# Temporal Ordering of Substitutions in RNA Evolution: Uncovering the Structural Evolution of the Human Accelerated Region 1

Maria Beatriz Walter Costa[a,b,c], Christian Höner zu Siederdissen[c], Dan Tulpan[d], Peter F. Stadler[c,e,f,g,h], Katja Nowick[a,b,c,i,j,*]

[a] *TFome Research Group, Bioinformatics Group, Interdisciplinary Center of Bioinformatics, Department of Computer Science, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany*
[b] *Paul-Flechsig-Institute for Brain Research, University of Leipzig, Jahnallee 59, D-04109 Leipzig, Germany*
[c] *Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, 04107 Leipzig, Germany*
[d] *National Research Council Canada, Information and Communication Technologies, 100 des Aboiteaux Street, Suite 1100, NB E1A7R1, Moncton, Canada*
[e] *University of Vienna, Institute for Theoretical Chemistry, A-1090 Vienna, Austria*
[f] *Max Planck Institute for Mathematics in the Science, 04103 Leipzig, Germany*
[g] *Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany*
[h] *Santa Fe Institute, Santa Fe NM 87501, USA*
[i] *Bioinformatics, Faculty of Agricultural Sciences, Institute of Animal Science, University of Hohenheim, Garbenstrasse 13, 70593 Stuttgart, Germany*
[j] *Freie Universität Berlin, Faculty for Biology, Chemistry, and Pharmacy, Institute for Biology, Königin-Luise-Strasse 1-3, 14195 Berlin, Germany*

## Abstract

The Human Accelerated Region 1, HAR1, is the most rapidly evolving region in the human genome. It is part of two overlapping long non-coding RNAs, has a length of only 118 nucleotides and features 18 human specific changes compared to an ancestral sequence that is extremely well conserved across non-human primates. The human HAR1 forms a stable secondary structure that is strikingly different from the one in chimpanzee as well as other closely related species, again emphasizing its human-specific evolutionary history. This suggests that positive selection has acted to stabilize human-specific features in the ensemble of HAR1 secondary structures. To investigate the evolutionary history of the human HAR1 structure, we developed a computational model that evaluates the relative likelihood of evolutionary trajectories as a probabilistic version of a Hamiltonian path problem. The model predicts that the most likely last step in turning the ancestral primate HAR1 into the human HAR1 was exactly the substitution that distinguishes the modern human HAR1 sequence from that of Denisovan, an archaic human, providing independent support for our model. The MutationOrder software is available for download and can be applied to other instances of RNA structure evolution.

*Keywords:* Human evolution, computational modeling, dynamic programming, non-coding RNA, secondary structure, data visualisation
*2010 MSC:* 92-04, 92-08

## 1. Introduction

Functional innovations at the phenotypic level are eventually the result of genetic changes. While most mutations are (nearly) neutral or even detrimental, occasionally they lead to innovations by affecting the expression pattern of genes or the sequence of the gene product itself. In the latter case, novel molecular and biological functions are thought to be the result of changes in the molecule's structure that in turn changes its interactions and thus its position within cellular networks. As a consequence, the mutant becomes subject to new selection pressures that may lead to rapid adaptive evolution [1]. Such scenarios are extremely difficult to model computationally, because it requires explicit models of structure formation, all relevant interactions in the network, and the functions of the network. In the special case of functional RNAs it is at least possible, however, to model the adaptation towards a target structure [2, 3]. In this contribution we ask to what extent the detailed history of recent adaptive evolution can be reconstructed from the knowledge of the current and ancestral structures of a rapidly evolving RNA element.

There are some regions on the genome that have accumulated many human specific changes while remaining

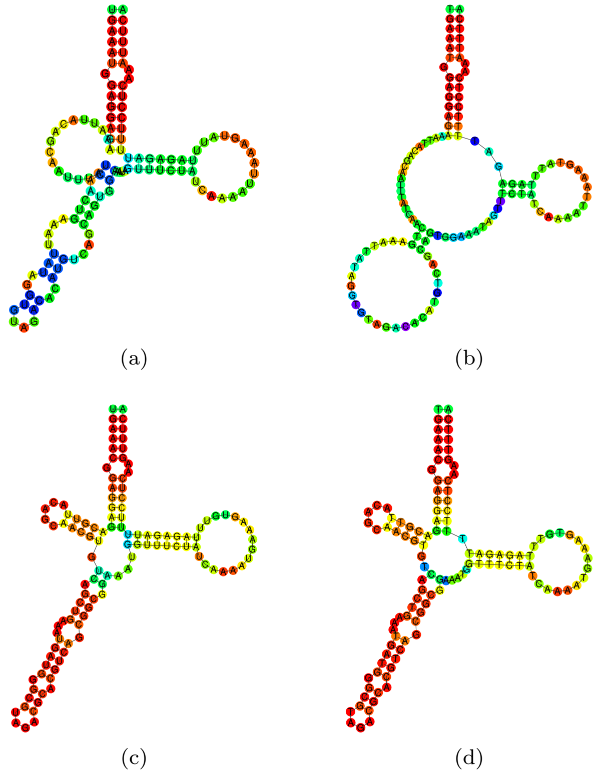
[*] Corresponding author

*Email addresses:* bia@bioinf.uni-leipzig.de (Maria Beatriz Walter Costa), choener@bioinf.uni-leipzig.de (Christian Höner zu Siederdissen), dan.tulpan@nrc-cnrc.gc.ca (Dan Tulpan), studla@bioinf.uni-leipzig.de (Peter F. Stadler), katja.nowick@fu-berlin.de (Katja Nowick)

constant in other closely related species. These are called human accelerated regions and are candidates for generating human specific traits [4]. The Human Accelerated Region 1 (HAR1) is the region with the most human specific changes in the primary sequence. It will serve here as our paradigmatic example. HAR1 is only 118 nucleotides long and contains 18 human specific substitutions, which corresponds to a rate of 0.025 substitutions per site per million years. This substitution rate is in humans two orders of magnitude larger compared to other amniotes, for which we expect only 0.27 rather than 18 substitutions in total [4]. HAR1 is located in a pair of overlapping long non-coding RNAs, HAR1F and HAR1R, both of which are very specifically expressed in Cajal-Retzius cells between the 7th and 19th gestational weeks. This is a crucial period for cortical neuron specification and migration. HAR1F and HAR1R were also reported to be co-expressed with reelin (RELN), a protein involved in the organisation of the laminar cortex of the brain [4]. HAR1R and HAR1F are direct targets of the RE1-silencing transcription factor (REST) in human but not in mouse [5], indicating a change in their regulatory interactions. Considering the highly specific expression pattern of HAR1 in Cajal-Retzius cells, HAR1F and HAR1R may have an important role in the correct organization of the developing human brain.

The secondary structure of HAR1 is conserved among vertebrates with the exception of humans. In humans, HAR1 forms a stable cloverleaf-like structure, that differs from the one of the other species, which was first supported by dimethyl-sulphate (DMS) treatment structure probing [4]. The predicted divergence of the human structure was afterwards confirmed by two independent empirical methods. Chemical and enzymatic probing [6] resulted in a hairpin-like structure for the chimpanzee sequence and a cloverleaf-like structure for the human, although the authors mentioned that the chimpanzee structure might also be able to form a hairpin structure. NMR spectroscopy confirmed the chimpanzee model but implied that the human structure contains two small hairpin domains connected by a flexible middle region [7].

Since HAR1 has been largely conserved among amniotes except in humans [4], we considered the chimpanzee version of HAR1 as the most likely ancestral version before humans and chimpanzees split from each other. Surprisingly, all 18 human specific substitutions replace an ancestral `A` or `T` with a `G` or `C`. In general `G-C` interactions are energetically more favorable than `A-T`, so that the substitutions are expected to lead to an overall stabilization of the RNA structure, which is in apparent contradiction with the empirically observed weakening of the ancestral hairpin structure in favor of a much more flexible human structure. A closer inspection, however, shows that the ancestral hairpin structure is only marginally stable and the human-specific substitutions have lead to a strategic stabilization of two of the three hairpins of the predicted cloverleaf structures (Fig. 1 (a,c)). This is clear when comparing the minimum free energy (MFE) structures of the ancestral and human versions, and especially when comparing the centroid structures (Fig. 1 (b,d)). While MFE structures are the most stable in the ensemble and require more energy to be broken, they are not the only ones occuring in the cell. The centroid structure can be interpreted as the most informative representative of the Boltzmann ensemble of possible structures, since it has the smallest average base pair distance to all alternatives. The ancestral centroid is much less stable than the human centroid, that is, the whole ensemble for human is structurally closer to its MFE and shows much less structural diversity.


(a)

(b)

(c)

(d)

Figure 1: Ancestral and human HAR1 structures as start and end points of an evolutionary simulation. Ancestral (a) minimum free energy (MFE) and (b) centroid structures, as the start point in our model, and human (c) minimum free energy and (d) centroid as the end point in our model. Nucleotides are colored according to their pairing frequency in the ensemble. Base pairs in shades of red occur in $\geq 90\%$ of all structures in the ensemble, while green to yellow denote increasing probabilities $\geq 50\%$. For unpaired nucleotides, colors toward red denote increasing unpairedness. The centroid structures contain base pairings that occur in more than 50% of the structures of each ensemble.

To understand how the human specific structure of HAR1 evolved, we developed a computational model. We revealed a strong reshaping of the HAR1 structure and the most likely last change of the 18 substitutions. Interestingly, all of the substitutions seem to drive HAR1 to a more stable ensemble. Moreover, the last predicted mutation separates the structures of the modern human from the archaic Denisovan hominin and recreates a stem that had been weakened on the evolutionary path from the ancestral sequence to Denisovan.

## 2. The Model

Our aim is to investigate the evolutionary history of a ncRNA that evolved from the ancestral structure to the extant structure, by reconstructing the statistically most likely order of substitutions. In this work we focused on HAR1, but the model provides a general framework that allows for statistical inference on the temporal ordering of a set of mutations between any two RNA sequences. Although the actual ancestral sequence is not known, it can be inferred with high confidence in cases such as HAR1, in which the sequence is conserved in the primate lineage, with the only exception in species of the genus *Homo*. Reconstructing the history of an RNA sequence, the human HAR1 in our case, can be phrased as an instance of a combinatorial optimization problem that is equivalent to the Hamiltonian path (HSP) problem [8]. It asks for a path on a graph in which each node is visited exactly once (similar to the travelling Salesman problem). In our setting, the substitutions under consideration are represented as the nodes of a complete graph, and each Hamiltonian path specifies one of the possible orders in which the substitutions may have occurred. The problem we face here *differs* from the usual HSP in the definition of a cost function, which depends on the entire *history*, i.e., on the nodes that have already been visited. Formally, the solution to such a problem falls into the class of histomorphisms [9]. In the following paragraph we describe the model and its assumptions in detail.

### *2.1. Mutation Ordering*

Given an ancestral sequence $x$, a secondary structure $S(x)$ and a corresponding derived extant pair $y$ and $S(y)$, we implicitly know the set $X$ of fixed substitutions from the alignment of sequences $x$ and $y$. What we are interested in is their temporal ordering. Each possible evolutionary path is therefore a permutation $\pi$ of $X$. The structure $S(y)$ of the extant sequence $y$ serves as a proxy for the selection target. This allows us to use a measure of structural distance to $S(y)$ as a proxy for fitness, i.e., substitutions that reduce the distance to $S(y)$ can be thought as adaptive and are quickly fixed, while substitutions that increase the distance to $S(y)$ are discouraged. Thus $f(u) = -d(S(u), S(y))$ serves as a fitness function. The fitness cost of an evolutionary path $\pi$ is then

$$f(\pi) = \sum_{i=2}^{|X|} \big(d(S(\pi_i), S(y)) - d(S(\pi_{i-1}), S(y))\big)_+ \qquad (1)$$

in which the sum only includes those steps in which the fitness decreases, that is the distance to the target increases. The likelihood of a path $\pi$ decreases exponentially with its fitness cost, i.e.,

$$\mathrm{Prob}[\pi] = e^{-\beta f(\pi)}/Z \qquad (2)$$

in which the "inverse temperature" $\beta$ is a scaling parameter measuring the stringency of selection, and $Z$ is a normalization factor.

There is some freedom in modelling the distance. In this contribution, we use the energies of centroid and MFE structures as well as the base pair distances between MFE and centroid structures. Conceivable other choices include variance or Kullback-Leibler distances measured for the base pairing probabilities, as used e.g. in `RNAsnp` [10].

Finding the most likely permutation, i.e., the one that minimizes $f(\pi)$ amounts to computing the Hamiltonian path from $x$ to $y$ with minimal total cost. This problem can be solved by a well-known exponential time dynamic programming algorithm [8, 11, 12], which is applicable in practice for a problem size of $n = 18$ fixed substitutions, as in the case of HAR1. As shown in [12], the use of ideas from algebraic dynamic programming makes it possible to also compute the posterior probabilities $P_{ij}$ for two fixed mutations $i$ and $j$ to be consecutive along a path. Using this matrix of posterior probabilities as the scoring function, the same recursive algorithm can be used to compute the Maximum Expected Accuracy (MEA) path.

The model also makes it simple to compute the probabilities $\pi_{ij}$ that the sequence of fixed substitutions started with position $i$ and terminated with position $j$. These quantities give access to the probability that $\pi_j = \sum_i \pi_{ij}$ of fixed substitution $j$ being the last one.

### *2.2. Intermediate and Backmutations*

Landscapes of an RNA structure tend to be rough, admitting drastic changes in response to a single substitution, and at the same time contain vast neutral plateaus [13, 2, 14]. Any exploration of the landscape based on a subsampling approach may easily not explore regions of high density or other properties of interest. Exhaustive algorithms therefore are preferrable, since they are guaranteed to model the landscape accurately. It is necessary, on the other hand, to restrict the landscape model in such a way that (i) the desired properties are still retained, and (ii) excluded extensions of the model or landscape are not likely to contain the most interesting regions. In the above algorithm, we excluded all unobserved substitutions (which also includes substitutions that might have been fixed in the population only for a short period of time). Thus, in the case of HAR1, only permutations of the 18 differences between human and the ancestral sequence are considered, instead of the unmanageable set of all possible paths that would also include backmutations.

This simplification makes our model computationally feasible, neglecting additional mutations, that might have been fixed only temporarily. However, in order to provide a more detailed view into these substitutions that were introduced and later reverted again, we also describe here a corresponding extension of the above algorithm. More precisely, we consider two cases: (i) A nucleotide that coincides in the ancestral and the extant state was replaced by one of the three alternatives somewhere along the path and later reverted back to its initial state ("back-mutation"). (ii) A nucleotide that has changed between ancestral and extant state has also changed to a different state from both the ancestral and the extant state and only later was substituted by the extant nucleotide ("intermediate mutation"). Within the framework of our model, we write the fitness cost of a path with a single "back-mutation" as

$$\begin{aligned} f(\pi, b_+, b_-) = &\sum_{i=2}^{b_+-1} \big(d(S(\pi_i), S(y)) - d(S(\pi_{i-1}), S(y))\big)_+ \\ &+ d(S(\pi^B_{b_+}), S(y)) - d(S(\pi_{b_+-1}), S(y))\big)_+ \\ &+ \sum_{i=b_+}^{b_--1} \big(d(S(\pi^B_i), S(y)) - d(S(\pi^B_{i-1}), S(y))\big)_+ \\ &+ d(S(\pi_{b_-}), S(y)) - d(S(\pi^B_{b_--1}), S(y))\big)_+ \\ &+ \sum_{i=b_-}^{|X|} \big(d(S(\pi_i), S(y)) - d(S(\pi_{i-1}), S(y))\big)_+ \end{aligned} \tag{3}$$

The backmutation occurs after $b_+$ steps in the permutation order have already been taken and is undone after $b_-$ steps have been taken. The permutation $\pi$ now includes the observed mutations as well as the steps $b_+$ and $b_-$. The RNA landscape changes with the introduction of the backmutation, which requires a temporary change into this new landscape, indicated by $\pi^B$. The additional terms are the incurred costs due to the switch. This new variant of the algorithm requires the solution of *three* Hamiltonian path problems, which are connected by edges denoted by a change of landscape due to the backmutation. The cost associated with the "intermediate mutation" case can be written in an analogous form.

These variants introduce a substantial additional cost in computation time. For sequences of length $n$ with $k$ mutations, there are $n-k$ backmutations with three possible nucleotide configurations each. The intermediate mutations introduce another $2k$ configurations. Given that $k \ll n$, the approximate increases in the running time *and* the number of sequences in the landscapes is $\approx 3n$. For HAR1 with $n = 118$ and $k = 18$, this amounts to a 336-fold increase in running time, as there are 300 backmutation configurations and 36 intermediate mutation configurations. Thus, instead of originally $2^{18} = 262\ 144$ sequences in the RNA landscape, the algorithm now has to handle 83 623 936 sequences: the original $2^{18}$ sequences plus the three nucleotide alternatives for the 100 positions subjected to backmutations plus the two nucleotide alternatives for the 18 positions subjected to intermediate mutations, amounting to $2^{18}+3\times100\times2^{18}+2\times18\times2^{18-1}$. The most cost-intensive contribution to the total running time is the evaluation of the RNA folding energy. Further extensions to include more than one "intermediate" or "back-mutation" are conceptually simple, but are not feasible in practice because the computational effort grows by another factor of $O(n)$ for each additional detour. Nevertheless, we can use the extended algorithm to check whether particular "intermediate" or "back-mutations" might have a dominant impact.

We note that both, the evaluation of the RNA folding energies for each member of the RNA landscape, and each individual calculation of the mutation order, are in fact parallel in nature. As a consequence, provided sufficient computational resources, a somewhat deeper exploration is possible.

### *2.3. Marginalization for Extended Models*

The cost associated with a permutation order $\pi$, Eq. 3, includes both the introduction of the backmutation $b_+$ and its eventual reversal $b_-$. The fully specified model $M(\pi, b_+, b_-, p, u) = f_{p,u}(\pi, b_+, b_-)$ yields the cost for a particular mutation order $\pi$, with a backmutation into and out of nucleotide $u$ at position $p$ in the RNA landscape; $b_+, b_-$ specify where the backmutation and its reversal occur along the sequence substitutions. We are most interested, however, not in the full details but rather in a more informative comparison of the models with and without a backmutation. To this end, the marginal likelihood is computed in Eq. 4

$$\int_{\pi,b_+,b_-} M(\pi, b_+, b_-, p, u) d\pi b_+ b_- =: M'(p,u) \tag{4}$$

for each pair $(p, u)$. This yields the evidence for a model that introduces a backmutation at $p$ of nucleotide $u$, but integrates out the exact order of mutations.

A more complex model will, in general, allow for a higher total evidence. This impact can be explicitly calculated for a penalty function of the type $f = g(\cdot)_+$ in which only the positive part of $g$ is taken into account. In such a case we have $\inf f = 0$ and given a permutation of $\{1 \dots n\}$ any permutation for which $f = 0$ holds provides a partial evidence of $\prod_1^n e^{-0} = 1$. The marginal likelihood for the original model (without backmutations) is then bounded by

$$\sum_{\pi\in\Pi} \prod_1^{|\pi|} e^{-0} = n! \tag{5}$$

or in the log-domain $\ln n!$, the maximal log-evidence in "nats" (units of information in the natural logarithm) for the model. The original model therefore yields at most $\approx 36.40$ nats, while a model incorporating backmutations could provide $\approx 42.34$ nats, a difference of about 6 nats due to the fact that the original model only deals with 18! permutations while the backmutation model deals with 20! permutations. This means that while each model incorporating a backmutation can be compared directly with any other, direct comparisons between different types of models (original, intermediate mutation, backmutation) need to be handled more carefully and we leave the choice of the penalty value open

(see Sec. 4.3).

## 3. Material and Methods

HAR1 has been extremely conserved in vertebrates. There is only one base that does not concur among non-human primates [4]. For this reason, we used the chimpanzee sequence as an approximation to the ancestral sequence. The HAR1 sequences were retrieved from the NCBI nucleotide database, including human, chimpanzee, and the archaic Denisovan. Independently of the causes of fixation, we restrict ourselves to the 18 fixed substitutions separating the human and ancestral sequences, and consider backmutations via a variant of our algorithm at the end of Section 4 section only. Since we assume that this rapid evolution was largely adaptive, back-mutations can be disregarded for now. We consider all subsets of the 18 observed fixed substitutions as potential intermediates.

Both minimum energy secondary structures and base pairing probabilities were computed using the `ViennaRNA` package [15] (Version 2.3.3) with the standard Turner energy model [16] for RNA secondary structures (dangling model `-d2`, and no lonely base pairs (`--noLP`)). These predictions were used to determine the structural and energetic differences between potentially consecutive mutations.

### *3.1. Visualisation of RNA Secondary Structures*

Since the comparison of the Boltzmann ensemble of two structures is more informative and yields more detailed insights than the comparison of MFE or centroid structures alone, we used superpositions of base pairing probability dot-plots with different colors for each species. While the combined dot-plots are useful to obtain a quick overview, they can be difficult to interpret.

The `CS`$^2$`-UPlot` [17] provides an alternative visualization representing the two main information components of an RNA secondary structure in two concentric graphical layers: the RNA sequence and the MFE and alternative base pairing possibilities. It uses `Circos` version 0.69-3 [18] and Perl version 5.022001 and combines base pairings with dot-plot values in a single graphical representation. It has the advantages of better highlighting similarities and differences than dot-plots and providing with the circular diagrams a graphical representation that is more intuitive to biologists.

### *3.2. Diversity of HAR1 in human populations*

To further investigate variability of the HAR1 region in human populations, we retrieved all reported SNPs in the 118 base pair HAR1 region using the ENSEMBL `Data Slicer` from the data set provided by the 1000 Genomes project [19]. We also checked the human genome for possible structural paralogs of the HAR1 region using Infernal [20], with no such paralogs being identified.

## 4. Results

### *4.1. Comparison Between Ancestral, Archaic and Modern Human Structures*

Centroid structures typically yield a better impression of the consensus of the equilibrium ensemble of secondary structures than the MFE structure. The centroid of the ancestral structure has a much more flexible space for base pairing (Fig. 1), and can form both a hairpin and a cloverleaf structure, which has been reported before [6]. In contrast, the human sequence has a more constrained set of energetically lower free energy structures, and hence exhibits a better defined, more stable cloverleaf-shaped structure (Fig. 1). This is consistent with the expectation of the increased GC content of the human sequence. We conclude that stabilization of the cloverleaf structure is a plausible model for how selection acted at the level of RNA structure.

The Denisovan HAR1 differs from its modern human counterpart only by a `T` instead of a `C` in position 47. The archaic human structure shares small stems with modern humans, which are only slightly shifted. However, the structural space of the Denisovan structure is still more diverse, featuring more base pairs that are less well-defined than in modern human, thus appearing more similar to the ancestral state (see Fig. 2). A corresponding dot-plot representation is shown in the Appendix.

In addition to uncovering the evolutionary path from the ancestral to the human version of HAR1, we also asked whether there are variants of HAR1 among modern humans. The 1000 Genomes Project [19] reports only three SNPs for HAR1: C47T, C52T and G113C, each occurring in less than 1% of the surveyed populations. The variant C47T (a change from Cytosine to Thymine at HAR1 position 47) is only present in South and East Asian populations and was not detected in African, American or European populations. Note that this is the exact same variant found in Denisovan. This is interesting, since Denisovans lived in the area ranging from Siberia to South East Asia and have inbred with modern humans who lived in the same area [21]. It may be a Denisovan HAR1 variant, still present in our species today and thus provide independent support for our suggestion that position 47 was one of the very last steps in the evolutionary reshaping of the HAR1 structure. It could however also be a case of parallel evolution that brought the ancestral variant back into modern humans.

While the variant C47T is shared with Denisovans, the other two variants seem to be novel, i.e. specific to modern humans. The variant C52T is exclusive to American and African populations, while the variant G113C is exclusive to Asian populations. These two variants seem to be novel, since they are not present in the ancestral nor in Denisovan. Position 52 is invariant in all amniotes, while position 113 has changed from an A in the ancestral to a G in Denisovan and modern humans. All three observed variants decrease the stability of the very stable wildtype

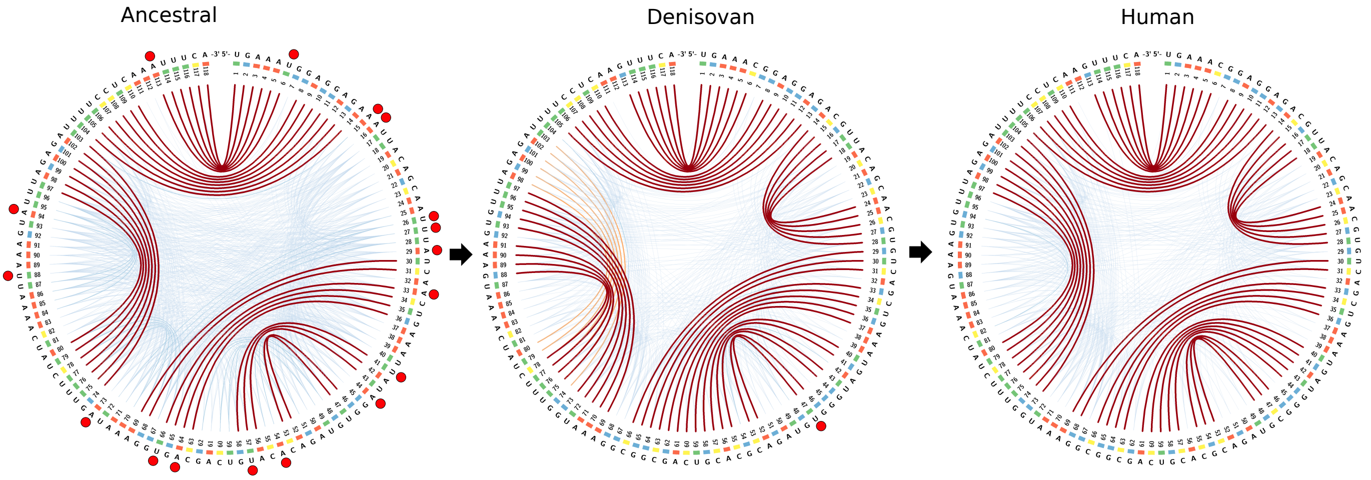


Figure 2: Comparison of the ancestral (left), Denisovan (middle) and modern human (right) ensembles of HAR1 secondary structures. The plots contain the sequence on the outer layer, the MFE base pairings in red lines and alternative base pairing possibilities in orange and blue, with orange base pairings being more likely than the blue ones. Mutations in relation to the modern human sequence are indicated in red dots. See Fig. 9 for a larger version.

human HAR1 ensemble. The variant at position 52 has the strongest impact, which can be seen especially in the centroid structure (Fig. 6). The MFE of variant 52 however still folds into a cloverleaf format (Fig. 6). The variant at position 47 destabilizes a small hairpin in the human centroid structure back to the Denisovan state. Despite these effects on the structure, no associations to diseases were reported for any of these three variants in the DisGeNET database [22]. Apart from the structural impact, any functional consequence caused by the variations could only be assessed by further experiments, which is out of scope for this project.

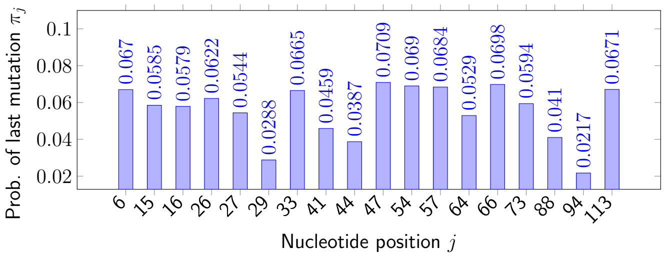


Figure 3: Probability $\pi_j$ for each mutation to be the last mutational event with $\beta = 1.0$. Nucleotide position 47 is the Denisovan-human mutation and has the highest posterior probability. Position 54 has $\geq 50\%$ posterior probability to pair with Position 47. The base pairs 44-57 and 33-66 are part of the same hairpin in human with $\geq 50\%$ probability.

## *4.2. Reconstructing the Evolution of HAR1*

We found qualitatively comparable features of the most likely pathways, even when using different models for the structural distances underlying the fitness model for evolutionary paths. Table 1 gives the number of solutions that all share the same equally optimal weight when all stability-gaining fixed mutations are assumed equally likely and mutations that decrease the stability of the structure are scored according to different criteria. We count the number of co-optimal paths for four different fitness models. The *MFE* and *centroid* models use the energy gain or loss between two structures and are essentially $\max(0, \Delta_E)$ while the *pairdist* models naturally model only losses. The basepair distance between the two structures is at least 0 and as high as the number of base pairs in the two structures. In all variants of the model there are large numbers of co-optimal permutations, suggesting that evolutionary paths along with monotonically increasing fitness were easy to find. It is particularly easy to find a large number of co-optimal paths using the MFE energy as fitness function, in which all mutational steps increase the fitness. Not surprisingly we find that there are more paths that stabilize the minimum free energy structure than paths that keep the centroid stable – and thereby stabilize a majority of the ensemble of structures.

We observed a large degree of redundancy, with many equivalent evolutionary trajectories, which leaves only small differences in the probabilities for the last mutations in the sequence.

Nevertheless, it is still interesting to note that the T to C transition in position 47, which separates Denisovan from modern human, is predicted as the most likely last step from our model (Fig. 3). Importantly, as the model only has information of the ancestral and the final states, but does not have information on the Denisovan state as a likely intermediate, we can interpret this result as a direct support for our modelling approach.

The large number of feasible paths makes it impossible to analyze them individually. Instead, we provide further summary statistics in the form of edge probability plots. These plots identify likely neighbours in the chronological ordering of the mutational events. Fig. 4 summarizes these probabilities for the pair distance fitness model based on centroid structures. In particular, the base pair 54 → 47 is recognized to have high probability in this chronological order. The `A` to `G` substitution at position 54 furthermore sets the stage for the `C/T` polymorphism, which, despite de-stabilizing the structure, maintains a similar ensemble of structures. Note that in ancestral as well as in Denisovan, position 54 was unpaired, while 47 paired with 52. In human, position 47 forms a pair with 54, while position 52 is now unpaired (Fig. 2). Interestingly, this rearrangement stabilizes the whole inner stem, which can be clearly seen in the centroid structures (Fig. 6(a) and (b)). We can conclude from the results that the last event in the evolution of the human structure stabilized the whole lower stem.

Summarizing the likely sequence of events as reconstructed by our approach, we arrive at the following plausible scenario. If the last stabilizing event occured in the lower stem, and the most downstream stem was already present in the ancestral structure, it is reasonable to conclude that the first big event that was fixed in the evolution of the human HAR1 formed the most upstream stem. This could have occured in one step with position $16A \rightarrow G$, which can already form a weak GU pair with position 26 (Fig. 2). With the surrounding bases also being able to form AU pairs (14 and 28, 15 and 27, 17 and 25, 18 and 24), a change in this position could have formed the third stem.

### *4.3. Impact of Back- and Intermediate Mutations*

In this section we investigate the (numerical) impact of including intermediate or backmutations in the model. In the following text, we focus on the variant including a single backmutation. The results hold analogously for an intermediate mutation. The histograms with the respective impact are given for both, back- and intermediate mutations (Fig. 5 and Fig. 8).

Fig. 5 shows the per-site and per-nucleotide impact of both, back- and intermediate mutations using the basepair difference model. For correct interpretation of the figure (and Fig. 8), note that the difference in model complexity as described in Sec. 2.3 gives different ln-evidence values for the original and extended models.

Table 1: Frequency of co-optimal permutations of the 18 human-specific fixed substitutions in HAR1 for different choices of the distance function in Eq. (1). The first two distance functions penalize increases in the folding energy computed for the minimum energy and centroid structures, resp. The number of base pairs that differ between the structure at each step and the human target is used as an alternative model. The third column gives the fraction of co-optimal paths among the 18 permutations.

| fitness model | # co-optimal path | fraction |
|---|---:|---|
| minimum energy | 3 931 510 681 533 | $6.14 \times 10^{-4}$ |
| centroid energy | 1 615 195 878 | $2.52 \times 10^{-7}$ |
| m.e. pairs | 17 338 903 092 | $2.71 \times 10^{-6}$ |
| centroid pairs | 2 239 218 | $3.50 \times 10^{-10}$ |

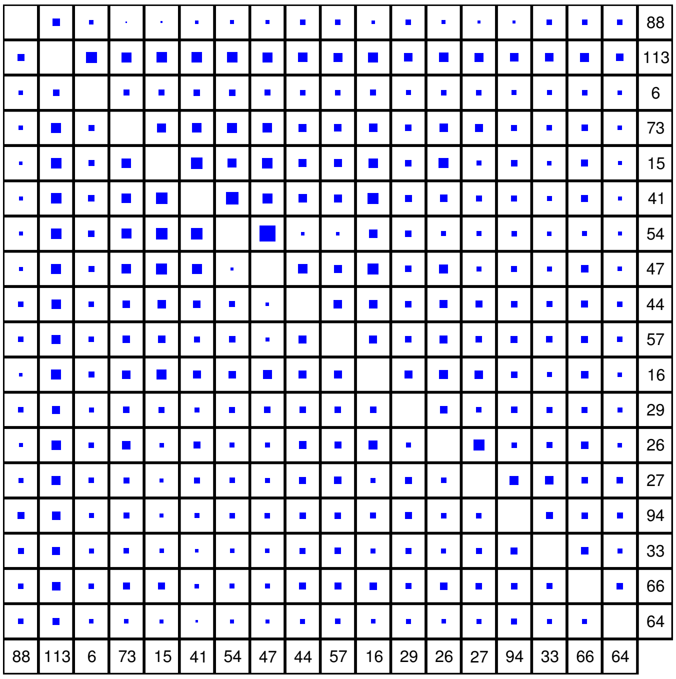


Figure 4: Probability $P_{kl}$ of mutation $k$ (row) to be followed by mutation $l$ (column) following mutation $k$ (row) using the pair distance fitness function on centroid structures. The pseudo-temperature is set to $\beta = 1.0$. Mutations are arranged in their order of appearance in the MEA path. The boxes are scaled as $1/(1 - \ln P_{kl})$ to highlight the uncertainties involved in determing the most likely evolutionary path. We note the high posterior probability for the sequence 54 → 47. The two nucleotides form a `GC` base pair in the human centroid structure that was produced as the last step in the evolutionary trajectory. The best weight of a trajectory ending with mutation 47 is about 1/8 of the trajectory shown here.

In the results, we have *not* penalized the probability of introducing a backmutation in addition to the cost given by the cost function (Eq. 3). The reason is that any such additional penalty can be moved out of the model completely. If one considers a model without any unobserved backmutations, say, 100× more likely, than one with unobserved backmutations, the additional penalty of $\ln 0.01 \approx -4.61$ is to be added to the given $\ln Z$ values in Fig. 5. The huge differences in observed log-evidence between models based on distance to the target structure (Fig. 5) and models based on changes in free energy (Fig. 8) is due to the possibility of finding permutations which only improve in energy, while structures always lead to change. Hence, energy-based models can come close to the maximal observable log-evidence, while basepair-distance based models cannot. In addition, any model that allows for negative costs to be assigned to beneficial steps can have arbitrarily high log-evidence values.

For HAR1 in particular, many of the possible $(p, u)$ pairs have modest impact. As an example, most backmutations falling into the range $81 \leq p \leq 96$, a hairpin loop, in the ancestral and human structure, have limited impact. Additional mutations in the stem of this hairpin, however,

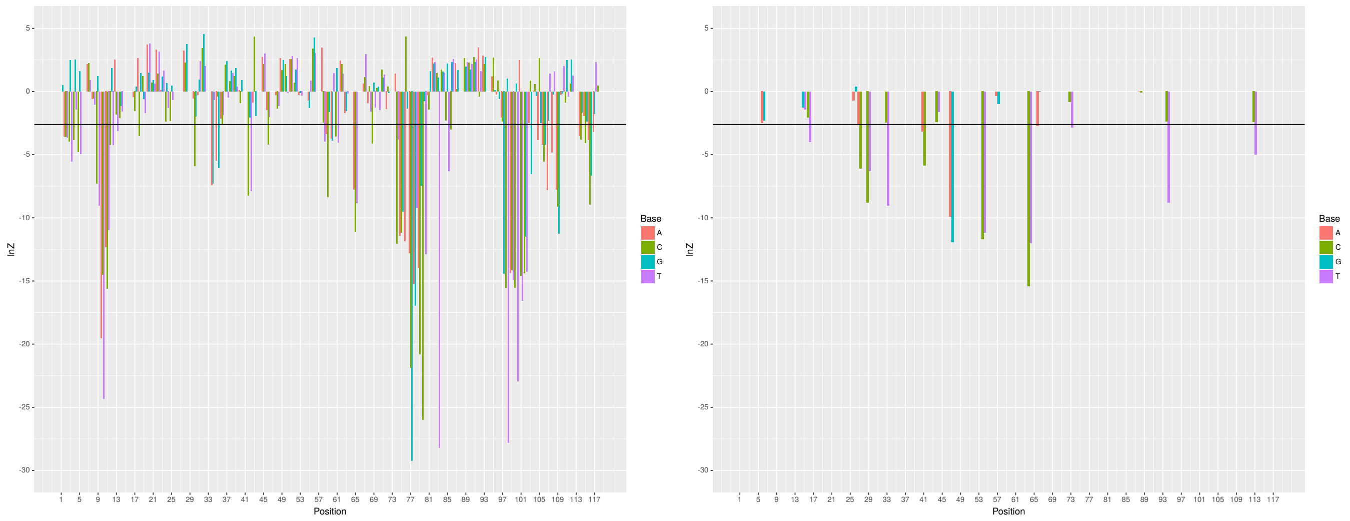


Figure 5: The impact of backmutations (left panel) and intermediate mutations (right panel) per site on the log-evidence ($\ln(Z)$) for the HAR1 sequence with the centroid basepair distance model. The $\ln(Z)$ values shown do not include any energetic penalty for the relative probability of including an additional mutation compared to the model with 18 observed mutations. The bold horizontal line is the $\ln(Z)$ for the original model not incorporating any non-fixed mutation. Left: backmutations at sites with known mutation. Right: intermediate mutations at sites with differing ancestral and extant nucleotides.

are extremely unlikely to have occurred. The difference of about 15 "nats" means that such a model has odds of only $1 : e^{15}$, i.e., $1 : 3 \times 10^6$ compared to a backmutation in the loop region. Overall, no single intermediate or backmutation would have lead to a possible evolutionary path that would be much more stable, given our cost functions.

## 5. Conclusion

Guided by HAR1 as the paradigmatic application, we have introduced here a suite of tools to investigate evolutionary trajectories of secondary structures in detail. We introduced a convenient visualization method for structural ensembles that enables intuitive insights into evolutionary changes of secondary structures at high resolution. A dynamic programming method makes it possible to compile exact statistics over possible evolutionary trajectories. Despite its exponential runtime the algorithmic approach is efficient enough to handle systems with up to at least 20 substitutions, which includes at least all moderate size structured RNAs. The approach proposed here can be used to test whether rapid changes are associated with altered selection pressures for novel RNA structures. Although beyond the scope of this contribution, the same type of model can also be used to evaluate adaptive evolution of protein structures – all that is needed is a distance measure to a target structure that correlates well with the actual fitness effects, i.e., that fitness is largely determined by structure.

Furthermore, an extension of our model provides statistics on the impact of intermediate mutation events that have not been observed as fixed in the extant species. This extension is computationally much more demanding compared to the variant that includes only observed mutations, but provides position-wise information on the impact of such mutation on the sequence.

Given a particular fitness model, it is quite possible to observe a comparatively large number of paths from the ancestral to the extant sequence that have equal cost. There are several reasons for this behaviour. First, consider a set of mutations of nucleotides that are predicted as unpaired in the ancestral and extant sequence. Given the Turner energy model [16], it is unlikely that such a mutation will lead to a change in the current structure. Mutations in a base pair that do not destroy the pair also have minor impact. As such, the order of these substitutions can be of minor impact *for any given order*. As a consequence, these have to be considered as equivalent since any relative order has the same overall effect. On a slightly larger scale, it stands to reason that a set of mutations that impact different base pairs in the same stem often yield different orders with the same cost. These arguments mirror those given to favor partition-function based models of RNA secondary structure prediction that provide posterior probabilities for base pairing. As with RNA secondary structure prediction, we here also advocate probabilistic answers as given by Fig. 3 for the last mutational event or Fig. 4 to determine "temporal neighbours".

A computational model assuming only selection against increasing divergence from the modern human target structure correctly identifies the single difference between human and Denisovan HAR1 as the most likely last step along the evolutionary trajectories. With that, we have shown that the rapid evolution of HAR1 from the last common ancestor between human and chimpanzee to the modern human sequence can be explained by directional selection for the more stable, modern secondary structure. It is likely that the stabilization of the lower stem of the human HAR1 structure was the last step in its evolution until now. Moreover the formation of the most upstream stem was an early step in the human HAR1 evolution since the last common ancestor with chimpanzee.

## 6. Software availability

The software to investigate evolutionary trajectories is available at: `http://hackage.haskell.org/package/MutationOrder`.

It includes both variants of the algorithm (with and without intermediate and backmutations), as well as a function to generate the required RNA landscape. Precompiled binaries are available on github:

`https://github.com/choener/MutationOrder/releases`.

The software to visualize and compare structures is in preparation as `CS`$^2$`-UPlot` web tool.

Supplemental RNA landscape data is available under `http://www.bioinf.uni-leipzig.de/publications/supplements/17-014`

## 7. Acknowledgements

We would like to thank Jan Aerts, Ryo Sakai, Eamonn Maguire, Henrike Indrischek and all contestants of the BioVis 2015 design challenge for providing valuable insights on RNA structural evolution. We also thank the anonymous reviewers for their detailed reviews, which led to numerous improvements of this work, including the incorporation of backmutations in a variant of our algorithm. This work was supported by CNPq Brasil/scholarship of Science without Borders (246039/2012-4) (MBWC), the Volkswagen Foundation within the initiative 'Evolutionary Biology' (KN), and the Deutsche Forschungsgemeinschaft as part of the SPP 1738 (MBWC & KN & PFS).

## References

[1] M. D. Laubichler, P. F. Stadler, S. J. Prohaska, K. Nowick, The relativity of biological function, Th. Biosci. 143 (2015) 143–147.

[2] P. Schuster, W. Fontana, P. F. Stadler, I. L. Hofacker, From sequences to shapes and back: A case study in RNA secondary structures, Proc. Roy. Soc. Lond. B 255 (1994) 279–284.

[3] M. A. Huynen, P. F. Stadler, W. Fontana, Smoothness within ruggedness: the role of neutrality in adaptation, Proc. Natl. Acad. Sci. (USA) 93 (1996) 397–401.

[4] K. S. Pollard, S. R. Salama, N. Lambert, M.-A. Lambot, S. Coppens, J. S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, et al., An RNA gene expressed during cortical development evolved rapidly in humans, Nature 443 (7108) (2006) 167–172.

[5] R. Johnson, N. Richter, R. Jauch, P. M. Gaughwin, C. Zuccato, E. Cattaneo, L. W. Stanton, Human accelerated region 1 noncoding RNA is repressed by REST in Huntington's disease, Physiological genomics 41 (3) (2010) 269–274.

[6] A. Beniaminov, E. Westhof, A. Krol, Distinctive structures between chimpanzee and human in a brain noncoding RNA, RNA 14 (7) (2008) 1270–1275.

[7] M. Ziegeler, M. Cevec, C. Richter, H. Schwalbe, NMR studies of HAR1 RNA secondary structures reveal conformational dynamics in the human RNA, ChemBioChem 13 (14) (2012) 2100–2112.

[8] R. Bellman, Dynamic programming treatment of the travelling salesman problem, J. ACM 9 (1962) 61–63.

[9] R. Hinze, N. Wu, Histo-and dynamorphisms revisited, in: Proceedings of the 9th ACM SIGPLAN workshop on Generic programming, ACM, 2013, pp. 1–12.

[10] R. Sabarinathan, H. Tafer, S. E. Seemann, I. L. Hofacker, P. F. Stadler, J. Gorodkin, `RNAsnp`: Efficient detection of local RNA secondary structure changes induced by SNPs, Hum. Mut. 34 (2013) 546–556.

[11] A. Bjorklund, Determinant sums for undirected hamiltonicity, in: Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on, IEEE, 2010, pp. 173–182.

[12] C. Höner zu Siederdissen, S. J. Prohaska, P. F. Stadler, Algebraic dynamic programming over general data structures, BMC Bioinformatics 16 (19) (2015) S2.

[13] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, P. Schuster, RNA folding landscapes and combinatory landscapes, Phys. Rev. E 47 (1993) 2083–2099.

[14] C. Flamm, I. L. Hofacker, P. F. Stadler, M. T. Wolfinger, Barrier trees of degenerate landscapes, Z. Phys. Chem. 216 (2002) 155–173.

[15] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, ViennaRNA package 2.0, Algorithms for Molecular Biology 6 (1) (2011) 1.

[16] Z. J. Lu, D. H. Turner, D. H. Mathews, A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation, Nucleic Acids Research 34 (17) (2006) 4912–4924.

[17] D. Tulpan, The circular secondary structure uncertainty plot (CS2-UPlot) - visualizing RNA secondary structure with base pair binding, `http://biovis.net/year/2015/papers/circular-secondary-structure-uncertainty-plot\-cs2-uplot-visualizing-rna-secondary.html` (Accessed: 2017-22-10).

[18] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, Circos: an information aesthetic for comparative genomics, Genome research 19 (9) (2009) 1639–1645.

[19] 1000 Genomes Project Consortium and others, A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74.

[20] E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster rna homology searches, Bioinformatics 29 (22) (2013) 2933–2935.

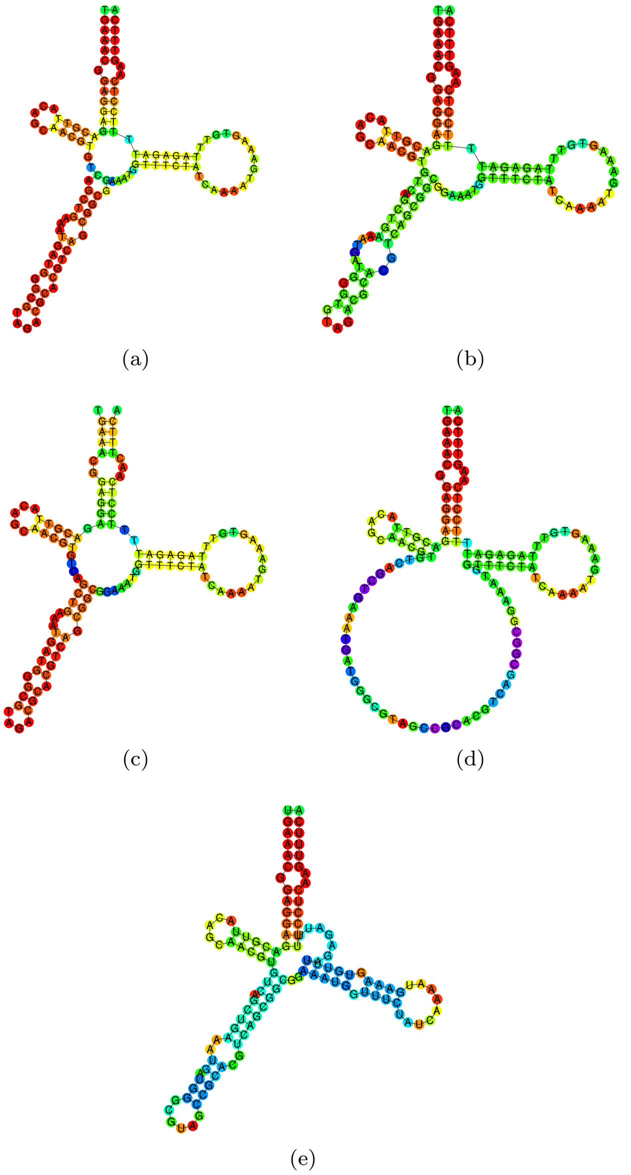
[21] M. Meyer, M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, J. Schraiber, F. Jay, K. Prüfer, C. de, Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, A high-coverage genome sequence from an archaic Denisovan individual, Science 338 (2012) 222–226.

[22] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, L. I. Furlong, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Res 45 (2017) D833–D839.

## Appendix

### A1 *Secondary Structures of Human HAR1 Variants*

Variations of HAR1 within the human species are rare and are found in less than 1% of human populations. Three variants of HAR1 have been reported to date and all cause changes to the wildtype structure (Fig. 6).



(a)

(b)

(c)

(d)

(e)

Figure 6: Wildtype and human variations of the HAR1 structure. (a) Wildtype centroid, (b) variant rs374630364 C47T centroid (the same as Denisovan), (c) variant rs544386774 G113C centroid, (d) variant rs183960348 C52T centroid and (e) variant rs183960348 C52T MFE.

### A2 *Comparative Dot-Plots*

Comparative dot-plots provide an alternative visualization of differences between the structural ensemble of two closely related sequences. The upper right triangle shows the base pairing probabilities in two colors, one for each input sequence. The lower left triangle displays the base pairs of the minimum energy structure.

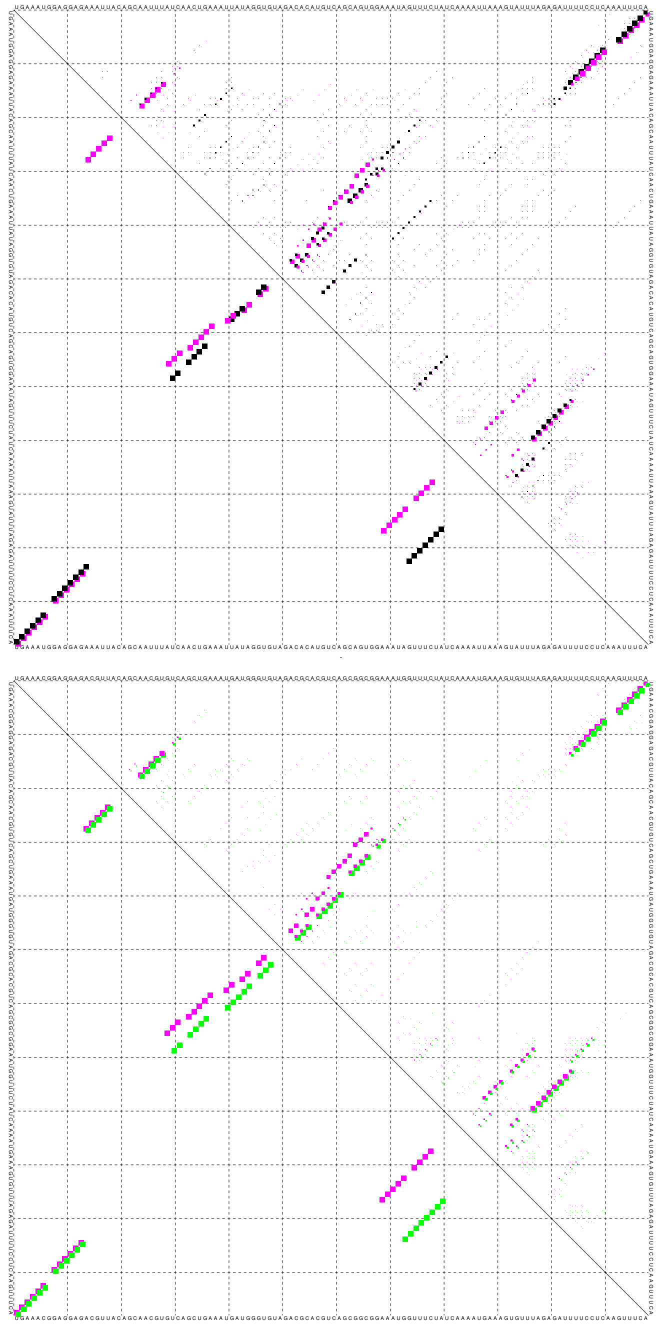


Figure 7: Base pairing patterns of the ancestral (black), Denisovan (magenta) and modern human (green) HAR1 sequences. The plots show the large difference between ancestral and Denisovan structures (left) and the more subtle differences between Denisovan and the modern human structure. Interestingly the 3'-most stem coincides in modern human and the ancestral state, but is shifted in Denisovan. On the other hand, the 5' part of the structure is already close to modern human in the Denisovan structure.

### *Impact of Back- and Intermediate Mutations*

Depicted in Fig. 8 is the impact of including a back- or intermediate mutation into the model. The maximal $\ln(Z)$ value for the original model is $\ln(18!) \approx 36.40$ "nats", $\ln(19!) \approx 39.34$ "nats" for intermediate mutations, and $\ln(20!) \approx 42.34$ "nats" for backmutations. Further details can be found in Sec. 4.3 in the main text.

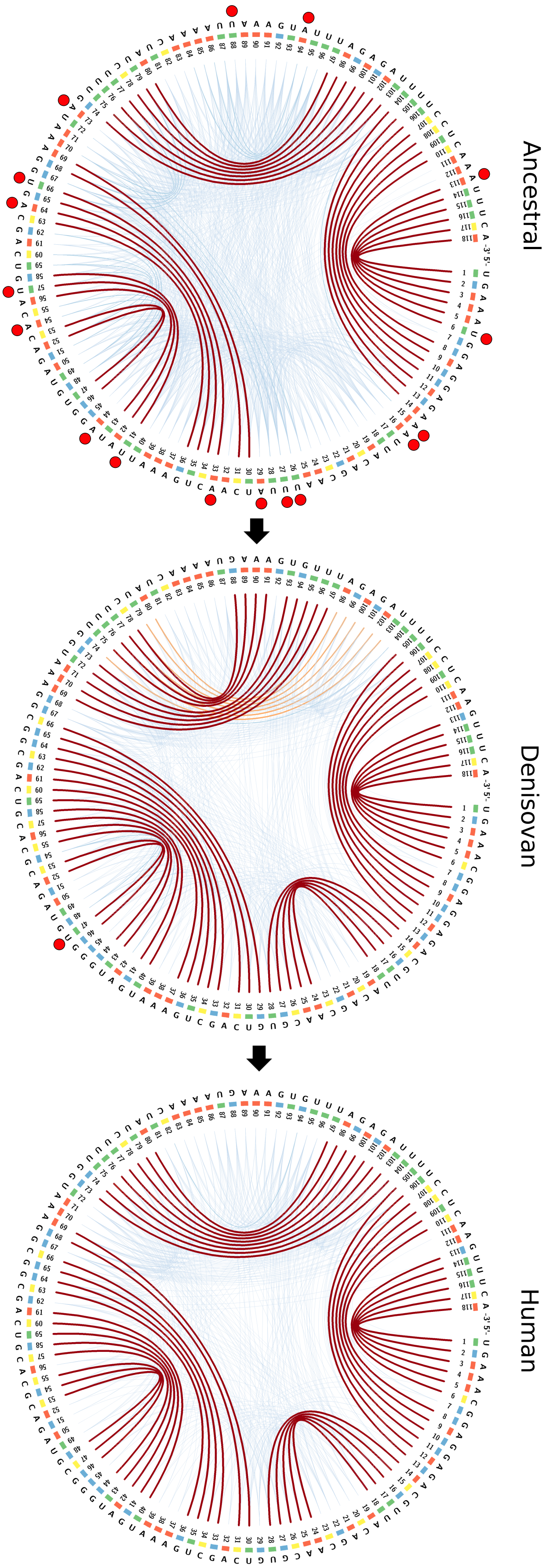


Figure 9: Comparison of the ancestral (left), Denisovan (middle) and modern human (right) ensembles of HAR1 secondary structures. The plots contain the sequence on the outer layer, the MFE base pairings in red lines and alternative base pairing possibilities in orange and blue, with orange base pairings being more likely than the blue ones. Mutations in relation to the modern human sequence are indicated with red dots.

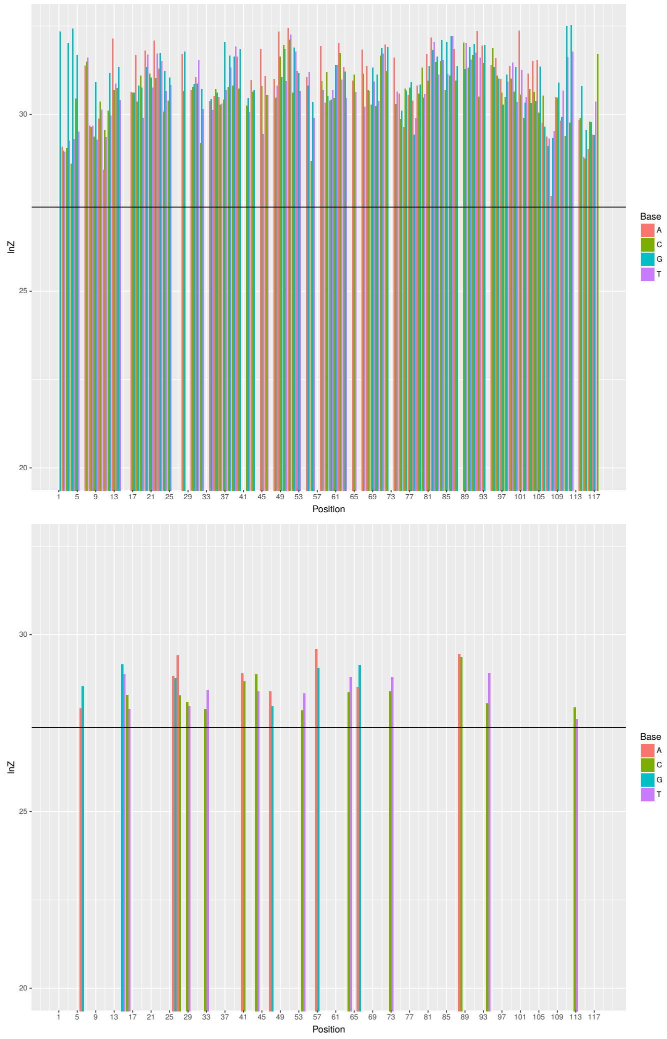


Figure 8: The impact of backmutations (top panel) and intermediate mutations (bottom panel) per site on the log-evidence ($\ln(Z)$) for the HAR1 sequence with the centroid energy model. The $\ln(Z)$ values shown do not include any energetic penalty for the relative probability of including an additional mutation compared to the model with 18 observed mutations. The bold horizontal line is the $\ln(Z)$ for the original model not incorporating any non-fixed mutation. Top: backmutations at sites with known mutation. Bottom: intermediate mutations at sites with differing ancestral and extant nucleotides.