Pergamon

Bioorganic & Medicinal Chemistry, Vol. 5, No. 6, pp 1097–1105, 1997

PII: S0968-0896(97)00045-X
0968-0896/97 $17.00 + 0.00

In Vitro Evolution Used to Define a Protein Recognition Site Within a Large RNA Domain

Amalia Sapag† and David E. Draper*
Department of Chemistry, Johns Hopkins University, Baltimore, MD 21218, U.S.A.

Abstract—A minimum of 460 nucleotides of 16S ribosomal RNA are needed to fold the target site for *E. coli* ribosomal protein S4, although a much smaller region within this large domain is protected from chemical reagents by the protein. Starting with a 531-nucleotide rRNA fragment, cycles of mutagenesis, selection with S4, and amplification ('in vitro evolution') were used to obtain a pool of 30 RNA sequences selected for S4 recognition but ~30% different from wild type. Numerous compensatory base pair changes have largely preserved the same secondary structure among these RNAs as found in wild-type sequences. A 20-base deletion and a single nucleotide insertion are among several unusual features found in most of the selected sequences and also prevalent among other prokaryotic rRNAs. Most of the compensatory base changes and selected features are located outside of the region protected by S4 from chemical reagents. It was unexpected that S4 would select for RNA structures throughout such a large domain; the selected features are probably contributing indirectly to S4 recognition by promoting correct tertiary folding of the region actually contacted by S4. The role of S4 may be to stabilize this domain (nearly one-third of the 16S rRNA) in its proper conformation for ribosome function. © 1997 Elsevier Science Ltd.

Introduction

A number of important cellular enzymatic activities are carried out by protein–RNA complexes; examples are RNAse P, telomerase, signal recognition particle, spliceosome and the ribosome.[1] Most of the RNA components of these complexes are known to adopt stable and compactly folded structures in the absence of proteins, and in all the cited examples the RNA is directly responsible for some (if not all) of the substrate recognition and catalytic capacity of the complex. How specific proteins bind these RNAs and modulate their activities are challenging questions, in part because the complexities of folding large (several hundred to several thousand nucleotide) RNA structures are not well understood.

Although the tertiary structures of most large RNAs remain unexplored, their secondary structures can be easily determined by phylogenetic comparisons. In this approach, the sequences of homologous RNAs from many organisms are aligned, and searches are made for pairs of compensatory base changes that preserve canonical base pairing. This is essentially the method used to propose the cloverleaf structure of tRNA: after two sequences became known, it was obvious that only the cloverleaf pairing was possible for both.[2] With ribosomal RNA sequence databases having hundreds to thousands of entries, detailed secondary structure maps of both large and small subunit rRNAs have been prepared, and suggestions of tertiary folding (e.g., loop–loop and triple base interactions) have been obtained.[3] Phylogenetic analysis takes advantage of an experiment, running for the last few billion years, in which RNAs have been randomly mutagenized and selected for a specific function. The same experiment, performed on a laboratory scale, has proved a very powerful method for investigating RNA–protein recognition. Starting from a large pool of RNA sequences generated by random synthesis or by mutagenesis of a specific sequence, those RNAs with specific affinity for a protein can be selected and analyzed, revealing the secondary structure required to form the protein recognition site. A number of proteins recognizing smaller RNA hairpin or internal loop structures (≲30 nt) have been examined this way, including T4 DNA polymerase,4 U1-snRNP-A,[5] R17 coat protein,[6] rho,[7] hnRNP-A1,[8] and Rev.[9–11] A typical pool of 10^{12}–10^{14} sequences will contain many variants of a given 30 nt structure with high affinity for a protein, but as the complexity of a structure recognized by a particular protein increases to encompass hundreds of nucleotides, the probability of finding RNAs with high affinity for the protein decreases rapidly. This limitation has been circumvented in selections for ribozymes by devising a true 'evolution' experiment in which rounds of mutagenesis and selection are alternated. In this way, sequences with very weak activity can be improved by mutagenesis, and highly active ribozymes of ~200 nt have been 'evolved' from random sequence pools.[12]

In the present work we extend this 'in vitro evolution' approach to the recognition between a large ribosomal RNA domain and *E. coli* ribosomal protein S4. Initial studies with nuclease-generated fragments of 16S rRNA (1542 nt) suggested that S4 recognizes about 500 nt near the 5′-terminus.[13,14] This result was confirmed by measurements with nested deletions of 16S rRNAs, which showed that a 460 nt region

†Current address: Laboratorio de Oncología Molecular y Celular, Facultad de Medicina, Pontificia Universidad Católica de Chile, Alameda 340, Santiago, Chile.

(positions 39–501) is essential for recognition.[15] This 5′-domain is shown in Figure 1. As the protein is only 205 amino acids in length, it is unlikely that it contacts a large fraction of this domain, and in fact chemical and enzymatic protection experiments have suggested that only a limited region near a junction of five helices participates in binding (the cluster of S4-induced protections from hydroxyl radical is shown in Fig. 1).[16] Although some hairpins can be deleted from within the domain without any effect on S4 binding, some small deletions and single base changes scattered through the domain adversely affect recognition, and it has not been possible to eliminate large portions of the domain and retain protein recognition (ref 17 and unpublished observations). Though these are negative results, they suggest that the structure of the entire 5′-domain is needed to maintain the recognition site in its correct conformation.

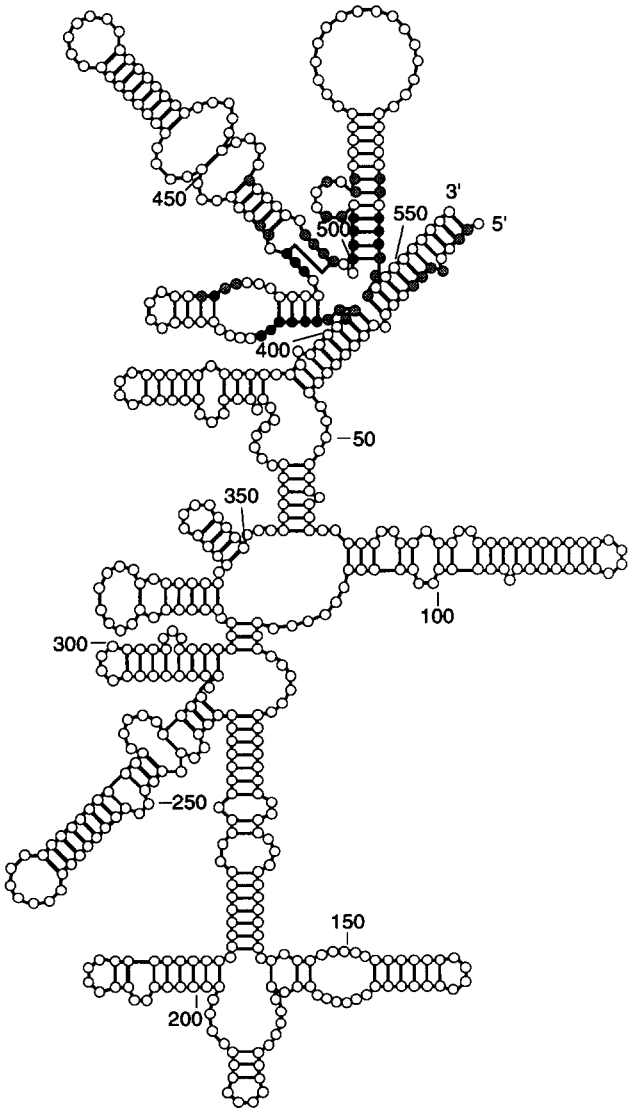

Figure 1. Phylogentically conserved secondary structure of the 5′-domain of 16S ribosomal RNA.[23] Black dots indicate sites strongly protected by S4 from hydoxyl radical reaction; gray dots indicate weakly protected nucleotides.[16]

Using site-directed mutagenesis to further define the minimum RNA structure needed to recognize S4 is feasible but tedious. We therefore looked to in vitro selection experiments to delineate which parts of the RNA secondary structure are essential for binding and which are irrelevant. We find that a remarkably large fraction of the domain secondary structure must remain intact for S4 recognition, and there are a number of intriguing parallels between selected sequences and the phylogenetic data base. Interactions with S4 may have been a strong constraint on the 5′-domain structure during evolution.

Materials and Methods

PCR reactions

Polymerase chain reactions were done essentially as described by Leung and coworkers.[18] The 100 μL reactions contained 16.6 mM $(NH_4)_2SO_4$, 67 mM Tris–HCl pH 8.8, 6.5 mM $MgCl_2$, 6.7 μM EDTA pH 8.0, 10 mM β-mercaptoethanol, 10% dimethylsulfoxide, 1 mM each dATP, dGTP, dCTP, dTTP, 1 ng of linearized plasmid, 80 pmol of each primer and 5 units of Taq DNA polymerase (BRL). Mutagenic reactions had only 0.2 mM dATP, 6.1 mM $MgCl_2$ and 0.5 mM $MnCl_2$. In some rounds of non-mutagenic PCR the concentration of $MgCl_2$ was reduced to 4 mM, which improved the yields significantly. Tubes of 0.5 mL were employed and the reaction mixture was covered with 100 μL of mineral oil. The reactions were started by a 2 min incubation at 94 °C and followed by 25 cycles of 94 °C for 1 min, 50 °C for 1 min, 70 °C for 4 min and a final step of 72 °C for 10 min. An Eppendorf thermocycler (Mycrocycler) was used to execute this program; the duration was ~4.5 h. The template used was a Sma I digest of the replicative form DNA of M13RV,[15] an M13 derivative carrying the sequence of the 5′-domain of *E. coli* 16S rRNA. Primer AS-34 (48-mer) anneals to positions 8–25 of the 16S rRNA sequence, provides a promoter for the RNA polymerase from phage T7, and carries an Eco RI site. Primer AS-35 (27-mer) anneals to positions 557–575 of the 16S rRNA sequence and carries a restriction site for Bam HI. All deoxynucleotides and ribonucleotides were purchased from Pharmacia. Oligonucleotides were synthesized on a Biosearch synthesizer using Millipore reagents and were purified through oligonucleotide purification cartridges from Applied Biosystems.

Processing of PCR reactions for transcription or cloning was done as follows. The reaction mix was drawn from under the mineral oil and extracted with $CHCl_3$. The DNA was then precipitated with sodium acetate and ethanol and resuspended in buffer. Digestion with EcoRI and BamHI (Stratagene) was followed directly by purification using the Magic PCR miniprep resin from Promega to remove the bulk of DNA under 300 nucleotides. This step improved the yield of full length product in the subsequent transcription reactions. The DNA was recovered in 50 μL of TE buffer.

Transcription

Transcriptions with T7 RNA polymerase were carried out in a total volume of 50 μL using one-third of the DNA product from one PCR reaction. Several reactions were done simultaneously to provide enough RNA for the selection reactions. Transcription reactions were precipitated with ethanol and then loaded on 6% polyacrylamide/50% urea gels and the band of full-length product was excised, frozen, crushed and soaked for 1–3 h at room temperature in 400 μL of 0.5 M sodium acetate in 10 mM Tris, 1 mM EDTA, pH 8.0 (TE buffer). The RNA recovered was further purified by NENsorb reverse phase cartridges (NEN). The yield of RNA was estimated by absorbance and concentrated for the selection reactions by precipitation.

Selection of RNA molecules that bind S4

RNA variants competent in S4 binding were separated from the bulk pool of RNA by filtration through a nitrocellulose filter. Protein and RNA were renatured separately prior to complex formation. Protein was diluted 20-fold in 30 mM Tris–HCl pH 7.6, 350 mM KCl, 43 mM 2-mercaptoethanol and incubated for 30 min at 37 °C, 5 min at room temperature and then transferred to ice. RNA was renatured in 30 mM Tris–HCl pH 7.6, 350 mM KCl, 20 mM $MgSO_4$ for 20 min at 42 °C, 10 min at room temperature and 10 min on ice. Appropriate dilutions were then carried out to achieve final concentrations of 30 mM Tris–HCl pH 7.6, 350 mM KCl, 8 mM $MgSO_4$, 25 mM 2-mercaptoethanol upon mixing of RNA and protein components. Total reaction volume was 50 μL and an incubation of 10 min on ice was allowed before filtration. The amount of RNA used typically varied between 50 and 200 pmol, and was in 10- to 15-fold molar excess over protein. After filtration under suction the filter (Schleicher and Schuell BA85, 25 mm) was cut in pieces of about 2 × 2 mm with a razor blade, transferred to an Eppendorf tube and soaked for 1 h at room temperature in 200 μL of TE and 400 μL of saturated phenol. Two extractions with ether followed and the RNA was precipitated in the presence of 20 μg tRNA (Sigma). The selected RNA was resuspended in the reverse transcription reaction mix.

Reverse transcription of selected RNA

cDNA synthesis was done in a total volume of 20 μL in the presence of 50 mM Tris–HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 10 mM dithiothreitol, 0.5 mM dNTPs and primer AS-35. This mix was incubated for 10 min at 55 °C and transferred to ice; 200 units of Superscript reverse transcriptase (BRL) were added and polymerization was allowed to proceed for 60 min at 45 °C. The tube was transferred to ice and the reaction stopped by addition of EDTA. The nucleic acids were precipitated and resuspended in 10 μL TE. A fraction of this material was then used in a PCR reaction to generate double stranded DNA complementary to the selected RNA.

Cloning and sequencing of variants

cDNAs were cloned at various stages of the selection in order to sequence individual molecules (Table 1). Double stranded DNA digested with Eco RI and Bam HI was run on a Nusieve 3% agarose gel to isolate full length molecules from the population of PCR products. The band was excised, melted at 70 °C for 10 min and the DNA recovered with Magic PCR miniprep resin (Promega). Standard ligation reactions for insertion into digested and dephosphorylated pUC-18 were carried out at room temperature using T4 DNA ligase from BRL. *E. coli* HB101 cells were transformed and recombinant clones were selected by growth in Hershey medium containing ampicillin (100 μg/mL). Plasmid DNA from individual clones was purified with the Magic DNA miniprep resin and protocol (Promega). The DNA obtained was suitable for automated sequencing.

Sequencing of clones was done on an Applied Biosystems sequencer model 373A using dye-primer chemistry. Each clone was sequenced from both directions in order to span the complete length of the inserts. The T7 primer and the -21M13 forward primer kits from Applied Biosystems were used to sequence from the Eco RI site and the Bam HI site, respectively. The length of insert sequence obtained was about 300 and 350 nucleotides for the T7 and -21M13 forward primers, respectively, resulting in an average overlap of ~110 nucleotides in the center of the molecule which allowed satisfactory checking of data in the stretches of highest uncertainty. Alignment of each sequence to the wild type 16S rRNA was done with the SeqEd 1.0.1 software package from Applied Biosystems using penalty parameters of 4–8–3 for mismatches, gaps and gap length, respectively. Manual adjustments were done according to the secondary structure map. All sequences from a particular set (i.e., from a particular stage of the iterative selection process) were compiled and final manual adjustments were made, especially in the placement of insertions and deletions.

Statistical analysis of mutation frequencies

Each set of sequences was subjected to a collective analysis as follows. In the compilation of aligned sequences the number of mutations in each stretch of 10 nucleotides was counted. This tally was taken from nucleotide 51 onwards since some sequences are

Table 1. Distribution of mutation levels among clones

Set	Low (%)	Medium (%)	High (%)	Clones sequenced
rm1	100.0			17
rm2	100.0			8
rm2s3	94.4		5.6	18
rm3s0	27.3		72.7	11
rm3s4	33.3	8.3	58.3	12
rm3s5	38.1	14.3	47.6	21
rm3s14	20.8	22.6	56.6	53

incomplete in the 5′-end. This count was then normalized per 10 sequences and will be referred to as the observed mutagenesis score (OMS). An expected mutagenesis of the wild type sequence, also expressed per 10 sequences and per 10-nucleotide stretch, was calculated using the frequency of base changes experimentally obtained for a pool subjected to two consecutive rounds of mutagenesis (pool rm2, Table 1). Specifically, the equation used to calculate this expected mutagenesis score (EMS) is:

$$\mathrm{EMS} = 10[A(A_m/A_0) + G(G_m/G_0) + C(C_m/C_0) + U(U_m/U_0)]M$$

where N (A, G, C or U) represents the frequency of each nucleotide in the 10-nucleotide stretch being scored; N_m is the frequency with which a base is mutagenized, obtained experimentally for a pool subjected to two consecutive rounds of mutagenesis (pool rm2, Table 1; $A_m = 0.453$, $G_m = 0.094$, $C_m = 0.113$, $U_m = 0.340$); N_0 is the fraction of nucleotide in the RNA before mutagenesis ($A_0 = 0.256$, $G_0 = 0.318$, $C_0 = 0.224$, $U_0 = 0.201$); and M is the overall level of mutagenesis for the RNA, 0.316 for the pool rm3s14 (high) (Table 1). The degree to which the observed mutagenesis is suppressed or enhanced over or under the expected mutagenesis can be expressed as OMS/EMS. When there is no suppression or enhancement this value equals 1. When there is suppression, values are <1 and by taking the reciprocal and applying a negative sign a suppression factor is obtained which is a measure of how many times the expected mutagenesis has been reduced. When there is enhancement of mutagenesis values will be positive and >1; these values are a measure of how much greater than expected the mutagenesis is.

Filter binding assays

Binding constants for particular pools and individual clones were calculated from filter binding assays carried out as described,[15] only substituting the renaturation protocol described above for filter selection. In the case of specific clones the plasmid DNA obtained by minipreps did not yield satisfactory RNA when used directly for in vitro transcription, so suitable templates were generated by amplifying the fragment of interest by PCR.

Results

Randomly mutagenized pools

To generate a pool of randomly mutated RNAs containing the S4 binding domain, a DNA fragment corresponding to the 5′-domain of *E. coli* 16S rRNA was randomly mutagenized by error-prone PCR using the procedure of Leung et al.[18] The mutagenized region comprised nucleotides A26–C556 (Fig. 1); PCR primers contained restriction sites and a T7 phage promoter to allow subsequent cloning of the DNA and transcription of the RNA fragment. The procedure was repeated twice to generate three pools with successively higher levels of mutations. RNA was transcribed from each of these pools and assayed for S4 binding (Fig. 2). As expected, the average affinity of each pool decreased after each round of mutagenesis. S4 has a fairly high nonspecific affinity for RNA which is close to the affinity of the third pool (K=1.3 μM^{-1}).[15] The second pool was therefore chosen to initiate the selection process.

It is important to know whether mutations in the initial pools are indeed random. Seventeen and eight clones were sequenced from the first and second pools, respectively, and overall mutagenesis levels of 0.58% and 1.37% were found; the mutations in each pool were randomly distributed throughout the sequence. There was significant bias in favor of transitions over transversions: 7.6:1 in the first pool and 4.9:1 in the second. This bias contrasts with the lack of any bias reported by Leung et al.;[18] however, others using the same reaction conditions have also found transitions significantly favored.[19]

It is noteworthy that a relatively low level of mutagenesis substantially decreases the S4 binding constant, as if a large portion of the 5′-domain is required for S4 recognition. Ribosomal proteins typically bind domains of less than 30 nucleotides[1,20,21] and are sensitive to mutations at even fewer sites. After two rounds of mutagenesis, the average number of mutations per RNA is about 7/531. Assuming there are 30 sensitive sites, less than half of the molecules should contain a deleterious mutation. Yet this level of mutagenesis reduces binding by nearly an order of magnitude, indicating that at least 90% of the molecules have a significantly impaired capacity to recognize S4.

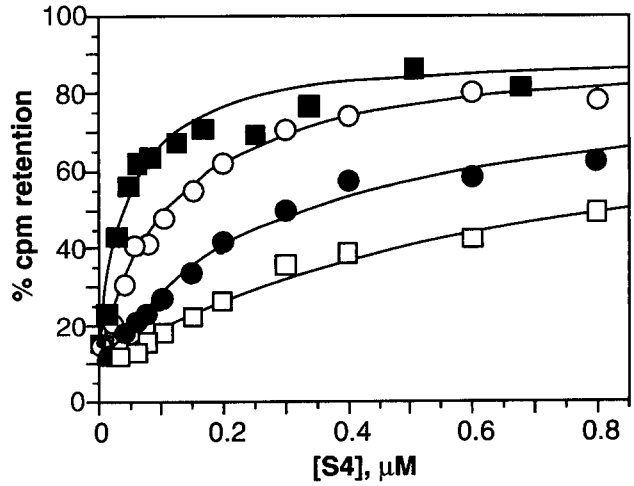

Figure 2. Binding curves for mutant pools of RNA. Curves were obtained from filter binding assays using RNA from three consecutive rounds of mutagenesis. Curves are least squares fits to the data. The background for the curves shown is 10%. ■: wild-type RNA (K = 18.3 μM^{-1}, plateau 91.0%); ○: pool rm1 (K = 8.6 μM^{-1}, plateau 91.7%); ●: pool rm2 (K = 3.3 μM^{-1}, plateau 85.6%); □: pool rm3 (K = 1.3 μM^{-1}, plateau 86.4%).

Selected pools

Successive rounds of selection/amplification were performed starting from the second random mutagenesis pool (termed rm2); about 10^{12} variants were used for the initial selection. Five rounds of selection for RNA molecules capable of binding S4, and therefore retained on a nitrocellulose filter after incubation with the protein, were carried out. The resulting pool, rm2s5, was then subjected to a third mutagenic cycle. Fourteen selection rounds ensued; these pools are called rm3s0–rm3s14. Selections were done using RNA:protein ratios of 10–15. The binding strength for pools rm2 and rm3s0 was approximately the same, ~3 μM^{-1}, and pools rm3s1 through rm3s14 exhibited binding constants ranging from 2.5 to 6 μM^{-1}, with a final K of ~5 μM^{-1} for rm3s14.

These selection results were initially disappointing, as we had hoped to obtain a more convincing increase in the average S4 affinity of the selected pools. However, the distribution of mutants in the selected pools was quite nonrandom, suggestive of strong selective pressure by S4. The mutagenesis rate was also much higher than expected, probably because the PCR amplification steps between selections had generated significant accumulation of mutants, so that a large number of mutations were selected. These findings are presented below, and confirm the suspected requirement of S4 for extensive portions of the 5′-domain.

Selected clones are of three types

A large number of clones from the final selection were sequenced, along with smaller numbers from other pools for comparison (Table 1). The level of mutagenesis tended to fall into one of three categories: low (0–10%), medium (11–14%), and high (29–34%). Clones which are heavily mutagenized display three areas where changes are particularly dense: nucleotides 70–98, 121–288 and 407–499, while clones of the medium type have only one or two of these dense centers and mutations in the low category are evenly distributed through the molecule. Histograms of these distributions are shown in Figure 3 for the 53 sequences from the rm3s14 pool. Of the 12 sequences in the medium class, 11 had multiple changes in the 407–499 area and the remaining one had dense mutagenesis in the other two sectors mentioned. The unusual distribution of mutation levels and how the unexpectedly high levels of mutation came about is commented on in the Discussion.

The information available in the group of selected RNAs with a low level of mutations is sparse, and the 'medium' group seems to show a subset of mutations appearing in the 'high' group. We therefore concentrate on the highly mutated class of sequences in the analysis that follows.

Conservation of secondary structure

An impressive number of complementary base pair changes was picked up by the selection process. For the 30 clones sequenced of the high mutagenesis type from pool rm3s14, an average of 35 pairs of complementary base changes per molecule was registered. These are pairs of nucleotides which are canonically paired (A–U, G–C, G–U) within helix segments in the wild type secondary structure (155 such pairs are drawn in Fig. 1) and appear as a different canonical pair in the mutant as a consequence of changes in both members of the pair (e.g., A–U→G–U was not counted). The number of noncomplementary base pair changes, in which both members of a canonical pair have been changed and a mismatch results, was also tabulated; for the rm3s14 set of the high mutagenesis type this number was 3.5 on average. If both nucleotides of a base pair are simultaneously changed the average chance of creating another canonical base pair is 41%; the observed percentage for this selected pool is 91%. For comparison, the effects of a mutation in only one member of a base pair were also counted: on average 12.7 base pairs per molecule were retained as canonical pairs while 8.1 were disrupted. In other words, in 61% of cases in which only one nucleotide is changed there is retention of complementarity; the level expected from random substitution is only 22%. Overall, whether by single or double changes, the number of pairs which continue to be such is four times greater than the number of disrupted base pairs. Figure 4 shows that the complementary base pair changes are extensively distributed through the molecule. It may be concluded that there has been strong selective pressure by S4 to maintain

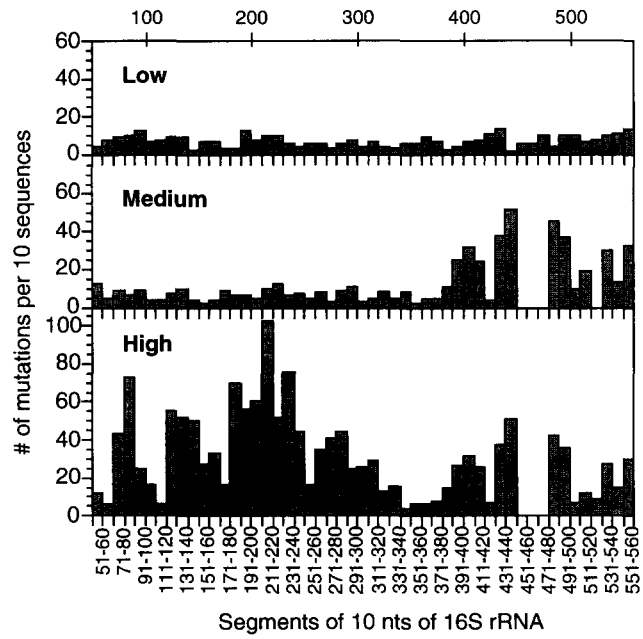

Figure 3. Frequency of mutations for selected pools. The mutation frequency for pool rm3s14 divided into three groups of overall mutagenesis level: low, medium, and high. Results are presented as number of mutations per 10 nucleotide segment in the sequence and normalized for 10 sequences.

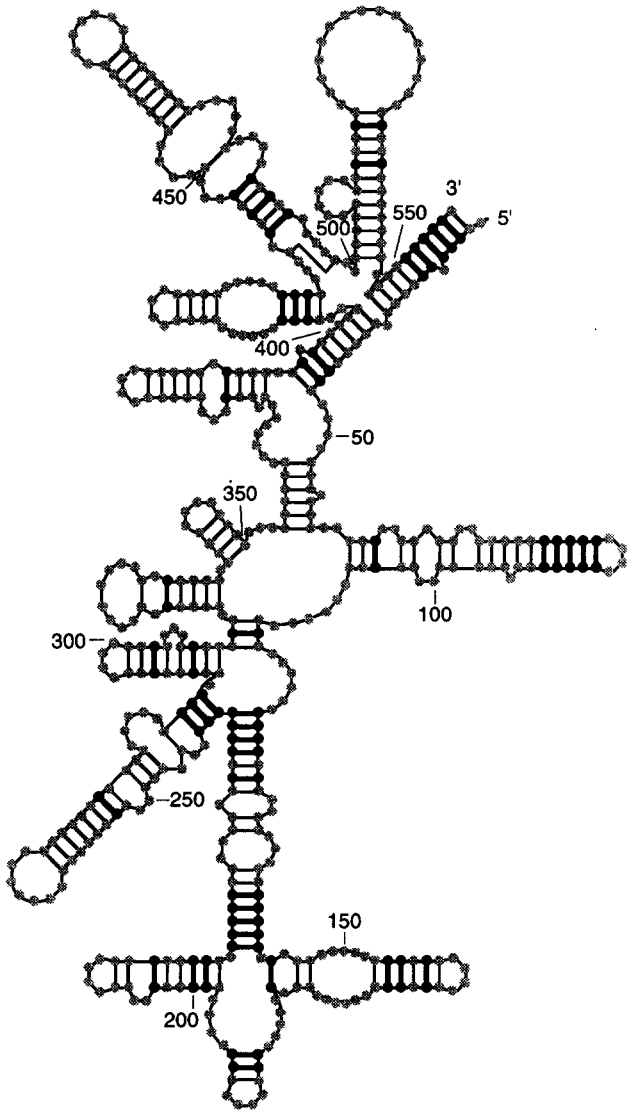

Figure 4. Compilation of complementary base pair changes. Nucleotides shown in black represent complementary base pair changes found in 30 clones of the high level of mutagenesis group from pool rm3s14.

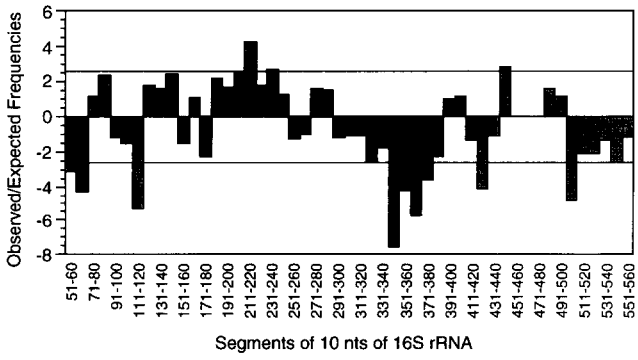

Figure 5. Mutagenesis suppression/enhancement factor. The degree to which mutagenesis has been suppressed or enhanced in reference to the expected value is shown for pool rm3s14 (high). Negative values indicate the times that mutagenesis is reduced; positive values indicate the number of times the level of mutagenesis exceeds the expected value. Lines mark one standard deviation.

secondary structure throughout much of the RNA domain.

Regions with low tolerance to mutagenesis

It is of interest to know whether the distribution of mutations in the selected RNAs is random; it might be expected that some regions would be less tolerant of base substitutions than others. This question was examined by first calculating the expected level of mutagenesis for 10-nucleotide intervals of the sequence based on the mutation frequencies registered for unselected clones from the rm2 pool, and then comparing this expectation with the observed level of mutations (see Materials and Methods for details of the calculation). Results are shown in Figure 5, expressed as degree of suppression or enhancement of mutagenesis relative to what would be expected if no selection had been applied. A 10 nucleotide window was chosen to give enough mutations to obtain statistically relevant

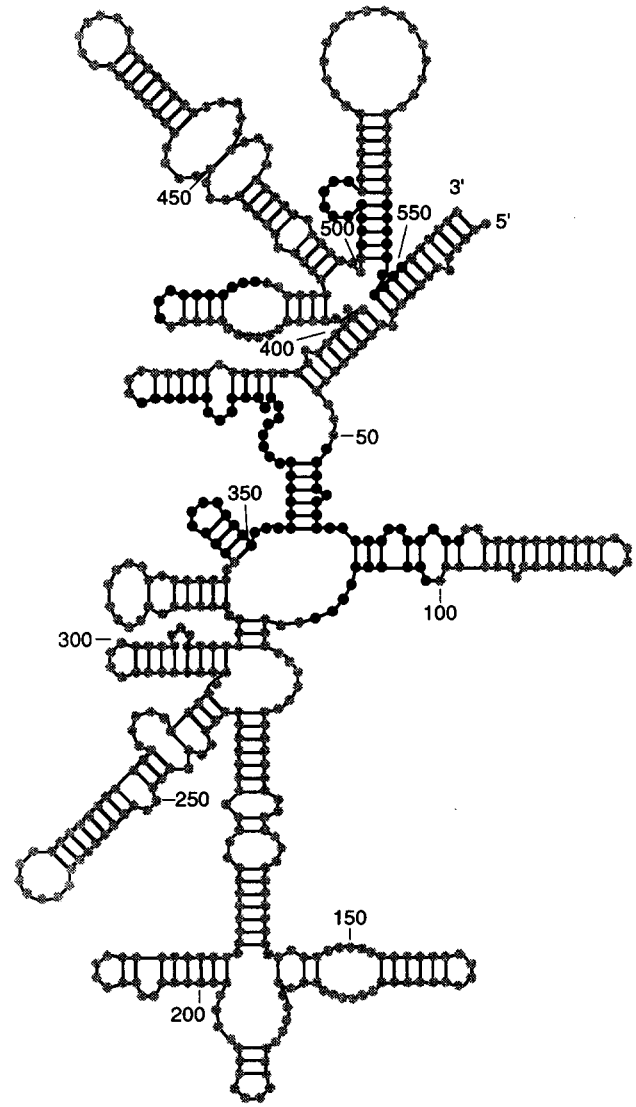

Figure 6. Regions where mutagenesis is poorly tolerated. Regions in which mutagenesis is suppressed more than 2.6 times as deduced from analysis of 10 nucleotide segments are shown in black.

results; the analysis is therefore limited to a fairly low level of resolution. Even with this limitation, a plot of regions with a significantly lower level of suppression than expected (Fig. 6) is very suggestive: mutations in the central region of the domain, centered on helix 52–58/354–359, must be particularly deleterious for S4 recognition. This region also shows remarkable phylogenetic conservation. Most rRNA helices are quite variable in sequence, but there is little variation of helix 52–58/354–359 and the A55 bulge in the phylogenetic record.[22,23] Additionally, chemical footprinting experiments[24] show enhanced reactivity for G362, A363 and A364 in the presence of S4, an indication that this segment undergoes conformational changes upon protein binding. This region may be key for folding the 5′-domain into the correct structure recognized by S4; perhaps it adopts a specific tertiary structure that correctly orients other helices of the domain.

Alterations in the 122–239 region

A number of intriguing loop or base pair substitutions were found in the set of highly mutated RNAs. Those substitutions found in a majority of this set of sequences are described in this and the following section (Figs 7 and 8).

In the hairpin centered at 210, a GA bulge and adjacent G–C pairs have been removed and substituted by a run of four G·U wobble pairs; all of the highly mutated set have the consensus sequence shown in Figure 7 or a close variant. Overall the helix is extended by one base pair, and the capping tetraloop that results from these sequence changes (GGAA) is completely different from the wild type loop (UUCG). Both tetraloop sequences adopt stable structures with noncanonical hydrogen bonds.[25,26] UUCG is present in 40% of bacterial sequences at this position, but 11% of bacteria have GCAA instead, which is a member of the same class of GNRA tetraloops as GAAA.[27]

The only two consistently observed insertions in the highly mutated RNAs are in the same region of the rRNA as the 210 helix. One is an insertion of a single purine that occurs at position 129/130 in most of the highly mutated RNAs (A: 25/30; G: 3/30) (Fig. 7). The identical insertion is common in the phylogenetic record: A is present at this position in most eubacteria and archaea and only subdivisions of the purple bacteria most closely related to *E. coli* lack it. The other insertion picked up during the selections is at position 193/194 (G: 29/30, A: 1/30), at the base of the 184–193 hairpin. In the phylogenetic data base, the 129/130 A insertion correlates with an extension of hairpin 184–193 from three to ~10 base pairs.[28] A structural interpretation of this correlation is not obvious, but it

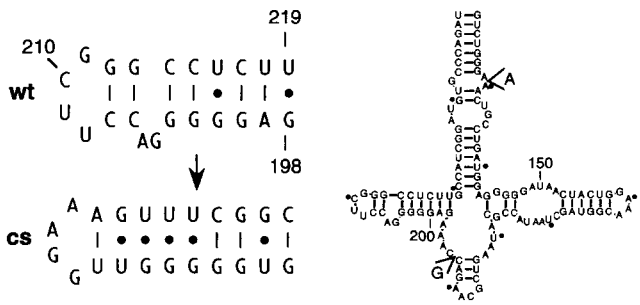

Figure 7. Wild-type (wt) and consensus (cs) sequences for the hairpin centered at 210 for pool rm3s14 (high) are shown on the left. The entire 122–239 *E. coli* sequence is shown at right, with two insertions found in most of the rm3s14 (high) sequences marked.

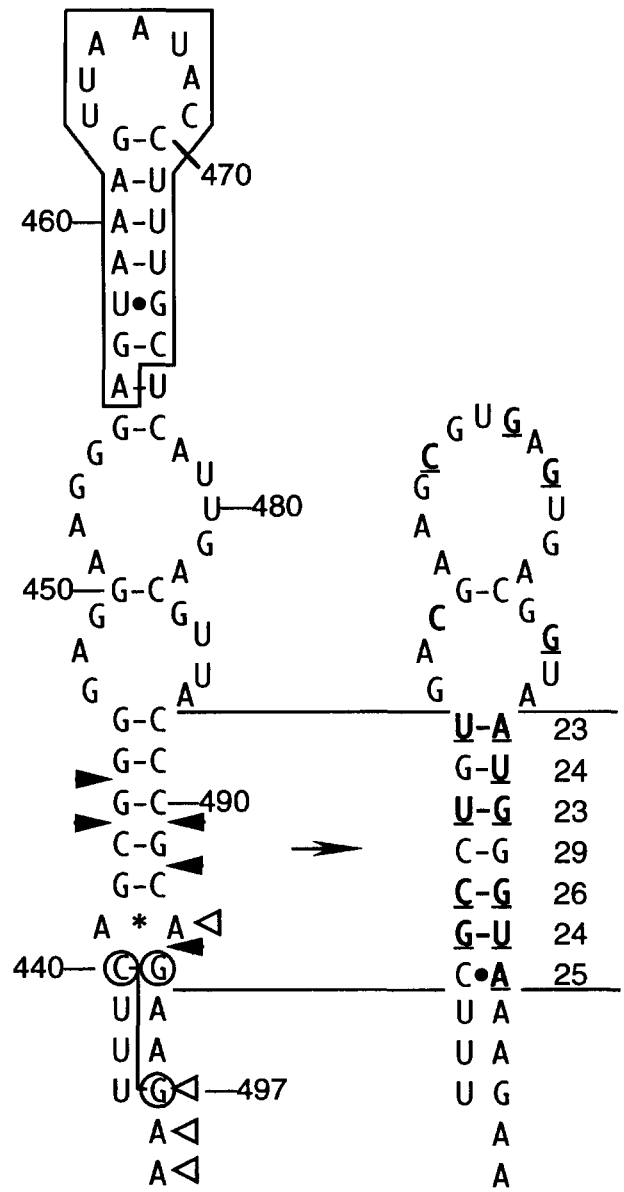

Figure 8. Changes in region 437–497. The consensus sequence for pool rm3s14 (high) is shown at the right, with nucleotides that are altered in most selected RNAs in bold-face type. The frequencies of occurrence (out of 30 sequences) of the consensus base pairs in the GC-rich stem are listed. The *E. coli* sequence is shown on the left with nucleotides in the selected sequences framed. Sites of S4 protection from V1 nuclease (black arrowheads) or single-strand specific chemical reagents (open arrowheads) are taken from ref 24. A triple covariance detected by phylogenetic analysis[3] is indicated by circled nucleotides.

shows a striking correspondence to the 129/130 and 193/194 insertions selected by S4.

437–498 extended hairpin

A 25 nucleotide deletion, from 456 to 475, was found in all RNAs of the highly mutated set (Fig. 8). Chloroplast rRNAs, which are closely related to the purple bacteria that include *E. coli*, have a remarkably similar deletion at 455–475 or 476; many other eubacteria and all archaea also carry approximately the same deletion. This deletion is particularly intriguing because a similar deletion in the context of the *E. coli* sequence adversely affects S4 binding,[17] suggesting that a compensating change may have occurred elsewhere in the selected RNAs. A mismatch in *E. coli* and many eubacteria occurs in the same extended hairpin, at A441–A493. This mismatch becomes a Watson–Crick base pair in the majority of sequences with a deletion of the distal helix (R. Gutell, personal communication); for example, 76% of archaea (which all have the deletion) have C441–G493, and the next most frequent combination (9%) is G441–C493. Among the 30 highly mutated RNAs, there are 24 with G·U, two with G·C, and one with A·U at 441 and 493, respectively.

The 437–446/485–497 stem is protected by S4 from enzymatic digestion and hydroxyl radical reaction, and is within the region that is the most likely candidate for direct contacts by S4. The conservation of this stem in S4-selected RNAs is thus of particular interest. Compensatory base changes have altered most of the bases through the stem in most of the highly mutated RNAs; the exception is C443–G491, which is unaltered in 29/30 sequences. C443–G491 is the most common base pair at this position in the eubacteria (42%), though neighboring base pairs show similar levels of conservation. It is worth noting that the mutation G494→A, which is frequently found, probably does not disrupt a base pair. C440, which opposes G494, covaries with G497 instead. The phylogenetic data are suggestive of a triple base interaction at these three positions; while G is the most common base at 494 when 440–497 is C–G, A or U are also found with C–G (R. Gutell, personal communication).

Discussion

Generation of high levels of mutation

The high level of mutations in some of the selected RNAs was unexpected. One such highly mutated molecule appeared among the 18 RNAs sequenced after rm2s3, and more than half had a high level by the end of the selections. It is unlikely that the random mutagenesis protocol, which after three rounds would give an average mutation rate of ~2%, could have generated sufficient numbers of molecules with base changes at 30% of the nucleotides to account for the selection results. We suspect that these highly mutated molecules gradually 'evolved' during the successive rounds of selection by two mechanisms. First, Taq polymerase is rather error prone, and many rounds of PCR had to be carried out between rounds of selection to obtain sufficient material to carry on to the next step. It is likely that we were inadvertently adding significant numbers of mutants at each round of selection, without having to use special reaction conditions. In other experiments in which mutagenesis was confined (by chemical synthesis) to 18 nucleotides of a 70 nt RNA, mutations at specific sites outside of the mutagenized region appeared midway through the selections and were present in all molecules by the end;[29] presumably these arose by PCR-generated errors.

The second factor that probably contributed to the mutagenesis is recombination during PCR. Mutations in any one region of the highly mutated molecules tend to belong to a family of related sequences, and examples of these families tend to occur in less highly mutated groups of selected RNAs. The unusual rearrangement in the 210 helix is a good example. It appears once among the moderately mutated RNAs as a region with a much higher density of mutations than the remainder of the molecule, but closely related 210 helix sequences also appear in most of the highly mutated RNAs, in combination with other regions having many mutations. During PCR, premature termination of polymerization will generate a partial copy that can serve as a primer in the next round of synthesis and (in effect) recombine two sequences. When deliberately encouraged, this kind of recombination has been shown to greatly expand the sequence space explored in selection experiments, and recombination occurs at a low level under standard PCR conditions.[30] PCR starting with partial length cDNAs, as is likely for the longer transcripts used in these experiments, can similarly 'recombine' sequences. A plausible scenario is that successive rounds of amplification introduced compensatory mutations that generated 'fit' regions for S4 binding, and that these regions recombined in subsequent rounds to produce the highly mutated RNAs. Protocols specifically designed to take advantage of these tendencies of PCR might be particularly effective in generating new sequences in large RNA domains.

S4-RNA recognition within the 5′-domain

Based on the relatively small region of the 5′-rRNA domain protected by S4 from hydroxyl radical, it might have been expected that the selection experiments described here would have yielded molecules with highly conserved sequences or secondary structures in a limited region coincident with the protection sites. It is therefore striking that compensatory mutations have been selected through much of the domain, as if a large fraction of the secondary structure is needed for S4 recognition.

We can imagine several reasons why sequences outside of the protected region might have been selected by S4.

First, the protected region may be much smaller than the actual contact region: S4 does not affect the reactivity of some contacted residues, or some contacted nucleotides are unreactive in the free RNA and therefore 'silent'. Given the number of chemical and enzymatic reagents used to probe the S4-rRNA complex,[16,24] it seems unlikely that any extensive contact with the RNA has been completely overlooked, though it is difficult to rule out this possibility entirely. A second possibility is that secondary structure outside of the protected region may be needed simply to keep these sequences from interfering with folding of the remainder of the molecule: accumulation of numerous mismatches from nucleotides 50–360 could very well cause the entire domain to fold into a quite different secondary structure. But if this were the case, we would expect that quite different secondary structures would appear in the 50–360 region as mutations accumulated, and no correspondence between the selected sequences and the phylogenetic data-base would be expected. Thus we do not think that this possibility can entirely explain the observed pattern of selection.

We think the most likely reason that a large fraction of the 5′-domain is selected by S4 is that the 50–360 region helps maintain the recognized domain in the appropriate conformation for S4 recognition via tertiary interactions. A specific framework of secondary and tertiary structures may be needed throughout the domain to preserve the S4 binding site. This possibility is consistent with site-directed mutation studies showing that eliminations of bulges or internal loops at several sites throughout the molecule have significant effects on S4 binding affinity (ref 17 and unpublished observations).

The similarity of the S4-selected RNAs to the phylogenetic data-base is noteworthy. Examples that have been mentioned are the insertion of an A at 129/130, the 456–475 deletion and the conversion of the A441–A493 mismatch to a canonical pair, and the reduced level of mutations within helix 52–58/354–59. The implication is that evolution of features throughout the 5′-domain has been constrained by interactions with S4 protein, and the conclusion that much of the 5′-domain structure contributes (directly or indirectly) to S4 recognition seems inescapable. A main role for S4 may be to stabilize a pre-existing tertiary structure in the 16S rRNA 5′-domain, and further examination of conserved regions within the 5′-domain for their roles in RNA structure and S4 recognition may be fruitful.

Acknowledgements

We thank Sarah Morse for providing T7 RNA polymerase and Peter Kebbekus for synthesis of DNA primers. We are grateful to Tom Schneider from the National Cancer Institute for the use of the ABI sequencer in his laboratory and to Denise Rubens for sharing her expertise in all aspects of automated sequencing. We also thank Robin Gutell (University of Colorado) for providing summaries of phylogenetic data and helpful discussions. This work was supported by NIH grant GM29048.

References

1. Draper, D. E. *Annu. Rev. Biochem.* **1995**, *64*, 593.
2. Madison, J. T.; Everett, G. A.; Kung, H. *Science* **1966**, *153*, 531.
3. Gutell, R. R.; Larsen, N.; Woese, C. R. *Microbiol. Rev.* **1994**, *58*, 10.
4. Tuerk, C.; Gold, L. *Science* **1990**, *249*, 505.
5. Tsai, D. E.; Harper, D. S.; Keene, J. D. *Nucleic Acids Res.* **1991**, *19*, 4931.
6. Schneider, D.; Tuerk, C.; Gold, L. *J. Mol. Biol.* **1992**, *228*, 862.
7. Schneider, D.; Gold, L.; Platt, T. *FASEB J.* **1993**, *7*, 201.
8. Burd, C. G.; Dreyfuss, G. *EMBO J.* **1994**, *13*, 1197.
9. Bartel, D. P.; Zapp, M. L.; Green, M. R.; Szostak, J. W. *Cell* **1991**, *67*, 529.
10. Giver, L.; Bartel, D.; Zapp, M.; Pawul, A.; Green, M.; Ellington, A. D. *Nucleic Acids Res.* **1993**, *23*, 5509.
11. Jensen, K. B.; Green, L.; MacDougal-Waugh, S.; Tuerk, C. *J. Mol. Biol.* **1994**, *235*, 237.
12. Ekland, E. H.; Szostak, J. W.; Bartel, D. P. *Science* **1995**, *269*, 364.
13. Mackie, G. A.; Zimmermann, R. A. *J. Mol. Biol.* **1978**, *121*, 17.
14. Ehresmann, C.; Stiegler, P.; Carbon, P.; Ungewickell, E.; Garrett, R. A. *Eur. J. Biochem.* **1980**, *103*, 439.
15. Vartikar, J. V.; Draper, D. E. *J. Mol. Biol.* **1989**, *209*, 221.
16. Powers, T.; Noller, H. F. *J. Biol. Chem.* **1995**, *270*, 1238.
17. Sapag, A.; Vartikar, J. V.; Draper, D. E. *Biochim. Biophys. Acta* **1990**, *1050*, 34.
18. Leung, D. W.; Chen, E.; Goeddel, D. V. *Technique* **1989**, *1*, 11.
19. Cadwell, R. C.; Joyce, G. F., *PCR Methods Applic.* **1992**, *2*, 28.
20. Draper, D. E. In *RNA-Protein Interactions*; Nagai, K.; Mattaj, I., Eds.; IRL Press, Oxford: 1995; pp 82–102.
21. Mougel, M.; Allmang, C.; Eyermann, F.; Cachia, C.; Ehresmann, B.; Ehresmann, C., *Eur. J. Biochem.* **1993**, *215*, 787.
22. Woese, C. R.; Gutell, R.; Gupta, R.; Noller, H. F., *Microbiol. Rev.* **1983**, *47*, 621.
23. Gutell, R. R.; Schnare, M. N.; Gray, M. W., *Nucleic Acids Res.* **1992**, *20*, S2095.
24. Stern, S.; Wilson, R. C.; Noller, H. F. *J. Mol. Biol.* **1986**, *192*, 101.
25. Heus, H. A.; Pardi, R. *Science* **1991**, *252*, 191.
26. Allain, F. H.-T.; Varani, G. *J. Mol. Biol.* **1995**, *250*, 333.
27. Woese, C. R.; Winker, S.; Gutell, R. R. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 8467.
28. Woese, C. R.; Gutell, R. R. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 3119.
29. Lu, M.; Draper, D. E. *Nucleic Acids Res.* **1995**, *23*, 3426.
30. Stemmer, W. P. C. *Nature (London)* **1994**, *370*, 389.

(*Received in U.S.A.* 2 *September* 1996; *accepted* 18 *February* 1997)