Distribution of Hammerhead and Hammerhead-like RNA Motifs Through the GenBank

Gerardo Ferbeyre, Véronique Bourdeau, Marie Pageau, et al.

*Genome Res.* 2000 10: 1011-1019
Access the most recent version at doi:10.1101/gr.10.7.1011

---

**References**

This article cites 26 articles, 13 of which can be accessed free at:
http://genome.cshlp.org/content/10/7/1011.full.html#ref-list-1

---

**Creative Commons License**

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

---

**Email Alerting Service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.
Distribution of Hammerhead and Hammerhead-like RNA Motifs Through the GenBank

Gerardo Ferbeyre,1,4,6 Véronique Bourdeau,2,4 Marie Pageau,2 Pedro Miramontes,3 and Robert Cedergren2,5

1Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724 USA; 2Département de Biochimie, Université de Montréal, Montréal, Québec, Canada H3C 3J7; 3Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, México

Hammerhead ribozymes previously were found in satellite RNAs from plant viroids and in repetitive DNA from certain species of newts and schistosomes. To determine if this catalytic RNA motif has a wider distribution, we decided to scrutinize the GenBank database for RNAs that contain hammerhead or hammerhead-like motifs. The search shows a widespread distribution of this kind of RNA motif in different sequences suggesting that they might have a more general role in RNA biology. The frequency of the hammerhead motif is half of that expected from a random distribution, but this fact comes from the low CpG representation in vertebrate sequences and the bias of the GenBank for those sequences. Intriguing motifs include those found in several families of repetitive sequences, in the satellite RNA from the carrot red leaf luteovirus, in plant viruses like the spinach latent virus and the elm mottle virus, in animal viruses like the hepatitis E virus and the caprine encephalitis virus, and in mRNAs such as those coding for cytochrome P450 oxidoreductase in the rat and the hamster.

The hammerhead ribozyme originally was discovered as a self-cleaving motif in viroids and satellite RNAs. These RNAs replicate using the rolling circle mechanism, which generates long multimeric replication intermediates. They use the cleavage reaction to resolve the multimeric intermediates into monomeric forms. The region able to self-cleave has three base paired helices (I–III) connected by two conserved single stranded regions and a bulged nucleotide (Forster and Symons 1987; for reviews see Symons 1992; Bratty et al. 1993; Birikh et al. 1997). The hammerhead ribozyme also seems to function in the generation of unit length sequences from multimeric transcripts of repetitive DNA sequences. Two of these RNAs have been characterized: one in several newt species (Epstein and Gall 1987) and the other one in three Schistosome species (Ferbeyre et al. 1998). Among the repetitive sequences of these two organisms, note that not all contained a bona fide hammerhead ribozyme. Indeed, many mutations also were found creating variants of the original motif. Overall, the rather limited distribution of this motif contrast with the simplicity of its secondary structure in which only a core of 14 nucleotides is absolutely required for cleavage.

We recently have conducted an extensive research of different RNA motifs in the GeneBank database (Bourdeau et al. 1999). The results showed that most of the motifs were distributed randomly among gene sequences suggesting that most RNA motifs originate by random drift. We now wish to extend these observations to the self-cleaving hammerhead ribozyme and its variants in which either an essential nucleotide in the single strand positions is allowed to be random or the identity of a conserved base pair from helices II and III is changed. We found that most of the hammerhead motifs are apparently underrepresented among gene sequences, but this comes from the bias of the GenBank for sequences with low CpG representation. We also report the finding of intriguing motifs in several repetitive sequences and mRNAs.

RESULTS

Searching for Self-cleaving RNA Motifs of the Hammerhead Type in the GenBank

The hammerhead ribozyme can be described by three helices separated by three single stranded regions of conserved nucleotides. There are three equivalent conformations of the self-cleaving hammerhead depending on which helix bears the 5’ and 3’ end of the motif. We named them HH-I, HH-II, and HH-III (Figure 1). The descriptors composed as input for the search program are presented beside each motif and described in the legend of Figure 1 (see also Methods). They were designed to detect any sequence with all the minimal nucleotide requirements to have some catalytic activity and with the possibility to fold like the hammerhead. In this context, it is expected that sequences will be found that combine several nonoptimum features

*These authors contributed equally to this work.
1Deceased.
6Corresponding author.
E-MAIL ferbeyre@cshl.org; FAX (516) 367 8454.
and be inactive for this reason, i.e., a non-GUC cleavage, a C in position 4, short helices, and long loops. It is also possible that they contain all the requirements for being catalytically active but the active confor-
mation is inaccessible because the RNA molecule that bears them folds into an alternative secondary structure.

The search for hammerhead self-cleaving motifs through the GenBank database (Benson et al. 1999) was performed using the program RNAMOT (Gautheret et al. 1990; Laferrière et al. 1994). The sequences detected with our descriptors are referred to as occurrences. The ability of the descriptors to identify the hammerhead motifs already characterized is illustrated in Table I. The program recognizes most of the known plant derived hammerheads (Symons 1997; see also http://callisto.si.usherb.ca/~jpperra/organisms.html; Bussière et al. 1996; Lafontaine et al. 1999) and all those present in satellite DNA sequences. Note that there is no known natural incidence of a hammerhead of the HH-II type.

Table 2 presents the frequencies of occurrences of potential hammerhead motifs in the different sections of the GenBank as well as the expected frequencies calculated from the number of occurrences obtained in a database of random sequences. In general, the number of occurrences observed are half of the frequency expected if our motifs were randomly distributed among the sequences of the GenBank. HH-I and HH-II detect twice as many motifs as HH-III because we designed the motifs in a way that Helix III had a 2–base pair requirement in HH-I and HH-II descriptors versus 3 base pairs in the HH-III descriptor (see Methods). This increase was predicted by the number of occurrences obtained in the random database.

The Frequency of Mutated Versions of the Hammerhead Self-cleaving RNAs

We also composed descriptors for variants of the hammerhead ribozyme motif. Substitutions were made by replacing, one at a time, each of the essential nucleotides located in the single stranded regions of the ribozyme core by N (boldface in Fig. 1) or by changing the

Figure 1  Structures and descriptors of the hammerhead self-cleaving ribozyme motifs. The three descriptors, HH-I, HH-II, and HH-III, are defined by which helix is at the 5’ end and named according to the helix number (Hertel et al. 1992). Each descriptor is composed of single stranded (s) and double stranded (H) regions. The regions first are named in order from 5’ to 3’ and then specified for their length (minimum:maximum), number of mismatches (in the case of H only), and presence of specific nucleotides. For example, HH-I consists of the following features: H1 s1 H2 s2 H2 s3 H3 s4 H3 s5 H1 where H1 is an helix of a fixed length of three base pairs with no mismatches and no specific nucleotides; H2 is also of three base pairs with no mismatches but with a starting G-C base pair; H3 is an helix of 2 base pairs beginning with an A-U base pair; s1 is a single stranded region of seven nucleotides exactly with a specific sequence; s2 varies between 0 and 100 undetermined nucleotides; and so on. The hammerhead-like motifs are the same as the three shown but with an “N” replacing one of the nucleotides in boldface or with a different identity of one of the base pairs in boldface. These motifs are named according to the original motif and the position of the mutation, e.g., HH-I-3 motif is as HH-I but with an N instead of a C at position 3; thus, HH-I-3 descriptor has a modified s1 as follows: s1 7-7 NYGANGA, similarly with HH-I-iAU, which is a HH-I motif with a A:U base pair in the Helix II instead of a G:C; thus, the descriptor HH-I-iAU has this particularity: H2 3:3 0 ANN:NNU. The cleaving site is after H17. (H) A, C, or U; (N) A, C, G, or U; (Y) C or U. See Methods for the basis of the sequence requirements.

Table 1. Known Hammerhead Motifs Identified in Our Search

<table>
<thead>
<tr>
<th>HH-I</th>
<th>HH-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avocado sunblotch viroid</td>
<td>Schistosoma mansoni DNA for repeated sequences</td>
</tr>
<tr>
<td>Tobacco Ringspot virus satellite RNA</td>
<td>Barley yellow virus satellite RNA</td>
</tr>
<tr>
<td>Ambystoma talpoideum satellite 2</td>
<td>Chrysanthemum chlorotic mottle viroid</td>
</tr>
<tr>
<td>Cryptobranchus alleganiensis satellite 2</td>
<td>Cherry small circular viroid-like RNA</td>
</tr>
<tr>
<td>Cyplops pyrgodactylus satellite 2</td>
<td>Lucerne transient streak virus RNA 2</td>
</tr>
<tr>
<td>Eurycea longicauda satellite 2</td>
<td>Peach latent mosaic viroid</td>
</tr>
<tr>
<td>Pheidole glutinosus satellite 2</td>
<td>Subterranean clover mottle virus satellite RNA</td>
</tr>
</tbody>
</table>
identity of each one of the 2 conserved base pairs of the hammerhead motif (also boldface in Fig. 1).

Table 2 presents the data on the distribution of the mutated variants of HH-I, HH-II, and HH-III from the single stranded region. It is expected that every mutant will increase the frequency of occurrences by a factor of four because we changed the requirements in every position from only one to all four nucleotides except in position 4 where C and U already were allowed and in the cleavage site where only G originally was excluded. Thus, in position 4 we expected to double the frequency, and in the cleavage site we expected a 25% increase. The results are mostly those anticipated based on these calculations. However the mutants of position 12 doubled the expected increase in all the orientations. This effect was not uniformly observed in the different subdivisions of the GenBank. Actually, most of the extra occurrences are located in the files containing ESTs (Expressed Sequence Tags) and mammalian sequences. These preferences were not observed in the random database in which the mutants showed the anticipated increase in their frequency in comparison with the original motif. The number of occurrences obtained in the virus section of the GenBank for the HH-III-8 variant was 722 instead of the 113 expected (HH-III has 3774 expected occurrences and viruses represent 3% of the GenBank). However, a quick analysis of the occurrences obtained with this descriptor revealed that most of them are the same motif repeated in 679 hepatitis C sequences.

Table 3 presents the frequencies obtained with the mutant hammerhead ribozymes using a different identity for the conserved base pair of helices II or III (positions 10.1:11.1 and 15.1:16.1, respectively). One striking observation is that all the mutants in Helix II (iiNN) have total occurrences two to six times higher than expected whereas the mutants in Helix III (iiiNN) have half the expected frequency. One more interesting point is the high number of occurrences obtained with the three orientations of the hammerhead ribozyme having a A:U base pair in Helix II (10.1:11.1) instead of the usual G:C.

The mutants in position 12 and the mutants of the conserved base pair of Helix II have in common that they disrupt the presence of a dinucleotide CpG in the resulting sequence. It is well known that CpG is underrepresented in vertebrate sequences (Karlin and Mrazek 1997). The GenBank is biased for those sequences mainly owing to human and rodent entries. In those files, the mutants that disrupt the CpG requirement have a higher frequency. To confirm that the overall frequency of the hammerhead motifs containing CpG dinucleotides is half of the expected one because of the low CpG content of vertebrate sequences, we built a new random database in which the frequency of CpG was reduced by half in favor of either AG, CA, CC, CT, GG, or TG to simulate the frequencies observed by Karlin and Mrazek (1997; see Methods). In this database, we observed an overall doubling of the original expected frequencies for all the motifs needing a CpG but not for the others (data not shown).

Still, the mutants with a A:U base pair in position 10.1:11.1 of the Helix II have a very high frequency in all three conformations of the motif: two to three times higher than expected even considering the CpG effect discussed above. So far, we have no explanation for this intriguing observation.

Finally, we made three more searches by changing the cleavage site from NUH to NIH based on the report of Kore et al. (1998) that such hammerheads were still active. We obtained for these new mutants a number of occurrences corresponding to half of what we expected according to the search in an equal A-C-G-T random database. Moreover, as for the previous motifs, the number of occurrences in the GenBank is comparable to the expected frequency according to the search in the reduced for CpG database. All the occurrences found in the GenBank are available in our web site at http://www.centrcn.umontreal.ca/~bourdeav/HH.

Some Intriguing Hammerhead Motifs that Might Have Functional Significance

This section presents a sample of motifs considered interesting either because of their location or because their structure is optimal for self cleavage. The hammerhead ribozyme occurs naturally in satellite RNAs, viroids, and transcripts from repetitive sequences. The probability of finding an active hammerhead should be higher among these genetic elements. Several potential hammerhead motifs were found in distinct families of repetitive DNA.

Hammerhead ribozymes were found in the satellite DNA from Dolichopoda schiavazzii (cricket) by using the HH-I descriptor (example in Figure 2A). Fourteen have a conserved HH-I motif and two have a HH-I-iiGU motif (G:U in position 10.1:11.1 instead of G:C). This ribozyme cleaves after CUA (A.A. Rojas, A. Vazquez-Tello, G. Ferbeyre, F. Venanzetti, L. Bachmann, B. Paquin, and R. Cedergren, in prep.). Helix I has the Gg:CC base pairs and the internal loop common to the hammerhead motifs in schistosomes (Ferbeyre et al. 1998) and newts (Pabon-Peña et al. 1991). It is noteworthy that among the 20 similar sequences submitted to GenBank, the four sequences not found through the search contained either mismatches in one of the helices or combined two point mutations.

A hammerhead-like motif was detected in the Kpn-13 family of human repetitive DNA by using the descriptor HH-I-4 (Fig. 2B). The motif is found in several ESTs containing Kpn-repetitive sequences (also known as L1-repetitive elements) indicating its expression at the RNA level. All the occurrences contain a
Table 2. Distribution of Hammerhead and Hammerhead-like Motifs in the Different Sections of the GenBank: Mutants of the Single Stranded Regions

<table>
<thead>
<tr>
<th>GenBank relative size</th>
<th>Total over expected</th>
<th>Expected (A = C = G = T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH-I</td>
<td>0.56</td>
<td>1635</td>
</tr>
<tr>
<td>HH-3</td>
<td>0.57</td>
<td>1780</td>
</tr>
<tr>
<td>HH-4</td>
<td>0.53</td>
<td>1798</td>
</tr>
<tr>
<td>HH-5</td>
<td>0.46</td>
<td>6488</td>
</tr>
<tr>
<td>HH-6</td>
<td>0.49</td>
<td>6973</td>
</tr>
<tr>
<td>HH-7</td>
<td>0.46</td>
<td>6883</td>
</tr>
<tr>
<td>HH-8</td>
<td>0.57</td>
<td>7081</td>
</tr>
<tr>
<td>HH-9</td>
<td>0.50</td>
<td>8831</td>
</tr>
<tr>
<td>HH-10</td>
<td>0.54</td>
<td>9111</td>
</tr>
<tr>
<td>HH-11</td>
<td>0.54</td>
<td>9171</td>
</tr>
<tr>
<td>HH-12</td>
<td>0.54</td>
<td>9171</td>
</tr>
<tr>
<td>HH-13</td>
<td>0.54</td>
<td>9171</td>
</tr>
<tr>
<td>HH-14</td>
<td>0.54</td>
<td>9171</td>
</tr>
<tr>
<td>HH-15</td>
<td>0.54</td>
<td>9171</td>
</tr>
<tr>
<td>HH-16</td>
<td>0.54</td>
<td>9171</td>
</tr>
</tbody>
</table>

The motifs are named as explained in Methods and in Figure 1. The “expected” number of occurrences was obtained by searching the different motifs in a database of 100,000 random sequences (equal representations of A, C, G, and T) and correcting the frequency to the relative size of the GenBank. The “total over expected” shows the ratio of occurrences obtained in the GenBank versus the expected ones according to the search performed in a random database with the size of the GenBank.

 pri) Primate sequence entries (from the two GenBank files); (rod) rodent sequence entries; (mam) other mammalian sequence entries; (vrt) other vertebrate sequence entries; (inv) invertebrate sequence entries; (pln) plant sequence entries (including fungi and algae); (bct) bacterial sequence entries; (rna) structural RNA sequence entries; (vrl) viral sequence entries; (phg) phage sequence entries; (syn) synthetic and chimeric sequence entries; (una) unannotated sequence entries; (est) EST (expressed sequence tag) (from 23 GenBank files); (pat) patent sequence entries; (sts) STS (sequence tagged site) sequence entries; (gss) GSS (genome survey sequence) sequence entries; (htg) HTGS (high throughput genomic sequencing) sequence entries.
<table>
<thead>
<tr>
<th>Section</th>
<th>pri</th>
<th>rod</th>
<th>mam</th>
<th>vrt</th>
<th>inv</th>
<th>pln</th>
<th>bct</th>
<th>rna</th>
<th>vrl</th>
<th>phg</th>
<th>syn</th>
<th>una</th>
<th>est</th>
<th>pat</th>
<th>sts</th>
<th>gss</th>
<th>htg</th>
<th>Total</th>
<th>over expected (A = C = G = T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH-I</td>
<td>51</td>
<td>21</td>
<td>0</td>
<td>14</td>
<td>176</td>
<td>119</td>
<td>130</td>
<td>0</td>
<td>37</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>184</td>
<td>45</td>
<td>4</td>
<td>17</td>
<td>108</td>
<td>916</td>
<td>0.56</td>
</tr>
<tr>
<td>HH-I-AU</td>
<td>569</td>
<td>100</td>
<td>30</td>
<td>36</td>
<td>822</td>
<td>545</td>
<td>426</td>
<td>3</td>
<td>58</td>
<td>7</td>
<td>8</td>
<td>0</td>
<td>1113</td>
<td>93</td>
<td>32</td>
<td>341</td>
<td>544</td>
<td>5475</td>
<td>2.56</td>
</tr>
<tr>
<td>HH-I-CG</td>
<td>343</td>
<td>59</td>
<td>12</td>
<td>20</td>
<td>203</td>
<td>117</td>
<td>512</td>
<td>1</td>
<td>24</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>669</td>
<td>69</td>
<td>15</td>
<td>66</td>
<td>221</td>
<td>2355</td>
<td>1.27</td>
</tr>
<tr>
<td>HH-I-GU</td>
<td>452</td>
<td>65</td>
<td>9</td>
<td>38</td>
<td>361</td>
<td>227</td>
<td>203</td>
<td>0</td>
<td>73</td>
<td>3</td>
<td>49</td>
<td>2</td>
<td>794</td>
<td>58</td>
<td>15</td>
<td>93</td>
<td>332</td>
<td>274</td>
<td>1.77</td>
</tr>
<tr>
<td>HH-II</td>
<td>606</td>
<td>124</td>
<td>21</td>
<td>82</td>
<td>406</td>
<td>364</td>
<td>260</td>
<td>2</td>
<td>165</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>941</td>
<td>43</td>
<td>15</td>
<td>175</td>
<td>436</td>
<td>3652</td>
<td>2.03</td>
</tr>
<tr>
<td>HH-II-AU</td>
<td>555</td>
<td>155</td>
<td>30</td>
<td>36</td>
<td>862</td>
<td>545</td>
<td>426</td>
<td>3</td>
<td>56</td>
<td>6</td>
<td>24</td>
<td>4</td>
<td>454</td>
<td>97</td>
<td>4</td>
<td>17</td>
<td>108</td>
<td>916</td>
<td>0.56</td>
</tr>
<tr>
<td>HH-II-CG</td>
<td>344</td>
<td>59</td>
<td>12</td>
<td>20</td>
<td>203</td>
<td>117</td>
<td>512</td>
<td>1</td>
<td>24</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>669</td>
<td>69</td>
<td>15</td>
<td>66</td>
<td>221</td>
<td>2355</td>
<td>1.27</td>
</tr>
<tr>
<td>HH-II-GU</td>
<td>452</td>
<td>65</td>
<td>9</td>
<td>38</td>
<td>361</td>
<td>227</td>
<td>203</td>
<td>0</td>
<td>73</td>
<td>3</td>
<td>49</td>
<td>2</td>
<td>794</td>
<td>58</td>
<td>15</td>
<td>93</td>
<td>332</td>
<td>274</td>
<td>1.77</td>
</tr>
<tr>
<td>HH-III</td>
<td>42</td>
<td>6</td>
<td>0</td>
<td>9</td>
<td>96</td>
<td>67</td>
<td>65</td>
<td>0</td>
<td>21</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>96</td>
<td>57</td>
<td>3</td>
<td>8</td>
<td>64</td>
<td>527</td>
<td>0.55</td>
</tr>
<tr>
<td>HH-III-AU</td>
<td>360</td>
<td>59</td>
<td>17</td>
<td>11</td>
<td>482</td>
<td>324</td>
<td>241</td>
<td>0</td>
<td>35</td>
<td>7</td>
<td>11</td>
<td>1</td>
<td>723</td>
<td>31</td>
<td>6</td>
<td>354</td>
<td>328</td>
<td>2900</td>
<td>5.05</td>
</tr>
<tr>
<td>HH-III-CG</td>
<td>203</td>
<td>37</td>
<td>4</td>
<td>9</td>
<td>98</td>
<td>58</td>
<td>101</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>469</td>
<td>8</td>
<td>6</td>
<td>42</td>
<td>132</td>
<td>1176</td>
<td>1.17</td>
</tr>
<tr>
<td>HH-III-GU</td>
<td>251</td>
<td>29</td>
<td>3</td>
<td>19</td>
<td>194</td>
<td>126</td>
<td>121</td>
<td>0</td>
<td>43</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>440</td>
<td>40</td>
<td>13</td>
<td>52</td>
<td>217</td>
<td>1552</td>
<td>1.73</td>
</tr>
<tr>
<td>HH-IIIUG</td>
<td>333</td>
<td>48</td>
<td>10</td>
<td>23</td>
<td>246</td>
<td>224</td>
<td>153</td>
<td>13</td>
<td>39</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>486</td>
<td>22</td>
<td>5</td>
<td>70</td>
<td>257</td>
<td>1930</td>
<td>2.24</td>
</tr>
<tr>
<td>HH-III-AUG</td>
<td>285</td>
<td>56</td>
<td>11</td>
<td>31</td>
<td>212</td>
<td>139</td>
<td>113</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>399</td>
<td>27</td>
<td>8</td>
<td>63</td>
<td>148</td>
<td>1522</td>
<td>2.02</td>
</tr>
<tr>
<td>HH-III-GCG</td>
<td>23</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>22</td>
<td>108</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>158</td>
<td>52</td>
<td>1</td>
<td>7</td>
<td>19</td>
<td>458</td>
<td>0.50</td>
</tr>
<tr>
<td>HH-III-GUC</td>
<td>78</td>
<td>14</td>
<td>24</td>
<td>15</td>
<td>38</td>
<td>66</td>
<td>100</td>
<td>0</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>173</td>
<td>27</td>
<td>2</td>
<td>18</td>
<td>38</td>
<td>608</td>
<td>0.57</td>
</tr>
<tr>
<td>HH-III-GUG</td>
<td>52</td>
<td>18</td>
<td>10</td>
<td>13</td>
<td>47</td>
<td>51</td>
<td>83</td>
<td>1</td>
<td>18</td>
<td>1</td>
<td>11</td>
<td>1</td>
<td>235</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>43</td>
<td>604</td>
<td>0.53</td>
</tr>
<tr>
<td>GenBank relative size</td>
<td>0.14</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>&lt;0.01</td>
<td>0.03</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.37</td>
<td>0.02</td>
<td>0.01</td>
<td>0.06</td>
<td>0.10</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

See legend of Table 2 for explanation of abbreviations.
disabling A at position 4, but one (AA564135) possesses a C. The latter motif is inactivated by a G per A substitution at position 12. Variants of this motif also are found in genomic clones containing Kpn repetitive sequences. Intriguingly, the L1 motif interrupting the dystrophin gene of a muscular dystrophy patient (accession number HSU09115) also has a disruption in Helix I. Four additional hammerhead-like motifs were found in the satellite DNA array from the rodent *Microtus chrotorrhinus* (accession number MICSATB, position 921–1079, not shown), in the repetitive DNA from the protozoan parasite *Theileria parva* (accession number S37077, position 84–223, not shown) with the descriptor HH-I-7 and in mouse repetitive DNA with descriptors for the HH-IIii/A and HH-IIIiiAU motifs (Fig. 2C,D). The first two motifs are predicted to be inactive because they contain A instead of G in position 12.

Viruses are good candidates for using catalytic RNA motifs. We have found several new intriguing hammerhead motifs in different viruses (Fig. 2E). Two

![Figure 2](image)

Figure 2  A–L show putative hammerhead motifs.
similar hammerhead ribozyme motifs were found in the 5’ untranslated region of two viruses of the I l ar vi
genus, family of Bromoviridae, which are single stranded positive RNA viruses. One motif is in the spin-
ach latent virus (accession number PMOVRNA3, position 252–331) and the other in the Elm mottle virus
(accession number SLUS7048, position 250–329) (Fig. 2F). Both motifs were found using the HH-III descrip-
tor. The region containing the hammerhead is highly conserved among these viruses. The hammerhead motif
found with HH-II in an RNA associated to carrot red luteovirus that is also very interesting because satellite
RNAs were the first molecules found to contain ham-
merhead ribozymes (Fig. 2F). This motif is predicted to
cleave after AUA. Mammalian viruses also contain po-
tential hammerhead ribozymes, and two of them found with HH-II are illustrated in Figure 2G,H, one in
the hepatitis E virus, and the other in the caprine en-
cephalitis virus.

Two hammerhead motifs in human mRNAs also
are presented in Figure 2J. Self-cleaving motifs in mrNA decay might regulate gene expression by promoting
RNA decay. The genes coding for the interferon-
induced DAP1 and the neuroleukin gene possess po-
tentially active hammerhead motifs found with HH-III
that are predicted to cleave after UUC and CUC, re-
spectively. Perhaps even more remarkable are the con-
served hammerhead motifs found in the genes coding
for NADPH-cytochrome P450 oxidoreductase both in
the rat and the hamster (Fig. 2K,L). All together, the
motifs presented here suggest that the hammerhead
ribozyme might have functions other than those pre-
viously suggested for satellite RNA and transcripts for
repetitive sequences.

DISCUSSION

We have used the search engine RNAMOT to scrutinize
the GenBank for potential self-cleaving hammerhead
ribozyme motifs. Our search extends earlier efforts
to find a subset of potential hammerheads in Escherichia
coli sequences (Ruffner et al. 1990). Because this motif
has relatively few structural constraints, we designed
an extensive set of descriptors for both the wild-type
motif and variants of its essential nucleotides. The re-
results show a wide distribution of potential hammer-
head-like motifs in all regions of the GenBank with a
higher frequency for the variants that do not require
the presence of a CpG dinucleotide in the final se-
quence of the motifs. This CpG dinucleotide in posi-
tions 11.1 and 12 is not absolutely required for self-
cleavage because other base pairs are acceptable in po-
sitions 10.1:11.1. We conclude that the reduction we
observed in the frequency of most hammerhead motifs
in this search is fortuitous.

We expect that most of the motifs found here are
inactive because we designed descriptors that include
mutations or nonoptimal features of the hammerhead
self-cleaving motif (Ruffner et al. 1990). However, our
results illustrate the possibility that natural sequences
might end up forming self-cleaving motifs by random
drift. In other words, it would be sufficient to mutate
one or two residues to activate the potential hammer-
head ribozymes described here. This is not only true for
the hammerhead ribozyme motif because other RNA
motifs can be found randomly in natural sequences
(Fontana et al. 1993; Reidys et al. 1997; Bourdeau et al.
1999).

The use of variants of the hammerhead ribozyme
was stimulated by previous work that showed that sat-
ellite DNA encoding hammerhead ribozymes is en-
riched with mutated variants of the motif (Zhang and
Epstein 1996; Ferbeyre et al. 1998). The ribozyme motif
found in the cricket satellite DNA follows this rule be-
cause 14 of the 20 sequences deposited until now in
the GenBank contains an active motif. Other mutant
hammerheads were found in different families of re-
petitive DNA by using descriptors for hammerhead-like
motifs, raising the possibility that other members of
these families, not yet sequenced, contain the active
motifs. The occurrence of hammerhead ribozymes in
transcripts of repetitive DNA from different species
suggests a functional role for the self-cleavage reaction
in the propagation and/or the metabolism of these
transcripts. We previously have proposed that self-
cleavage might limit the expansion of repetitive se-
quen
ces through the genome by retrotransposition
(Ferbeyre et al. 1998). This model predicts that recent
insertions of these elements will contain disabling mu-
tations in the hammerhead motif. The family of L1
repetitive elements for example contains mutated ver-
sions of the hammerhead and members of this family
still retrotranspose in humans, sometimes causing ge-
netic diseases (Holmes et al. 1994). Another intriguing
possibility is that viroids and satellite RNAs originated
from transcripts of repetitive sequences when these
transcripts parasitizes a viral replication machinery.
Subsequently, they might jump from one organism to
another using the virus as a vector, and as a result their
distribution will cross phylogenetic barriers.

Many ESTs and mRNAs were found here to possess
hammerhead-like motifs. To test any role of the ham-
nerhead motifs identified in this work, we need a com-
bination of biochemical and genetic analysis. Our
group has finished the characterization of hammer-
head motifs in repetitive DNA of Schistosome (Ferb-
beyle et al. 1998) and the cricket (A.A. Rojas, A.
Vazques-Tello, G. Ferbeyre, F. Venanzetti, L. Bach-
mann, B. Paquin, and R. Cedergren, in prep.). All the
occurrences we found in the GenBank are available at
our web site (URL: http://www.centrcn.umontreal.ca/
-bourdeav/HH) for those interested in finding where
“hammers” can cut.
METHODS

The pattern searching for RNA secondary structures was performed by RNAMOT (Gautheret et al. 1990; Laferrérie et al. 1994). The inputs for this program are nucleotide sequences, and a descriptor file defining the structural motif to be searched. RNAMOT reports all the occurrences of the motif as well as its positions along the sequence. Two of the three helices defining the hammerhead self-cleaving motif are closed by loops. The remaining helix connects the motif to the rest of the RNA molecule. As a result, there are three ways of defining a self-cleaving hammerhead ribozyme motif. We have built descriptors for these three different orientations of the motif taking into account the following constraints (Fig. 1):

1. Three nucleotides in Helix I. Helix I has no specific nucleotide requirements although the hammerhead motif found in the newt and in Schistosomes possess a conserved GG:CC base pairing, three nucleotides downstream from the cleavage site as well as an internal loop farther downstream (Fabon-Peña et al. 1991; Ferbeyre et al. 1998).

2. The conserved sequence CYGANGA. This sequence is part of the catalytic core of the ribozyme and is entirely conserved with the exception of position 7. In the latter, although all nucleotides are accepted, the preferred ones are U then G or A and finally C. More recently, position 4 was reported to accept also U, so we have included this feature in our search (Ambros and Flores 1998).

3. Three nucleotides in Helix II. There is a strong preference for a R.Y base pair in positions 10.1:11.1, but the pair G:C confers the better activity and was the only one allowed in our original descriptors.

4. The conserved sequence GAA is absolutely required for catalysis. In the X-ray model of the hammerhead, nucleotides G12 and A13 form two reverse Hoogsteen G-A base pairs with nucleotides A9 and G8, respectively, whereas A14 form a non-Watson Crick base pair with N7 (Scott et al. 1995).

5. Helix III requires an A:U base pair which is also of non Watson Crick type and a minimum of one more pair in two of the orientations (HH-I and HH-II). When the helix is open as in HH-III, two more pairs are required.

6. The cleavage site was defined as NUH (H is any nucleotide but G). However, natural ribozymes contain GUC, GUA, AUA, and AUC because they allow the highest reaction rates (Shimayama et al. 1995; Ferbeyre et al. 1998).

7. The loops closing the helices were allowed to have from 0 to 100 nucleotides.

Sixty-three additional mutants also were included in the study. These were derived from the original motifs shown in section 1734 solely to indicate this fact.

The search was performed in the July 15, 1998 release of GenBank sequence database (National Center for Biotechnology Information-GeneralBank flat file release 108.0). Searches were performed on both strands and all occurrences of motifs involving unidentified bases denoted by N in the database were disregarded. A Power Challenge XL with 32 CPUs IP 19, R4400, 150-MHz processor (3072 Mbytes) running UNIX IRIX 6.2 was used.

ACKNOWLEDGMENTS

We thank Bruno Paquin for valuable comments, and to NSERC of Canada which financed this project. V.B. holds a doctoral fellowship from NSERC of Canada. The late R.C. was Richard Ivey Scholar of the Canadian Institute for Advanced Research (CIAR) program in Evolutionary Biology. We acknowledge previous efforts from Dr. Daniel Gautheret to search for hammerhead sequences with RNAMOT in our laboratory. In addition, we thank Bernard Lorazo, Daniel Raymond and André Fourrier of the DITER (Direction des infrastructures technologiques d’enseignement et de recherche) at the Université de Montréal for their assistance. P.M. wishes to thanks the hospitality of the Université de Montréal and the Institute of Physics, UNAM.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


1018 Genome Research www.genome.org
Looking for Hammerhead Ribozymes in the GenBank


Received December 17, 1999; accepted in revised form May 3, 2000.