A visual RNA Motif Editor for generating Thermodynamic Matchers

Janina Reeder

Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany

ABSTRACT

The search for RNA motifs, i.e. specific RNA secondary structures, has been a field of research for some years. Many different more or less specialised motif search tools have already been developed. In our opinion, many of these share one common disadvantage: the input of the motif must be provided in specific textual code which is not always straightforward. Recall what RNA motifs truly are, namely compact, well-defined two-dimensional objects of a specific orientation. So, our idea is to implement an editor that allows for visual input of these objects based on intuitive drag-and-drop techniques. Then, the motif is translated into a thermodynamic matcher. Such a matcher is a program for a specific secondary structure that uses thermodynamic scoring parameters when searching for a fitting candidate. Ultimately, our goal is to develop a GUI-based visual editor which allows for the automatic generation of thermodynamic matchers.

BACKGROUND

The world of RNA can roughly be divided into two classes: coding and noncoding RNAs. While coding RNAs (mRNA) are the template for translated proteins, noncoding RNAs constitute molecules with a function of their own. This function usually relies on their secondary structure, often in combination with short sequence motifs. Some well-known classical textbook examples of such functional RNAs are transfer and ribosomal RNAs. Yet, with the currently emerging hot field of microRNAs and other small RNAs, the types and number of such RNA genes is expanding rapidly (Lee and Ambros, 2001). A comprehensive collection of noncoding RNA families is presented e.g. in the Rfam database (Griffiths-Jones *et al.*, 2003).

Since many of the known or assumed functions of noncoding RNAs lie in the different stages of regulation, they play an important, for a long time underestimated role in the overall concept of molecular biology. Thus, it is not surprising that there is a strong interest in searching for secondary structure motifs of potential medical or scientific importance. A well-known example for an RNA motif of regulatory function is the iron responsive element (IRE). It can be located in both 5' and 3' UTRs and serves as the binding site for the IRE - binding protein. As the interaction of the protein with the binding site despends strongly on both RNA sequence and structure (Jaffrey *et al.*, 1993), the IRE serves as an ideal and well-studied candidate for finding sites of regulatory importance based on a consensus RNA motif (see Figure 1).

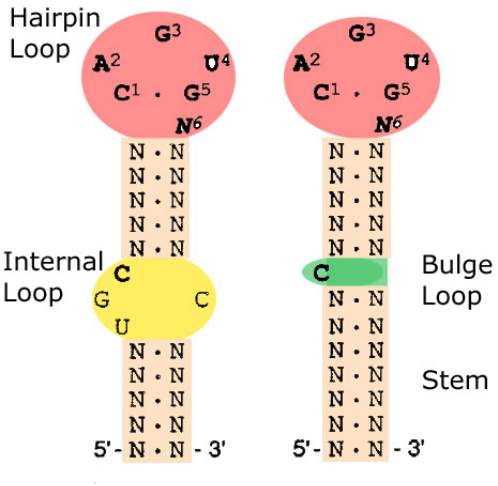

Fig. 1. The IRE exists in two different consensus forms: one with a single C bulge and one with an internal loop, also referred to as the "ferritin" bulge.

There are already quite a few search tools, some specific for certain types of noncoding RNAs (tRNAscan-SE: Lowe and Eddy (1997)), others more general (RNAMotif, HyPa, PatScan). At large, these tools require the definition of a secondary structure motif and a nucleotide sequence as input. The specified motif is then searched by folding the nucleotide sequence according to the known basepairing rules and evaluating it against the motif. Here, we will focus only on the general motif search tools, as this is what we are aiming at as well. While all of those produce more or less satisfying results, in our opinion all of them can be improved from at least one perspective.

The program RNAMotif (Macke *et al.*, 2001) depends on user interaction in two essential steps: the motif itself has to be specified in an awk-like language and the scoring function

has to be provided by the user (also a default scoring can be chosen). While the motif language is not highly complex, the user has to study it carefully in order to be able to use the program. He has to transform what is intuitively clear to him/her into an abstract code. The same holds for the scoring function which is not used directly when producing the hits of the search phase, but rather as an additional filter afterwards.

A much more powerful tool is the program HyPa (Gräf *et al.*, 2001). It searches for hybrid patterns, not just RNA structure motifs. Here, the user can search for sequence and structure similarity and he/she can specify arbitrary characteristics, such as thermodynamic constraints. But, in order to do so, it requires even more complicated input of the user in a special declarative pattern description language (Strothmann *et al.*, 2000). The search strategy is based on known (PatScan, RNAMotif) and new pattern matching algorithms.

PatScan (Dsouza *et al.*, 1997) is a pattern matcher for protein or nucleotide patterns. Here, input also must be made in textual code according to a set of rules. Any scoring function must be incorporated within the pattern of interest.

Other tools are available that do not ask the user to specify a motif himself, but rather use the results of multiple sequence alignments or database information for a search procedure (Infernal: Eddy (2002)). Of course, this does not require any complicated input from the user, but it restricts the application of the program to only those motifs which are already known.

Two main fields of improvements

From the methodological point of view, these programs do not take the true biochemical background into account. Instead, they produce hits based on the input of the user in both motif description and scoring parameters. Even though our current knowledge of the underlying biochemical or even biophysical processes of RNA folding and interactions is still not perfect, we cannot do any better than basing our search evaluation on currently known thermodynamic properties. While the user can specify thermodynamic parameters as the scoring scheme of some search programs, to our knowledge, no motif search program based solely and automatically on these parameters is available as of today. Therefore, we propose to develop thermodynamic matchers which are search programs specific for a certain structure motif that incorporate the folding of the query sequence according to thermodynamic laws as the essential step of the search phase. These matchers are based upon the ADP (Algebraic Dynamic Programming) technique which allows for the separation of the definition of the search space (the RNA motifs) and the scoring algebra (the thermodynamic properties). This enables us to implement the scoring algebra once for all possible matchers without the user having to spend any time or thought on it. The grammar of the matcher restricts the folding of the query sequence to the specified motif, i.e. we do not allow the query sequence to fold into the best possible shape and filter it, but rather force it into a certain structure and rate this one. We obtain an executable search program simply by defining the motif in an ADP grammar and thereby reducing the search space. Yet, while this sounds straightforward and simple to do, to actually write the required code is not trivial and demands expert knowledge. It cannot be asked of a user, e.g. a biologist, searching for a certain motif in a number of query sequences. The interested reader can find information on the use of ADP for pattern matching algorithms on mixed sequence and secondary structure motifs in RNA in (Meyer and Giegerich, 2002).

Here, the second hindrance of current search programs comes into play: the user has to provide information on the motif in textual code. Our approach is to develop a graphical user interface through which he/she can design an RNA motif without the need to learn any textual code or know anything about ADP. This motif will then be translated automatically into the appropriate code, thereby producing an executable thermodynamic matcher specific for this motif.

THE VISUAL EDITOR

Visual Building Blocks

The concept of the visual editor is based on the fact that RNA secondary structures are composed of a very limited set of “graphical” building blocks. In general, only 6 different types of building blocks are sufficient to design any potential structure: stems, internal loops, bulge loops, hairpin loops, multiloops and single stranded regions. This does of course exclude any structures with tertiary interactions such as pseudoknots or “kissing hairpins”. Nevertheless, the inclusion of some simple classes of pseudoknots is thinkable. Using these building blocks, imagine them to be like puzzle pieces or domino stones, we can define an RNA motif by placing them next to each other. And this is what the editor we are presenting here basically does.

Implementation

The editor is implemented in Java, relying on Java 2D Graphics for the visual interface through which the user can construct his/her motifs. A general view of the editor in its current state is shown in Figure 2. Since there is such a clean distinction of an RNA secondary structure into its building blocks, we can completely abstract from the underlying biological concepts. That means, the Java program has a biological data level that keeps track of any relevant information and restrictions, and a corresponding independent graphical level which is responsible for what the user sees.

The user can select which type of building block to include into the current structure and move it around with the mouse cursor. If the block is moved into close proximity of an open end of an already existing structure part, it will snap into the correct connecting position. Then, it can be set down by the user simply by pressing the mouse button. Also, the addition

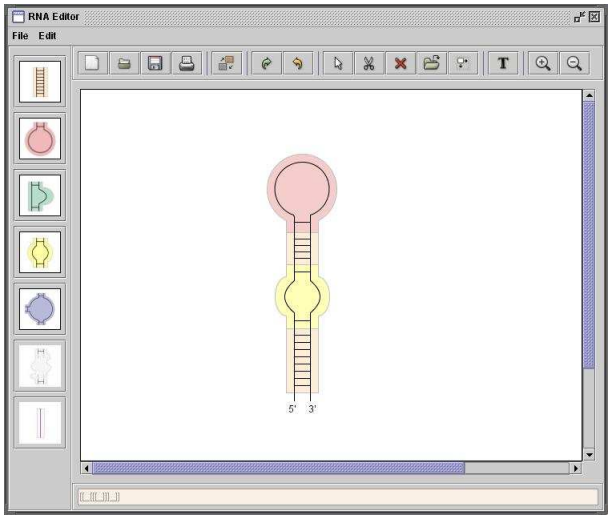

Fig. 2. On the left side of the editor are the buttons for the different building blocks, on top are those for user interaction. Here, an IRE-element type motif was constructed.

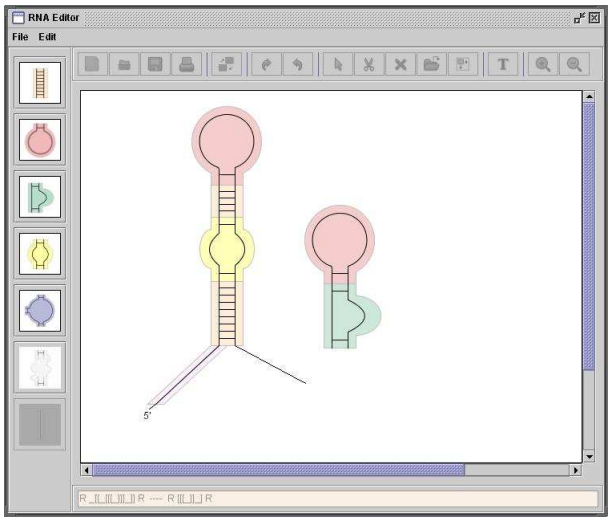

Fig. 3. When a single strand is included, it is first shown as a solid block. Once it is connected to a structure, it is drawn as a line connecting the structure with the mouse cursor. By pressing the mouse button, the user then can choose the outline of the single strand shape.

of new, remote structure parts is possible. Different closed parts of a structure, i.e. those with only one open end remaining, can later be connected or extended by single strands. The shape of a single strand is not fixed, but rather freely adjustable (see Figure 3). All other building blocks though have a standard shape, that is only adjusted according to size specifications by the user.

Many kinds of user interaction, such as rotating elements of the structure, zooming in and out, deleting or removing an element of the structure, provide ample means to create the secondary structure motif the user has in mind. For each element of the structure, the user can open an edit interface through which he/she can specify internal information such as the size of the element (the length of the contained RNA sequence) or a sequence motif.

In general, the editor makes sure that only feasible structures are designed by, e.g. not allowing for closed structures or keeping track of size restrictions. The editor gives an online translation of the current motif into the abstract shape notation (Giegerich *et al.*, 2004) which is based upon the well-known Vienna Strings. Here, for the represented RNA sequence any unpaired base is indicated by a dot and a basepair by an opening and a corresponding closing bracket; for the IRE: "(((((...(((((.....))))).)))))". With shapes, the abstractions is even greater as we do not show every individual base, but rather only stretches of unpaired bases as "_" and a stretch of basepairs as "[" "]"; for the IRE: "[_[_]_]". The advantage of the shape notation is its highly compact, yet precise representation of an RNA secondary structure. It is the link between the abstract graphical level of the program and the corresponding biological information level. It shows the one-to-one relationship between the visual building blocks and the composition of a secondary structure. While the abstract representation is of no direct use, it offers a notation widely accepted in the world of RNA computing. Also, it currently serves as a means to secure that the internal data resembles the visual structure specified by the user.

In the end, the same mechanism used for the translation of the building blocks into the shapes can be applied for our ultimate goal: the automatic translation into a thermodynamic matcher. While the matchers themselves are not part of the work presented here, keep in mind that this one-to-one translation of structures into matchers is only possible because the grammar of the matchers is made up of the same building blocks we are using in our editor. All is based on the fact that RNA secondary structures are composed of a limited set of distinct elements.

FUTURE WORK: WEBSERVER INTEGRATION AND IMPROVEMENTS

In order to provide the editor through our webserver, we decided to use XML as a secure communication language in between the editor and the executable ADP grammar of the thermodynamic matchers. Therefore, for each building block of the RNA motif we have a corresponding XML element storing all the necessary information. The translation of the building blocks from our editor into XML code is current work.

Once the translation into XML is completed, the editor will be integrated into the webserver generating XML output. We will provide an XMLtoADP script that translates the output into the language of thermodynamic matchers. This script is

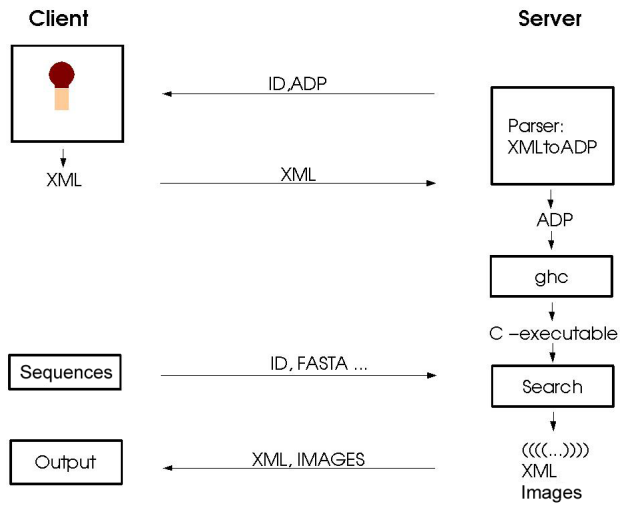

Fig. 4. The client (RNA editor) produces XML output which is sent to the webserver. Here, a parser generates ADP code from the XML file and returns the ADP grammar together with an id to the client. Also, an executable c-program is compiled from the ADP code for efficiency reasons. Then, the user (client) sends the nucleotide sequence(s) and the id to the server and the actual search phase is invoked. Finally, the result is presented and/or send back to the client in form of Vienna strings, XML and/or images.

not straightforward, especially since we are in parallel also still developing and enhancing the concept of thermodynamic matchers in general. Figure 4 gives an overview of the communication protocols of the program and the server.

Apart from these essential future steps to achieve the goal of this thesis, a lot of work still remains in implementing useful features of the editor. We will have to keep the editor up to date with further enhancements of the underlying thermodynamic matchers, but also a large amount of possible improvements from the user-friendliness side can still be made. Not only general design aspects can always be improved, but also new functionality is still desirable. Some examples for future improvements are the ability to specify groups of building blocks (e.g. either a bulge or an internal loop at a certain position), to allow for alternative input formats or to provide several editing interfaces.

REFERENCES

Dsouza, M., Larsen, N. and Overbeek, R. (1997) Searching for patterns in genomic data. *Trends in Genetics*, **13**, 497–498.

Eddy, S. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**.

Giegerich, R., Voss, B. and Rehmsmeier, M. (2004) Abstract Shapes of RNA. *Nucleic Acids Res.*, **32**, 4843–4851.

Gräf, S., Steger, G., Strothmann, D. and Kurtz, S. (2001) HyPaLib: a Database of RNAs and RNA Structural Elements defined by Hybrid Patterns. *Nucleic Acids Res.*, **29**, 196–198.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S. R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.

Jaffrey, S., Haile, D., Klausner, R. and Harford, J. (1993) The interaction between the iron-responsive element binding protein and its cognate RNA is highly dependent upon both RNA sequence and structure. *Nucleic Acids Res.*, **21**, 4627–4631.

Lee, R. and Ambros, V. (2001) An extensive class of small RNAs in caenorhabditis elegans. *Science*, **294**, 862–864.

Lowe, T. and Eddy, S. (1997) tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.*, **25**, 955–964.

Macke, T. J., Ecker, D. J., Gutell, R. R., Gautheret, D., Case, D. A. and Sampath, R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.

Meyer, C. and Giegerich, R. (2002) Matching and significance evaluation of sequence-structure patterns in RNA. *J. Physical Chemistry*, **216**, 1–24.

Strothmann, D., Kurtz, S., Gräf, S. and Steger, G. (2000) The Syntax and Semantics of a Language for Describing Complex Patterns in Biological Sequences. Report 2000–06, Technische Fakultät, Universität Bielefeld.