

**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 oleh *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

TITLE PAGE .....	i
ORIGINAL LITERARY WORK DECLARATION.....	ii
ABSTRACT.....	iii
ABSTRAK.....	v
ACKNOWLEDGEMENT .....	vii
TABLE OF CONTENTS.....	viii
LIST OF FIGURES .....	x
LIST OF TABLES .....	xvi
LIST OF ABBREVIATIONS.....	xix
<b>CHAPTER 1 INTRODUCTION AND LITERATURE REVIEW .....</b>	<b>1</b>
1.1 Synteny Analysis.....	1
1.2 Orthologous Gene Identification.....	3
1.3 Objectives of Study .....	7
<b>CHAPTER 2 METHODOLOGY .....</b>	<b>8</b>
2.1 Identification of Orthologous Genes by Ad hoc BLAST.....	9
2.2 Identification of Orthologous Genes by InParanoid.....	10
2.3 Identification of Synteny Blocks by Orthocluster.....	12



2.4	Visualisation by Circos .....	13
<b>CHAPTER 3</b>	<b>RESULTS .....</b>	<b>15</b>
3.1	Orthologous Relationships .....	15
3.2	Conserved Synteny Blocks.....	18
3.3	Non-conserved Synteny Blocks .....	25
3.4	Visualisation.....	33
<b>CHAPTER 4</b>	<b>DISCUSSION .....</b>	<b>39</b>
<b>CHAPTER 5</b>	<b>CONCLUSIONS .....</b>	<b>45</b>
<b>REFERENCES</b> .....		<b>46</b>
<b>APPENDIX</b> .....		<b>49</b>

## LIST OF FIGURES

- Figure 1.1:** In this hypothetical gene tree, Gene A in an ancestral species ‘A’ undergoes a gene duplication event giving rise to A1 and A2 genes. After that, speciation event occurs leading to two lineages ‘B’ and ‘C’. The genes C2 and C3 in the C genome are inparalogs since their gene duplication occurred after speciation and they are orthologous to the B2 gene because they share a common ancestral gene A2. B1 is an outparalog of the B2, C2 and C3 genes. .... 4
- Figure 2.1:** Clustering of additional orthologs (inparalogs). Each circle represents sequence from species B or species C. B2 and C2 are the original seed-ortholog pair with an inparalog score of 1.0 (all inparalogs are clustered around this pair). Other inparalogs, C3, are scored according to the relative similarity to the seed-inparalog, C2. The score is the reverse distance between pairwise comparison of sequences, in this case, C2 is relatively more similar to B2 than C3, thus C3 receives a lower inparalog score (0.7). Sequences outside the circle are classified as outparalogs; thus C1 and B1 form a cluster of their own. C1 and B1 are orthologous to each other. .... 12
- Figure 3.1:** Heatplot of one-to-one orthologous relationships between *O. sativa* (row) and *A. thaliana* (column) from InParanoid and ad hoc BLAST. The colours of each block represent the number of orthologous relationships between the chromosome of *O sativa* and *A. thaliana*. .... 17
- Figure 3.2:** Venn diagram for the three types of orthologous relationships: orthologous relationships shared by both methods, InParanoid-specific and ad hoc BLAST-specific orthologous relationships. .... 18

**Figure 3.3:** Distribution of conserved synteny block numbers for *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Barplots on top panel are distribution of conserved synteny block numbers identify from orthologous gene datasets of InParanoid. Barplots on bottom panel are distribution of conserved synteny block numbers identify from orthologous gene datasets of ad hoc BLAST. The height of the bar represents the number of conserved synteny blocks for each chromosome..... 20

**Figure 3.4:** Distribution of conserved synteny blocks size in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Boxplots on left panel are distribution of conserved synteny blocks sizes identify from orthologous gene datasets of InParanoid. Boxplots on right panel are distribution of conserved synteny blocks sizes identified from orthologous gene datasets of ad hoc BLAST. The y-axis is plotted in log-scale of base 10. .... 21

**Figure 3.5:** Distribution of gene numbers of each conserved synteny block in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Barplots on top panel are distribution of gene numbers of each conserved synteny block identify from orthologous gene datasets of InParanoid. Barplots on bottom panel are distribution of gene numbers of each conserved synteny block identify from orthologous gene datasets of ad hoc BLAST. The colours of the bar represent the number of genes of the conserved synteny block. The height of the bar represents the number of conserved synteny blocks with specified number of genes within the block..... 22

**Figure 3.6:** Coverage of conserved synteny blocks on each chromosome in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc

BLAST method. Scatterplots on top panel are coverage of conserved synteny blocks on each chromosome identify from orthologous gene datasets of InParanoid. Scatterplots on bottom panel are coverage of conserved synteny blocks on each chromosome identify from orthologous gene datasets of ad hoc BLAST. The numbers in the scatterplot represent the number of chromosome of each species. The genome sizes of each chromosome were plotted on x-axis while the coverage of conserved synteny blocks on each chromosome was plotted on y-axis..... 23

**Figure 3.7:** Distribution of conserved synteny blocks between *O. sativa* (row) and *A. thaliana* (column) identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. The colours of the each block represent the number of conserved synteny blockd between the chromosome of *O sativa* and *A. thaliana*. ..... 24

**Figure 3.8:** Distribution of non-conserved synteny block numbers for *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Barplots on top panel are distribution of non-conserved synteny block numbers identify from orthologous gene datasets of InParanoid. Barplots on bottom panel are distribution of non-conserved synteny block numbers identify from orthologous gene datasets of ad hoc BLAST. The height of the bar represents the number of conserved synteny blocks for each chromosome..... 28

**Figure 3.9:** Distribution of non-conserved (50% mismatches) synteny blocks size in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Boxplots on right panel are distribution of non-conserved synteny blocks sizes identify from orthologous gene datasets of

InParanoid. Boxplots on left panel are distribution of non-conserved synteny blocks sizes identified from orthologous gene datasets of ad hoc BLAST. Result for 10% mismatches was excluded because the result was same with conserved synteny block. The y-axis is plotted in log-scale of base 10. .... 29

**Figure 3.10:** Distribution of gene numbers of each non-conserved (50% mismatches) synteny block in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Barplots on top panel are distribution of gene numbers of each non-conserved synteny block identify from orthologous gene datasets of InParanoid. Barplots on bottom panel are distribution of gene numbers of each non-conserved synteny block identify from orthologous gene datasets of ad hoc BLAST. The colours of the bar represent the number of genes of the non-conserved synteny block. The height of the bar represents the number of conserved synteny blocks with specified number of genes within the block. Result for 10% mismatches was excluded because the result was same with conserved synteny block. .... 30

**Figure 3.11:** Coverage of non-conserved synteny blocks on each chromosome in *O. sativa* and *A. thaliana* identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. Scatterplots on top panel are coverage of non-conserved synteny blocks on each chromosome identify from orthologous gene datasets of InParanoid. Scatterplots on bottom panel are coverage of non-conserved synteny blocks on each chromosome identify from orthologous gene datasets of ad hoc BLAST. The numbers in the scatterplot represent the number of chromosome of each species. The genome sizes of each chromosome were plotted on x-axis while the coverage of non-conserved synteny blocks on each chromosome was plotted on y-axis. .... 31

<b>Figure 3.12:</b> Distribution of non-conserved synteny blocks between <i>O. sativa</i> (row) and <i>A. thaliana</i> (column) identify from orthologous gene datasets of InParanoid and ad hoc BLAST method. The colours of the each block represent the number of non-conserved synteny blockd between the chromosome of <i>O sativa</i> and <i>A. thaliana</i> . .....	32
<b>Figure 3.13:</b> Circos image of conserved synteny block identified from orthologous gene datasets of ad hoc BLAST.....	35
<b>Figure 3.14:</b> Circos image of conserved synteny block identified from orthologous gene datasets of InParanoid. ....	36
<b>Figure 3.15:</b> Circos image of non-conserved synteny block identified from orthologous gene datasets of ad hoc BLAST.....	37
<b>Figure 3.16:</b> Circos image of non-conserved synteny block identified from orthologous gene datasets of InParanoid.....	38
<b>Figure 4.1:</b> Example of orthologous relationships identified by InParanoid (solid lines) and ad hoc BLAST (dotted lines) when highly similar genes are involved.....	40
<b>Figure 4.2:</b> Example shows the orthologous relationships identified by InParanoid (solid lines) and ad hoc BLAST (dotted lines). Filtering criteria has cause ad hoc BLAST identified less number of orthologous relationships.....	41
<b>Figure 4.3:</b> Example of orthologous relationships identified by ad hoc BLAST (dotted lines) and InParanoid (solid lines). InParanoid identified a lot more orthologous relationships than ad hoc BLAST. ....	44

**Figure 7.1:** In this example, blocks A1 in genome 1 and A2 in genome 2 are composed of four genes. The order of the genes in each block is the same, and each pair of genes has the same orientation. .... 49

**Figure 7.2:** In this example, blocks A1 in genome 1 and A2 in genome 2 are composed of four genes. The order of the genes in each block is the same, but each pair of genes has different orientation. .... 49

**Figure 7.3:** In this example, blocks A1 in genome 1 and A2 in genome 2 are composed of four genes. The order of the genes in block A1 is inverted with respect to that in block A2, and each pair of genes has the same orientation. .... 49

**Figure 7.4:** In this example, blocks A1 in genome 1 and A2 in genome 2 are composed of four genes. The order of the genes in block A1 is inverted with respect to that in block A2, and each pair of genes has different orientation..... 50

**Figure 7.5:** Blocks conformed of the 7 ortholog genes with no in-map mismatches and no out-map mismatches..... 50

**Figure 7.6:** Blocks conformed of 6 ortholog genes. For block A1, one in-map mismatch (g3) and one out-map mismatch exists..... 50

## LIST OF TABLES

<b>Table 3.1:</b> Orthologous relationships identified from ad hoc BLAST and InParanoid.....	
.....	15
<b>Table 3.2:</b> Number of orthologous protein identify by InParanoid method with different confidence value and above the cutoff. Inparalogs with confidence value less than 0.05% was not shown by default setting of InParanoid. ....	16
<b>Table 3.3:</b> Comparison of synteny block returned from OrthoCluster for the orthologous genes dataset identified from InParanoid and ad hoc BLAST. Common block is the block identify by both orthologous gene identification method. There are some blocks in the same genome position from both methods but block size is larger on either one of the method (larger block size in InParanoid or ad hoc BLAST). Some blocks are method-specific block which only appear in either InParanoid or ad hoc BLAST. Result for 10% mismatches was excluded because the result was same with conserved synteny block. ....	33
<b>Table 7.1:</b> One-to-one orthologous relationships between <i>O. sativa</i> (row) and <i>A. thaliana</i> (column) from InParanoid and ad hoc BLAST. Numbers in parenthesis represent one-to-one orthologous relationships result from ad hoc BLAST method.....	51
<b>Table 7.2:</b> Conserved synteny blocks and their corresponding genomic coverage, size and range in <i>O. sativa</i> identified from orthologous gene datasets of InParanoid and ad hoc BLAST method. Numbers in parenthesis represent conserved synteny blocks result from ad hoc BLAST method. ....	52



**Table 7.3:** Conserved synteny blocks and their corresponding genomic coverage, size and range in *A. thaliana* identified from orthologous gene datasets of InParanoid and ad hoc BLAST method. Numbers in parenthesis represent conserved synteny blocks result from ad hoc BLAST method. .... 53

**Table 7.4:** Distribution of perfect synteny blocks between *O. sativa* chromosomes (rows) and *A. thaliana* chromosomes (columns) identified from orthologous gene datasets of InParanoid and ad hoc BLAST method. Numbers in parenthesis represent conserved synteny blocks result from ad hoc BLAST method. Mitochondria chromosome (ChrM) of *A. thaliana* was excluded because it doesn't have any synteny blocks with *O. sativa*. .... 54

**Table 7.5:** Non-conserved synteny blocks for mismatches (10% / 50%) and their corresponding genomic coverage, size and range in *O. sativa* identified from orthologous gene datasets of InParanoid and ad hoc BLAST method. Numbers in parenthesis represent conserved synteny blocks result from ad hoc BLAST method. Result for 10% mismatches was excluded because it was same with the result of conserved synteny block. .... 55

**Table 7.6:** Non-conserved synteny blocks for mismatches (50%) and their corresponding genomic coverage, size and range in *A. thaliana* identified from orthologous gene datasets of InParanoid and ad hoc BLAST method. Numbers in parenthesis represent conserved synteny blocks result from ad hoc BLAST method. Result for 10% mismatches was excluded because it was same with the result of conserved synteny block. .... 57

**Table 7.7:** Distribution of non-conserved synteny blocks for mismatches (10% / 50%) between *O. sativa* chromosomes (rows) and *A. thaliana* chromosomes (columns) identified from orthologous gene datasets of InParanoid and ad hoc BLAST method. Numbers in parenthesis represent conserved synteny blocks result from ad hoc BLAST method. Mitochondria chromosome (ChrM) of *A. thaliana* was excluded because it doesn't have any synteny blocks with *O. sativa*. Result for 10% mismatches was excluded because it was similar with the result of conserved synteny block. .... 58

## 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