NESTED TANDEM REPEAT COMPUTATION AND ANALYSIS

A thesis presented in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

in Computational Biology

at Massey University

Atheer Matroud

2013

Abstract

Biological sequences have long been known to contain many classes of *repeats*. The most studied repetitive structure is the *tandem repeat* where many approximate copies of a common segment (the *motif*) appear consecutively. In this thesis, a complex repetitive structure is investigated. This repetitive structure is called a *nested tandem repeat*. It consists of many approximate copies of two motifs interspersed with one another.

This thesis is a collection of published and in progress papers. Each paper addresses a computational problem related to the analysis of nested tandem repeats. Nested tandem repeats have been observed in the intergenic spacer of the ribosomal DNA gene in *Colocasia esculenta*. The question of whether such repeats can be found elsewhere in biological sequence databases is addressed and `NTRFinder`, a software tool to detect nested tandem repeats, is described. Another problem that arises after detecting a nested tandem repeat is the alignment of the nested tandem repeat region against its two motifs. An algorithm that guarantees an optimal solution to this problem is introduced. After detecting nested tandem repeats and identifying their structures, the identification of the motif boundaries is an unsolved problem which arises not only in nested tandem repeats but in tandem repeats as well. Heuristic solutions to this problem are implemented and tested. In order to compare two tandem repeat sequences an algorithm that aligns a hypothetical ancestral sequence of both sequences against each sequence is presented. This algorithm considers substitutions, deletions, and unidirectional duplication, namely, from ancestor to descendant.

Acknowledgements

I am sincerely thankful to my supervisors Prof. Mike Hendy and Dr. Chris Tuffley for their continuous support from day one through my candidature. It has been a privilege to work with both of them. Without their guidance and patience this work would have not been possible. I am forever grateful for their continuous support. I would also like to thank A/Prof. David Bryant for his support and help in my studies after I moved to Dunedin.

I would like to thank the Allan Wilson Centre for Molecular Ecology and Evolution for funding my study. I would also like to thank the Institute of Fundamental Sciences for the financial support they have provided to me.

I would like to thank Prof. Jens Stoye for the fruitful discussions I had with him during my visit to his lab and during our meeting in Campo Grande. Many thanks to A/Prof. Nadia El-Mabrouk for her support of my visit to her lab and for her advice. I would also like to thank A/Prof. Eric Rivals for his suggestions and comments.

I am grateful to Prof. Hamish Spencer for giving me his office to use. I am also thankful to the Department of Zoology staff for their hospitality. I am thankful to the Mathematics and Statistics department members at the University of Otago for hosting me and being so friendly.

I would like to thank all my colleagues in my office at Massey University and all colleagues in the "boffin lounge" for their company, help and support.

Last but not least, I am very grateful to my wife Zainab for her patience and strong support throughout my study. I am thankful to my children, Qaswar, Jaafar, and Durar for being such a great inspiration to me. I am grateful to my father, my mother, Naseer, Abeer, Akeel, and Aseel for their continued support.

Contents