Volume 17 Number 12 1989 Nucleic Acids Research

Detection of common motifs in RNA secondary structures

Hanah Margalit, Bruce A.Shapiro[1], Amos B.Oppenheim[2] and Jacob V.Maizel Jr[1]

Laboratory of Mathematical Biology, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, [1]Frederick Cancer Research Facility, Frederick, MD 21701, USA and [2]Department of Molecular Genetics, Hebrew University, Hadassah Medical School, Jerusalem 91010, Israel

Received January 9, 1989; Revised and Accepted May 19, 1989

ABSTRACT

We describe a novel computerized system for comparison of RNA secondary structures and demonstrate its use for experimental studies. The system is able to screen a very large number of structures, to cluster similar structures and to detect specific structural motifs. In particular, the system is useful for detecting mutations with specific structural effects among all possible point mutations, and for predicting compensatory mutations that will restore the wild type structure. The algorithms are independent of the folding rules that are used to generate the secondary structures.

INTRODUCTION

Possible relationship between RNA secondary structure and its biological function has been demonstrated in several cases. Molecules with the same function have the potential to fold into very similar secondary structures although differing in their primary sequences, indicating the importance of those secondary structures to the function. Such an example are the ribosomal RNA molecules (reviewed in 1).

mRNA secondary structure probably plays a role in translation regulation. Translation initiation may be affected by the mRNA structure around the initiation site (reviewed in 2-4). The more unstructured the regions of the initiation codon AUG and the Shine-Dalgarno sequence, the more accessible they are for the interaction with the ribosome to initiate translation. Ganoza et al. (5) compared the potential secondary structure of prokaryotic mRNA translation start regions to that of regions with an internal AUG. They showed that AUG's in start positions appear to be single stranded while the internal AUG's are embedded in highly structured regions. Involvement of mRNA secondary structure in translation initiation was demonstrated in several cases. Mutations that disrupt a potential secondary structure in the AUG region led to increased level of the translated protein (6,7). Tessier et al. (8) studied 8 variants in the initiation region of the human immune interferon gene and showed a correlation between the presence/absence of potential stem-loop structure and decreased/increased level of translation, as expressed in E. coli. Zagorska et al. (9) induced irreversible unfolding of the mRNA by using O-methylhydroxylamine and showed that the interaction of E.coli ribosomes with the mRNA depends on its secondary structure. Methoxyamine reacts specifically with cytosines, preventing them from base-pairing and leading to irreversible unfolding of the RNA. They showed that the unfolded RNA was about 50 times more active in ribosome binding than the native RNA. Hall et al. (10) showed the effects of mutations in the initiation region of E.coli lamB mRNA on the translation levels. While the two single mutations in their study decreased significantly the translation efficiency, the double mutant restored the synthesis almost to the wild type level. The effects of the single mutations and double mutant were attributed to changes in the secondary structure of the initiation region. Translation regulation by proteins may involve primary sequence specific patterns as well as higher structures that serve as recognition features for the protein. Studies on the interaction between the ribosomal protein L1 and the L11/L1 operon mRNA illustrate the importance of RNA secondary structure for such interactions (11-13). L1 autoregulates its synthesis by inhibiting the translation of its own mRNA. Nomura et al. (12) suggested that L1 recognizes a site on the mRNA that is structurally similar to its binding site on 23S ribosomal RNA. They supported this idea by analyzing various mutations in this region, showing that mutations with a potential for disruption of the structure had an effect on the level of inhibition (11). A correlation between mRNA secondary structure and translation regulation by proteins was also shown in other systems such as R17 coat protein interaction with the replicase initiation site (14,15), and threonyl tRNA synthetase interaction with its own mRNA (16-19).

In most of the studies described above the secondary structures were not proven by biochemical means but were generated on the basis of free energy considerations (20), with the predicted structure being the one with the minimum free energy. Indeed, the experimental structural data to evaluate the accuracy of these predictions is limited (19, 21, 22). Despite this shortcoming, the ability to correlate the experimental results of translation regulation to the secondary structures predicted by the computer is encouraging.

In the above studies the different structures that result from mutations were compared manually, thus, limiting the number of structures that could be examined. Also, such a non-automated examination is possible when looking for local effects of the mutations on the structures. When long sequences are studied and long range interactions play a role in the determination of the structure, a manual examination of those structures is even more limited. An automated system that could predict mutations with certain structural effects could be a useful tool for studies on structure-function relationship. We have developed a computerized system with algorithms that are able to compare a large number of secondary structures, clustering similar structures and pointing to common sub-structures. For a given sequence the computer generates all the possible sequences that could result from a point mutation in each position. The secondary structures of all the sequences are derived automatically and a comparison and clustering of the structures is performed. Also, the system is able to look for a specific motif that is suspected to be related to the function, such as an unstructured Shine-Dalgarno region or a feature for protein recognition. A computerized system such as this enables screening of a very large number of sequences quickly, pointing to the mutations which affect the structure, and thus, assisting the experimentalist in the process of choosing the interesting sequences (mutations) to study. In addition, this system is useful for detecting conserved structures among many molecules (for example, among molecules with a functional similarity). We describe below the characteristics of the automated system and demonstrate its use on specific examples.

Methods

The system that we describe here is one part of an extended system to be described in detail elsewhere (23). The core of the system described here is an algorithm that compares and clusters secondary structures by their morphological entities (23,24). The algorithm uses the secondary structures as its input, represents each secondary structure as an ordered tree and then clusters the trees based on a distance metric. The program expects to read the secondary structure input as a region table which includes information on the base-paired sub-sequences, encoding thereby the information on the secondary structure of the whole molecule. The information in the region table is organized in triplets where each triplet consists of the 5' end of the base-paired region, the 3' end of this region and the size of the region. The region table information is then translated into an ordered tree, where the nodes of the tree are the morphological entities of the structure and the arcs represent base-paired regions. The morphological entities of the structure are the different types of loops: hairpin, internal, bulge, and multibranch. A hairpin loop is formed when a double stranded region has an intervening single strand. Internal loops are defined by two double stranded regions separated by two intervening single strands, each of them is at least one base long. A bulge loop is a specific type of an internal loop in which one of the single strands is of length zero. A multibranch loop is a loop from which two or more double stranded regions are emanating.

The tree clustering is done by two different schemes. By one scheme the trees are represented as strings of characters that stand for the different morphological entities (B, I, M, and H for bulge, internal, multibranch and hairpin, respectively). The clustering is achieved by using a multiple sequence alignment algorithm on the strings of characters (24,25). The output of this procedure is a clustering of sequences of similar structures. By the second scheme (26,27), the pairwise distances between the ordered trees representing the secondary structures are computed. The method that is used to calculate the distances between the trees that represent the secondary structures is analogous to the scheme to calculate the distance between two sequences. One may assign a cost function that is associated with node insertion, deletion and relabeling, where the nodes represent the morphological entities of hairpin, loop (bulge, internal, or multibranch), and base-paired region. The minimum editing sequence to transform one tree to another is then determined. When the morphological entities are compared, the trees contain just the morphological information, described above. However, another scheme, similar to the second one, is able to consider not just the morphological entities but also their sizes as well as the sizes of the base-paired regions. For this purpose the nodes of the trees contain specific information concerning loop and base-paired region sizes, and the size differences between corresponding loops and base-paired regions are added to the penalties of insertion, deletion or relabeling. The distances between structures can then be used in a clustering algorithm to generate a taxonomy tree. Once the pairwise distances between structures have been computed, the structures are clustered based upon the measure:

$$D(a,b) = \min_{a_0,,,a_n} \{\max_{0<i<n}[d(a_i,a_{i+1}) | a_0=a,\ a_n=b,\ a_0,,,a_n]\}$$

where d(x,y) represents the pairwise tree distance. A detailed description of this algorithm is given in (26).

In addition, the system is able to search for the presence or absence of a certain motif and to extract structures according to the occurrence of this motif.

The different features of the above algorithms can be applied on any set of secondary structures, provided by the user in the form of region tables. When the researcher does not have a specific way to fold the sequence (according to his experimental results or phylogenetic data) and wishes to look for the minimum free energy structures, the input can be the primary sequences. The FOLD program (20), which is also incorporated into the system, will then generate the secondary structures with minimum free energy for all the sequences in the set in a non-interactive fashion. The FOLD program may use any set of free energy values. In the examples provided below we use the refined free energy values that were published by Freier et al. (28).

One way to study the effect of the structure on the biological activity is by using mutations or substitutions and examining the correlation between their structural and functional effects. Our system is useful for such studies that are aimed to identify mutations with certain structural effects. For a specific input sequence, a computer program generates all the possible point mutations. Upstream from the start position the program generates all the three possible substitutions in each position. Downstream from the AUG, the user can choose between two options: either to make all possible substitutions or to make only those substitutions that conserve the amino acids. The number of mutated sequences depends, of course, on the length of the input sequence. The computer is programmed to generate up to 10**5 mutated sequences. This number can be increased, if necessary, by a simple change in the program. The system will then generate the secondary structures for all the mutated sequences and cluster together sequences of similar structure, or search for a specific motif that is suspected to be relevant to the function. If there is already a known mutation with a suspected structural effect, the system is useful for locating compensatory mutations that will restore the wild type structure. In that case, the input sequence includes the point mutation and the computer will generate the double mutants by substituting the other positions.

The system is currently distributed across several different machines, but is controlled by a Symbolics 3675 LISP machine, making the distributed aspect transparent to the user. Algorithms that are run on the Symbolics or on other hardware systems are invoked by the Symbolics by a mouse click on menu items. This approach takes advantage of those algorithms which are useful even though they physically reside on different machines. The structure alignment and tree comparison algorithms run on a SUN and are written in C. The mutation generation algorithm runs on a VAX and is written

GCCUUUUGUUUUUAUGGGCCUUGCCCGUAAAACGAUUUUUUAUAUCACGGGGAGCCUCUC
AGAGGCGUUAUUACCCAACUUGAGGAAUUUAUAAUGGCUAAGA

FIG. 1
The leader region of L11 mRNA. The Shine-Dalgarno sequence and the AUG are underlined.

in FORTRAN. The structure generation, motif searches, and taxonomy presentation algorithms are written in LISP and run on the Symbolics. At present, an effort is being made to incorporate all the components of the system onto a SUN workstation.

RESULTS

We demonstrate the capabilities of the computerized system on some simple examples.

As the first example we examine the L11 operon of E. coli. The L11 mRNA has been studied extensively by Nomura's group who suggested that some secondary structure features of that region are important for its translation regulation by L1 protein (11,12,21). They showed that mutations with a potential effect on the structure affect the regulation and suggested that a structure of a stem-loop-stem, in which the stems are GC rich and the loop includes a critical adenine, is responsible for the recognition by L1. We would like to demonstrate how the computerized system is able to analyze all the possible mutations in that region, point out the ones with potential effect on the secondary structure, and cluster all the ones that are similar to the wild type structure in one group.

The input sequence is from the leader region of L11 mRNA and is 103 nucleotides long (Fig. 1). The Shine-Dalgarno region starts at position 82 and the AUG starts at position 94. We run the mutation program on nucleotides 1-80 because this is the region of interest. Since all the mutations are upstream from the AUG, all the three possible substitutions are made in each position. Thus, we generate a set of 241 sequences (240 mutations and the wild type sequence). Kearney and Nomura (21) studied experimentally the secondary structure of that region and suggested a structure that is very similar to the minimum free energy structure that is predicted by the FOLD algorithm. Also, the features that are suspected to be important for the interaction with L1 are predicted by the FOLD algorithm, and therefore we felt confident to do our structural analysis on the computer generated structures. 241 region tables that represent the secondary structures are generated

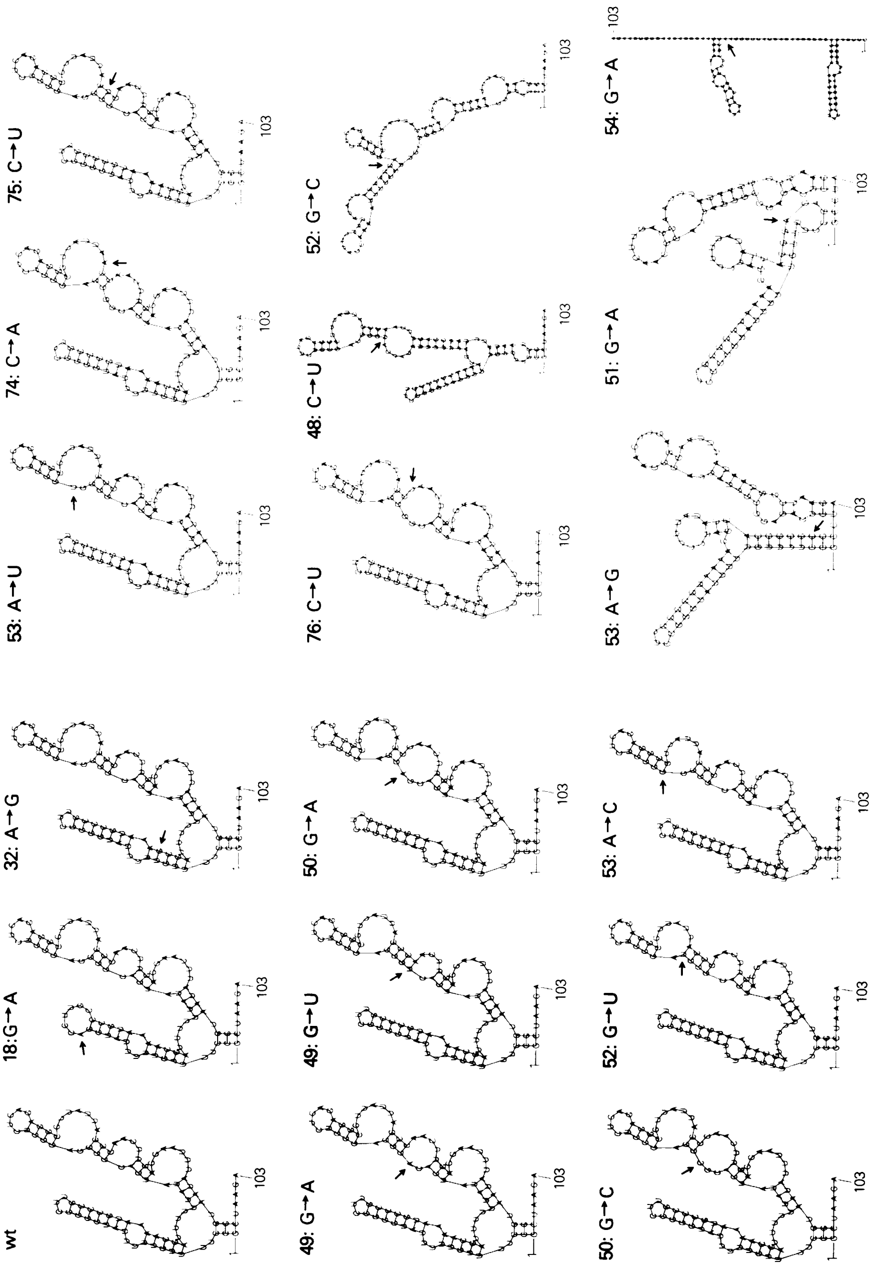
TABLE I

SUMMARY OF KNOWN MUTATIONS IN THE L11 mRNA LEADER REGION

POSITION OF MUTATION	NUCLEOTIDE CHANGE	REPRESSION RATIO	STRING REPRESENTATION OF SECONDARY STRUCTURE
W.T.		8.45	(N(M(IIIH)(IH)))
32	A→G	6.51	(N(M(IIIH)(IH)))
75	C→U	3.37	(N(M(IIIH)(IH)))
76	C→U	5.89	(N(M(IIIH)(IH)))
18	G→A	2.73	(N(M(IIIH)(IH)))
52	G→C	2.38	(NIII(M(H)(IH)))
52	G→U	2.32	(N(M(IIIH)(IH)))
74	C→A	1.92	(N(M(IIIH)(IH)))
49	G→A	1.89	(N(M(IIIH)(IH)))
54	G→A	1.47	(N(IIBH)(IH)))
49	G→U	1.32	(N(M(IIIH)(IH)))
50	G→C	1.30	(N(M(IIIH)(IH)))
51	G→A	1.28	(N(IIIH)(I(M(H)(H))))
53	A→G	1.24	(N(IIIH)(M(H)(IH)))
50	G→A	1.23	(N(M(IIIH)(IH)))
53	A→C	1.23	(N(M(IIIH)(IH)))
53	A→U	1.16	(N(M(IIIH)(IH)))
48	C→U	1.08	(N(I(M(IIH)(H)))

The mutation data is taken from a study by Thomas and Nomura (11). The mutations below the line abolish regulation and those above the line affect regulation without completely eliminating it, as specified by Thomas and Nomura in Fig. 2. of reference 11. The repression ratio gives a measure of the effect of each mutation on the translational regulation by L1 (11). The last column describes the structures by their morphological entities where B, I, M, and H stand for bulge, internal, multibranch and hairpin loops. The parentheses separate the different substructures. Note that the ordering is such that the morphological structures that are closest to the 5' end of the molecule are on the right side of each representation.

automatically and clustered. As mentioned above, the clustering can be done either by calculating the inter-distances of the trees that represent the secondary structures, or by running a multiple alignment algorithm on the


wt
103
18:G→A
103
32: A→G
103
53: A→U
103
74: C→A
103
75: C→U
103
49: G→A
103
49: G→U
103
50: G→A
103
76: C→U
103
48: C→U
103
52: G→C
103
50: G→C
103
52: G→U
103
53: A→C
103
53: A→G
103
51: G→A
103
54: G→A
103

strings of characters that represent the secondary structure trees. For the L11 sequence, the two methods give the same clustering but by two different representations. The tree clustering represents the distances among the structures, where a distance of 0 means that the structures are identical and hence, clustered together. The other method gives an ordered list of the structures according to their string alignment and is more visually informative. One can learn which features are different between two clusters of structures or what are the common features that maintain structures in one cluster. Assuming that the structure generated by the computer represents the structure in nature, mutations with a structural effect can be then chosen for further experimental study.

Since the outputs for 241 structures are too long to be summarized here we will demonstrate the use of the system on the set of mutations studied by Thomas and Nomura (11). In table I there is a summary of these mutations: the position of the mutation, the nucleotide change, the repression ratio as a measure of the effect of the mutation on the regulation by L1 and the string representation of the secondary structure. The string representation can easily be understood by examining, for example, the structure of the wild type (Fig. 2) and its string representation (TABLE I). It includes a multibranch loop (M); on one branch there are three internal loops and a hairpin loop (IIIH) and on the other branch there are an internal loop and a hairpin loop (IH). We align the structures by their string representations (TABLE II) and generate the cluster tree (Fig. 3). This is a clustering that is based on the morphological entities of the secondary structures. Twelve mutations are clustered with the wild type, and assuming a correlation between the secondary structure and the biological activity, we would not predict a significant difference in activity between the wild type and these mutations. Two of those indeed had a very small effect on the regulation, five were found to affect the regulation without completely eliminating it, and the other five completely abolished regulation. Hence, if we were looking for the most significant substitutions we could have missed these last five. However, assuming a role for A in the internal loop (11) will exclude 53:A→C, 53:A→U from this list and reduce the number of false predictions to 3. Among the 5

FIG. 2
Secondary structure predictions for the L11 wild type sequence (shown in Fig. 1) and the 17 mutations (summarized in TABLE I). The positions and types of the mutations are indicated.

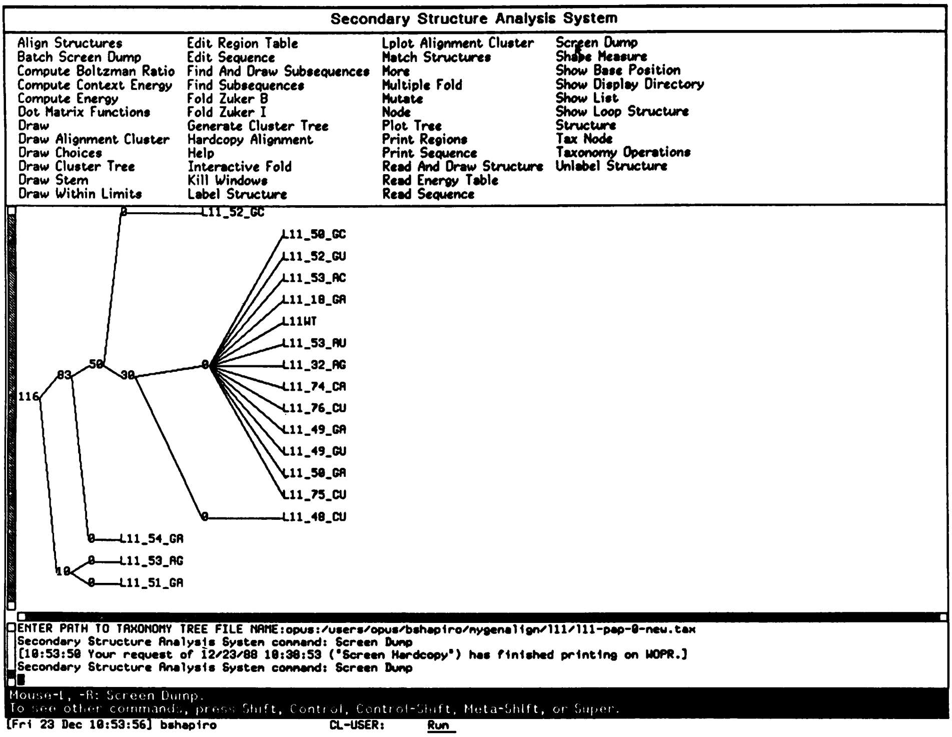
TABLE II
ALIGNMENT OF SECONDARY STRUCTURES

(N (IIbH)(IH))	L11 54 GA
(N (M(IIIH)(IH)))	L11 WT
(N (M(IIIH)(IH)))	L11 18 GA
(N (M(IIIH)(IH)))	L11 32 AG
(N (M(IIIH)(IH)))	L11 49 GA
(N (M(IIIH)(IH)))	L11 49 GU
(N (M(IIIH)(IH)))	L11 50 GA
(N (M(IIIH)(IH)))	L11 50 GC
(N (M(IIIH)(IH)))	L11 52 GU
(N (M(IIIH)(IH)))	L11 53 AC
(N (M(IIIH)(IH)))	L11 53 AU
(N (M(IIIH)(IH)))	L11 74 CA
(N (M(IIIH)(IH)))	L11 75 CU
(N (M(IIIH)(IH)))	L11 76 CU
(N I (M(IIH)(H)))	L11 48 CU
(N III (M(H)(iH)))	L11 52 GC
(N(IIIH) (M(H)(H)))	L11 53 AG
(N(IIIH)(i(M(H)(H))))	L11 51 GA

Alignment of the secondary structures of L11 mutations by their string representation. Small letters indicate "mismatches" in the aligned structures.

mutations that are not clustered with the wild type, only one is a false prediction (52:G→C).

As was mentioned above we ran the analysis for 241 structures and clustered them as described above. Among those, 148 were found to have the same structural features as the wild type. The structures of the other 92 mutations deviate to different degrees from the wild type structure. We can further narrow this group by searching for a motif of interest, for example, adenine in the loop or a stem of length 3 preceding this loop, and concentrating on the structures that miss this motif as candidates for further study.



```
ENTER PATH TO TAXONOMY TREE FILE NAME:opus:/users/opus/bshapiro/nygenalign/l11/l11-pap-0-new.tax
Secondary Structure Analysis Systen connand: Screen Dump
[10:53:50 Your request of 12/23/88 10:30:53 ("Screen Hardcopy") has finished printing on WOPR.]
Secondary Structure Analysis Systen connand: Screen Dump
Mouse-L, -R: Screen Dump.
To see other commands, press Shift, Control, Control-Shift, Meta-Shift, or Super.
[Fri 23 Dec 10:53:56] bshapiro                CL-USER:       Run
```

FIG. 3
A taxonomy tree representation for the clustering of the different mutations. This representation is generated by an algorithm that computes the pairwise distances between the ordered trees representing the secondary structures (26).

After detecting a mutation that affects the regulation, compensatory mutations can be searched similarly. The input sequence includes one point mutation and all the possible second mutations are generated by the computer. By clustering the structures, the double mutants that restore the wild type structure can be found. The next example illustrates the analysis of compensatory mutations.

For the second example we analyze the sequence of the lamB gene of E. coli that was studied by Hall et al. (10) (Fig. 4a). The wild type Shine - Dalgarno sequence is partially contained in an unstable stem. The single mutants, 701 and 708, with reduced translation efficiency, may enable base pairing that stabilizes this stem and makes the Shine-Dalgarno sequence less accessible to the ribosome. In the double mutant, 701-708, with almost wild type activity, the formation of this stable stem is probably prevented and a

```
a    701                        708
      G                          U
      ↑                          ↑
AAUGACUCAGGAGAUAGAAUGAUGAUUACUCUGCGCAAA


b          10        20
  AAUGACU   G  A  GAAU
         CAG AG UA    G
         GUC UC AU    A
  -AAACGC   -  -  UAGU
 30


c          10        20
  AA  A     G  A  GAAU
    UG CGCAG AG UA    G
    AC GCGUC UC AU    A
  AA  -     -  -  UAGU
               30
```

FIG. 4
a)The region of translation initiation of the lamB gene. Underlined are the Shine-Dalgarno sequence and the AUG. The mutations 701 and 708 studied in Hall et al.(18) are indicated. b) The secondary structure with the minimum free energy for the wild type. c) Predicted secondary structure of the mutant 701.

structure similar to that of the wild type is formed. We show below how the computerized system can automatically detect those double mutations with a compensatory effect on the structure, including the double mutant 701-708 that appears in the study.

The examined sequence is the region around the translation start position of the lamB gene (Fig. 4a), of length 39 nucleotides. The postulated structure for the wild type sequence by the FOLD program is presented in Figure 4b. Our input sequence includes the mutation 701 (U→G in position 7). In the predicted secondary structure the Shine-Dalgarno sequence is included in a more stable stem (Fig. 4c). Note that the structures that are predicted here by the FOLD program differ slightly from the structures that Hall et al. (10) predicted manually.

Since Hall et al. (10) did not require amino acid conservation for the mutations downstream from the AUG we also do not include such a requirement. The program introduces all the possible point mutations (three substitutions

```
         10        20
AAUGAC  -AG   A  GAAU
      GC   GAG UA    G
      CG   CUC AU    A
———AAA  AGU   -  UAGU
           30
(C→A in position 34)
```

```
       10       20
AA  A  AG   A  GAAU
 UG CGC  GAG UA    G
 AC GCG  CUC AU    A
AA  -   -A   -  UAGU
         30
(U→A in position 32)
```

```
         10        20
AAUGAC  -AG   A  GAAU
      GC   GAG UA    G
      CG   CUC AU    A
———AAA  GGU   -  UAGU
           30
(C→G in position 34)
```

```
       10       20
AA  A  AG   A  GAAU
 UG CGC  GAG UA    G
 AC GCG  CUC AU    A
AA  -   -G   -  UAGU
         30
(U→G in position 32)
```

```
       10        20
AA  A  AAG   A  GAAU
 UG CG   GAG UA    G
 AC GC   CUC AU    A
AA  -  -GU   -  UAGU
         30
(C→A in position 8)
```

```
       10        20
AA  A  CAG   A  GAAU
 UG CG   GAG UA    G
 AC GC   CUC AU    A
AA  -  -AU   -  UAGU
         30
(G→A in position 33)
```

FIG. 5
Structures of the double mutants with a structural compensatory effect identified by the algorithm.

in each position and therefore a total of 117 sequences of double mutants). The true revertant and AUG mutations are also generated automatically but can be ignored afterwards in the analysis. The goal of the computer is to find all the mutations that are compensatory, i.e., restore the wild type structure. We can either use the tree clustering schemes and locate the structures that are similar to the wild type (as shown in the previous example) or look for an essential motif that occurs in the wild type. For the lamB case the program that clusters similar structures according to their morphological entities is less useful, because the differences between the mutants and wild type are due to more refined details of the structure and not to general differences in the morphological entities. In such a case the program may be misleading, detecting structures with morphological entities similar to those of the wild type but with the Shine-Dalgarno sequence included in a stable stem. An extreme example is the double mutant that includes mutation 701 and an A→U substitution in position 28:

```
        10
AA  A     -     AGAA
  UG CGCAG GAGAU
  AC GCGUC CUUUA    U
AA  -     U     GUAG
         30         20
```

The string representations of both the wild type and the double mutant is BBH. The algorithm that aligns the structures by their string representation will cluster this mutation together with the wild type and therefore will list it as a possible compensatory mutation although the Shine-Dalgarno sequence here is completely buried in the two stems. Therefore, in the lamB system, where the length of the stem in the Shine-Dalgano region is suspected to be critical, we should either use the second way of tree comparisons which takes into account the size of the morphological entities or look for a certain motif that appears in the wild type. Since the claim was that the difference between the wild type and the mutant was in the stability of the stems in the Shine-Dalgarno region, we decided to look for double mutants with such a motif. In other words, we look for a sub-structure around the Shine-Dalgarno sequence that will not be more stable than this sub-structure in the wild type (which had a free energy value of -4.3 kcal/mol). The program locates 6 such potential compensatory mutations (Fig. 5). Mutation 708, C→A in position 34 which was predicted and proven by Hall et al. (10) to have a compensatory effect on the structure is detected by the system. In addition, the system identifies potential compensatory mutations in other positions : 8, 32, and 33. The structures of these double mutants are very similar to that of the wild type and, actually, would be interesting to test whether their effect on the activity is similar to that of the double mutant 701- 708.

When repeating the analysis with mutant 708 as the input sequence and searching for compensatory mutations, mutation 701 and other 9 mutations in positions 7, 8, 32, and 33 are located as candidates.

DISCUSSION

The ability to compare a large number of RNA secondary structures by computer is very useful for several purposes: 1) For the identification of a consensus structure in molecules with the same function - the program can compare either the structures that are generated by the FOLD algorithm or structures that are provided by the user, and by clustering the structures detect a consensus structure. 2) For evaluation of the effect of point mutations on the secondary structure, detecting point mutations that disrupt or double mutants that restore the wild type structure - the program automatically generates all possible mutated sequences and their structures by the FOLD algorithm. Common structures are clustered together enabling the researcher to detect mutations with structural effects of interest. 3) For comparison of sub-optimal structures that are generated by the FOLD algorithm - frequently it is important not only to examine the optimal structure, which is the one with the minimum free energy, but also sub-optimal structures with free energy values close to the minimum value. The program can compare these sub-optimal structures and detect consensus features in them.

In addition to comparisons and clustering, one of the algorithms in the system can search for the presence of a structural motif of interest. The motif can be defined either as a sequence specific or as a sequence independent feature. The motif can be either purely a morphological entity, a morphological entity of specific size, or a region of certain stability (as evaluated by free energy computations).

We mainly concentrated in this report on the use of the computerized system for mutation analysis, demonstrating how the programs can be used to predict mutations with structural effects for further studies of structure-function relationship. We presented examples to clarify the potential of the system, however, its great advantage is in its ability to analyze a very large number of sequences very fast, so that practically, the researcher can examine all possible point mutations in a sequence. For example, the analysis of 241 L11 structures was completed in less than 1 hour: the folding of 241 sequences needed about 36 minutes of CPU time, the comparison and clustering took about 15 minutes and the other steps (generation of the mutations and translation of the secondary structures into strings of characters) required only a few minutes.

The computerized system is of great practical value assuming that the computer predicted secondary structures are correct. The shortcoming of the system concerns the validity of the computer predicted structures. It is not clear to what degree the structures that are derived by free energy computations reflect the native secondary structure of the RNA. The best way to evaluate the secondary structure is to compare the structures with the minimum free energy that FOLD predicts to known structures. However, the secondary structures of very few sequences have been determined experimentally by biochemical means, and the interpretations of these biochemical results is some times ambiguous. Nevertheless, the few structures for which biochemical data is available (21,22) agree with the computer predicted structures quite well. Attempts to improve the predictions are continuously in progress, either by using more accurate free energy values (28, 29, 30), or by performing different variations on the folding algorithm like following the folding chain as it grows and searching for conserved sub-structures (31,32). While there is no doubt as to the importance of developing better predictive algorithms, it is important to note that the correctness of the algorithms presented here is independent of the folding algorithms. The computerized system is able to efficiently cluster together similar secondary structures and to search over a very large number of structures for a structural motif of interest. At present, when the input data is only primary sequence information, the system analyzes the secondary structures predicted by free energy computations since this is the best predictive scheme available. As our understanding of the folding process expands and better structure prediction algorithms are developed the practical value of our system will be even greater.

ACKNOWLEDGMENT
We thank Kathye Currey for reading the manuscript.

REFERENCES

1) Noller, H.F. (1984) Ann. Rev. Biochem. 53, 119-162.
2) Stormo, G.D. (1986) In : Maximizing Gene Expression Resnikof, W.R. and Gold, L. (eds.) , Butterworth Publishers, Stoneham, MA. pp. 195-224.
3) Buel, G. and Panayotatos, N. (1986) In: Maximizing Gene Expression Resinokof, W.R. and Gold, L. (eds.) , Butterworth Publishers, Stoneham, MA. pp. 345-363.
4) Gold, L. and Stormo, G. (1987) In: Escherichia coli and Samonella Typhimirium. Cellular and Molecular Biology Neidhardt, F.C. (ed.) American Society of Microbiology, Washington, D.C. pp. 1302-1307.
5) Ganoza, M.C., Kofoid, E.C., Marliere, P. and Louis, B.G. (1987) Nuc. Acids. Res 15, 345-360.
6) Wood, C.R., Ross, M.A., Patel, T.P. and Emtage, J.S. (1984) Nuc. Acids. Res 12, 3937-3950.
7) Buell, G.N., Schulz, M.F., Selzer, G., Chollet, A., Movva, N.R., Semon, D., Escanez, S. and Kawashima, E. (1985) Nuc. Acids. Res 13, 1923-1938.
8) Tessier L.-R., Sondermeyer, P., Faure, T., Dreyer, D., Benavente, A., Villeval, D., Courtney, M. and Lecocq J.-P. (1984) Nuc. Acids. Res. 12, 7663-7675.
9) Zagorska, L., Chroboczek, J., Klita, S. and Szafranski, P. (1982) Eur. J. Biochem. 122, 265-269.
10) Hall, M.N., Gabay, J., Debarbouille, M. and Schwartz, M. (1982) Nature 295, 616-618.

11) Thomas, M.S. and Nomura, M. (1987) Nucleic Acids Research 15, 3085-3096.

12) Nomura, M., Yates, J.L., Dean, D. and Post, L. (1980) Proc. Natl. Acad. Sci. USA 77, 7084-7088.

13) Baughhman, G. and Nomura, M. (1984) Proc. Natl. Acad. Sci. USA 81, 5389-5393.

14) Romaniuk, P.J., Lowary, P., Wu, H.-N., Stormo, G. and Uhlenbeck, O.C. (1987) Biochemistry 26, 1563-1568.

15) Wu, H.-N. and Uhlenbeck, O.C. (1987) Biochemistry 26, 8221-8227.

16) Springer, M., Plumbridge, J.A., Butler, J.S., Graffee, M., Dondon, J., Mayaux, J.F., Fayat, G., Lestienne, T., Blanquet, S. and Grunberg-Manago, M. (1985) J. Mol. Biol. 185, 93-104.

17) Springer, M., Graffee, M., Butler, J.S. and Grunberg-Manago, M. (1986) Proc. Natl. Acad. Sci. USA 83, 4384-4388.

18) Butler, J.S., Springer, M., Dondon, J. and Grunberg-Manago, M. (1986) J. Bacteriol. 165, 198-203.

19) Moine, H., Romby, P., Springer, M., Grunberg-Manago, M., Ebel, J.-P., Ehresmann, C., and Ehresmann, B. (1988) Proc. Natl. Acad. Sci. USA 85, 7892-7896.

20) Zuker, M. and Stiegler, J. (1981) Nuc. Acids. Res. 9, 133-148.

21) Kearney, K. and Nomura, M. (1987) Molec. Gen. Genet. 210:60-68.

22) Inoue, T. and Ceck, T.R. (1988) proc. Natl. Acad. Sci. USA 82, 648-652.

23) Shapiro, B.A., in preparation.

24) Shapiro, B.A. (1988) Comp. App. in the Biol. Sci. 4, 387-393.

25) Sobel, E. and Martinez, H.M. (1986) Nuc. Acids. Res 14, 363-374.

26) Shapiro, B.A. and Zhang, K., in preparation.

27) Zhang, K., and Shasha, D. (1988) Society of Industrial and Applied Mathematics, in press.

28) Freier, S.M., Ryszard, K., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T., and Turner, D.H. (1986) Proc. Natl. Acad. Sci. USA 83, 9373-9377.

29) Turner, D.H., Sugimoto, N., Jaeger, J.A., Longfellow, C.E., Freier, S.M., and Kierzek, R. (1987) Cold Spring Harbor Symp. Quant. Biol. 52, 123-133.

30) Turner, D.H., Sugimoto, N. and Freier, S.M. (1988) Ann. Rev. Biophys. Biophys. Chem. (1988) 17, 167-192.

31) Le, S., Currey, K.M., Nussinov, R., and Maizel, J.V. (1987) Computer. Biomed. Res. 20, 563-582.

32) Le, S., Chen J-H., Nussinov, R., and Maizel, J.V. (1988) Comp. App. in the Biol. Sci. 4, 337-344.