A GUIDED DYNAMIC PROGRAMMING APPROACH FOR DNA SEQUENCE SIMILARITY SEARCH

MOHD NORDIN BIN ABDUL RAHMAN

DOCTOR OF PHILOSOPHY
UNIVERSITI MALAYSIA TERENGGANU
MALAYSIA

2009

A GUIDED DYNAMIC PROGRAMMING APPROACH FOR DNA SEQUENCE SIMILARITY SEARCH

MOHD NORDIN BIN ABDUL RAHMAN

Thesis Submitted in Fulfillment of the Requirement for the Degree of Doctor of Philosophy in the Faculty of Science and Technology Universiti Malaysia Terengganu

March 2009

Abstract of thesis presented to the Senate of University Malaysia Terengganu in fulfillment of the requirement for the degree of Doctor of Philosophy

A GUIDED DYNAMIC PROGRAMMING APPROACH FOR DNA SEQUENCE SIMILARITY SEARCH

MOHD NORDIN BIN ABDUL RAHMAN

March 2009

Chairperson : Professor Md Yazid bin Mohd Saman, Ph.D.

Members : Associate Professor Aziz bin Ahmad, Ph.D.
Professor Abu Osman bin Md Tap, Ph.D.

Faculty : Science and Technology

DNA sequence similarity search is an important task in computational biology applications. Similarity search procedure is executed by an alignment process between query and targeted sequences. An optimal alignment process based on the dynamic programming algorithms has shown to have O(n × m) time and space complexity. Heuristics algorithms can process a fast DNA sequence alignment, but generate low comparison sensitivity. The biologists frequently demand for optimal comparison result so that the perfect structure of living beings evolution can be constructed. This task becomes more complex and challenging as the sizes of public sequence databases get very large and are increasing exponentially each year. The aim of this study is to develop

a guided dynamic programming approach for an efficient DNA sequence similarity search.

Incremental research and development paradigm has been employed in developing of the model. The model developed consists of three main processes: (i) filtering, (ii) classification and reduction and (iii) parallelization. The main goal of the first two processes is to reduce the size of DNA sequence dataset and subsequently minimized the iterations of dynamic programming algorithm. Parallel processing technique in the third process is used to increase the efficiency of all processes in the model. Five supporting elements which are BioJava, public sequence database, rough sets theory, string matching algorithms and parallel computing technique have been used to accomplish the model.

The filtering process involved the automaton based exact string matching technique and named as F-R-A model. The time and space complexity of this technique is O(n) for preprocessing and O(m + k) for searching process is very effective for DNA sequence exact matching. The filtering process has significantly discarded irrelevant DNA sequences in the database from being computed by the dynamic programming algorithm. The experimental results show that more than 80% of dynamic programming algorithm iterations have been minimized.

Rough sets theory has been used to enhance the process in the F-R-A model. The theory provides an indiscernibility relation technique which is used to classify and reduct the database according to the definition of 'equivalence'. The indiscernibility relation technique removes the superfluous DNA sequences from the dataset. Only the DNA sequences in the reduct set are considered for an alignment process. The similarity search results produced by this reduct set will represent the whole DNA sequences in all equivalence classes defined. The experimental results demonstrate that the F-R-A model is improved up to 4 to 9%.

Finally, the improved F-R-A model has been reengineered to be executed under a parallel processing environment. A PC-based cluster system is used to implement the parallel version. The system architecture consists of eight PCs connected with an Ethernet network under a master-worker paradigm. The MPJ Express software is utilized as a communication interface protocol between the machines. The experimental results show that the parallel F-R-A model has achieved the reasonable speedups factor of and performance efficiency.

Abstrak tesis yang dikemukakan kepada Senat Universiti Malaysia Terengganu sebagai memenuhi keperluan untuk ijazah Kedoktoran Falsafah

PENDEKATAN PENGATURCARAAN DINAMIK TERPANDU UNTUK GELINTARAN KESAMAAN JUJUKAN DNA

MOHD NORDIN BIN ABDUL RAHMAN

Mac 2009

Pengerusi : Professor Md Yazid bin Mohd Saman, Ph.D.

Ahli : Professor Madya Aziz bin Ahmad, Ph.D.
Professor Abu Osman bin Md Tap, Ph.D.

Fakulti : Sains dan Teknologi

Gelintaran kesamaan jujukan DNA merupakan satu operasi penting di dalam aplikasi komputeran biologi. Tatacara bagi gelintaran kesamaan ini dilaksanakan melalui proses penjajaran antara jujukan pertanyaan dan jujukan sasaran. Proses penjajaran optimum berasaskan alkhwarizmi pengaturcaraan dinamik telah menunjukkan kekompleksan masa dan ruang O(n × m). Alkhwarizmi heuristik boleh memproses penjajaran jujukan DNA dengan pantas, tetapi menjanakan sensitiviti perbandingan yang rendah. Ahli biologi selalunya memerlukan keputusan perbandingan yang optimum dengan itu struktur evolusi benda-benda hidup yang sempurna dapat dibina. Proses ini menjadi semakin sukar dan mencabar dengan saiz pangkalan data jujukan sangat besar dan meningkat secara eksponen setiap tahun.

Matlamat kajian ini adalah untuk membangunkan satu pendekatan pengaturcaraan dinamik terpandu untuk gelintaran kesamaan jujukan DNA yang efisien.

Paradigma penyelidikan dan pembangunan peningkatan telah digunakan bagi membangunkan model ini. Model yang dibangunkan ini mengandungi tiga proses utama iaitu: (i) penapisan (ii) pengkelasan dan pengurangan (iii) pemprosesan selari. Matlamat utama bagi dua proses yang pertama adalah untuk mengurangkan saiz set data jujukan DNA dan seterusnya lelaran pengaturcaraan dinamik dapat diminimumkan. Dalam proses yang ketiga, teknik pemprosesan selari digunakan bagi meningkatkan keefisienan keseluruhan proses di dalam model. Lima elemen sokongan telah digunakan iaitu BioJava, pangkalan data jujukan awam, teori set kasar, alkhwarizmi padanan rentetan dan teknik pemprosesan selari bagi menyempurnakan keseluruhan model.

Proses penapisan melibatkan teknik padanan tepat rentetan aksara berasaskan automata dan dinamakan sebagai F-R-A model. Dengan kekompleksan masa dan ruang O(n) untuk prapemprosesan dan O(m + k) bagi proses gelintaran maka ianya sangat efektif bagi padanan tepat jujukan DNA. Proses penapisan ini dapat menyingkirkan jujukan DNA di dalam pangkalan data yang tidak relevan daripada diproses oleh

alkhwarizmi pengaturcaraan dinamik. Hasil eksperimen menunjukkan lebih 80% lelaran alkhwarizmi pengaturcaraan dinamik dapat diminimumkan. Teori set kasar telah digunakan untuk menambahbaik proses di dalam model F-R-A. Teori ini menyediakan satu teknik hubungan ketidakbolehbeza yang mana ianya digunakan bagi tujuan pengkelasan dan pengurangan pangkalan data berasaskan takrif 'kesamaan'. Teknik hubungan ketidakbolehbeza dapat mengeluarkan jujukan DNA yang berlebihan daripada set data. Hanya jujukan DNA yang berada di dalam set pengurangan akan dipertimbangkan di dalam proses penjajaran. Hasil gelintaran kesamaan daripada set pengurangan ini mewakili keseluruhan jujukan DNA di dalam kelas-kelas 'kesamaan' yang telah ditakrifkan. Hasil eksperimen menunjukkan prestasi model F-R-A dapat dipertingkatkan kepada 4 hingga 9%.

Akhirnya, model F-R-A yang ditambahbaik ini telah diperikayasaankan untuk dilaksanakan pada persekitaran pemprosesan selari. Satu sistem kluster berasaskan komputer peribadi digunakan bagi melaksanakan model F-R-A selari. Senibina sistem mengandungi lapan buah komputer peribadi yang dirangkaikan dengan rangkaian *Ethernet* di bawah paradigma *master-worker*. Perisian *MPJ Express* digunakan sebagai protokol antaramuka komunikasi antara mesin. Hasil

eksperimen menunjukkan model F-R-A selari ini dapat mencapai pencepatan dan kecekapan prestasi yang munasabah.