Form 3

Abstract of Thesis

Name (Kai WANG)	
Title	Development of a transcription factor database useful for developmental biology, and decoding the maternal and zygotic transcriptome of the appendicularian, *Oikopleura dioica* （発生生物学研究のための転写因子データベースの開発とオタマボヤを用いた母性と胚性のRNA-Seq解析）

Genes encoding transcription factors that constitute gene-regulatory networks and maternal factors accumulating in egg cytoplasm are two classes of essential genes that play crucial roles in developmental processes. Transcription factors control the expression of their downstream target genes by interacting with *cis*-regulatory elements. Maternal factors initiate embryonic developmental programs by regulating the expression of zygotic genes and various other events during early embryogenesis. This manuscript documents the transcription factors of 77 metazoan species as well as human and mouse maternal factors. I improved the previous method for prediction of transcription factors using a statistical approach adding Gene Ontology information to Pfam based identification of transcription factors. This method detects previously un-discovered transcription factors. The novel features of this database are:

1) It includes both transcription factors and maternal factors, although the number of species, in which maternal factors are listed, is limited at the moment.
2) Ontological representation at the cell, tissue, organ, and system levels has been specially designed to facilitate developmental studies. This is the unique feature in our database and is not available in other transcription factor databases.

I developed, a user-friendly web interface, REGULATOR (http://www.bioinformatics.org/regulator/), which can help researchers to efficiently identify, validate, and visualize the data in the database. Using this web interface, users can browse, search, and download detailed information on species of interest, genes, transcription factor families, or developmental ontology terms.

In the second part of this thesis, I performed transcriptome analysis for the appendicularian, *Oikopleura dioica* (*O. dioica*), which is a planktonic chordate. It has been used as a novel model species in development biology and evolutionary studies. The most significant character of this species is rapid development and short life cycle (only 5 days). Recently, many bioinformatics resources are publicly available, such as the genome sequence and microarray data in the OikoBase. However, transcriptome information is still not complete, especially as to the next-generation sequencing data. In this study, we carried out transcriptome analysis using RNA sequencing data of a Japanese population collected from egg and larval stages. The major findings of this study are:

1. Via mapping our reads (Japanese population) to the reference genome sequence deposited in the OikoBase (Norwegian population), an extreme low reads mapping rate were found, which due to the significant sequence variations between the Japanese population and the Norwegian population.
2. After *de novo* transcriptome assembly, 16,423 proteins (belongs to 12,136 known unigenes) of Japanese population have homologies with that of the Norwegian population (OikoBase). These proteins corresponding to 95.4% of the protein-encoding genes deposited in OikoBase.
3. Via comparing the 4,136 one-vs-one bidirectional best hits (BBH) at both the nucleotide and protein levels, the sequence similarity between Japanese and Norwegian population was estimated to be 91.0% in nucleotide level and 94.8% in amino acid level.
4. Additional 175 novel protein-encoding genes were found. Among these novel genes, 144 of them were not predicted in the gene models deposited in OikoBase, but they can be found in the Norwegian reference genome; whereas 31 unigenes were not found in the OikoBase reference genome due to the low sequence similarity.
5. I found approximately 63% unigenes were expressed in egg-stage, whereas 99% were observed in larval-stage; Analysis of differentially expressed genes (DEG) using a fold change threshold of four for the total 12,311

unigenes, we found 3,772 of them were up-regulated and 1,336 were down-regulated. Gene ontology (GO) analyses showed distinct gene activities in these two developmental stages.

6. The previously reported mRNA 5' *trans*-spliced leader was also detected by our method. While, it was observed in 40.8% of the total unique transcripts, which is much more frequently happed than the previous estimation. This *trans*-spliced leader showed preferential linkage to adenine at the 5' ends of the downstream exons.
7. By comparing the *trans*-spliced mRNAs between egg and larval stage, as well as the down-regulated and up-regulated groups, we found *trans*-spliced mRNAs were more frequently observed in egg compared to larva.

Our data of transcriptome assembly will provide an additional resource for studies of Japanese population. These findings would be useful for better understanding of the development as well as evolutionary history of *O. dioica*.

様式 7

論文審査の結果の要旨及び担当者

氏　名	(Kai Wang)	
	(職)	氏　名
論文審査担当者	主 査　教 授	西田　宏記
	副 査　教 授	安永 照雄
	副 査　教 授	蘇 智慧

論文審査の結果の要旨

本論文は二部構成になっており、第一部では、７７種の動物の公開ゲノム配列から転写因子をコードしている遺伝子を予測し、それを元に発生生物学的研究に有用と考えられる転写因子のデータベースを作成したことを報告している。第二部では脊索動物に属するワカレオタマボヤを用いて RNA-seq を行い、卵と幼生で発現している遺伝子の比較及び特徴付けを行ったことを報告している。申請者は動物の胚発生をより深く理解することを目的として本研究を行った。

生物の胚発生では、時間軸に沿って様々な遺伝子が発現するが、その発現は DNA 結合能を持った転写因子と呼ばれる蛋白質によって制御されている。Wang 氏は、７７種の動物の公開ゲノム配列から独自に改良したバイオインフォーマティクス的方法を用いて転写因子をコードしている遺伝子を予測し、それをデータベース（REGULATOR 、http://www.bioinformatics.org/regulator/）として公開した。このデータベースは、発生生物学的研究に有用であると期待できる。このデータベースは、発生学者が利用しやすいように工夫がなされており、また、様々の情報とのリンクが張られている。この成果は、BMC Bioinformatics 誌に論文として出版された。

次に、Wang 氏はそのバイオインフォーマティクスに長けた能力を活かし、脊索動物に属するワカレオタマボヤ（*Oikkopleura dioica*）の胚発生の研究を行った。次世代 DNA シークエンサーにより RNA-seq を行い、卵と幼生で発現している遺伝子を網羅的に検出し、その発現量を解析した。その結果、ノルウェーと日本のワカレオタマボヤの間には、ある程度の DNA 配列の違いがあること、卵と幼生期の両方をあわせると全遺伝子の 95%程度が発現していることがわかった。また、これまでに予測されていなかった遺伝子が 175 個存在することや、Trans-splice leader 配列が 40%近くの RNA のついており、それはもっぱら卵に存在する母性 RNA についていることなどがわかった。これらの成果は、Development Genes and Evolution 誌に論文として出版された。

以上のように申請者は、バイオインフォーマティクスを用いた大規模データの解析に熟達しており、優れた能力を持っているといえる。その能力を駆使して行われた上記の研究の新奇性は極めて高いと考えられた。

よって、本論文は博士（理学）の学位論文として十分価値あるものと認める。