Explorations In Searching Compressed Nucleic Acid And Protein Sequence Databases And Their Cooperatively-Compressed Indices

Paul Gardner-Stephen

Bachelor of Science,

Flinders University of South Australia

A Thesis Submitted for the Degree of Doctor of Philosophy

Flinders University

School of Informatics and Engineering

Adelaide, South Australia

2008

(Submitted 20 December 2007)

Acknowledgements

Before I acknowledge those who have helped me make it through my candidature, I wish to thank the unbroken chain of people who have believed in me, encouraged me, nurtured and supported me, and without whom I would not have made it to the starting line: Mrs Jan Stephen, my mother, who first kindled my interest in science, and Mr. Keith Gardner, my father; Mr. Peter Brune, a primary school teacher, who believed that I would be someone great one day (perhaps I will get there yet); Dr. Allen Hodson, a high school teacher who taught me more lessons about life and statistics than I suspect he realises; Dr. Todd Rockoff, a university lecturer who kindled my interest in hardware design; Mr. Michael Renz of Germany, who taught me my first lessons about the commercial world; Mr. Murray Rogers, who took a scruffy young university student and made a respectable (but still scruffy) Systems Administrator; and Prof. Greg Knowles, who had enough faith in me to take me on as his student.

For their support throughout my candidature, I wish to thank again: Mr. Murray Rogers for graciously releasing my time so that I could study; and Prof. Greg Knowles for his support and investment as my supervisor. In addition to these people I wish to thank: Prof. Janet Verbyla as head of the School of Informatics & Engineering, and my colleagues (academic staff, general staff and fellow candidates) for their support and friendship during my candidature, and also my friends who have not abandoned me, even though “as busy as a Gardner-Stephen” has become a proverb among them.

A special mention must go to Ms. Fran Banytis of the Staff Development & Training Unit, for the way she has invested much of her self into my candidature. Her encouragement and support throughout my candidature has been as valuable as it has been unexpected.

Finally I would like to thank Dr Dione Gardner-Stephen, my beautiful wife who not only loved me enough to trade in her perfectly good maiden name for one that no-one can spell, but believed in me and supported me through the thick and thin of candidature, not only with the insight and understanding of another doctoral candidate, but in the way that only a loving wife can. I love you much more than words can express.

Through the actions of all these people I see the guiding hand of a God of love, and who has been gracious enough, not merely to bring me to this place, but to invite me to know Him personally. That God is Jesus Christ of Nazareth, and to Him I say “thank you” for the contributions that each of you have made in my life.

Abstract

Nucleic acid and protein databases such as GenBank are growing at a rate that perhaps eclipses even Moore's Law of increase in computational power[1]. This poses a problem for the biological sciences, which have become increasingly dependant on searching and manipulating these databases. It was once reasonably practical to perform exhaustive searches of these databases, for example using the algorithm described by Smith and Waterman, however it has been many years since this was the case. This has led to the development of a series of search algorithms, such as FASTA, BLAST and BLAT, that are each successively faster, but at similarly successive costs in terms of thoroughness.

Attempts have been made to remedy this problem by devising search algorithms that are both fast and thorough. An example is CAFE, which seeks to construct a search system with a sub-linear relationship between search time and database size, and argues that this property must be present for any search system to be successful in the long term.

This dissertation explores this notion by seeking to construct a search system that takes advantage of the growing redundancy in databases such as GenBank in order to reduce both the search time and the space required to store the databases and their indices, while preserving or increasing the thoroughness of the search.

The result is the creation and implementation of new genomic sequence search and alignment, database compression, and index compression algorithms and systems that make

[1] More accurately, Moore's Law predicts that the capacity for transistors on an integrated circuit will double approximately every two years. In practice, due to the efforts of computer architects this has translated into a roughly corresponding increase in computation throughput.

progress toward resolving the problem of reducing search speed and space requirements while improving sensitivity. However, success is tempered by the need for databases with adequate local redundancy, and the computational cost of these algorithms when servicing un-batched queries.

"I Paul Gardner-Stephen, certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text."

Candidate:

Paul Gardner-Stephen

Contents

List of Figures

List of Tables