

## SUMMARY: REPETITIVE DNA DETECTION AND CLASSIFICATION

### 1 Paper Reference

Jie Zheng, S. Lonardi, Discovery of repetitive patterns in DNA with accurate boundaries. *IEEE Symposium on Bioinformatics and Bioengineering (BIBE)*, 2005.

### 2 Abstract

This paper looks at the problem of identifying the boundaries of repetitive sequences and describes a new approach to this identification that takes into consideration both the length and the frequency of repeats. The definition is derived using a bottom-up approach, starting with the basic building block of repeats. These blocks need to occur a set minimum number of times, and cannot be shorter than a certain length. In addition, they cannot be comprised of additional basic blocks - hence the authors refer to them as *elementary repeats*.

The authors also make a distinction between exact and approximate elementary repeats. The former are repeats that are maximal in length and do not contain non-trivial substrings with a different distribution of occurrences, while the latter differ in that they do not need to appear 'exactly' in the string sequence under consideration. This makes approximate repeats more difficult to identify. As repeats can be simply represented by the pairing [*left boundary, right boundary*], the challenge of identifying repeats comes down to finding their boundaries. The authors thus outline two algorithms (one for exact repeats, the other for approximate repeats) that scan an input sequence string and decide whether or not each position in the sequence is a repeat boundary or not.

The algorithm for exact repeats finds boundaries by detecting positions where there is a change in the occurrence of seeds, or *l-mers* in the input sequence, where *l* is the threshold that sets what a non-trivial substring is. The algorithm for approximate repeat identification follows a similar methodology but extends the first algorithm by using similarity between occurrence lists of two successive *l-mers* to first establish how likely it is that they belong to the same approximate repeat.

The experimental results indicated that accuracy for both exact and approximate repeats with average repeat lengths 50 or 200 was very high - with sensitivity values higher than 98% and true positives higher than 95%, although they found this dropped for longer approximate repeats as these were harder to detect.

### 3 Discussion

The approach the authors of this paper take to identifying *de novo* repeats differs from other repeat identification methods as it does not require a library of repeat families and neither does it need try to output all repeated regions. In addition a number of other methods have ignored repeat frequency, and only focused on maximizing the length of the repeats. The authors believe that without taking both length and frequency into consideration, the definition of repeats is not as biologically meaningful.

The two papers covered in class are similar to this paper as the methods they outline also concern detection of repeat boundaries. A.L. Price et al also use *l-mers* - they describe a dynamic programming algorithm that identifies repeat families by using these high frequency *l-mers* as seeds and then tries to progressively extend the seeds into longer sequences according to alignments inferred between the input sequence and occurrences in the genome. The algorithm described by R.C. Edgar and E.W. Myers on the other hand finds boundaries by using distinctive patterns of local alignments induced by four classes of repetitive structure - terminal repeats, tandem arrays, pseudosatellites and dispersed families. Their method thus aimed to attain high specificity at the sacrifice of some sensitivity.

Results of this paper were not however able to be compared with other results, since other methods aim to identify maximal or longest repeats and the notion of *elementary repeats* is unique only to this paper's methodology.