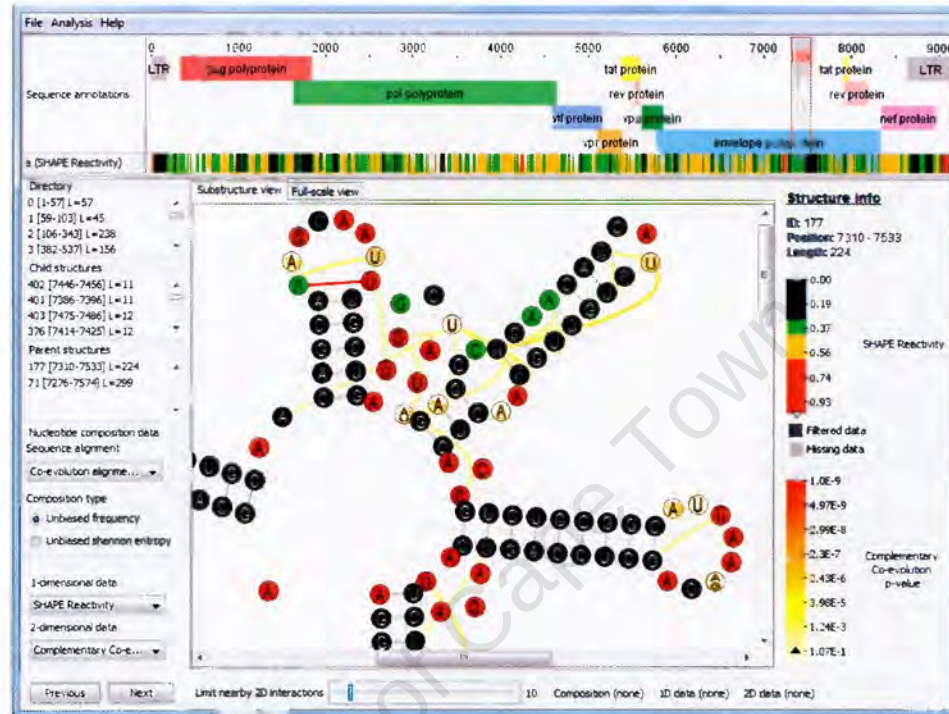


Figure 2.2: Screenshot of the RRE substructure in DOOSS. A screenshot from DOOSS showing a portion of the NL4-3 HIV-1 Rev response element (RRE) with SHAPE reactivities overlaid on the structure. SHAPE reactivities are experimentally inferred values that provide information about nucleotide accessibility facilitating the inference of the secondary structure of an RNA sequence. Notice that nucleotides with low SHAPE reactivities (i.e. low accessibility, black and green) tend to occur in the stem regions of the secondary structure, whereas nucleotides with high SHAPE reactivities (yellow and red) tend to occur in loop regions of the structure.

In this chapter a new tool is developed, called NAVA (Nucleic Acid Visualisation and Analysis). NAVA is an extension of DOOSS which provides a much larger set of tools for analysing secondary structures within sequences, providing both prediction and identification of biologically relevant substructures. The need addressed by NAVA is that there is no single tool currently available that is capable of simultaneously predicting, analysing and visualising nucleic acid secondary structures. Typically users have to rely on various different tools to analyse nucleic acid sequences and these often have different input and output formats, forcing the user to convert from one set of file formats to another. NAVA alleviates this problem by supporting a wide-range of file formats and includes all steps necessary in a typical secondary structure prediction and analysis workflow, such as: sequence alignment, prediction of secondary structures from a sequence or an alignment of sequences, performing various types of site-by-site rates analyses and visualisation of the generated secondary structures (in much

the same way that DOOSS does). This software tool is designed to be extensible so that other developers can plug-in their own applications, making them accessible from NAVA's graphical user interface (GUI). These applications can be existing command-line binaries or native Java applications and it has been ensured that there are relatively simple ways for developers to programmatically add their own custom applications together with user interface components for integration into the NAVA GUI. By enabling the prediction, analysis and visualisation of secondary structures in a simple, regularly updated computer program we hope it will become the most useful tool of its type.

In addition to NAVA, a method was also developed for identifying regions of conserved substructures within larger nucleic acid secondary structures. Substructures within large nucleic acid molecules often have specific functions and selective processes are expected to preserve such functionally important substructures (Hofacker *et al.*, 1998; Collier *et al.*, 2002; Pedersen *et al.*, 2006). The ability to identify which substructures are structurally conserved amongst multiple homologous sequences provides evidence as to which substructures are likely to be functionally relevant. For this reason a substructure conservation analysis was developed for automatically detecting conserved substructures amongst a group of aligned sequences. This analysis provides quantitative information about the degrees of structural conservation at positions within an alignment of related sequences and produces a list of corresponding consensus substructures representing the conserved substructures amongst the sequences being analysed. We believe that this substructure conservation analysis will be helpful in assisting researchers to identify previously uncharacterised conserved biologically relevant substructures.

Methods

Nucleic Acid Visualisation and Analysis

Requirements and availability

NAVA (Nucleic Acid Visualisation and Analysis) is written in the Java programming language and requires that Java Runtime Environment 1.7 or later be installed on the user's computer. Java is a cross-platform programming language meaning that programs written in Java can generally be run on a number of different operating systems, such as Windows, MacOS and Linux. However, NAVA is currently only available for Windows as a number of key third party software components within NAVA require a Windows operating system in order to run. Having said that, we intend to make NAVA compatible with both Windows and Linux operating systems in a future release. The NAVA application is available for download from our website, along with a video tutorial: <http://www.computingforbiology.org/software>. The NAVA source code is licensed under GNU GPLv3 and is available for download from: <https://github.com/MichaelGolden/NAVA>.

NAVA workflow overview

Here NAVA is briefly described in terms of a workflow (Figure 2.4). The workflow of NAVA is essentially divided into two components: a "data input and data generation" component and a "data visualisation and exploration" component. These two components are available from a "Data input" tab and a "Visualisation" tab, respectively (Figure 2.3).

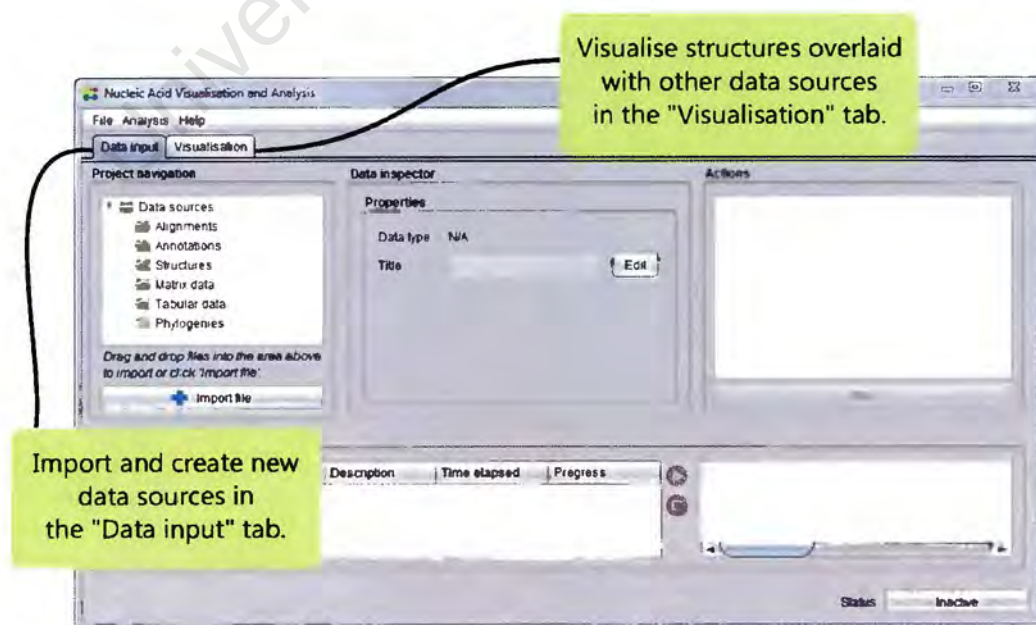


Figure 2.3: NAVA data input and visualisation tabs. NAVA is divided into main components: a "data input and data generation" component and a "data visualisation and exploration" component. These components are available from a "Data input" tab and a "Visualisation" tab, respectively.

The starting point for any analysis is the creation of a new NAVA project. Once a project has been created, external data files can be imported into the project in the "Data input" tab. NAVA automatically detects the data source type of an imported file and creates a data source object representing that file. Data sources can be used with built-in applications to generate other data sources. For example, an alignment data source can be used in conjunction with the built-in PPfold application to predict a consensus secondary structure which can later be visualised.

Once some data sources have been imported or generated the user can explore those data sources using the data visualisation and exploration component of NAVA. In order to visualise data the user needs to open the visualisation tab. NAVA is designed such that the user can switch freely between the data input and visualisation tabs at any time. In order to visualise a particular data source, a data overlay needs to first be created from that data source. A data overlay is either a secondary structure or a piece of data that can be graphically overlaid on top of a secondary structure, thus permitting the visualisation of data in the context of the secondary structure.

In addition to the visual analysis of secondary structures, different types of automated ranking analyses can be performed on a combination of structure and data overlays. These data analyses use statistical tests to provide quantitative insights into the biological relevance of either an entire secondary structure, or particular individual substructures. For publication purposes NAVA allows substructure visualisations and other graphical elements to be exported in a number of high-quality image formats.

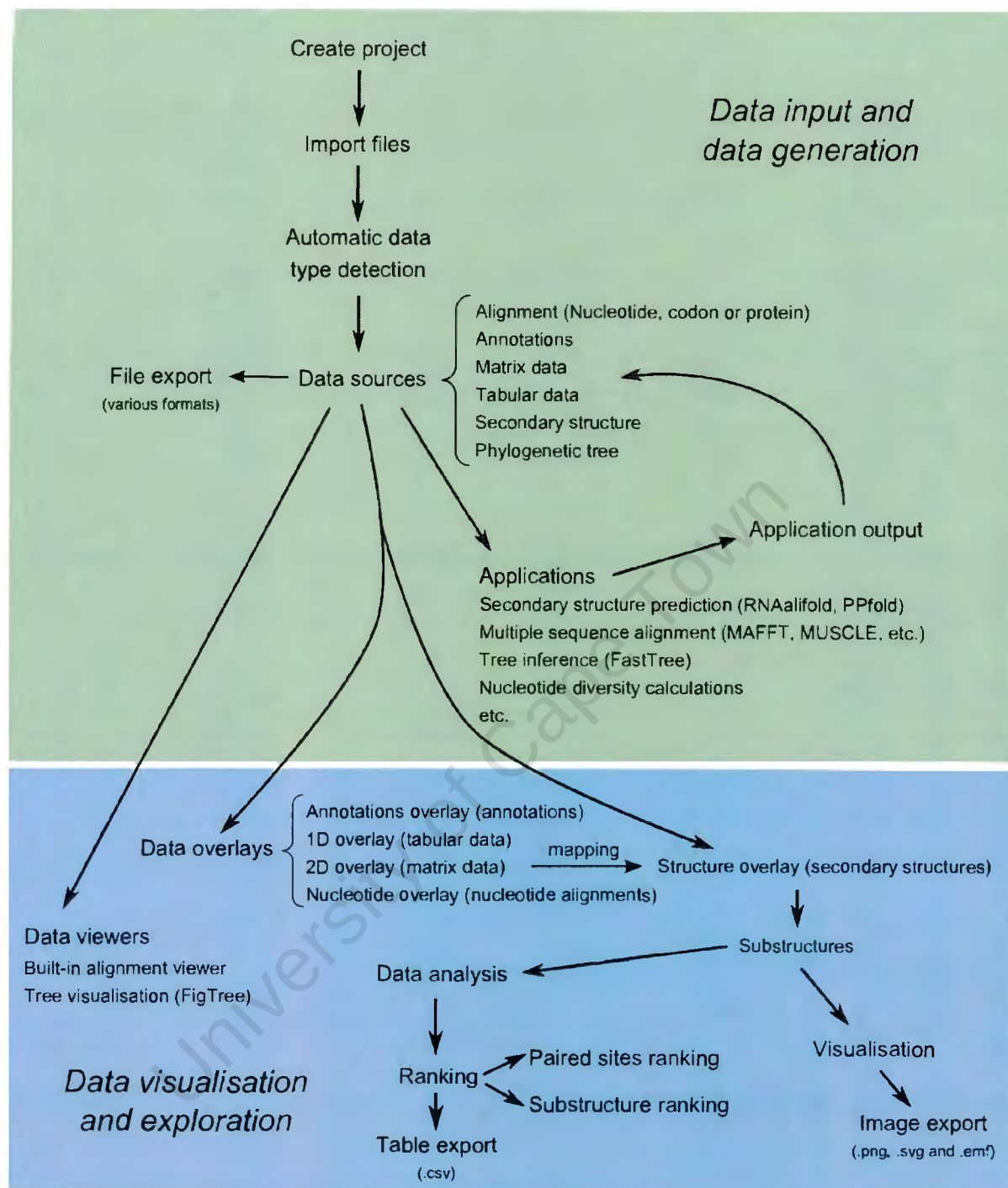


Figure 2.4: Flowchart describing NAVA workflow. The *data input and data generation* component of NAVA allows the user to import files (Import files). NAVA automatically detects the data source type of the file (Automatic data type detection) and creates a data source object representing that file (Data sources). Data sources can be used with built-in applications to generate other data sources (Applications), for example: an alignment data source can be used by PPfold to predict a consensus secondary structure. Once some data sources have been inputted or generated the user can then explore those data sources using the *data visualisation and exploration* component of NAVA. For some data source types the user has options to view data sources of that type using a data viewer (Data viewers) or export them to a variety of file formats (Data export). Data sources are used to create *data overlays* (Data overlays). A data overlay is either a secondary structure or a piece of data that can be graphically overlaid on top of a secondary structure to view data in the context of nucleic acid secondary structures (Visualisation). Different types of automated ranking analyses can be performed on a combination of a structure overlay and a data overlay (Ranking). These data analyses are designed to give insights into the biological relevance of an entire secondary structure or particular substructures within that secondary structure. Substructure visualisations and other visual elements, such as data legends, can be exported in a variety of image formats (Image export) for publication purposes. Generated tabular data from a data analysis, such as rankings, can also be exported from NAVA for external use (Table export).

Data input and output

The *data input and data generation* component facilitates import and export of various data files. NAVA will automatically detect the file format of imported files by attempting to parse it using various file parsers. In cases where a file is identified as having more than one possible data format or data type NAVA will prompt the user to select the correct one. NAVA has support for a wide variety of file formats so as not to require the user to have to convert their data to a specific format before importing it into NAVA. Once a data file is imported it becomes a NAVA *data source*, which can be one of six abstract data types: alignments (nucleotide, coding or protein alignments), annotations (e.g. gene names and locations), matrix data (e.g. matrix representing nucleotide interactions), tabular data (e.g. Excel spreadsheets containing probing information at individual nucleotide sites), secondary structures (which defines base-paired nucleotides in a sequence) and phylogenetic trees (Figure 2.5). NAVA uses its own internal representation to represent data sources of different data types. For example, if the user imports a matrix,

NAVA will convert this matrix into its internal compact matrix format so as to allow NAVA to do fast lookups on the matrix while it is stored on disk rather than in memory.

NAVA also allows for the export of data sources. This is useful for converting an imported data source to another file format or for the export of data sources generated by NAVA for external use by other software. NAVA supports a variety of formats for both data source import and export (Table 2.1).

Applications and data generation

NAVA has a number of built-in applications that the user can access from the GUI both to view imported data and to generate new data. A benefit of having an internal representation for the different data types is that it makes it easy to manage transparent piping of data to the different built-in applications, as each application has different data format input requirements. NAVA ensures that data sources are transparently and automatically converted to the correct format before they are processed by an application. This means that the user does not have to concern themselves with having to convert from one data format to another.

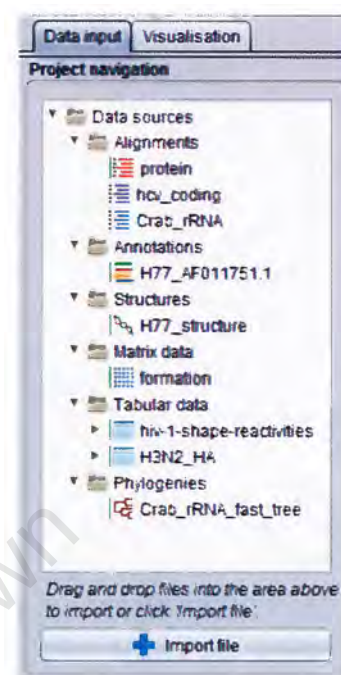


Figure 2.5: NAVA project navigation panel. This panel provides a graphical tree of data sources contained within the current project. Each data source is categorized into 1 of 6 data type categories.

Table 2.1: Input (import) and output (export) file formats supported by NAVA for different data types.

Data type	Input file formats supported	Output file formats supported
Alignments	FASTA; GenBank/GB; EMBL; Zuker; Phylip3.2; Phylip4; Nexus; Clustal; MSF; IG/Stanford; NBRF; DNASTrider; Fitch; Olsen; PIR/CODATA; ASN.1	FASTA; Phylip4; Nexus; Clustal
Annotations	GenBank; EMBL	Annotations output not yet supported
Secondary structures	Vienna Dot Bracket; Connect; Tab-delimited helix; bpseq	Vienna Dot Bracket; Connect
Matrix data	Dense matrix (tab or comma-separated list of lines representing matrix rows); Coordinate-list sparse matrix (tab and comma-separated)	Dense matrix (tab and comma-separated); Coordinate-list sparse matrix (tab and comma-separated)
Tabular data	Comma-separated values; Excel spreadsheet	Comma-separated values
Phylogenies	Newick tree format	Newick tree format

NAVA has a variety of applications that the user can use both to generate new data sources from existing data sources, and to view data sources (Table 2.2). For example, there are three different multiple sequence alignment applications currently available (MAFFT, MUSCLE and ClustalW) that allow the user to align a set of nucleotide or amino acid sequence data sources. The aligned data sources can be used to generate other data sources, such as phylogenetic trees (using FastTree) or secondary structure consensus predictions (using PPfold or RNAalifold).

NAVA makes use of pre-existing applications by programmatically executing these and passing user-specified data sources to them in the expected formats. Once one of these third party applications has completed running, NAVA will import the generated output files and convert these to NAVA's own internal formats. This means that a plethora of existing open-source software tools can be plugged into NAVA without having to write implementations from scratch in Java. Having these built-in applications means that much of the data generation that is typically required in a secondary structure analysis workflow can be done directly from the NAVA graphical user interface without having to install or run specialised applications on the user's computer.

NAVA has been designed so that it is easy for developers with access to the source code to incorporate different pre-existing or new applications. Developers can do this by extending NAVA's *Application* class interface (an interface is a Java programming construct that is like a template that developers wishing to add an application are required to conform to). An additional Java interface assists developers in creating their own graphical dialogs that allow the user to specify application parameters for a given run of a particular application and also specifies how NAVA is allowed to run an application. For example, it is possible to specify the

number of CPUs to use, pause and resume options, and user feedback (such as percentage progress).

Table 2.2: Currently available NAVA applications.

Application name	Data source Input	Output	Description
PPfold (Sükösd <i>et al.</i> , 2011)	Nucleotide or codon alignment	Secondary structure	Predicts a consensus RNA secondary structure from a set of aligned sequences using a Stochastic-Context Free Grammar (SCFG) approach.
RNAalifold (Bernhart <i>et al.</i> , 2008)	Nucleotide or codon alignment	Secondary structure	Predicts a consensus RNA secondary structure from a set of aligned sequences using a thermodynamic approach.
MAFFT alignment (Kato and Toh, 2008)	Nucleotide, codon or protein alignment	Alignment	Performs a multiple sequence alignment using MAFFT - MAFFT is fast and reasonably accurate.
MUSCLE alignment (Edgar, 2004)	Nucleotide, codon or protein alignment	Alignment	Perform a multiple sequence alignment using MUSCLE - MUSCLE is accurate, but somewhat slow.
CLUSTALW alignment (Thompson <i>et al.</i> , 1994)	Nucleotide, codon or protein alignment	Alignment	Perform a multiple sequence alignment using ClustalW - ClustalW is an older multiple sequence alignment tool which is no longer state-of-the-art.
Simple nucleotide diversity	Nucleotide or codon alignment	Tabular data	Calculates nucleotide diversity at each alignment site.
Maximum Expected Accuracy Prediction (Knudsen and Hein, 2003)	Base-pair probability matrix	Secondary structure	Given a base-pairing probability matrix, returns the secondary structure that maximises the expected number of correctly predicted positions.
FastTree (Price <i>et al.</i> , 2010)	Nucleotide, codon or protein alignment	Phylogeny	Exceptionally fast and accurate approach for inferring a phylogenetic tree from a set of aligned sequences.
FigTree (Rambaut, 2012)	Phylogeny	Opens a phylogenetic tree (phylogeny) in FigTree.	Graphical viewer and editor for phylogenetic trees.

Data visualisation and exploration

Once some data sources are available the user can begin exploring the data. The first step in doing so is the creation of *data overlays* from existing data sources, there are five categories of data overlay: structure, one-dimensional, two-dimensional, nucleotide, and annotation overlays. Data overlays are necessary in order to define how a specific data source should be visualised or analysed together with other data overlays. For example, the user may import a secondary structure data source that they wish to visualise, but before that secondary structure can be visualised a structure overlay needs to be created from it so that NAVA can acquire additional information informing it how to appropriately visualise the structure and other data overlays that may be overlaid on the structure.

Mapping between data overlays

Every data overlay that is created requires a corresponding mapping source that informs NAVA how to map data between structure overlays and the four other types of overlay. A mapping source is a sequence or an alignment of sequences, where the nucleotide or column positions correspond to positions in the data, for example: the sequence associated with a particular secondary structure. Mapping sources are most often the sequences or the alignments that were used to generate that data source. For nucleotide and annotation overlays the user does not need to specify the mapping source, because the data source itself (a sequence alignment in the case of nucleotide overlays) is the exact mapping source. However, in the case of one-dimensional and two-dimensional data overlays, the user is required to explicitly specify a mapping source. By associating a mapping source with each data overlay, NAVA can automatically map data onto various different structure overlays that the user may specify. For example, the user may have multiple different HCV secondary structure overlays that they want to visualise with a specific one-dimensional data overlay. This automatic mapping is achieved by performing a profile-profile alignment between the structure overlay mapping source and the data overlay mapping source (see Supplementary Figure 1 for an explanation). A natural requirement for mapping between two data overlays is that the mapping sources must be homologous, i.e. accurately alignable.

Types of data overlays

Structure overlays

Structure overlays are NAVA's most important type of overlay. The four other types of data overlay are mapped to structure overlays. Structure overlays represent the coordinates of base-paired and single-stranded sites in a nucleic acid secondary structure and are derived from secondary structure files such as connect format and Vienna dot-bracket format files. The user is expected to specify additional information about structure overlays, such as the conformation of the sequence (whether it be linear or circular - circular sequences do not have 5' and 3' ends, i.e. the 5' and 3' ends are covalently linked) and the sequence or alignment (the mapping source) that corresponds to the secondary structure (Figure 2.6). Optionally, a set of criteria for generating a list of substructures from the full secondary structure can also be specified by the user (Figure 2.7), these criteria determine the minimum and maximum size of substructures and whether substructures are allowed to overlap with other substructures in the generated list.

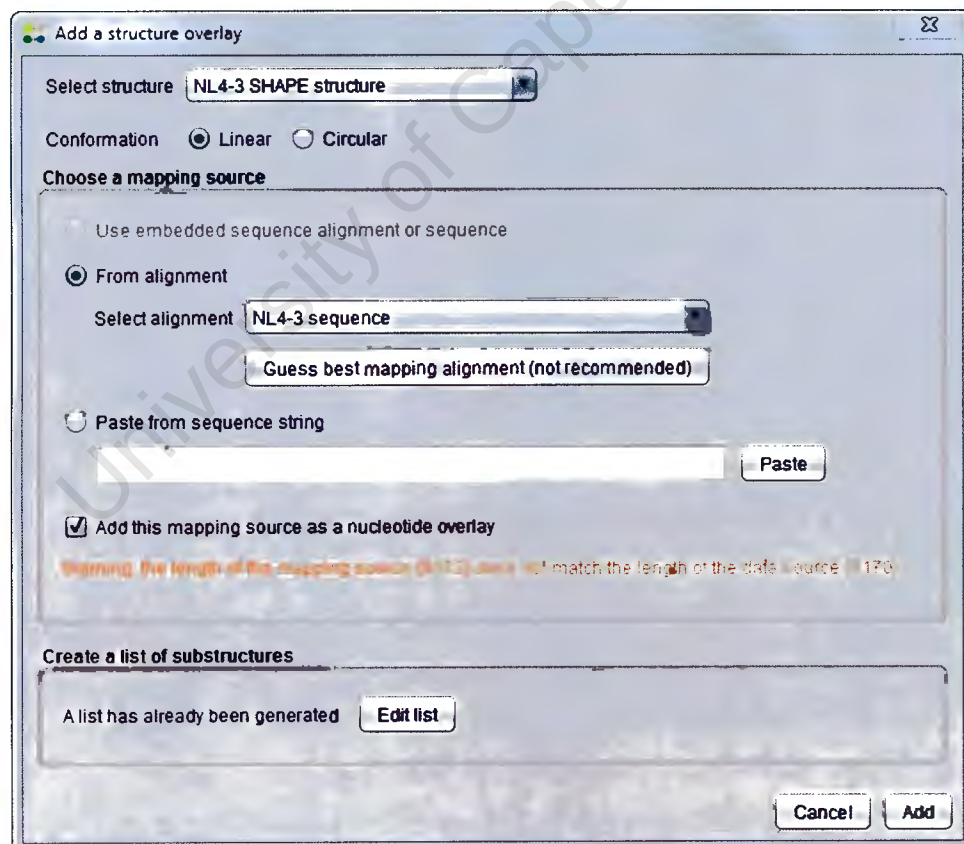


Figure 2.6: The "Add a structure overlay" dialog. Before a secondary structure can be visualised the user first needs to create a structure overlay. Creating a structure overlay compels the user to specify some information about a nucleic acid secondary structure so that NAVA knows how to display that structure appropriately.

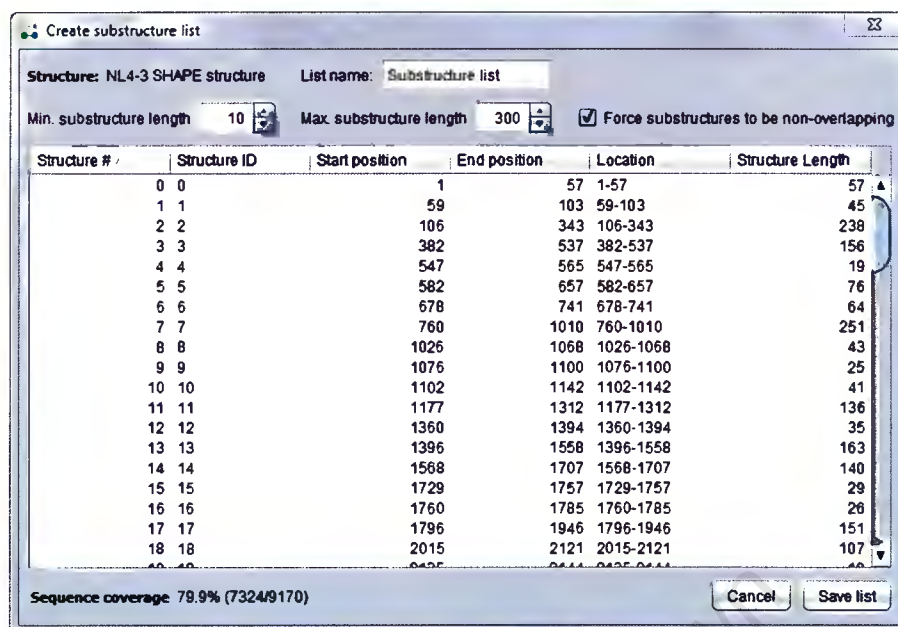


Figure 2.7: The "Create substructure list" dialog. This dialog is available from the add/edit structure overlay dialog and allows user to specify how the full secondary structure associated with a given structure overlay is subdivided into a number of substructures for visualisation and analysis purposes. This dialog gives user the option to specify the minimum and maximum size of the substructures and whether substructures in the list are allowed to overlap other substructures in the list (specified using the "Force substructures to be non-overlapping" option).

One-dimensional data overlays

One-dimensional data overlays allow for the visualisation of single nucleotide or single codon resolution data. In NAVA, these types of data are derived from columns within comma-separated value (.csv) or Excel spreadsheet files. NAVA terms these types of data "tabular data sources". Tabular data sources are used to create one-dimensional data overlays. SHAPE accessibility scores (Figure 2.11 B) are an example of a one-dimensional data overlay. SHAPE is an experimental technique that is used to predict base-paired nucleotide positions in a RNA sequence by measuring nucleotide accessibility (Wilkinson *et al.*, 2006). SHAPE reactivity scores are used to infer the secondary structure of a nucleic acid molecule and are a good example of information that the user may want visualise as a secondary structure overlay. For every one-dimensional data overlay the user is expected to define a colour legend, which specifies how numeric values corresponding to nucleotides are depicted on the structure (Figure 2.8).

1. Select a data field

Data table: NL4-3 SHAPE reactivities

Data field: SHAPE reactivity

Title: SHAPE reactivity

2. Specify how the values are positioned

☒ Natural (1, 2, 3, ...) ☒ First line is header

☐ From field: Position

☐ Positions are codon positions

3. Map the data values to the structure

Mapping source: NL4-3 sequence

The length of the mapping source (9173) matches the length of the data source (9173)

Guess best mapping source (not recommended)

4. Specify the numeric range

Minimum value: 0.0 e.g. 0.01 or 1e-2

Maximum value: 1.0 Reset

How should values out of this range be treated?

☐ As missing data ☒ Clamped

5. Choose how the data values are displayed

Select a transform: Linear

Data preview

Positi...	M...	Value
1	G	
2	G	
3	T	
4	C	
5	T	
6	C	
7	T	
8	C	
9	T	
10	G	
11	G	
12	T	0.07
13	T	0.01
14	A	0.09
15	G	0.34
16	A	2.87
17	C	0.08
18	C	0.14
19	A	0.03
20	G	0.03
21	A	1.28
22	T	1.37
23	C	3.63
24	T	0.22
25	G	0.05
26	A	0.05
27	G	0
28	C	0
29	C	0.24

Figure 2.8: The "Add a 1-dimensional data overlay" dialog. When adding a 1-dimensional data overlay the user first must select the table and column field they want to retrieve data values from. They also need to specify information about how the data are positioned (e.g. whether there is a header in the table, if the table numbering starts at zero or one, etc.). A "data preview" panel on the right displays how the selected data field values correspond to nucleotide positions in the selected mapping source. The user should ensure that the data values correspond to the correct nucleotide positions. For visualisation purposes the user needs to specify a colour legend, which is a mapping between numeric values and colours. The values will be displayed on the selected structure overlay as coloured nucleotides indicating the corresponding numeric value at each nucleotide position within the secondary structure.

Two-dimensional data overlays

Two-dimensional data overlays are for visualisation of two-dimensional quantitative data relating to pairs of nucleotides, such as the distance in ångströms between nucleotides within the predicted or experimentally determined tertiary structure, the degree to which pairing between the nucleotides is conserved, predicted base-pairing probabilities (Figure 2.11 C) or probabilities that a pair of nucleotides is co-evolving. Wherever a pair of nucleotides in the sequence has a corresponding numeric value within a certain range, it is depicted as a coloured line (also with a user-definable colour gradient) between the nucleotides. Two-dimensional data overlays rely on matrix data, where a particular element i,j of a matrix corresponds to an interaction, i.e. an interaction between a nucleotide i and a nucleotide j . Matrices used to represent two-dimensional data overlays can be sparse, which captures the fact that some pairs of nucleotides may have no interactions associated with them. The creation of a two-dimensional data overlay is demonstrated in Figure 2.9.

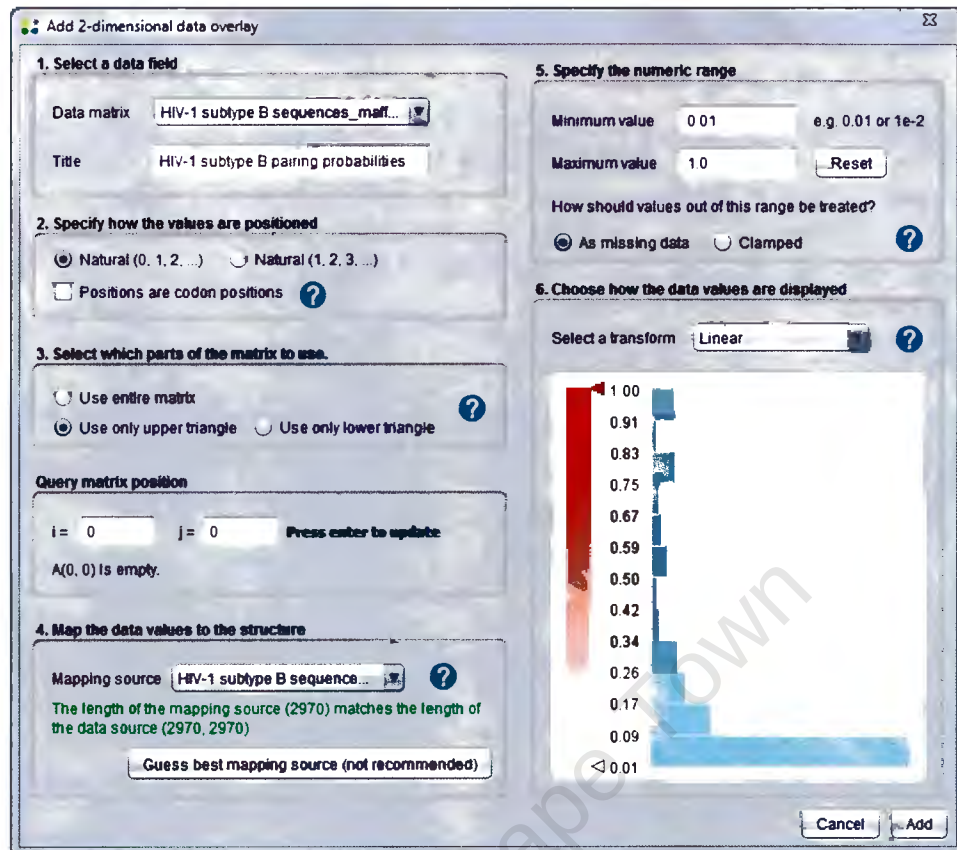


Figure 2.9: The "Add a 2-dimensional data overlay" dialog. When adding a 2-dimensional data overlay the user first must specify the matrix they want to retrieve interaction data from. They also need to specify information about how the data is positioned (i.e. whether the matrix format numbering starts at zero or one) and which part of the matrix they want to retrieve data from: often the user may just want to retrieve data from one half of a matrix, such as when a matrix is symmetric and the same value is represented by both the x,y element and the y,x element. Doing so will prevent the same interaction from being rendered twice on the secondary structure. In addition to this, the user needs to select a colour legend for visualisation purposes. The colour legend maps numeric values to colours which are displayed on the secondary structure as coloured lines linking nucleotides wherever there is a corresponding interaction.

Nucleotide data overlays

Nucleotide data overlays are pieces of data that represent nucleotide composition. These types of overlay are derived from nucleotide sequence alignments, where the columns of the alignment are used to determine nucleotide composition at each nucleotide position within a structure. Nucleotide composition data are displayed in the form of a sequence logo at each nucleotide position (depicted by a circle) within a structure. Using different sequence alignments consisting of related groups of sequences (for example, particular HIV genotypes) allows one to visualise patterns of nucleotide variation at particular positions within a secondary structure (Figure 2.11 D).

Annotation overlays

Annotation overlays are a special type of overlay that represents sequence locations of biological features such as genes, structural RNA, variation information, exons and introns. Annotation overlays are derived from GenBank or EMBL files. Annotation overlays are not displayed visually on the secondary structure, but rather are displayed in a linear fashion in a separate annotations panel above the secondary structure visualisation (Figure 2.10). By moving the mouse over the annotations panel, the user can navigate the different substructures present in a secondary structure. This, for example, allows the user to view all the substructures that are present in a given gene. By right-clicking on the annotations panel the user can add new annotations from an annotation source, rearrange the visual elements representing the different annotations by shifting them up or down, or remove annotations which they don't wish to have displayed on the panel.

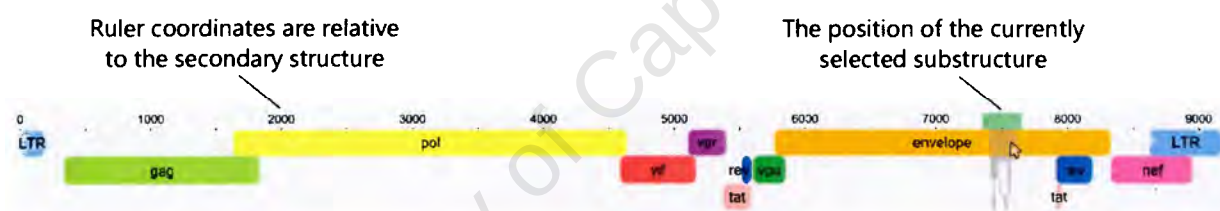


Figure 2.10: A graphical depiction of the NL4-3 HIV-1 genome annotations in the NAVA annotations panel. These annotations were derived from the NL4-3 GenBank entry (GenBank ID: AF324493, <http://www.ncbi.nlm.nih.gov/nuccore/AF324493>) and automatically mapped to the SHAPE determined NL4-3 HIV-1 secondary structure by NAVA. When the user moves their mouse over the panel a transparent set of light gray blocks appears at the position of the mouse cursor, indicating possible substructures at that position in the secondary structure. NAVA highlights the currently selected substructure in the annotations panel in light green.

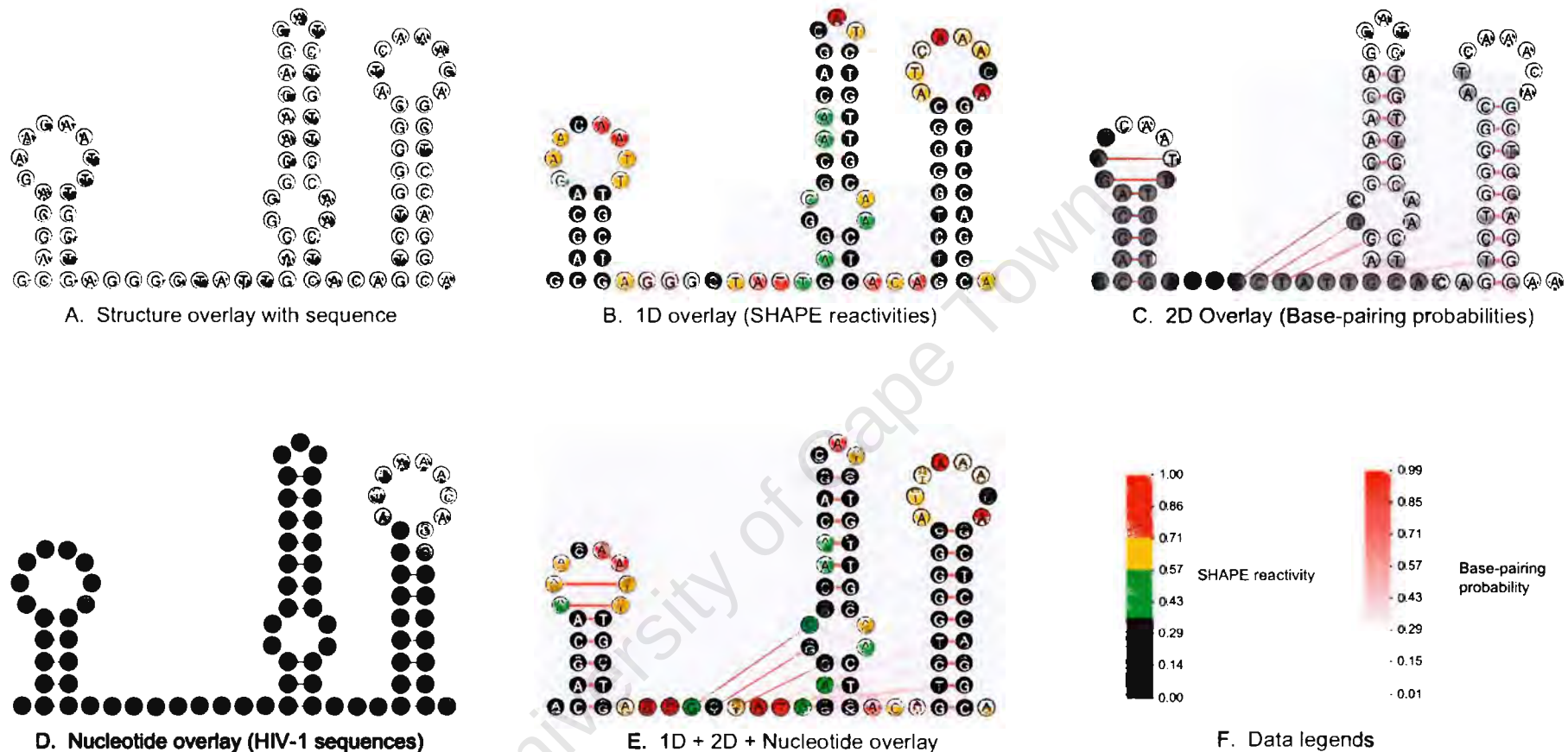


Figure 2.11: Examples of NAVA overlays demonstrated for a HIV-1 substructure. The most basic type of overlay is a structure overlay (A). Structure overlays are combined with other overlays to produce structure visualisations (B, C, D and E). 1D data overlays (B) reflect numeric data that correspond to individual nucleotides in a structure. In this example SHAPE reactivities (a proxy for structural accessibility) are overlaid on the structure. Notice how nucleotides in the base-paired regions tend to have lower SHAPE reactivities (an enrichment of black nucleotides implying lower accessibility). This is a good example of how overlaying data in the context of the structure can be insightful since low SHAPE reactivities are used to infer base-paired nucleotide positions. 2D data overlays (C) corresponding to interactions between pairs of nucleotides. In this example base-pairing probabilities derived from a group of computationally inferred HIV-1 secondary structures are overlaid on the experimentally inferred structure. Although most of the strong base-pairing probabilities correspond to base-paired nucleotides in the structure, base-pairing probabilities corresponding to unpaired nucleotides suggest alternative base-pairings. Nucleotide overlays (D) reflect information about nucleotide composition or sequence variation amongst a group of aligned sequences using sequence logos inscribed inside the circles that represent nucleotide positions. The nucleotide overlay used here (D) is based on an alignment of HIV-1 sequences. Notice how single-stranded nucleotide positions tend to be more variable than those in the base-paired regions. This type of pattern is anticipated because there should be fewer selective constraints on unpaired nucleotides than there are on paired nucleotides. As one would expect, NAVA allows the different types of data overlays to be combined with one another (E). Finally, data legends are used to represent the numeric values corresponding to the colors used by the different overlays.

Paired sites association ranking

A paired sites association ranking test available from the "Analysis" menu ranks individual substructures belonging to a given structure overlay by comparing degrees of numeric values obtained from a user-specified data source at paired and unpaired nucleotide positions. For each substructure two lists of numeric values are compared: one corresponding to paired nucleotide positions and one corresponding to unpaired nucleotide positions. The two lists are compared using a Wilcoxon-rank sum test which assesses statistically whether values in the paired sites list are significantly higher (or lower) than values in the unpaired sites list. This test produces a z-score and a corresponding p-value for each substructure indicating the degree of association, for example: a substructure with a large negative z-score and a p-value < 0.05 would indicate that paired sites are significantly associated with smaller numeric values and, conversely, that unpaired sites are significantly associated with larger numeric values. Substructures are ranked based on the degree of detected association by sorting substructures on the p-value obtained using the Wilcoxon-rank sum test, where lower p-values indicate more statistically significant associations (either significantly lower or higher numeric values than expected at paired sites).

Substructure ranking

A substructure ranking test is available from the "Analysis" menu in NAVA which ranks individual substructures belonging to a given structure overlay by comparing degrees of numeric values obtained from a user-specified data source in a given substructure to numeric values in the full sequence. For each substructure two lists of numeric values are compared: one corresponding to data values in the substructure and one corresponding to data values in the full sequence. The two lists are compared using a Wilcoxon-rank sum test which assesses statistically whether values in the substructure are significantly higher (or lower) than values in the full sequence list. This test produces a z-score and a corresponding p-value for each substructure indicating the degree of association. For example, a substructure with a large negative z-score and a p-value < 0.05 would indicate that numeric values within the substructure are significantly lower than numeric values in the full sequence. Substructures are ranked based on the degree of detected association by sorting substructures on the p-value obtained using the Wilcoxon-rank sum test, where lower p-values indicate higher degrees of significance, i.e. greater deviations from the expectation (either significantly lower or significantly higher data values within the substructure than are expected by chance).

Substructure conservation analysis

Measuring structural similarity

The function of a nucleic acid molecule is largely determined by its structure. Consequently, if two or more molecules share a similar structure it is likely that they share a similar function (Hofacker *et al.*, 1998; Collier *et al.*, 2002; Pedersen *et al.*, 2006). There are a number of secondary structure similarity metrics available that allow one to measure structural similarity (V Moulton *et al.*, 2000). In the context of nucleic acid secondary structure these metrics are most commonly used to benchmark the predictive accuracy of secondary structure prediction programs. In other words, given a set RNA sequences and set of "true" secondary structures for these (typically determined by experimental techniques), these metrics can be used to test how accurately a given secondary structure prediction matches the true structure (Anderson *et al.*, 2013). A different use of similarity metrics, which this section will focus on, is detecting structural conservation amongst two or more nucleic acid secondary structures.

To assess the similarity between two or more secondary structures, some form of distance or similarity metric is required. Similarity metrics are typically normalised to be values between 0 and 1 (or 0% and 100%), where 1 indicates that two structures are identical and 0 indicates that two secondary structures are completely distinct. Distance metrics can often be transformed into equivalent similarity metrics by applying an appropriate normalisation. For consistency and ease of understanding only similarity measures will be used throughout this section. The similarity metric employed depends on the type of secondary structure information that is available, for example: it may be that one only has information about which nucleotides are base-paired within a particular sequence; on the other hand, one may have complete information about the base-pairing probabilities between all pairs of nucleotides within a sequence. It is most frequently the case that only information about which nucleotides are predicted to base-paired within a particular sequence is available. This type of information is typically represented using dot-bracket notation (see Figure 2.12 for a description). Given that this is the most commonly available secondary structure information, the metrics discussed below will be limited to only those that use this type of base-pairing information.

..((...))...(((...)))

Figure 2.12: Dot bracket representation for nucleic acid secondary structures. Matching brackets represent base-paired nucleotide positions and dots represent unpaired (singled-stranded) nucleotide positions. The two brackets coloured in blue are an example of a matching pair of open- and closed- brackets corresponding to a base-pairing in the sequence.

An unsophisticated metric between two secondary structures would be to just count the proportion of base-pairs that match (Vincent Moulton *et al.*, 2000). Clearly, this is quite a coarse-grained approach as it requires base-pairs to match exactly, however, this is the predominant method used for benchmarking secondary structure predictions. The way this metric is used in benchmarking is to measure sensitivity and positive predictive value (PPV), which are defined as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

A more straight-forward interpretation of sensitivity and PPV is as follows:

Sensitivity = number of correctly predicted base-pairings (TP) compared to the total number of base-pairings in experimental structure (TP + FN).

PPV = number of correctly predicted base-pairings (TP) compared to the total number of predicted base-pairings (TP + FP).

It is trivial to increase the sensitivity of a particular secondary structure prediction method by inflating the number of base-pairings it predicts (i.e. over-predicting base-pairs), however, this will on average result in an increase in the number false positives (FP) and a corresponding decrease in PPV. Therefore any particular method should aim to score high on both metrics, for this reason the two metrics are "averaged" by taking their harmonic mean (termed the F-score):

$$F\text{-score} = \frac{2}{\frac{1}{\text{Sensitivity}} + \frac{1}{PPV}} = \frac{2 \times TP}{(2 \times TP) + FN + FP}$$

The F-score provides a fairer test for benchmarking predictive accuracy, although both sensitivity and PPV are usually reported alongside the F-score in benchmarking studies in order to more fully describe the behavior of a particular method.

One can see how these measures could also be used for detecting structural conservation between two structures by treating one structure as the experimental structure and one as the predicted structure. However, it could also be argued that considering only exact base-pair matches (as these measures do) is not appropriate in the context of determining structural similarity, as one might also want to consider base-pairings that are shifted up or down in the

sequence in order to take account of overall structural features that might be necessary for a particular function.

Mountain metrics

The mountain similarity metric is a measure of secondary structure similarity based on the mountain representation of secondary structures, which leads naturally from the dot-bracket representation for secondary structures (see Figure 2.13 for a description of the mountain representation).

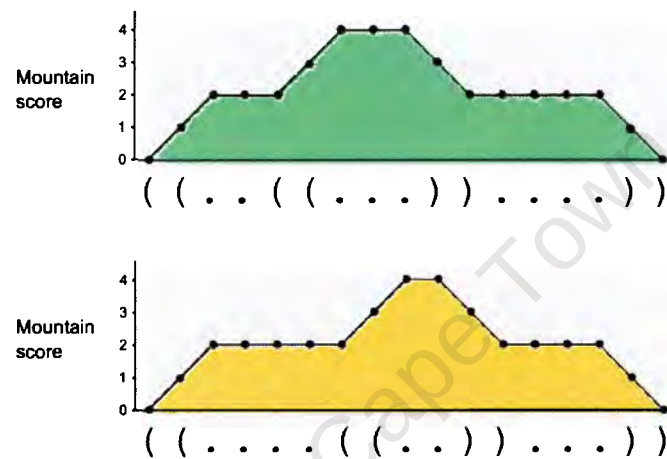


Figure 2.13: Mountain representation for secondary structures. Two secondary structures represented using their dot-bracket representations and corresponding mountain representations. The mountain score at a particular sequence position corresponds to the number of base-pairs that enclose that position. For example, the third sequence positions in both structures are enclosed by two base-pairs and therefore both of those positions have a mountain value of 2. Note that the two structures presented in the figure are qualitatively similar although their central base-pairings are offset from one another.

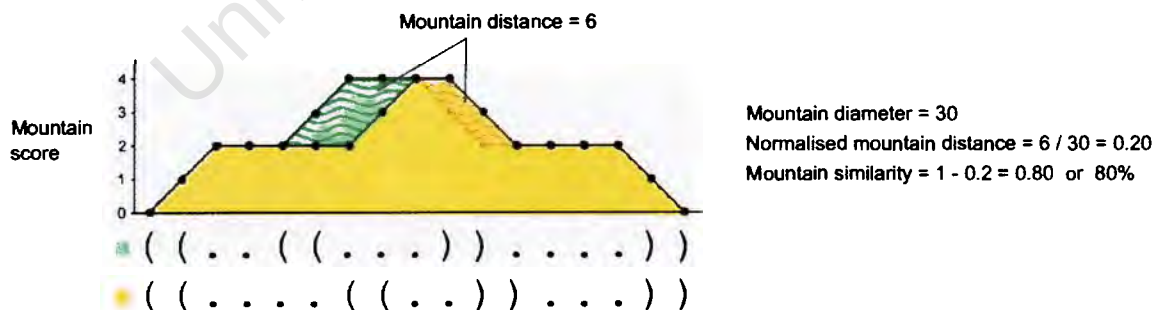


Figure 2.14: Mountain similarity calculation. The mountain distance can be calculated for any two secondary structures of equal length (i.e. equal number of nucleotides). By overlaying the mountain plots corresponding to each of the two structures the mountain distance is calculated as the amount of non-overlapping area (the wavy area in the figure). The mountain distance is then divided by the mountain diameter which acts a normalising constant. The mountain diameter is the maximum possible mountain distance between two structures of a given length. In this example each structure has a length of 17 nucleotide and a corresponding mountain diameter of 30 (there is a simple formula for calculating the mountain diameter as a function of structure/sequence length; Vincent Moulton *et al.*, 2000). The normalised mountain distance can be converted to a mountain similarity by subtracting the normalised mountain distance from 1, in this example $1 - 0.20$ gives a mountain similarity of 0.80 or 80%.

Using the mountain representations for secondary structures it is easy to graphically derive the mountain similarity metric (see Figure 2.14 for an explanation, for the mathematical exposition refer to Vincent Moulton *et al.*, 2000). The mountain similarity metric provides a far more nuanced measure of structural similarity than the simple base-pair metrics described above, because it better captures structural features than base-pair metrics that heavily penalize base-pairings that are slightly offset in the sequence, for example: shifting the central base-pairings between two secondary structures as in Figure 2.14 would be heavily penalized by simple base-pair metrics (resulting in an F-score of 0.5 or 50% in this example), whereas the mountain similarity metric captures that fact that these two secondary structures are qualitatively similar (resulting in a mountain similarity measure of 0.80 or 80%): a desirable property when considering structural conservation.

Identification of conserved substructures

For small secondary structures (<300 nucleotides in length) testing whether two secondary structures are conserved (or similar) is fairly straightforward using the metrics described above. However, for larger structures such as those formed by single stranded RNA and DNA virus genomes (for example, HCV has a genome of ~9600 nucleotides in length), measures of structural similarity become less useful because it is often the case that only a few regions of the analysed sequences will be structurally conserved. However, it will often be the case that individual base-pairings between nucleotides that are close to one another in the primary sequence (<300 nucleotides apart) will be better preserved than those between more separated nucleotides. This should result in the occurrence of identifiable groups of base-pairings within what we have referred to above as substructures. Discrete substructures often have particular functions such as, for example, the Internal Ribosomal Entry Sites (IRESs) present in HCV and other viruses (Tsukiyama-Kohara *et al.*, 1992; Mokrejs *et al.*, 2010) and the cis-acting replication elements (CREs) found in picornaviruses (Mason *et al.*, 2002) and other viruses (Raman and Brian, 2005; Paul *et al.*, 2000). Given the conserved nature and potential functional relevance of such substructures, a reliable means of identifying these within the context of large (>1000 nt long) secondary structures would be very useful.

Mountain similarity measure for substructures

In order to detect conserved substructures the mountain similarity measure was extended to consider substructures. Given a pair of substructures, the mountain similarity measure produces a value from 0.0 - 1.0 indicating the similarity between the two structures. A mountain similarity of 1.0 indicates that the two structures are identical and a mountain similarity of 0.0 indicates they are completely distinct.

The specific problem addressed here was measuring the similarity between two substructures, denoted $s_{1(a,b)}$ and $s_{2(a,b)}$, that belong to a pair of aligned parent structures s_1 and s_2 , respectively, where a indicates the start position of a substructure in the full-length sequence (or structure) and b indicates the end position in the full-length sequence (or structure). Extending the mountain similarity measure to the case of substructures (i.e. $M_{sim}(s_{1(a,b)}, s_{2(a,b)})$) is trivial, it simply involves considering only the subset of nucleotide positions belonging to that substructure in the full structure. However, a problem that arises when considering substructures is the possibility of nucleotide positions within a substructure being paired with nucleotides positions outside of a substructure, i.e. in the full structure. These "hanging base-pairs" can lead to negative values of the mountain similarity measure, however, this is not considered problematic as the mountain measure is still bounded above by 1 and in testing it was found that values of the mountain similarity measure for substructures were negative only when substructures were particularly dissimilar - a fact implying that such substructures could be safely ignored for the purposes of detecting structural conservation.

Conservation permutation test

Although the mountain similarity measure provides a similarity score between 0.0 and 1.0 indicating the similarity between two structures or two substructures, an interpretation of this score is not completely straightforward. Whereas for bigger substructures mountain similarity measures of 0.7 might denote significant degrees of similarity, for smaller substructures a score of 0.7 might be non-significant. In order to provide a more meaningful interpretation of mountain similarity measures a permutation test was devised that calculates a p-value indicating the degree to which similarity between two substructures within a given genome $s_{1(a,b)}$ and $s_{2(a,b)}$, denoted $M_{conservation}(s_{1(a,b)}, s_{1(a,b)})$, is expected, using the substructures at other positions in the sequence as the null distribution.

Given two substructures $s_{1(a,b)}$ and $s_{2(a,b)}$, the lengths of $s_{1(a,b)}$ and $s_{2(a,b)}$ (denoted $|s_{1(a,b)}|$ and $|s_{2(a,b)}|$, respectively) are assumed to be much less than the length of the full structures. For the purposes of this study it was assumed that there was at least one order of magnitude difference in length between the substructures and the full structures. The test works by generating n random permutations of $s_{1(a,b)}$, where a permutation of $s_{1(a,b)}$ is denoted $s_{1^d(a,b)}$ and is equal to $s_{1(a+d, b+d)}$. Note that $|s_{1(a,b)}| = |s_{1^d(a,b)}| = |s_{2(a,b)}| = |s_{2^d(a,b)}|$. The d in the substructure notation indicates the random amount by which a substructure position has been shifted up or down in the full structure, i.e. the substructure takes the set of base-pairings at its new random position $(a + d)$ to $(b + d)$ in the full structure. For each of the n permutations the mountain similarity score is calculated between the permuted structure, $s_{1^d(a,b)}$, and $s_{2(a,b)}$, i.e. $M_{sim}(s_{1^d(a,b)}, s_{2(a,b)})$. The

fraction of permuted similarities ($M_{sim}(s_1^d(a, b), s_2(a, b))$) that are strictly greater than the actual similarity ($M_{sim}(s_1(a, b), s_2(a, b))$) is taken to be the p-value:

$$M_{conservation}(s_1(a, b), s_2(a, b)) = \frac{1}{n} \sum_{i=0}^n 1_M(s_1^d(a, b), s_2(a, b))$$

Where $1_M(s_1^d(a, b), s_2(a, b))$ is the indicator function, which returns a 1 if the permuted structure s_1^d is more similar to s_2 than the real structure s_1 is to s_2 , otherwise it returns a zero:

$$1_M(s_1^d(a, b), s_2(a, b)) = \begin{cases} 1, & M_{sim}(s_1^d(a, b), s_2(a, b)) > M_{sim}(s_1(a, b), s_2(a, b)) \\ 0, & \text{otherwise} \end{cases}$$

Note: It is ensured that the p-value generated by $M_{conservation}(s_1(a, b), s_2(a, b))$ is $\geq \frac{1}{n}$, by including $d = 0$ ($s_1^0(a, b) = s_1(a, b)$) as one of the random permutations, this is necessary in order to generate a correct p-value (otherwise it would be possible to obtain a p-value = 0). This implies that the minimum p-value is bounded below by $1/n$, so it is not possible to detect degrees of conservation below a particular p-value threshold for a given number of permutations. For the purposes of this study 1000 permutations were used, which allowed p-values as low as 0.001 to be detected.

Detecting pairwise secondary structure conservation

Now that a measure of structural conservation between two substructures has been defined it can be used to detect substructure conservation between a pair of aligned secondary structures s_1 and s_2 . This is achieved by running a sliding window of size w across a pair of secondary structures and measuring $M_{conservation}$ at each position. The window is initially placed at positions $(0, w)$ in the aligned pair of structures, s_1 and s_2 , and moved one nucleotide at a time to the end of the alignment, where each window is a substructure. This produces a list of p-values which provides the degree of structural conservation within each window along a pairwise alignment of two secondary structures. In this study a window size $w = 75$ was used for all analyses.

Alignment of secondary structures

To detect substructure conservation as outlined above an alignment of secondary structures is necessary, whether it be a pairwise alignment or a multiple alignment. In order to align a set of secondary structures, the corresponding sequences was taken and a multiple nucleotide sequence alignment performed using MAFFT (Kato and Toh, 2008). The dot bracket representation for each secondary structure was then taken and wherever a gap was inserted into the corresponding sequence during the multiple sequence alignment, a gap (or equivalently an unpaired base) is inserted into the secondary structure. Whilst, this is not a true structural

alignment based on purely structural information, it is suitable wherever sequences are similar enough to be aligned. For more diverse sets of sequences it is recommended that a structural alignment is performed, however, this is substantially more computationally demanding than a nucleotide sequence alignment (Eddy, 2002; Torarinsson *et al.*, 2007). After the alignment procedure any column positions of the alignment where 20% or more of the characters were gaps were masked - this prevented unpaired nucleotides that were created during the sequence alignment process from being detected as conserved stretches of unpaired nucleotides.

Detecting secondary structure conservation within and between different species

This method of detecting pairwise conservation can be extended to detect conservation within a group of 3 or more sequences. This is done by taking a multiple alignment of secondary structures and performing a pairwise secondary structure conservation test between every pair of structures, thereby generating multiple p-values for each pairwise test at each column position of the alignment. The pairwise p-values corresponding to each column position of the alignment are taken and the median p-value for each column calculated. The median p-value at each column is a measure of structural conservation within a group of sequences at a window surrounding that position. The interpretation of the median p-value at a position is as follows: a median p-value of p indicates that at least 50% of pairwise comparisons had a p-value of p or lower. For example, if it was found that the median p-value at a particular position was 0.05 this would mean that in at least 50% of pairwise comparisons, a conservation p-value of 0.05 or lower was detected at that position. This would be considered statistically significant evidence of a conserved substructure at that position.

Results

Nucleic Acid Visualisation and Analysis application

Presented here is an example analysis using NAVA and the SHAPE-inferred HIV-1 NL4-3 genomic secondary structure (Watts *et al.*, 2009). This example makes use of five pre-existing data sources namely: the HIV-1 NL4-3 SHAPE-inferred genomic secondary structure, the corresponding SHAPE reactivities, the NL4-3 sequence, the NL4-3 sequence annotations and a set of 50 HIV-1 subtype B sequences. NL4-3 is an HIV-1 subtype B sequence which is compared in this example to an alignment of 50 other subtype B sequences. In particular, this analysis is focused on the Rev Response Element (RRE) which is an exceptionally well-characterised and biologically relevant substructure in the HIV-1 *env* gene.

Out of the 50 sequences used in this example only the NL4-3 sequence has an experimentally determined secondary structure. Because of this, the PPfold program (Sükdösd *et al.*, 2011), which is integrated into NAVA, was used to computationally predict a consensus structure from the alignment of 50 HIV-1 subtype B sequences in order to provide further computational results for inclusion in the analyses. PPfold provides not only a consensus structure prediction, but also base-pairing probabilities which can give information about possible alternative base-pairing interactions. In this example both a visualisation of the NL4-3 RRE and a ranking of substructures in the NL4-3 genome were performed using experimentally- and computationally-derived data sources in order to demonstrate some of NAVA's key features.

Importing data

After creating a new project the first step is to import at least one data file into the "Data input" tab. Data files can be imported by 1) clicking "Import file" and using NAVA's built-in file browser or 2) dragging and dropping files from the user's computer. In this example five data files are imported by dragging them from a folder on the user's computer and dropping them into the "Project navigation" panel (Figure 2.15). Four of these files correspond to HIV-1 NL4-3 sequence, namely: the NL4-3 sequence itself, the SHAPE-determined NL4-3 secondary structure (Watts *et al.*, 2009), a table containing the SHAPE reactivities and a GenBank file containing annotations specific to the NL4-3 sequence. The fifth data file that was imported was a set of 50 HIV-1 subtype B sequences. When a data file is imported into NAVA, the data type of that file is automatically detected and it is subsequently considered a "data source" (Figure 2.16). Data files that are detected as alignment data sources are further automatically classified into 1 of 3 alignment types, namely: nucleotide, codon and amino alignments based on their contents. The user can change this selection from the GUI by selecting the data source in the navigator panel and selecting a new type from the "Data inspector" panel to the right.

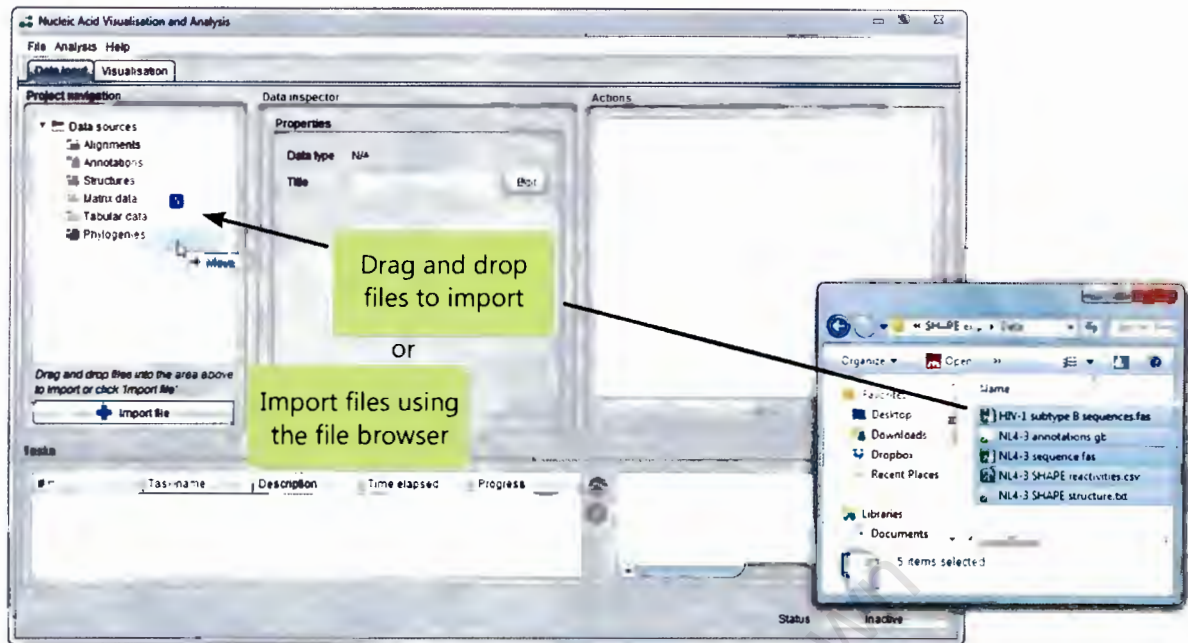


Figure 2.15: Importing data files. Data files can be imported from the user's computer either by clicking "Import file" and using the built-in file browser, or by dragging and dropping files into the "Project navigation" panel in the "Data input" tab.

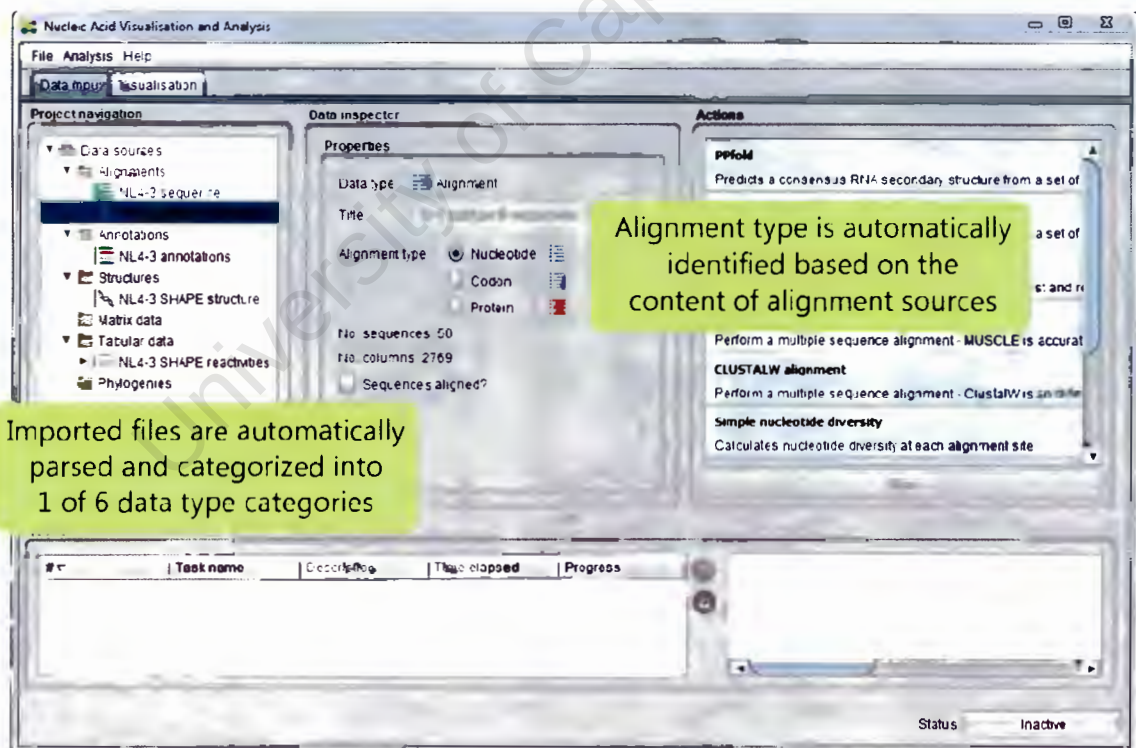


Figure 2.16: Automatic data type detection. When a data file is imported the content of that file is used to infer its data type and is subsequently categorized into 1 of 6 data type categories (alignments, annotations, structures, matrix data, tabular data and phylogenies). Alignments data files are further categorized into 1 of 3 alignment types: nucleotide, codon and protein alignments based on their contents.

Creating new data sources from existing data sources

NAVA has a number of built-in applications (available in the "Actions" panel in the "Data input" tab) that can be used to create new data sources from existing data sources. In this example a set 50 HIV-1 subtype B sequences that have been imported into NAVA were taken and first nucleotide aligned using the built-in MAFFT application (Kato and Toh, 2008). The generated alignment source was then selected and the built-in PPfold application was run in order to perform a secondary structure prediction. The resulting PPfold output files (a secondary structure, a base-pairing probability matrix and a table containing site-by-site pairing probabilities) appear in the "Project navigation" panel after the PPfold application has completed running (Figure 2.17). Application statuses and any output can be viewed as they are running in the "Tasks" panel in the lower half of the GUI. Tasks are multi-threaded so that single applications can take advantage of multiple CPU cores (if available on the user's computer) and also such that multiple tasks can run in parallel at the same time. If the number of threads reserved for a given task exceeds the number of available CPU cores, tasks will wait in a task queue until a thread slot becomes available. Permissible actions/applications are filtered when the user selects a particular data source.

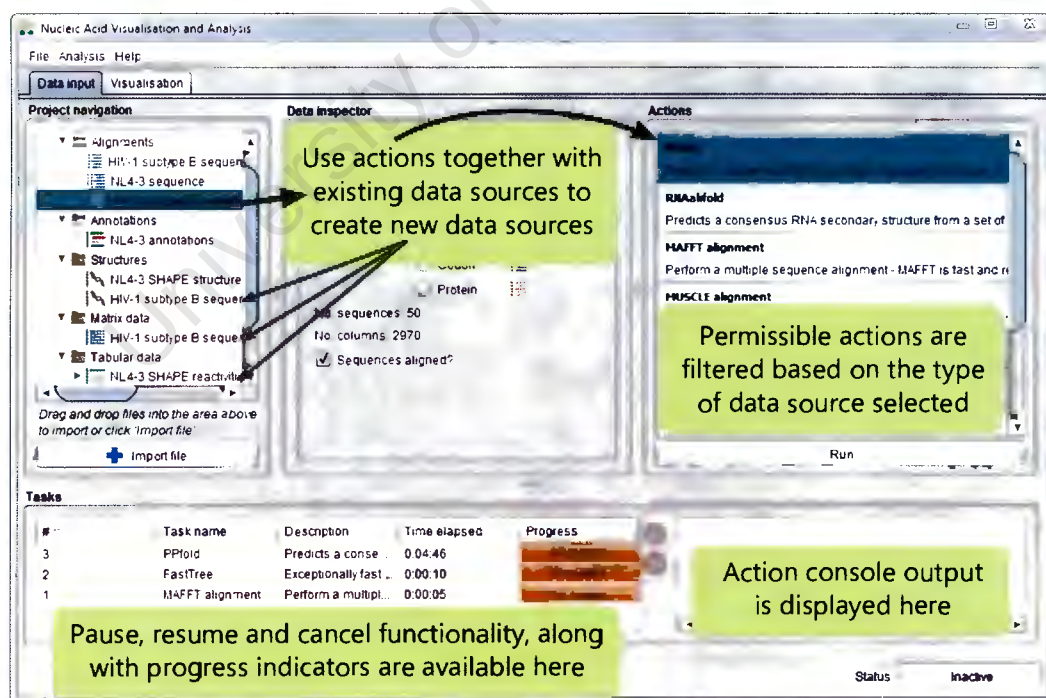


Figure 2.17: NAVA applications. NAVA has a number of built-in applications that can be used to create new data sources from existing data sources. In the example above, an alignment source (50 MAFFT-aligned HIV-1 subtype B sequences) was used in conjunction with the built-in PPfold application to perform a secondary structure prediction. The resulting output files (in this case: a secondary structure, a base-pairing probability matrix and a table containing site-by-site pairing probabilities) appear in the "Project navigation" tree panel once the PPfold application has completed running.

Substructure visualisation

Once a secondary structure data source is available in the "Data input" tab the user can visualise substructures within that secondary structure by switching over to the "Visualisation" tab. The first step in doing so is to create a structure overlay from a secondary structure data source that the user wishes to visualise. In this example the SHAPE-inferred NL4-3 HIV-1 secondary structure source is used together with three other data sources, namely: NL4-3 SHAPE reactivities, base-pairing probabilities generated by a PPfold structure prediction based on an alignment of 50 HIV-1 subtype B sequences and nucleotide variation data corresponding to the same alignment of 50 HIV-1 subtype B sequences. These different data sources are visualised by first creating a corresponding data overlay for each which is then automatically mapped to the structure overlay by NAVA. Various options are available to specify how data overlays are overlaid on the secondary structure, for example: how data should be mapped to the structure and the colour range used to denote different data values. These options are available when adding a data overlay and can be modified at any time. In this example we have replicated the colour legend used in the Watts *et. al.* 2009 study which experimentally inferred the NL4-3 secondary structure using the SHAPE methodology.

By inspecting the resulting visualisation (Figure 2.18) of a portion of the RRE substructure it is apparent that base-paired nucleotides tend to have lower SHAPE reactivities than unpaired nucleotides. This result is anticipated as low SHAPE reactivities were used to infer base-paired nucleotides within the NL4-3 sequence. When comparing the computationally inferred base-pairing probabilities (HIV-1 subtype B pairing probabilities) to the base-pairings inferred by SHAPE it is evident there is a strong correspondence between these (red lines in Figure 2.18). This correspondence suggests that HIV-1 subtype B sequences share a structural conformation that is highly similar to that inferred by the SHAPE experiment for NL4-3.

It is, however, also noteworthy that some of the computationally inferred base-pairing probabilities do not correspond to the SHAPE-inferred base-pairings. This may reflect the inaccuracy of the computationally predicted base-pairing probabilities, an imprecise mapping between the two data sources or, most plausibly: alternative base-pairing interactions. Although PPfold assumes that each of the 50 sequences used to determine base-pairing probabilities has the same set of secondary structure configurations, this is not a biologically realistic scenario. It is likely that the 50 sequences used in the PPfold structure prediction have some base-pairing interactions that differ both from one another and from the SHAPE inferred NL4-3 secondary structure. Given that most of the 50 HIV-1 subtype B sequences used in this analysis were found circulating in nature and are probably viable, it is expected that if these alternative PPfold

inferred RRE base-pairing interactions really do exist, then they likely do not have a large impact on HIV viability.

Finally, sequence logos used to depict nucleotide variation visually demonstrate that nucleotide positions with high degrees of variability tend to occur within the loop regions of the depicted substructure. This result makes sense given that unpaired nucleotide positions are expected to have more freedom to vary without disrupting the overall structural conformation.

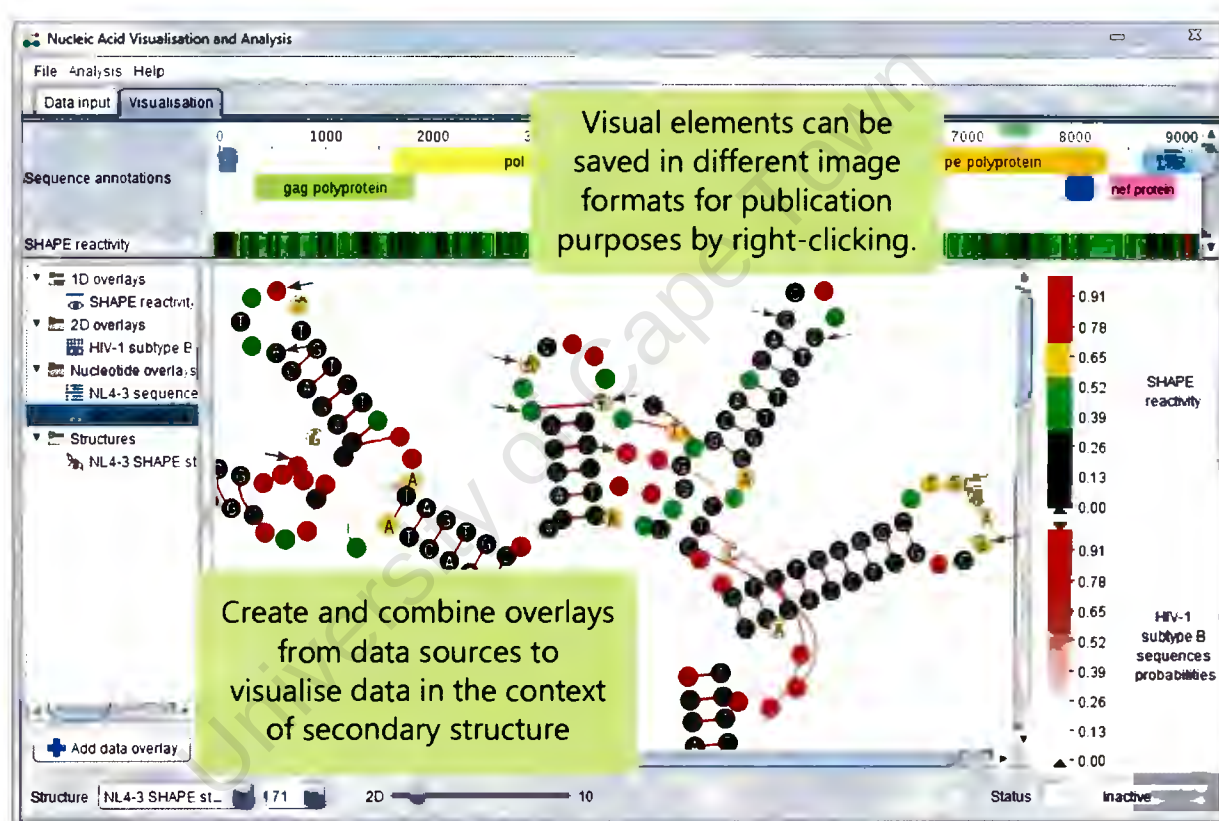


Figure 2.18: Substructure visualisation. In this example a portion of the Rev Response element (RRE) in the SHAPE-inferred HIV-1 NL4-3 genomic secondary structure is visualised in NAVA. The secondary structure has a one-dimensional data overlay corresponding to SHAPE reactivities (coloured nucleotides), a two-dimensional data overlay depicting the base-pairing probabilities of 50 inferred HIV-1 subtype B genomic secondary structures (red lines joining pairs of nucleotides), and a nucleotide composition overlay representing nucleotide variation derived from an alignment of the 50 HIV-1 subtype B sequences (depicted using sequence logos within the circles - nucleotides with high degrees of variation are annotated using gray arrows on this figure). All visualisations, including the current substructure, the data legends, the graphical colour bar and annotations can be saved in raster (.png) and vector (.svg and .emf) graphics formats for publication purposes.

Analysis of substructures

There are currently two ways of quantitatively analysing secondary structures in NAVA. These are called the "Paired sites comparison ranking" and "Substructure ranking" tests. Whereas the paired sites comparison ranking test compares sequence-wide user-specified data source values at paired and unpaired nucleotide positions, the substructure ranking test compares these same user-specified data source values between individual substructures and the remainder of the sequence. In this example we perform both a paired sites comparison ranking test and a substructure ranking test using data values corresponding to SHAPE reactivities at individual nucleotide positions.

The results of the paired sites comparison reveal that SHAPE reactivities are significantly lower at paired nucleotide positions than at unpaired nucleotide positions throughout the entire NL4-3 SHAPE structure (full sequence, Figure 2.19, $p\text{-value} = 6.49 \times 10^{-115}$ and $z\text{-score} = -57.385$), which is expected given that nucleotides that are unpaired are meant to display higher degrees of SHAPE reactivity than those which are base-paired. Interestingly, the substructure with the most significant difference in SHAPE reactivities between paired and unpaired nucleotides (substructure ID 72, $p\text{-value} = 4.71 \times 10^{-22}$ and $z\text{-score} = -11.892$) is the RRE.

The substructure ranking test was then used to rank substructures in the NL4-3 structure based on their associated degrees of SHAPE reactivity. As expected based on the results of the paired sites comparison ranking test, the substructure with an associated SHAPE reactivity that deviated most from that of the remainder of the HIV genome was the RRE (substructure ID 72, $p\text{-value} = 1.48 \times 10^{-15}$ and $z\text{-score} = -8.708$). Such low degrees of SHAPE reactivity within the RRE suggest that it is highly stable and are consistent with, the RRE being one of the most biologically important RNA structural elements within the HIV-1 genome.

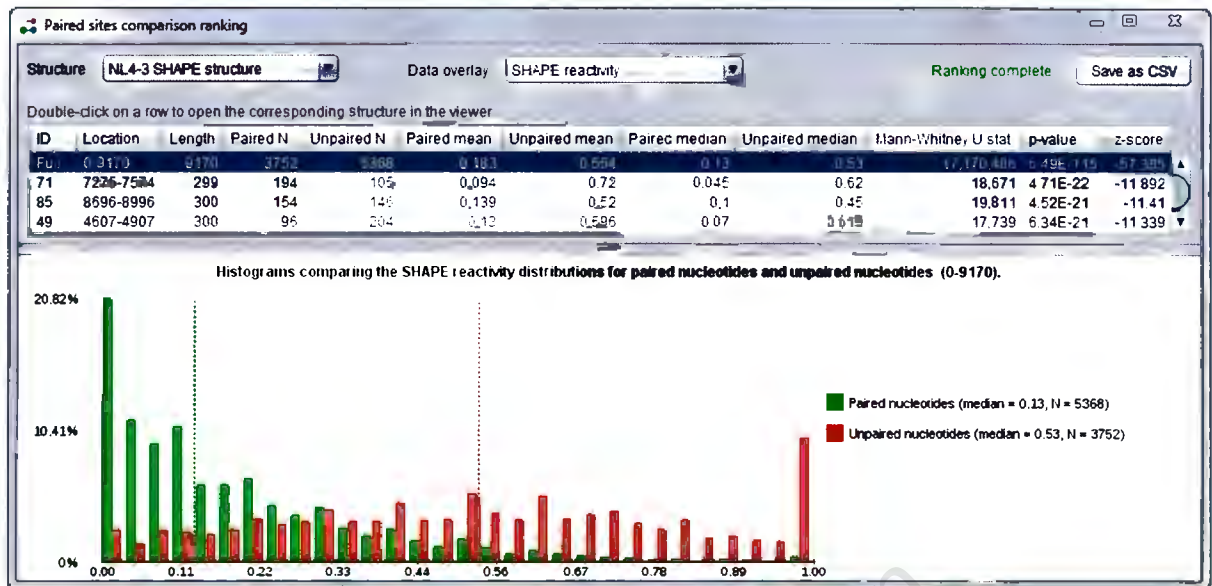


Figure 2.19: Paired sites comparison ranking. This example reveals that SHAPE reactivities are significantly lower at paired nucleotide positions than at unpaired nucleotide positions throughout the entire NL4-3 genome, which is entirely expected given that high SHAPE reactivity values are an indication of nucleotide accessibility and that these values were used to infer the NL4-3 secondary structure.

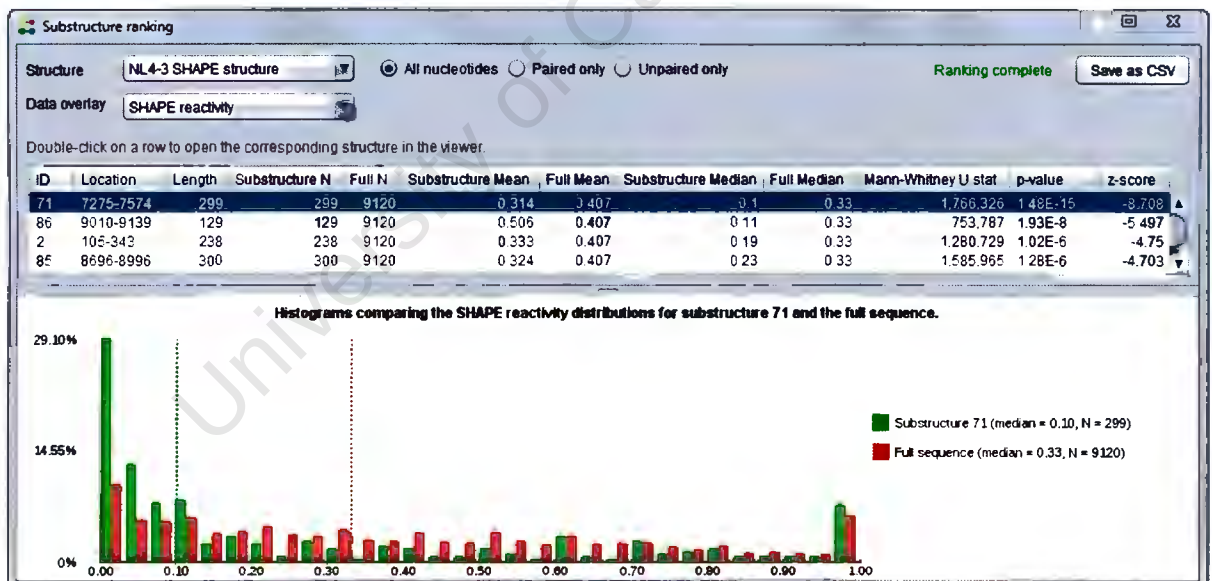


Figure 2.20: Substructure ranking. Using the NL4-3 SHAPE-inferred structure and the corresponding list of SHAPE reactivities, the ranking analysis ranks the HIV-1 RRE element (ID 72) as the substructure having the lowest over-all degree of SHAPE reactivity out of 86 substructures identified in the HIV-1 NL4-3 genome. A low degree of SHAPE reactivity suggests that the RRE is the most stable substructure in the NL4-3 genome, which is to be expected given that the RRE is one of the most biologically important RNA elements in the HIV-1 genome.

Detection of conserved substructures in HIV-1

In order to illustrate the usefulness of the substructure conservation analysis tool, the methods described in this chapter were used to detect conserved substructures amongst different groups of HIV-1 sequences. Three groups were analysed: HIV group 1M sequences (containing the viruses driving the global HIV epidemic), HIV group 10 (a divergent group of HIV variants not usually seen outside of West-central Africa) and HIV-2 (an even more divergent group of HIV variants also mostly found in West Africa; Reeves and Doms, 2002). The phylogenetic relationships between these three groups are depicted in the phylogenetic tree in Figure 2.21. Each of these groups consisted of 50 sequences, providing 150 sequences in total. Secondary structure predictions were performed independently for each sequence using an optimized version of UNAFold that significantly reduces the amount of time taken to fold a sequence (Rizk and Lavenier, 2009; Markham and Zuker, 2008) and the sequences and corresponding secondary structures were aligned using the MAFFT alignment tool (Kato and Toh, 2008). Three different comparisons were made using the substructure conservation analysis. The first looked at substructure conservation only within HIV group 1M sequences (Figure 2.21, top panel), the second looked at substructure conservation between HIV group 1M sequences and HIV group 10 sequences (Figure 2.21, middle panel), and the final looked at substructure conservation between in HIV-1 (Groups 1M and 10) versus more distantly related sequences from HIV-2 (Figure 2.21, bottom panel). Conservation was calculated using the conservation permutation test and the degrees of conservation across the different groups of HIV-1 groups are graphed using the conservation z-scores which are calculated by the test (note that although the test itself generates p-values, these are presented here using an equivalent z-score representation). Also depicted is the mountain similarity within each of the groups at each position along the genome. It is difficult to discern patterns of structural conservation using the mountain similarity metric - for this reason the conservation permutation test is used to provide a more meaningful indication of structural conservation. The rationale for plotting z-scores instead of the equivalent p-values is that z-scores provide a more natural way of graphically plotting degrees of significance, where lower z-scores indicate more significant degrees of structural conservation.

Substructure conservation across HIV species

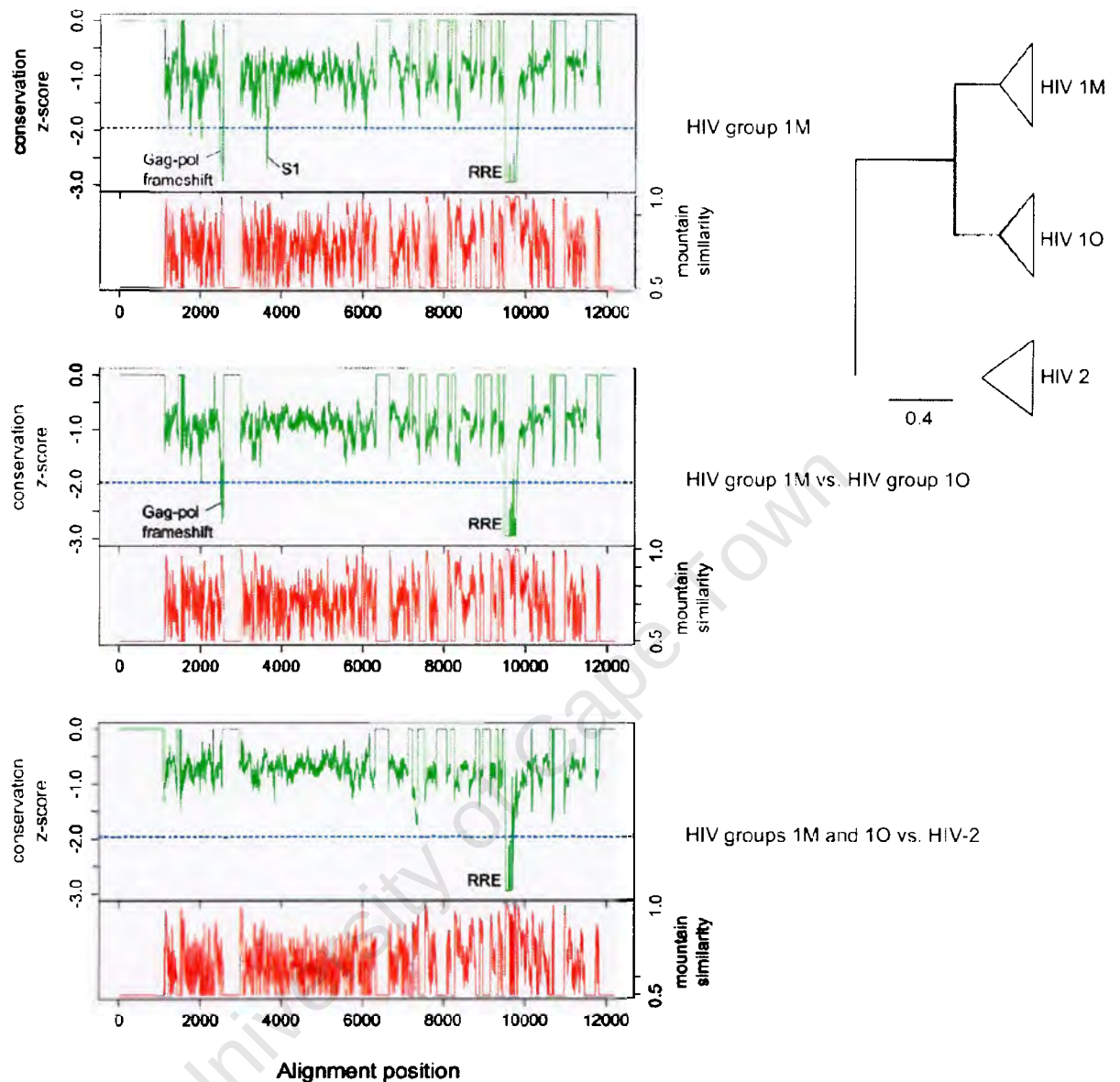


Figure 2.21: HIV-1 and HIV-2 genomic RNA substructure conservation analysis. Secondary structures were analysed for evidence of structural conservation within and between various evolutionarily distinct groups of HIV sequences (HIV group 1M, HIV group 10 and HIV group 2). The bottom half of each graph (mountain similarity) represents the 75 nucleotide long sliding window mountain similarity scores centered at each position along the genome. Mountain similarity values close to 1.0 represent high degrees of structural similarity and values <0.6 represent high degrees of structural dissimilarity. The upper half of each graph (conservation z-score) depict the results of the conservation permutation test across the entire input alignment, positions on the graph below the threshold z-score of -1.96 (dashed blue line, equivalent to a p-value of 0.05) are considered to be structurally conserved within a 75 nucleotide window surrounding that position. Three substructures were detected as conserved within HIV group 1M sequences (top panel): the Gag-pol frameshift, an uncharacterised substructure (S1) and the Rev Response Element (RRE). Between HIV group 1M and HIV group 10 sequences (middle panel) only the Gag-pol frameshift and the RRE display evidence of structural conservation at the selected 75nt long scanning resolution. Between the more diverged HIV group 1M, group 10 and HIV-2 sequences (bottom panel) only the RRE was detectably conserved at the 75nt scanning resolution.

The results of the first analysis revealed three detectably conserved substructures amongst HIV group 1M sequences (Figure 2.21, top panel). These three substructures corresponded to the RRE (Rev Response Element), the Gag-pol frameshift element and a substructure, S1, which appears to have not been previously. Between HIV group 1M and group 10 sequences we found that both the Gag-pol frameshift element and the RRE were detectably conserved, whereas the uncharacterised S1 structure was not (Figure 2.21, middle panel). The third and final analysis which looked at HIV sequences from group 1M and 10 versus more distantly related HIV-2 sequences found that only the RRE was detectably conserved (Figure 2.21, bottom panel). This is just one example of how this substructure conservation analysis can be used to detect conserved genomic substructures. In Chapter 3 a more complete analysis of genomic secondary structure conservation in HIV and number of flaviviruses is performed.

Conclusions

A software tool, called NAVA, was developed to simplify the inference, visualization and analysis of nucleic acid secondary structures. NAVA has a graphical user interface that assists in the import and generation of data, the visualization of secondary structures with overlaid data and the identification of biologically functional secondary structural elements within nucleic acid sequences approximately 10,000 nucleotides long. NAVA aids in the identification of structural elements with likely biological functions using statistical tests that can rank substructures based on, for example, their associated degrees of conservation or the nucleotide substitution characteristics of their constituent nucleotides. Visual elements, including structure visualisations, and data elements can be exported in wide variety of formats.

This chapter demonstrated how NAVA, together with experimentally- and computationally-inferred data sources, could be used to provide insights into substructures present in the genome of HIV. For example, experimentally-determined SHAPE reactivities were used to rank 81 substructures in the HIV-1M NL4-3 sequence and determined that the RRE was the substructure with the lowest over-all degree of SHAPE reactivity. This suggests that the RRE is not only one of the most stable substructures in the NL4-3 genome, but also corroborated the validity of the ranking analyses given that the RRE is known to be crucial to the survival of HIV. In addition to this, visual inspection of nucleotide compositions at individual nucleotide positions derived from an alignment of HIV-1 subtype B sequences, demonstrated that nucleotide positions within single-stranded regions of the RRE tolerate higher degrees of nucleotide variability than base-paired regions.

In addition to NAVA, a methodology for identifying conserved substructures amongst a set of aligned secondary structures was developed along with a graphical tool for doing so. The tool quantifies degrees of structural conservation at positions along an alignment of secondary structures by performing pairwise structural comparisons. This facilitates the detection of conserved substructures amongst groups of related sequences. The usefulness of this tool was demonstrated by the automated identification of two well-characterised biologically functional substructures within the HIV genome (the gag-pol frameshift element and the RRE) and a third substructure (S1) which does not appear to have been previously characterized. By providing an automated means of identifying such conserved substructures, we anticipate that this tool will assist researchers in focusing their experimental analysis of genomic substructures on those that are most likely to have biological relevance.

Given that both the substructure conservation tool and NAVA provide related functionality it is intended that the substructure conservation analysis will eventually be directly incorporated into NAVA so as to allow researchers to automatically generate a structure overlay from an alignment of nucleotide sequences so as to allow identification and direct visualization in NAVA of substructures with significant degrees of structural conservation.

It is additionally intended that NAVA be extended to enable cross-platform support. Currently, most of the built-in applications bundled with NAVA run only under the Windows operating system, implying that much of NAVA's functionality is available for use on Windows only. However, by compiling and bundling versions of these applications capable of running under the Linux operating system it should be relatively straightforward to make a Linux version of the program.

In addition to cross-platform support, additional applications will be added to NAVA that will further facilitate secondary structure analyses. In particular it is anticipated that the most computationally intensive of these new tools, such as parametric maximum likelihood coevolution and substitution rate estimation methods will, rather than running directly on the user's computer, run on remote cloud computing resources accessible over the Internet. These cloud-based applications will have the added advantage of automatically allowing NAVA to support multiple operating systems, as the application will run on a pre-configured machine which may be running a different operating system to that on the user's computer, yet will still be able to communicate with the version of NAVA installed on the user's computer via Java's networking features. Ideally, the most computationally intensive of these third party applications would be specifically configured to take advantage of parallelization offered by remote computer clusters and graphical processing units.

3 Biologically relevant secondary structures in single-stranded RNA viruses

Introduction

Large single-stranded nucleotide sequences such as cellular RNA molecules and single-stranded DNA and RNA virus genomes, form complex secondary structures through Watson-Crick base-pairing between bases on the same DNA/RNA strand. Both computational folding (Markham and Zuker, 2008) and experimental techniques (Watts *et al.*, 2009) are available for determining the structural conformations of such nucleic acid molecules, and these have been used to identify many thousands of potentially functional secondary structural elements within gene transcripts and viral genomes. To date, however, only a small subset of these elements has been proven to have any biological functionality. Rather than this being because biologically functional secondary structures are rare, it is because the techniques that are used to identify secondary structures typically give little or no indication of whether the identified structures are biologically important.

Besides using mutational studies to directly test the fitness impacts of secondary structural disruptions (Parkin, M Chamorro, *et al.*, 1992; Honda *et al.*, 1996; Vassilaki *et al.*, 2008), it is also possible to use comparative structural and evolutionary analyses of related sequences to computationally infer the biological relevance of secondary structures (Hofacker *et al.*, 2002). For example, evidence that a secondary structural element is conserved within a set of homologous sequences can provide compelling evidence that the element is biologically relevant (Hofacker *et al.*, 2002; Semegni *et al.*, 2011). In addition, the evolutionary processes operating both to conserve these structures and to maintain the coding potential of the nucleic acid sequences from which they are composed, can also be used to identify structures with likely biological relevance (Tuplin *et al.*, 2002, 2004; Simmonds *et al.*, 2004; Davis *et al.*, 2008; Ngandu *et al.*, 2008; Sanjuán and Bordería, 2011; Simmonds *et al.*, 2008).

A large-scale analysis was carried out to identify potential biologically functional secondary structures within the single-stranded RNA genomes of Human Immunodeficiency Virus (HIV) and seven different flavivirus species: Hepatitis C virus (HCV), Dengue virus (DENV), West Nile virus (WNV), Japanese Encephalitis virus (JEV), Tick-borne Encephalitis virus (TBEV), Bovine Viral Diarrhea virus (BVDV) and Classical Swine Fever virus (CSFV). Secondary structures were inferred computationally for each individual sequence in each virus dataset and these were used to broadly study the relationships between secondary structure and evolutionary

dynamics. Overall, detectable differences in patterns of mutation and selection between base-paired and non-base-paired sites suggested that many of the studied genomes contained large numbers of biologically functional structural elements. Detailed analyses were performed on individual HIV, HCV, DENV, WNV, JEV, and TBEV structural elements that were both particularly well conserved between related virus lineages and displayed evolutionary dynamics that were strongly suggestive of their having some selectively advantageous biological function.

Methods

Dataset construction

An automated pipeline was written in the Java programming language for generating virus datasets. This pipeline takes as input a GenBank file containing a large number of entries for a specific virus species (100 - 2000 entries), where each entry typically consists of a full genome nucleotide sequence (or a large portion thereof) and the corresponding gene locations within each sequence. The GenBank files were downloaded from NCBI nucleotide database (<http://www.ncbi.nlm.nih.gov/nuccore/>) based on their organism classification, for example: one GenBank file contained all HIV group 1M sequences in NCBI, whereas another contained just the HIV group 1M subtype B sequences. The pipeline scans the entries and automatically selects a list of genes that are common to the majority of sequences in a particular GenBank file by using the provided gene annotations. At least 75% of the entries were required to contain a specific gene annotation in order for that gene annotation to be added to the overall list of gene annotations for a given file. All entries that did not contain the complete set of gene annotations were discarded. In runs of this pipeline, this procedure discarded no more than 20% of entries in datasets that were processed. Any entries containing modified, patent, or chimaeric sequences were discarded based on keywords contained in the GenBank file in order that we only had naturally occurring sequences in our dataset. For each GenBank file a number of sequence alignments were generated: one alignment containing the full-length sequences (full-alignments) and separate alignments containing nucleotide gene sequences (gene-alignments) corresponding to each of the individual genes within a particular virus species. From each of these full- and gene-alignments, subsets of these alignments were generated based on different size cut-offs, for example: subsets containing 50, 100, 200, 300 and 400 sequences for both the full- and gene-alignments. The sequences in each of these smaller alignments were chosen in order to maximise the total amount of sequence diversity, for example: sequences very similar or identical to other sequences in the dataset were likely to be discarded. This was done in

order to maximise the amount of information contained within each of these smaller alignments. These smaller more information-dense alignments were generated because certain downstream analyses have alignment size restrictions due to computational limitations. The alignments containing the full-alignments were automatically nucleotide-aligned using MAFFT (Kato and Toh, 2008) and the gene-alignments were translated to protein alignments and protein-aligned using MAFFT and then back-translated into codons once again in order to obtain codon alignments for of each gene-alignment. Once this automated process was completed, any dataset that was used in a downstream analysis was manually curated to remove any obvious alignment errors or unsuitable sequences that were missed by the automated pipeline.

Prediction of nucleic acid secondary structures

Secondary structures were computationally predicted for individual sequences for every virus dataset, consisting of 100-200 sequences each, using an optimized version of UNAFold's hybridssmin program (Markham and Zuker, 2008; Rizk and Lavenier, 2009). This optimized folding program is faster on single-core usage and is also parallelised to take advantage of multi-core CPUs thus enabling a much large number of sequences to be folded in a relatively short amount of time than the original hybridssmin. All virus sequences were folded using a linear RNA conformation at 37°C using the program's default settings.

Inference of nucleotide substitution rates

HyPhy (Hypothesis testing of phylogenies, Kosakovsky Pond *et al.*, 2005) was used to analyse substitution rates at individual nucleotide positions in several of our datasets. To infer phylogenetic trees for each of the alignment datasets, FastTree (Price *et al.*, 2010) together with a general time reversible nucleotide substitution model with 20 rate categories (GTR-CAT) was used. The model used to analyse substitution rates in our HyPhy analyses was also a GTR model.

Inference of synonymous substitution rates

Synonymous substitution rates were inferred for individual codon sites within each of the codon-aligned gene-alignments. Given a particular amino acid, there are typically a number of synonymous codons that can encode that amino acid (for example: arginine, leucine and serine are each encoded by six different codons). One might expect therefore that where an amino acid can be encoded by more than one codon, the different synonymous codon options would on average be used with roughly equal frequency to encode a particular amino acid. However, in biological systems the rate of synonymous codon substitutions can often deviate for a variety of reasons, for example: selection acting to maintain the underlying nucleotide sequence in a particular region can result in a decrease in the rate of synonymous codon substitutions

because of the fitness advantages provided by maintaining certain patterns of nucleotides in that region.

To infer synonymous substitution rates FUBAR (A Fast, Unconstrained Bayesian AppRoximation for inferring selection; (Murrell *et al.*, 2013), a tool publicly available on the Datamonkey server (Delpont *et al.*, 2010), was used. FUBAR takes a codon alignment of sequences and a phylogenetic tree and infers both synonymous substitution rates and non-synonymous substitution rates at individual codon sites in the alignment. FUBAR uses the Muse and Gaut 1994 codon model (Muse and Gaut, 1994) in order to model synonymous and non-synonymous substitution rates. Typical implementations of this model are slow due to having to perform likelihood optimizations over many possible combinations of non-synonymous and synonymous substitution rate values. FUBAR on the other hand is an approximate algorithm which pre-computes values of the synonymous and non-synonymous rates on a fixed grid, greatly speeding-up the analysis. The accuracy of FUBAR compares well with non-approximate methods, which typically can only work with smaller alignments (<100 sequences). Because FUBAR is fast, it is fairly easy to run a large number of datasets and datasets with a large number of sequences, which means that our study was able to make better use of the vast number of sequences that are available in public databases for the viruses being analysed. Using more sequences means making use of more information and therefore maximizing the accuracy of the inferred synonymous substitution rates.

It is likely that many of the sequences being analysed here were recombinant. It was therefore necessary to take steps to explicitly account for recombination within these analyses (D. Martin *et al.*, 2011). Accordingly, prior to performing selection analyses the Genetic Algorithm for Recombination Detection (GARD; Kosakovsky Pond *et al.*, 2006) method was used to detect recombination breakpoint sites. These sites were then used to partition the alignment into “mostly recombination free” (it is unlikely that every recombination event was detected and accounted for) sub-alignments. For each of these sub-alignments a phylogenetic tree was inferred and these trees were collectively used as inputs for FUBAR.

For each gene-alignment FUBAR provides a list of inferred synonymous substitution rates corresponding to each codon position. Synonymous substitution rates can be used as evidence of selection acting on regions of the nucleotide sequence. This type of information is useful because one way in which selection can act to maintain a particular nucleic acid substructure at a position is by maintaining the pattern of nucleotides in the region of the substructure, thereby maintaining the nucleotide interactions that comprise that substructure. Low synonymous substitution rates provide evidence of such selection (see Figure 3.1 for a description of this

process). This has successfully been used as evidence of selection acting to preserve secondary structures in a number of studies of virus genomes (Tuplin et al., 2004; Simmonds et al., 2008; Sanjuán and Bordería, 2011). Evidence of low synonymous substitution rates is particularly indicative of selective maintenance of the underlying nucleotide sequence within protein-coding regions as opposed to evidence of nucleotide conservation, which is strongly influenced by selection acting at the protein-coding level.

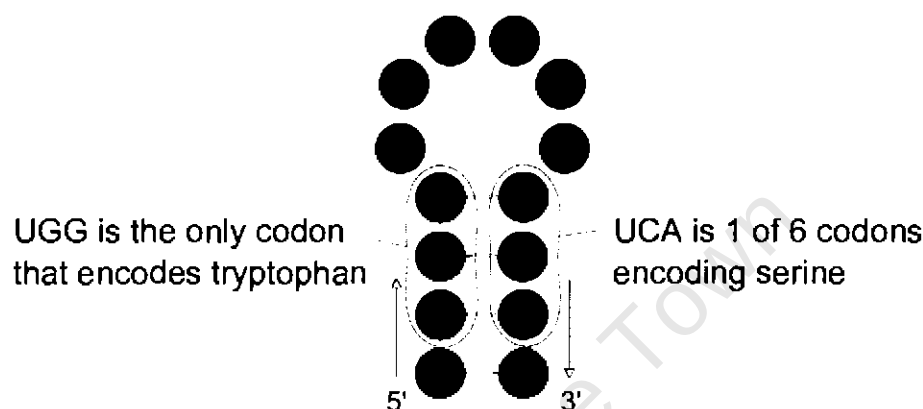


Figure 3.1: Constraints imposed by the selective maintenance of a secondary structure on synonymous codon substitutions. This figure depicts a secondary structure in a coding region of a sequence. For illustrative purposes two codons are circled that base-pair with one another in the structure. The codon on the left (UGG) encodes tryptophan - the only codon in the genetic code that encodes tryptophan. The codon on the right (UCA) encodes the amino acid serine which can be potentially encoded by six synonymous codons (UCU, UCC, UCA, UCG, AGU, AGC). Out of the six possible codons encoding serine, there are only two that will maintain the amino acid at that position and the exact pattern of Watson-Crick base-pairing complementarity seen in this structure, those codons being UCA and UCG, this is because both of these codon triplets can base-pair with UGG in the correct order. One might therefore expect that if both the protein sequence and the secondary structure were under strong selective maintenance that these two codons would occur at a much higher frequency at this position in a set of homologous sequences, even though there are six codons that encode the amino acid encoding serine. This type of selection can be indirectly detected by measuring the rate of codon substitutions within a codon alignment of sequences using methods such as FUBAR (Murrell *et al.*, 2013) or PARRIS (Scheffler *et al.*, 2006), where a low synonymous substitution rate might indicate selection acting to maintain the nucleotide sequence in order to maintain the secondary structure configuration at that position.

Inference of non-synonymous substitution rates

For each gene-alignment FUBAR also provides a list of inferred non-synonymous substitution rates corresponding to each codon position. Non-synonymous substitution rates indicate the rate at which substitutions at codon positions result in an amino acid change (i.e. a non-synonymous change) at that position.

Inference of nucleotide coevolution

	AT	TA	CG	GC	AA	AC	...
AT	*	0	0	0	$\beta\pi_A/\lambda$	$\alpha\pi_C/\lambda$	...
TA	0	*	0	0	$\beta\pi_A/\lambda$	0	...
CG	0	0	*	0	0	0	...
GC	0	0	0	*	0	$\mu/4\lambda$	...
AA	$\beta\pi_T\lambda$	$\beta\pi_T\lambda$	0	0	*	$\beta\pi_C$	...
AC	$\alpha\pi_T\lambda$	0	0	$\alpha\pi_C\lambda$	$\beta\pi_A\lambda$	*	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 3.2: Muse 95 paired character model. The Muse model is represented by a modified HKY85 rate matrix for paired character evolution. Entries in the matrix represent the rate of change from one pair of nucleotides to another. Rates involving two nucleotides changes in one step are considered impossible and are therefore represented by zeros. A * in a row or column implies that the row or column sums to 1. λ is used to account for pairing and appears in the numerator when a change involves going from an unpaired state to a paired state (e.g. AA → AT) and in the denominator when going from a paired state to an unpaired state. Paired states are those that correspond to Watson-Crick base-pairs (GC and AU base-pairs) and the GU wobble-base-pair. π_i represents the base frequencies (where $i=A, C, G$ or T). Adapted from Muse 1995.

A coevolution HyPhy script, kindly provided by Art Poon, was used to detect nucleotide coevolution consistent with evolutionary maintenance of base-pairing within secondary structures. Given a nucleotide alignment and a corresponding phylogenetic tree, the script is able to detect coevolution of nucleotide pairs by estimating rates of nucleotide substitution at pairs of alignment sites. For each nucleotide pair the method compares the Muse (1995) paired character model against an independent sites model – HKY85. The Muse paired character model is a modified HKY85 model, which instead of using the standard 4x4 nucleotide substitution matrix of the HKY85 model, employs a 16x16 transition rate matrix. Each element in this matrix represents the relative probability of changing from one pair of character states to another (e.g. AA → AT, CT → CA, etc.) (see Figure 3.2). Where paired sites are favoured (in a stem, for example) it is expected that a change from a paired state to an unpaired state will occur at a lower frequency than a change from a paired state to a paired state. In order to account for such pairing the model introduces a pairing parameter λ . Rates in the matrix which are associated with the formation of a pairing (i.e. changes that result in a complementary nucleotide pair, e.g. AG → AT) are multiplied by λ and those that result in a mispairing (e.g. AT → CT) are multiplied by $1/\lambda$. When fitting this model to a given nucleotide alignment and tree, pairs of sites favouring evolution of base-pairing should have an associated $\lambda > 1$. Setting $\lambda = 1$ in the Muse paired character model reduces it to a HKY85 independent sites model. This nesting permits one to perform a likelihood ratio test (LRT) by comparing the likelihood scores obtained for each model. In a situation where evolution of paired sites is favoured, the Muse model should yield a significantly higher likelihood score than the HKY85 model. Where neither avoidance nor

formation of paired sites is favoured, the two models should yield similar likelihood scores. The LRT allows one to calculate a p-value which indicates whether the Muse model provides a significantly better fit to the data. A low p-value indicates the null model (HKY85) is rejected in favour of the Muse model and therefore a model of coevolution is favoured for the pair of sites under consideration (Muse 1995).

In much the same way that recombination can undermine selection analyses, it can also undermine coevolution analyses (D. Martin *et al.*, 2011). When assessing the possibility of coevolution of sites within regions exchanged by recombination, the coevolution analysis performed here will tend to incorrectly infer that nucleotide polymorphisms within the exchanged fragment (i.e. nucleotide differences between the incoming and outgoing fragments) are all substitutions that occurred at the same time and which are therefore much more likely to be coevolving.

It was therefore necessary to directly account for recombination in the coevolution analyses that were carried out. In order to do this, recombination detection was performed on each of our full datasets using RDP 4.20 (Martin *et al.* 2010) (with default settings) and a recombination-free alignment was produced within which portions of recombinant sequences derived from different parental sequences were separated into two sequences. This procedure results in an expanded alignment containing more sequences than were present in the input dataset. A 500 nucleotide sliding window was then moved across the recombination-free expanded alignment in one nucleotide steps and the N longest sequences from each window selected, where the length of a sequence is defined as the number of non-gap characters it contained and N is the number of sequences in the original alignment prior to recombination analysis. The first 500 nucleotides from each of the sequences in each window were then saved in a separate alignment file. In order to reduce the total number of files and to reduce the number of trees computed, adjacent windows with the same N longest sequences were combined, for example: if the windows starting at positions 10 - 15 contained the same N longest sequences, nucleotides 10 - 515 (15 + window size) from those N sequences in the full-alignment were saved to the same alignment file. Following that, all sequences which were identical to at least one other sequence in the alignment partition were discarded, leaving one remaining copy. This was done because the presence of duplicate sequences provided no extra information and therefore would only serve to slow down the analyses. For each of the generated alignment files maximum likelihood trees were inferred using FastTree (Price *et al.*, 2010).

Coevolution analyses were then performed in HyPhy (Pond *et al.*, 2005) using the coevolution script for each separate alignment file and its corresponding tree. Nucleotide pairs where at least one nucleotide in a pair was contained in an invariant column (a column where all non-gap characters were the same) were ignored during the analyses, because detecting nucleotide coevolution at sites where no mutations have been observed is not possible. Ignoring these nucleotides also greatly increased the speed of the analysis.

Paired sites association test

An association test was designed to measure how numeric values are associated with paired and unpaired sites amongst a given set of secondary structures. The test works by taking an alignment of secondary structures and counting the number of times a particular column position in the alignment is paired across all the secondary structures, if 50% or more of the positions in a particular column of the alignment are paired, that column is classified as a "paired site" otherwise that column is classified as an "unpaired site". A set of numeric values (e.g. nucleotide substitution rates) is then mapped to the same alignment of secondary structures, i.e. each column of the alignment is assigned a numeric value. A comparison between numeric values that are assigned to the "paired sites" and "unpaired sites" sets is performed using a Wilcoxon-rank sum test which assesses statistically whether values associated with the paired sites set are significantly higher (or lower) than values assigned to the unpaired sites set. This test produces a z-score and a corresponding p-value indicating the degree of significance with a large negative z-score and a p-value < 0.05 indicating that paired sites are significantly associated with smaller numeric values and, conversely, that unpaired sites are associated with larger numeric values. For example, using this test one can assess whether paired sites tend to have lower (or higher) nucleotide substitution rates in comparison to unpaired sites (see Figure 3.3).

Distribution of nucleotide substitution rates at paired and unpaired sites in HIV 1M subtype C sequences

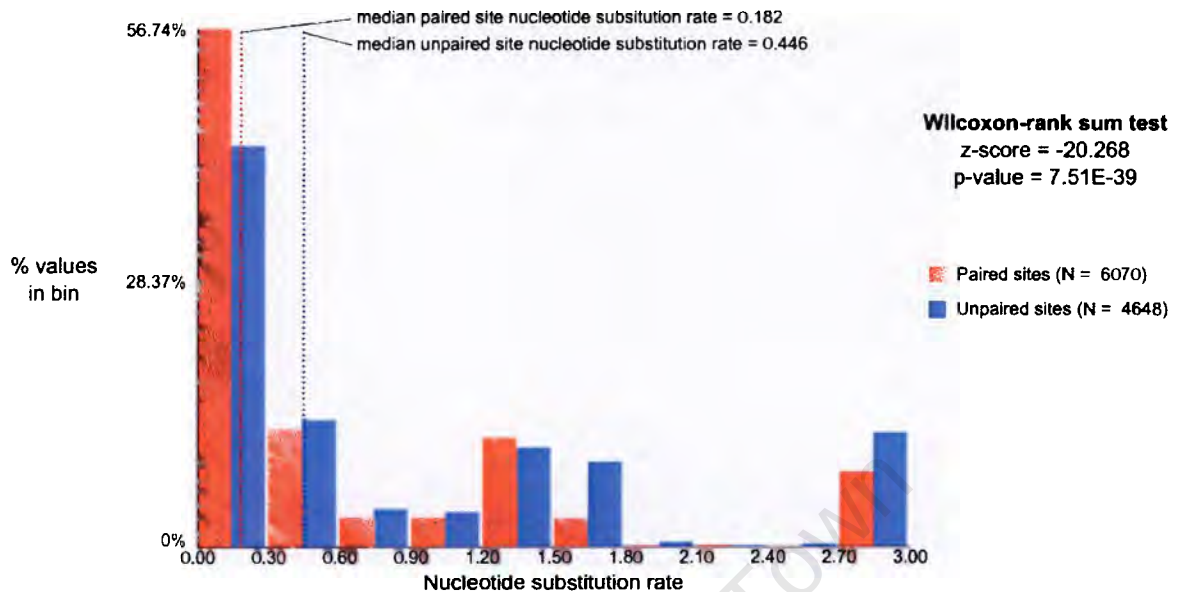


Figure 3.3: Example demonstrating the "Paired sites association test". This histogram illustrates how the distributions of inferred nucleotide substitution rates at paired and unpaired nucleotide positions in HIV 1M subtypes C sequences are compared in order to detect associations between nucleotide substitution rates and paired and unpaired sites. It is visually apparent from that the distribution of nucleotide substitution rates corresponding to paired sites tends to contain lower values (i.e. is left-shifted) than the distribution of nucleotide substitution rates corresponding to unpaired sites (median rate = 0.182 for paired sites and median rate = 0.446 for unpaired sites) in HIV 1M subtype C sequences. The difference between these distributions is tested using a Wilcoxon-rank sum test which is a statistical test analogous to a Student's t-test, but which does not assume that that data are distributed normally, i.e. it is a non-parametric test. This test provides a z-score indicating directionality (i.e. whether values in the paired set tend to be lower than values in the unpaired set) and a p-value indicating whether the difference is statistically significant.

Tests of synonymous and non-synonymous substitution rates at paired and unpaired codon positions

Codon positions within gene-alignments were analysed for degrees of lower (or higher) than expected synonymous or non-synonymous substitution rates at paired and unpaired codon positions using a Wilcoxon-rank sum test in a similar manner to that outlined in the "Paired sites association test" described above. Five different tests were performed: firstly a separate test was performed for each of the three nucleotide positions within a codon (Tests 1 - 3 corresponding to codon positions 1, 2 and 3, respectively). A separate test for each nucleotide in a codon was performed because each of the three nucleotides in a codon can be classified as being paired or unpaired at the same time, whereas there is only a single synonymous and non-synonymous rate assigned to each codon. A nucleotide position was classified as being paired if at least 50% of secondary structures in the alignment were paired at that position; otherwise it was classified as being unpaired. Test 4, the "All codon positions" test, considered the combined

set of codon positions 1, 2 and 3: a factor that might statistically invalidate this test because the same rate is repeated 3 times for each codon. A fifth test, the "Nucleotides paired ≥ 2 " test, accounted for this problem by counting the number of paired nucleotides within a codon, thereby classifying each codon as either "paired" or "unpaired"; if 2 or 3 nucleotide positions were paired (using the same pairing criteria as above) within a particular codon that codon was considered to be "paired", otherwise that codon was considered to be "unpaired".

Paired sites correlation test

A paired sites correlation analysis was designed in order to test whether rates (be they synonymous substitution rates or nucleotide substitution rates) at base-paired sites within a secondary structure were correlated with one another (see Figure 3.4). One might expect that sites which are physically base-paired with one another might have substitution rates that are detectably correlated. Factors that could give rise to such correlations might include: (1) selective processes that act similarly at both positions of a particular base-pair within a structure given that that both positions are under similar degrees of selective maintenance; and (2) basal mutational processes which have a similar effect on both positions due to the close physical proximity of base-paired nucleotides.

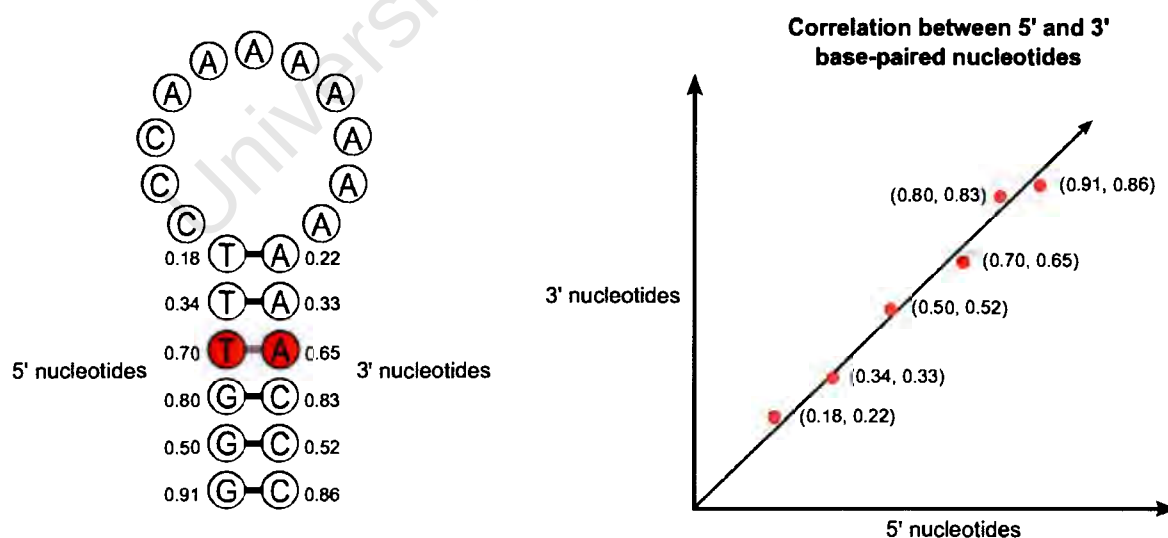


Figure 3.4: Detecting correlations between base-paired sites. The above figure gives an example of the types of correlations the "Paired sites correlation analysis" aims to detect. Given a particular set of numeric values assigned to individual nucleotide positions within a structure, e.g. nucleotide substitution rates (in this example only the numeric values at base-paired positions are displayed), a Spearman's rank correlation test is performed which calculates whether the numeric values of base-paired nucleotides are correlated. In particular, correlations between the numeric values of 5' nucleotides in a base-pairing and their 3' prime base-pairing partners are tested. In this example there is a strong positive correlation between 5' nucleotides and their 3' base-pairing partners.

Correlations between base-paired nucleotides were tested using two different tests: a simple regression test and a permutation-based regression test. The simple regression test was performed by taking each of the two numeric values corresponding to all base-paired nucleotide positions within a predicted secondary structure and computing the Spearman's rank correlation, which provided both a correlation value and a p-value (Spearman, 1904). The Spearman's rank correlation test is similar to the standard Pearson's linear correlation test, but does not assume that the two variables being considered are linearly correlated. Instead it assumes the more general case where the relationship between the two variables can be described by a monotonic function, i.e. a function that either increases or decreases, but not necessarily linearly.

Using this simple regression test, the detected correlations were highly significant on most of the datasets tested. Whilst this is clearly interesting, a potential problem with the detected correlations may be that they are simply capturing the fact that base-paired nucleotides tend to be separated from one another in the sequence by only a small number of nucleotides (empirically one can actually clearly observe that base-paired nucleotides have a strong tendency to be close to one another in the sequence; Supplementary Figure 2). This would result in detectable correlations between the nucleotide substitution rates of paired nucleotides when, for example, nearby nucleotides simply have similar nucleotide substitution rates.

In order to test whether the rate correlation was due to the nearness of corresponding base-paired positions within the sequence or whether there was also a contribution from the physical base-pairing of sites in the predicted secondary structures, a permutation test was designed. The permutation test works as follows: pairs of rates corresponding to each pair of base-paired positions in the inferred secondary structure were listed and the Spearman's rank correlation calculated as in the simple regression test. To account for the proximity of sites corresponding to their distance in terms of number of nucleotide between them, the ordered lists of rates were taken and permuted simply by shifting them in order up- or down-stream by a random number of positions in the sequence. Spearman's rank correlations were then re-calculated with these repositioned rates where both the nucleotide distances between paired nucleotides and regional variations in nucleotide substitution rates are maintained but where the correspondence between base-pairing and nucleotide substitution rates at base-paired sites is broken. Calculating the Spearman's rank correlation for each of 2000 randomly permuted lists produces a p-value for the null hypothesis that rates at paired sites are correlated solely due to their close proximity and not due to physical base-pairing. This can be calculated as the proportion of times that the correlation coefficients determined for the permuted datasets is greater than that determined for the non-permuted datasets.

Both the simple regression test and the permutation test were performed separately on each computationally inferred secondary structural element that was inferred for a dataset, which resulted in multiple results for each dataset. The reported result for a particular dataset was the result containing the median p-value, implying that at least 50% of results for a particular dataset showed similar or strong signals of association as the reported result.

Testing for coevolution of base-paired sites

Pairs of columns in the full-genome alignments of HIV, HCV, JEV and DENV were analysed for evidence of maintaining base-pairing complementarity via nucleotide coevolution. Coevolving pairs of columns were inferred using the methods described in "Inference of nucleotide coevolution". A pair of columns where nucleotide coevolution favouring base-pairing complementarity was inferred with a p-value of 0.05 or smaller was considered "coevolving", otherwise the pair was considered "not coevolving". For each alignment a coevolution matrix was formed where each element i,j in the matrix corresponded to a pair of columns i,j in the alignment. An element i,j in each matrix was filled with a 1 if a particular column i,j was inferred to be coevolving complementarily or -1 otherwise. Only pairs of column i,j where $|i-j| \leq 100$ were considered due to the lengthy running time of the coevolution analysis, implying that elements in the matrix where $|i-j| > 100$ were empty (these cells were filled with zeros to indicate this). Secondary structures corresponding to each sequence in a particular alignment were used to generate a base-pair count matrix for each alignment of secondary structures. Pairs of nucleotide positions in the alignments that were base-paired in at least 25% of the sequences were considered "paired", otherwise they were considered "unpaired". The nucleotide coevolution matrix and the base-pair count matrices for each virus dataset were overlaid and the number of simultaneously "paired" and "coevolving", "paired" and "not coevolving", "unpaired" and "coevolving", and "unpaired" and "not coevolving" pairs of positions were counted, whilst empty cells were not included in the counts. These counts were used to form a chi-squared contingency table, where expected values were calculated and a chi-square tested performed in order to specifically test whether base-paired positions tend to be coevolving complementarily more than can be accounted for by chance.

Detection of conserved nucleic acid substructures

The methods described in "Chapter 2: Substructure conservation analysis" were used to detect substructures that were conserved between various groups of viruses. The viruses were compared in two different ways: HIV, HCV and DENV sequences were each analysed separately, i.e. only virus sequences within each particular species were compared against one another, whereas JEV, TBEV and West Nile virus, which are from different but related species, were placed in the same group and compared against one another. Comparisons between specific subtypes were also performed (e.g. HIV 1M subtype B vs. HIV 1M subtype C) for particular virus species, but comparisons within specific subtypes were not because one would expect high degrees of structural conservation within the same subtype due to sequence similarity alone. Evidence of structural conservation amongst more distantly related subtypes provides compelling evidence that selective maintenance of a particular substructure has taken place, which is more suggestive of a substructure's biological relevance, given that it appears to have been maintained since the last common ancestor of the two diverged viral lineages.

For each group of viruses analysed a large set of aligned secondary structures (75 - 200 structures) was used. Each conserved substructure that was detected was assigned a z-score by the substructure conservation analysis which gives an indication of the extent to which a substructure is conserved. A substructure with a large negative z-score (less than -1.96, corresponding to a p-value < 0.05) is considered to be significantly conserved.

Large-scale investigation of nucleic acid secondary structures

A number of different analyses were used to investigate the large-scale relationships between mutation and secondary structure in HIV and a number flavivirus species (namely: HCV, DENV, WNV, JEV, TBEV, BVDV and CSFV). Three different types of associations with secondary structures were investigated: nucleotide substitution frequencies, synonymous codon substitution rates and nucleotide coevolution. Previous studies have shown that single-stranded parts of structured nucleic acid molecules are more prone to chemical damage (Lindahl and Karlstrom, 1973; Lindahl and Nyberg, 1974; Loeb and Preston, 1986) and strand-cleavage (Parthasarathi *et al.*, 1995) and therefore one would expect to find that base-paired nucleotide positions within structured nucleic acids will tend to undergo less frequent mutation. In addition to this, base-paired positions that are part of biologically relevant secondary structures may be expected to accumulate fewer observable nucleotide substitutions than unpaired positions due to the constraints imposed by selective processes acting to preserve interactions that maintain biologically relevant secondary structures. Selection favouring the maintenance of

base-pairing interactions within secondary structures could also potentially be detected by analysing synonymous substitution rates at codon sites that contain constituent nucleotides that are inferred to be base-paired within secondary structures. One would expect codons that contain nucleotides that are base-paired within biologically important secondary structural elements to accumulate fewer synonymous substitutions than those that have no base-paired nucleotides. Selection favouring the maintenance of biologically important secondary structures might additionally result in a situation where such base-paired nucleotides tend to detectably coevolve with their pairing partners. Specifically, the deleterious effect that a mutation at a given site may have on a base-pairing interaction within a biologically important secondary structure could be reversed by the occurrence of a compensatory mutation at the other site within the base-pairing.

Detailed investigation of nucleic acid secondary structures

An investigation was performed in order to identify biologically relevant conserved substructures within a number of single-stranded RNA viruses (namely: HIV, HCV and DENV) and between three different flavivirus species (the mosquito-borne flaviviruses WNV and JEV and the tick-borne flavivirus, TBEV). Regions of structural conservation (i.e. conserved substructures) were identified and patterns of mutation (e.g. degrees of synonymous substitution rates or nucleotide coevolution) within identified substructures and associations between these were tested to determine the potential biological relevance of a particular substructure. These tests were performed in a similar fashion to the large-scale analysis, but were restricted to the regions of the substructures in question. Individual substructures and data were visualized using NAVA (described in Chapter 2) and graphical representations of structures were edited using the Inkscape program (<http://sourceforge.net/projects/inkscape/>).

Results and Discussion

Large-scale analysis of secondary structures in various RNA viruses

Associations between nucleotide substitution rates and paired sites

Virus species were analysed for genome-wide associations between paired nucleotide positions and lower than expected (or higher than expected) nucleotide substitution rates (Table 3.1) using the "Paired sites association test". In 19 out of 22 cases base-paired nucleotide positions were found to have lower than expected nucleotide substitution rates ($p\text{-value} < 0.05$, $z\text{-score} < -1.96$). In the 3 remaining cases statistically significant differences could not be found supporting either lower or higher than expected rates. Studies have shown that single-stranded parts of nucleic acid molecules are more prone to chemical damage (Lindahl and Nyberg, 1974) and strand-cleavage (Parthasarathi *et al.*, 1995). In light of these studies, single-stranded nucleotide positions are expected to have higher basal mutational rates than base-paired nucleotide positions, which is precisely what our results support across almost all our virus datasets. In addition to single-stranded nucleotide positions having higher basal mutations rates, there is another possible effect which one should consider, namely selection acting to preserve base-paired positions within biologically important secondary structures. One expects selection acting to preserve a particular base-pairing to constrain the nucleotide possibilities at the base-pairing positions, thus reducing the observable nucleotide substitution rates at those positions.

Table 3.1: Analysis of nucleotide substitution rates at paired and unpaired nucleotide positions in various viral coding sequences.

Dataset	Paired nucleotide rate median	Unpaired nucleotide rate median	z-score	p-value ¹
Human Immunodeficiency Virus (HIV) group 1M	0.368	0.453	-2.035	4.18×10^{-2}
Human Immunodeficiency Virus (HIV) group 1M subtype b	0.179	0.413	-13.721	1.74×10^{-25}
Human Immunodeficiency Virus (HIV) group 1M subtype c	0.182	0.446	-20.268	7.51×10^{-39}
Human Immunodeficiency Virus (HIV) group 1M subtype d	0.432	0.585	-12.610	3.21×10^{-23}
Human Immunodeficiency Virus (HIV) group 1M subtype g	0.230	0.230	-10.306	1.63×10^{-18}
Human Immunodeficiency Virus (HIV) group 10	0.216	0.378	-11.374	1.07×10^{-20}
Human Immunodeficiency Virus (HIV) group 2	0.233	0.276	-10.203	2.64×10^{-18}
Hepatitis C Virus (HCV) genotype 1	0.208	0.209	-0.503	0.615
Hepatitis C Virus (HCV) genosubtype 1a	0.077	0.216	-4.227	2.37×10^{-5}
Hepatitis C Virus (HCV) genosubtype 1b	0.163	0.184	-3.202	1.36×10^{-3}
Hepatitis C Virus (HCV) genotype 2	0.234	0.234	0.372	0.710
Hepatitis C Virus (HCV) genotype 6	0.288	0.198	1.189	0.234
Dengue (serotypes 1- 4)	0.356	0.559	-9.983	7.41×10^{-18}
Dengue serotype 1 (DENV1)	0.216	0.237	-4.520	1.90×10^{-43}
Dengue serotype 2 (DENV2)	0.191	0.191	-0.32	1.80×10^{-43}
Dengue serotype 3 (DENV3)	0.310	0.347	-7.658	6.18×10^{-54}
Dengue serotype 4 (DENV4)	0.353	0.378	-5.698	6.20×10^{-50}
West Nile virus (WNV)	0.123	0.323	-7.727	2.71×10^{-15}
Japanese encephalitis virus (JEV)	0.302	0.462	-10.895	1.40×10^{-8}
Tick-borne encephalitis virus (TBEV)	0.215	0.215	-5.974	2.35×10^{-9}
Bovine viral diarrhea virus (BVDV)	0.889	0.889	-10.915	9.28×10^{-20}
Classical swine fever virus (CSFV)	0.320	0.320	-22.792	$5.30E \times 10^{-44}$

¹Viral lineages where nucleotide substitution rates at paired sites are significantly lower than those at unpaired sites (negative z-score and p-value < 0.05) are highlighted in blue. The p-values are calculated using a Wilcoxon-rank sum test, which compares the two distributions of values, not just the corresponding medians of the distributions. As a result, it is possible that the two distributions being compared can have the very similar or the same medians, but yet the Wilcoxon-rank sum test may still generate a significant p-value which would indicate that the underlying distributions are significantly different (the values in one distribution are significantly less than, or significantly greater than, values in the other distribution).

Associations between synonymous and non-synonymous substitution rates and paired sites

HIV group 1 sequences and a number flavivirus sequences were analysed for large scale associations between paired codon positions and lower than expected (or higher than expected) synonymous substitution rates (Table 3.2). Low synonymous substitution rates provide evidence of selection acting to preserve the nucleotide sequence (see Figure 3.1 earlier in this chapter for an explanation). For each of the individual datasets that were tested the patterns of association between base-pairing and synonymous substitution rates were largely consistent for each of the five tests performed. HIV group 1 and Hepatitis C Virus sequences showed strong associations between base-paired nucleotide positions and significantly lower than expected synonymous substitution rates, which is exactly what one might expect for these two viruses given that they both contain a number of substructures which are known to be both evolutionarily conserved and biologically relevant. Lower than expected synonymous substitution rates are considered to be evidence of selection acting to conserve the underlying nucleotide sequence, which in this case is likely a consequence of selection acting to maintain base-pairing interactions in biologically relevant secondary structures. Similar associations were also detected between paired codon positions and lower than expected synonymous substitution rates in WNV, JEV and TBEV. The fact that WNV and JEV yielded similar results is reassuring given that they are closely related viral species and share a number of conserved substructures.

Interestingly, in DENV BVDV and CSFV the reverse association was found in that that predicted paired sites within codons were associated with higher than expected synonymous substitution rates. The fact that we do not detect an association between paired sites and lower than expected synonymous rates in DENV is perhaps not entirely unexpected given that, with the exception of a single structural element within the Core gene (substructure 16-308, see Figure 3.7), there is little evidence of biologically relevant conserved substructures in DENV. However, even when this single conserved substructure was analysed in isolation across all DENV serotypes (DENV 1-4), there was no detectable association between paired codon positions and synonymous substitution rates. It is noteworthy that, when this structure was analysed in the DENV serotypes individually, some significant associations between paired codon positions and lower than expected synonymous substitution rates were found in DENV1, DENV3 and DENV4 (p-values < 0.05) – a result consistent with this structural element potentially having a biological function.

Nevertheless, the strong association between paired sites and higher than expected synonymous substitution rates throughout the DENV, BVDV and CSFV coding regions still

requires explanation. This association is puzzling because in the absence of selection to conserve biologically relevant secondary structural elements, variations in synonymous substitution rates are expected to be primarily influenced by basal mutational processes such that paired sites might still be expected to experience lower than expected synonymous substitution rates simply because these sites would be less susceptible to mutation. In support of this, analyses of nucleotide substitution frequencies indicated that DENV, BVDV and CSFV genomic sites that are predicted to be base-paired have significantly lower nucleotide substitution frequencies than sites which are predicted to be unpaired (p-values = 7.41×10^{-18} , 9.28×10^{-20} and 5.30×10^{-44} respectively; Table 3.1). This apparently contradictory result likely stems from the fact that when one considers non-synonymous substitution rates rather than synonymous substitution rates, across all the DENV, BVDV and CSFV datasets these are also significantly lower at paired sites than at unpaired sites (p-values = 8.51×10^{-14} , 1.72×10^{-4} , 8.48×10^{-16} respectively; Table 3.2).

In fact, a similar pattern of non-synonymous rates being lower at paired sites than at unpaired sites was also found in HIV, WNV, JEV and TBEV (p-values = 7.99×10^{-5} , 7.39×10^{-7} , 1.02×10^{-3} , 1.12×10^{-8} respectively; Table 3.2). This suggests that, rather than selection favoring the preservation of functional secondary structures within the coding regions of these viruses it is acting primarily to preserve functional amino acid sequences. Whereas synonymous substitutions enable the preservation of key amino acids at particular sites while still permitting some variation in the underlying codon sequences, selection favoring the reduction of any mutations at all these codon sequences would have much the same effect. It is plausible therefore that in these viruses, selection might favor base-pairing within certain amino acid encoding sequences simply because it reduces mutation frequencies within these sequences. . Also, the fact that synonymous substitutions at paired nucleotide positions are apparently more tolerable in DENV, BVDV and CSFV than in the other flaviviruses considered here implies that unlike in these other viruses, selection acting on DENV, BVDV and CSFV genomes might favor a diversity of secondary structural configurations over the conservation of specific structural elements – possibly as a mechanism of avoiding detection by host immune defenses, such as detection by siRNAs (Sessions *et al.*, 2009).

Table 3.2: Analysis of both synonymous substitution rates and non- synonymous substitution rates at paired and unpaired codon positions in various viral coding sequences.

Dataset	Test	Paired synonymous rate median	Unpaired synonymous rate median	z-score ¹	p-value ¹	Paired non-synonymous rate median	Unpaired non-synonymous rate median	z-score ¹	p-value ¹
Human Immunodeficiency Virus (HIV) group 1	Codon position 1	0.386	0.465	-4.464	8.04×10^{-6}	0.039	0.087	-4.417	1.00×10^{-5}
	Codon position 2	0.393	0.470	-5.429	5.66×10^{-8}	0.039	0.09	-4.057	4.96×10^{-5}
	Codon position3	0.406	0.449	-2.442	1.46×10^{-2}	0.042	0.066	-1.726	8.44×10^{-2}
	All codon positions	0.395	0.456	-7.131	4.91×10^{-12}	0.041	0.075	-5.891	3.86×10^{-9}
	Nucleotides paired ≥ 2	0.395	0.472	-4.725	2.30×10^{-6}	0.04	0.093	-3.945	7.99×10^{-5}
Hepatitis C Virus (HCV) genotype 1	Codon position 1	0.779	0.788	-2.028	4.26×10^{-2}	0.047	0.047	0.424	6.71×10^{-1}
	Codon position 2	0.769	0.796	-5.883	4.05×10^{-9}	0.047	0.047	-0.616	5.38×10^{-1}
	Codon position3	0.766	0.801	-7.600	5.42×10^{-13}	0.047	0.047	-0.201	8.41×10^{-1}
	All codon positions	0.771	0.797	-8.920	1.10×10^{-15}	0.047	0.047	-0.233	8.16×10^{-1}
	Nucleotides paired ≥ 2	0.768	0.800	-6.703	3.66×10^{-11}	0.047	0.047	-0.677	4.99×10^{-1}
Hepatitis C Virus (HCV) genosubtype 1a	Codon position 1	0.715	0.733	-0.927	3.54×10^{-1}	0.002	0.003	0.78	4.36×10^{-1}
	Codon position 2	0.672	0.783	-6.508	9.19×10^{-11}	0.002	0.047	-2.113	3.46×10^{-2}
	Codon position3	0.634	0.804	-11.507	5.75×10^{-21}	0.002	0.003	-0.911	3.62×10^{-1}
	All codon positions	0.676	0.783	-10.875	1.12×10^{-19}	0.002	0.004	-1.305	1.92×10^{-1}
	Nucleotides paired ≥ 2	0.667	0.792	-7.559	6.57×10^{-13}	0.002	0.047	-2.017	4.37×10^{-2}
Hepatitis C Virus (HCV) genosubtype 1b	Codon position 1	0.749	0.763	-2.575	1.00×10^{-2}	0.047	0.047	-0.326	7.45×10^{-1}
	Codon position 2	0.736	0.798	-6.896	1.49×10^{-11}	0.047	0.047	-0.776	4.38×10^{-1}
	Codon position3	0.718	0.807	-10.168	3.10×10^{-18}	0.047	0.047	-1.104	2.69×10^{-1}
	All codon positions	0.737	0.796	-11.274	1.72×10^{-20}	0.047	0.047	-1.267	2.05×10^{-1}
	Nucleotides paired ≥ 2	0.737	0.802	-7.316	2.05×10^{-12}	0.047	0.047	-0.901	3.67×10^{-1}

Dataset	Test	Paired synonymous rate median	Unpaired synonymous rate median	z-score ¹	p-value ¹	Paired non-synonymous rate median	Unpaired non-synonymous rate median	z-score ¹	p-value ¹
Hepatitis C Virus (HCV) genotype 2	Codon position 1	0.721	0.733	-0.951	3.42×10^{-1}	0.046	0.047	0.407	6.84×10^{-1}
	Codon position 2	0.718	0.737	-2.713	6.66×10^{-3}	0.046	0.047	-1.05	2.94×10^{-1}
	Codon position3	0.717	0.746	-4.836	1.32×10^{-6}	0.046	0.047	0.177	8.60×10^{-1}
	All codon positions	0.718	0.740	-4.892	9.96×10^{-7}	0.046	0.047	-0.29	7.72×10^{-1}
	Nucleotides paired ≥ 2	0.720	0.738	-2.554	1.06×10^{-2}	0.046	0.048	-0.822	4.11×10^{-1}
Hepatitis C Virus (HCV) genotype 6	Codon position 1	0.796	0.796	0.841	4.00×10^{-1}	0.007	0.047	1.765	7.75×10^{-2}
	Codon position 2	0.793	0.805	-3.064	2.18×10^{-3}	0.007	0.046	1.535	1.25×10^{-1}
	Codon position3	0.794	0.804	-1.968	4.91×10^{-2}	0.007	0.047	0.232	8.16×10^{-1}
	All codon positions	0.795	0.801	-2.427	1.52×10^{-2}	0.007	0.047		
	Nucleotides paired ≥ 2	0.795	0.801	-1.501	1.33×10^{-1}	0.007	0.047	1.184	2.36×10^{-1}
Dengue (serotypes 1-4)	Codon position 1	0.822	0.811	2.672	7.54×10^{-3}	0.054	0.056	-4.635	3.57×10^{-6}
	Codon position 2	0.822	0.809	3.789	1.51×10^{-4}	0.054	0.057	-7.711	3.22×10^{-13}
	Codon position3	0.820	0.814	2.025	4.28×10^{-2}	0.054	0.057	-6.963	1.08×10^{-11}
	All codon positions	0.821	0.811	4.889	1.01×10^{-6}	0.054	0.057	-11.118	3.57×10^{-20}
	Nucleotides paired ≥ 2	0.822	0.807	3.644	2.68×10^{-4}	0.054	0.057	-7.994	8.51×10^{-14}
Dengue serotype 1	Codon position 1	0.790	0.777	0.691	4.90×10^{-1}	0.024	0.025	-6.237	3.29×10^{-10}
	Codon position 2	0.795	0.764	2.088	3.68×10^{-2}	0.024	0.026	-9.76	2.11×10^{-17}
	Codon position3	0.799	0.756	3.248	1.16×10^{-3}	0.024	0.025	-5.968	2.44×10^{-9}
	All codon positions	0.794	0.761	3.461	5.38×10^{-4}	0.024	0.025	-12.66	2.54×10^{-23}
	Nucleotides paired ≥ 2	0.796	0.757	2.917	3.53×10^{-3}	0.024	0.026	-8.889	1.27×10^{-15}

Dataset	Test	Paired synonymous rate median	Unpaired synonymous rate median	z-score ¹	p-value ¹	Paired non-synonymous rate median	Unpaired non-synonymous rate median	z-score ¹	p-value ¹
Dengue serotype 2	Codon position 1	0.863	0.835	2.385	1.71×10^{-2}	0.017	0.018	-5.871	4.33×10^{-9}
	Codon position 2	0.863	0.826	2.613	$8. \times 10^{-3}$	0.017	0.018	-11.293	1.57×10^{-20}
	Codon position3	0.863	0.837	0.658	5.11×10^{-1}	0.017	0.018	-6.923	1.31×10^{-11}
	All codon positions	0.863	0.834	3.242	1.19×10^{-3}	0.017	0.018	-13.878	8.31×10^{-26}
	Nucleotides paired ≥ 2	0.863	0.826	2.444	1.45×10^{-2}	0.017	0.018	-10.523	5.85×10^{-19}
Dengue serotype 3	Codon position 1	0.814	0.801	-0.167	8.67×10^{-1}	0.025	0.027	-6.417	1.41×10^{-10}
	Codon position 2	0.822	0.769	1.5	5.93×10^{-1}	0.023	0.027	-13.774	1.36×10^{-25}
	Codon position3	0.817	0.786	0.494	6.21×10^{-1}	0.025	0.027	-6.922	1.31×10^{-11}
	All codon positions	0.818	0.785	2.118	3.42×10^{-1}	0.023	0.027	-15.603	2.51×10^{-29}
	Nucleotides paired ≥ 2	0.820	0.773	1.951	5.10×10^{-2}	0.023	0.027	-12.879	9.09×10^{-24}
Dengue serotype 4	Codon position 1	0.809	0.804	0.355	7.23×10^{-1}	0.025	0.026	-4.48	7.48×10^{-6}
	Codon position 2	0.81	0.804	1.386	1.66×10^{-1}	0.025	0.027	-10.493	6.74×10^{-19}
	Codon position3	0.811	0.803	-0.443	6.58×10^{-1}	0.025	0.026	-4.439	9.05×10^{-6}
	All codon positions	0.81	0.804	0.711	4.77×10^{-1}	0.025	0.027	-11.168	2.83×10^{-20}
	Nucleotides paired ≥ 2	0.809	0.804	0.509	6.10×10^{-1}	0.025	0.027	-8.827	1.69×10^{-15}
West Nile virus	Codon position 1	0.858	0.86	-1.354	1.76×10^{-1}	0.015	0.015	-1.801	7.17×10^{-2}
	Codon position 2	0.858	0.861	-0.927	3.54×10^{-1}	0.015	0.015	-4.66	3.17×10^{-6}
	Codon position3	0.857	0.862	-3.357	7.89×10^{-4}	0.015	0.016	-5.232	1.67×10^{-7}
	All codon positions	0.858	0.861	-3.268	1.08×10^{-3}	0.015	0.015	-6.713	3.51×10^{-11}
	Nucleotides paired ≥ 2	0.858	0.86	-1.228	2.19×10^{-1}	0.015	0.015	-4.951	7.39×10^{-7}

Dataset	Test	Paired synonymous rate median	Unpaired synonymous rate median	z-score ¹	p-value ¹	Paired non-synonymous rate median	Unpaired non-synonymous rate median	z-score ¹	p-value ¹
Japanese encephalitis virus (JEV)	Codon position 1	0.808	0.813	-1.041	2.98×10^{-1}	0.025	0.025	-1.736	8.26×10^{-2}
	Codon position 2	0.807	0.816	-2.334	1.96×10^{-2}	0.025	0.025	-3.617	2.98×10^{-4}
	Codon position3	0.807	0.816	-2.953	3.14×10^{-3}	0.024	0.025	-4.248	2.16×10^{-5}
	All codon positions	0.807	0.815	-3.645	2.67×10^{-4}	0.025	0.025	-5.51	3.60×10^{-8}
	Nucleotides paired ≥ 2	0.808	0.815	-1.858	6.31×10^{-2}	0.025	0.025	-3.286	1.02×10^{-3}
Tick-borne encephalitis virus (TBEV)	Codon position 1	0.836	0.861	-2.654	7.96×10^{-3}	0.023	0.022	-0.431	6.66×10^{-1}
	Codon position 2	0.839	0.853	-1.537	5.27×10^{-2}	0.022	0.023	-7.23	3.08×10^{-12}
	Codon position3	0.843	0.843	0.617	5.37×10^{-1}	0.022	0.023	-4.706	2.53×10^{-6}
	All codon positions	0.839	0.853	-2.351	1.87×10^{-2}	0.022	0.023	-7.105	5.55×10^{-12}
	Nucleotides paired ≥ 2	0.838	0.861	-2.810	4.96×10^{-3}	0.022	0.023	-5.712	1.12×10^{-8}
Bovine Viral Diarrhea virus	Codon position 1	0.844	0.841	0.310	7.57×10^{-1}	0.065	0.073	-2.882	3.95×10^{-3}
	Codon position 2	0.846	0.835	3.312	9.26×10^{-4}	0.066	0.074	-4.707	2.51×10^{-6}
	Codon position3	0.845	0.839	2.246	2.47×10^{-2}	0.068	0.072	-2.025	4.28×10^{-2}
	All codon positions	0.845	0.838	3.358	7.85×10^{-4}	0.066	0.073	-5.576	2.47×10^{-8}
	Nucleotides paired ≥ 2	0.845	0.837	2.022	4.32×10^{-2}	0.066	0.073	-3.757	1.72×10^{-4}
Classical Swine Fever virus	Codon position 1	0.871	0.867	0.627	5.31×10^{-1}	0.044	0.046	-4.281	1.86×10^{-5}
	Codon position 2	0.874	0.856	2.843	4.40×10^{-3}	0.043	0.047	-11.378	1.05×10^{-20}
	Codon position3	0.871	0.869	0.536	5.92×10^{-1}	0.044	0.046	-4.689	2.74×10^{-6}
	All codon positions	0.872	0.864	2.298	2.16×10^{-2}	0.044	0.046	-11.747	1.86×10^{-21}
	Nucleotides paired ≥ 2	0.873	0.859	1.962	4.98×10^{-2}	0.043	0.047	-8.975	8.48×10^{-16}

¹Tests where the rates at paired sites are significantly lower than those at unpaired sites (negative z-score and p-value < 0.05) are highlighted in blue and tests where the rates at paired sites are significantly higher than those at unpaired sites (positive z-score and p-value < 0.05) are highlighted in green.

Synonymous substitution rate paired sites correlation analysis

A number of different viruses were tested for correlations between the synonymous substitution rates at base-paired sites (Table 3.3). HIV, HCV, DENV, WNV and JEV yielded strong evidence of such correlations using our simple regression test. When we applied our permutation test, however, only Hepatitis C Virus (HCV) yielded any evidence that these correlations could not simply be accounted for simply by base-paired sites tending to have similar substitution rates due simply to their close proximity in the presence of regional variations in synonymous substitution rates.

The strong associations detected for the HCV datasets analysed suggests that a measurable degree of the detectable synonymous substitution rate correlation between codons pairs that contain constituent nucleotides which are base-paired with one another is likely attributable to those codons evolving under similar selective constraints favouring the maintenance of the structural elements that they are a part of. While strongly suggesting that a substantial proportion of base-pairings that are predicted to form in HCV genomes really do exist, it also implies that many of the structural elements within which these base-pairings occur are probably biologically functional. Again, however, an alternative explanation that excludes the role of selection is that similar basal mutational processes might simply be at play between physically paired codons.

The fact that there is less evidence in the other virus groups analysed here of codon-pairs containing base-paired codons having correlated synonymous substitution rates does not imply secondary structures in these other viruses do not make a substantial contribution to their fitness. It also does not imply that basal mutation processes do not tend to be similar for the pair of nucleotides that are base-paired. It simply implies that, if such associations exist, they are simply undetectable in these other virus groups.

Table 3.3: Table of different viruses and their synonymous substitution rate correlations.

Dataset (Synonymous substitution rates)	Spearman's rank correlation ¹	Spearman's rank correlation p-value ¹	Permutation test p- value ²
HIV group 1	0.194	9.61×10^{-35}	0.33500
HCV 1	0.209	1.28×10^{-59}	0.00565
HCV 1a	0.210	9.82×10^{-60}	0.00045
HCV 1b	0.213	5.71×10^{-62}	0.00060
HCV 2	0.171	2.10×10^{-39}	0.02670
HCV 6	0.166	1.38×10^{-34}	0.20800
Dengue (serotypes 1- 4)	0.033	7.98×10^{-3}	0.28400
Dengue serotype 1	0.052	8.92×10^{-5}	0.06400
Dengue serotype 2	0.016	0.194	0.49800
Dengue serotype 3	0.012	0.113	0.46900
Dengue serotype 4	0.000	0.334	0.62000
West Nile virus	0.035	4.27×10^{-3}	0.29300
Japanese encephalitis virus (JEV)	0.046	1.62×10^{-4}	0.22300
Tick-borne encephalitis virus (TBEV)	0.022	0.070	0.35100
Bovine Viral Diarrhea virus	0.008	0.231	0.65900
Classical Swine Fever virus	0.008	0.173	0.41400

¹The simple test which uses Spearman's correlation to calculate the degree of correlation between pairs of synonymous substitution rates corresponding to base-paired sites. **Tests with p-values below the significance threshold of 0.05.**

²The "Permutations test p-value" takes into account the distance between base-paired sites. **Tests with p-values below the significance threshold of 0.05.** A significant permutation test p-value is a good indication that the rates at inferred base-paired nucleotides in a sequence are correlated due to the base-pairing interactions and not just the distance between nucleotides along the sequence.

Associations between coevolving sites and base-paired sites

HIV, HCV, JEV and DENV were analysed for associations between coevolving nucleotide pairs and base-paired positions. A position i,j was considered base-paired if it was paired in at least 25% of sequences within a particular virus dataset, otherwise that position was considered unpaired. Our analysis was restricted to just four viruses due to the long running time of the coevolution detection method. Each of the four viruses showed highly significant genome-wide associations between sites that were predicted to be base-paired and sites which were detectably coevolving (Table 3.4). Specifically, for all four groups of viruses nucleotide polymorphisms that were anticipated to maintain base-pairing complementarity occurred significantly more frequently at sites that were predicted to be base-paired than could be accounted for by chance. This suggests that in all of the virus datasets tested, there are likely a number of specific base-pairs that are being selectively maintained, and therefore that many of

the predicted structural elements containing these base-pairs have biological functions. The presence of a highly statistically significant association in DENV is somewhat unexpected given that there appear to be very few conserved biologically relevant substructures in DENV, however, it should be noted that there are also correspondingly fewer nucleotide positions that reach the 25% base-pairing cut-off in DENV relative to HIV, HCV and JEV, suggesting that although there are fewer paired positions identified as conserved, those that are identified as conserved are likely biologically relevant given that there still remains a strong association with coevolution.

Table 3.4: Contingency tables and corresponding chi-squared association tests for genome-wide paired and coevolving nucleotide pairs in four viruses.

Human Immunodeficiency Virus (HIV) group 1				Hepatitis C Virus (HCV)			
	Unpaired	Paired	Total		Unpaired	Paired	Total
Not coevolving	277 213 (277 146.9)	251 (317.1)	277 464	Not coevolving	323 275 (323 156.2)	509 (627.8)	323 784
Coevolving	115 185 (115 251.1)	198 (131.9)	115 383	Coevolving	143 115 (143 233.8)	397 (278.2)	143 512
Total	392 398	449	392 847	Total	466 390	906	467 296
$\chi^2 = 47.0$ p-value = 7.11×10^{-12}				$\chi^2 = 73.3$ p-value < 1.00×10^{-16}			

Japanese Encephalitis Virus (JEV)				Dengue virus (DENV)			
	Unpaired	Paired	Total		Unpaired	Paired	Total
Not coevolving	416 753 (416 650.6)	1 472 (1 574.4)	418 225	Not coevolving	394 903 (394 846.4)	253 (309.6)	395 156
Coevolving	68 877 (68 979.4)	363 (260.6)	69 240	Coevolving	96 099 (96 155.6)	132 (75.4)	96 231
Total	485 630	1 835	487 465	Total	491 002	385	491 387
$\chi^2 = 47.0$ p-value = 7.01×10^{-12}				$\chi^2 = 52.9$ p-value = 3.54×10^{-13}			

Unbracketed numbers represent observed counts and bracketed numbers represent the corresponding expected counts.

Detailed analysis of conserved substructures in various RNA viruses

The prediction of nucleic acid secondary structures is imperfect, whether it be prediction based on experimental information or purely computational prediction based on nucleotide sequence information alone. The UNAFold program (Markham and Zuker, 2008), which was used throughout this section to predict secondary structures, is an example of a thermodynamic approach which finds the secondary structure that minimises the free energy of given a sequence. The secondary structure with the minimum free energy is considered to be the most probable structure at equilibrium; however, there typically exist multiple highly likely structural conformations a nucleic acid molecule can form. The assumption that the minimum free energy structure reflects the true secondary structure is therefore incorrect given that a single nucleotide sequence can fold into more than one conformation. Having said that, it has been recognized that the secondary structures within the set of most probable conformations typically closely resemble one another in that they share a significant proportion of base-pairings (Mathews, 2006). Another limitation of the thermodynamic approaches is that they cannot predict secondary structures containing non-nested base-pairings, i.e. pseudoknots (Rivas and Eddy, 1999; Staple and Butcher, 2005).

The metastable nature of nucleic acid secondary structures together with the inability to predict pseudoknots and other simplifying assumptions that the underlying thermodynamic models employ, imply that computational secondary structure predictions will only be approximations of the underlying biological reality. These inaccuracies are important to keep in mind when considering the analysis presented here, which aims to identify conserved substructures in the genomes of a number of virus species. It is likely that the predicted structural conformations of the identified substructures will be inaccurate. Having said that, even if the predicted structural conformations of the identified substructures is imprecise, additional supporting evidence provided here may give some assurance that at least the positions of the identified biologically relevant substructures have been accurately identified.

Conserved HIV secondary structural elements

Human Immunodeficiency Virus (HIV) type 1 group M subtypes B, C, D and G, HIV type 1 group O, and HIV group 2 full genome sequences were analysed for evidence of biologically relevant conserved secondary structural elements (see Figure 3.5). The largest and most strongly conserved of these predicted structural elements (conservation z-score = -2.930, p-value = 0.003, substructure 7289-7609, Supplementary Figure 4) was the well-characterized Rev Response Element. The RRE plays a crucial role in the regulating the export of HIV virions (Heaphy *et al.*, 1990). Unsurprisingly, it was found to be conserved across all of the HIV genotypes tested. The fact that our conservation test ranks this substructure as having the

largest degree of conservation of all detected substructures within HIV-1 is an encouraging validation of our methodology. Furthermore, synonymous substitution rates in the region of this substructure are significantly lower than expected (z-score = -9.416, p-value = 4.69×10^{-21}) which provides good evidence of selective maintenance of the underlying nucleotide sequence and hence base-pairings within this substructure. A strong association between computationally predicted base-pairings and nucleotide coevolution (z-score = -3.681, p-value = 0.0002) was also detected, providing evidence of the selective maintenance of base-pairings within this structural element.

The second most widely conserved secondary structural element detected in HIV genome was the gag-pol frameshift element (substructure 1576-1695, Supplementary Figure 4). This substructure regulates the ratio of HIV gag/gag-pol synthesis (Parkin, Mario Chamorro, *et al.*, 1992). It was detected as conserved across all HIV-1 genotypes tested (conservation z-score -2.011, p-value = 0.044), but not between the HIV-1 genotypes and HIV-2. There is, however, evidence of a different substructure conserved within HIV-2 at a nearby position and it is possible that this substructure is also responsible for gag-pol frameshifting in HIV-2 (Hung *et al.*, 1998). Unfortunately, synonymous substitution rates could not be reliably estimated for codon sites that correspond with this structural element because it is present within two overlapping expressed reading frames, however, a strong association between computationally predicted base-pairings within this structural element and nucleotide coevolution (z-score = -7.017, p-value = 2.27×10^{-12}) was detected, providing evidence of selective maintenance of base-pairings within this element.

The third most strongly conserved secondary structural element was detected in the *pol* gene (substructure 2393-2545, conservation z-score = -2.321, p-value = 0.020, Supplementary Figure 3). This substructure was only conserved amongst the three HIV-1M subtypes, C, D, and G which are all found circulating in African populations. Visualisation of this presently uncharacterised secondary structural element reveals a 30 nucleotide-long stem with a number of intermediary bulges (substructure 2393-2545, Supplementary Figure 3). The biological functionality of this structural element are further supported by the fact that codons overlapping this structural element have lower than expected synonymous substitution rates (z-score = -2.321, p-value = 0.020) and nucleotides that are predicted to be base-paired display evidence of complementary coevolution (z-score = -8.383, p-value = 1.36×10^{-14}). A corresponding structural element was not identified in the SHAPE-determined NL4-3 secondary structure; this is to be expected as NL4-3 is a HIV-1 subtype B isolate. We strongly recommend this structural element for further investigation.

Another four secondary structural elements were detectably conserved between two or more of the groups of analysed HIV datasets, these include two that were not detected in the NL4 structure model of Watts *et al* , 2009 (substructure 859-946 and substructure 1123-1201).

Conserved HCV secondary structural elements

HCV genotype 1, 2, 3, 4 and 6 genomes were analysed for evidence of conserved secondary structural elements (Figure 3.6). HCV genotype 5 was excluded from this analysis due a lack of available full genome sequences. As with HIV, HCV has been extensively studied with regards to secondary structure and therefore this section will mainly serve as a validation of our approach, rather than being intended to reveal the likely biological relevance of previously uncharacterised secondary structural elements.

The largest conserved secondary structural element detected across all the HCV genotypes analysed was detected in the Core gene (conservation z-score = -2.385, p-value = 0.017, substructure 334-924, Supplementary Figure 5). A number of stem-loop structures within this structural element have likely biological functions, including a stem-loop structure 443 nucleotides downstream of the Core gene start codon, denoted SL443 by Tuplin *et al.*, 2002 that is thought to be involved in HCV translational regulation (Ito and Lai, 1999) and which may also be involved long-range interactions with the 5' end of the genome (Honda *et al.*, 1996). Also a number of other stem-loops within this substructure have been found to be important for viral proliferation both in *in vivo* and in cell-culture and (Vassilaki *et al.*, 2008). Although the Core gene codons that overlap this region have significantly lower than expected synonymous substitution rates (z-score = -22.206, p-value = 3.00×10^{-109}), this is likely primarily attributable to positions ~126 to ~432 falling within a region of overlap between two genes that are expressed in different reading frames (Branch *et al.*, 2005). These overlapping open reading frames would be expected to strongly constrain synonymous substitutions within this region since synonymous substitutions in one frame will likely be non-synonymous in the other. This means that the low synonymous substitution rates detected in these two regions are not necessarily due to selection acting to preserve functionally relevant base-pairing interactions. What is notable, however, is that even when this region is excluded from the analysis, synonymous substitution rates are much lower than expected within the remaining codons that overlap this structure (Figure 3.6). Finally, a strong association between computationally predicted base-pairings and nucleotide coevolution was detected (z-score = -9.011, p-value = 2.04×10^{-19}), providing evidence of selective maintenance of biologically important base-pairings within this structural element.

The structural element detected in HCV with the most significant conservation score (z-score = -3.067, p-value = 0.002) was found in the NS5B gene bordering on the 3' non-coding region (substructure 8997-9407, Supplementary Figure 7). The potential biological functionality of this substructure is supported by codons within this substructure having significantly lower than expected synonymous substitution rates (z-score = -13.671, p-value = 1.51×10^{-42}). As before, there is evidence that a stem-loop structure in this region, denoted 5BSL3.2 by You et al. (2004), binds the RNA-dependent RNA polymerase encoded by NS5B. This function is particularly plausible given that the structural element that contains this stem-loop is well-placed to bind the polymerase following its translation (You *et al.*, 2004). Within this structural element, 5BSL3.2 forms part of a larger cruciform substructure, denoted 5BSL3 by You et al (2004), with an as yet undetermined function. It has nevertheless been proposed that these substructures are simply part of a genome-scale ordered RNA structure (GORS) that may be associated with viral persistence via modulation of dsRNA recognition pathways (Simmonds *et al.*, 2004; Tuplin *et al.*, 2004).

Amongst the other most highly conserved structural elements detected, five were within the NS5B gene and have been previously characterised but still have no known function (see You *et al.*, 2004; Tuplin *et al.*, 2004).

Conserved DENV secondary structural elements

DENV serotypes 1, 2, 3 and 4 all display similar degrees of diversity and have large numbers of available full genome sequences: factors making this virus species ideal for studying evidence of secondary structure conservation. Presented in Figure 3.7 is a summary of conserved substructures detected amongst Dengue serotypes 1 - 4.

Of the five structural elements that were conserved between two or more of these serotypes, the largest is the only one that is conserved across all four serotypes (with a conservation z-score = -2.029 and an associated p-value = 0.035), this structural element overlaps parts of both the 5' non-coding region and the 3' part of the Core gene (substructure 16-308, Supplementary Figure 8). Both the fact that codons overlapping this structural element within the Core gene have significantly lower than expected synonymous substitution rates (z-score = -14.903, p-value = 3.15×10^{-50}) and a statistically significant association between computationally predicted base-paired positions and complementary nucleotide coevolution (z-score = -8.511, p-value = 7.48×10^{-15}) suggests that base-pairings within this substructure are under selective maintenance. It is known that the 5' half of this structural element (corresponding to position ~25-140 in NC_001477) is likely to be biologically functional in DENV in that it is known to direct translation start codon selection in DENV serotype 2 (Clyde and Harris, 2006). Although

the stem-loop structures that are predicted to form in the 3' half of this structural element have no known functions, the presence of highly conserved base-pairings within these stem-loop structures strongly suggest that they are also biologically functional.

Secondary structural elements conserved across WNV, JEV, TBEV genomes

WNV, JEV and TBEV are three closely related Flavivirus species that were analysed to determine whether they possess shared secondary structural elements (Figure 3.8). This revealed three structural elements that are conserved between WNV and JEV, and two that are conserved between JEV and TBEV. As in DENV (which is closely related to these viruses), the largest and most conserved of the inferred structural elements found straddling the interface between the 5' non-coding region and the Core gene (substructure 74-308; conservation z-score = -2.357, p-value = 0.018; Supplementary Figure 10). This substructure has a number of stem-loop elements which while being strongly conserved in WNV and JEV, are relatively less conserved in TBEV. Synonymous substitution rates in Core gene codons overlapping the 3' half of this structural element are significantly lower than expected which additionally suggests that maintenance of this structure has been selectively favoured. Although the function of this structural element has not been well-established in either WNV or JEV, the substructure is likely homologous to a substructure found in the same position in DENV and is also thought to influence translation start site recognition (Lindenbach and Rice, 2003).

The structural element with the most significant conservation score (z-score = -2.432, p-value = 0.015) was found conserved within the envelope genes of WNV and JEV (substructure 2048-2183, Supplementary Figure 11). This structural element has not previously been characterised and contains a long stem with a number of loop regions. Despite moderate degrees of divergence between WNV and JEV with respect to the nucleotide sequence of the stem, a high proportion of computationally inferred Watson-Crick base-pairings have apparently been maintained since the divergence of these species. Although codons overlapping this substructure did not have significantly lower than expected synonymous substitution rates, the presence of highly conserved base-pairings in this substructure (as indicated by red lines joining base-paired nucleotide positions in substructure 2048-2183 of Supplementary Figure 11) strongly suggest that this substructure is biologically functional and warrants further study.

Finally, a structural element in the 3' non-coding region (substructure 10881-10961, Supplementary Figure 11) that is conserved between JEV and TBEV (Figure 3.8, conservation z-score = -2.102, p-value = 0.036) has been previously identified in JEV and is likely involved in minus-strand RNA synthesis in that it has been found to form a complex with replicase proteins expressed from the NS3 and NS5 genes (Chen *et al.*, 1997). The fact that this substructure has

the same structural conformation and position despite these two flavivirus species differing in nucleotide sequence at ~30% of genomic sites, suggests that the structure may also have a biological function in TBEV.

University of Cape Town

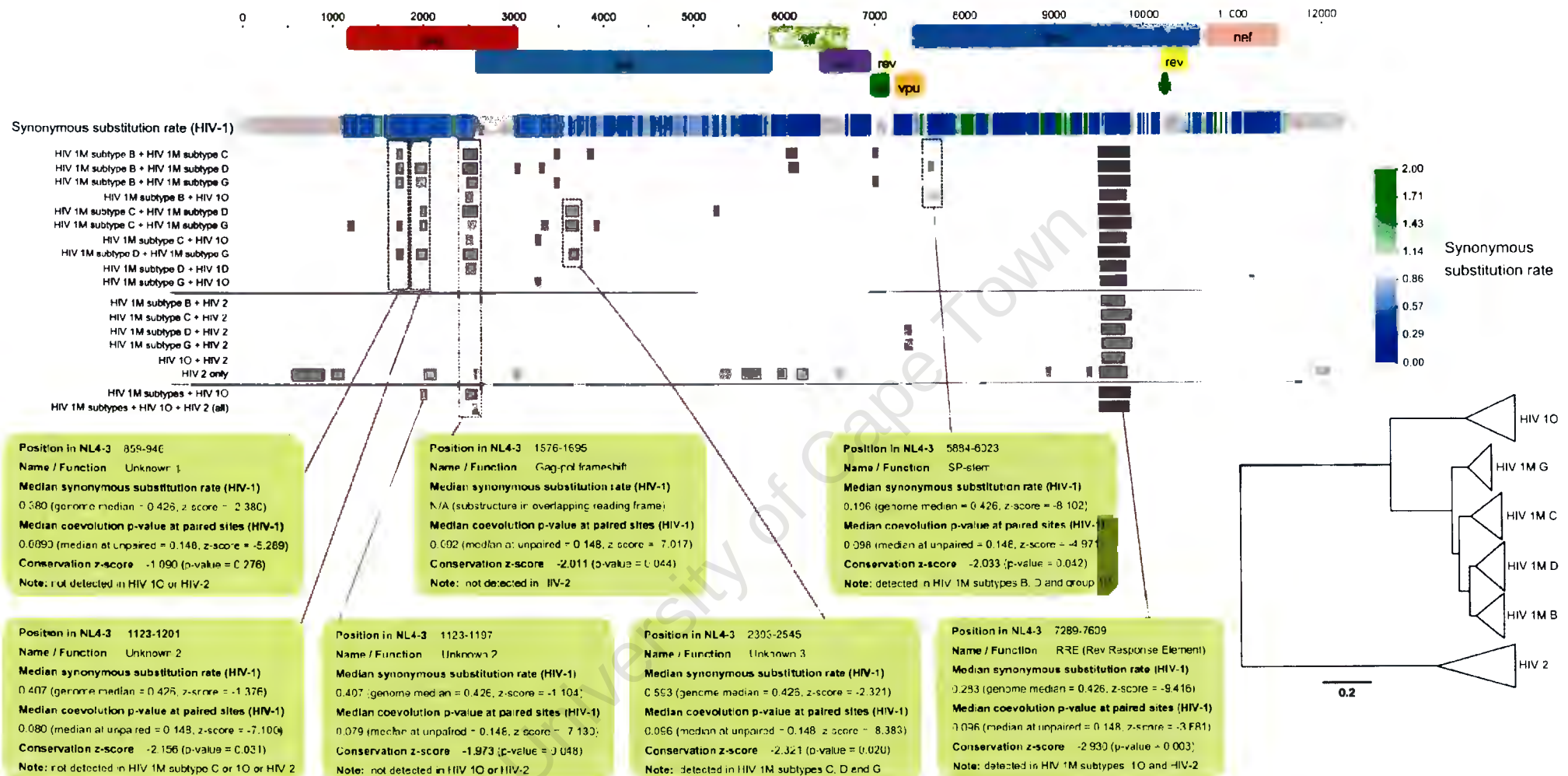


Figure 3.5: Summary of conserved substructures detected across HIV subtypes. This figure depicts substructures detected as conserved in several HIV group 1M subtypes (B, C, D and G), HIV group 10 and HIV-2. There is strong evidence that the REV response element is conserved in both HIV-1 and HIV-2 sequences. The gag-pol frameshift substructure that regulates the ratio of gag/gag-pol synthesis is detectably conserved across all HIV 1M and 10 subtypes tested, but not in HIV-2. A substructure in the Gag gene located at position (859-946) is detectably conserved in all HIV group 1M subtypes but not group 10. The SP-stem is detected in HIV 1M subtype B as expected and in HIV 1M subtypes C and D. Large negative z-scores (less than -1.96) for "Median coevolution p-value at paired sites" indicates a significant correspondence between predicted base-paired positions and complementary nucleotide coevolution. A large negative "Conservation z-score" (less than -1.96) provides statistically significant evidence that a substructure is structurally conserved.

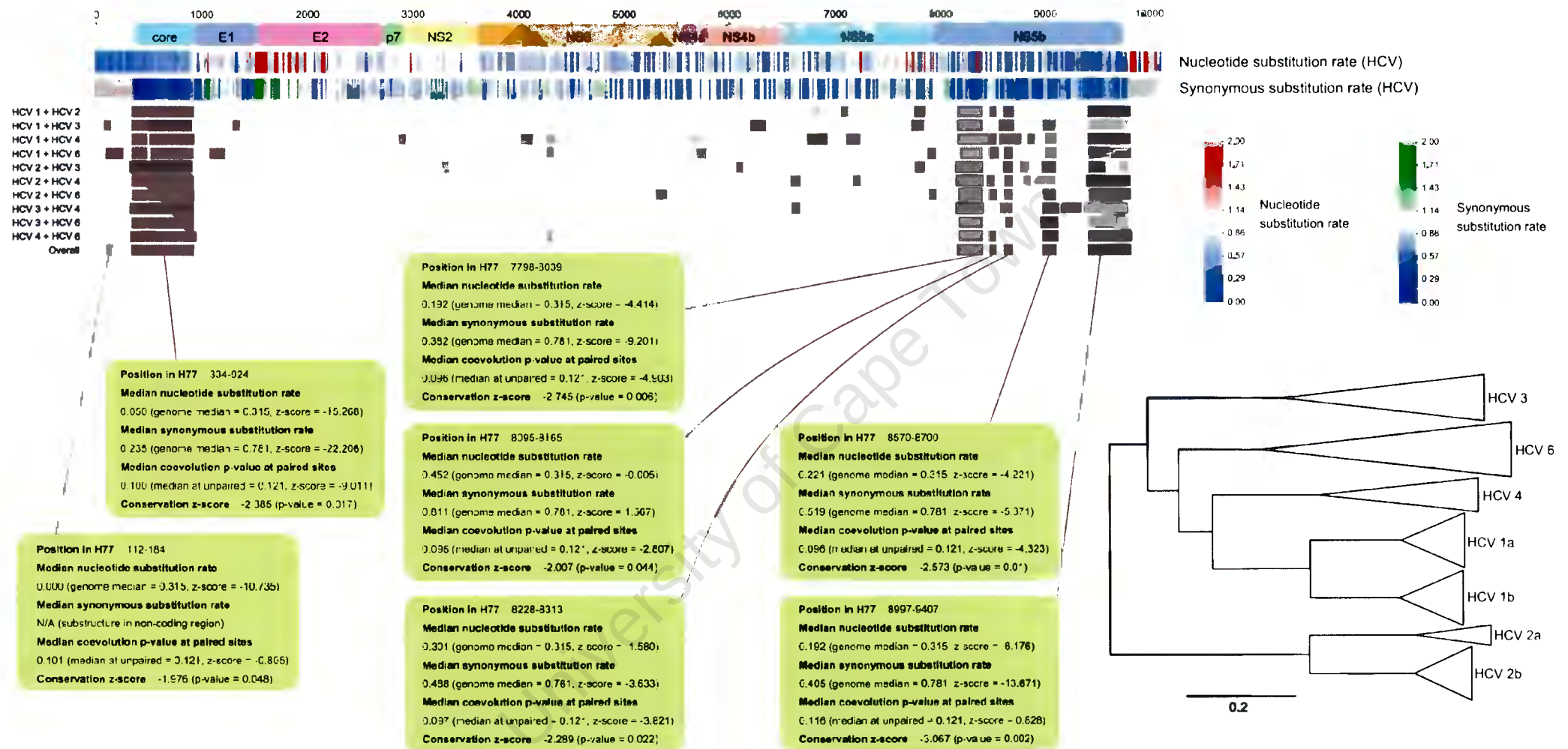


Figure 3.6: Summary of conserved substructures detected amongst several HCV genotypes. This figure depicts substructures detected as conserved in five HCV genotypes (1, 2, 3, 4 and 6). Where there is a gray block in one of the first 5 rows it indicates that a substructure is detectably conserved between 50% of two HCV genotype sequence pairs. The final row (overall) indicates substructures that are conserved between at least 50% of all the HCV genotype sequence pairs in our dataset. Using this information it is possible to differentiate where different substructures are conserved within different HCV genotypes. For example, the substructure at position 8570-8700 in NS5b is detectably conserved in all tested pairs of genotypes, except between genotypes 1 and 2. Large negative z-scores (less than -1.96) for "Median coevolution p-value at paired sites" indicates a significant correspondence between predicted base-paired positions and complementary nucleotide co-evolution. A large negative "Conservation z-score" (less than -1.96) provides statistically significant evidence that a substructure is structurally conserved.

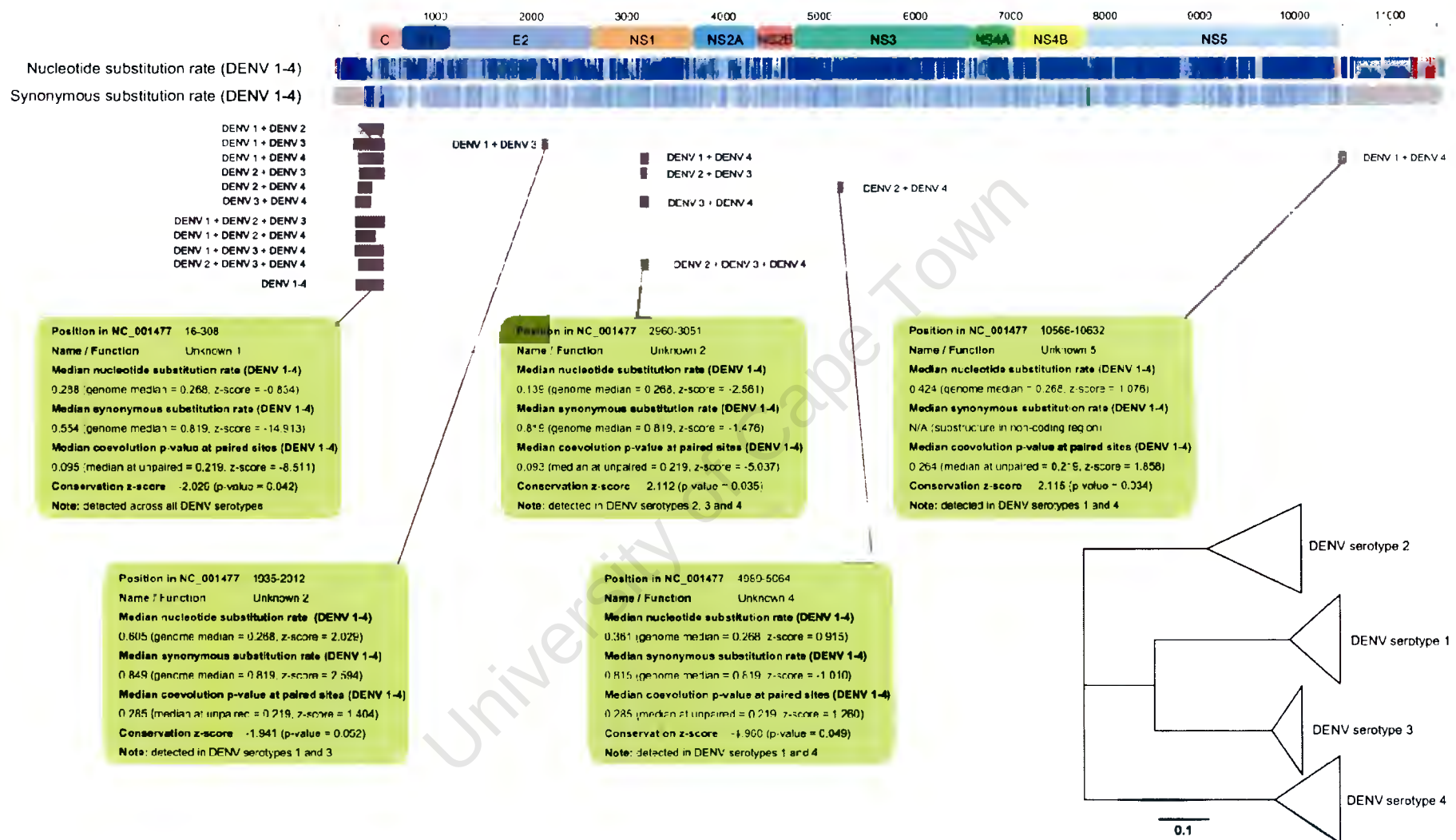


Figure 3.7: Summary of conserved substructures detected amongst Dengue serotypes 1 - 4. This figure depicts substructures detected as conserved in four Dengue virus serotypes. Overall there is only a single substructure which detected as conserved across all four serotypes (substructure 16-308); this substructure is supported by significantly lower than expected synonymous substitution rates. Large negative z-scores (less than -1.96) for "Median coevolution p-value at paired sites" indicate a significant correspondence between predicted base-paired positions and complementary nucleotide co-evolution. A large negative "Conservation z-score" (less than -1.96) provides statistically significant evidence that a substructure is structurally conserved.

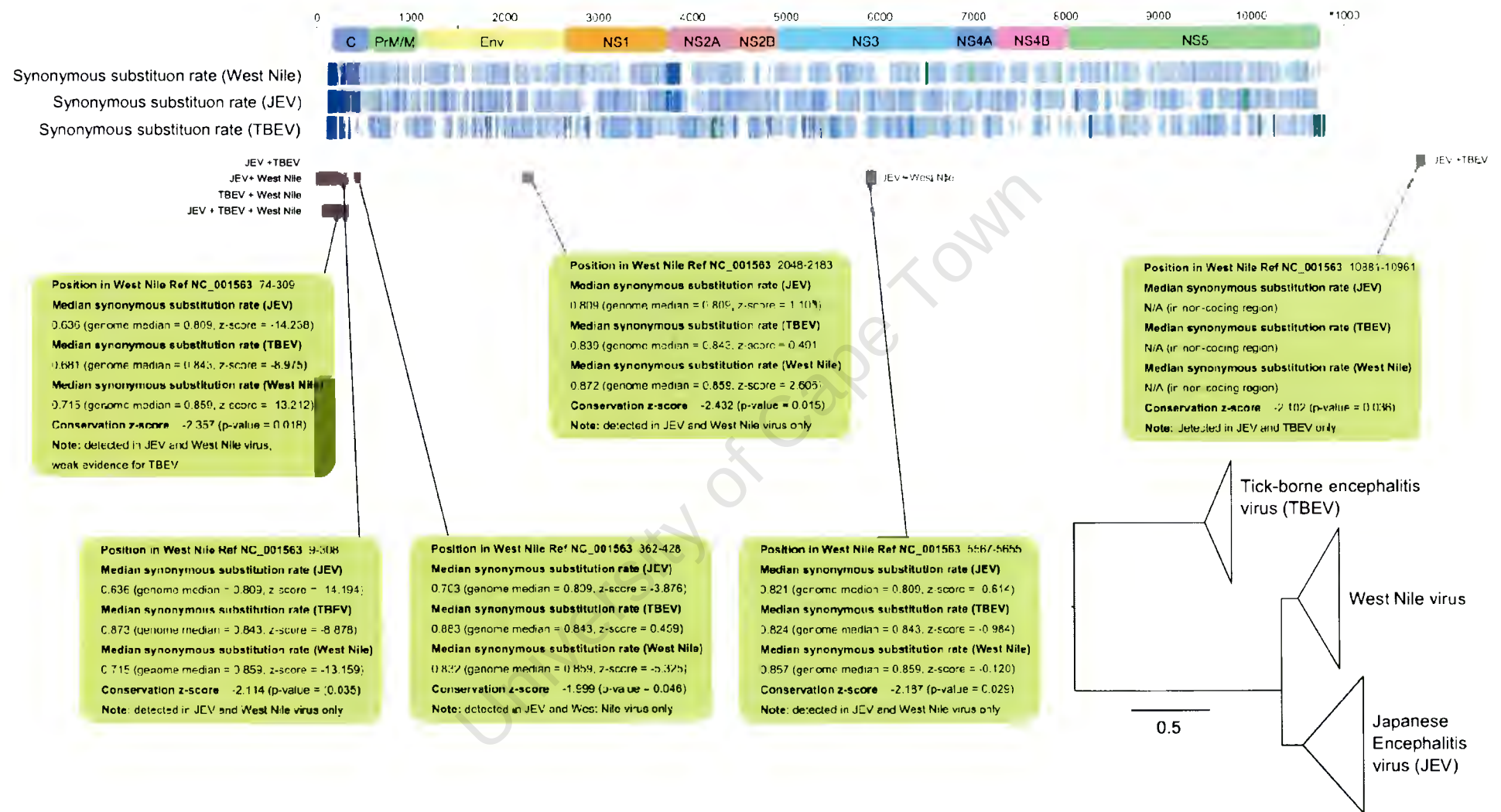


Figure 3.8: Summary of conserved substructures detected between West Nile Virus (WNV), Japanese encephalitis virus (JEV) and tick-borne encephalitis virus (TBEV). Two tick-borne viruses (WNV and JEV) and one mosquito-borne virus (TBEV) were tested for evidence of conserved substructures. A number of substructure were detected as conserved between WNV and JEV, which is expected given that they are phylogenetically quite closely related. Two substructures were detected as conserved between the much more distantly related JEV and TBEV and no substructures were detected as conserved between TBEV and WNV. Large negative conservation z-scores (less than -1.96) provide statistically significant evidence that a substructure is structurally conserved. A large negative "Conservation z-score" (less than -1.96) provides statistically significant evidence that a substructure is structurally conserved.

Conclusions

Both the genome-wide and substructure focused analysis revealed overwhelming evidence of genomic secondary structures impacting the mutational and evolutionary dynamics of both HIV and a number different flaviruses species. There was evidence in almost all virus species tested that paired nucleotide positions were significantly associated with lower than expected nucleotide substitution rates. Similarly, in HIV, HCV, WNV, JEV, and TBEV codons containing base-paired nucleotides had significantly lower synonymous substitution rates than codons without base paired nucleotides. One plausible explanation for all of these results is that base-paired nucleotides are less susceptible to mutation and strand cleavage than are unpaired nucleotides (Lindahl and Nyberg, 1974; Parthasarathi *et al.*, 1995). Another explanation for base-pairing associated reductions in substitution rates is that they are the product of selection acting to preserve base-pairing interactions within biologically relevant secondary structures.

The synonymous substitution rates of base-paired DENV codons (i.e. codon-pairs comprised of nucleotides that base-paired with one another) in HIV, HCV, DENV, WNV and JEV were also all detectably correlated but in all of these datasets other than HCV, this correlation was likely driven by the close spatial proximity on the genome of paired sites coupled with regional variations in nucleotide substitution rates. Therefore, although our results up until this point provided valuable independent confirmation of the likely existence of a substantial fraction of the base-pairing interactions that we have computationally inferred and are consistent with the action of natural selection favouring the maintenance of these interactions, they cannot definitively exclude the possibility that the detected substructures are conserved simply due to fewer mutations occurring at base-paired sites than at unpaired sites. Much stronger evidence of natural selection acting to conserve base-pairing interactions was a highly significant tendency in all of the virus datasets tested for base-paired nucleotides to coevolve in a manner that maintained base-pairing complementarity. This implies that the structural configurations of at least some of the secondary structural elements detected within the analysed viral genome datasets have an impact viral fitness and, therefore, are likely to have defined biological functions.

Detailed analysis of individual secondary structural elements revealed that whereas there are a number of highly conserved structural elements within HIV and HCV genomes, there were very few such structures detectable within the DENV, WNV, JEV and TBEV genomes. The fact that many of these conserved structures in HIV and HCV have been previously characterized and have known biological relevance suggests that many of the similarly conserved but currently uncharacterized substructures that we have identified in these species (but particularly in HIV)

might also have important biological functions. Structural elements such as the one that is conserved within the *pol* gene of HIV-1M subtype C, D and G genomes (substructure 2393-2545 in Supplementary Figure 3) certainly warrant further investigation.

Our detection of far fewer highly conserved structural elements in DENV than in HIV and HCV may indicate that individual functionally relevant genomic secondary structures play a far smaller role in DENV biology than they do in the biology of HCV and HIV. Although it is clear that the genomes of all three species likely form secondary structures and there exists strong evidence that in all three species natural selection likely favors the complementary coevolution of base-paired nucleotides within these substructures, it is also apparent from our analyses that the nature of this selection likely differs between DENV and these other two species. Specifically, whereas in all three species synonymous substitution rates at base-paired sites are generally lower than at unpaired sites, in DENV coding regions this pattern is apparently completely attributable to non-synonymous substitutions being significantly lower at base-paired sites with synonymous substitution rates actually being significantly higher at base-paired sites than at unpaired sites. This suggests that in DENV protein evolution has priority to the point where one of the prime functions of base-pairing in DENV may be to reduce the basal rates of non-synonymous substitutions at base-paired sites.

Collectively our evolutionary analyses suggest that while numerous selectively maintained biologically functional structural elements occur within the genomes of HIV and various flavivirus species, many of the less conserved predicted structural elements within these genomes – and possibly most of those occurring in DENV, BVDV and CSFV genomes – have no uniquely advantageous molecular configuration that is specifically favored by natural selection. In such cases, selection may simply be acting either at the level of individual nucleotide pairs so as to reduce basal mutation rates at particular key codon sites, or on the level of individual codons so as to ensure that base-pairing does not reduce non-synonymous substitution rates. Alternatively, at the whole genome scale selection favoring the avoidance of intra-cellular targeting by siRNAs may simply be acting to reduce the number of unpaired genomic sites. This type of large-scale non-conserved (but nevertheless evolutionarily favored) genomic secondary structure, called genome-scale ordered RNA (or GORS), has previously been hypothesized to facilitate the long-term persistence of ssRNA viruses (including many studied here) that cause chronic infections (Simmonds *et al.*, 2004; Davis *et al.*, 2008). Our results are entirely consistent with the coexistence within individual genomes of a continuum of structural elements ranging from those with specific biological functions that have evolutionarily conserved configurations, to those that simply form as a consequence of selection favoring either reduced substitution rates in specific genome regions or increased protection against anti-viral cellular defenses.

might also have important biological functions. Structural elements such as the one that is conserved within the *pol* gene of HIV-1M subtype C, D and G genomes (substructure 2393-2545 in Supplementary Figure 3) certainly warrant further investigation.

Our detection of far fewer highly conserved structural elements in DENV than in HIV and HCV may indicate that individual functionally relevant genomic secondary structures play a far smaller role in DENV biology than they do in the biology of HCV and HIV. Although it is clear that the genomes of all three species likely form secondary structures and there exists strong evidence that in all three species natural selection likely favors the complementary coevolution of base-paired nucleotides within these substructures, it is also apparent from our analyses that the nature of this selection likely differs between DENV and these other two species. Specifically, whereas in all three species synonymous substitution rates at base-paired sites are generally lower than at unpaired sites, in DENV coding regions this pattern is apparently completely attributable to non-synonymous substitutions being significantly lower at base-paired sites with synonymous substitution rates actually being significantly higher at base-paired sites than at unpaired sites. This suggests that in DENV protein evolution has priority to the point where one of the prime functions of base-pairing in DENV may be to reduce the basal rates of non-synonymous substitutions at base-paired sites.

Collectively our evolutionary analyses suggest that while numerous selectively maintained biologically functional structural elements occur within the genomes of HIV and various flavivirus species, many of the less conserved predicted structural elements within these genomes – and possibly most of those occurring in DENV, BVDV and CSFV genomes – have no uniquely advantageous molecular configuration that is specifically favored by natural selection. In such cases, selection may simply be acting either at the level of individual nucleotide pairs so as to reduce basal mutation rates at particular key codon sites, or on the level of individual codons so as to ensure that base-pairing does not reduce non-synonymous substitution rates. Alternatively, at the whole genome scale selection favoring the avoidance of intra-cellular targeting by siRNAs may simply be acting to reduce the number of unpaired genomic sites. This type of large-scale non-conserved (but nevertheless evolutionarily favored) genomic secondary structure, called genome-scale ordered RNA (or GORS), has previously been hypothesized to facilitate the long-term persistence of ssRNA viruses (including many studied here) that cause chronic infections (Simmonds *et al.*, 2004; Davis *et al.*, 2008). Our results are entirely consistent with the coexistence within individual genomes of a continuum of structural elements ranging from those with specific biological functions that have evolutionarily conserved configurations, to those that simply form as a consequence of selection favoring either reduced substitution rates in specific genome regions or increased protection against anti-viral cellular defenses.

4 Protein and RNA structure is associated with recombination in HIV-1

Introduction

This chapter investigates the impacts of both RNA secondary structure and protein tertiary structure on genetic recombination in Human Immunodeficiency Virus group 1 (HIV-1). Recombination is an evolutionary process involving the exchange of genetic material between nucleotide sequences (Lai, 1992). It is important from an evolutionary perspective because within individual genomes it can both remove harmful mutations, and facilitate the accumulation of beneficial mutations by potentially creating novel combinations of nucleotide polymorphisms which can, in turn, enable far wider exploration of a sequence space than is achievable by mutation alone (Muller, 1964; Felsenstein, 1974).

In retroviruses such as HIV, the predominant form of recombination that occurs is called "copy-choice recombination". This type of recombination involves a process termed "template switching" which takes place at the start of the replication cycle when RNA copies of the HIV genome are being reverse transcribed into DNA (Jetzt *et al.*, 2000; Zhang *et al.*, 2000; Simon-Loriere and Holmes, 2011). Retroviruses, such as HIV, are "diploid viruses" implying that every virus particle contains two genome copies and template switching between these copies during reverse transcription is a natural occurrence during the virus replication cycle (Billeter *et al.*, 1974; Beemon *et al.*, 1974; Onafuwa *et al.*, 2003; Chin *et al.*, 2005; Song *et al.*, 2007). If the two co-packaged HIV genomes are genetically different then template switching during this process will yield a recombinant genome.

It is well-established that secondary structures within template RNA molecules can have a large influence on copy-choice mechanisms of recombination. The first mechanism whereby RNA structures affect copy-choice recombination is their influence on which genomes get co-packaged within the same viral particle. RNA structures in the 5' regions of HIV genomes facilitate the dimerization of RNA genomes via inter-genome base-pairing interactions (Onafuwa *et al.*, 2003; Flynn *et al.*, 2004; Flynn and Telesnitsky, 2006; Zhuang *et al.*, 2006; Chin *et al.*, 2005; Song *et al.*, 2007). By regulating the frequencies at which heterogeneous RNA genomes form dimers and become co-packaged, these secondary structures will influence which pairs of genomes get an opportunity to recombine. The second way in which RNA structures can influence recombination is by directly influencing the template switching process itself. RNA structures can induce pausing of reverse transcriptases: a factor which can increase the

probability of template switching occurring at particular positions within structured RNA molecules (Roda *et al.*, 2002; Moumen *et al.*, 2003; Galetto *et al.*, 2004). This presents a mechanism via which RNA structures can direct the positions along the genome at which recombination breakpoints will most frequently occur (Simon-Loriere *et al.*, 2010).

Although the capacity to recombine can provide viral species with a number of evolutionary benefits, many of the individual recombination events that occur between any particular pair of viruses will be deleterious; especially if they occur between distantly related genomes (Martin *et al.*, 2006; van der Walt *et al.*, 2008; Lefeuvre *et al.*, 2009). By bringing together divergent genome fragments that have largely independent evolutionary histories, recombination can potentially cause disruptions in coevolved intra-genome interaction networks (Voigt *et al.*, 2002; Martin *et al.*, 2005; Lefeuvre *et al.*, 2007). Examples of intra-genome interactions include: base-pairing interactions in RNA structures, sequence specific protein-DNA interactions, interactions between proteins (inter-protein interactions) and interactions between amino acids within three-dimensional protein folds (intra-protein interactions). Natural selection should disfavour the survival of recombinant genomes in which such interactions are disrupted and it is therefore expected that patterns of recombination evident within circulating viruses might display evidence of such selection. It has in fact been demonstrated in *in vitro* protein evolution experiments that the most viable of the chimaeric proteins that are expressed from recombinant genes tend to have lower degrees of predicted folding disruption relative to wild-type proteins than do randomly generated chimaeras (Voigt *et al.*, 2002). Importantly, similar observations were made when extending this approach to the analysis of chimaeric virus proteins that both occur naturally (Lefeuvre *et al.*, 2007) and emerge during evolution experiments (D. P. Martin *et al.*, 2011). An obvious explanation of these tendencies is that recombinants expressing chimaeric proteins in which certain necessary intra-protein amino acid interactions are maintained will have a higher likelihood of replicating and surviving, whereas those that don't will be purged by selective processes.

Besides potentially disrupting intra-protein amino acid interactions, it is similarly possible that whenever biologically functional nucleic acid secondary structures are present within virus genomes, recombination could disrupt nucleotide-nucleotide interactions within these. It is expected that recombinants in which biologically functional secondary structures are undisrupted should be more viable than those in which they are disrupted and, therefore, that natural recombinant genomes might display lower degrees of predicted secondary structural disruption than is expected if the maintenance of secondary structures had no evolutionary significance. While evidence of selection acting against recombinants with disrupted genomic nucleic secondary structures has been observed amongst recombinant virus genomes arising

during *in vitro* recombination experiments (D. P. Martin *et al.*, 2011), it remains to be determined whether such selection has a detectable impact on patterns of virus recombination that arise under natural conditions.

Here it is determined whether evidence exists within the recombination patterns of naturally occurring HIV recombinants of selection disfavours the survival of HIV-1M recombinants with either recombinationally disrupted intra-protein interactions or recombinationally disrupted RNA secondary structures. This chapter uses fairly simple permutation-based tests that rely on the simulation of natural and random recombinants and the approximation of protein and nucleic acid folding disruption in these recombinants to test whether natural recombinants display significantly lower degrees of disruption than is expected by chance.

Methods

Dataset construction

A set of aligned HIV-1 sequences was obtained that had been previously analysed in great detail to characterise inter- and intra- subtype recombination events (i.e. to estimate recombination breakpoint locations and identify both the recombinant sequences and sequences resembling their parents; Archer *et al.*, 2008). This alignment contained only the protein-coding regions of the analysed HIV-1M sequences and is hereafter referred to as the coding-region or CR dataset. The 434 recombination events detected within the 274 sequences included within this dataset were all manually checked with a range of breakpoint localisation and recombinant sequence identification tools available within the program RDP4 (Martin *et al.*, 2010) to yield a fairly accurate list of recombination events. A recombination event is considered here to be a recombinant sequence, a set of two breakpoints in that recombinant sequence and the identities of sequences in the dataset that most closely resemble the parental sequences that recombined in order to form the recombinant sequence. For any particular recombination event we differentiate between the parental viruses by referring to the sequence in the dataset that most closely resembles the sequence that provided the biggest fraction of the recombinant's genome as the "major" parent and the sequence resembling that which provided the smaller fraction as the "minor" parent.

In addition to the above dataset, a "full genome" HIV-1M dataset was constructed (hereafter referred to as the FG dataset) which included non-coding sequences for most of the isolates included in the CR dataset. The 274 sequences contained within the FG dataset correspond to the 274 sequences contained within the CR dataset except with the inclusion the 5' and 3' non-

coding regions, which were excluded in the CR dataset. Sequences in this dataset were aligned using MUSCLE (Edgar, 2004) with default settings. The FG dataset was produced to test the impact on recombination patterns of nucleic acid secondary structures within both the coding and non-coding regions of HIV-1M genomes.

PDB files containing the three-dimensional atomic coordinates of different protein crystal structures were downloaded from the Protein Data Bank (PDB, Bernstein *et al.*, 1977). The PDB files corresponding to protein structures used in the recombination protein-fold disruption test are presented in Supplementary Table 2.

Recombination Analysis

An automated RDP 4.20 analysis (Martin *et al.*, 2010) was performed to detect recombination events evident within the FG dataset. Only recombination events detectable with 6 or more of the seven independent recombination detection methods implemented in RDP 4.20 were taken into account in subsequent analyses. 400 recombination events were detected amongst the 274 full genome sequences present in the FG dataset using this automated RDP analysis.

Simulating recombinants

Given the breakpoint positions and parental sequences associated with a particular recombination event, an *in silico* generated recombinant sequence with breakpoint positions corresponding to those of a real recombinant was produced from the minor and major parental sequences. *In silico* generated recombinants of this type were called “mimic” or M-recombinants in that although they resembled actual recombinants at the moment when these were generated, they were not expected to be identical to these actual recombinants. We refer hereafter to the 5' breakpoint in this M-recombinant as the “start position” of the recombination event, the 3' breakpoint as the “end position” of the recombination event and the number of sites differing between the major and minor parents between these two positions as the “event-length”. For each of the M-recombinants, multiple simulated recombinant sequences, called S-recombinants, were then generated from the same major and minor parental sequences and with the same event-length but with a randomly selected start position (such that the end position is simply determined by the event-length). Keeping the event-length constant between the M- and S-recombinants ensures that these all have exactly the same genetic distance from both the major and minor parental sequences – something that is crucial for our permutation-based tests of recombination-induced protein and nucleic acid structural disruption. See Figure 4.1 for a diagrammatic representation of this procedure.

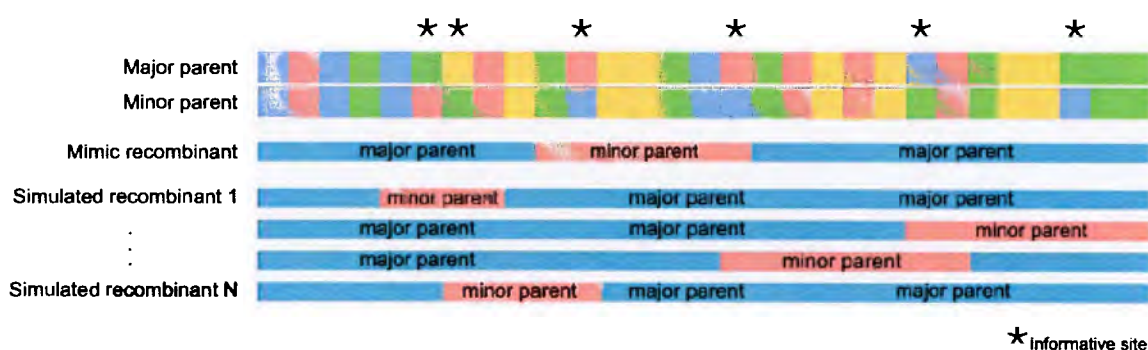


Figure 4.1: Diagrammatic representation of generating simulated recombinants. For a particular recombination event specifying a major parent, a minor parent, and a pair of recombination breakpoint locations delineating a fragment of sequence derived from the minor parent (containing in this particular case two nucleotides that vary between the major and minor parents), an *in silico* mimic of the real recombinant sequence is created using the minor and the major parent sequences. Following that, a set of N simulated recombinants is generated in a similar way to the mimic recombinant, but using random starting and ending positions, whilst maintaining the same number of variable nucleotides between the randomized breakpoint sites as occur in the mimic recombinant. In this example the mimic and simulated recombinant all have two sites between the 3' and 5' breakpoints that are not identical between the parental sequences.

Protein folding disruption tests

This test is based on that presented by Lefeuvre *et al.* 2007, and was performed using RDP 4.20 (Martin *et al.*, 2010). The test compares M- and S-recombinant sequences to their minor and major parental sequences in order to evaluate whether degrees of protein fold disruption, estimated using the SCHEMA method (Voigt *et al.*, 2002), are significantly lower in the M-recombinants than in the S-recombinants. Specifically, this test evaluates whether recombinant sequences display a degree of protein fold disruption that is lower than can be accounted for by chance under random recombination in the absence of natural selection. As input SCHEMA takes a PDB protein structure file and parental amino acid sequences that are homologous to those within the PDB file. Assuming identical folding of the major and minor parental amino acid sequences to that represented within the PDB file, the SCHEMA method identifies pairs of potentially interacting amino acid residues as any pair of amino acids with any atoms that are within 4.5 ångströms of one another. The amino acid contact map thus generated is then used to determine the degree of expected fold disruption within a chimaeric protein that is expressed from a recombinant gene. At all pairs of potentially interacting amino acid sites where the parental amino acid sequences differ from one another (i.e. polymorphic sites), SCHEMA counts the number of these where one amino acid of the potential interacting pair comes from the minor parent and the other comes from the major parent. This number, called the E-score, has previously been shown to be directly correlated with degrees of protein fold disruption (Voigt *et al.*, 2002).

In order to test whether recombinant protein sequences avoid protein fold disruption more than can be accounted for by chance, random recombination events were simulated in the manner described above - each of the real recombination events was shifted up or down along the nucleotide sequence by a random amount, whilst maintaining the same number of polymorphic nucleotides (termed "informative sites") as in the real recombinant. E-scores were calculated for each of the M-recombinants and their corresponding sets of S-recombinant proteins and a permutation test was performed which counted the fraction of times that real individual M-recombinant proteins were less disrupted (i.e. had a lower "E" score) than 100,000 simulated recombinant proteins. This fraction corresponds to the p-value that the real recombination events that are collectively represented amongst all the M-recombinants are no less disruptive of protein folding than are those represented amongst corresponding sets of S-recombinants: For a particular gene a p-value <0.05 therefore suggests a $>95\%$ confidence that recombination events detectable under natural conditions within that gene are less disruptive of protein folding than would be expected in the absence of either (1) selection disfavoring the survival of viruses that express chimaeric proteins with disrupted folds or (2) mechanistic factors that cause recombination events to occur most frequently at genomic sites where they will have minimal impact on protein folding.

Nucleic acid fold-disruption tests

Tests for recombination induced RNA secondary structure disruption were performed that were similar to those used to test for protein tertiary structure disruption. For each detected recombination event, one M- and ten S-recombinant sequences were generated and each of these sequences computationally folded along with the major and minor parental sequences using an optimised version of the UNAFold program (Rizk and Lavenier, 2009; Markham and Zuker, 2008). Differences in the folds between the parental sequences and those of their corresponding M- and S-recombinant sequences were quantified to determine whether the secondary structures of the M-recombinant sequences were significantly more similar to those of the parental sequences than were those of the S-recombinants. We used a variety of different metrics to quantify the similarity between the M- or S-recombinants and the parental sequences: (1) the average structural similarity of the folded M/S recombinant sequences to each of the parents were calculated using the mountain similarity metric (V Moulton *et al.*, 2000) by computing the metric between the folded M/S recombinant sequence and each of the two folded parental sequences and taking the average, (2) a base-pair disruption score was calculated as the number of positions that were paired in one or both of the parents but where the position in the M/S recombinant was either unpaired or did not correspond exactly to a base-pairing present at that position in at least one of the parents. A third (3) and fourth (4)

disruption score metric similar to (2) was calculated by considering each of the parents individually: (3) counted the number of base-pairings present in the minor parent that were not present in the M/S recombinant and (4) counted the number of base-pairings present in the major parent that were not present in the M/S recombinant. For each of these metrics we obtained (1) a list of similarity scores for the M-recombinants and (2) a list of similarity scores for the S-recombinants. These lists were compared using a two-tailed Wilcoxon-rank sum test to determine whether the similarity scores of the M-recombinants were significantly different than those of the S-recombinants. In these tests a higher similarity scores for the M-recombinants coupled with an associated p-value <0.05 would indicate with $>95\%$ confidence that recombination events detectable within HIV genomes are less disruptive of RNA folding than would be expected in the absence of either (1) selection disfavoring the survival of viruses in which recombination has disrupted RNA folding or (2) mechanistic factors that cause recombination events to occur most frequently at genomic sites where they will have a reduced impact on RNA folding.

Tests for associations between recombination breakpoints and structured genome regions

In addition to testing the impact of recombination on the disruption of RNA secondary structure, the impact of secondary structure on the locations of detectable recombination breakpoints was also investigated. Five different tests of this were performed (Tests 1-5). Tests 1-3 used all 274 sequences in the alignment of HIV-1 sequences to calculate the pairing probability at each column position. Test 1 calculated the average pairing probability within a 70 nucleotide window centred around each of the breakpoint positions in the M-recombinants which provided a distribution of averages which was compared against the distribution of averages for the S-recombinants using a Wilcoxon-rank sum test. Tests 2 and 3 were performed in a similar manner as test 1, except test 2 focused solely on the 5' recombination breakpoints and test 3 focused on solely on the 3' recombination breakpoints. Whereas Tests 1-3 used the entire alignment to calculate alignment-wide probabilities, tests 4 and 5 were computed using sequence-specific pairing proportions within a window centred around each of the breakpoints to obtain a measure of structure. In particular, test 4 was computed by calculating the proportion of paired nucleotides within a 70 nucleotide window centred around each of the breakpoint positions in each of the identified minor parents which resulted in a distribution of values which was compared against a simulated distribution where the breakpoint positions had been randomly permuted in the minor parents. Test 5 was performed in the same way as test 4, but made use of the identified major parent sequences instead.

Results and Discussion

Avoidance of protein fold disruption in HIV-1 recombinants

Recombination that occurs between divergent genome fragments having largely independent evolutionary histories can potentially disrupt coevolved intra-genomes interactions such as those occurring between amino acids within a three-dimensional protein fold (i.e. intra-protein interactions). Here chimaeric proteins resembling those expressed by actual viruses were tested to determine whether they displayed lower degrees of predicted folding disruption than those of randomly generated protein chimaeras. Actual recombinants sampled from nature would be expected to display less disruption of intra-protein interactions than simulated recombinants either if natural selection disfavoured the survival of recombinants expressing misfolded proteins, or recombination breakpoints tend to coincidentally occur most frequently at sites where it has minimal impact on protein folding.

Figure 4.2A illustrates degrees of intra-protein amino-acid—amino-acid interaction disruption (where higher E-values equate with greater disruption) that are predicted to occur within various HIV-1 proteins between HIV variants that have previously been observed to recombine (Figure 4.2B). The average degrees of predicted recombination induced folding disruption in the gp120 and Nef proteins are appreciably higher than those predicted in the other HIV-1 proteins for which extensive high resolution structural data is available. Consistent with the notion that these proteins may be particularly sensitive to recombination-induced folding disruption is the fact that, in actual HIV-1M recombinants, breakpoints only very rarely occur at sites where they are anticipated to have a maximally disruptive effect on the folding of these proteins. A plausible explanation for gp120 and Nef being particularly sensitive to recombination-induced folding disruption relative to the other HIV-1M proteins examined here is that they are less conserved than these other proteins (see amino acid substitution rate plot in Figure 4.2A) and recombinant versions of gp120 and Nef will therefore tend to have many more potentially disruptive amino acids combinations.

In order to more rigorously test whether the chimaeric proteins that are expressed by HIV-1M recombinants tend to display lower degrees of protein folding disruption than can be accounted for under random recombination in the absence of selection against misfolded protein chimaeras, individual HIV-1M proteins were analysed using a previously described permutation-based “avoidance of protein folding disruption” test (Lefeuvre *et al.*, 2007). This revealed that, for five out of the eight analysed HIV-1 proteins, intra-protein amino acid interactions were significantly less disrupted than could be accounted for by chance in natural

HIV-1M recombinants that displayed recombination breakpoints in the genes encoding these proteins (Table 4.1).

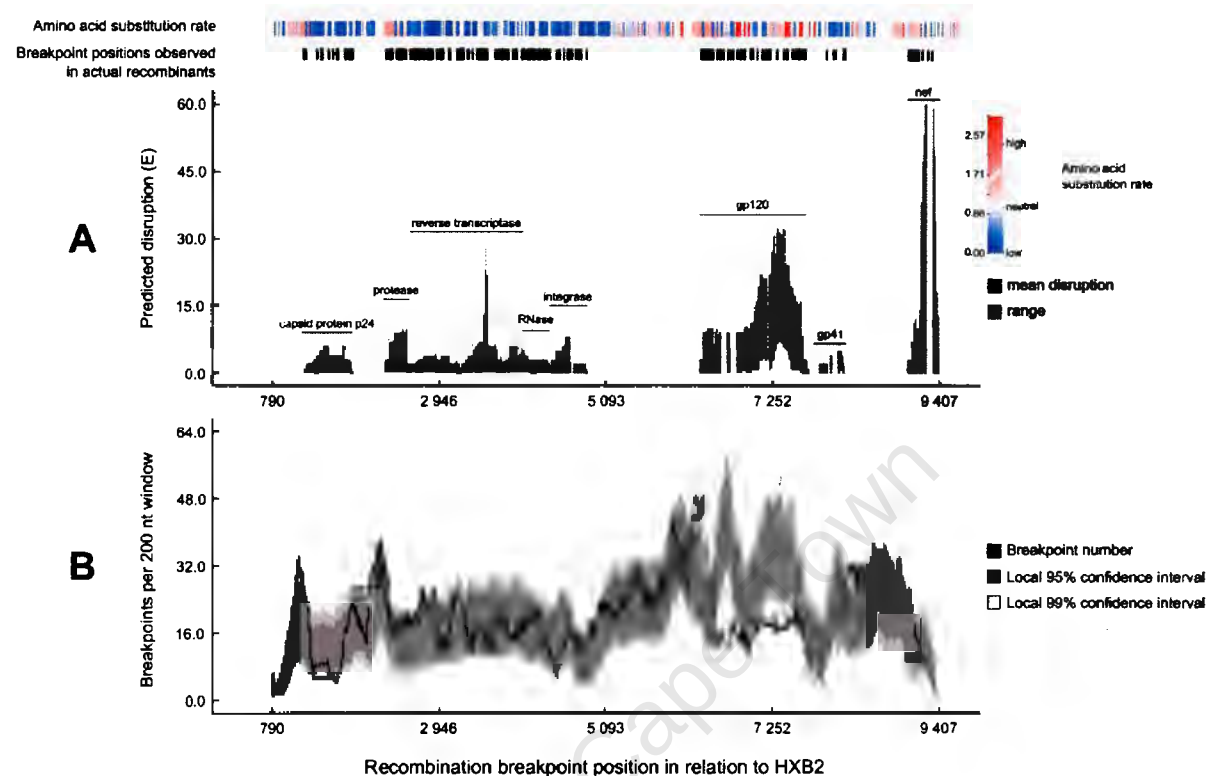


Figure 4.2A: The predicted sensitivity of HIV-1M proteins to recombination-induced disruption. Depicted are the means (black lines) and ranges (gray backgrounds) of predicted degrees of recombination-induced folding disruption in various HIV-1 proteins (those for which suitable atomic resolution three dimensional structures are available). For all genome regions that had associated protein structure data, all conceivable single breakpoint recombinants were simulated using parental sequences that resembled as closely as possible the parental sequences of actual recombinant viruses with detectable recombination breakpoints in these genome regions. Amino acid substitution rates and breakpoint positions occurring in these actual HIV-1 recombinants are displayed at the top of the figure.

Figure 4.2B: Recombination breakpoint density plot illustrating breakpoint positions detected across 434 detectable HIV-1M recombination events (After Archer *et al.*, 2008). Note firstly that the peaks of the plots in A indicate recombination breakpoint positions that are predicted to have the greatest disruptive effects on protein folding, and secondly that across actual known recombinant HIV-1M genomes these “disruptive breakpoint positions” tend to correspond in plot B with regions of low recombination breakpoint density.

The same avoidance of protein folding disruption test was also performed using a completely automated RDP recombination analysis of the same CR dataset to assess how robust the disruption test was to the recombination dataset used (Supplementary Table 3). Despite this re-analysis yielding slightly weaker p-values, four of the five proteins that had previously displayed significant evidence of protein fold disruption avoidance using the manually curated recombination dataset once again displayed evidence of significantly less recombination-induced disruption than expected (p-values < 0.05). The fifth protein, P24 (the capsid protein), also yielded a low p-value (0.0967) which is suggestive of it too displaying lower than expected recombination-induced folding disruption.

This demonstration that datasets assembled using fully automated alignment and recombination detection steps can be effectively used to detect evidence that selection may disfavour the survival of recombinants that express chimaeric proteins with disrupted folds is important in that it indicates that such analyses could be productively carried out with a minimum of user supervision. Although in such automated analyses the accuracies of both the input sequence alignments and the recombination breakpoint distribution maps that are inferred are likely decreased relative to analyses based on manually supervised alignment and recombination breakpoint detection, these factors apparently only slightly decrease the power of the fold-disruption test.

There are two non-mutually exclusive potential explanations why breakpoints in natural recombinants might occur at genomic sites where they minimise protein fold disruption. The most obvious of these explanations is that the expression of a misfolded chimaeric protein is expected to have a negative impact on a virus' fitness such that recombination patterns observable amongst viruses sampled from nature will largely reflect the consequences of selection disfavouring the survival of viruses that express such proteins. The less obvious, but not less plausible, explanation is that HIV-1M genomes are mechanistically predisposed to accumulate recombination breakpoints at locations that minimise the chances of low-viability recombinants arising. Specifically, RNA secondary structures within HIV-1M genomes have a strong influence both on where recombination breakpoints are likely to occur (Simon-Loriere *et al.*, 2010) and on how proteins are likely to fold (Watts *et al.*, 2009). It is therefore possible that RNA structures occurring both at the junctions of different genes and at sites encoding the boundaries between discrete protein domains, "direct" recombination to preferentially occur at locations in genes where it will have a minimal impact on protein folding (Simon-Loriere *et al.*, 2010).

Table 4.1: Degrees of protein fold disruption in natural and simulated HIV-1 recombinants.

Protein	Number of breakpoints	Mean E-score of M-recombinants	Mean E-score of S-recombinants	p-value ¹
Capsid protein P24	18	0.059	0.478	0.01060
Protease	16	1.875	1.971	0.50419
Reverse transcriptase	67	0.321	1.216	0.00000
RNase	22	0.818	1.044	0.32595
Integrase	21	0.211	0.868	0.01886
gp120	51	5.465	9.365	0.00000
gp41	5	1.000	1.954	0.16177
Nef	14	2.091	5.502	0.01848

¹The p-value is the probability that mimic recombination breakpoints do not tend to avoid disrupting protein folding. Recombinant proteins with a p-value less than 0.05, [REDACTED], are considered to be significantly less disrupted than can be accounted for by chance - 5 out of 8 HIV-1 proteins analysed here were detected as such, whilst the other 3 displayed no significant tendency in either direction.

Although three of the eight proteins analysed here displayed no detectable signal of lower-than-expected recombination-induced fold-disruption, in at least one case (gp41) this may simply be due to low numbers of recombination breakpoints having been detected within the protein: A fact which reduces our ability to detect a signal in this protein. In this regard it is noteworthy that for all three of these proteins (gp41, RNase and Protease), the mean estimated fold disruption in the M-recombinants were consistently lower than that of the S-recombinants (compare the E-scores in Table 4.1) which suggests that given either more data or a more powerful fold disruption tests, it might be possible to demonstrate that these proteins too display lower than expected degrees of recombination-induced fold disruption.

Co-localisation of recombination breakpoint positions and RNA secondary structures

It has previously been shown during selection-free *in vitro* recombination experiments involving the HIV-1M *env* gene (which encodes gp120 and gp41) that recombination breakpoints tend to very obviously co-localise with RNA secondary structures (Simon-Lorier *et al.*, 2010). Although less obvious, a similar tendency was also detected accross the entire genomes of natural HIV-1M recombinants (which have obviously also been subject to natural selection). Crucially, both the *env* and full-genome associations were detected based on verified RNA secondary structures identified within a single complete HIV genome using the SHAPE method (Watts *et al.*, 2009).

Various association tests were devised both to determine whether similar associations could be detected using only computationally inferred HIV genomic RNA secondary structures, and to further characterise factors underlying the colocalisation of recombination breakpoints and secondary structures. Application of these tests to the FG dataset (for which completely automated alignment and recombination detection had been carried out) clearly revealed that recombination breakpoints displayed a significant tendency to fall at sites computationally inferred to be involved in base-pairing ($p = 0.014$; Test 1 in Table 4.2). This trend seemed more attributable to the 3' recombination breakpoint which displayed a significant tendency to occur at base-paired sites, whereas the the 5' recombination breakpoint only displayed a slight tendency to occur at base-paired sites (p -valued = 0.041 and 0.134 respectively; Tests 3 and 2 in Table 4.2). It was found, however, that base-pairing at the breakpoint sites of the major parent (i.e. the parent providing sequences upstream of the 5' breakpoint and downstream of the 3'

breakpoint) is more strongly associated with breakpoint positions than is base pairing at the breakpoint site in the minor parent (p-values = 0.020 and 0.131 respectively; Tests 4 and 5 in Table 4.2). An analysis of the CR dataset (for which a rigorous manual alignment and recombination detection were carried out) revealed similar but weaker signals of recombination breakpoints tending to occur at base-paired positions (Supplementary Table 4).

These observations are entirely consistent with expectations given that copy-choice recombination is the predominant mechanism of recombination that is thought to occur in HIV. Specifically, it seems plausible that recombination is initiated when secondary structure within an RNA molecule that is being reverse transcribed either stalls and/or causes the reverse transcriptase to disengage the “major parent” RNA. If reengagement occurs on the minor parent RNA a recombination breakpoint will occur at a site corresponding to a structured region in the major parent but not necessarily at a structured region in the minor parent. However, given that the breakpoints do not obviously co-localise with base-paired sites in the minor parent it is apparent that switching of the reverse transcriptase back to its original template (as very frequently occurs during HIV recombination, Moumen *et al.*, 2003; Galetto and Negroni, 2005) is possibly not as strongly influenced by secondary structure.

Table 4.2: Breakpoint position tests performed on the full-genome (FG) recombination dataset.

Test ¹	Median mimic value	Median simulated value	z-score	p-value	Comments
1. HIV-1 alignment-wide pairing probability near recombination breakpoints	0.600	0.588	2.451	0.014	Recombination breakpoints occur significantly more frequently within predicted paired regions in the alignment than can be accounted for by chance.
2. HIV-1 alignment-wide pairing probability near 5' recombination breakpoints	0.593	0.580	1.500	0.134	5' recombination breakpoints occur slightly more frequently within predicted paired regions in the alignment.
3. HIV-1 alignment-wide pairing probability near 3' recombination breakpoints	0.610	0.598	2.045	0.041	3' recombination breakpoints occur significantly more frequently within predicted paired regions in the alignment than can be accounted for by chance.
4. Minor sequence pairing probability near recombination breakpoints	0.612	0.594	1.512	0.131	Slight tendency for recombination breakpoints to occur within predicted paired regions in minor parent sequences.
5. Major sequence pairing probability near recombination breakpoints	0.614	0.586	2.327	0.020	Significant tendency for recombination breakpoints to occur within predicted paired regions in major parent sequences than can be accounted for by chance.

¹Described in "Tests for associations between recombination breakpoints and structured genome regions".

Some evidence of RNA structure disruption in HIV-1M recombinants

Besides influencing where recombination is most likely to occur, RNA structures could also influence which recombinants are likely to survive following the occurrence of recombination. If, for example, a recombination breakpoint occurs within the sequence of a biologically functional hairpin structure it is possible that nucleotide differences between the parental genomes will cause destabilisation of the structure, and, consequently, a reduction of the resulting recombinant's fitness. A pair of tests similar to that used to investigate recombination-induced protein folding disruption was devised to test whether recombination-induced RNA folding disruption has had a detectable influence on recombination breakpoint distributions that are detectable in natural HIV-1M recombinants.

Whereas the first of these tests (Test 1 in Table 4.3) measured whether recombinant sequences tended to maintain the overall computationally predicted structural conformations of their parental sequences, tests 2-4 examined whether individual base-pairs within the minor and major parental secondary structures were maintained in the recombinant secondary structure (Test 2-4 in Table 4.3).

In an analysis of the full-genome recombination dataset (FG; Table 4.3), none of these tests yielded any evidence that recombination events occurring within natural HIV-1M recombinants tended to yield recombinant genomes with lower degrees of RNA structure disruption than would be expected under random recombination in the absence of selection against the survival of genomes carrying recombinationally disrupted RNA structures. In an attempt to increase the strength of these tests they were focused exclusively on regions of the genome containing two known biologically relevant conserved secondary structures: the Gag-pol frameshift structural element and the Rev Response element. 3 recombination breakpoints were detectable within the former region and 3 were detectable within the latter region. When considering these two regions separately there was no convincing evidence (z-score = -1.480, p-value = 0.138 and z-score = -0.325, p-value = 0.745, respectively) that detected recombination events in these regions were less disruptive of these secondary structures than could be accounted for by chance, nor was there evidence when jointly considering these two regions (z-score = -0.682, p-value = 0.495).

Table 4.3: Secondary structure fold-disruption tests performed on the full-genome (FG) recombination dataset.

Test	Median mimic value	Median simulated value	z-score	p-value	Comments
1. Recombinant average structural similarity to parents	0.951	0.915	0.411	0.681	No significant tendency for folds of mimic recombinant sequences to be more similar to folds of both their parental sequences than are the folds of simulated recombinant sequences.
2. Disruption score (number of positions that are paired in either of the parents where there is not a base-pairing in the recombinant matching at least one of the parental base-pairings)	783	923	-0.650	0.515	No significant tendency for mimic recombinant sequences to have folds which are more similar to their minor and major parent sequences than are the folds of the simulated recombinant sequences.
3. Disruption score (minor parent only, number of base-pairs in minor parent disrupted in recombinant)	3948	3966	-0.198	0.843	No significant tendency for folds of mimic recombinant sequences to be more similar folds of their minor parent sequences than are the folds of the simulated recombinant sequences.
4. Disruption score (major parent only, number of base-pairs in major parent disrupted in recombinant)	1307	1408	-0.550	0.582	Slight tendency for folds of mimic recombinant sequences to be closer the folds of their major parent sequences than are the folds of the simulated recombinant sequences

Given that the CR dataset (Table 4.4) yielded stronger evidence than the FG dataset (Table 4.3) that observable recombination breakpoints had been influenced by recombination-induced protein folding disruption, the RNA fold disruption tests were further applied to this dataset. For this dataset test 2 (which considered base-pair disruption in comparison to both the minor and major parent folds) indicated that recombination events were detectably less disruptive of base-pairings than could be accounted for by chance (z-score = -1.993, p-value = 0.046, Table 4.4). Using test 3 and 4, which considered base-pair disruption separately in comparison to the minor and major parent folds, respectively, only slight evidence was found that base-pairs were less disrupted in the M-recombinants than in the S-recombinants, however, the corresponding p-values (p-value = 0.645 for minor parent base-pair disruption and p-value = 0.127 for major parent base-pair disruption, Table 4.4) do suggest that base-pairs in M-recombinants have a greater tendency to be less disrupted when compared to the predicted major parent folds, more so than when compared to the predicted minor parent folds. This is not entirely surprising given that the major parent sequences contribute a larger fraction of sequence than the minor parent sequences to the M/S-recombinant sequences and therefore are expected to have a bigger impact on their folding.

Table 4.4: Secondary structure fold-disruption tests performed on the coding-region (CR) recombination dataset.

Test	Median mimic value	Median simulated value	z-score	p-value	Comments
1. Recombinant average structural similarity to parents	0.395	0.272	1.360	0.174	No significant tendency for folds of mimic recombinant sequences to be more similar to folds of their parental sequences than are the folds of simulated recombinant sequences.
2. Disruption score (number of positions that are paired in either of the parents where there is not a base-pairing in the recombinant matching at least one of the parental base-pairings)	546	633	-1.993	0.046	Significant tendency for folds of mimic recombinant sequences to be more similar the folds of both their minor and major parent sequences than are the folds of the simulated recombinant sequences.
3. Minor parent disruption score (number of base-pairs in the minor parent not present in the recombinant)	3954	3940	-0.461	0.645	No significant tendency for folds of mimic recombinant sequences to be more similar folds of their major parent sequences than are the folds of the simulated recombinant sequences.
4. Major parent disruption score (number of base-pairs in the major parent not present in the recombinant)	766	900	-1.527	0.127	Slight tendency for folds of mimic recombinant sequences to be closer the folds of their major parent sequences than are the folds of the simulated recombinant sequences

While it is possible that selection favouring the maintenance of biologically functional RNA secondary structures has only a small influence on patterns of recombination in HIV, it is also possible that the tests that were applied here were simply not powerful enough to easily detect such an influence. In this regard the tests that we have performed have four potentially important shortcomings: (1) the actual parental sequences of naturally occurring recombinants were not used in these tests and it is entirely possible that with these in hand subtle structural differences between the actual and simulated recombinant genomes would have been clearer; (2) the individual recombination events that were analysed involved single pairs of 5' and 3' recombination breakpoints and were probably not representative of natural recombinants which frequently have more than two detectable breakpoints; (3) the accuracy of computational secondary structure prediction is not perfect and it is likely therefore than incorrectly inferred base-pairing interactions decayed (at least slightly) the power with which disruptions of actual base-pairing interactions could be estimated; (4) the possible inclusion of non-viable viruses within the set of analysed sequences could have decreased the power of our tests because it would have violated the implicit assumption that all of the analysed HIV genomes were reasonably fit and were therefore likely free of recombination-induced RNA structure disruption.

Conclusions

Recombination events detectable in the coding regions of five out of eight of the HIV-1 proteins analysed here were found to be less disruptive of intra-protein amino-acid—amino-acid interactions than could be accounted for by chance. This result is what we would anticipate if chimaeric proteins that are expressed by recombinants are sometimes misfolded and cause decreased viral fitness. Natural selection would act against recombinants expressing such proteins and, as a result, viruses sampled from nature would tend to have recombination breakpoints at sites that had minimal impact on protein folding.

Similar analyses of RNA folding in recombinants yielded only marginal evidence that maintenance of RNA secondary structure had any appreciable impact on patterns of recombination detectable within natural HIV-1M recombinants. In fact the only signal that natural selection might be acting against recombinants with misfolded RNA was detectable with our CR dataset which had been aligned and analysed for recombination with rigorous manual supervision. Although this does provide some evidence that selection favours the survival of HIV-1M recombinants in which base-pairing interactions within biologically important RNA secondary structures is maintained, it was somewhat surprising given both the large numbers of recombination events considered and the pervasiveness of secondary structures within HIV-1M genomes, that this affect was not clearer. Even when specifically focusing on two HIV-1 genomic secondary structures with well-established biological functions (the gag-pol frameshift element and the rev-response element) no clear evidence was discernible of natural recombinants displaying lower degrees of structural disruption than simulated recombinants.

Despite the weak signal of natural recombinants avoiding the disruption of nucleotide-nucleotide interactions within biologically important nucleic acids secondary structures, we did detect a strong signal of recombination breakpoints tending to occur more in structured than in unstructured parts of HIV-1 genomes. This result is consistent with previous studies which have found that particularly structured regions of the HIV-1 genome tend to both co-localise with recombination breakpoint positions and be associated with gene and protein domain boundaries. It has been proposed that, by “directing” recombination breakpoints to domain boundaries, secondary structures within the HIV-1M genome may maximise the chances that recombinant genomes will express properly folded chimaeric proteins (Simon-Lorier *et al.*, 2010). Therefore, besides selection favouring the survival of recombinants with minimally disrupted protein and RNA folding, it is plausible that in HIV-1 selection has favoured the formation of secondary structures at genomic sites where recombination will have minimal impact on protein and RNA folding.

5 Incorporating alignment uncertainty into RNA secondary structure prediction

Introduction

Accounting for evolutionary processes that preserve the functionality of biologically relevant RNA secondary structures can significantly enhance the accuracy with which such structures can be predicted (Knudsen and Hein, 1999). It is for this reason that tools such as PPfold (Sükösd *et al.*, 2011) and RNAalifold (Bernhart *et al.*, 2008) use aligned sets of sequences to make RNA secondary structure predictions. RNA secondary structure prediction from multiple sequences is termed comparative secondary structure prediction (Hofacker *et al.*, 2002). The increased accuracy achievable with such approaches stems from the simple fact that structural conservation in most cases requires the maintenance of base-pairing complementarity. The underlying assumption required for these approaches to work is that the sequences used must have either the same or very similar secondary structures.

Paradoxically, in order for this approach to be effective, the sequences used must have enough variation to guide prediction (Gutell *et al.*, 2002). Pairs of homologous nucleotides in an alignment that co-vary in lockstep with one another as if to maintain their Watson-Crick base-pairing complementarity provide good evidence that nucleotides at particular positions are base-paired with one another within a secondary structure. Selection will act to preserve base-pairing complementarity between pairs of positions in sequences where base-pairing is important for the maintenance of a biologically functional structure (Hancock *et al.*, 1988; Dutheil *et al.*, 2005).

Just as too little variation between analysed sequences can preclude the detection of such “complementarily coevolving” nucleotide pairs, too much variation between sequences can also be problematic, particularly for approaches that work using a multiple alignment of sequences (Gardner and Giegerich, 2004). The reason for this is that sequences need to be similar enough that homologous nucleotides can be correctly identified and aligned using a sequence alignment program. Two or more nucleotides in different sequences are considered to be homologous if they share a common ancestral nucleotide, this is reflected in a multiple sequence alignment by placing them in the same column (Bradley *et al.*, 2009).

Accurately aligning a set of sequences can itself be a challenging problem (Hein *et al.*, 2000; Bradley *et al.*, 2009) that can become particularly acute during comparative RNA secondary structure prediction since the quality of an alignment will affect the accuracy with which

secondary structures can be predicted using that alignment (for example, consider an alignment where the sequences are randomly aligned). Most currently available multiple sequence alignment methods such as those implemented in the computer programs ClustalW (Thompson *et al.*, 1994), MAFFT (Kato and Toh, 2008) and MUSCLE (Edgar, 2004) rely on score-based heuristics rather than attempting to model the underlying biological process of substitution, insertion and deletion. For many of the “heuristic” alignment methods this frequently causes systematic biases in the locations where insertion and deletion events are invoked to align nucleotides. Such biases can lead to the alignment of non-homologous nucleotides (Lunter *et al.*, 2008; Hein *et al.*, 2000). It is for this reason that lower quality alignments will likely yield less accurate comparative secondary structure predictions than will higher quality alignments.

Statistical alignment (Hein *et al.*, 2000; Anderson *et al.*, 2013) provides a solution to many of the issues encountered with heuristic multiple sequence alignment approaches. It models sequence evolution as a stochastic process involving sequence insertions, deletions and character substitutions, which defines a probability distribution over the alignments of the sequences. Using techniques such as Expectation Maximisation or Markov Chain Monte Carlo (MCMC) one can aim to either maximise the likelihood of the alignment, or generate a sample of alignments that is representative of the underlying probability distribution. One major drawback of the statistical alignment approach is that it is computationally inefficient: for large alignments a large number of iterations need to be taken to ensure both that the MCMC algorithm converges to the correct target distribution, and that one can be certain that the resulting set of alignment samples is statistically representative. Despite this drawback, two major advantages of statistical alignment over computationally more efficient heuristic alignment approaches are the use of biologically realistic models rather than score-based heuristics, and the production of a sample of different similarly plausible alignments rather than a single consensus alignment. This second advantage is particularly important in the context of comparative secondary structure prediction in that it allows one to explicitly account for the alignment when identifying complementarily coevolving nucleotide pairs.

In this chapter we describe StatAlign 2.0 (Arunapuram *et al.*, 2013), an approach that combines secondary structure prediction with statistical alignment. We show that predicting secondary structures from large numbers of similarly plausible multiple sequence alignments that are generated by a statistical alignment approach enables more accurate secondary structure prediction. Finally, we introduce an improved method for evaluating the reliability of a given secondary structure prediction.

Methods

Structure prediction from large numbers of multiple sequence alignments

Alignment sampling

StatAlign (Novák *et al.*, 2008) is a statistical alignment software package that allows joint sampling of multiple alignments, phylogenetic trees and evolutionary parameters. The insertion-deletion model used is a modified version of the TKF92 model (Thorne *et al.*, 1992). Substitutions are modelled using a nucleotide substitution model of the users choosing - for our purposes we use the HKY85 model. Using StatAlign it is possible to generate a statistically representative sample of similarly plausible alignments. Specifically, alignments of a particular set of sequences are sampled by StatAlign at a rate that is proportional to their likelihood under the specified evolutionary model (see Figure 5.1 for a graphical depiction). For example, suppose there are 2 possible alignments, Σ_1 and Σ_2 , for a given set of sequences (in reality there will be an enormous number of possible alignments) and suppose that under the model Σ_2 is twice as likely as Σ_1 , then if one were able to generate a large number of samples, the frequency of alignments Σ_1 and Σ_2 in the final sample would approximate $1/3$ and $2/3$, respectively.

Typically users want some form of consensus alignment that summarises the model's "best guess" at what the true alignment looks like. Under a maximum likelihood approach the alignment with the highest probability would be chosen as the consensus alignment (alignment Σ_2 in the above example). Another approach would be to use an algorithm called *maximum posterior decoding*, which takes contributions from all alignments in a given sample and aims to individually maximise the accuracy at each alignment column. This has the advantage of using all available information instead of placing all of one's confidence in a single alignment. Regardless of the approach used, a consensus alignment is typically what most users will use for downstream analyses. For example, a good choice of alignment on which to perform a comparative secondary structure prediction might be the maximum posterior decoding alignment. In StatAlign 2.0 rather than using a consensus alignment for comparative secondary structure prediction, we use the individual alignment samples directly: i.e. we perform secondary structure predictions using large numbers of multiple sequence alignments. To achieve this two different approaches were developed: (1) *base-pairing averaging* and (2) *phylogenetic averaging*.

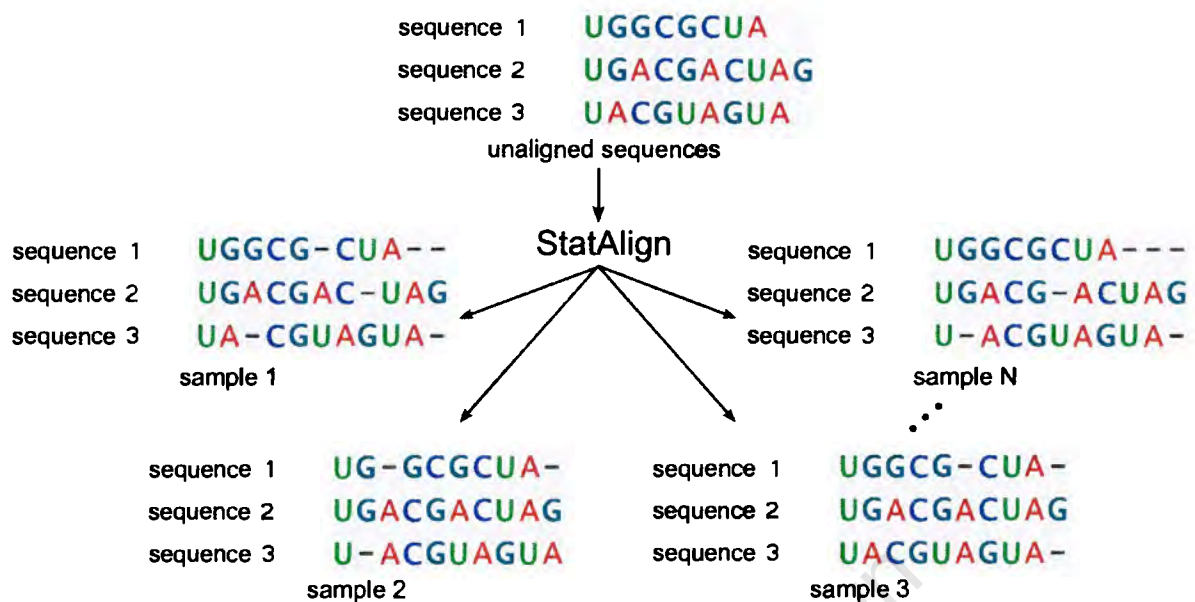


Figure 5.1: Sampling alignments with StatAlign. Using StatAlign it is possible to generate a statistically representative sample of alignments by randomly sampling alignments at a rate proportional to their likelihood under the model. In each of the alignment samples the ordering of the nucleotide characters within each of the sequences remains exactly the same but gaps are inserted at different positions. The positions of gaps reflect the model's hypotheses about nucleotide homologies. Two or more nucleotides in different sequences are considered to be homologous if they share a common ancestral nucleotide. The overarching aim of multiple sequence alignment is to accurately predict these nucleotide homologies.

Base-pairing averaging

In this approach StatAlign is used to produce a set of alignment samples. For each alignment within the sample an RNA secondary structure prediction is performed to produce a base-pairing probability matrix. The corresponding entries of each of the base-pairing probability matrices derived from each of the original sampled alignments are averaged to yield a single 'averaged' base-pairing probability matrix. In this way all the information relating to predicted RNA secondary structures from different alignment samples is combined.

Averaging the corresponding entries of the base-pairing probabilities matrices is not as simple as it may first appear. The main problem is that different alignment samples typically have different lengths due to the introduction of variable numbers of gap characters, such that different base-pairing probability matrices (each corresponding to a particular alignment sample) might have different dimensions (the dimensions of these square matrices are equal to the lengths of their corresponding alignments). Clearly, it is not possible to average matrices of different dimensions since it is not possible to sum the matrices.

In order to work around this problem a method was devised to project the matrices in such a way that they will all have the same dimensions. For a given set of input sequences the sequence containing the greatest number of non-gap characters is chosen as the reference sequence. Each

base-pairing probability matrix is then projected by deleting columns and rows of the matrix corresponding to gap positions in the reference sequence of the alignment sample used to generate that matrix. This ensures that all of the matrices produced from any particular alignment sample will all have dimensions equal to the number of non-gap characters in the reference sequence. It is then trivial to calculate an averaged base-pairing probability matrix (simply by summing the corresponding entries and dividing the sum by the total number of matrices/samples).

It should be pointed out here that each of the individual base-pairing probability matrices produced during this process represents the ensemble of possible secondary structures for a particular sampled alignment. The information contained within a base-pairing probability matrix can itself be used to derive a consensus secondary structure prediction for all of the sequences in the alignment. Using an averaged base-pairing probability matrix means that we combine information from many different secondary structure predictions, which, by explicitly accounting for alignment uncertainty, will hopefully result in a more accurate consensus secondary structure prediction. The algorithm used to derive a consensus secondary structure prediction from the base-pairing probability matrix is termed maximum expected accuracy (MEA; B. Knudsen and J. Hein, 2003). MEA aims to find the secondary structure that best maximises the expected number of correctly predicted paired and unpaired sites.

Phylogenetic averaging

Our second approach which also makes use of multiple alignment samples is to average over the phylogenetic probabilities used by PPfold. The phylogenetic probabilities are the prior base-pairing probabilities that are calculated by PPfold using the input alignment. PPfold uses these prior base-pairing probabilities together with a pre-specified grammar (Lari and Young, 1990; Knudsen and Hein, 1999) to infer the most-probable secondary structure. Usually the phylogenetic probabilities calculated are based on a single alignment, in this approach we calculate the phylogenetic probability matrix based on large numbers of multiple sequence alignments. Therefore, an alternative method for accounting for alignment uncertainty is then to (1) calculate the phylogenetic probabilities of each of the alignment samples, (2) project these to obtain identically dimensioned phylogenetic probability matrices and (3) average over these matrices in a similar way to that carried out for base-pairing probability matrices in the *base-pair averaging* method described above.

Reliability measure of secondary structure prediction accuracy

A modified reliability score, called the "pairs-only reliability score" was developed. This score is a measure of the model's belief in how accurate a particular secondary structure prediction is with respect to the F-score measure of predictive accuracy. This improved reliability score is calculated simply by averaging over the posterior base-pairing probabilities that correspond to base-paired sites in the consensus structure:

$$\text{Pairs-only reliability score} = \frac{\sum_{(i,j) \text{ base-paired}} P(i,j)}{N}$$

Where $P(i,j)$ is the posterior base-pairing probability of nucleotides i and j as predicted by the model and N is the number base-paired positions. This reliability score is a modified version of the PPfold reliability score. Whereas the PPfold version averages over the posterior base-pairing probabilities of both base-paired and non-base-paired sites, the score described here considers only base-paired sites; in much the same way as the F-score which only considers base-paired sites.

Results and Discussion

Predicting secondary structure from large numbers of multiple sequence alignments

Extensive benchmarking was carried out in order to test the predictive accuracy of the two novel approaches of secondary structure prediction from large numbers of multiple sequence alignments (Table 5.1). Predictive accuracy was measured against 44 experimentally determined structures from the RFAM database (Griffiths-Jones *et al.*, 2005) using the F-score measure of predictive accuracy. First, secondary structure predictions were performed on manually aligned versions of the 44 RFam datasets, referred to as "reference alignments" and benchmarked. These reference alignments were manually aligned based on experimentally determined RNA structures, resulting in high quality "gold standard" alignments. The average predictive accuracies on these reference alignments was much higher than both the novel and standard single alignment secondary structure prediction approaches, which was expected. Re-aligning these reference alignments, as was done using during benchmarking of the standard single alignment approaches, discards information relating to the manual alignment of the sequences which results in poorer quality alignments, however, these poorer quality alignments will be typical of what is available when performing a secondary structure prediction, because

usually one will not have access to structural information which was used to guide the reference alignments in the first place. It is therefore reassuring that both PPfold and RNAalifold achieve high predictive accuracy on the reference alignments (Average F-scores = 0.710 and 0.746, respectively) given that these alignments should accurately reflect the true nucleotide homologies, which should result in less alignment uncertainty being introduced into the secondary structure prediction.

The best single alignment method from the perspective accurate secondary structure prediction was MAFFT which produced alignments yielding better predictive accuracy with both PPfold (Average F-score = 0.567) and RNAalifold (Average F-score = 0.599) than those produced by MUSCLE, ClustalOmega and StatAlign posterior-decoding. This may suggest that the MAFFT alignment tool is the most accurate in being able to infer the underlying nucleotide homologies, however, it has been found that there are biases amongst multiple sequence alignment programs that lead some programs to achieve better performance on certain types of alignments (Golubchik *et al.*, 2007) suggesting that one should not place too much importance in this result.

The benchmark results for our two novel approaches show that the phylogenetic averaging method, which uses an averaged phylogenetic probability matrix that is supplied to PPfold to perform a structure prediction, is superior to all other methods, achieving substantially higher predictive accuracy (Average F-score = 0.614). Whereas the base-pair averaging method, which averages over base-pairing probability matrices, also achieves higher predictive accuracy than the alignment-based methods when using RNAalifold (Average F-score = 0.605). The predictive accuracy of the base-pair averaging method using PPfold is slightly lower than that achieved with MAFFT alignments and only slightly higher than that achieved with the StatAlign maximum posterior decoding alignments.

Table 5.1: Average F-scores (signifying predictive accuracy) for RNA secondary structure prediction benchmarks on 44 re-aligned RFam alignments.

	PPfold	RNAalifold
Reference alignments	0.710	0.746
MUSCLE alignments	0.539	0.572
MAFFT alignments	0.567	0.599
Clustal Omega alignments	0.554	0.598
StatAlign maximum posterior decoding alignments	0.561	0.584
StatAlign: base-pairing averaging	0.562	0.605
StatAlign: phylogenetic averaging	0.614	N/A

Whilst, both our methods for predicting secondary structures from large numbers of multiple sequence alignments improve significantly on the predictive accuracy of both PPfold and RNAalifold in comparison to when PPfold and RNAalifold are applied to single alignments produced by a variety of popular multiple sequence alignment tools, it is worthwhile exploring some of the advantages and disadvantages of both these methods and to compare them to one another. A disadvantage of the base-pair averaging approach is that time-consuming secondary structure prediction is required for each alignment sample: A factor, limiting the number of samples that one might want to consider. An advantage of this approach over the phylogenetic averaging approach is that any secondary prediction method that generates a base-pairing probability matrix can be employed, giving the approach greater flexibility.

While an advantage of the phylogenetic-averaging approach over the base-pair averaging approach is that it only requires the prediction of a single secondary structure, a significant disadvantage is that can only work in conjunction SCFG-based folding methods such as that implemented in PPfold. The reason for this is that these methods use a set of prior base-pairing probabilities (phylogenetic probabilities) to calculate a final structure, whereas thermodynamic folding methods such as RNAalifold do not.

For both the base-pairing averaging and phylogenetic averaging approaches as the number of sequences or their lengths increase, the statistical alignment step will eventually become more computationally limiting than the secondary structure prediction step. The reason for this is that secondary structure prediction using PPfold or RNAalifold has an algorithmic time complexity of $O(N^3)$ where N is the length of the alignment which is considered tractable, whereas statistical alignment using StatAlign has a time complexity which is the product of the sequence lengths, i.e. $O(N^m)$, where m is the number of sequences. In practice StatAlign 2.0 can predict secondary structures for alignments of 10 sequences, each 1000 nucleotides long, in ~2 hours on a standard desktop CPU. In order decrease the time required by StatAlign to fold alignments of this length, code optimizations and parallelisation of time consuming routines are currently under development.

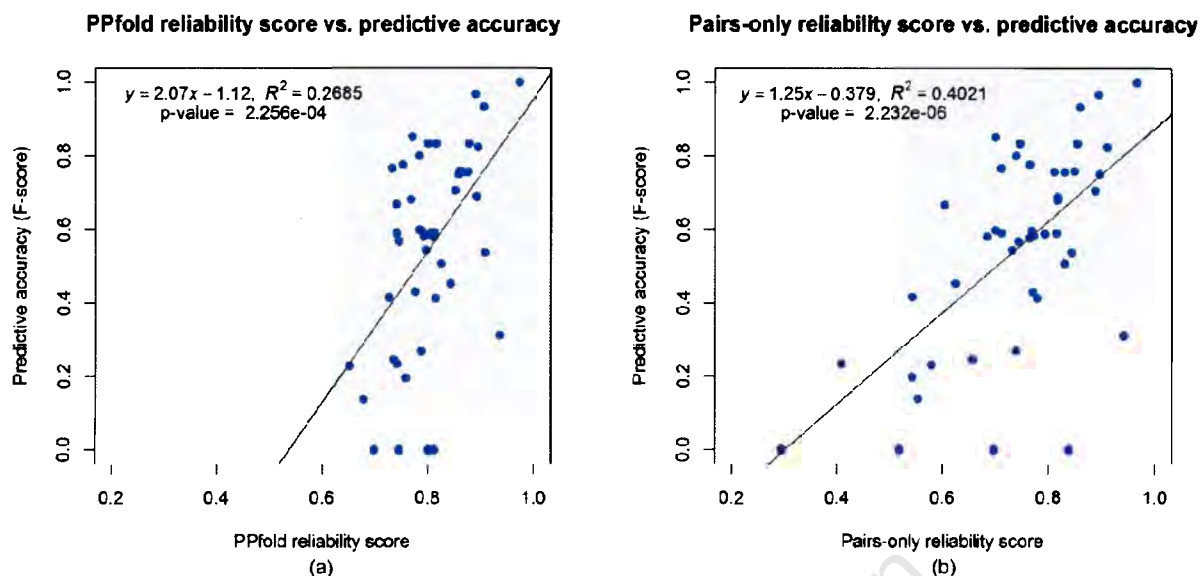


Figure 5.2: Correlation of two reliability scores, the PPfold reliability score (a) and the pairs-only reliability score (b), with predictive accuracy (F-score) on 44 structure predictions. Both the PPfold (a) and pairs-only (b) reliability scores are significantly correlated with predictive accuracy (p-values of 2.256×10^{-4} and 2.232×10^{-6} , respectively), however, the pairs-only reliability score is substantially more strongly correlated (it has a higher R^2 value and therefore accounts for more of the variance).

A novel reliability measure of secondary structure prediction accuracy

In addition to improving the predictive accuracy of secondary structure prediction by incorporating alignment uncertainty, a novel statistic, called "the pairs-only reliability score," was devised for evaluating the predictive accuracy of a secondary structure model. This new statistic was necessary because it was found that PPfold's existing reliability score did not correlate particularly well with the F-score. The purpose of a reliability score is to give a user some guidance with respect to how much confidence they should place in a particular secondary structure prediction: i.e. the reliability measure acts as a predictor of the F-score, which can usually only be calculated if the true structure is known.

A comparative linear regression analysis of the old PPfold reliability score and the new pairs-only reliability score based on the 44 benchmarking datasets indicated that the new score did indeed correlate better with the F-score than did the old score (Figure 5.2). Specifically the R^2 value of 0.4021 for the correlation between the "pairs-only reliability score" and the F-score was substantially increased relative to the R^2 value of 0.2685 yielded by that of the PPfold reliability score vs. the F-score. This improvement therefore clearly advocates the use of the pairs-only reliability score over that of the old PPfold reliability score.

Conclusions

In conclusion, we found that incorporating alignment uncertainty into a secondary structure prediction approach by way of considering multiple appropriately sampled multiple sequence alignments it was possible to significantly improve the predictive accuracy of secondary structure prediction over approaches that consider only single sequence alignments at a time. When using the novel phylogenetic averaging approach to predict secondary structures by the PPfold method, an 8.2% improvement in the F-score (0.614 vs. 0.567) was achieved over that obtained using a single alignment produced by MAFFT (the best performing of the four multiple sequence alignment methods tested). This improved predictive accuracy was achievable by simply taking into account alignment uncertainty. It should be emphasised, however, that the phylogenetic averaging and base-pairing averaging approaches described here are slower than standard secondary structure prediction approaches which scale in only the length of the sequences (time complexity of $O(N^3)$). The new approaches scale in both the length of sequences and the number of sequences (time complexity of $O(N^m)$). Any major improvements in speed can only come from speeding up the statistical alignment portion of StatAlign.

It should also be noted that our benchmarking results are based on sets of non-coding RNA sequences. This is due to a lack of available experimentally-determined secondary structures of sequences containing coding regions. It would be interesting to see whether we achieve higher degrees of accuracy over existing approaches on full-length viral sequences using our novel approaches, although due to the $O(N^m)$ time complexity of our approaches we would be limited in either the length or the number of sequences used.

In addition to improving the accuracy of secondary structure prediction, an improved reliability score, called the pairs-only reliability score was devised. This score measures the degree of confidence that should be placed in the predictive power of a particular secondary structure prediction. This score was better correlated with the F-score measure of predictive accuracy than PPfold's existing reliability score, and it therefore provides better information on whether or not to trust a particular structure prediction. It should be noted that this improved reliability score is not only limited to use with the secondary structure prediction approach outlined in the chapter, but can be used with any nucleic acid secondary structure prediction approach that generates a base-pairing probability matrix and corresponding consensus secondary structure prediction.

6 Summary

In Chapter 2 of this thesis a graphical tool, called NAVA, was developed for visualising and analysing secondary structures within nucleic acid sequences. This tool integrates various existing pieces of software into a graphical user interface that allows a user to both analyse nucleic acid sequences for the presence of secondary structures with likely biological relevance, and visualise these secondary structures in the context of relevant additional data such as synonymous substitution rates and degrees of nucleotide conservation.

Using NAVA, the secondary structure of the SHAPE-inferred NL4-3 HIV-1 genome was visualised overlaid with SHAPE reactivities, as well as base-pairing probabilities corresponding to a computational secondary structure prediction of 50 HIV-1 subtype B sequences and nucleotide composition information corresponding to the same 50 sequences. This example demonstrated how diverse types of data can be visually overlaid on the same secondary structure providing researchers with useful biological insights into nucleic acid secondary structures. Furthermore, an automated ranking of a list of identified NL4-3 HIV-1 secondary structure elements was performed from the NAVA graphical user interface and it was found that the HIV-1 Rev response element had the lowest degrees of SHAPE reactivities amongst all the elements present in the NL4-3 HIV-1 genome, suggesting that this element was the most stable element in the HIV-1 genome, which was credible given its well-established biological functionality.

A short-coming of NAVA is that it currently only runs on the Windows operating system, despite being programmed in JAVA which supports multiple different operating systems. Therefore in the future we intend to add cross-platform support. Moreover, additional applications will be added to NAVA that will further facilitate secondary structure analyses. In particular it is anticipated that the most computationally intensive of these new tools, such as parametric maximum likelihood coevolution and substitution rate estimation methods will, rather than running directly on the user's computer, run on remote cloud computing resources accessible over the Internet. These cloud-based application will have the benefit of automatically supporting multiple operating systems, as well reducing the computational burden placed on the user's computer by taking advantage of the parallelization offered by remote computer clusters and graphical processing units.

In addition to NAVA, a tool for identifying regions of conserved nucleic acid secondary structure evident amongst an aligned set of sequences was developed and was used to automatically identify two well-characterized HIV-1 secondary structure elements, namely the gag-pol

frameshift element and Rev Response element. The successful identification of these two elements as being structurally conserved provided a positive validation of this tool.

In Chapter 3, the substructure conservation analysis, developed in the Chapter 2, revealed a number of highly conserved structural elements within HIV and HCV genomes, whilst there were very few such structures detectable within the DENV, WNV, JEV and TBEV genomes. Many of the conserved structures in HIV and HCV had been previously characterized and have known biological relevance, this provided a further validation of this tool and also suggested that similarly conserved but currently uncharacterized substructures that were identified in DENV, WNV JEV and TBEV might also have important biological functions.

A large-scale analysis revealed that the genomes of HIV, HCV, DENV and WNV species likely form secondary structures and there exists strong evidence that in these four virus species that natural selection likely favors the complementary coevolution of base-paired nucleotides within these structures. Even though evidence of selection favouring the maintenance of secondary structures was found in all four species, our detection of far fewer highly conserved structural elements in DENV than in HIV and HCV may indicate that individual functionally relevant genomic secondary structures play a far smaller role in DENV biology than they do in the biology of HCV and HIV. It was apparent from our large-scale analyses that the nature of this selection likely differs between DENV and these other species. Specifically, whereas in HIV, HCV, WNV, JEV and TBEV synonymous substitutions rates at base-paired sites are generally lower than at unpaired sites, in DENV, BVDV and CSFV coding regions this pattern is apparently completely attributable to non-synonymous substitutions being significantly lower at base-paired sites with synonymous substitution rates actually being significantly higher at base-paired sites than at unpaired sites. This suggested that in DENV, BVDV and CSFV protein evolution has priority to the point where one of the prime functions of base-pairing may be to reduce the basal rates of non-synonymous substitutions at base-paired sites.

Collectively our evolutionary analyses suggest that while numerous selectively maintained biologically functional structural elements occur within the genomes of HIV and various flavivirus species, many of the less conserved predicted structural elements within these genomes – and possibly most of those occurring in DENV, BVDV and CSFV genomes – have no uniquely advantageous molecular configuration that is specifically favored by natural selection. In such cases, selection may simply be acting either at the level of individual nucleotide pairs so as to reduce basal mutation rates at particular key codon sites, or on the level of individual codons so as to ensure that base-pairing does not reduce non-synonymous substitution rates. Alternatively, at the whole genome scale selection favoring the avoidance of intra-cellular

targeting by siRNAs may simply be acting to reduce the number of unpaired genomic sites. This type of large-scale non-conserved (but nevertheless evolutionarily favored) genomic secondary structure, called genome-scale ordered RNA (or GORS), has previously been hypothesized to facilitate the long-term persistence of ssRNA viruses (including many studied here) that cause chronic infections (Simmonds *et al.*, 2004; Davis *et al.*, 2008). Our results were entirely consistent with the coexistence within individual genomes of a continuum of structural elements ranging from those with specific biological functions that have evolutionarily conserved configurations, to those that simply form as a consequence of selection favoring either reduced substitution rates in specific genome regions or increased protection against anti-viral cellular defenses.

In Chapter 4, the relationship between genetic recombination and both RNA secondary structure and protein tertiary structure in natural HIV-1M recombinants was investigated. We found that in five out of eight HIV-1M proteins analysed that recombination events were less disruptive of intra-protein amino-acid—amino-acid interactions than could be accounted for by chance, which is the result we anticipated from viruses sampled from nature where natural selection is expected to remove recombinants expressing misfolded chimaeric proteins.

A similar analyses of computationally-inferred RNA secondary structure in recombinants yielded only marginal evidence that the maintenance of RNA secondary structure had any appreciable impact on patterns of recombination detectable within natural HIV-1M recombinants, which was somewhat surprising given both the large numbers of recombination events considered and the pervasiveness of biologically relevant secondary structures within HIV-1M genomes. It is plausible that recombinants are less disruptive of nucleic acid secondary structure, but we were unable to detect this signal due to a number of possible methodological shortcomings that were outlined in this chapter.

Despite the weak signal of natural recombinants avoiding the disruption of nucleotide-nucleotide interactions within biologically important nucleic acid secondary structures, we did detect a strong signal of recombination breakpoints tending to occur more in structured than in unstructured parts of HIV-1 genomes. This result is consistent with previous studies which have found that particularly structured regions of the HIV-1 genome tend to both co-localise with recombination breakpoint positions and be associated with gene and protein domain boundaries (Simon-Loriere *et al.*, 2010).

In the final chapter of this thesis, Chapter 5, two novel approaches for predicting secondary structures from a set of related sequences were developed. It was found that by incorporating alignment uncertainty into a secondary structure prediction approach by way of considering

multiple statistically sampled multiple sequence alignments it was possible to significantly improve the predictive accuracy of secondary structure prediction over approaches that consider only single sequence alignments at a time. The reason why these two new approaches were not used in this study to predict the secondary structures of the virus genomes analysed here was due to severe computational limitations brought about by the statistical alignment procedure used to statistically sample multiple sequences. In particular, the length of these virus genomes (8500 - 11000 nucleotides) made these approaches too computationally time-consuming to run. The reason for this is that standard single alignment secondary structure prediction approaches scale in only the length of the sequences (time complexity of $O(N^3)$), whereas the two new approaches scale in both the length of sequences and the number of sequences (time complexity of $O(N^m)$), this is directly a consequence of the statistical alignment step having a time complexity of $O(N^m)$.

Finally, in addition to improving the accuracy of comparative secondary structure prediction, an improved reliability score, called the pairs-only reliability score was devised. This score measures the degree of confidence that should be placed in the predictive accuracy of a particular secondary structure prediction. This score was better correlated with the F-score measure of predictive accuracy than PPfold's existing reliability score, and it therefore provides better information on whether or not to trust a particular structure prediction.

7 References

- Agniswamy,J. *et al.* (2012) Terminal interface conformations modulate dimer stability prior to amino terminal autoprocessing of HIV-1 protease. *Biochemistry*, **51**, 1041–50.
- Anderson,J.W.J. *et al.* (2013) Quantifying variances in comparative RNA secondary structure prediction. *BMC Bioinformatics*, **14**, 149.
- Archer,J. *et al.* (2008) Identifying the important HIV-1 recombination breakpoints. *PLoS Comput. Biol.*, **4**, e1000178.
- Arunapuram,P. *et al.* (2013) StatAlign 2.0: combining statistical alignment with RNA secondary structure prediction. *Bioinformatics*, **29**, 654–5.
- Beemon,K. *et al.* (1974) Evidence for Crossing-Over Between Avian Tumor Viruses Based on Analysis of Viral RNAs. *Proc. Natl. Acad. Sci.*, **71**, 4254–4258.
- Bernhart,S.H. *et al.* (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
- Bernstein,F.C. *et al.* (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.*, **80**, 319–324.
- Bieth,E. *et al.* (1990) A study of the dimer formation of Rous sarcoma virus RNA and of its effect on viral protein synthesis in vitro. *Nucleic Acids Res.*, **18**, 119–127.
- Billeter,M.A. *et al.* (1974) The Nucleotide Sequence Complexity of Avian Tumor Virus RNA. *Proc. Natl. Acad. Sci.*, **71**, 3560–3564.
- Bradley,R.K. *et al.* (2009) Fast statistical alignment. *PLoS Comput. Biol.*, **5**, e1000392.
- Branch,A.D. *et al.* (2005) The hepatitis C virus alternate reading frame (ARF) and its family of novel products: the alternate reading frame protein/F-protein, the double-frameshift protein, and others. *Semin. Liver Dis.*, **25**, 105–17.
- Brucoleri,R.E. and Heinrich,G. (1988) An improved algorithm for nucleic acid secondary structure display. *Bioinformatics*, **4**, 167–173.
- Budkowska,A. *et al.* (2011) Synonymous mutations in the core gene are linked to unusual serological profile in hepatitis C virus infection. *PLoS One*, **6**, e15871.
- Chen,C.J. *et al.* (1997) RNA-Protein Interactions: Involvement of NS3 , NS5 , and 3' Noncoding Regions of Japanese Encephalitis Virus Genomic RNA. **71**.
- Cheng,N. *et al.* (2012) Coevolution in RNA molecules driven by selective constraints: evidence from 5S rRNA. *PLoS One*, **7**, e44376.
- Chin,M.P.S. *et al.* (2005) Identification of a major restriction in HIV-1 intersubtype recombination. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 9002–7.
- Clever,J. *et al.* (1995) RNA secondary structure and binding sites for gag gene products in the 5'packaging signal of human immunodeficiency virus type 1. *J. Virol.*, **69**.

- Clyde,K. and Harris,E. (2006) RNA secondary structure in the coding region of dengue virus type 2 directs translation start codon selection and is required for viral replication. *J. Virol.*, **80**, 2170–82.
- Collier,A.J. *et al.* (2002) A conserved RNA structure within the HCV IRES eIF3-binding site. *Nat. Struct. Biol.*, **9**, 375–80.
- Contreras,A.M. *et al.* (2002) Viral RNA Mutations Are Region Specific and Increased by Ribavirin in a Full-Length Hepatitis C Virus Replication System. *J. Virol.*, **76**, 8505–8517.
- Darlix,J.L. *et al.* (1990) Cis elements and trans-acting factors involved in the RNA dimerization of the human immunodeficiency virus HIV-1. *J. Mol. Biol.*, **216**, 689–699.
- Darty,K. *et al.* (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–5.
- Daugherty,M.D. *et al.* (2010) Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nat. Struct. Mol. Biol.*, **17**, 1337–42.
- Davis,M. *et al.* (2008) Bioinformatic and physical characterizations of genome-scale ordered RNA structure in mammalian RNA viruses. *J. Virol.*, **82**, 11824–36.
- Delpont,W. *et al.* (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, **26**, 2455–7.
- Ding,Y. *et al.* (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.*, **32**, W135–41.
- Do,C.B. *et al.* (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–8.
- Dutheil,J. *et al.* (2005) A model-based approach for detecting coevolving positions in a molecule. *Mol. Biol. Evol.*, **22**, 1919–28.
- Eddy,S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.
- Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
- Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–7.
- Felsenstein,J. (1974) The evolutionary advantage of recombination. *Genetics*, **78**, 737–56.
- Flynn,J.A. *et al.* (2004) Nonrandom dimerization of murine leukemia virus genomic RNAs. *J. Virol.*, **78**, 12129–39.
- Flynn,J.A. and Telesnitsky,A. (2006) Two distinct Moloney murine leukemia virus RNAs produced from a single locus dimerize at random. *Virology*, **344**, 391–400.
- Galetto,R. *et al.* (2004) The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot in vivo. *J. Biol. Chem.*, **279**, 36625–32.
- Galetto,R. and Negroni,M. (2005) Mechanistic features of recombination in HIV. *AIDS Rev.*, **7**, 92–102.

- Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
- Golden,M. and Martin,D. (2013) DOOSS: a tool for visual analysis of data overlaid on secondary structures. *Bioinformatics*, **29**, 271–2.
- Golubchik,T. *et al.* (2007) Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol. Biol. Evol.*, **24**, 2433–42.
- Griffiths-Jones,S. *et al.* (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–4.
- Gutell,R.R. *et al.* (2002) The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.*, **12**, 301–310.
- Haasnoot,J. *et al.* (2007) RNA interference against viruses: strike and counterstrike. *Nat. Biotechnol.*, **25**, 1435–43.
- Hancock,J. *et al.* (1988) Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster*. *Mol. Biol. Evol.*, **5**, 393–414.
- Heaphy,S. *et al.* (1990) HIV-1 regulator of virion expression (Rev) protein binds to an RNA stem-loop structure located within the Rev response element region. *Cell*, **60**, 685–93.
- Hein,J. *et al.* (2000) Statistical alignment: computational properties, homology testing and goodness-of-fit. *J. Mol. Biol.*, **302**, 265–79.
- Hofacker,I. *et al.* (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–66.
- Hofacker,I.L. *et al.* (1998) Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.*, **26**, 3825–3836.
- Honda,M. *et al.* (1996) Stability of a stem-loop involving the initiator AUG controls the efficiency of internal initiation of translation on hepatitis C virus RNA. *RNA*, **2**, 955–68.
- Hung,M. *et al.* (1998) Importance of ribosomal frameshifting for human immunodeficiency virus type 1 particle assembly and replication. *J. Virol.*, **72**, 4819–24.
- Isel,C. *et al.* (1995) Initiation of Reverse Transcription of HIV-1: Secondary Structure of the HIV-1 RNA/tRNA^lmbopopnbop|Lys|clobop|3 (Template/Primer) Complex. *J. Mol. Biol.*, **247**, 236–250.
- Ito,T. and Lai,M.M. (1999) An internal polypyrimidine-tract-binding protein-binding site in the hepatitis C virus RNA attenuates translation, which is relieved by the 3'-untranslated sequence. *Virology*, **254**, 288–96.
- Jetzt,A.E. *et al.* (2000) High Rate of Recombination throughout the Human Immunodeficiency Virus Type 1 Genome. *J. Virol.*, **74**, 1234–1240.
- Ji,J. *et al.* (1994) Mutagenicity and pausing of HIV reverse transcriptase during HIV plus-strand DNA synthesis. *Nucleic Acids Res.*, **22**, 47–52.
- Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.*, **9**, 286–98.

- Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.
- Knudsen,B. and Hein,J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.
- Kosakovsky Pond,S.L. *et al.* (2006) GARD: a genetic algorithm for recombination detection. *Bioinformatics*, **22**, 3096–8.
- Lai,M.M. (1992) RNA recombination in animal and plant viruses. *Microbiol. Rev.*, **56**, 61–79.
- Lari,K. and Young,S.J. (1990) The estimation of stochastic context-free grammars using the Inside-Outside algorithm. *Comput. Speech Lang.*, **4**, 35–56.
- Lefeuivre,P. *et al.* (2007) Avoidance of protein fold disruption in natural virus recombinants. *PLoS Pathog.*, **3**, e181.
- Lefeuivre,P. *et al.* (2009) Widely conserved recombination patterns among single-stranded DNA viruses. *J. Virol.*, **83**, 2697–707.
- Lindahl,T. and Karlstrom,O. (1973) Heat-induced depyrimidination of deoxyribonucleic acid in neutral solution. *Biochemistry*, **12**, 5151–5154.
- Lindahl,T. and Nyberg,B. (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry*, **13**, 3405–3410.
- Lindenbach,B.D. and Rice,C.M. (2003) Molecular biology of flaviviruses. *Adv. Virus Res.*, **59**, 23–61.
- Loeb,L. and Preston (1986) Mutagenesis by apurinic/apyrimidinic sites. *Annu. Rev. Genet.*, **20**, 201–30.
- Lukavsky,P.J. *et al.* (2000) Structures of two RNA domains essential for hepatitis C virus internal ribosome entry site function. *Nat. Struct. Biol.*, **7**, 1105–10.
- Lunter,G. *et al.* (2008) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, **18**, 298–309.
- Mandal,M. and Breaker,R.R. (2004) Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.*, **5**, 451–63.
- Markham,N.R. and Zuker,M. (2008) UNAFold: Software for Nucleic Acid Folding and Hybridization. In, Keith,J.M. (ed), *Bioinformatics : Structure, Function and Applications*. Humana Press, Totowa, NJ, pp. 3–31.
- Martin,D. *et al.* (2011) Analysing recombination in nucleotide sequences. *Mol. Ecol. Resour.*, **11**, 943–55.
- Martin,D. *et al.* (2010) RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics*, **26**, 2462–3.
- Martin,D. *et al.* (2005) The evolutionary value of recombination is constrained by genome modularity. *PLoS Genet.*, **1**, e51.
- Martin,D.P. *et al.* (2011) Complex recombination patterns arising during geminivirus coinfections preserve and demarcate biologically important intra-genome interaction networks. *PLoS Pathog.*, **7**, e1002203.
- Martin,G. *et al.* (2006) Selection for recombination in structured populations. *Genetics*, **172**, 593–609.

- Mason, P.W. *et al.* (2002) Identification and Characterization of a cis-Acting Replication Element (cre) Adjacent to the Internal Ribosome Entry Site of Foot-and-Mouth Disease Virus. *J. Virol.*, **76**, 9686–9694.
- Mathews, D.H. (2006) Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, **359**, 526–32.
- Mathews, D.H. and Turner, D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.
- Mathews, D.H. and Turner, D.H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**, 270–8.
- Mitsui, Y. *et al.* (1970) Physical and enzymatic studies on poly d(I-C)-poly d(I-C), an unusual double-helical DNA. *Nature*, **228**, 1166–9.
- Mokrejs, M. *et al.* (2010) IRESite--a tool for the examination of viral and cellular internal ribosome entry sites. *Nucleic Acids Res.*, **38**, D131–6.
- Moore, M.D. and Hu, W.S. (2009) HIV-1 RNA dimerization: It takes two to tango. *AIDS Rev.*, **11**, 91–102.
- Moulton, V. *et al.* (2000) Metrics on RNA secondary structures. *J. Comput. Biol.*, **7**, 277–92.
- Moulton, Vincent *et al.* (2000) Metrics on RNA secondary structures. *J. Comput. Biol.*, **7**, 277–92.
- Moumen, A. *et al.* (2003) Evidence for a mechanism of recombination during reverse transcription dependent on the structure of the acceptor RNA. *J. Biol. Chem.*, **278**, 15973–82.
- Muller, H.J. (1964) The relation of recombination to mutational advance. *Mutat. Res. Mol. Mech. Mutagen.*, **1**, 2–9.
- Murrell, B. *et al.* (2013) FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.*, **30**, 1196–205.
- Muse, S. and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, **11**, 715–24.
- Ngandu, N.K. *et al.* (2008) Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. *Virol. J.*, **5**, 160.
- Nielsen, M.H. *et al.* (2005) Molecular strategies to inhibit HIV-1 replication. *Retrovirology*, **2**, 10.
- Novák, A. *et al.* (2008) StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*, **24**, 2403–4.
- Onafuwa, A. *et al.* (2003) Human Immunodeficiency Virus Type 1 Genetic Recombination Is More Frequent Than That of Moloney Murine Leukemia Virus despite Similar Template Switching Rates. *J. Virol.*, **77**, 4577–4587.
- Pancera, M. *et al.* (2010) Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 1166–71.
- Parkin, N., Chamorro, M. *et al.* (1992) Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: demonstration by expression in vivo. *J. Virol.*, **66**, 5147–5151.

- Parkin,N., Chamorro,Mario, *et al.* (1992) Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: demonstration by expression in vivo. *J. Virol.*, **69**, 2101–2109.
- Parthasarathi,S. *et al.* (1995) Genetic rearrangements occurring during a single cycle of murine leukemia virus vector replication: characterization and implications. *J. Virol.*, **69**, 7991–8000.
- Paul,A. V. *et al.* (2000) Identification of an RNA Hairpin in Poliovirus RNA That Serves as the Primary Template in the In Vitro Uridylation of VPg. *J. Virol.*, **74**, 10359–10370.
- Pedersen,J.S. *et al.* (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
- Plotkin,J.B. and Kudla,G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*, **12**, 32–42.
- Pond,S.L.K. *et al.* (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, **21**, 676–9.
- Pornillos,O. *et al.* (2009) X-ray structures of the hexameric building block of the HIV capsid. *Cell*, **137**, 1282–92.
- Price,M.N. *et al.* (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
- Proctor,J.R. and Meyer,I.M. (2013) COFOLD: an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Res.*, **41**, e102.
- Raman,S. and Brian,D.A. (2005) Stem-loop IV in the 5' untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. *J. Virol.*, **79**, 12434–46.
- Rambaut,A. (2012) FigTree: a graphical viewer of phylogenetic trees.
- Reeves,J.D. and Doms,R.W. (2002) Human immunodeficiency virus type 2. *J. Gen. Virol.*, **83**, 1253–1265.
- Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
- Rivas,E. and Eddy,S. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–68.
- Rivas,E. and Eddy,S. (2000) The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics*, **16**, 334–340.
- Rizk,G. and Lavenier,D. (2009) GPU accelerated RNA folding algorithm. *Comput. Sci. 2009*, 1–13.
- Roda,R.H. *et al.* (2002) Strand transfer occurs in retroviruses by a pause-initiated two-step mechanism. *J. Biol. Chem.*, **277**, 46900–46911.
- Roy,S. *et al.* (1990) A bulge structure in HIV-1 TAR RNA is required for Tat binding and Tat-mediated trans-activation. *Genes Dev.*, **4**, 1365–1373.
- Sanjuán,R. and Bordería,A. V (2011) Interplay between RNA structure and protein evolution in HIV-1. *Mol. Biol. Evol.*, **28**, 1333–8.

- Scheffler, K. *et al.* (2006) Robust inference of positive selection from recombining coding sequences. *Bioinformatics*, **22**, 2493–9.
- Semegni, J.Y. *et al.* (2011) NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. *Bioinformatics*, **27**, 2443–5.
- Sessions, O.M. *et al.* (2009) Discovery of insect and human dengue virus host factors. *Nature*, **458**, 1047–50.
- Siebert, S. and Backofen, R. (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, **21**, 3352–9.
- Sievers, F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
- Simmonds, P. *et al.* (2008) Bioinformatic and functional analysis of RNA secondary structure elements among different genera of human and animal caliciviruses. *Nucleic Acids Res.*, **36**, 2530–46.
- Simmonds, P. *et al.* (2004) Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *RNA*, **10**, 1337–51.
- Simon-Loriere, E. *et al.* (2010) RNA structures facilitate recombination-mediated gene swapping in HIV-1. *J. Virol.*, **84**, 12675–12682.
- Simon-Loriere, E. and Holmes, E.C. (2011) Why do RNA viruses recombine? *Nat. Rev. Microbiol.*, **9**, 617–26.
- Song, R. *et al.* (2007) HIV-1 viral RNA is selected in the form of monomers that dimerize in a three-step protease-dependent process; the DIS of stem-loop 1 initiates viral RNA dimerization. *J. Mol. Biol.*, **371**, 1084–98.
- Spearman, C. (1904) The proof and measurement of association between two things. *Am. J. Psychol.*, **15**, 72–101.
- Staple, D.W. and Butcher, S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol.*, **3**, e213.
- Su, H.-P. *et al.* (2010) Structural basis for the inhibition of RNase H activity of HIV-1 reverse transcriptase by RNase H active site-directed inhibitors. *J. Virol.*, **84**, 7625–33.
- Sükkösd, Z. *et al.* (2011) Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars. *BMC Bioinformatics*, **12**, 103.
- Tafer, H. *et al.* (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.
- Tan, K. *et al.* (1997) Atomic structure of a thermostable subdomain of HIV-1 gp41. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 12303–8.
- The Jmol Team (2007) Jmol: an open-source Java viewer for chemical structures in 3D. *jmol.sourceforge.net*.
- Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–80.

- Thorne, J.L. *et al.* (1992) Inching toward reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.*, **34**, 3–16.
- Torarinsson, E. *et al.* (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–32.
- Tsukiyama-Kohara, K. *et al.* (1992) Internal ribosome entry site within hepatitis C virus RNA. *J. Virol.*, **66**, 1476–83.
- Tu, X. *et al.* (2010) Structural basis of HIV-1 resistance to AZT by excision. *Nat. Struct. Mol. Biol.*, **17**, 1202–9.
- Tuplin, A. *et al.* (2004) Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. *J. Gen. Virol.*, **85**, 3037–47.
- Tuplin, A. *et al.* (2002) Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. *RNA*, 824–841.
- Vassilaki, N. *et al.* (2008) Role of the hepatitis C virus core+1 open reading frame and core cis-acting RNA elements in viral RNA translation and replication. *J. Virol.*, **82**, 11503–15.
- Voigt, C.A. *et al.* (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.*, **9**, 553–8.
- Van der Walt, E. *et al.* (2008) Viable chimaeric viruses confirm the biological importance of sequence specific maize streak virus movement protein and coat protein interactions. *Virol. J.*, **5**, 61.
- Wang, J. (1979) Helical Repeat of DNA in Solution. *Proc. Natl. Acad. Sci.*, **76**, 200–203.
- Wang, J. *et al.* (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, **282**, 680–686.
- Wang, J.Y. *et al.* (2001) Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *EMBO J.*, **20**, 7333–43.
- Watson, J. and Crick, F. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, **171**, 737–738.
- Watts, J.M. *et al.* (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–6.
- Westerhout, E.M. *et al.* (2005) HIV-1 can escape from RNA interference by evolving an alternative structure in its RNA genome. *Nucleic Acids Res.*, **33**, 796–804.
- Wilkinson, K. a *et al.* (2008) High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol.*, **6**, e96.
- Wilkinson, K. a *et al.* (2006) Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.*, **1**, 1610–6.
- Xayaphoummine, A. *et al.* (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.*, **33**, W605–10.
- Yoshida, K. *et al.* (1997) Mutations of the HIV Type 1 V3 Loop under Selection Pressure with Neutralizing Monoclonal Antibody NM-01. *AIDS Res. Hum. Retroviruses*, **13**, 1283–1290.

- You, S. *et al.* (2004) A cis-acting replication element in the sequence encoding the NS5B RNA-dependent RNA polymerase is required for hepatitis C virus RNA replication. *J. Virol.*, **78**, 1352–66.
- Zhang, J. *et al.* (2000) Most Retroviral Recombinations Occur during Minus-Strand DNA Synthesis. *J. Virol.*, **74**, 2313–2322.
- Zhuang, J. *et al.* (2006) High rate of genetic recombination in murine leukemia virus: implications for influencing proviral ploidy. *J. Virol.*, **80**, 6706–11.

University of Cape Town

8 Supplementary material

Supplementary Table 1: Top 20 out of 422 HIV-1 SHAPE structures with lower than expected SHAPE reactivities (ranked by ascending z-scores). The top ranking structure (Structure ID 71) corresponds to the Rev response element. Where a substructure corresponds to a substructure within another substructure in the table it is marked with a *.

Rank	Structure ID	Location	Structure Length	Structure Median	Genome Median	z-score	p-value	Region / Structure
1	71	7276-7575	299	0.1	0.33	-8.71	1.49×10^{-15}	RRE (Heaphy <i>et al.</i> , 1990)
2	193	9074-9135	61	0.01	0.33	-7.58	2.98×10^{-13}	3' UTR
3	177	7310-7534	224	0.11	0.33	-6.46	5.71×10^{-11}	RRE*
4	283	9083-9127	44	0.02	0.33	-5.75	4.56×10^{-9}	3' UTR
5	86	9011-9140	129	0.11	0.33	-5.49	1.96×10^{-8}	3' UTR
6	214	1645-1673	28	0.045	0.33	-5.07	2.00×10^{-7}	Gag-pol frameshifting region (Parkin, Mario Chamorro, <i>et al.</i> , 1992)
7	277	8723-8752	29	0.06	0.33	-4.85	6.10×10^{-7}	
8	2	106-344	238	0.19	0.33	-4.75	1.03×10^{-6}	
9	348	248-271	23	0.07	0.33	-4.70	1.28×10^{-6}	Ψ locus "kissing loop" (Clever <i>et al.</i> , 1995)
10	85	8697-8997	300	0.23	0.33	-4.70	1.29×10^{-6}	
11	189	8700-8804	104	0.185	0.33	-4.62	1.96×10^{-6}	
12	287	243-278	35	0.09	0.33	-4.52	3.11×10^{-6}	
13	1	59-104	45	0.13	0.33	-4.44	4.54×10^{-6}	
14	0	1-58	57	0.075	0.33	-4.36	6.62×10^{-6}	TAR (Roy <i>et al.</i> , 1990)
15	196	134-179	45	0.12	0.33	-4.33	7.61×10^{-6}	Primer-binding site*
16	15	1729-1758	29	0.13	0.33	-4.31	8.26×10^{-6}	
17	34	3336-3359	23	0.09	0.33	-4.16	1.62×10^{-5}	
18	341	8727-8748	21	0.05	0.33	-4.15	1.69×10^{-5}	
19	89	125-224	99	0.14	0.33	-4.14	1.74×10^{-5}	Primer-binding site (Isel <i>et al.</i> , 1995)
20	267	7337-7504	167	0.13	0.33	-4.12	1.87×10^{-5}	RRE*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Data overlay			T	-	-				T	T	T								
Structure								T						T	-	T			G

1. Unaligned sequences from the data overlay mapping alignment and the structure overlay mapping alignment, respectively. Note that data overlay mapping alignment has a different length than that of the structure overlay mapping alignment.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Data overlay			T	N	N				T	T	T								
Structure				G	G			T						T	N	T			

2. Unaligned sequences with gaps replaced by N's. Replacing gapped positions with N's allows sequence alignment to proceed without losing information about gap positions in the original mapping alignments.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Data overlay	-	-	-	-	-			T	N					T	-	T	T		
Structure								T						T	N	T			

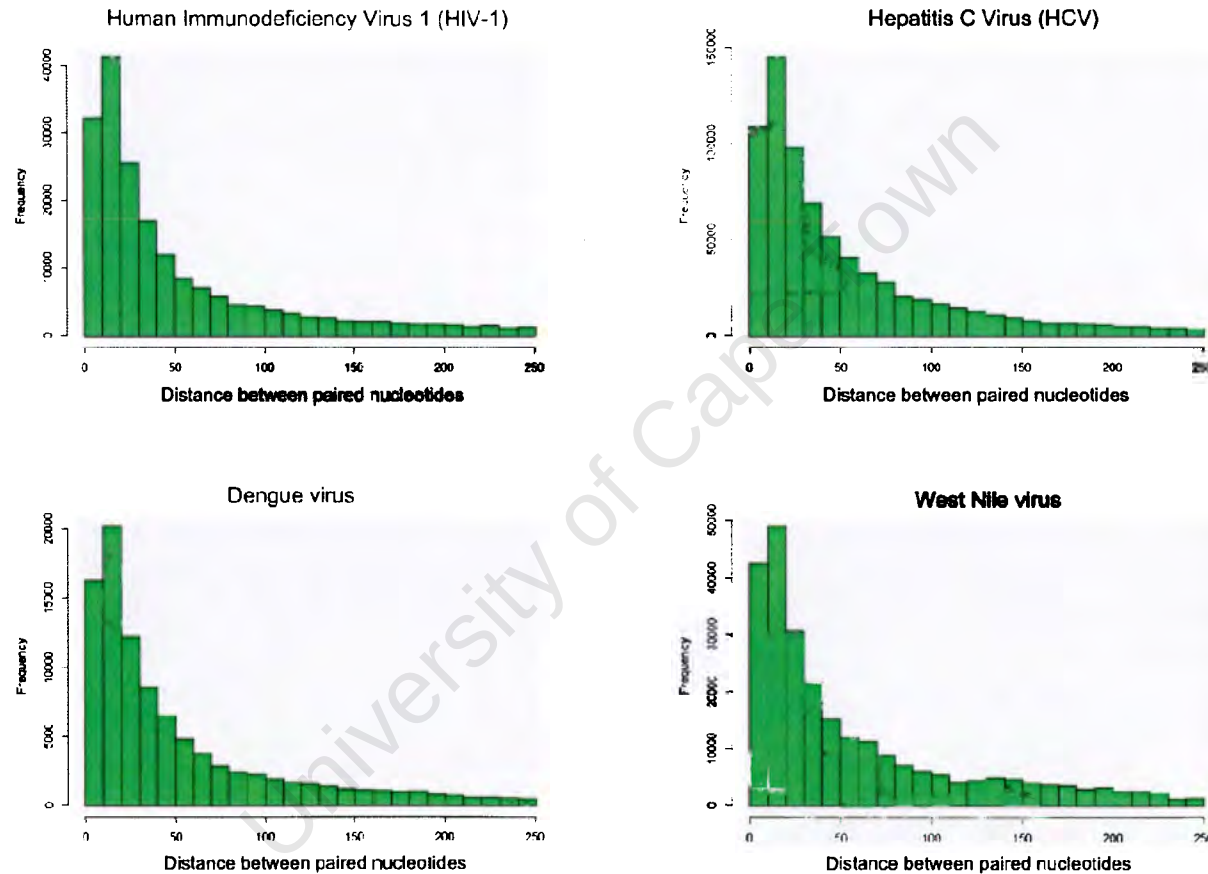
3. Sequences after alignment. Aligning the sequences matches up homologous regions, which defines the mapping between the data overlay and the structure overlay.

Data overlay	1	2	3	4	5	6	7	8	9	10	11	12	13						
Structure	6	7	8	9	10	11	12	13	14	16	17	18	19						

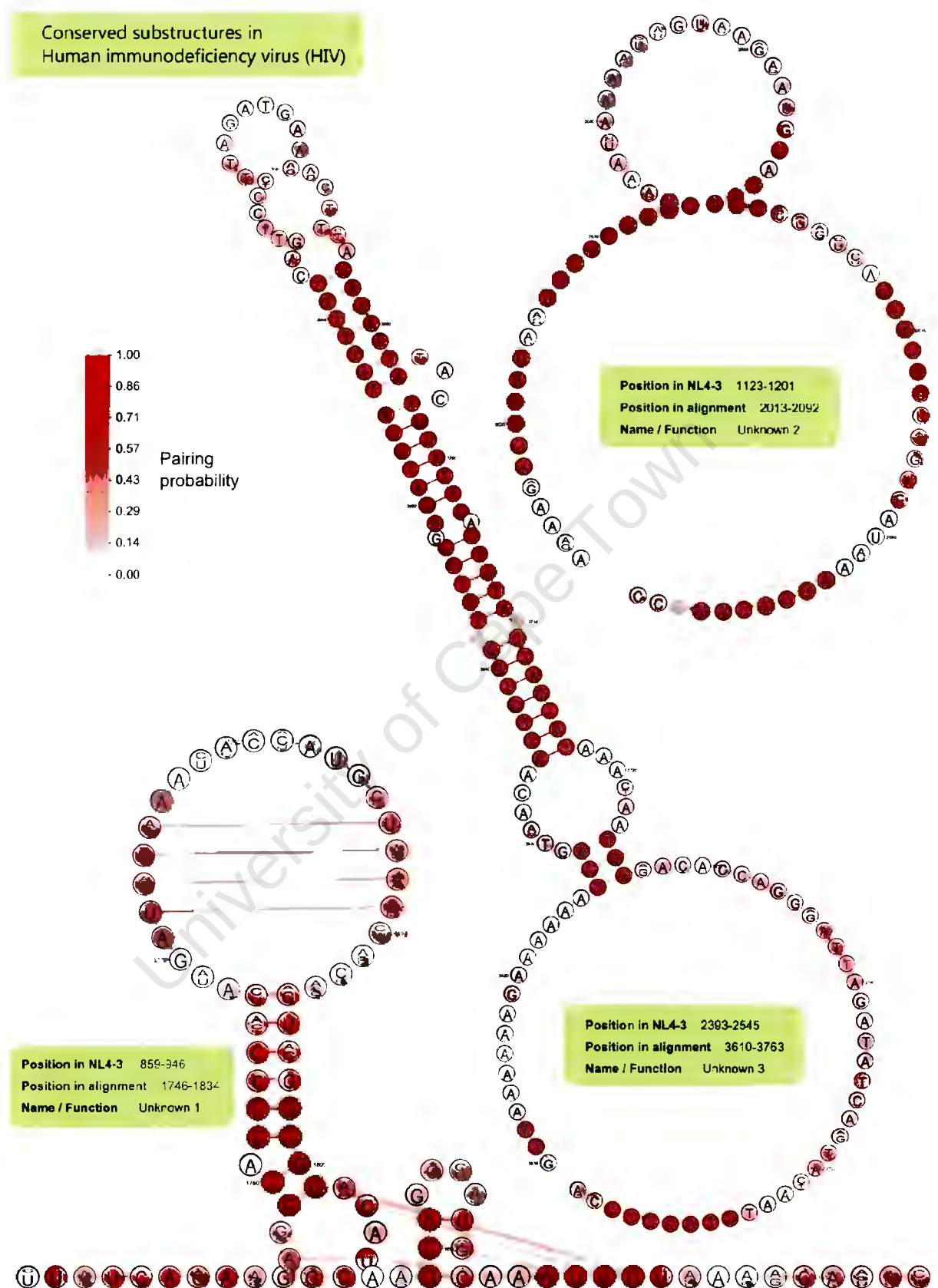
Mapping lookup table. The first row specifies which positions in the data overlay map to which positions in the structure overlay (in the second row).

Supplementary Figure 1: Example of how a data overlay mapping alignment and a structure overlay mapping alignment are used to create a mapping between a data overlay and a structure overlay.

Distributions of distances between base-paired nucleotides in different viruses

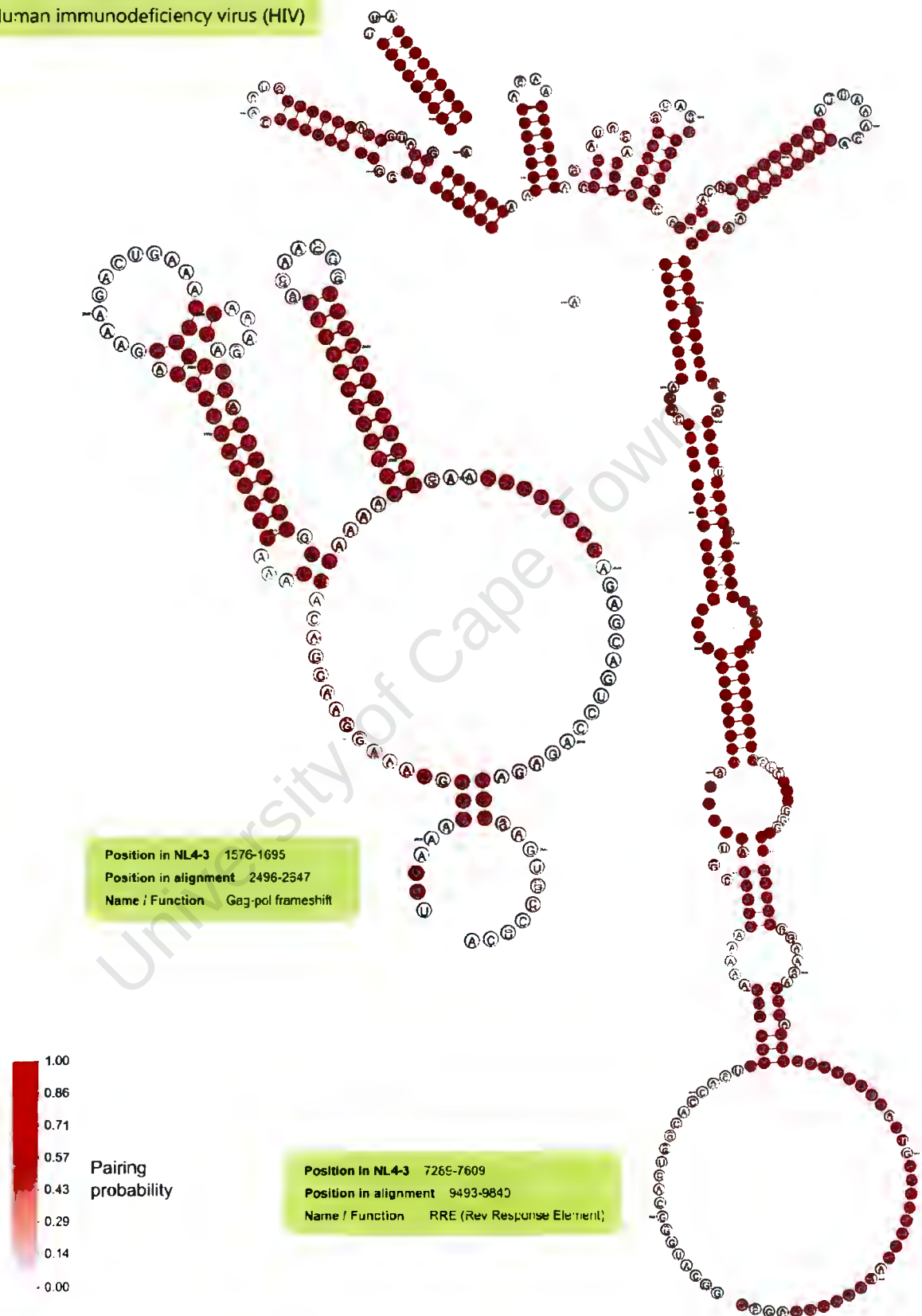


Supplementary Figure 2: Distributions of distances between base-paired nucleotides in different viruses. Each of these histograms shows a similar distribution of distances between base-paired nucleotides in different computationally inferred virus secondary structures. It is evident that nucleotide positions further apart in the sequence are less frequently base-paired with one another. There is an exception for those nucleotides that are very close one another in the sequence (<10 nucleotides), this is presumably because nucleotides which are very close to one another (<5 nucleotides) will have to form a very tight loop in order to base-pair which is energetically unfavourable.



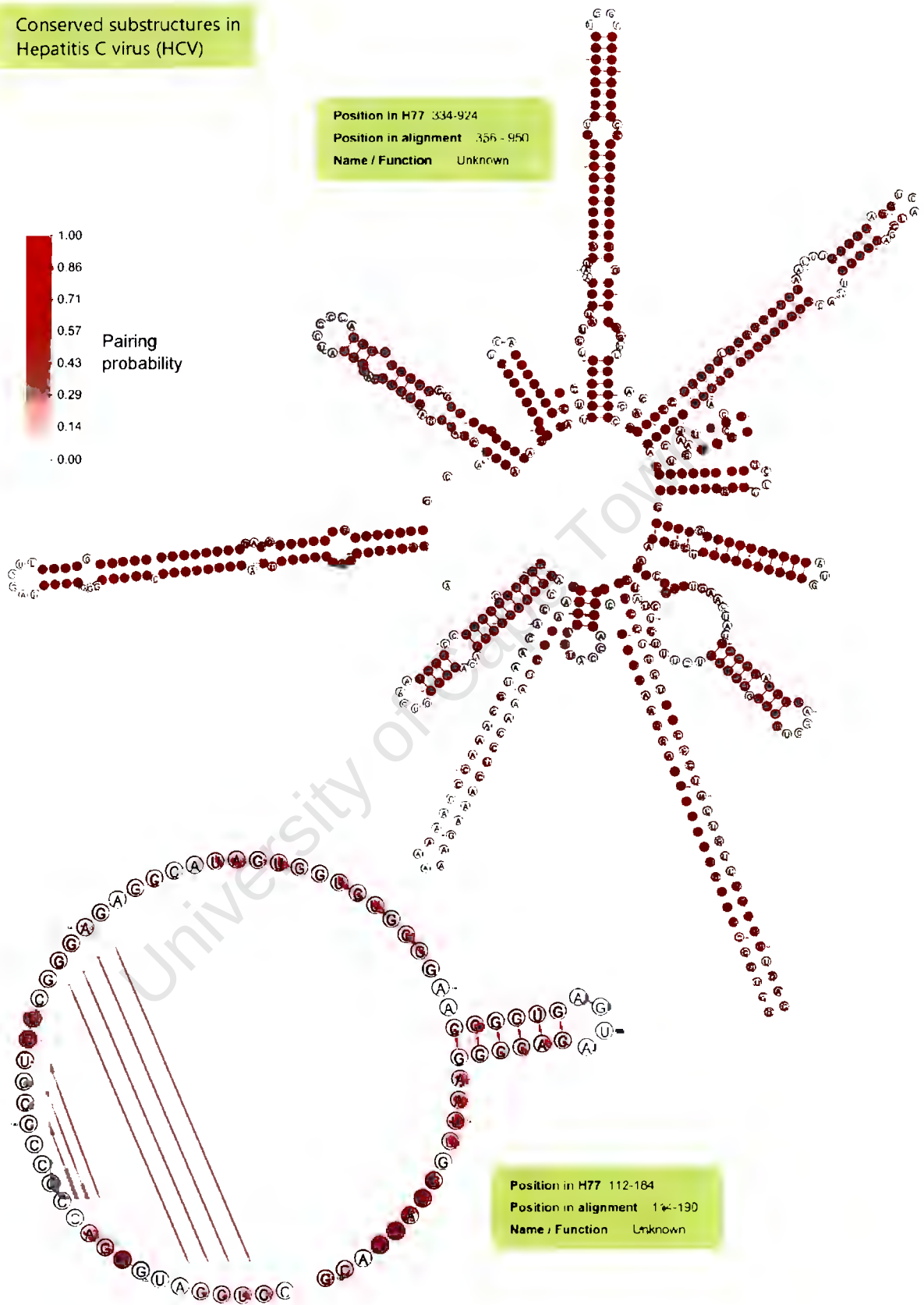
Supplementary Figure 3: Substructure visualisations for conserved substructures in Human Immunodeficiency Virus (HIV) page 1. "Pairing probability" refers to the proportion of positions (white-red nucleotides) or pairs of positions (white-red lines) that were paired in the alignment of sequences.

Conserved substructures in
Human immunodeficiency virus (HIV)



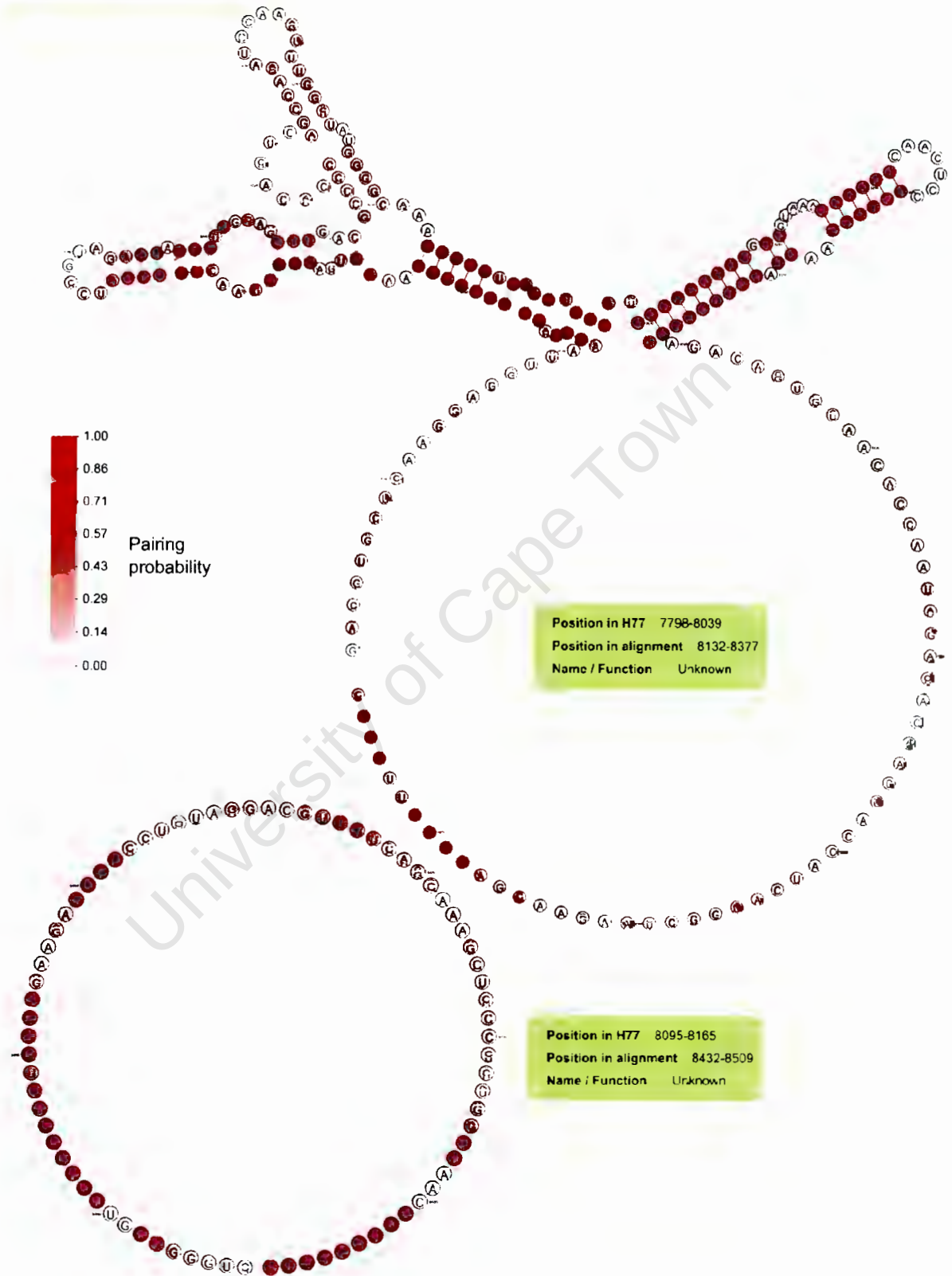
Supplementary Figure 4: Substructure visualisations for conserved substructures in Human Immunodeficiency Virus (HIV) page 2. "Pairing probability" refers to the proportion of positions (white-red nucleotides) or pairs of positions (white-red lines) that were paired in the alignment of sequences.

Conserved substructures in
Hepatitis C virus (HCV)



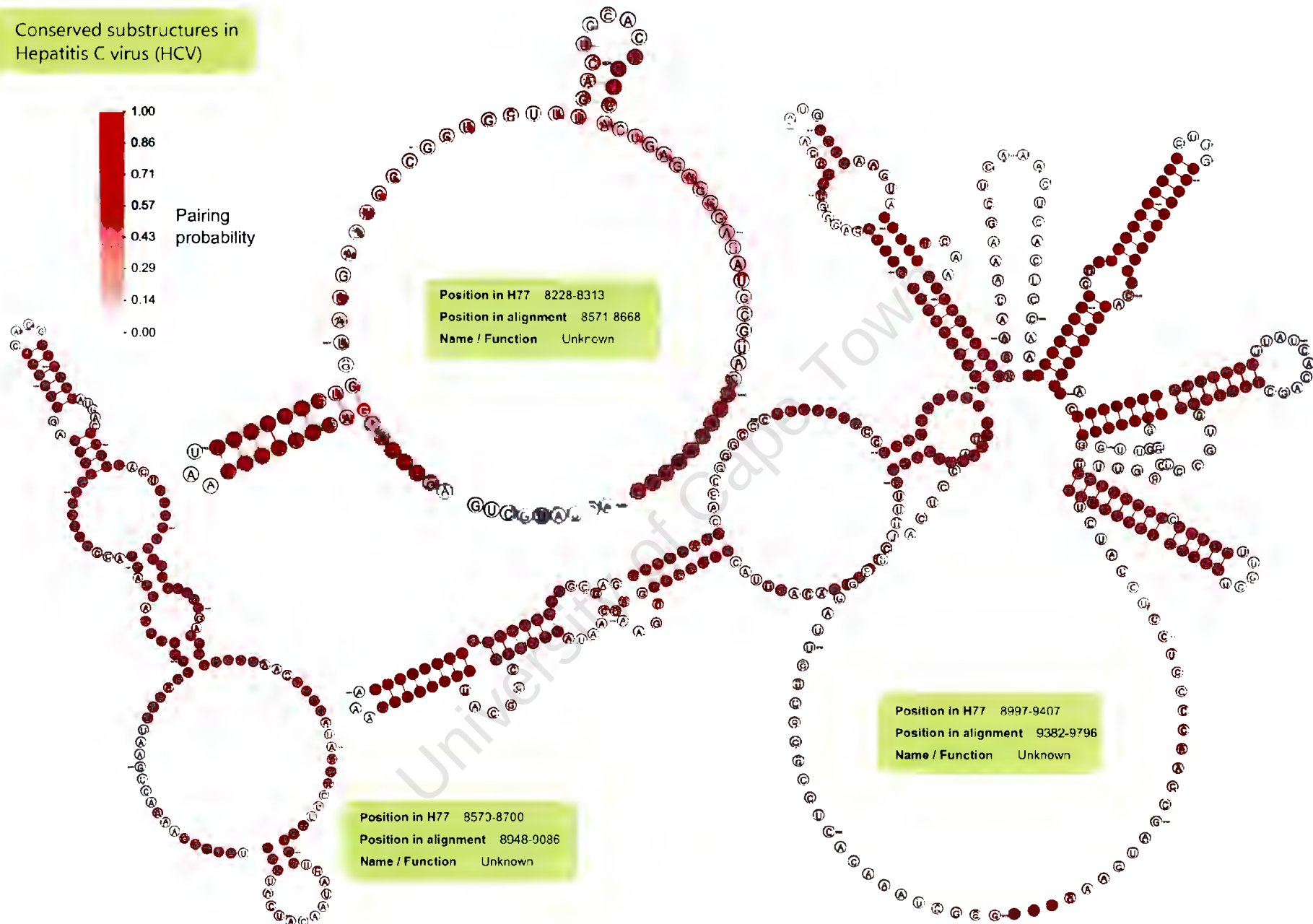
Supplementary Figure 5: Substructure visualisations for conserved substructures in Human Immunodeficiency Virus (HIV) page 3. "Pairing probability" refers to the proportion of positions (white-red nucleotides) or pairs of positions (white-red lines) that were paired in the alignment of sequences.

Conserved substructures in
Hepatitis C virus (HCV)



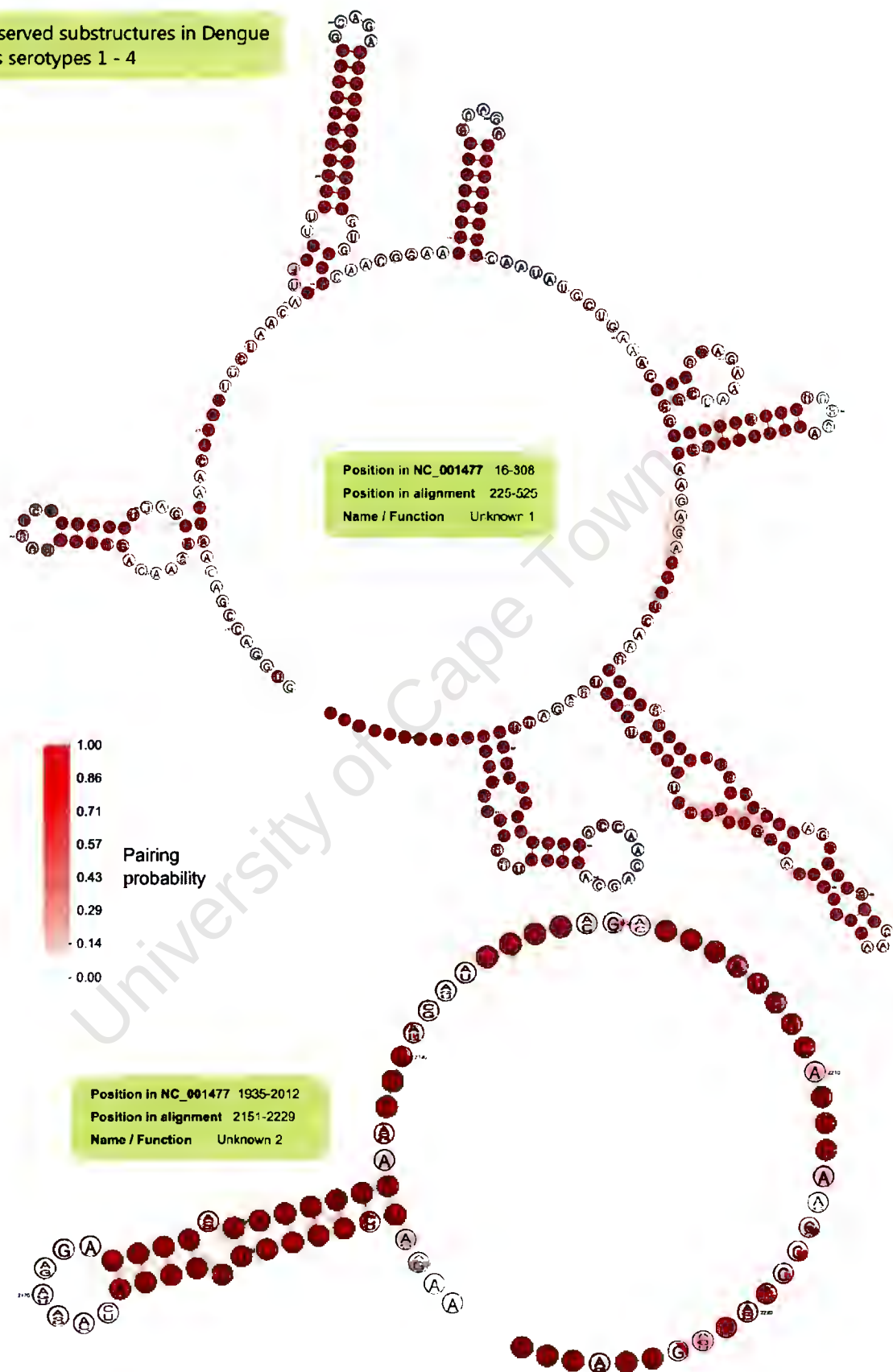
Supplementary Figure 6: Substructure visualisations for conserved substructures Hepatitis C Virus (HCV)
page 1. "Pairing probability" refers to the proportion of positions (white-red nucleotides) or pairs of positions (white-red lines) that were paired in the alignment of sequences.

Conserved substructures in Hepatitis C virus (HCV)



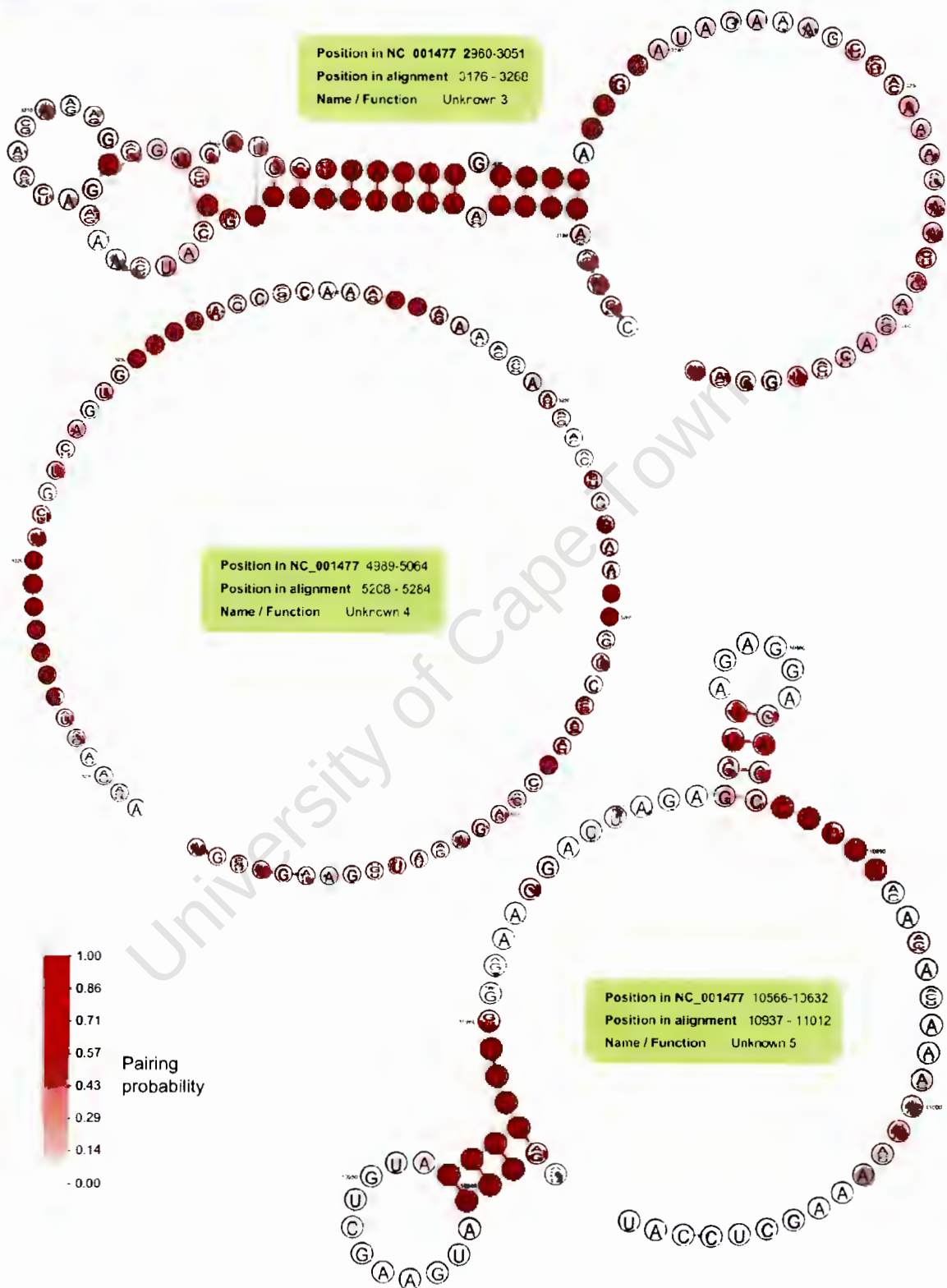
Supplementary Figure 7: Substructure visualisations for conserved substructures Hepatitis C Virus (HCV) page 2. "Pairing probability" refers to the proportion of positions (white-red nucleotides) or pairs of positions (white-red lines) that were paired in the alignment of sequences.

Conserved substructures in Dengue virus serotypes 1 - 4



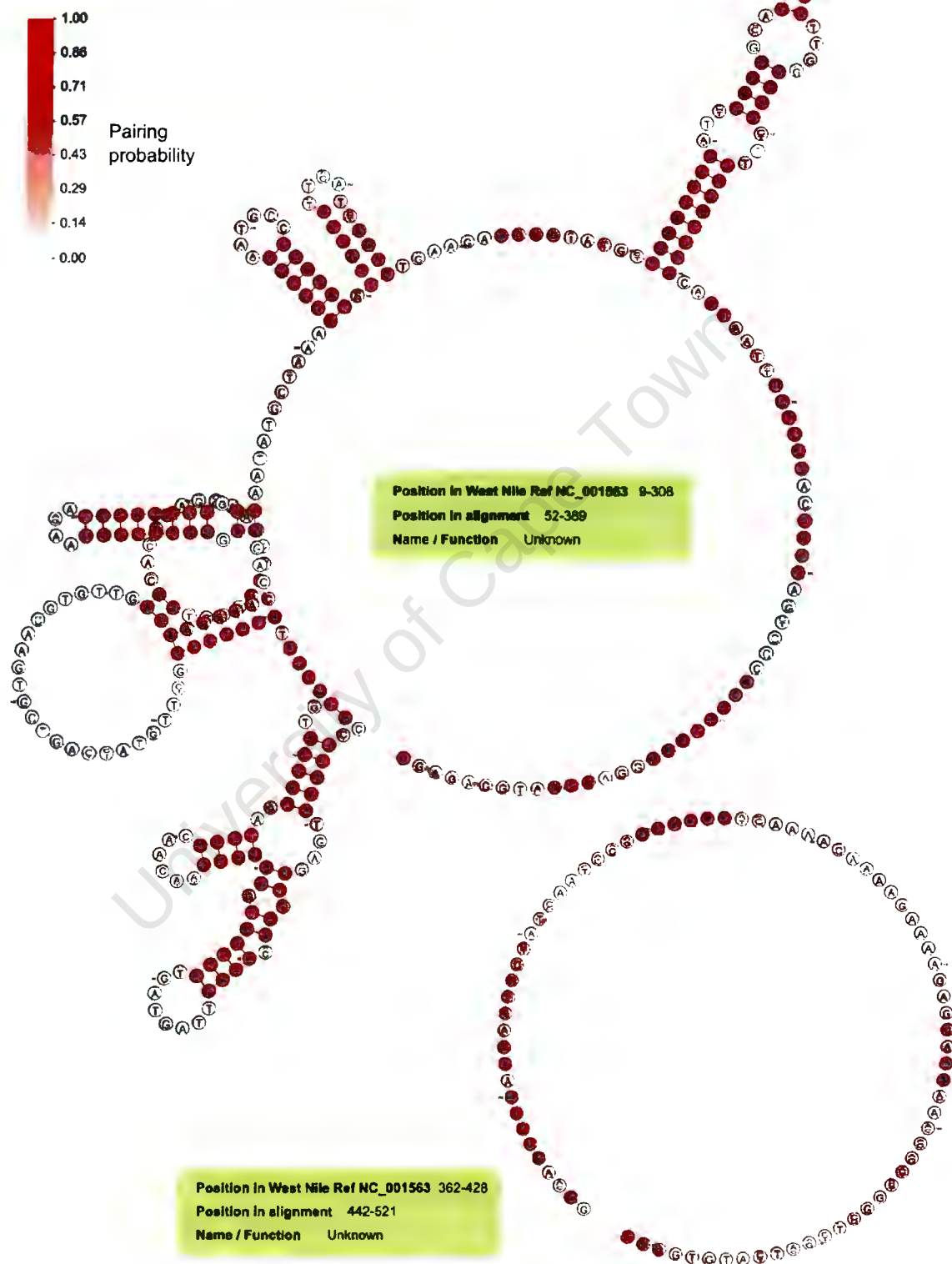
Supplementary Figure 8: Substructure visualisations for conserved substructures in Dengue virus page 1. "Pairing probability" refers to the proportion of positions (white-red nucleotides) or pairs of positions (white-red lines) that were paired in the alignment of sequences.

Conserved substructures in Dengue virus serotypes 1 - 4



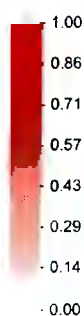
Supplementary Figure 9: Substructure visualisations for conserved substructures in Dengue virus page 2. "Pairing probability" refers to the proportion of positions (white-red nucleotides) or pairs of positions (white-red lines) that were paired in the alignment of sequences.

Conserved substructures in West Nile virus,
Japanese encephalitis virus (JEV)
and Tick-borne encephalitis virus (TBEV)



Supplementary Figure 10: Substructure visualisations for conserved substructures in West Nile Virus, Japanese Encephalitis Virus (JEV) and Tick-borne Encephalitis Virus (TBEV) page 1.

Conserved substructures in West Nile virus,
Japanese encephalitis virus (JEV)
and Tick-borne encephalitis virus (TBEV)

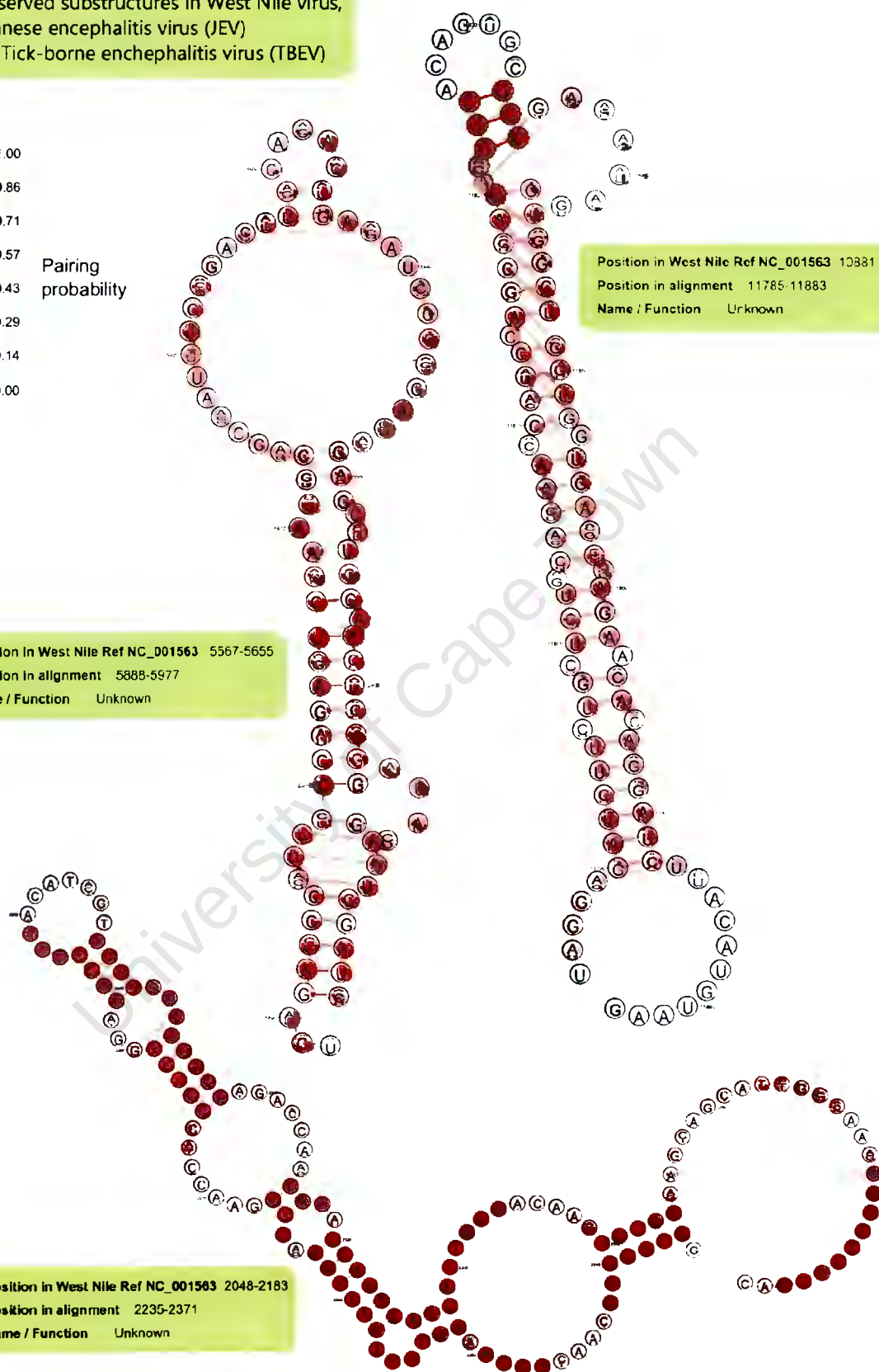


Pairing
probability

Position in West Nile Ref NC_001563 5567-5655
Position in alignment 5888-5977
Name / Function Unknown

Position in West Nile Ref NC_001563 10381-10961
Position in alignment 11785-11883
Name / Function Unknown

Position in West Nile Ref NC_001563 2048-2183
Position in alignment 2235-2371
Name / Function Unknown



Supplementary Figure 11: Substructure visualisations for conserved substructures in West Nile Virus, Japanese Encephalitis Virus (JEV) and Tick-borne Encephalitis Virus (TBEV) page 2.

Supplementary Table 2: List of PDB structure files used in the recombination protein-fold disruption test.

Protein name	PDB ID	Citation
Capsid P24	3H4E	(Pornillos <i>et al.</i> , 2009)
gp41	1SZT	(Tan <i>et al.</i> , 1997)
gp120	3JWD	(Pancera <i>et al.</i> , 2010)
Integrase	1K6Y	(Wang <i>et al.</i> , 2001)
Nef	4D8D	Unpublished
Protease	3TKG	(Agniswamy <i>et al.</i> , 2012)
Rev binding protein	2OLM	Unpublished
RNase	3LP3	(Su <i>et al.</i> , 2010)
Reverse transcriptase	3KLF	(Tu <i>et al.</i> , 2010)

Supplementary Table 3: Degrees of protein fold disruption in natural and simulated HIV-1 recombinants (full-genome analysis).

Protein	Number of breakpoints	Mean E-score of real recombinants	Mean E-score of simulated recombinants	p-value
Capsid protein P24	12	0.600	0.836	0.09667
Protease	8	2.500	2.068	0.52960
Reverse transcriptase	62	0.634	1.482	0.00036
RNase	19	1.056	1.297	0.30090
Integrase	12	0.455	1.280	0.02478
gp120	40	7.031	9.673	0.00131
gp41	2	0.000	1.366	0.12217
Nef	11	0.750	3.603	0.02321

Supplementary Table 4: Breakpoint position tests for the coding-region (CR) recombination dataset.

Test ¹	Median mimic value	Median simulated value	z-score	p-value ²	Comments
1. HIV-1 alignment-wide pairing probability near recombination breakpoints	0.593	0.582	2.092	0.036	Recombination breakpoints occur significantly more frequently within predicted paired regions in the alignment than can be accounted for by chance.
2. HIV-1 alignment-wide pairing probability near 5' recombination breakpoints	0.579	0.573	1.440	0.150	5' recombination breakpoints occur slightly more frequently within predicted paired regions in the alignment.
3. HIV-1 alignment-wide pairing probability near 3' recombination breakpoints	0.604	0.589	1.660	0.097	3' recombination breakpoints slightly more frequently within predicted paired regions in the alignment.
4. Minor sequence pairing probability near recombination breakpoints	0.612	0.594	1.512	0.131	Slight tendency for recombination breakpoints to occur within predicted paired regions in minor parent sequences.
5. Major sequence pairing probability near recombination breakpoints	0.600	0.587	1.455	0.146	Slight tendency for recombination breakpoints to occur within predicted paired regions in major parent sequences.

List of co-authored publications

- Martin,D. et al. (2011) Recombination in eukaryotic single stranded DNA viruses. *Viruses*, 3, 1699–738.
- Golden,M. and Martin,D. (2013) DOOSS: a tool for visual analysis of data overlaid on secondary structures. *Bioinformatics (Oxford, England)*, 29, 271–2.
- Arunapuram,P. et al. (2013) StatAlign 2.0: combining statistical alignment with RNA secondary structure prediction. *Bioinformatics (Oxford, England)*, 29, 654–5.
- Anderson,J.W.J. et al. (2013) Quantifying variances in comparative RNA secondary structure prediction. *BMC bioinformatics*, 14, 149.
- Bernardo,P. et al. (2013) Identification and characterization of a highly divergent geminivirus: evolutionary and taxonomic implications. *Virus research*.
- Muhire,B.M. et al. (2013) Evidence of pervasive biologically functional secondary-structures within the genomes of eukaryotic single-stranded DNA viruses. *J. Virol.*, JVI.03031–13–.