A COMPARATIVE STUDY OF TWO ORTHOLOGOUS GENE IDENTIFICATION METHODS ON SYNTENY BLOCK INFERENCE

CHOW KINGSLEY

(SGJ100006)

SUBMITTED TO
INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA

IN PARTIAL FULFILMENT

OF THE REQUIREMENTS FOR

THE DEGREE OF MASTER OF BIOINFORMATICS

2012

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: CHOW KINGSLEY (I.C/Passport No: 850118-14-6277)

Registration/Matric No: SGJ100006

Name of Degree: MASTER OF BIOINFORMATICS

A Comparative Study of Two Orthologous Gene Identification Methods On Synteny Block Inference ("this Work"):

Field of Study: Bioinformatics, Synteny, Orthologs, Inparalogs

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any Copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:

Designation:

ABSTRACT

A synteny block is a set of orthologous genes that share the same relative ordering on the chromosomes of two species. Synteny analysis at the genome scale is a powerful means of identifying orthologs in a set of genomes of interest for downstream phylogenetic studies. OrthoCluster is a data mining tool for inferring synteny blocks among multiple genomes. Before using OrthoCluster to infer synteny blocks, orthologous gene relationships between the species of interest have to be identified first. In this study, we evaluated the effects of two different orthologous gene identification methods: InParanoid and ad hoc BLAST, on the number, size and content of synteny blocks returned by OrthoCluster using the genomes of *Oryza sativa* and *Arabidopsis thaliana*.

Results show that InParanoid identified 22 124 orthologous relationships while ad hoc BLAST identified 14 928. Subsequently, OrthoCluster identified 942 conserved synteny blocks that contain no mismatches using input from InParanoid. These synteny blocks cover 1234 genes (5.97 Mb) in *O. sativa* and 1403 genes (2.76 Mb) in *A. thaliana*, respectively. With input from ad hoc BLAST, OrthoCluster detected just 314 conserved synteny blocks, which cover 427 genes (2.3 Mb) in *O. sativa* and 435 genes (1.1 Mb) in *A. thaliana*. Allowing mismatches within a synteny block, OrthoCluster identified 1510 non-conserved synteny blocks from InParanoid input, which cover 3509 genes (25.1 Mb) in *O. sativa* and 3648 genes (9.06 Mb) in *A. thaliana*. Only 589 non-conserved synteny blocks were detected using ad hoc BLAST input, with 1335 genes (8.22 Mb) in *O. sativa* and 1257 genes (3.32 Mb) in *A. thaliana*.

InParanoid identified about 50% more orthologous genes compared to ad hoc BLAST. This subsequently led to OrthoCluster detecting at least 2 times more synteny blocks (conserved / non-conserved), and about 3 times more genes. This result suggests that synteny blocks inferred by OrthoCluster are highly dependent on the method and parameters used for identifying orthologous gene relationships.

ABSTRAK

Blok sinteni ialah satu set gen ortolog yang berkongsi arahan relatif yang sama antara kromosom dua species. Analisis sinteni pada genom merupakan satu cara yang ampuh untuk mengenal pasti ortolog dalam satu set genom untuk kajian filogenetik seterusnya. *OrthoCluster* ialah satu perisian untuk mengesan blok sinteni antara genom-genom. Sebelum menggunakan *OrthoCluster* untuk mengesan blok sinteni, hubungan ortolog gen antara pelbagai species yang berkaitan perlu dikenalpasti terlebih dahulu. Dalam kajian ini, kami menilai perbezaan antara dua kaedah pengenalan gen ortolog: *InParanoid* dan *ad hoc* BLAST, dari segi bilangan, saiz dan kandungan dalam blok sinteni yang dikenalpasti oleh *OrthoCluster*, dengan menggunakan genom *Oryza sativa* dan *Arabidopsis thaliana*.

Hasil menunjukkan *InParanoid* mengenal pasti 22 124 hubungan ortolog manakala *ad hoc* BLAST mengenal pasti 14 928 hubungan ortolog. Seterusnya, *OrthoCluster* mengenal pasti 942 blok sinteni abadi yang tidak mengandungi ketakpadanan dengan menggunakan input daripada *InParanoid*. Blok sinteni yang dikenalpasti masing-masing mengandungi 1234 gen (5.97 Mb) pada *O. sativa* dan 1403 gen (2.76 Mb) pada *A. thaliana*. Dengan input daripada *ad hoc* BLAST, *OrthoCluster* mengesan hanya 314 blok sinteni abadi yang masing-masing mengandungi 427 gen (2.3 Mb) pada *O. sativa* dan 435 gen (1.1 Mb) pada *A. thaliana*. Dengan membenarkan ketakpadanan dalam blok sinteni, *OrthoCluster* mengenal pasti 1510 blok sinteni tak abadi daripada *InParanoid,* meliputi 3509 gen (25.1 Mb) pada *O. sativa* dan 3648 gen (9.06 Mb) pada *A. thaliana*. Hanya 589 blok sinteni tak abadi dikesan dengan menggunakan input daripada *ad hoc* BLAST, meliputi 1335 gen (8.22 Mb) pada *O. sativa* dan 1257 gen (3.32 Mb) pada *A. thaliana*.

InParanoid mengenal pasti kira-kira 50% lebih gen ortolog berbanding dengan *ad hoc* BLAST. Dengan demikian, *OrthoCluster* mengesan sekurang-kurangnya 2 kali lebih banyak blok sinteni (abadi / tak abadi), dan kira-kira 3 kali lebih banyak gen. Keputusan ini menunjukkan bahawa blok sinteni yang dikenalpasti olen *OrthoCluster* sangat bergantung kepada kaedah dan parameter yang digunakan untuk mengenal pasti hubungan gen ortolog.

ACKNOWLEDGEMENT

I wish to express my gratitude and appreciation to my supervisor, Dr. Khang Tsung Fei, for his invaluable guidance and supervision on science knowledge, personal skills and excellent thinking throughout of this project. I would like to thank the Ministry of Higher Education of Malaysia for financing my study through the MyBrain15 programme. My sincere thanks to my entire course mates for their guidance supports, assistance and co-operation in this project. Finally, I would like to thank my family members and friends for their moral support, tolerance and personal encouragement throughout these years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AL	-	aligned length
BBH	-	bidirectional best-hit
BLAST	-	Basic Local Alignment Search Tool
CALIP	-	cumulative alignment length identity percentage
CALP	-	cumulative alignment length percentage
COG	-	Cluster of Orthologous Group
eggNOG	-	evolutionary genealogy of genes: Non-supervised Orthologous Groups
HSP	-	high-scoring segment pairs
IRGSP	-	International Rice Genome Sequencing Project
kb	-	kilobase
KOG	-	cluster of euKaryotic Orthologous Group
Mb	-	megabase
OMA	-	ortholog matrix project
RAP	-	Rice Annotation Project
RSD	-	reciprocal smallest distance
ssp	-	subspecies
TAIR	-	The Arabidopsis Information Resource