قسم علم الحشرات
كلية العلوم

اتجاهات المعلوماتية الحياتية والبيولوجية الحسابية في تعريف الأحماض النووية الريبوزية الدقيقة فى الحشرات

رسالة مقدمة لقسم علم الحشرات- كلية العلوم-جامعة عين شمس
كجزء متمم للحصول على درجة الماجستير في العلوم (علم الحشرات)

لـ م

منى جابر عبد العزيز محمود شعلان

بكالوريوس في العلوم (علم الحشرات)
معيدة بقسم علم الحشرات- كلية العلوم- جامعة عين شمس

تحت إشراف

أ.د. مجدي جبريل شحاتة
أستاذ علم الحشرات الطبية
قسم علم الحشرات- كلية العلوم
جامعة عين شمس

د. عماد إبراهيم خاطر
مدرس علم الحشرات
قسم علم الحشرات- كلية العلوم
جامعة عين شمس

آ. ياسر محمد عبد اللطيف
مدرس علوم الحاسب- قسم علم الرياضيات
كلية العلوم- جامعة عين شمس

وتتوافق هذه الخطة المقترحة مع الاستراتيجية البحثية للقسم وخاصة بناء قاعدة معلومات وإنشاء بنك للحمض النووي للحشرات و تطبيقاته الواسعة في دراسة الحشرات وبرامج مكافحة الآفات

القاهرة ٢٠١٠

Department of Entomology
Faculty of Science
Ain Shams University

Bioinformatics and Computational Biological Approaches to Identify Insect microRNAs

A thesis submitted to the Department of Entomology,
Faculty of Science, Ain Shams University

In partial fulfillment of the requirements for the award of the
M.Sc. degree in Entomology

By

Mona Gaber Abd-El-Aziz Mahmoud Shalaan

(B.Sc., Entomology)

Demonstrator, Department of Entomology,
Faculty of Science, Ain Shams University

SUPERVISORS

Prof. Magdi Gebril Shehata

Professor of Medical Entomology
Faculty of Science
Ain Shams University

Dr. Emad Ibrahim Khater

Lecturer of Entomology
Faculty of Science
Ain Shams University

Dr. Yasser Mohammed Abdel-Lateef

Lecturer of Computer Science, Department of Mathematics
Faculty of Science, Ain Shams University

This thesis is in fulfillment of the Department of Entomology research strategy, to establish entomologic databases and a DNA bank with their enormous applications in insect research and pest control programs

Cairo, 2010

Biography

Name: Mona Gaber Abd-El-Aziz Mahmoud Shalaan.

Date and place of birth: October 18th, 1984, Cairo, Egypt.

Degree awarded: B.Sc. (Entomology)

Department: Entomology

Faculty: Science

University: Ain Shams

Date of Graduation: June, 2006.

Occupation: Demonstrator in Department of Entomology, Faculty of Science, Ain Shams University.

Date of registration for M.Sc. Award: April 14th, 2008.

Publications:

Mona G. Shalaan, Enas H. Ghallab, Yasser M. Abdel-Latif, Magdi G. Shehata and Emad I. Khater.2010.Development of a local entomological database for education and research using simulation (virtual) methods. (submitted)

APPROVAL SHEET

Title: Bioinformatics and Computational Biological Approaches to Identify Insect microRNAs.

Name of candidate: Mona Gaber Abd-El-Aziz Mahmoud Shalaan

Submitted to: Department of Entomology, Faculty of Science, Ain Shams University.

Supervisors:

Prof. Dr. Magdi Gebril Shehata, Professor of Medical Entomology, Faculty of Science, Ain Shams University.

Dr. Emad Ibrahim Khater, Lecturer of Entomology, Faculty of Science, Ain Shams University.

Dr. Yasser Mohammed Abdel-Lateef, Lecturer of Computer Science, Department of Mathematics, Faculty of Science, Ain Shams University.

Head of Entomology Department:

Prof. Dr. Akila El Shafei,

Faculty of Science,

Ain Shams University

Acknowledgment

I thank **ALLAH** WHO showered me with enormous blessings, the knowledge, power and everything I have to accomplish this piece of work.

I would like to express my deepest thanks and appreciation to **Prof. Magdi G. Shehata**, professor of medical entomology, Faculty of Science, Ain Shams University, for his scientific guidance, direct supervision and his kind encouragement throughout my study.

Great heartfelt appreciation to **Dr. Emad Ibrahim M. Khater**, lecturer of medical entomology, Faculty of Science, Ain Shams University for suggesting the research point and training on all aspects of the study. His usual scientific advice, great supervision and kind moral support have inspired and guided me from the initial phase of this thesis and throughout my work. His valuable efforts made possible the achievement of this work.

Warmest thankfulness goes to **Dr. Yasser M. Abdel-Latif**, lecturer of computer sciences, Mathematics Department, Faculty of Science, Ain Shams University, for his guidance and his greatest efforts to make this work real, starting from understanding different mathematical aspects of molecular biological processes and building a customized entomological database.

Special thanks to **Dr. George. F. Mayhew**, the Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin, USA, for providing the text editor (GVim) programme to access and read the sequence files of the mosquito *Culex p. quinquefasciatus* genome.

Last but not least, I am indebted to **Prof. Akila El-Shafei**, head of the Department of Entomology, Faculty of Science, Ain Shams University and every one my professors, colleagues and friends who supported me to accomplish this work.

LIST OF ABBREVIATIONS

aga: *Anopheles gambiae*

ame: *Apis mellifera*

bmo: *Bombyx mori*

Cx.p.quinquefasciatus: *Culex pipiens quinquefasciatus*

Cscore: conservation score

dme: *Drosophila melanogaster*

dmo: *Drosophila mojavensis*

dps: *Drosophila pseudoobscura*

dya: *Drosophila yakuba*

dwi: *Drosophila willistoni*

dvi: *Drosophila virilis*

dsi: *Drosophila simulans*

dse: *Drosophila sechellia*

dpe: *Drosophila persimilis*

dgr: *Drosophila grimshawi*

der: *Drosophila erecta*

dan: *Drosophila ananassae*

Ds RNA: double-stranded RNAs.

Hid: head involution defective

isc: *Ixodes scapularis*

Kb: kilobase

lmi: *locusta migratoria*

MFE: minimum folding free energy

mRNA: messenger RNA

miRNA: microRNA

miRNA *: opposite miRNA sequence

ncRNA: non coding RNA

nt: nucleotide (s)

ORF: open reading frame

Pre-miRNA: precursor miRNA

pri-miRNA: primary microRNA

PCR: polymerase chain reaction

qRT-PCR: quantitative real-time reverse transcription-polymerase chain reaction.

RNAi: RNA interference

RISC: RNA- induced silencing complex

SVM: support vector machine

tca: *Tribolium castaneum*

tca: *Tribolium castaneum*

U: Uracil

UTR: untranslated region (s)

3’UTR: untranslated region of a messenger RNA following the coding sequence.

Abstract

Mona Gaber Abd-El-Aziz Mahmoud Shalaan. Bioinformatics and Computational Biological Approaches to Identify Insect microRNAs. Faculty of Science, Ain Shams University, 2010.

MicroRNAs (miRNAs) are a large family of 21-22 nucleotides (nts) long non- protein-coding small RNAs, which are involved in regulation of mRNA translation and expression in the cell. Mature miRNAs are processed from a longer sequence known as pre-miRNAs, which can form stem-loop hairpin secondary structure. Mature miRNAs unite with multi-protein complexes known as RNA-induced silencing complex (RISC), which binds to specific sequences in target mRNAs. This binding triggers the translational repression or degradation of many mRNAs.

Due to the very short sequence of miRNAs and their repeats in genomes, bioinformatics and comparative genomics have been the method of choice to identify them, with subsequent validation by biochemical and molecular cloning. The integration of theoretical *in silico* and experimental miRNAs identification approaches helps to overcome the problems of misidentification (false positives) and missed identification (false negatives) of these tiny genes. These also enabled researchers to distinguish them from other small RNAs scattered in the genome.

A large number of miRNAs has been identified in many insect species including the fruit fly *Drosophila*, the mosquito *Anopheles*

gambiae, the red flour beetle *Tribolium castaneum* and the silk worm moth *Bombyx mori*. Identification of miRNAs critical for insect development will help develop new tools to control serious disease vectors and pests. In this study we extended the process of miRNAs identification to the southern house mosquito *Culex pipiens quinquefasciatus*, a serious vector of filariasis and arboviral diseases in the world and Egypt. The genome of this mosquito is in the final stages of completion and raw sequence files were accessed from www.vectorbase.org.

Identification of miRNAs from the genome of *Cx. p. quinquefasciatus* (Cpq) was carried out using two different predictions methods. The first method depended on the identification of Cpq-miRNAs directly from the whole genome. A region of nearly 280000 nts from the whole genome was finished. To facilitate the prediction process, this region was divided into 28 microcontigs (micons) each of 10,000 nts in length. Three different overlapping windows between micons were tested: 50-nts, 250-nts and 500-nts. The 50-nts window was the best one; it produced predictions more than the other overlapping windows. From five of these 28 micons, 51 pre-mirs were predicted including 9 pre-mirs (17.9%) that are considered as novel mirs.

The second identification method was by homology search using identified miRNAs to search for homologous pre-mirs in *Cx. quinquefasciatus* genome. By this method the following miRNAs were identified as a test set, including Cpq-mir-1, Cpq-let-7, Cpq-mir-263b, Cpq-mir-276, Cpq-mir-307, Cpq-mir-315, Cpq-mir-7, Cpq-mir9c, Cpq-bantam and Cpq-mir-87. These mirs conformed to the general

miRNAs criteria and with >90% homology at the mature sequences in insect mirs.

Prediction of secondary structure stem-loop formation of all pre-mirs identified by the 2 methods was carried out by using MFold programme. The following mirs were identified by both methods: mir-315, mir-7, mir-9c, bantam and mir-87.

To identify all Cpq-miRNAs, it was necessary to build a customized database that contains the whole genome sequence of the test mosquito *Cx. p. quinquefasciatus* and a reference genome, such as that of *Drosophila* or *An. gambiae*. We were able to finish programming of the primary functions and processes such as gene splicing and translation. However, it was difficult to finish more complex functions needed for miRNAs large-scale prediction (within the time-frame of the MSc thesis). This task is a future objective, in addition to the experimental characterization of identified *Culex* miRNAs.

Identification of *Culex* miRNAs opens the field for the functional analysis of these genes to understand their role in mosquito development. The results of such research will pinpoint novel targets for mosquito control.

Key words: Bioinformatics, computational biology, *Culex* mosquitoes, microRNAs.

Contents

LIST OF TABLES AND FIGURES

I. TABLES