Agnieszka Danek
Silesian University of Technology, Institute of Computer Science, Poland
e-mail: agnieszka.danek@polsl.pl
Rafał Pokrzywa
Silesian University of Technology, Institute of Computer Science, Poland
e-mail: rafal.pokrzywa@polsl.pl

Algorithm for Searching for Approximate Tandem Repeats based on the Burrows-Wheeler transform

Genomic sequences tend to contain many types of repetitive structures of different length, either interspersed or tandem. Tandem repeats play an important role in the gene expression and transcription regulations. They can be used as markers for DNA mapping and DNA fingerprinting. Some, when occurring in increased, abnormal number, are known to be the cause of inherited diseases. All functions of tandem repeats in genomic sequences are still not well defined and understood. However, growing biological databases together with tools for efficient identification of these repeats may lead to discovery of their specific role or correlation with particular symptoms or diseases.

Perfect tandem repeat consists of successive duplications of some motif. Typically tandem copies are approximate due to mutations. Hence approximate tandem repeat (ATR) can be defined as a consecutive, inexact copies of some motif. In our considerations we are assuming that two such successive repeats must be of equal lengths and can differ only by an established number of mismatches. Dissimilarity of these two approximate copies is measured using Hamming distance between them. We are interested in finding approximate tandem repeat when each repeated motif is similar enough to the adjacent duplicate.

Algorithm presented is an enhancement of a method for finding perfect tandem repeats in DNA sequences based on Burrows-Wheeler transform (BWT). It uses its intermediate results, groups of particular sequences repeated within the whole input string, to find candidates for double ATR — that is the first stage of searching. The second stage consists of investigating found candidates and accepting or rejecting them as a pair of ATRs. Finally, in last stage, located double ATRs are extended to contain as much successive, similar copies, as possible.

In the first stage the input string is converted according to BWT. This, together with some auxiliary arrays, allows to make use of the alphabetically sorted array of input string suffixes, without the need of storing the whole suffix array structure. The algorithm finds the range of positions of the repeated pattern in the suffix array. It starts with the empty pattern P and recursively appends, in front of P, characters from the considered alphabet. This approach uses the results from the previous iteration to calculate a range of positions for a longer pattern and it is done in a constant time, according to the idea of Ferragina and Manzini. Two sequences from the range of repeated patterns are considered a candidate for a double approximate tandem repeat if they lay close enough to each other within the input string, in particular, if it is possible that they will form an approximate tandem repeat with established, maximum dissimilarity. To limit the number of

redundant candidates the algorithm makes use of the property of two strings of length n and with Hamming distance h between them, which states that two such strings have always a common, matching substring at corresponding positions of length $\lfloor \frac{n}{h+1} \rfloor$ at least. Hence, repeated patterns of length d are used to search only for ATRs of length n that satisfies the equation $d = \lfloor \frac{n}{h+1} \rfloor$ for all acceptable h. Additionally, as positions of previously found ATRs are known, qualifying as a candidate the ATR discovered before is avoided.

In the next stage Hamming distance between found pairs of candidates is measured (checking all possible alignments of found candidates) and if it satisfies the assumptions, the double approximate tandem repeat is reported. In the third, final stage, Hamming distance is measured between marginal motif of found ATR and a neighboring string. As long as it is not greater than the assumed maximum, the ATR is extended in that direction.

The developed algorithm exploits the advantages offered by the BWT algorithm and the suffix array data structure to return ATRs from the input string, assuming that any two consecutive copies within ATR differ at most by a provided Hamming distance.

Acknowledgement: This work was supported by the European Union from the European Social Fund.

References

[1] R. Pokrzywa, *Application of the Burrows-Wheeler Transform for searching for tandem repeats in DNA sequences* Int. J. Bioinform. Res. Appl. vol. 5 (4) (2009) 432–446.

[2] R. Pokrzywa and A. Polański, *BWtrs: A tool for searching for tandem repeats in DNA sequences based on the Burrows-Wheeler transform* Genomics 96 (2010) 316–321.

[3] M. Burrows and D.J. Wheeler, *A block-sorting lossless data compression algorithm* SRC Research Report 124, Digital Equipment Corporation, Palo Alto, California, May 10 1994.

[4] P. Ferragina and G. Manzini, *Opportunistic data structures with applications* Proceedings of the 41st Annual Symposium on Foundations of Computer Science, 2000, pp. 390–398.