BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 15 2013, pages 1900–1901
doi:10.1093/bioinformatics/btt299

Sequence analysis

Advance Access publication May 24, 2013

Triplex: an R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences

Jiří Hon[1], Tomáš Martínek[1], Kamil Rajdl[2] and Matej Lexa[2,*]

[1]Department of Computer Systems, Faculty of Information Technology, Brno Technical University, Božetěchova 2, 61266 Brno and [2]Department of Information Technology, Faculty of Informatics, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Upgrade and integration of triplex software into the R/Bioconductor framework.

Results: We combined a previously published implementation of a triplex DNA search algorithm with visualization to create a versatile R/Bioconductor package 'triplex'. The new package provides functions that can be used to search Bioconductor genomes and other DNA sequence data for occurrence of nucleotide patterns capable of forming intramolecular triplexes (H-DNA). Functions producing 2D and 3D diagrams of the identified triplexes allow instant visualization of the search results. Leveraging the power of Biostrings and GRanges classes, the results get fully integrated into the existing Bioconductor framework, allowing their passage to other Genome visualization and annotation packages, such as GenomeGraphs, rtracklayer or Gviz.

Availability: R package 'triplex' is available from Bioconductor (bioconductor.org).

Contact: lexa@fi.muni.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 3, 2013; revised on May 21, 2013; accepted on May 22, 2013

1 INTRODUCTION

DNA sequence analysis and annotation are important steps in uncovering the molecular basis of life. Although protein-coding sequences have been intensively studied in the past, recent focus has shifted toward the less-known biological functions encoded in intergenic DNA, as well as the study of structural and regulatory aspects of genetic information packaging in chromosomes. Tools for the necessary sequence analysis of non-coding sequences are less common than their gene-centered counterparts. We have recently formulated and implemented an algorithm to detect potential triplex-forming sequences in genomes (Lexa *et al.*, 2011). Such sequences have been implicated as important players in several key processes, such as transcriptional regulation (Walter *et al.*, 2001) or DNA recombination (Rooney and Moore, 1995).

Triplex DNA forms when a third strand of nucleotides is allowed to align with a Watson–Crick duplex using Hoogsteen bonds to stabilize the nascent structure (Soyfer and Potaman, 1995). H-DNA is a form of DNA where triplexes form intramolecularly, without the participation of other DNA molecules (Htun and Dahlberg, 1989).

Currently, several research groups reported on their efforts to map triplex-forming sites in known genomes, as well as on the development of tools to carry out such searches. Hoyne *et al.* (2000) used pattern recognition tools to search for homopurine/homopyrimidine stretches in DNA as likely triplex formation sites. Cer *et al.* (2012) created a non-B DNA search tool (nBMST) that includes mirror repeat detection functionality to identify potential triplexes. Buske *et al.* (2012) and Lexa *et al.* (2011) created triplex detection procedures allowing for a small percentage of imperfections in the sequences, leading to higher sensitivity of searches. Often, the tools exist as stand-alone software or web tools, which led us to the idea to integrate triplex search, visualization and genome annotation into a unified Bioconductor software package in R for increased flexibility.

Here, we describe *triplex*, demonstrating its use in sequence analysis of sample data, focusing on functions integrating it with the rest of the R/Bioconductor suite. Of the aforementioned softwares, only *triplex* provides specialized H-DNA searching. The other software treats H-DNA as general mirror repeats and lacks fine-grained or configurable mismatch evaluation (nBMST), focuses on a different class of triplexes (Hoyne *et al.*, 2000) or provides general results that need to be further filtered to identify H-DNA (triplexator), requiring several orders of processing time more than *triplex*. The software by Lexa *et al.* (2011) used to create the package was improved by (i) integration into R/Bioconductor, (ii) elimination of recognized bugs in scoring and alignment and by (iii) providing base pair information, either as text/variables or visualizations.

We performed a simple comparison of nBMST and triplexator programs with *triplex* (see Supplementary Material). It showed that reported (CT)n and (TA)n mirror repeats coincide with H-DNA found by *triplex*. Triplexator returned several longer patterns reported by *triplex* in fragments, a problem that may depend on precise settings, although we found computation time and memory use increased significantly at such attempts. This is likely caused by triplexator design to find any combinations of triplex-forming sequences, not only local patterns leading to H-DNA.

*To whom correspondence should be addressed.

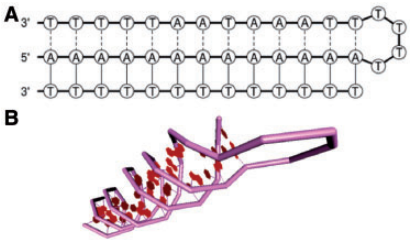

Fig. 1. (**A**) 2D diagram and (**B**) 3D model of one of the best scored triplex

2 THE SOFTWARE

The R triplex package is essentially an R interface to the underlying C implementation of a dynamic-programming search strategy of the same name (Lexa *et al.*, 2011). The main functionality of the original program was to detect the positions of subsequences in a much larger sequence capable of folding into an intramolecular triplex (H-DNA) made of as many canonical nucleotide triplets as possible. We extended this basic functionality to include the calculation of exact base pairing in the triple helices. This allowed us to include visualization, showing the exact base pairing in 1D, 2D or 3D (see Section 3). The created package takes advantage of the existing Bioconductor infrastructure. For example, the triplex search method uses the *DNAString* object as input. As a result, all available genomes (*BSgenomes* objects) can be easily analyzed. As for the output, identified triplexes are stored in data objects of a class based on *XStringViews*. Thus, all other libraries or methods working with *IRanges* can be applied to triplexes as well. Alternatively, the results can be transformed into *GRanges* objects that enable further possibilities, such as visualization of genome tracks using *GenomeGraphs* or export of results to the GFF3 annotation format.

3 USAGE EXAMPLE

In the following example, we load a genomic sequence from one of the *BSGenome* packages, identify potential triplexes with length over eight triplets of nucleotides and score ≥17, create two different visualizations of the best-scored triplex. Finally, we export the identified positions into a genome annotation track (via a GFF3 file) and store the sequences in a FASTA file.

I) Load necessary libraries and genomes.

```
> library(triplex)
> library(BSgenome.Celegans.UCSC.ce10)
```

II) Search for potential triplex positions and display the results.

```
> t <- triplex.search(Celegans[["chrX"]],
+                     min_score=17,min_len=8)
> t
  Triplex views on a 17718866-letter DNAString subject
subject: CTAAGCCTAAGCCTAAGCCTAAGCCTAA...TAGGCTTAGGCTTAGGCTTAGGCTTAGG
triplexes:
           start width score  pvalue ins type s
    [1]      762    28    17 6.5e-04   0    4 - [TCTAAAAGACACACAATTTAGAAAAAAA]
    [2]     1160    26    17 3.7e-04   0    7 + [ACAAAAACTTCATCAACAAGAAAAAA]
    ...
[20033] 17715172    29    17 3.7e-04   0    6 + [AAAAAAAAGTGAAAAAAACTGAATTTCAT]
[20034] 17718247    27    17 3.7e-04   0    6 + [AAAAAAAAACACTTAAACATAAAACTA]
```

III) Sort the results by score and display the best-scoring non-trivial triplex. Graphical output is shown in Figure 1.

```
> ts <- t[order(score(t),decreasing=TRUE)]
> triplex.diagram(ts[1])
> triplex.3D(ts[1])
```

IV) Export the results as GFF3 and FASTA files.

```
> library(rtracklayer)
> export(as(t, "GRanges")," test.gff", version="3")
> writeXStringSet(as(t, "DNAStringSet"), file="test.fa",
+                 format="fasta")
```

4 CONCLUSION

We present a new R/Bioconductor package that integrates our previously defined algorithm for identification of triplex-forming sequences with two new methods of their visualization (2D diagram and 3D model). The created package uses existing Bioconductor infrastructure in such way that available genomes (*BSGenomes*) can easily be used as input. The identified triplexes can be further analyzed as *IRanges* or *GRanges* objects (and optionally exported into GFF3 or FASTA file). In connection with R language and existing libraries for statistical analysis, the package represents powerful tool for molecular biologists interested in analysis of non-canonical DNA structures such as triplexes.

Funding: Framework of IT4Innovations project (CZ.1.05/1.1.00/02.0070) funded by the EU Operational Programme 'Research and Development for Innovations', MSMT Grants (No.0021630528) 'Security-Oriented Research in Information Technology', and (LA09016) 'Participation of CR in ERCIM', and BUT grant (FIT-S-11-1) 'Advanced secured, reliable and adaptive IT'.

Conflict of Interest: none declared.

REFERENCES

Buske,F.A. *et al.* (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic. *Genome Res.*, **22**, 1372–1381.

Cer,R. *et al.* (2012) *Searching for Non-B DNA-Forming Motifs Using nBMST (Non-B DNA Motif Search Tool)*. Curr. Protoc. Hum. Genet., Chapter 18. Unit 18.7.1–22.

Hoyne,P.R. *et al.* (2000) Searching genomes for sequences with the potential to form intrastrand triple helices. *J. Mol. Biol.*, **302**, 797–809.

Htun,H. and Dahlberg,J.E. (1989) Topology and formation of triple-stranded H-DNA. *Science*, **243**, 1571–1576.

Lexa,M. *et al.* (2011) A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics*, **27**, 2510–2517.

Rooney,S.M. and Moore,P.D. (1995) Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells. *Proc. Natl Acad. Sci. USA*, **92**, 2141–2144.

Soyfer,V.N. and Potaman,V.N. (1995) *Triple-Helical Nucleic Acids*. Springer-Verlag, New York.

Walter,A. *et al.* (2001) Evidence for a DNA triplex in a recombination-like motif: I. recognition of Watson-Crick base pairs by natural bases in a high-stability triplex. *J. Mol. Recognit.*, **14**, 122–139.