BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–8

# A dynamic programming algorithm for identification of triplex-forming sequences

Matej Lexa [1]*, Tomáš Martínek [2], Ivana Burgetová [2], Daniel Kopeček [1] and Marie Brázdová [3]

[1]Department of Information Technology, Faculty of Informatics, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic
[2]Department of Computer Systems, Faculty of Information Technology, Brno Technical University, Božetěchova 2, 61266 Brno, Czech Republic
[3]Department of Biophysical Chemistry and Molecular Oncology , Institute of Biophysics, Academy of Sciences of the Czech Republic v.v.i., Královopolská 135, CZ-612 65 Brno, Czech Republic

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

**ABSTRACT**

**Motivation:** Current methods for identification of potential triplex-forming sequences in genomes and similar sequence sets rely primarily on detecting homopurine and homopyrimidine tracts. Procedures capable of detecting sequences supporting imperfect, but structurally feasible intramolecular triplex structures are needed for better sequence analysis.

**Results:** We modified an algorithm for detection of approximate palindromes, so as to account for the special nature of triplex DNA structures. From available literature we conclude that approximate triplexes tolerate two classes of errors. One, analogical to mismatches in duplex DNA, involves nucleotides in triplets that do not readily form Hoogsteen bonds. The other class involves geometrically incompatible neighboring triplets hindering proper alignment of strands for optimal hydrogen bonding and stacking. We tested the statistical properties of the algorithm, as well as its correctness when confronted with known triplex sequences. The proposed algorithm satisfactorily detects sequences with intramolecular triplex-forming potential. Its complexity is directly comparable to palindrome searching.

**Availability:** Our implementation of the algorithm is available at http://www.fi.muni.cz/~lexa/triplex as source code and a web-based search tool. The source code compiles into a library providing searching capability to other programs, as well as into a stand-alone command-line application based on this library.

**Contact:** lexa@fi.muni.cz

**Supplementary Information:** Links to additional data and figures available at the journal's web site.

## 1 INTRODUCTION

Triplexes are local structural variants of DNA, wherein the molecule adopts a specific secondary structure differing from a canonical duplex by the recruitment of a third DNA strand. The third strand binds to the duplex by Hoogsteen or reverse Hoogsteen bonds with stringency of the same order of magnitude as duplex-forming strands for the most stable nucleotide combinations (reviewed by Frank-Kamenetskii and Mirkin, 1995). Depending on the source of the third strand, triplex DNA can be $intrastrand$ and $interstrand$, or $intramolecular$ and $intermolecular$. The third strand may just come from the other strand of the same DNA duplex or from a completely different DNA molecule, as is the case with triplex-forming oligonucleotides (Knauert and Glazer, 2001). Nucleotides in the middle strand of a triplex have Watson-Crick base-pairing to one nucleotide and Hoogsteen or reverse Hoogsteen pairing to another nucleotide. Together they form a triplex-forming triplet (also called triad) (Soyfer and Potaman, 1995; Mirkin and Frank-Kamenetskii, 1994). Depending on the orientation of the third strand, we distinguish $parallel$ and $antiparallel$ triplexes, named according to the orientation of the third strand in respect to the central strand. Figure 1 shows eight types of $intramolecular$ triplex structures considered in this paper. A given sequence on the $(+)$ strand of a DNA molecule can possibly support all eight types, but necessarily, only one of the types will be formed at any particular moment. In DNA triplexes, there is a requirement for neighboring triplets to be isomorphic, otherwise the potential triplet would be under strain, hindering the binding of the third strand (Thenmalarchelvi and Yathindra, 2005; Rathinavelan and Yathindra, 2006). Regardless of orientation and geometry, the middle nucleotide is generally a purine-containing one, to support the extra hydrogen bonds needed to bind the third nucleotide.

Because the middle nucleotide is almost invariably one with a purine base, attempts to correlate sequence with triplex-forming properties usually involve detection of homopurine and homopyrimidine tracts in the analyzed sequence. For example Gaddis *et al*. (2006) created a web-based program that identifies target sequences for triplex-forming oligonucleotides. The program identifies homopurine stretches that are allowed to be occasionally interrupted by a pyrimidine. While this is an appropriate method for detection of strong triplex-forming signals, we consider this to be an oversimplification. Numerous papers have reported the existence of imperfect triplexes (Xodo *et al*., 1993; Roberts and Crothers, 1991; Mergny *et al*., 1991), including cases where the authors

*to whom correspondence should be addressed

© Oxford University Press 2011.

deliberately changed individual nucleotides to observe the effects of such change. Changes resulting in the formation of non-canonical triplets did not necessarily disrupt the entire triplex. It is conceivable that many of the imperfect triplexes may still have similar biological activity to their ideal counterparts. One possible explanation for the existence of imperfect triplexes is that they may allow an overlap between the structural signal and some other sequence feature, such as nucleosome positioning pattern or a regulatory protein-binding sequence. Kinniburgh (1989) proposed a triplex structure containing a single deletion to explain his experimental results. Additionally, analyzed sequences may contain errors, including occasional deletions and insertions.

The existence of triplex DNA has been repeatedly associated with important biological processes at molecular level, making them an attractive target in sequence analysis. Most of the observed associations suggest roles in mutagenesis, recombination and gene regulation. Non-B DNA structures, including DNA triplexes, have been shown to cause deletions, expansions and translocations in both prokaryotes and eukaryotes (Raghavan *et al*., 2005). Their distribution is not random and often colocalizes with sites of chromosomal breakage (Zhao *et al*., 2010). Triplex structures can block the replication fork and result in double-stranded breaks (Dixon *et al*., 2008). Unlike other non-canonical structures, triplex-forming sequences are found frequently in promoters and exons and have been found to be involved in regulating the expression of several disease-linked genes (Wang and Vasquez, 2004). In some cases, the mutagenesis induced by such sequences is enhanced by their transcription (Belotserkovskii *et al*., 2007), possibly via transcriptional arrest.

Sequence-structure relationships of triplexes were brought into a small number of computational tools for identifying relevant sequences in genome sequences. Schroth and Ho (1995) analyzed the occurrence of inverted and mirror repeats in three genomes. Hoyne *et al*. (2000) analyzed the $E.coli$ genome for intrastrand triplex sequences. Another recent work (Cer *et al*., 2010) created a web-based catalog of non-B DNA sequences in major mammalian genomes. Their definition of triplex covers the most stable canonical triplexes made of G.GC/A.AT and C.GC/T.AT triplets, but leaves little room for possible errors. Jenjaroenpun and Kuznetsov (2009) created a web-based analysis tool for triplex target sequences.

Intramolecular triplex DNA (also called H-DNA) has been shown to exist both $in\ vivo$ and $in\ vitro$ (Hanvey *et al*., 1988). Its formation also depends on the topological state of the given DNA molecule. While sequences supporting canonical triplets, such as $(CT(T))_n$ and $(GA(A))_n$ tracts, form triplexes readily, imperfect triplexes may require special conditions, such as low superhelical density, or certain pH to form. $In\ vitro$, superhelical density and pH can be easily controlled. $In\ vivo$, pH is tightly controlled by the cell, while the topological state of any stretch of genomic DNA is generally unknown, but presumed to be under regulatory control as well. This uncertainty is the main reason for using the term ”triplex-forming sequence” or ”triplex-forming potential”, which hints that while the sequence should be capable of forming a triplex, it may only be formed under special circumstances.

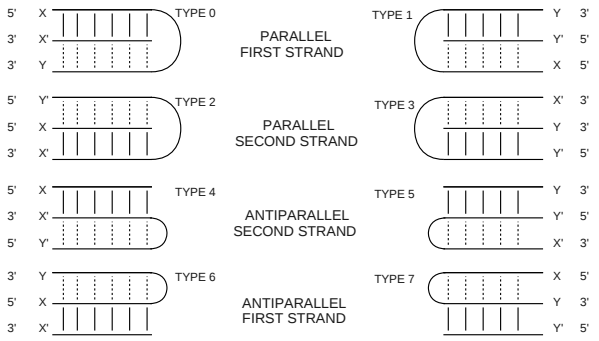


Fig. 1: Eight types of triplexes that are detected in separate runs of the algorithm for a given region. Numbering of types is shown as used in the accompanying software (see Supplemental information). Watson-Crick base-pairing is shown by vertical bars. X and Y are two nucleotides on the same strand that will form a triplet. The eight possible triplets are: Y.X’X, Y’.XX’, Y’.X’X, Y.XX’, X.Y’Y, X’.YY’, X’.Y’Y and X.YY’ (N’ - a nucleotide complementary to N; ”.” - Hoogsteen or reverse Hoogsteen bond)

## 2 APPROACH

Based on available literature, we assume there are two important classes of sequence-based imperfections (errors) destabilizing potential triplex structures.

- Base-pairing mismatch
- Geometrical mismatch

A base-pairing mismatch occurs upon the formation of a nucleotide triplet that does not support strong Hoogsteen or reverse Hoogsteen bonds. The ability to form the bond and its strength is related to the number of hydrogen bonds that can be made between the $2^{nd}$ and $3^{rd}$ strand base. In this paper, we present an algorithm that is based on scores assigned to base triplets. The scores are meant to approximate energy contributions of individual triplets, but at the same time to be simple enough to support rapid searching that could be used as pre-filtering, preceding detailed energy calculations on the candidate sequences.

A geometrical mismatch occurs when directly neighboring triplets in a structure are not isomorphic. This places extra stress on the backbone of the third DNA strand preventing it from creating optimal hydrogen bonds. According to Thenmalarchelvi and Yathindra (2005), conformational changes necessitated by triplet non-isomorphism are found to induce an alternative zig-zag backbone structure for the third-strand in special cases. Accordingly, we made our algorithm favor triplet combinations that are either isomorphic or made of non-isomorphic pairs that could form a zig-zag shape by canceling their geometric effect on the third-strand backbone.

We currently ignore other known factors of triplex DNA formation, such as the competition between alternative structures (Rippe *et al*., 1992), 4th strand (the strand which is not part of the predicted triplex) secondary structure, effects of C+ distribution (Seidman and Glazer, 2003; James *et al*., 2003) and other distortions caused by electrostatic forces (Kang *et al*., 1992; Tan

and Chen, 2006). Most of these factors depend non-trivially on the environment (Plum *et al*., 1995). Since the algorithm does not consider the environment, we focus primarily on sequence-coded effects and the resulting constraints which can be computed using the information from primary structure. Destabilizing effects of loop lengths that differ from the optimum of about 5 nucleotides (Haasnoot *et al*., 1986) and the overall length of the triplex (Tan and Chen, 2006) are partly accounted for, since these parameters can be set as hard limits in our implementation, to narrow the search space.

## 3 METHODS

**Datasets** To evaluate the algorithm on selected datasets, we prepared a set of sequences to work with (all about 4.7 Mbp to match the size of $E.coli$ genome): i) a random nucleotide sequence, ii) $E.coli$ K-12 MG1655 complete genome (the 1995 U00096.1 version to be able to compare our results to previous publications) iii) $E.coli$ K-12 MG1655 complete genome (the current U00096.2 version for proper positioning in genome browsers), iv) a randomized nucleotide sequence of the same $E.coli$ genome v) a part of the human chromosome 5 sequence (positions 144635154 to 149340649) and vi) a randomized version of the same human sequence. For the human randomized sequence we also generated a triplex-seeded version with 418 triplex-forming sequences from literature inserted at positions approximately 10000bp apart. All the sequence data is available as supplementary data and can also be downloaded from $http://www.fi.muni.cz/{\sim}lexa/triplex$. Random sequences were generated with equal probability for all four bases, randomized sequences were prepared with an in-house algorithm seqmix-0.2 (see Supplementary information).

**Molecular simulations of triplets** To obtain objective information about isomorphic groups we analyzed the angle and radius formed by C1 atoms of triplet nucleotides as defined in Thenmalarchelvi and Yathindra (2005). The groups were determined using the following procedure. First, the structures of all considered triplets were constructed using the NAB language from AmberTools 1.4 and their potential energy surface was explored for local minima by moving and rotating the third (Hoogsteen) base in the plane formed by the other two bases. The energy function was parametrized using the *ff99bsc0* set (Perez *et al*., 2007). The obtained local minima were filtered according to the values of the C1 angle ($t$) and the ratio $|WH|/|CH|$, where $|WH|$ represents the distance between the C1 atoms of the Hoogsteen pair and $|CH|$ represents the distance between the C1 atoms of the mutually unpaired bases. Filtering thresholds were derived from measurements on a set of real structures, namely the structures 135D, 149D, 1BCB, 1D3X (PDB identifiers). The specific thresholds used were $70 \leq t \leq 130$ and $0.54 \leq |WH|/|CH| \leq 0.88$. From the resulting set of local minima, the structure with the lowest potential energy was selected as the source of the parameters $t$ and $r$ (the radius of the circle formed by the C1 atoms). Finally, the groups were established by performing cluster analysis using Ward's method and euclidean distance between the ($t$,$r$) vectors. These results were interpreted to obtain isomorphic groups in Table 1, detailed results are available as Supplementary information.

**Testing overview** We tested our implementation for correctness and usability. Clearly, the algorithm will only be useful, if it is capable of identifying potential triplex-forming sequences in a genomic background with a reasonable success rate. To test the implementation in this respect, we performed statistical tests on real and randomized sequences, a sequence recovery test on the triplex-seeded sequences, we compared our solution to previously published results for the $E.coli$ genome (Hoyne *et al*., 2000) and a currently published non-B DNA database (Cer *et al*., 2010).

**Statistical tests** The statistical tests served to find parameters for the distribution of scores on randomized sequences and establish a proper threshold above which candidate hits should be considered significant. The distribution of scores was modeled according to principles used for evaluating BLAST results and other sequence similarity scores (Altschul *et al*., 1994; Korf *et al*., 2003), since the alignment of a DNA strand against itself is statistically similar to aligning two different sequences. This treatment allowed us to fit the score distribution with an extreme value distribution function and fit the parameters $\lambda$ and $\mu$ as described by Korf *et al*. (2003). To carry out the calculation we used a function from hmmer-2.3.2 source code (Eddy, 1997).

**Recovery tests** The recovery tests evaluated how many of the introduced triplex-forming sequences were recovered for a selected significance threshold (P-value) from different backgrounds sequences. We used the commonly used characteristics for such experiments: specificity (precision), sensitivity (recall), $F_2$ measure and accuracy (Manning *et al*., 2008). The algorithm was tested against our triplex-seeded sequence and a database of non-B DNA (Cer *et al*., 2010).

$E.coli$ **tests** We compared our tool and its performance on the $E.coli$ genome sequence to the results published by Hoyne *et al*. (2000). Additionally, we calculated the genome positioning of program output in respect to known $E.coli$ genes, counting the frequency with which predicted triplexes fell inside the gene, outside any genes or intersected with them. Distance to the closest gene was calculated as shown in Figure 6.

## 4 THE ALGORITHM

Our approach to search for approximate triplexes is based on a dynamic programming (DP) algorithm to search for approximate palindromes that can be traced back to Landau and Vishkin (1986). The relationship between triplex DNA and palindromes stems from the fact, that one of the DNA strands in the triplex must fold back onto itself, either for Hoogsteen base-pairing or for reverse Hoogsteen base-pairing, depending on the type of triplex that is to be formed (parallel or antiparallel) and the nucleotide sequence present at the site in question. We will call the part of the triplex that folds back onto itself *self-recognizing*.

A DP matrix is constructed so that one side represents the original sequence, while the other contains the same sequence written backwards (see Figure 2). With such setup, the main antidiagonal of the DP matrix represents the $n$ possible central starting positions for the self-recognizing parts of triplexes with an odd number of nucleotides in the loop. The neighboring antidiagonal contains the other $n-1$ possible starting sites for the triplexes with even number of nucleotides in their loops. Naturally, diagonals starting at any of these positions represent potential triplexes. If we fill the cells representing the starting positions with zeros, we can start filling the DP matrix along the diagonals. At each position $[i,j]$ of the DP matrix, we compare the symbols at positions $i$ and $j$ in the original sequence. If they represent a pair present in triplex-forming triplets (tabulated in Table 1), they are evaluated with positive score. In opposite case, they are penalized with a negative score value. The numbers entered represent the best score in the subsequence evaluated so far.

The necessity for a dynamic programming algorithm comes from the possibility to insert gaps into the triplexes, where symbols in some positions have no symbols to pair up with in the other arm of the self-recognizing sequence. In terms of the described algorithm, this means moving from one diagonal to a neighboring one when calculating the score. At any position, three possibilities are evaluated:

1. Extending the existing triplex along the diagonal - *match* or *mismatch*,
2. Inserting a gap at position *i* of the original sequence - *insertion*,
3. Inserting a gap at position *j* of the original sequence - *deletion*.

The solution that leads to the maximum score value is recorded in the DP matrix, while the other possibilities are discarded.

In comparison to a similar algorithm for approximate palindrome detection, we have introduced three important modifications. First, we redefined the concept of match and mismatch. Instead of being made up

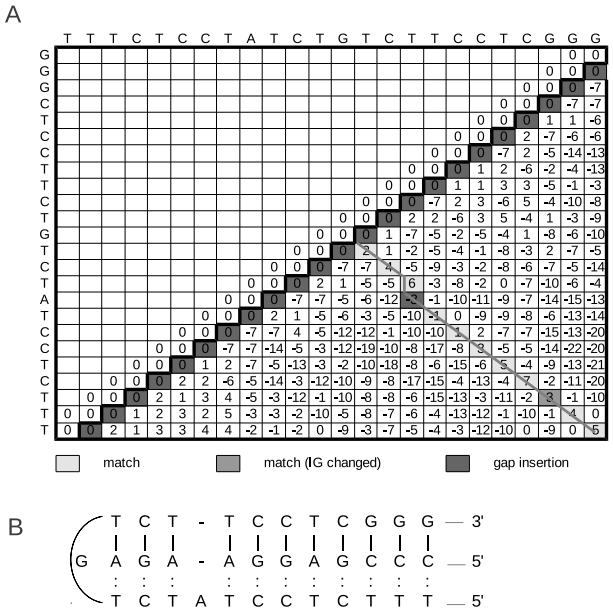


Fig. 2: Triplex detection by the dynamic programming (DP) algorithm demonstrated on the string $tttctcctatctgtcttcctcggg$: A) The DP matrix with calculated score values. Because of space limitations, loop size was forced to 1. B) Triplex alignment. Hoogsteen bonds are shown by semicolons.

**Table 1.** Triplex scoring of canonical and less usual triplets. The final score values for both Hoogsteen and reverse-Hoogsteen bonds are in accordance with tables 4.1 and 4.2 in Soyfer and Potaman (1995). Isomorphic groups shown here are based on residual twist calculations using molecular dynamics simulations with the $nbd$ program (AmberTools). . - Hoogsteen bp; : - Watson-Crick bp; $tts$ - tabulated triplet score.

| Triplex type | Triplet H.WC:WC | Score ($tts$) | Isomorphic group | References |
|---|---|---|---|---|
| PARALLEL | T.A:T | 2 | a | (2) |
| | T.G:C | 1 | a | (3) |
| | C.G:C | 2 | a | (1,2) |
| | G.G:C | 1 | b | (7) |
| | G.T:A | 2 | b | (4) |
| | T.C:G | 1 | b | (7) |
| ANTIPARALLEL | A.A:T | 2 | c | (2,5) |
| | A.G:C | 1 | d | (5,6) |
| | T.A:T | 2 | c | (2,5) |
| | T.C:G | 1 | e | (6,8) |
| | C.A:T | 1 | d | (6,7,9) |
| | G.G:C | 2 | e | (2,5) |

1) Walter *et al.* (2001) 2) Goni *et al.* (2004) 3) Ghosal and Muniyappa (2006) 4) Gowers and Fox (1998) 5) Mirkin and Frank-Kamenetskii (1994) 6) Raghavan and Lieber (2007) 7) Soyfer and Potaman (1995) 8) Beal and Dervan (1992) 9) Dayn *et al.* (1992)

by pairs of nucleotides with only two possible base-pairs, triplexes can be thought of as sequences of triplets with many possible combinations of nucleotides in the triplet. There are 16 possible base-pairs for parallel DNA strands and another 16 for antiparallel strands. For these reasons, we constructed a general similarity matrix instead of using a single match rule and score.

Second modification brings geometrical considerations into the algorithm, making certain sequences of triplets less desirable than others. This is similar to the nearest-neighbor scoring used in duplexes, although we are not as much concerned about base stacking as we are about the geometry of the third strand and its ability to position itself for optimal hydrogen bonding. As discussed by Thenmalarchelvi and Yathindra (2005); Rathinavelan and Yathindra (2006), some combinations disrupt the backbone geometry. We therefore decided to divide the triplets into isomorphic groups. Groups of triplets from one group are more likely to form stable triplexes than other sequences. Our modification assigns the information about isomorphic groups to the last computed DP matrix cell on each diagonal. When calculating a new cell, we lower the score if the newly evaluated triplet belongs to a different isomorphic group than the preceding one. The score calculation is

$$S[i,j] = max \begin{cases} S[i,j-1] + gp \\ S[i-1,j] + gp \\ S[i-1,j-1] + tts[a,b] + nip \end{cases} \quad (1)$$

where $a$, $b$ are characters at appropriate row and column, $tss$ is tabulated triplet score, $gp$ is gap penalty and $nip$ is no-isomorphism penalty.

The third consideration is to account for all the possible ways a triplex can form from a given sequence, i.e. which three strands combine together and in which orientation (Figure 1). There are always eight ways that can give rise to a intramolecular triplex at a given position, since there are two strands that can serve as the third strand, each having two ends that can loop back onto the double-stranded region and in each of these cases it can attach on either side of the duplex in a parallel or antiparallel fashion, forming Hoogsteen and reverse Hoogsteen bonds respectively. In order to detect all types of triplexes the computation is repeated eight times with scoring matrices specific for parallel and antiparallel triplexes.

## 4.1 Scoring Function

We evaluate the combinations based on their ability to form Hoogsteen base-pairs, tabulating the 32 values as complementarity scores. One way to populate such table is to consider all canonical triplets to represent a match and everything else a mismatch. Because the ability to form Hoogsteen bonds depends partly on the environment of the given nucleotide, we took a semi-empirical approach, giving all canonical triplets a match score of 2, scanning triplex literature for examples of less usual triplets and giving those a score of 1, while all other combinations are scored as a mismatch (see Table 1). Other approaches leading to a better scoring scheme are certainly possible, but beyond the scope of this paper.

## 4.2 Triplex loop detection

The algorithm introduced in this section has been designed to detect the best candidates for triplex formation. To avoid the inclusion of free-strand and loop nucleotides into the overall score for a particular triplex (because these nucleotides do not participate in Watson-Crick or Hoogsteen base-pairing), our calculations use a technique composed of a combination of local and global alignment.

In terms of the DP matrix, potential loops always begin at the main antidiagonal, extending up to $l_{loop_{max}}$ (user-defined algorithm parameter), using Equation 1 to calculate new values. The first $2l_{loop_{max}}$ antidiagonals are therefore calculated by a technique similar to the one used in Smith-Waterman local sequence alignment. In this part, we allow the score of a growing triplex to grow or decline. However, if the density of errors is

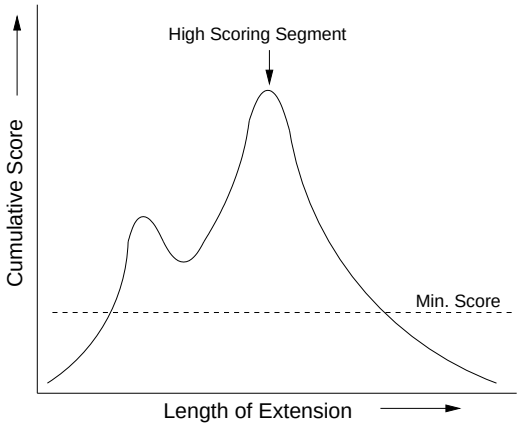


Fig. 3: Detection of high scoring segments.

high enough to bring the score into the negative territory (potential loop occurrence), we do not allow the score to become negative.

Once the calculations exit the area of a potential loop, the calculations continue in a global alignment mode. This way the algorithm can detect high-quality triplex candidates without considering errors that fall within potential loops.

### 4.3 Triplex detection

The best triplexes in the DP matrix can be identified as those reaching the highest score. To allow detection of such *high scoring segments* (HSS) during the calculation, we use a technique similar to the one used in the BLAST program. Once the score rises above a preset threshold value, the region responsible for the score is considered a potential triplex. The score is monitored (allowed both to increase and decrease) until it falls below a preset threshold. The sequence from the beginning (the first antidiagonal) up to the maximum score becomes the HSS of the potential triplex (see Figure 3).

A number of filtration mechanisms can be applied to the step of HSS segment detection. One of the problems we had to deal with (causing false HSS detection), was the transfer of scores from neighboring diagonals. In the presence of a high-quality triplex sequence, neighboring diagonals adopt its high score by introduction of an extra insertion or deletion. We therefore check for such cases and only report genuine HSS scores and not the neighboring derivatives.

Further filtration is carried out based on statistical significance of the results, eliminating all short or low-quality potential triplexes below a user-defined E-value or P-value threshold (see Results for details on P-value calculations on experimental datasets). A pair of filtering programs (prefilter_gff.c and filter_gff.c, see Supplementary information) were used to filter out results not supporting a local score maximum (meaning there is a better result nearby).

### 4.4 Time and space complexity

Time complexity: the calculation of the entire triangle of the DP matrix has $n^2/2$ steps. However, when analyzing real or random sequences, the likelihood of finding a potential triplex decreases with its length (see Results for a detailed description of this effect). Therefore, for most practical purposes we only need to evaluate a limited number of antidiagonals, say $2l$, where $l$ is the maximal length of detected triplexes. Time complexity thus becomes $O(2ln)$.

Space complexity: With respect to data dependencies, only the values for the last two antidiagonals are necessary for calculation. Thus the space complexity of our algorithm is $O(2n)$.

Both simplifications/efficiency enhancements used to derive the time and space complexities allow us to easily extend the algorithm to perform an *incremental calculation*. If upon completion of the calculation we find that the number of antidiagonals was not sufficient, leaving several potential triplexes unresolved, we can pick up the score values from the last two diagonals and continue in the calculations in another $2l$ antidiagonals.

## 5 RESULTS AND DISCUSSION

We subjected the algorithm to increasing levels of scrutiny to verify the validity of our searching procedures, fine-tune some of the parameters and establish the biological relevance of selected results.

Initial experiments were directed towards establishing reasonable mismatch and insertion/deletion penalties. The penalties have to be high enough to allow for a negative average score per triplet (Korf *et al.*, 2003). Without any rigorous optimization, we found the combination $mismatch\ -7$, $insertion$ or $deletion\ -9$, $no\_isomorphism\ -5$ to fulfill these criteria and work reasonably well on all sequences.

Identification of a higher number of potential triplexes in real-world sequences compared to random and randomized sequences is the first confirmation that the patterns we are collecting using this approach are not random, but rather specific combinations with a possible function that are less frequent in random sequences.

For a rigorous test of non-randomness of the identified candidates, we tested our implementation of the algorithm against a set of 4.7MBp DNA sequences from $E.coli$ and human genomes, their randomized version and a triplex-seeded randomized $E.coli$ genome (see Methods). For each of the sequences, we used the program to identify all potential triplexes and their scores. Since an incrementally detected triplex-forming sequence must obey similar rules as an incrementally growing sequence alignment (only with different base-pairing rules), we would expect the obtained scores to obey an extreme value distribution described by Altschul *et al.* (1994).

$$P(S > x) = 1 - e^{-e^{-\lambda * (x-\mu)}} \tag{2}$$

We used a maximum likelihood method described by Eddy (1997) to fit our scores to this function. The resulting values of $\lambda$ and $\mu$ are given in Table 2. Figure 4 shows a graphical representation and corresponding parameter values of triplex scores for the different datasets used. Clearly, randomized sequences have a lower content of high-scoring sequence patterns. Also, human sequences seem to be richer in potential triplex-forming sequences, comparable in density to the artificially seeded $E.coli$ sequence with one triplex sequence per every 10000bp.

We used the $\lambda$ and $\mu$ values to derive statistical thresholds for searching (Table 2).These are different for parallel and antiparallel triplexes, since the two use a different similarity matrix, resulting in different score distributions.

Next, we analyzed the non-B DNA database triplex predictions (Cer *et al.*, 2010) and our triplex-seeded sequence containing 421 inserted triplexes with artificial mismatches and insertions. Our program preferentially recovered the positions of known triplex sequences. Figure 5 shows sensitivity, specificity, accuracy and $F_2$ measure for these two sets. F measure is the harmonic mean of sensitivity and specificity. $F_2$ measure is its commonly used modification, which gives higher priority to recall. $F_2$ measure values above 40% are satisfactory, given that 100% of potential

**Table 2.** The results of fitting an Extreme Value Distribution function to score distribution data obtained from randomized sequences of $E.coli$ and human genomes. The threshold shown here for reference purposes is the score above which less than 10 sequences were found in randomized data. Precise E-values and P-values can be calculated from values of $\lambda$ and $\mu$ according to Equation 2.

| Randomized sequence data | $\lambda$ | $\mu$ | threshold |
|---|---|---|---|
| $e.coli$ | 0.91 | 6.00 | 20 |
| human chr5 | 0.84 | 6.28 | 21 |

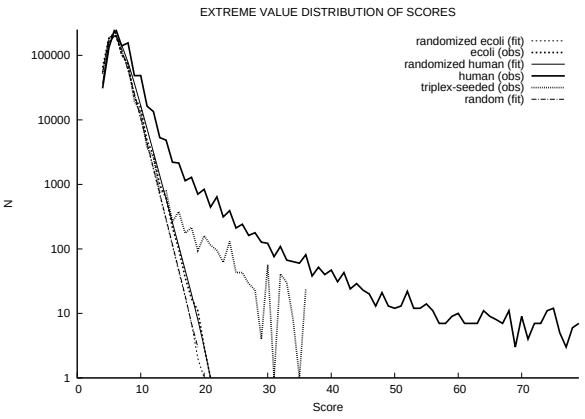


Fig. 4: Log-scale extreme value distribution functions for $E.coli$ (dashed line), human (solid line) and triplex-seeded datasets (dotted line) compared to background random sequences (thin lines), including a random sequence, randomized $E.coli$ and human sequences. A maximal likelihood fit to the random sequences is available in Table 2. While the $E.coli$ genome contains potential triplex sequences only slightly above background levels, the human genome seems to be rich in such sequences with density similar to the triplex-seeded dataset.

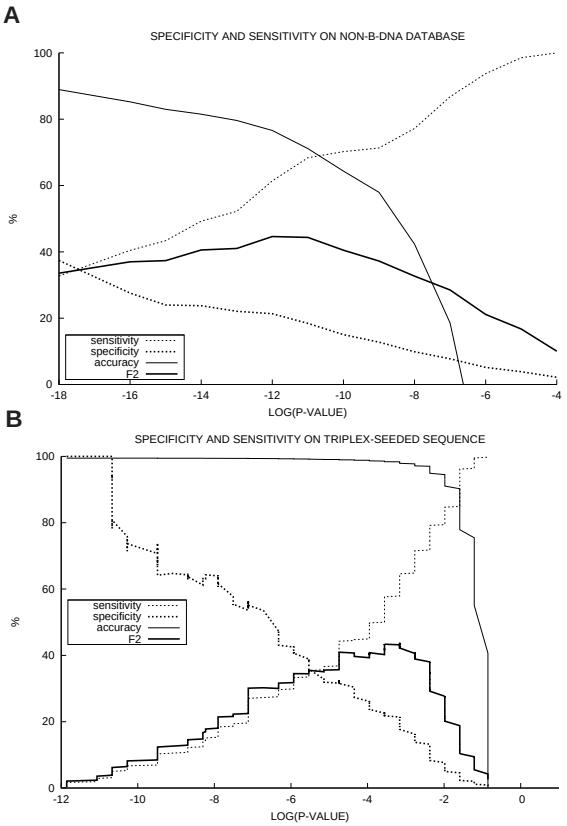


Fig. 5: Sensitivity, specificity, accuracy and $F_2$ measure calculated for A) the non-B DNA database (Cer *et al*., 2010); B) the triplex-seeded dataset. The figures show that the best matches obtained with the described algorithm and settings are entirely made up of the seeded sequences. At lower P-values we start picking up some sequences from the background sequence, acceptable results before accuracy drops sharply are achieved for P-values of less than $1.0 * 10^{-2}$

triplexes are recovered with a P-value better than 0.01. Some loss of performance on triplex-seeded data is understandable, since mismatches and insertions/deletions were introduced in sequences as short as 6bp.

One of the detected sequences, is a well-studied triplex from human metallothionein-I promoter (Bacolla and Wu, 1991). This sequence was the second highest-scoring sequence in the triplex-seeded data, scoring 34 with a P-value of $5.10^{-9}$. Interestingly, we detected two high-scoring subsequences within the MT-I promoter potential triplex, supporting the view of Bacolla and Wu (1991) and Becker and Maher (1998) that alternative triplex structures may be formed at this specific site.

For an alternative evaluation of the validity of our algorithm we analyzed the $E.coli$ genome for triplex-forming sequences and compared the results with those described in Hoyne *et al*. (2000). They searched for potential intrastrand triplex (PIsT). The PIsT element requires the consecutive occurrence of all three triplex-forming blocks of nucleotides, while potential intramolecular triplex (PImT) element requires the consecutive occurrence of just two triplex-forming blocks (the third block is provided by the parallel strand). Thus, every PIsT element by definition contains also a PImT element.

For each of the 25 PIsT elements presented in Hoyne *et al*. (2000) we are able to identify the corresponding PImT element in $E.coli$ genome with appropriate parameter settings. The score of these elements range from the value of 6 to the value of 20 and the corresponding P-values vary from $4.7*10^{-1}$ to $2.9*10^{-6}$. The best potential triplex element in $E.coli$ genome found by our algorithm scored 21 with a P-value of $1.2 * 10^{-6}$.

Finally, we examined some of the identified potential triplex sites for biological relevance. Producing a GFF file with results enabled us to view them in the UCSC Genome Browser. Here, we noticed a possible relationship to known $E.coli$ genes. To test this, we counted the number of predicted triplexes falling within genes, outside genes or less than 100bp from gene boundaries (Figure 7A).

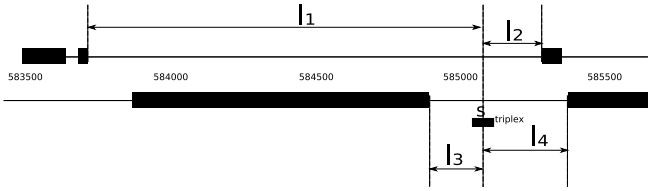


Fig. 6: The definition of the closest gene as used in the numerical experiment. For each triplex we identified its center $S$ (rounded up for even triplexes), and calculated the distances $l_1$, $l_2$, $l_3$ and $l_4$ to the closest upstream and downstream gene borders on both DNA strands. The minimum of these four values was used.

We also calculated the number of predicted triplexes occuring at different distances from the closest gene (Figure 6) and calculated the ratio of this value to randomly placed positions. There seems to be some preference for potential triplexes to occur in the -50 to -160 region of known genes (Figure 7B). Given the relatively high P-value at which this effect was still visible, it is possible that it is not directly related to the presence of triplexes, but rather a result of shared sequence characteristic between triplexes and regulatory sequences, such as their underlying palindromic nature.

Another observation showed these positions to be clustered at boundaries of evolutionarily poorly conserved regions. A quick literature search revealed a possible connection. Non-B DNA structures are likely to pose a physical barrier to transcriptional apparatus, causing possible transcriptional arrest at such sites (Young *et al.*, 1991). Transcriptional arrest has been directly linked to increased mutation rate (Belotserkovskii *et al.*, 2007), which could explain some aspect of the above-mentioned positioning in genomes.

While the main purpose of this paper is to present the algorithm itself, a more detailed analysis of the best parameter settings and performance with specific DNA sequences is needed to further increase confidence in this kind of sequence analysis.

Because of the increased complexity of scoring, the outlined procedure for scoring individual triplets within the DP matrix cannot be easily extended to take advantage of suffix arrays as is done with palindromes, to further speed up computation.

Overall, we consider it an advantage that triplex identification can be mapped to a well-researched family of DP algorithms and possibly take advantage of approaches aimed originally at other problems, such as sequence alignment.

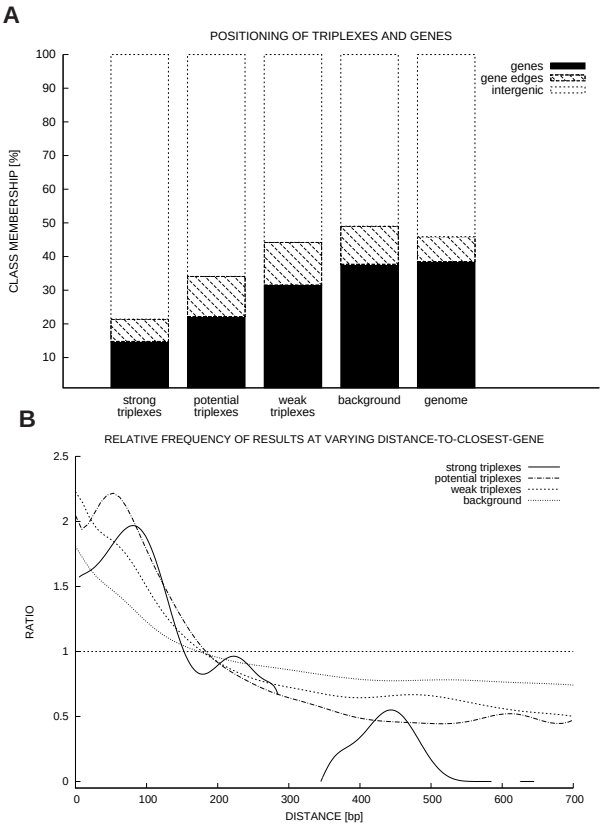


Fig. 7: Graphs showing how potential triplexes identified by the program are positioned in respect to genes in $E.coli$. A) The percentage of triplexes in the results falling inside genes, intersecting with a gene or falling within intergenic segments of the genome. Bars are shown for results of decreasing specificity (from left to right); B) The relative abundance of high-scoring sequences at different distances from nearby genes (relative to randomly placed positions). Both figures were generated after applying the following cutoffs to the results: top 122 (strong triplex), top 1391 (potential triplex), top 15300 (weak triplex), top 106623 (background) and random selection of positions (genome).

## 6 CONCLUSION

We present a novel approach to identifying triplex-forming sequences in genomes and other DNA sequence data. The approach is presented in the form of an algorithm based on previously published algorithms for detection of palindromes. The novelty stems from the adaptation of DP for use with triplexes instead of relying on simpler identification of homopurine and homopyrimidine tracts, which are most appropriate for detection of perfect triplexes. We implemented our algorithm as a program written in C, using a reasonable set of parameters based on published data. The test runs of this program are encouraging, suggesting that the algorithm can provide high speed searches with increased sensitivity for approximate triplex-forming sequences.

## ACKNOWLEDGEMENT

*Funding*: This work has been carried out with the support of grants No. 204/08/1560 and No. 301/10/2370 from the Czech Grant Agency, MSMT Research Grant No.0021630528 – Security-Oriented Research in Information Technology, BUT grants FIT-S-11-1 – Advanced secured, reliable and adaptive IT and FIT-S-11-2 – Recognition and presentation of multimedia data.

*Conflict of Interest*: none declared.

## REFERENCES


Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) Issues in searching molecular sequence databases, *Nat Genet*, **6**, 119–129.

Bacolla, A. and Wu, F.Y-H. (1991) Mung bean nuclease cleavage pattern at a polypurine-polypyrimidine sequence upstream from the mouse metallothionein-I gene, *Nucleic Acids Res*, **1**, 1639–1647.

Beal, P.A. and Dervan, P.B. (1992) The influence of single base triplet changes on the stability of a pur.pur.pyr triple helix determined by affinity cleaving, *Nucleic Acids Res*, **20**, 2773–2776.

Becker, N.A. and Maher III, L.J. (1998) Characterization of a polypurine/polypyrimidine sequence upstream of the mouse metallothionein-I gene, *Nucleic Acids Res*, **26**, 1951–1958.

Belotserkovskii, B.P., De Silva, E., Tornaletti, S., Wang, G., Vasquez K.M. and Hanawalt, P.C. (2007) A Triplex-forming Sequence from the Human c-MYC Promoter Interferes with DNA Transcription *J Biol Chem*, **282**, 32433–32441.

Cer, R.Z., Bruce, K.H., Mudunuri, U.S., Yi, M., Volfovsky, N., Luke, B.T., Bacolla, A., Collins, J.R. and Stephens, R.M. (2010) Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes, *Nucleic Acids Res*, **39**, D383–D391.

Dayn, A., Samadashwily, G.M. and Mirkin, S.M. (1992) Intramolecular DNA triplexes: Unusual sequence requirements and influence on DNA polymerization, *Proc Natl Acad Sci USA*, **89**, 11406–11410.

Dixon, B.P., Lu, L., Chu, A. and Bissler, J.J. (2008) RecQ and RecG helicases have distinct roles in maintaining the stability of polypurine.polypyrimidine sequences, *Mutat Res*, **643**, 20–28.

Eddy, S.R. (1997) Maximum likelihood fitting of extreme value distributions, doi:10.1.1.31.4013 (ftp://ftp.genetics.wustl.edu/pub/eddy/papers/evd.ps).

Frank-Kamenetskii, M.D. and Mirkin, S.M. (1995) Triplex DNA structures, *Annu Rev Biochem*, **64** 65–95.

Gaddis, S.S., Wu, Q., Thames, H.D., DiGiovanni, J., Walborg, E.F., MacLeod, M.C. and Vasquez, K.M. (2006) A web-based search engine for triplex-forming oligonucleotide target sequences, *Oligonucleotides*, **16**, 196–201.

Ghosal, G., Muniyappa, P. (2006) Hoogsteen base-pairing revisited: resolving a role in normal biological processes and human diseases, *Biochem Biophys Res Commun*, **343**, 1–7.

Goni, J.R., de la Cruz, X. and Orozco, M. (2004) Triplex-forming oligonucleotide target sequences in the human genome, *Nucleic Acids Res*, **32**, 354–360.

Gowers, D.M. and Fox, K.R. (1998) Triple helix formation at $(AT)_n$ adjacent to an oligopurine tract, *Nucleic Acids Res*, **26** 3626–3633.

Haasnoot, C.A.G., Hilbers, C.W., van der Marel, G.A., van Boom, J.H., Singh, U.H., Pattabiraman, N. and Kollman, P.A. (1986) On loop folding in nucleic acid hairpin-type structures, *J Biomol Struct Dyn*, **3**, 843–857.

Hanvey, J.C., Shimizu, M. and Wells, R.D. (1988) Intramolecular DNA triplexes in supercoiled plasmids, *Proc Natl Acad Sci USA*, **85** 6292–6296.

Hoyne, P.R., Edwards, L.M., Viari, A. and Maher, L.J. (2000) Searching genomes for sequences with the potential to form intrastrand triple helices, *J Mol Biol*, **302**, 797–809.

James, P.L., Brown, T. and Fox, K.R. (2003) Thermodynamic and kinetic stability of intermolecular triple helices containing different proportions of C+?GC and T?AT triplets, *Nucleic Acids Res*, **31**, 5598–5606.

Jenjaroenpun, P. and Kuznetsov, V.A. (2009) TTS Mapping: integrative WEB tool for analysis of triplex formation target DNA sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome, *BMC Genomics*, **10**, Suppl 3:S9.

Kang, S.M., Wohlrab, F. and Wells, R.D. (1992) Metal ions cause the isomerization of certain intramolecular triplexes, *J Biol Chem*, **267**, 1259–1264.

Kinniburgh, A.J. (1989) A cis-acting transcription element of the c-myc gene can assume an H-DNA conformation, *Nucleic Acids Res*, **17**, 7771–7778.

Knauert, M.P. and Glazer, P.M. (2001) Triplex forming oligonucleotides: sequence-specific tools for gene targeting, *Hum Mol Genet*, **10** 2243–2251.

Korf, I., Yandell, M. and Bedell, J. (2003) BLAST, O'Reilly & Associates, Inc., Sebastopol, 368 pages.

Landau, G.M. and Vishkin, U. (1989) Fast parallel and serial approximate string matching, *J Algorithms*, **10** 157–169.

Manning, C.D., Raghavan, P. and Schütze, H. (2008) Introduction to information retrieval, Cambridge University Press, Cambridge, 496 pages.

Mergny, J.L., Sun, J.S., Rougée, M., Montenay-Garestier, T., Barcelo, F., Chomilier, J. and Hélène, C. (1991) Sequence specificity in triple helix formation: experimental and theoretical studies of the effect of mismatches on triplex stability, *Biochemistry*, **30**, 9791–9798.

Mirkin, S.M. and Frank-Kamenetskii, M.D. (1994) H-DNA and related structures, *Annu Rev Biophys Biomol Struct*, **23**, 541–576.

Perez, A., Marchan, I., Svozil, D., Sponer, J., Cheatham, T.E., Laughton, C.A. and Orozco, M. (2007) Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of $\alpha/\gamma$ Conformers, *Biophys J*, **92**, 3817–3829.

Plum, G.E., Pilch, D.S., Singleton, S.F. and Breslauer, K.J. (1995) Nucleic acid hybridization: triplex stability and energetics, *Annu Rev Biophys Biomol Struct*, **24**, 319–350.

Raghavan, S.C. and Lieber, M.R. (2007) DNA structure and human diseases, *Front Biosci*, **12**, 4402–4408.

Raghavan, S.C., Chastain, P., Lee, J.S., Hegde, B.G., Houston, S., Langen, R., Hsieh, C.-L., Haworth, I.S. and Lieber, M.R. (2005) Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation, *J Biol Chem*, **280**, 22749–22760.

Rathinavelan, T. and Yathindra, N. (2006) Base triplet nonisomorphism strongly influences DNA triplex conformation: effect of nonisomorphic G* GC and A* AT triplets and bending of DNA triplexes., *Biopolymers*, **82**, 443–61.

Rippe, K., Fritsch, V., Westhof, E. and Jovin, T.M. (1992) Alternating d(G-A) sequences form a parallel-stranded DNA homoduplex, *EMBO J*, **11**, 3777–3786.

Roberts, R.W. and Crothers, D.M. (1991) Specificity and stringency in DNA triplex formation, *PNAS USA*, **88**, 9397–9401.

Schroth, G.P. and Ho, P.S. (1995) Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA, *Nucleic Acids Res*, **23**, 1977–1983.

Seidman, M.M. and Glazer, P.M. (2003) The potential for gene repair via triple helix formation, *J Clin Invest*, **112**, 487–494.

Soyfer,V.N. and Potaman, V.N. (1995) Triple-helical nucleic acids, Springer-Verlag, Heidelberg, 360 pages.

Tan, Z.J. and Chen, S.J. (2006) Nucleic acid helix stability: effects of salt concentration, cation valence and size, and chain length, *Biophys J*, **90**, 1175–1190.

Thenmalarchelvi, R. and Yathindra, N. (2005) New insights into DNA triplexes: residual twist and radial difference as measures of base triplet non-isomorphism and their implication to sequence-dependent non-uniform DNA triplex, *Nucleic Acids Res*, **33**, 43–55.

Walter, A., Schütz, H., Simon, H. and Birch-Hirschfeld, E. (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.

Wang, G. and Vasquez, K.M. (2004) Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells, *PNAS USA*, **101**, 13448–13453.

Xodo, L.E., Alunni-Fabbroni, M., Manzini, G. and Quadrifoglio, F. (1993) Sequence-specific DNA-triplex formation at imperfect homopurine-homopyrimidine sequences within a DNA plasmid, *Eur J Biochem*, **212**, 395–401.

Young, S.L., Krawcyzk, S.H., Matteucci, M.D. and Toole, J. (1991) Triple helix formation inhibits transcription elongation in vitro, *PNAS USA*, **88**, 10023–10026.

Zhao, J., Bacolla, A., Wang, G. and Vasquez, K.M. (2010) Non-B DNA structure-induced genetic instability and evolution, *Cell Mol Life Sci*, **67**, 43–62.