Defining 5S rRNA Structure Space: Point Mutation Data Can Be Used to Predict the Phenotype of Multichange Variants

Madhavi Nayar and George E. Fox*

1Department of Biology and Biochemistry, University of Houston

*Corresponding author: E-mail: fox@uh.edu.

Associate editor: Dan Graur

Abstract

A portion of the 5S ribosomal RNA (rRNA) structure space in the vicinity of the Vibrio proteolyticus 5S rRNA sequence is explored in detail with the intention of establishing principles that will allow a priori prediction of which sequences would be valid members of a particular RNA structure space. Four hundred and one sequence variants differing from the V. proteolyticus 5S rRNA wild-type sequence in 1–7 positions were characterized using an in vivo assay system. Most significantly, it was found that in general, the phenotypic effects of single changes were independent of the phenotypic effect of a second change. As a result, it was possible to use the new data in conjunction with results from prior studies of the same RNA to develop “truth tables” to predict which multiple change variants would be functional and which would be nonfunctional. The actual phenotype of 93.8% of the multichange variants studied was consistent with the predictions made using truth tables thereby providing for perhaps the first time an upper limit estimate of how frequent unexpected interactions are. It was also observed that single changes at positions involved in secondary structure were no more likely to be invalid than changes in other regions. In particular, internal changes in long standard stems were in fact almost always tolerated. Changes at positions that were hypervariable in the context of an alignment of related sequences were, as expected, usually found to be valid. However, the potential validity of changes that were idiosyncratic to a single lineage of related sequences when placed in the V. proteolyticus 5S rRNA context was unpredictable.

Key words: 5S rRNA, sequence space, mutagenesis, neutral network, adaptive landscape.

Introduction

Typically, functional RNAs or proteins are represented in nature by multiple alternative primary sequences. In principle, if a particular primary sequence satisfies all the structural constraints associated with a molecule’s functional requirements, then that primary sequence will have high fitness for the function. In the case of a particular RNA or protein, the set of all such primary sequences that exist, or could exist, can together be considered to comprise a neutral network or structure space for the molecule of interest (Schuster et al. 1994).

Complete knowledge of a molecule’s structure space would in principle provide considerable evolutionary insight. For example, by comparing sequences that are present in extant organisms with those in the structure space, one could determine how much of the structure space has already been explored. Knowledge of the structure space would also make it possible to appreciate the sequence variations that are observed in aligned sequence sets. When constructing phylogenetic trees, realistic ancestral sequences could be required and probable historical paths identified. It might even be possible to make assessments of the most likely future events. To this end, considerable effort has focused on understanding RNA structure spaces (Fontana and Schuster 1998a, 1998b; Schuster et al. 1994; Huynen 1996; Huynen et al. 1996).

In practice, predicting the fitness of macromolecules is a complex process. One reason for this difficulty is that macromolecules can tolerate imperfections that are generally not expected. Another hurdle in predicting the fitness of the macromolecules is that the properties of molecules may vary with physical conditions such as temperature and pH with the result that some sequences may only intermittently belong to the structure space. Typically, macromolecules interact with one another and in order to predict the validity of these macromolecules as members of a structure space, it becomes necessary to understand the structures of molecules in relation to one another as well. Most of the theoretical work done with RNA structure space has been based on the analyses of primary sequence and standard Watson–Crick secondary structure without considering other types of interactions or interactions with other molecules.

In order to make further progress, it would be useful to have a realistic RNA structure space in which membership in or lack thereof of any particular sequence can be predicted from obtainable information. Such information would potentially include primary sequence comparisons, atomic resolution structural data, and point mutation data. To this end, we have begun efforts to establish a local region of the 5S ribosomal RNA (rRNA) structure space using Vibrio proteolyticus 5S rRNA expressed in Escherichia coli cellular context as a realistic model RNA structure space. The available data include multiple primary sequences from related organisms, atomic resolution structural information on the closely related E. coli 5S rRNA (Ban et al. 2000; Szymanski et al. 2002; Schuwirth et al. 2005), and its...
associated ribosomal proteins (L5, L18, and L25), as well as multiple point mutation studies on the *V. proteolyticus* 5S rRNA (Lee et al. 1993; Lee et al. 1993, 1997; Yang 1994; Ban et al. 2000; Zhang et al. 2003).

It is hypothesized here that the phenotypic effect of a point mutation is likely to be independent of the effect of a second point mutation unless together they recreate a Watson–Crick pair. If this is a reasonable conjecture, then single point mutation data can be effectively utilized to better understand the types of changes that are likely to be allowed and to predict the effect of multiple mutations. To examine this conjecture, we herein, experimentally generate large numbers of multichange variants of the *V. proteolyticus* 5S rRNA sequence, some of which likely belong to the SS rRNA structure and some that likely do not. It is found that the phenotype of approximately 94% of these multichange variants can be predicted from knowledge of the behavior of the point mutants that comprise the multichange variant.

Materials and Methods

In Vivo Mutagenesis

Variants of 5S rRNA were created and characterized as described previously (Zhang et al. 2009). The method utilizes an *E. coli* strain, EMG2ΔBDHG, in which five of the eight 5S rRNA genes have been deleted (Ammons and Rampersad 2001). This strain, although viable, grows very slowly. The growth rate increases significantly when it is transformed with a plasmid that provides a functional 5S rRNA gene. Pools of plasmid PCV5S (Zhang et al. 2009), each carrying a different variant of the *V. proteolyticus* 5S rRNA gene, were used to transform the ΔBDHG strain. *Vibrio proteolyticus* 5S rRNA is functional in *E. coli* (Lee et al. 1993) and has been shown by transcriptional analysis to have little or no effect on the cellular environment (Lee et al. 1993). Multiple site-directed mutations in the *V. proteolyticus* 5S rRNA insert of the plasmid were created. Following transformation, individual colonies were selected and the identity of the 5S rRNA-specific variant carried was determined by sequencing the 5S rRNA region of the plasmid insert. Growth rates were determined for each variant. Colony size was also examined. Variants in which the plasmid carries a useful 5S rRNA typically produce much larger colonies after overnight incubation.

Construction of Variant *V. proteolyticus* 5S rRNA Genes

The Quik-Change Multi Site-Directed Mutagenesis Kit from Stratagene Cloning Systems (La Jolla, CA) was used to create libraries of 5S rRNA variants within which each variant carried 1–7 mutations. The *V. proteolyticus* 5S rRNA gene of plasmid PCV5S was the target for mutagenesis. Degenerate primers that consist of a mixture of four oligonucleotides that are identical except at the targeted positions were utilized. Two or three degenerate primers each directing mutations in 1–3 positions were used for each mutagenesis reaction (fig. 1). The positions mutated could be contiguous or noncontiguous as long as the Tm of the primer is such that it can readily bind to the DNA. Mutant plasmids were isolated from *E. coli* clones and checked by sequencing for the presence of desired mutations. A total of 28 libraries were created and ultimately 401 unique variants that carried 1–7 mutations were isolated. Mutagenesis was done in 28 separate sets, each set using 1–3 degenerate oligonucleotide primers. Each set was named with one or two alphabetical characters (e.g., A, C etc.). Each of the variants that were isolated from these sets was named alphanumerically (AJ10, C12 etc.).

DNA Sequencing

DNA sequencing was carried out by SeqWright DNA Technology Services, Houston, TX. The S3 sequencing primer (5′-CGTCCCCGGGATTGTC-3′) for 5S rDNA was used in all cases.

Hanahan Method of Transformation

A modified Hanahan method (Hanahan 1983) was used for preparation and transformation of frozen stocks of competent cells. Several well-isolated colonies were resuspended in 1 ml of Luria–Bertani (LB) and used to inoculate 50 ml of LB medium in a 1-l flask. The cells were grown at 37 °C for 2.5–3.0 h until the culture reached an OD 600 of 0.40. The culture was transferred to sterile ice-cold 50-ml polypropylene tubes and was cooled for 10 min in ice. Cells were pelleted by centrifugation at 2,700 × g for 10 min at 4 °C. The pellet was gently resuspended in 20 ml of ice-cold frozen storage buffer (FSB) transformation buffer and stored in ice for 20 min. The cells were recovered by centrifugation at 2,700 × g for 4 °C. The pellet was resuspended in 4 ml of ice-cold FSB transformation buffer. Dimethyl sulfoxide (DMSO; 140 μl) was added and gently mixed. The suspension was stored in ice for 15 min and an additional 140 μl of DMSO was added to the suspension, which was then returned to ice. Aliquots (200 μl) of suspension were dispensed to chilled sterile microfuge tubes and stored at −80 °C until needed.

Double-stranded plasmid DNA was added (25 ng per 50 μl of competent cells) to competent cells in chilled Falcon 2059 tubes, and the mixture was incubated in ice for 30 min. The cells were heat shocked at 42 °C for 90 s and immediately cooled in ice for 2 min. Eight hundred microliters of LB medium was added to the mixture and the cultures incubated at 37 °C or 1 h at 250 rpm. An appropriate volume of the transforming mixture was transferred to LB agar plates with appropriate antibiotic and incubated overnight at 37 °C.

![Fig. 1. Construction of variant 5S rRNA plasmid.](image-url)
Growth Rate Determination

In all cases, 60 ml of yeast extract-tryptone medium was inoculated with 100 μl of an overnight (16 h) culture also grown in YT medium, and bacterial growth rates were determined by monitoring the optical density of the culture at 600 nm (OD_{600}) with respect to time. Strain EMG2 (wild-type), EMG2ΔBDHG (partial knockout), and EMG2ΔBDHG-pCV5S (the partial knockout strain transformed with plasmid pCV5S harboring the wild type or a mutant V. proteolyticus 5S rRNA gene) were grown at 37 °C and ~200 rpm in an incubator shaker (New Brunswick Scientific Co., Inc., Edison, NJ). One milliliter samples were removed every ~20 min for OD_{600} determination. The exponential phase of the growth curve is specified by

\[ N_t = N_{t0} \times 2^{(t - t0)/\tau}, \]

where \( N_t \) is the cell number at time \( t \), \( N_{t0} \) the cell number at time \( t0 \), and \( \tau \) is the doubling time of the culture. The cell number doubles every \( \tau \) min. By rearranging Equation (1), the doubling time \( \tau \) can be calculated as

\[ \tau = \log_{10} N_t \times (t - t0)/\log_{10} (N_t/N_{t0}) \]

If \( \log \) (base 10) of OD_{600} versus time is plotted for the exponential portion of the growth curve and the plotted points are subjected to a linear regression analysis, the slope of the resulting line is defined as the growth rate \( (r) \). It is given by the expression (Monod 1949)

\[ \log_{10} N_t = r(t - t0) + \log_{10} N_{t0} \]

\[ r = \log_{10} (N_t/N_{t0})/(t - t0) \]

Comparison between Equations (2) and (4) makes it clear that the doubling time \( (\tau) \) and the growth rate \( (r) \) have this relationship: \( \tau r = \log_{10} 2 \).

Colony Size

Fifty milliliters of YT medium were inoculated with 100 μl of an overnight culture also grown in YT and allowed to grow at 37 °C for 1 h at 250 rpm. One milliliter of the culture was removed and the OD_{600} measured. Based on the density, the culture was appropriately diluted and plated on YT agar plates. The plates were incubated at 37 °C for 18 h and the colony size measured in units using a ruler under a magnifying glass (1.58 mm corresponds to 1 unit). If a variant displayed inconsistency in colony size or appearance it was considered that contamination might be present and it was not included in subsequent studies.

Characterization of 5S rRNA Variants

Previous studies used site-directed mutagenesis to produce specific variants of V. proteolyticus in the E. coli cellular context. The individual variants were then characterized by their effect on growth rate using a competition assay (Lenski 1988a, 1988b) and the extent to which the variant was incorporated into 50S subunits, 70S ribosomes, and in some cases polysomes (Hedenstierna et al. 1993; Lee et al. 1993). Using this approach, most variants were found to be of one of three types. “Green” variants had minimal effect on growth rate and incorporated successfully into 50S subunits and 70S ribosomes and thus were considered “valid.” “Red” variants failed to incorporate into 50S subunits or 70S ribosomes and thus were considered “invalid.” “Blue” variants produced a 5S rRNA that was rapidly degraded and thus it was typically uncertain whether the 5S rRNA would be functional if degradation were not occurring. Finally, a small number of variants, regardless of their incorporation behavior, were classified as red because the change clearly had a deleterious effect on growth rate. Previously, all variants from each category were reexamined with the current method in order to establish a correlation between the two approaches. It was found (Zhang et al. 2009) that growth rate studies alone could distinguish the Green mutants from the Blue or Red but the Blue could not be fully distinguished from the Red types. In particular, strain EMG2ΔBDHG grew with an average doubling time of 50 min. When it was transformed with plasmid pCV5S carrying a wild-type V. proteolyticus 5S rRNA gene, the average doubling time was reduced to 30 min. Variants that have a doubling time of 32 min or less correspond to the Blue and Red mutations of earlier studies and can be considered to be valid members of the 5S rRNA sequence space. Those with a longer doubling time correspond to the Blue and Red mutations of earlier studies and although they typically offer some benefit to the 5S rRNA starved cells are likely not valid members of the 5S rRNA sequence space.

Results

Doubling Time Studies

Four hundred and one unique variants that carry 1–7 mutations in the V. proteolyticus 5S rRNA gene were constructed and characterized as described herein. Based on the previous doubling time studies (Zhang et al. 2009) and the new results obtained here, each variant is initially classified as either a valid (less than 33 min) or invalid (34 min or more). Valid mutations are thought to belong to the 5S rRNA sequence space, whereas the invalid variants presumably do not. The majority of the likely functional variants have a doubling time between 27 and 32 min, whereas the likely nonfunctional variants have a doubling time between 34 and 44 min (figs. 2 and 3). The results obtained were reproducible as indicated by the fact that on nine occasions, the same mutation was created in more than one experiment with the same assignment obtained in every case except that of the double mutant A63G, C90U. That variant gave average doubling times of 33.25 and 31, which placed it in the uncertain category in the first case, but the viable category in the second. It was therefore considered to be valid. The results were also highly consistent with earlier site-directed mutagenesis studies. Twenty point mutations examined here were previously studied using site-directed mutagenesis (see supplementary table 1, Supplementary Material online). Six of these were previously assigned to be valid or invalid and the same result was
obtained for five of these. The remaining 14 were previously found to be rapidly degraded and hence were poorly incorporated. Such variants might actually be valid when they avoid degradation and in fact one of them, A19C, was found to be valid by the doubling time measure. The results for all 401 unique variants are summarized in table 1 and given in detail in supplementary table 2 (Supplementary Material online).

Colony Size of 5S rRNA Variants

The colony size of 5S rRNA variants were also examined with the expectation that ΔBDHG cells that receive a functional 5S rRNA will grow well and develop into relatively large colonies, whereas cells that receive a nonfunctional 5S rRNA will not grow well and will develop into much smaller colonies. Although all the likely functional variants gave comparatively larger colonies (1.5–2.5 unit), some of the nonfunctional variants also produced fairly large colonies (0.5–2.0 unit). We have not used colony size as a metric in the current work. However, the results indicate that if one wanted to focus exclusively on either functional or nonfunctional variants then colony size could be used as an effective screen for the variants of interest. In particular, small colonies reliably indicated nonfunctional variants but the reverse was not always true (fig. 4).

Point Mutation Assignments

Sixty-three variants exhibited only one change. This allowed direct assessment of the validity/invalidity of those variants. In addition, in some cases, it was possible to deduce likely validity/invalidity from the behavior of multi-change variants. Finally, previous studies using traditional-directed mutagenesis methods (Lee et al. 1993, 1997; Yang 1994; Ban et al. 2000; Zhang et al. 2003) had established the validity/invalidity of specific variants at RNA positions that occur in one or more of the multiple change variants examined here. In total, the phenotype of 170 point mutations at 78 different positions in the V. proteolyticus 5S rRNA sequence were utilized in the current study as listed in supplementary table 3 (Supplementary Material online).

Mutation Location

5S rRNA is highly structured and one expects that much of that structure must be preserved to maintain a functional molecule. The standard secondary structure of 5S rRNA (Fox and Woese 1975; Ban et al. 2000) consists of four duplex regions that typically contain Watson–Crick (WC) or G-U (GU) wobble pairs. A fifth stem, in what was originally referred to as loop E, consists of two normal pairs and five

![Figure 2](https://example.com/figure2.png)

**Fig. 2.** Average doubling time of various 5S rRNA variants. Green indicates the average doubling time for variants with growth rate of 32 min or less. Red indicates the average doubling time for variants with growth rate of 34 min or more. WT is the wild-type strain without plasmid pCV5S, pCV5S is the partial knockout strain—EMG2 Δ BDHG with pCV5S, and K. out is the partial knockout strain—EMG2 Δ BDHG without any plasmid. Error bars indicate the maximum and minimum values of doubling time for each variant.

![Figure 3](https://example.com/figure3.png)

**Fig. 3.** Relationship between the number of variants and the doubling time of the variants. Green indicates functional variants, red indicates nonfunctional variants, and yellow indicates “uncertain variants” that could be considered either functional or nonfunctional.

![Figure 4](https://example.com/figure4.png)

**Fig. 4.** Chart showing the average colony size of variants from each class. The error bars indicate the maximum and minimum values for colony size in each category. Strain nomenclature is as in figure 2. One colony size unit = 1.58 mm.

<table>
<thead>
<tr>
<th>Number of Mutations</th>
<th>Total Number of Variants Characterized</th>
<th>Valid</th>
<th>Invalid</th>
<th>% Valid</th>
<th>% Invalid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63</td>
<td>30</td>
<td>33</td>
<td>48</td>
<td>52</td>
</tr>
<tr>
<td>2</td>
<td>102</td>
<td>44</td>
<td>58</td>
<td>44</td>
<td>56</td>
</tr>
<tr>
<td>3</td>
<td>108</td>
<td>31</td>
<td>77</td>
<td>29</td>
<td>71</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>18</td>
<td>63</td>
<td>22</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>3</td>
<td>34</td>
<td>8</td>
<td>92</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 1.** Relation between the Number of Mutations and the Validity of the Variants.
Immediately adjoining bases might also be a special case because as neighboring bases frequently interact by base stacking. Twenty immediately adjoining double variants were examined here. Ten were valid and 10 invalid. The properties of 17 of the 20 were successfully predicted from the behavior of the individual point changes. The three that were not predicted were all expected to be invalid but were in fact valid.

Base Conservation
It is usually assumed that RNA sequence positions that are hypervariable in a data set of related sequences will likely accept most variants. Previously, 75 Vibrio SS rRNA sequences were compared (Szymanski et al. 2002) and a number of hypervariable positions were identified. Six of these, table 3, all of which occur in single-stranded regions of the secondary structure, were examined in the current study. All but one of the 15 changes made at these six positions resulted in a viable variant. Thus, the remaining changes at single-stranded positions were viable in 45.7% of the cases.

Another way to obtain insight from other Vibrio sequences is to examine the validity of variants that arose independently in at least two Vibrio lineages. The expectation in this case is that the mutation having arise multiple times and been accepted is likely to be generally acceptable in sequences that are in the Vibrio family. This was examined previously and in fact proved to be the case as such sequence changes were in fact found to have high probability of being accepted in the V. proteolyticus SS rRNA context (Zhang et al. 2003). The obvious next question is whether mutations uniquely found in a single Vibrio lineage would have a high expectation of being viable in the V. proteolyticus context? Apparently not, as seven such sequence changes were examined here with three being valid (C12U, A25U, and G67A) and four being invalid (G18A, U35A, C90A, and C108U). However, it should be noted that unique changes also may be the result of sequencing error.

Effect of Multiple Mutations on Cell Growth Rate
Of the 401 unique variants characterized in the present study, 31.4% (126 of 401) were found to likely be valid. As expected, the number of likely valid SS rRNA variants was found to decrease as the number of mutations in the variant increased. In particular, it was observed that variants with three or more mutations largely tend to be nonfunctional (table 1). However, the number of likely valid mutations actually decreased more slowly than one would expect from a simple extrapolation of the one event results, which would have predicted only 23% of the double mutants and 11% of the triple mutants to be functional. Nevertheless, if less than optimal, but valid, variants accumulate, this might lead to a gradual loss of validity. A specific example is AJ2 (U22A, U87C, U88C, U89C) in which all four point changes are individually valid, but when all are present, the doubling time falls into the gray zone with an average of 33 min.

Fig. 5. Vibrio proteolyticus SS rRNA secondary structure.

nonstandard pairs (non-Watson–Crick, NWC). These secondary structure features are verified by crystallographic studies of the bacterial 50S ribosomal subunit, which includes the SS rRNA molecule (Ban et al. 2000; Schuwirth et al. 2005). In the case of the 120 nucleotide V. proteolyticus SS rRNA, 80 positions are involved in base pairing, 10 of which involve NWC pairs. The secondary structure of V. proteolyticus SS rRNA is shown in figure 5.

One might suspect that point mutations that change one of the bases of a pair would typically be deleterious, but this is in fact not necessarily the case. As table 2 shows the overall effect of changes in the duplex regions is similar to what is seen in other regions of the molecule. In each case, approximately 60% of the changes are actually viable. A closer examination, however, shows that changes in the interior of long helices (helices I and IV) that do not contain bulges or NWC interactions, for example, helices I and IV, are far more likely to be viable. In fact, 26 of the 29 changes in these two regions were viable. When taking this into account in the remaining regions of secondary structure, 47.6% of the mutations gave viable phenotypes.
Truth Tables

In the case of multichange variants, one can hypothesize that most changes are accepted or rejected independently of what is happening at other positions except when two positions are associated by base pairing. If this is true, one could use knowledge about the phenotype of individual point mutations to create a “truth table” for multichange variants. For example, a truth table for a double mutant should have the following structure: valid/valid = valid; invalid/valid = invalid, valid/invalid = invalid, and invalid/invalid = invalid. For larger numbers of mutations if any one mutation is deleterious, then the variant is expected to be invalid. In the case of positions thought to be associated with a base pair, one might expect invalid/invalid, invalid/valid, and valid/invalid to all produce valid if the pairing is recreated. Truth tables of this type were constructed for each of the 28 experiments. The results for experiment AI are shown in figure 6. The remaining truth tables are provided as supplementary table 4 (Supplementary Material online).

Of the 337 multiple change variants that were analyzed using the truth tables, 303 exhibited the predicted phenotype. In 14 cases it was not possible to unambiguously predict a phenotype because the separate behavior of one or more of the changes could not be deduced. The overall success of prediction was 93.8%. Of the failed predictions, 13 were predicted to be invalid but were actually valid, whereas seven expected to be valid were actually invalid (table 4).

Discussion

The most common experimental approach to rapidly explore RNA sequence spaces uses in vitro selection methods such as systematic evolution of ligands by exponential enrichment (SELEX) (Tuerk and Gold 1990). Although these methods have considerable practical utility, they are not the perfect tools to determine the biological complexities of an RNA sequence space. First of all, they are not the perfect tools to determine the biological complexity of an RNA sequence space. Moreover, there is no assurance that a particular sequence has actually been tested. By taking advantage of the relation between colony size and variant viability shown here, one can eliminate the latter problem because one can obtain a good sampling of both valid and invalid variants. These can then be subjected to detailed study.

The in vivo approach used here also has the advantage of considering the variants in a realistic cellular context, albeit E. coli. Although a Vibrio context would in principle be better, it is not a serious limitation. Earlier studies showed that ten different Vibrio wild-type sequences were entering both ribosomes and polysomes in E. coli ribosomes with typically no ill effect on growth rate. In fact, many of them actually improved fitness slightly (Lee et al. 1993). Transcrip- tome studies revealed no significant changes in gene expression when the V. proteolyticus wild-type 5S rRNA was being actively expressed in E. coli. The only significant difference in expression involved a gene associated with degradation of nucleic acids, a possible response to the presence of excess 5S rRNA (Tucker et al. 2005). In addition, the Vibrio 5S rRNA conserves all the known secondary structural features of E. coli 5S rRNA.

The present studies represent considerable progress in predicting the structure of real RNA sequence spaces. The reader will perhaps find the most surprising result is that individual changes in residues involved in secondary structure are not significantly more likely to be invalid than changes in the single-stranded regions, table 2. This likely reflects the fact that the stability of the various secondary structure features is such that the first mismatch mutation need not always be deleterious. This is clearly true in the interior of extended stems. It is noteworthy in this regard that 6 of the 7 variants that were expected to be valid, but actually were not, had two or more individually valid changes in a stem region. In contrast, many of the single-stranded positions are actually involved in tertiary

<table>
<thead>
<tr>
<th>Position</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>U</th>
<th>Viable Variants</th>
<th>Nonfunctional Variants</th>
<th>Not Determined</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>5</td>
<td>2</td>
<td>58</td>
<td>10</td>
<td>A,U,G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>4</td>
<td>1</td>
<td>16</td>
<td>54</td>
<td>C,G</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>32</td>
<td>29</td>
<td>13</td>
<td>A,C,U</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>87</td>
<td>8</td>
<td>3</td>
<td>14</td>
<td>50</td>
<td>C,G</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>1</td>
<td>32</td>
<td>29</td>
<td>13</td>
<td>A,C</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>31</td>
<td>9</td>
<td>2</td>
<td>33</td>
<td>C,U</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>
interactions or interactions with ribosomal proteins, which is likely why many of them resist change.

From the perspective of information obtainable from sequence comparisons using moderately to closely related sequences of the same molecule, there are now several useful rules for inferring the likely validity of variants without resorting to direct experimental studies. First, any changes at hypervariable positions are likely to be valid as demonstrated here, Table 3. The reason that bases at certain positions are highly tolerant to change is simply that they are not involved in the secondary and tertiary structure interactions that largely define the molecule's structure. Second, as previously shown, changes that have occurred independently multiple times among closely related sequences are likely to be valid in all the related sequences (Zhang et al. 2003). Third, as found here, any single change in the interior of helical stems devoid of complicating factors such as NWC interactions and bulge bases is very likely to be valid. This likely reflects the fact that the overall stability of the helical element is such that its structural integrity can be maintained in the presence of a mismatch. Finally, although not addressed by the mutation set described here, when a standard base pair is broken by a mutation, the expectation is that a double mutation that restores the pairing will likely be valid even if one or both of the single mutations are invalid. However, there are likely to be exceptions, as replacement of A19U64 with C19G64 has previously been shown to result in a variant that does not incorporate into ribosomes and therefore was considered to be invalid (Lee et al. 1997). Thus, in the future, it will be important to develop a mutation data set with a large number of changes of this type to properly test this argument.

The single most significant result of the present effort relates to the ability to predict the phenotype of multichange variants from the individual change data. It is well known in evolutionary discussions that a usually innocuous change may have unexpected consequences when an apparently unrelated change elsewhere occurs. This is certainly the case, but how common is this? Possibly for the first time, the results presented here begin to provide some quantitative insight to the frequency of this phenomenon. Given that 93.8% of the predictions made by the truth tables were successful, it is clear that this type of unexpected interdependence is rare. To the extent they exist, unexpected interdependence is likely found among the 13 variants that are expected to be invalid but are actually valid. Future work to better understand why these variants are actually valid will be helpful. From the studies presented here, it is apparent that the number of random mutations has a direct correlation with the validity of the sequence. This effect can be explained in terms of the cumulative effect of mutations on the 5S rRNA structure. With large

Table 4. Summary of Truth Table Results.

<table>
<thead>
<tr>
<th>Total Number of Multiple Variants Analyzed Using Truth Table</th>
<th>337</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variants in which phenotype was determined by one or more undefined bases</td>
<td>14</td>
</tr>
<tr>
<td>Total number of variants in which phenotype predictions could be made</td>
<td>323</td>
</tr>
<tr>
<td>Variants expected to be functional but are nonfunctional</td>
<td>7</td>
</tr>
<tr>
<td>Variants expected to be nonfunctional but are functional</td>
<td>13</td>
</tr>
<tr>
<td>Total number of variants that showed expected phenotype</td>
<td>303</td>
</tr>
</tbody>
</table>
numbers of changes, the structure is most likely to be altered to an extent that it can no longer perform the normal functions of the molecule. Thus, as a result more than five changes will almost always result in a nonfunctional variant. In order to populate a significant region of RNA structure space, it therefore will be useful to utilize multiple known \( \text{Vibrio} \) sequences as starting points rather than just one.

In summary, it is shown here that unexpected sequence interdependencies are at best rare. As a result, it is likely that it will be possible to utilize point mutation data to make realistic enumerations of the structure space in the immediate vicinity of any \( \text{Vibrio} \) 5S rRNA and ultimately other RNAs as well. Toward this end, after examining the 3D structure, natural sequence variation, and the effect of mutations at specific sequence positions, several heuristic rules have been deduced that will help in predicting the membership of specific sequences in the 5S rRNA structure space. The major advantage of having rules that allow accurate prediction of validity and invalidity is that it will ultimately allow the development of computational tools to that can produce a detailed and realistic characterization of structure spaces encompassing millions of variants, which is likely to remain impossible with an experimental approach. Such tools might employ learning algorithms that can be tested by holding out from the analysis some of the data or collecting new data.

**Supplementary Material**

Supplementary tables 1–4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

This work was funded in part by grants to G.E.F. from the Georgia Institute of Technology Center for Ribosomal Evolution and Adaptation and the Center for Bionanotechnology and Environmental Research at Texas Southern University (NASA Cooperative Agreement NNX08B4A47A) and the Robert A. Welch Foundation (E-1451).

**References**


