196–198 Nucleic Acids Research, 2001, Vol. 29, No. 1

HyPaLib: a database of RNAs and RNA structural elements defined by hybrid patterns

Stefan Gräf, Dirk Strothmann[1], Stefan Kurtz[1] and Gerhard Steger*

Institut für Physikalische Biologie, Geb 26.12.U1, Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, D-40225 Düsseldorf, Germany and [1]Technische Fakultät, Universität Bielefeld, Germany

Received September 1, 2000; Revised and Accepted November 6, 2000

ABSTRACT

The database, called HyPaLib (for Hybrid Pattern Library), contains annotated structural elements characteristic for certain classes of structural and/or functional RNAs. These elements are described in a language specifically designed for this purpose. The language allows convenient specification of hybrid patterns, i.e. motifs consisting of sequence features and structural elements together with sequence similarity and thermodynamic constraints. We are currently developing software tools that allow a user to search sequence databases for any pattern in HyPaLib, thus providing functionality which is similar to PROSITE, but dedicated to the more complex patterns in RNA sequences. HyPaLib is available at http://bibiserv.techfak.uni-bielefeld.de/HyPa/.

INTRODUCTION

Determining or finding a protein gene in genomes of eubacteria or archaea is thought to be easy; due to the presence of introns and other complexities it is more difficult to do so in eukaryotes. On the contrary, finding an RNA gene (not a mRNA gene) in genomic DNA is a challenging and difficult task in most cases. To support this task, it would be very helpful to have a general tool that allows the user to define his/her own RNA-related patterns or to use an existing library of RNA-related patterns and to search for such patterns in biological sequences. In this paper, we describe our 'Hybrid Pattern Library' (HyPaLib, for short), which contains annotated structural elements characteristic for certain classes of structural and/or functional RNAs. These elements are described in a language HyPaL (for Hybrid Pattern Language) specifically designed for this purpose. The language allows convenient specification of hybrid patterns, i.e. motifs consisting of sequence features and structural elements together with sequence similarity and thermodynamic constraints. We are currently developing software tools that allow a user to search sequence databases for any hybrid pattern in HyPaLib, thus providing functionality which is similar to PROSITE (1), but dedicated to the more complex patterns in RNA sequences. HyPaLib is available at http://bibiserv.techfak.uni-bielefeld.de/HyPa/.

FEATURES OF THE LANGUAGE HyPaL

HyPaL is a declarative language; that is, the user specifies what to search without specifying how to search it. Essential properties of the language are:

- HyPaL is modular: patterns can be reused as part of descriptions of more complex patterns.
- HyPaL contains approximative elements: specification of a pattern with a (numerical) tolerance interval allows for searches of related (homologous, paralogous) sequences with insertions, deletions and mutations.
- HyPaL allows for: (i) user-defined scoring functions, and (ii) user-defined constraints (based on a set of standard functions).

For example, the free energy of a specified structure is an allowed search criterium (c.f. Fig. 1). The availability of such scoring functions and constraints is the major difference between HyPaL (and its associated search tool) and other languages/search tools known from the literature.

HyPaL provides many different linguistic elements. For a complete definition of the language, see http://bibiserv.techfak.uni-bielefeld.de/HyPa/. In the following, we list only the most important elements, and give corresponding examples.

- Simple sequences: e.g. a nucleic acid ATUGCYT or a protein sequence LWYMN.
- Wildcards: the symbol . matches any character.
- Character classes: the expression [ACG] matches one of the characters A, C or G.
- Concatenation: AC[GT]C.G AAGGT means that AC[GT]C.G is followed by AAGGT.
- Disjunction: ACCG | A[CG]AGGT matches ACCG or A[CG]AGGT.
- Repetitive elements: p{3,5} matches p three to five times.
- Spacers: p1 <3,5> p2 specifies a gap of length 3–5 between p1 and p2.
- Spacers with overlaps: p1 <–3,14> p2 specifies that p1 and p2 may overlap up to three positions or there is a gap of length at most 14.
- Pprofiles or consensus matrices: [[10,5,4,3,6],[6,11,5,4,2]]>15 specifies scores for the characters A, [UT], G, C and gap, at the first and second position, and a total score of 15 to be exceeded. Thus this profile matches the sequences AA, AT, TT and –T.

*To whom correspondence should be addressed. Tel: +49 211 81 14927; Fax: +49 211 81 15167; Email: steger@biophys.uni-duesseldorf.de

```
ID secis
DT 28/08/2000
DE Selenocysteine Insertion Sequence (SECIS); common stem-loop structure in
DE the 3' untranslated region of selenoprotein mRNAs
KW SECIS
HP secis =
HP @stem1 @iloop1 @quartet @stem2 @aloop @cstem2 @cquartet @iloop2 @cstem1
HP << @energy(@stem1 (@iloop1 @iloop2) @cstem1) < -30 &  # both energy
HP    @energy(@stem2 (@aloop) @cstem2) < -21             # constraints must hold
HP
HP where @stem1:=.{7,}      # 7+ bps
HP       @iloop1:=.{2,6} A  # internal loop of 6 to 14 nts; 5' part ends with A
HP       @quartet:=UGA.     # 4 non-WC bps
HP       @stem2:=.{11,12}   # 11 or 12 bps
HP       @aloop:=AA .{5,15} # hairpin loop of 7 to 17 nts; starts with AA
HP       @cstem2:=@^(@stem2)[3,1,1] # rev. compl. of stem2, at most 3 mismatches
HP                                  # and 1 bulge on either side (i.e. 1 ins./del.)
HP       @iloop2:=.{3,7}
HP       @cquartet:=.GA.
HP       @cstem1:=@^(@stem1)[1,1,1] # rev. compl. of stem1, at most 1 mismatch
HP                                  # and 1 bulge on either side
RE Kryukov, G.V. et. al. (1999) J. Biol. Chem. (274), 33888-33897.
//
```

Figure 1. Example of a HyPaLib entry. Pattern for the Selenocysteine Insertion Sequence (SECIS). In addition to the necessary primary and secondary structure constraints, the pattern contains thermodynamic constraints to exclude a large number of false negative matches in databank searches (15).

Table 1. Description of HyPaLib items

Kind	Description	Feature[a]	Remark
ID	Identifier	mu	Unique name of a hybrid pattern
DT	Date	mu	List of creation/modification dates in format DD/MM/YYYY
DE	Description	m	Informal description or comment
KW	Keyword		List of keywords useful for database search
HP	Hybrid Pattern	u	Formal definition in the syntax of HyPaL
NO	EMBL-Number		Release of EMBL according to which hybrid pattern was verified
AL	Alignment		Aligned sequences and properties describing their origins
CS	Consensus Sequence	u	Consensus sequence of aligned sequences ('-' is used to align the consensus sequence with the consensus structure)
ST	Consensus Structure	u	Valid secondary structure in Vienna format (bracket-dot notation)
PM	Parameters		Special parameters
RE	References		List of references

[a]m means a mandatory item and u means a unique item.

- References to sequence parts by variables: @v:=(AT.....GC) binds the variable v to the sequence matching the pattern AT.....GC.
- Reverse complement of sequences: @^(@v) matches the sequence that is the reverse complement of the sequence bound to the variable v.
- Thermodynamic constraints depending on variables: @energy(@v) < –15 means that the thermodynamic free energy of the structure bound to the variable v has to be lower than –15 kJ/mol.
- Logical operators may be used to combine constraints.

FORMAT OF HyPaLib

HyPaLib is available as a plain text file formatted according to the general rules of the EMBL and related databases. We also provide a version in HTML format with hyperlinks to sequence and citation databases. Each entry in HyPaLib describes an annotated RNA or DNA pattern (for an example see Fig. 1) on different lines called items. An item consists of two uppercase letters specifying the kind of item, followed by the information stored for that item (see Table 1 for a description of the items).

CURRENT STATUS OF HyPaLib

HyPaLib contains sequential and structural elements characteristic for different classes of RNA. In the following, we enumerate the classes and give a few examples of corresponding HyPaLib entries.

- Simple patterns of DNA or RNA sequences with specific biological function like Pribnow box, –35 region, promoter region or definition of codons and stop codons. Most examples are adopted from Mehldau *et al*. (2). Because of their

low specificity these patterns should mainly be used as building blocks for construction of more complex patterns [for reviews on further important components of the 'kit' of RNA structural elements see Batey *et al.* (3), Conn and Draper (4) and Moore (5)].

- Simple patterns of RNA secondary structure like hairpin, pseudoknot, clover leaf, attenuator, TMV 3′ end. Examples are taken from Searls *et al.* (6); these patterns are not very specific.
- Patterns for RNA protein binding motifs like the secondary structure of the four Rev-binding elements (RBE) (7) or the putative Tat-binding elements (TBEs) in viruses linked to human immunodeficiency infections (8). Patterns are taken from Bourdeau *et al.* (9).
- Patterns describing chemically and catalytically active motifs like hammerhead ribozyme, UV-sensitive loop E, leadzyme and DNAzyme. Patterns are adopted from Bourdeau *et al.* (9).
- Patterns describing small molecule-binding RNA motifs like aptamers for valine and neomycin. Patterns are adopted from Bourdeau *et al.* (9).
- Profiles of small nuclear RNAs like 5′ end of U2, 5′ end of U3 or the central part of 7S RNA. These patterns are derived from alignments and structures in the uRNA Database (10) and the Signal Recognition Particle Database (11), respectively. After inclusion of additional sequences and refinement using ClustalX (12) and ConStruct (13) these patterns give no false positive results with searches in the EMBL database except for non-annotated entries.
- Patterns describing motifs with thermodynamic constraints like hairpin C (14), a structural feature from mRNA of prion proteins, or the selenocysteine insertion sequence element (Fig. 1). Both patterns contain thermodynamic constraints to exclude a large number of false negative matches.

Specification of most patterns is a time consuming task, because consensus sequences or structures from text books and older references are unspecific in most cases due to the enormous growth of recent EMBL or GenBank releases. To cover a more complete set of relevant RNA patterns and to make the library more useful, we would like to encourage the reader to suggest other hybrid patterns (in any format). Please contact the corresponding author.

ACKNOWLEDGEMENTS

We thank R. Giegerich for stimulating discussions and for his support. This work was supported by grants from the Deutsche Forschungsgemeinschaft (STE465/4 and KU1257/1) and Fonds der Chemischen Industrie.

REFERENCES

1. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
2. Mehldau,G. and Myers,E.W. (1993) A System for Pattern Matching Applications on Biosequences. *Comp. Appl. Biosci.*, **9**, 299–314.
3. Batey,R.T., Rambo,R.P. and Doudna,J.A. (1999) Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed. Engl.*, **38**, 2326–2343.
4. Conn,G.L. and Draper,D.E. (1998) RNA structure. *Curr. Opin. Struct. Biol.*, **8**, 278–285.
5. Moore,P.B. (1999) Structural motifs in RNA. *Annu. Rev. Biochem.*, **68**, 287–300.
6. Searls,D.B. (1993) The computational linguistics of biological sequences. In Hunter,L. (ed.), *Artificial Intelligence and Molecular Biology*. AAAI Press, Menlo Park, CA.
7. Leclerc,F., Cedergren,R. and Ellington,A.D. (1994) A three-dimensional model of the Rev-binding element of HIV-1 derived from analyses of aptamers. *Nat. Struct. Biol.*, **1**, 293–300.
8. Ferbeyre,G., Bourdeau,V. and Cedergren,R. (1997) Does HIV tat protein also regulate genes of other viruses present in HIV infection? *Trends Biochem. Sci.*, **22**, 115–116.
9. Bourdeau,V., Ferbeyre,G., Pageau,M., Paquin,B. and Cedergren,R. (1999) The distribution of RNA motifs in natural sequences. *Nucleic Acids Res.*, **27**, 4457–4467.
10. Zwieb,C. (1997) The uRNA database. *Nucleic Acids Res.*, **25**, 102–103.
11. Zwieb,C. and Samuelsson,T. (2000) SRPDB (Signal Recognition Particle Database). *Nucleic Acids Res.*, **28**, 171–172. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 169–170.
12. Jeanmougin,F., Thompson,J.D., Gouy,M., Higgins,D.G. and Gibson,T.J. (1998) Multiple sequence alignment with Clustal X. *Trends Biol. Sci.*, **23**, 403–405.
13. Lück,R., Gräf,S. and Steger,G. (1999) *ConStruct*: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res.*, **27**, 4208–4217.
14. Lück,R., Steger,G. and Riesner,D. (1996) Thermodynamic prediction of conserved secondary structure: application to the RRE element of HIV, the tRNA-like element of CMV and the mRNA of prion protein. *J. Mol. Biol.*, **258**, 813–826.
15. Kryukov,G.V., Kryukov,V.M. and Gladyshev,V.N. (1999) New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J. Biol. Chem.*, **274**, 33888–33897.