Nucleic Acids Research, 1995, Vol. 23, No. 21 ***4471–4479***

Alignment editing and identification of consensus secondary structures for nucleic acid sequences: interactive use of dot matrix representations

Jeffrey P. Davis, Nebojsa Janjić, David Pribnow[1] and Dominic A. Zichi*

NeXstar Pharmaceuticals Inc., 2860 Wilderness Place, Boulder, CO 80301, USA and [1]Oregon Health Sciences University, 3181 SW Sam Jackson Park Road, Portland, OR 97201, USA

Received June 21, 1995; Revised and Accepted September 20, 1995

ABSTRACT

We present a computer-aided approach for identifying and aligning consensus secondary structure within a set of functionally related oligonucleotide sequences aligned by sequence. The method relies on visualization of secondary structure using a generalization of the dot matrix representation appropriate for consensus sequence data sets. An interactive computer program implementing such a visualization of consensus structure has been developed. The program allows for alignment editing, data and display filtering and various modes of base pair representation, including co-variation. The utility of this approach is demonstrated with four sample data sets derived from *in vitro* selection experiments and one data set comprising tRNA sequences.

INTRODUCTION

Randomized oligonucleotide libraries contain molecules with a wide spectrum of potential functional properties. High efficiency screening of these libraries with SELEX (Systematic Evolution of Ligands by EXponential enrichment) technology has recently led to identification of oligonucleotides with unique binding and catalytic features (1–7). In most SELEX experiments iterative rounds of selection–amplification are carried out until the majority of molecules in the enriched pool share the functional property for which they were selected. In a typical successful SELEX experiment cloning and sequencing of molecules from the enriched pool results in a collection of sequences. As a consequence of the functional relatedness among clones, these molecules can typically be classified into families based upon similar primary structure (3,8–11). Functional relatedness also implies that a secondary structural motif may be common to members of a family.

While basic elements of the consensus primary structure in a similar sequence set are recognized either by inspection or with the use of multiple sequence alignment algorithms (12; B.Javornik and D.A.Zichi, unpublished results), recognition of conserved secondary structural elements is often problematic, especially in a large data set (13,14). The search for a common structural motif is often complicated by the existence of many individual sequences with mutually exclusive potential secondary structures that may have similar predicted free energies of folding. Indeed, analyzing a set of sequences by examining optimal and suboptimal foldings from a Zuker–Turner secondary structure prediction program (15,16) rarely results in identification of a common secondary structure. In addition, such a scheme will not identify certain well-known secondary structural motifs, such as pseudoknots and G-quartets, which have been found in SELEX experiments (8,17). In this report we describe the use of cumulative dot matrix representation (18) as an aid in determining consensus secondary structure for functionally related nucleic acids with similar primary structures.

The utility of dot matrix representation for visualizing secondary structures of oligonucleotides for single sequences has been recognized previously (14,18). In brief, the rows and columns of the structure matrix, **M**, represent the oligonucleotide sequences written in a 5′→3′ direction. A dot is placed on an individual matrix element M_{ij} if bases i and j are a potential Watson–Crick base pair (the definition of thermodynamically favorable base pairing can easily be expanded to include non-canonical base pairs, e.g., G-U pairs). Double helical (stem) regions are then easily recognized as runs of sequential dots perpendicular to the diagonal. Such a representation is symmetric about the diagonal, so that the two symmetric halves of the matrix can be used for different display purposes. We present a generalization of this dot matrix display to visualize consensus secondary structures within similar sequence sets that were initially aligned by sequence identity considerations alone. This initial alignment can then be altered with the goal of optimizing the extent of conserved base pair formation. The method described here combines the use of the high resolution color graphics capabilities of modern microcomputers with the well-known ability of the human eye to recognize sophisticated patterns. In essence, the viewer is presented with an image of aligned and overlaid matrix representations of all secondary structure possibilities within the sequence set. In this composite image color is used to highlight the regions where consensus base pairing is observed. In the supporting computer software we provide interactive features that allow for editing of the overall sequence alignment, modification of various

* To whom correspondence should be addressed

parameters that affect the display of consensus secondary structure, display editing that allows progressive filtering or pruning of the display from complex (all inclusive) to simpler (showing only the most conserved and most stable regions), easy identification of mutually exclusive structures, detection of base pairing co-variation and detection of G-quartet structures.

METHODOLOGY

Input format

Functionally related oligonucleotide sequences are first arranged in sets aligned according to primary structure similarity using, for example, the Feng–Doolittle (12) or CLUSTER multiple sequence alignment algorithm. (CLUSTER is a program that performs multiple sequence alignment with re-optimization of gap placement within the growing consensus alignment. The algorithm consists of two parts: sequence alignment and clustering. Sequence alignment is done taking gaps into account with a dynamic programing algorithm. An alignment cost, normalized by the number of sequences, is computed based upon the sum of the pairwise alignments for all sequences in the consensus. Sequences are clustered into families by first allowing all sequences to define cluster centers and then systematically merging the closest, or most similar, centers based on their alignment costs. A new center is computed at each step by optimizing the alignment for gap placement. This is done by successively removing each sequence within the cluster followed by re-alignment until no changes in the alignment occur. Clustering ends when no two centers are similar enough to be combined.) The relative offsets and gap placements contained in the aligned sequence file are preserved upon input. In SELEX experiments the random oligonucleotide region is flanked by defined (fixed) sequence regions which are required for amplification by the polymerase chain reaction (1,2). Only the evolved (initially random) regions are considered in alignment, but the fixed regions are re-introduced prior to generation of the secondary structure matrix.

Consensus structure matrix

Elements of the consensus structure matrix can be computed for any well-defined secondary structure. We have currently implemented three different measures of secondary structure, namely Watson–Crick base pairing, base pairing co-variation and G-quartet formation.

The structure matrix for expanded Watson–Crick base pairing (standard Watson–Crick including G-U wobble base pairs) is computed according to equation **1**

$$M_{ij} = 1/N \, \Sigma_n \, \Sigma_{b:b'} \, C(b:b') \, \delta(a_{n,i} - b) \, \delta(a_{n,j} - b') \qquad (1)$$

where $a_{n,i}$ designates the nucleotide at position i of sequence n in the N-sequence multiple alignment and $b:b'$ indicates one of the four Watson–Crick or two G:U base pairs. $C(b:b')$ is a general coefficient that can, for example, reflect the energy of base pair formation and $\delta(a_{n,i} - b)$ is the Kronecker delta function, equal to 1 if $a_{n,i} = b$, otherwise equal to 0. We impose a minimum length hairpin loop to be 2 nt long, so all diagonal elements within two positions of the main diagonal are set to zero. In the following we set $C(b:b') = 1$ for all base pairs $b:b'$, so M_{ij} varies from 0, indicating *no* base pair formation between positions i and j in any of the N sequences, to 1, indicating that *all* sequences form a base pair between positions i and j. For this definition of **M** the consensus matrix simply represents the degree to which base pairing is conserved within a set of similar sequences.

Detection of base pairing co-variation (change in concert at two positions in a sequence according to Watson–Crick rules) is of incomparable value for secondary structure predictions (13,14). In the absence of experimental data, co-variation analysis provides the most reliable means of detecting base pairs in RNA. It is useful therefore to include a structure matrix representing base pairing co-variation determined by the following mutual information content formula (14),

$$M_{ij} = 1/2 \, (\Sigma_{b:b'} f_{bi,b'j} \log_2[f_{bi,b'j}/f_{bi} f_{b'j}]) \qquad (2)$$

where

$$f_{bi} = 1/N \, \Sigma_n \, \delta(a_{n,i} - b)$$

and

$$f_{bi,b'j} = 1/N \, \Sigma_n \, \delta(a_{n,i} - b) \, \delta(a_{n,j} - b')$$

The f_{bi} term is the fraction of sequences which have a base b at position i in the multiple alignment and $f_{bi,b'j}$ is the fraction of sequences which form a $b:b'$ base pair at positions i and j. Here, complete co-variation of Watson–Crick base pairs at positions i and j results in $M_{ij} = 1$, while *either* no base pairing *or* no co-variation of the structure at positions i and j results in $M_{ij} = 0$. Intermediate values of M_{ij} are possible for cases where less than complete co-variation is observed. Although we limit the co-variation analysis to expanded Watson–Crick base pairs in the current version of the program, extending this analysis to include co-variation of any type would be straightforward. However, general co-variation analysis requires substantially more data than considered here in order to generate meaningful statistical information.

Detection of possible G-quartet structures is also readily accomplished by computing M_{ij} from equation **3**

$$M_{ij} = 1/N \, \Sigma_n \, \{\delta(a_{n,i} - G) \, \delta(a_{n,j} - G) \times \theta[\delta(a_{n,i+1} - G) \, \delta(a_{n,j-1} - G) + \delta(a_{n,i-1} - G) \, \delta(a_{n,j+1} - G)]\} \qquad (3)$$

where $\theta(x)$ is the Heaviside step function, equal to 1 for $x > 0$ and zero otherwise. Here M_{ij} is the fraction of potential G:G quadruplex base pairs that is bracketed by an additional G:G base pair either above or below. Of course, although not sufficient to describe a G-quartet, this condition is necessary. The final step in identification of the G-quartet is made visually on the screen, as is a stem structure in the previous examples.

It is useful to note that **M** could also be input from other application programs, most notably the energy matrix from a Zuker–Turner free energy evaluation (15,16). This feature, for example, would allow for selection of consensus structures based upon energetic considerations. In the current implementation we only consider the relatively simple structural measures outlined above for several reasons. First, free energy considerations are most appropriate for solution conformations; in the case of SELEX data sets, however, the consensus structure is likely to correspond to the *bound* conformation and not a solution minimum free energy structure (see below). In addition, during thermodynamic energy minimization many potential secondary structures are lost from consideration, even though some suboptimal foldings may be identified. Our measures of structure are all inclusive. Finally, since the program is used primarily as an interactive alignment tool, we require rapid calculation of the structural measures. Thermodynamic

algorithms are currently too computationally demanding for interactive use on personal computers.

In order to facilitate visual detection of conserved secondary structural elements, we use a color-coded representation of the 'strength' of the consensus at each matrix position. This color coding is used for viewing all measures of structure defined above. Red indicates $M_{ij} = 1.0$ (completely conserved secondary structural element defined by **M**), purple is $M_{ij} = 0.75$, dark blue is $M_{ij} = 0.5$, light blue is $M_{ij} = 0.25$ and white indicates $M_{ij} = 0.0$ (complete absence of that secondary structure). For M_{ij} between these values colors are generated continuously from these end points. A strong consensus stem structure will be recognized as a contiguous red line perpendicular to the main diagonal. For a given sequence set the consensus matrix pattern is clearly a signature of the consensus secondary structure. Quigley *et al.* (18) have provided a fairly complete description of the types of secondary structures represented by various matrix patterns.

From equations **1–3** it is obvious that **M** is symmetric about the diagonal M_{ij}. This symmetry diagonal divides the matrix display into two triangles that contain the same information and can therefore be used for different display purposes. The upper right triangle contains all the unfiltered data computed from the current alignment. The lower left triangle can then be used to display filtered data (see below).

Output display

The primary output of the program is a color graphical display of the consensus structure matrix along with the current multiple sequence alignment. The interactive features of the program, as well as a detailed description of the display, are summarized in Figure 1.

Display filters

The full set of data embodied in the consensus matrix is generally quite dense, creating a crowded display. In order to facilitate identification of strong consensus structures a set of two filters for **M** has been provided. The filtered data appears in the lower left triangle, while the unfiltered data remains displayed in the upper right triangle. Currently there are two filters available, a consensus strength filter and a stem length filter. Both filters have values that may be adjusted interactively.

The consensus strength filter sets a threshold value on M_{ij} for display. Only those M_{ij} points with values above the threshold will be displayed in the lower left triangle. In the case of equation **1** the threshold value reflects the fraction of sequences in the multiple alignment that form a base pair at positions i and j. In the case of equation **2** the threshold value reflects the extent of base pair co-variation at the individual sites. In equation **3** the threshold value reflects the fraction of sequences in the multiple alignment that potentially form quadruplex G:G base pairs consecutively with a stack height of at least two. The second filter sets a minimum stem length value, between one and six, that serves as the lower limit for the number of contiguous base pairs that make up a stem. It should be noted that any point that appears in the filtered display matrix must satisfy *both* of the above filters. The two filters, however, can be adjusted separately so that the consensus strength and the stem length can be examined independently.

In addition to the filters that modulate the matrix display, a filter that controls the threshold level for displaying nucleotide base conservation is provided (shown to the left of the structure

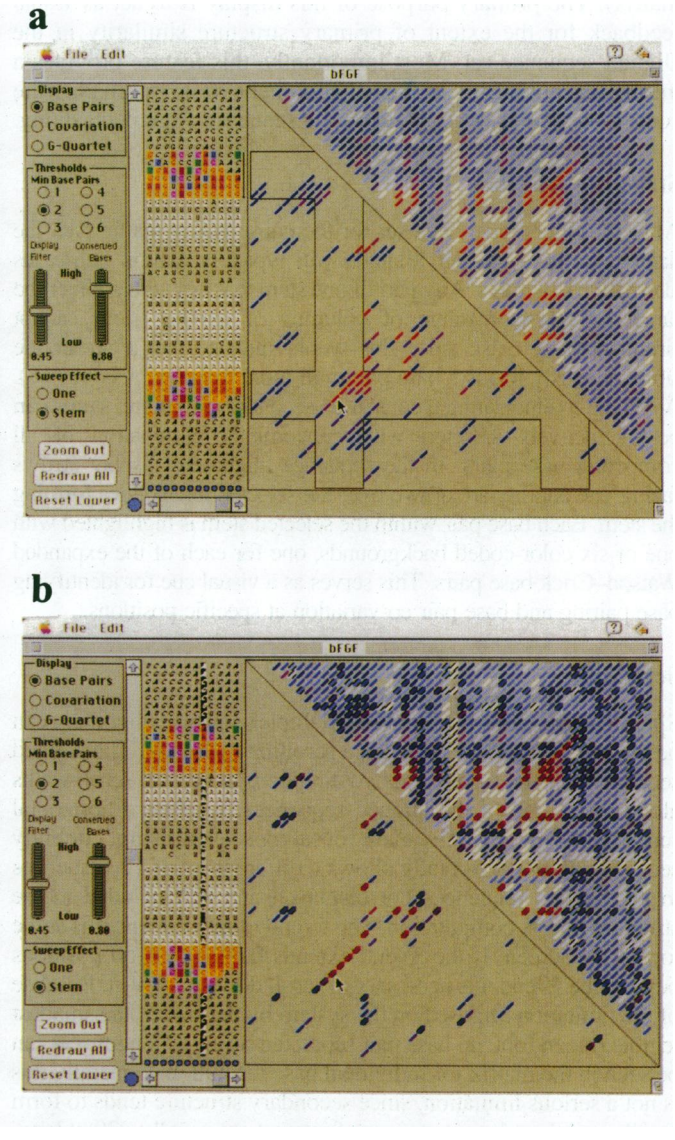

Figure 1. (**a**) Consensus structure matrix program display for RNA ligands which bind to bFGF. The upper right triangle shows the unfiltered results. The lower left triangle displays the filtered data, at 65% for consensus base pairs and a minimum stem length of 2 bp. Selecting the stem designated by the arrow in the lower left triangle results in removal of all competing structures contained within the polygon outlined in black within the structure matrix [see (b) below]. An option for removing competing structures 1 bp at a time is also available under the sweep effect menu. The aligned sequences with the 5′-end at the top are shown on the left hand side of the consensus matrix. Bases conserved in at least 95% of the sequences are shown in black on a white background. Base pairs resulting from the selected stem are displayed in one of six color-coded backgrounds. Bases in the fixed sequence regions are shown in italics. Filters and other display controls are placed to the left of the sequences and are only shown in this figure. The current version of the program reads a sequence input file containing up to 40 individual sequences and each sequence may be up to 200 residues in length. (**b**) Illustration of the display feature for alignment editing which highlights a single chosen sequence. The sequence set and filter settings are as in (a) after pruning competing structures for the 6 bp stem designated above. The chosen sequence is highlighted by changing all of its bases to white on a black background, with the exception of those base pairs involved in previously selected stems, which remain color-coded as described earlier. In both halves of the structure matrix all of the displayed base pairs to which the selected sequence contributes are outlined in black, allowing for easy identification of its possible secondary structural elements.

matrix). The primary purpose of this display is to act as visual feedback for the extent of primary structure similarity in the aligned sequence set. Most importantly, this feature provides a means for monitoring sequence alignment while optimizing the consensus secondary structure during sequence alignment editing.

Structure selection

After the filters have been set for the consensus structure data the display in the lower left triangle will typically exhibit numerous alternative structures. Any part of one stem that lies within either the range of rows or range of columns of another stem cannot simultaneously exist, since this would indicate that a nucleotide simultaneously forms a base pair with more than one residue (18). As an aid to elucidating a consensus secondary structure, stems can be interactively selected with concomitant elimination of all competing base pairs. In the sequence alignment display arrows facing one another are drawn over the selected nucleotides forming the stem. Each base pair within the selected stem is highlighted with one of six color-coded backgrounds, one for each of the expanded Watson–Crick base pairs. This serves as a visual cue for identifying base pairing and base pair co-variation at specific positions.

Scrolling

Since in some cases not all of the multiple sequence alignment can be viewed at once, we included a scrolling operation. The aligned sequences can be scrolled in two directions; a vertical scroll walks along the length of the aligned sequences, whereas a horizontal scroll walks into the alignment revealing sequences not currently seen. Scrolling horizontally allows a different set of 12 sequences to be visible. Sequences not displayed are still included in the structure matrix computation. The 5′-end of the sequences is at the top of the screen. If the 3′-end extends farther than 50 positions beyond the 5′-end it is necessary to scroll down to view it. Because of the limitation imposed on the system by the size of the smallest legible screen font, no base pair separated by an i–j length >50 can be seen in the matrix for individual base resolution. Generally this is not a serious limitation, since secondary structure tends to form locally and the relevant regions of interest are usually <30 nt long. Scrolling down the positions in the sequences corresponds to sliding the window down the symmetry diagonal of the matrix. Figure 2 demonstrates that moving the matrix window along the diagonal of a sequence set whose length is >50 necessarily omits from view the structure due to bases in the upper right and lower left triangles of the complete matrix. However, a zoom option allows for display of the entire matrix at reduced resolution.

Alignment editing

Since the original alignment of a set of related sequences is generally done with no consideration of secondary structure, it is usually necessary to alter the alignment in order to optimize the consensus secondary structure. The program allows for interactive editing of the alignment by providing a means for both repositioning one sequence relative to the remaining set and for gap placement within a sequence. Both editing functions occur within the window containing the current alignment. Upon rearrangement of the multiple alignment both halves of the structure matrix display are recomputed and redrawn.

In order to facilitate alignment editing it is desirable to view the contribution to the structure matrix of an individual sequence in the context of the others. This allows one to determine whether the secondary structure of an individual sequence can be brought into register with the main consensus secondary structure by either gap placement or by repositioning the sequence. This feature is illustrated in Figure 1b. Finally, sequences that are more distantly related in primary structure to the rest of the members can be temporarily removed from the set and their contribution to the matrix calculation excluded.

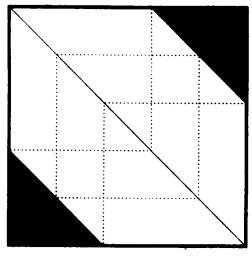

Figure 2. Hexagonal section of the consensus structure matrix visible while scrolling up and down through sequences longer than 50 nt. Only the square matrices indicated by the dotted lines can be seen through normal scrolling. Employing the zoom feature allows one to view the entire matrix at the expense of being able to clearly read the reduced font size imposed on the sequence area.

Implementation

The software was written in ANSI C for a Macintosh personal computer. An executable version of the program is available on a 3.5 inch floppy disk. Contact JPD at davis@nexstar.com for more information.

RESULTS AND DISCUSSION

Our method relies on the tacit assumption that all molecules within a family have a common three-dimensional shape which is supported by a common secondary structure that allows key functional groups to adopt similar spatial positions. This is well appreciated for proteins and some nucleic acid molecules, such as tRNA or rRNA. Indeed, short oligonucleotides, because of their propensity to form stacked base pairs, are generally more structured than peptides of comparable sequence length. The key groups required for high affinity binding, for example, are typically manifest by high sequence conservation within a family of molecules. Identification of these consensus sequences, accomplished with existing multiple sequence alignment algorithms, allows us to focus on these regions for identification of conserved secondary structure. Examples presented below, as well as many others not discussed here (13), support the notion that residues conserved at the primary structure level are presented in a similar three-dimensional context, which in turn governs the functional properties of these molecules.

Nucleic acid molecules of the size described here (~20–30 bases for truncated SELEX molecules) are capable of assuming a multitude of secondary structures, generally spanning a wide range of free energies of folding. Some groups of conformations may exist in rapid equilibrium with one another and some may be kinetically trapped. A complex function of thermodynamic stability and kinetics of folding and conformational interconversion determines the overall conformational repertoire of any given sequence. In SELEX molecules in which the most populated solution conformer coincides with the active conformation have a

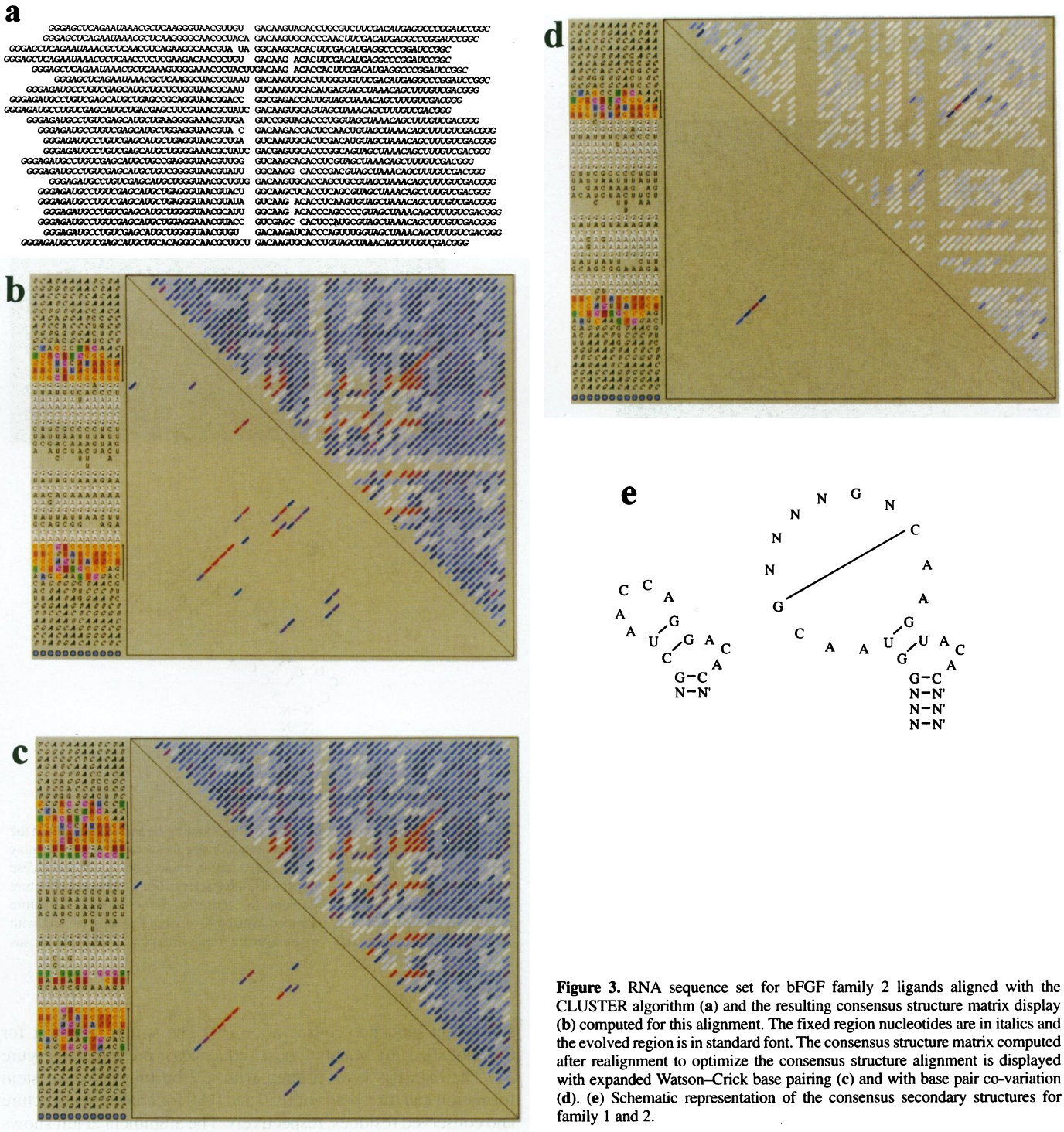

Figure 3. RNA sequence set for bFGF family 2 ligands aligned with the CLUSTER algorithm (**a**) and the resulting consensus structure matrix display (**b**) computed for this alignment. The fixed region nucleotides are in italics and the evolved region is in standard font. The consensus structure matrix computed after realignment to optimize the consensus structure alignment is displayed with expanded Watson–Crick base pairing (**c**) and with base pair co-variation (**d**). (**e**) Schematic representation of the consensus secondary structures for family 1 and 2.

selection advantage over those in which competing, inactive conformations predominate. However, molecules in which inactive conformations predominate can still be selected efficiently, as long as the active conformation is accessible. Therefore, one cannot deduce the active conformation by examining only the most thermodynamically stable set of conformations predicted for any one sequence. Indeed, one cannot reliably determine the most stable solution conformation for a single sequence either. A set of similar sequences that share a functional property is considerably more informative, since the active conformation must be accessible to all members, while alternative secondary structures will occur more randomly. Amplification of the secondary structure of the active conformation (signal) displayed in a matrix format in the background of competing structures (noise) is the basis of the method described here.

To illustrate the utility of our program for identifying consensus secondary structure and refining multiple sequence alignments we first examine sequence sets derived from four different

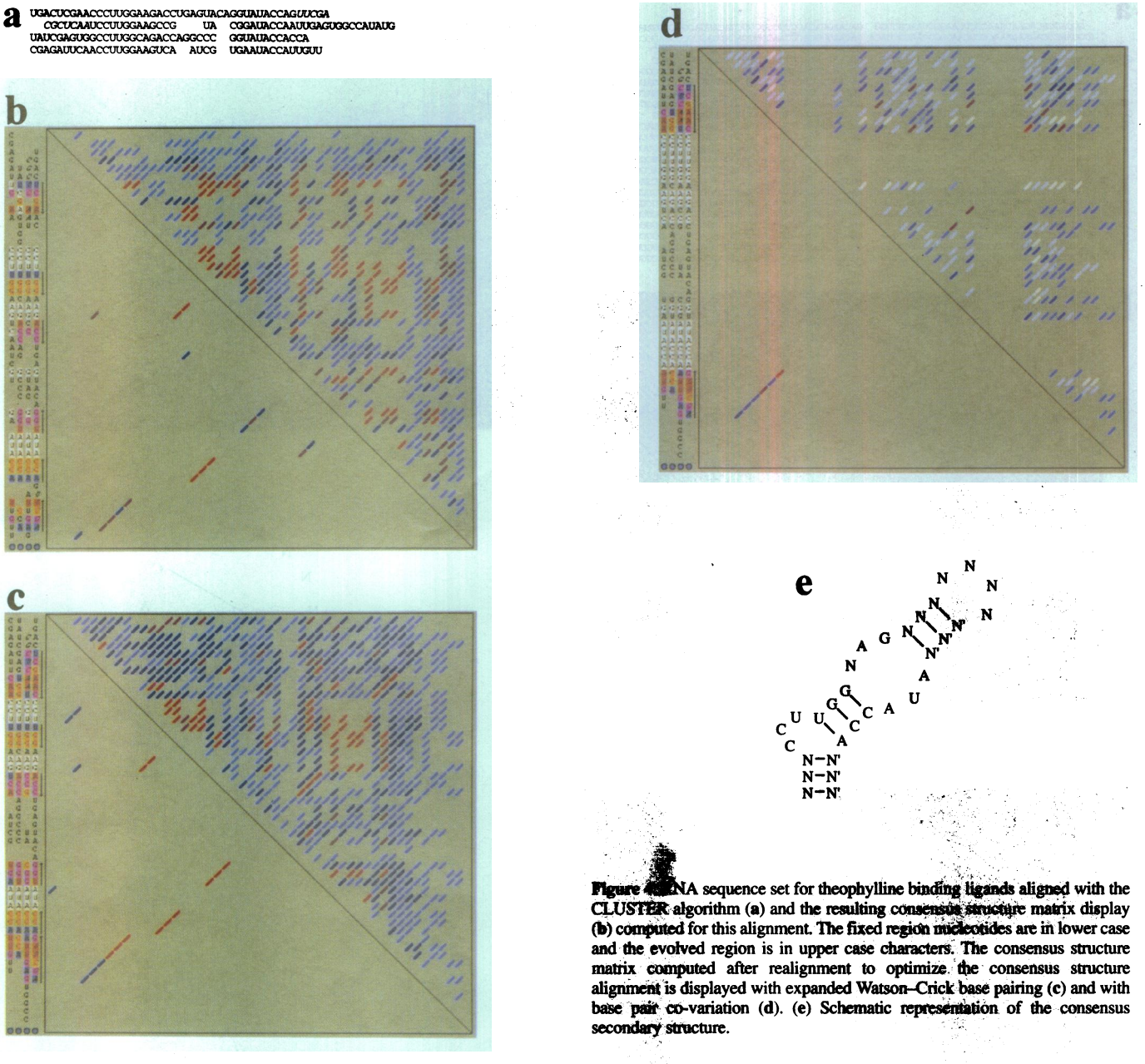

Figure 4. RNA sequence set for theophylline binding ligands aligned with the CLUSTER algorithm (**a**) and the resulting consensus structure matrix display (**b**) computed for this alignment. The fixed region nucleotides are in lower case and the evolved region is in upper case characters. The consensus structure matrix computed after realignment to optimize the consensus structure alignment is displayed with expanded Watson–Crick base pairing (**c**) and with base pair co-variation (**d**). (**e**) Schematic representation of the consensus secondary structure.

SELEX experiments. In each case we have used the CLUSTER algorithm to classify sequences into distinct families according to their primary structure similarity. As a first example we consider a group of RNA sequences that were selected for high affinity binding to basic fibroblast growth factor (bFGF) (9). Two distinct ligand families have been identified based on sequence and secondary structure similarities. Family 1 ligands are characterized by a consensus secondary structure motif that has a variable length stem with a 3 nt bulge followed by a highly conserved 5 nt loop. Most family 1 ligands bind to bFGF with dissociation constants of 3–20 nM. Family 2 is the larger of the two sequence sets (22 sequences; see Fig. 3a) and contains the highest affinity ligands, some of which bind to bFGF with dissociation constants as low as 0.2 nM (9). A common secondary structure motif for family 2 sequences has been identified by inspection and consists of a variable length stem closing a 19–22 nt loop that contains significant sequence conservation (9). The structure matrix for the published alignment of these sequences is presented in Figure 3b. The data has been filtered with a 2 bp minimum for stem formation and thresholds of 0.50 and 0.80 for consensus structure and conserved residues, respectively. The alignment at left shows several distinct regions of highly conserved bases, highlighted in white. The stem is easily identified as the long run of five contiguous base pairs which show strong consensus and co-variation. Several possible structures are evident within the loop region. The short stem of 2 nt, which results in the formation of a bulge within the main stem, allows this family of molecules to adopt a structural motif more similar to family 1. The revised alignment, in which this motif is enhanced, is presented in Figure 3c and d, with both a base pair and co-variation display. The former shows an extension of the primary stem by 1 bp and the latter indicates significant co-variation is present in this stem. The

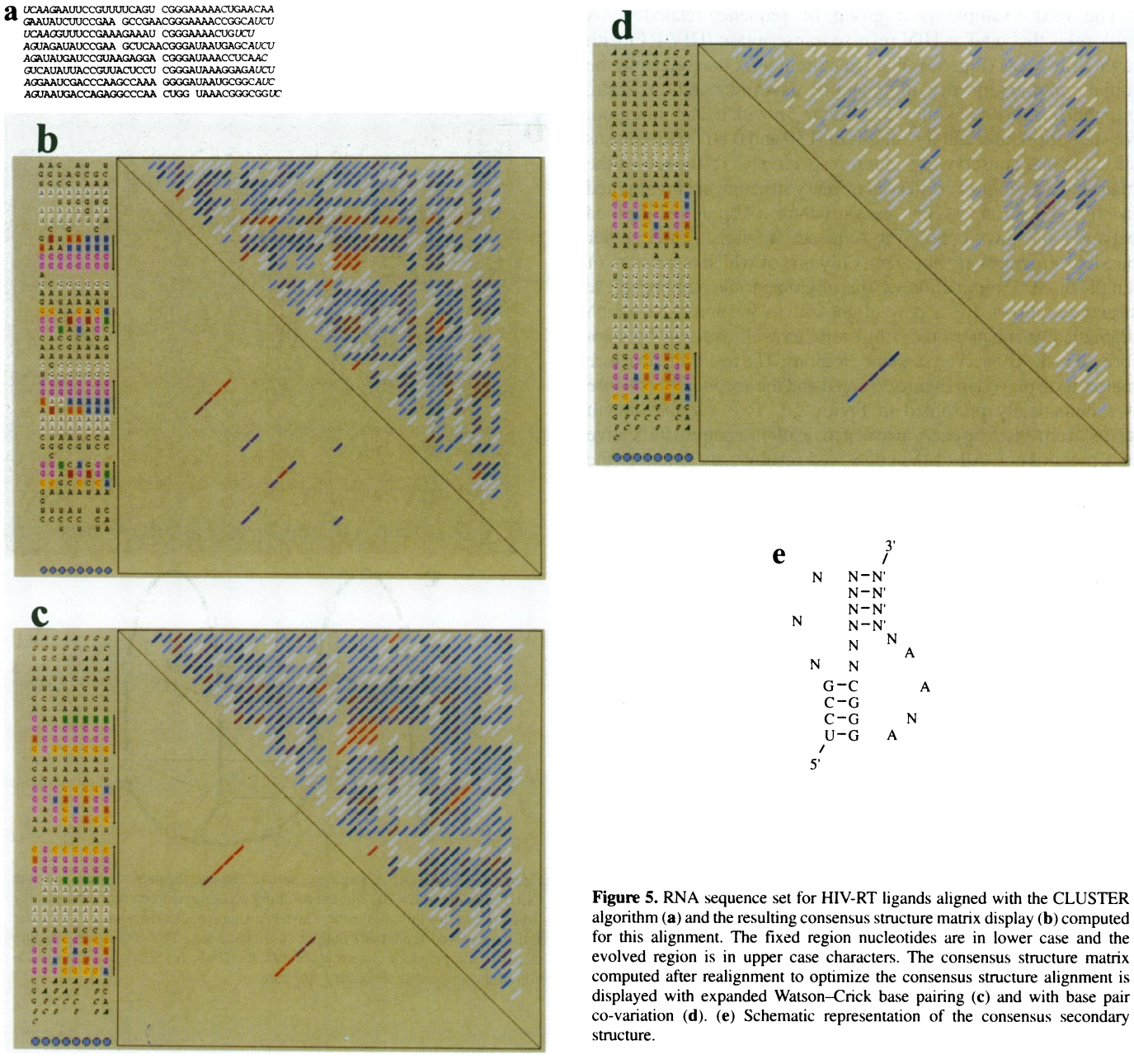

Figure 5. RNA sequence set for HIV-RT ligands aligned with the CLUSTER algorithm (**a**) and the resulting consensus structure matrix display (**b**) computed for this alignment. The fixed region nucleotides are in lower case and the evolved region is in upper case characters. The consensus structure matrix computed after realignment to optimize the consensus structure alignment is displayed with expanded Watson–Crick base pairing (**c**) and with base pair co-variation (**d**). (**e**) Schematic representation of the consensus secondary structure.

resulting secondary consensus structure is schematically displayed in Figure 3e, which includes, for comparison, the consensus structural motif of family 1 molecules. The sequence conservation in the bulge and loop regions for the two motifs is worth noting. Clearly, these two sets of sequences appear to share a structural motif which was recognized only after an examination of the consensus structure matrix presented here.

The next example we consider is a set of RNA molecules that were selected for their ability to bind the bronchodilator theophylline with high affinity (low micromolar range) and in enormous preference over caffeine (>10 000-fold lower affinity), a compound which differs from theophylline by a single methyl group at the N-7 position (19). Figure 4a displays the sequence set and Figure 4b shows the structure matrix resulting from the initial alignment, displayed to the left. In this case the lower left triangle has been filtered with a 3 bp minimum for stem formation and thresholds of 0.45 and 0.80 for consensus structure and conserved residues, respectively. The CLUSTER alignment algorithm finds two distinct regions of highly conserved bases, highlighted in white. The most conserved secondary structure is a short stem of length three formed by six absolutely conserved residues. Hence, no base pair co-variation is observed in this stem. Two other less conserved stems encompassing more variable positions also appear in the lower left triangle. With this as the starting point, we are able to optimize the degree of consensus structure formation by refining the original alignment (Fig. 4c and d). Strong consensus structures for all three stems are now observed. The two outermost structures are seen to co-vary significantly (see Fig. 4d). This co-variation is a strong indication of structure and is not obvious in the original sequence alignment (Fig. 4b). The resulting secondary structural motif for this set of molecules embodied in the matrix is shown in Figure 4e.

The next example is a group of sequence-related RNA molecules that bind to HIV reverse transcriptase (HIV-RT) with high affinity (8) (Fig. 5a). The structure matrix derived from the initial multiple sequence alignment (B. Javornik and D.A. Zichi, unpublished results) is shown in Figure 5b with a 2 bp minimum for stem formation and thresholds of 0.50 and 0.80 for consensus structure and conserved residues, respectively. To the left of the matrix two regions of conserved primary structure are highlighted in white. The strongest consensus structure, a 4 bp stem, is formed between the two conserved sequence regions. Three other possible structures are apparent, only one of which has a stem of length three. Examination of the alignment surrounding these structures allows for a new alignment to be proposed which dramatically enhances the 3 bp stem, as well as the 4 bp stem formed in the conserved sequence regions. The resulting structure matrix is displayed in Figure 5c and d and the secondary structure is schematically presented in Figure 5e. Two variable length stems, from 3 to 5 bp each, are seen to result in a pattern indicative of a pseudoknot. All of the conserved residues are found in the upper stem, while the lower stem has a variable composition with extensive base pairing co-variation. This is clearly illustrated in Figure 5d, where the structure is viewed as a co-variation matrix. Since base pair co-variation occurs much less frequently than base pairing, the co-variation structure matrix is less crowded. With lower background, positions where base pair co-variation is significant are readily detected, even without filtering, as evidenced in the upper right triangle of Figure 5d.

As our last SELEX example we present the consensus structure results for a set of sequences containing modified ribose moieties at pyrimidine nucleosides, namely a 2′-amino substituted for the 2′-hydroxyl group of RNA. The SELEX procedure was used to isolate 2′-amino-pyrimidine RNA molecules (Fig. 6a) which bind to human neutrophil elastase (20). These molecules are quite G-rich and were determined to be consistent with folding into a G-quartet motif. The resulting G-quartet structure matrix is displayed in Figure 6b. There are many regions of conserved sequence, primarily G residues, that appear in the alignment. The characteristic pattern of G-quartet formation, a set of at least two contiguous Gs interacting with three other sets of contiguous Gs, takes the form of a structure triangle in the lower left triangle of Figure 6b, comprising six pairs of G-G bases.

Our final example illustrates the utility of our cumulative dot matrix method for application to molecular taxonomy. We have computed the consensus structure matrix for a set of 40 tRNA sequences, all of which have a well-known 'clover leaf' secondary structure. The 40 aligned sequences were selected at random from the EMBL tRNA database of over 2000 sequences (21). The resulting consensus dot matrix is presented in Figure 7a and b for base pair and co-variation displays. The secondary structure comprising four stems is clearly illustrated by the four red diagonals of 3–7 bp in length observed in Figure 7a. In addition to these secondary structure interactions, several known tertiary contacts are also found, namely base pairs at positions 8/14, 18/55 and 19/56. These are not unique, however; several other potential base pairs which do not correspond to any known tertiary contacts can be observed in Figure 7a. The co-variation display is seen to highlight the four stems of tRNA dramatically and a new stem not seen in the base pair display is observed. This corresponds to the extra arm seen in some tRNAs in the so-called variable region. No tertiary contacts are seen in Figure 7b, in contrast to a similar analysis in Gutell *et al.* (14). This is most likely due to our limited sampling size of 40 sequences.

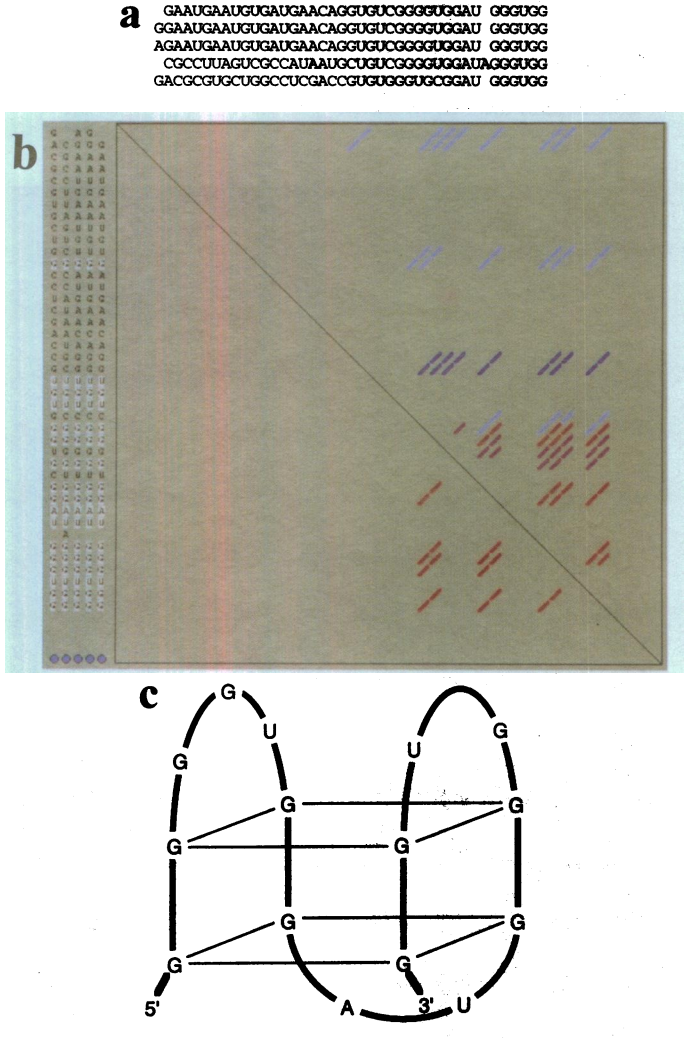

Figure 6. 2′-NH_2 RNA sequence set for elastase ligands aligned with the CLUSTER algorithm (**a**) and the resulting consensus structure matrix display (**b**) computed for this alignment. The fixed region nucleotides are in lower case and the evolved region is in upper case characters. The consensus structure matrix is computed for G-quartet identification only. (**c**) Schematic representation of the consensus secondary structure.

CONCLUSIONS

We have introduced an interactive computer program that generates a composite dot matrix representation of secondary structure elements from a set of functionally related oligonucleotides. The composite image facilitates visual detection of conserved secondary structure motifs within similar sequence sets. The complete pattern displaying all possible base pairings can be readily pruned with two progressive filters to provide the user with a simplified figure displaying only the most stable and conserved regions. Additional simplification of the structure matrix image can be achieved by eliminating mutually exclusive structures. Alignment editing provided within the program facilitates refinement of the consensus secondary structure.

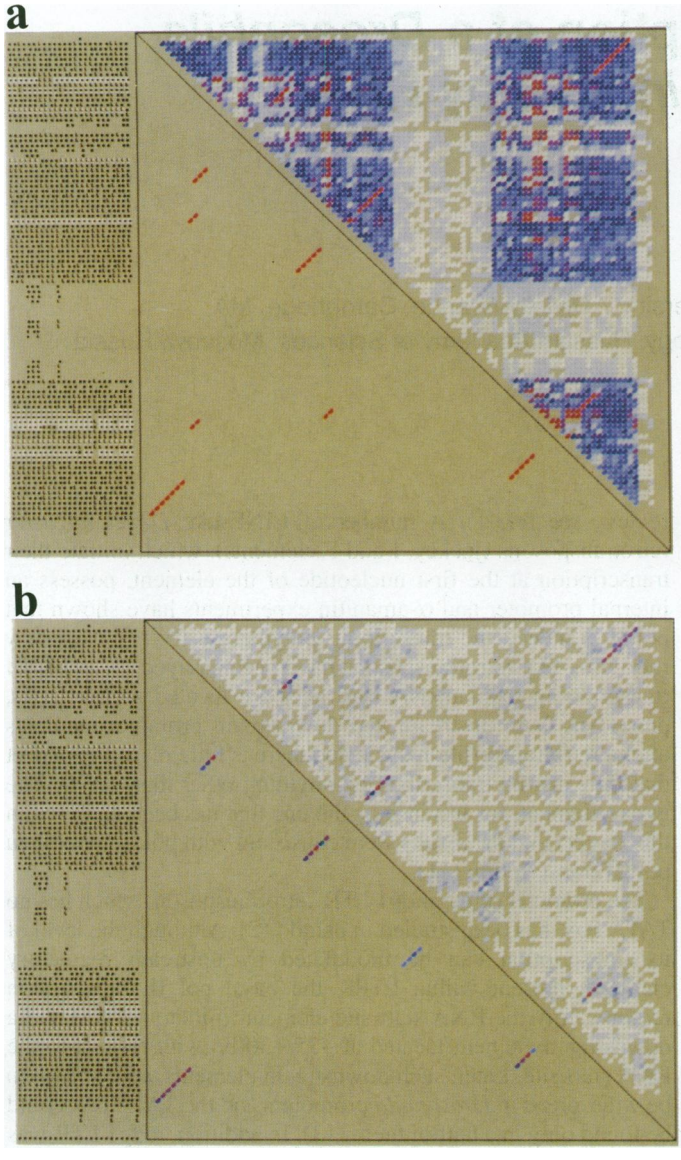

Figure 7. Consensus structure matrix (**a**) and base pair co-variation (**b**) displays computed for 40 aligned tRNA sequences.

The utility of the program was illustrated with four examples from SELEX experiments. In each case the initial multiple sequence alignment provided a strong enough consensus structure to allow optimization of the alignment and identification of a unique conserved secondary structural motif. Although the consensus secondary structures depicted in Figures 3–6 are generally in accord with those previously reported (8,9,19,20), additional conserved structure became apparent on examination of the matrix of one of the sequence sets (Fig. 3). In general, the use of the consensus matrix should expedite the process of identifying consensus secondary structure motifs. More importantly, this method allows for simultaneous examination of all possible structures, reducing the likelihood of overlooking a significant component of the consensus secondary structure. Finally, it is important to emphasize that although most examples presented here were derived from SELEX experiments, this method can clearly be applied to any functionally related oligonucleotide sequences, such as phylogenetic data sets. This was illustrated with a set of tRNA sequences, where the well-known clover leaf secondary structure for the aligned molecules was clearly elucidated.

ACKNOWLEDGEMENTS

We are grateful to Drs Gary Stormo (University of Colorado) and Craig Tuerk (Morehead State University) for many helpful discussions.

REFERENCES

1 Tuerk,C. and Gold,L. (1990) *Science*, **249**, 505–510.
2 Ellington,A. and Szostak,J. (1990) *Nature*, **346**, 818–822.
3 Gold,L., Tuerk,C., Allen,P., Binkley,J., Brown,D., Green,L., MacDougal,S., Schneider,D., Tasset,D. and Eddy,S. (1993) In Gespeland,A. and Atkins,J. (eds), *The RNA World*. Cold Spring Harbor Laboratory Press, Plainview, NY, Chapter 19, pp. 497–509.
4 Pan,T. and Uhlenbeck,O.C. (1992) *Nature*, **358**, 560–563.
5 Beaudry,A.A. and Joyce,G.F. (1993) *Science*, **257**, 635–641.
6 Bartel,D.P. and Szostak,J.W. (1994) *Science*, **261**, 1411–1418.
7 Illangasekare,M., Sanchez,G., Nickles,T. and Yarus,M. (1995) *Science*, **267**, 643–647.
8 Tuerk,C., MacDougal,S. and Gold,L. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 6988–6992.
9 Jellinek.,D., Lynott,C. K., Rifkin,D.B. and Janjic,N. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 11227–11231.
10 Jellinek,D., Green,L.S., Bell,C. and Janjić,N. (1994) *Biochemistry*, **33**, 10450–10456.
11 Kubik,M.F., Stephens,A.W., Schneider,D., Marlar,R.A. and Tasset,D. (1994) *Nucleic Acids Res.*, **22**, 2619–2626.
12 Feng,D.-F. and Doolittle,R.F. (1987) *J. Mol. Evol.*, **25**, 351–360.
13 Woese,C.R. and Pace,N.R. (1993) In Gespeland,A. and Atkins,J. (eds), *The RNA World*. Cold Spring Harbor Laboratory Press, Plainview, NY, Chapter 4, pp. 91–117.
14 Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) *Nucleic Acids Res.*, **20**, 5785–5795.
15 Jaeger,J.A., Turner,D.H. and Zuker,M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 7706–7710.
16 Jaeger,J.A., Turner,D.H. and Zuker,M. (1990) *Methods Enzymol.*, **183**, 281–306.
17 Bock,L., Griffin,L., Latham,J., Vermaas,E. and Toole,J. (1992) *Nature*, **355**, 564–566.
18 Quigley,G.J., Gehrke,L., Roth,D.A. and Auron,P.E. (1984) *Nucleic Acids Res.*, **12**, 347–366.
19 Jenison,R.D., Gill,S.C., Pardi,A. and Polisky,B. (1994) *Science*, **263**, 1425–1428.
20 Lin,Y., Qiu,Q., Gill,S.C. and Jayasena,S.D. (1994) *Nucleic Acid Res.*, **22**, 5229–5234.
21 Steinberg,S., Misch,A. and Sprinzl,M. (1993) *Nucleic Acid Res.*, **21**, 3011–3015.