Development of a Local Entomological Database for Education and Research using Simulation (Virtual) Methods

Emad I. Khater [1,] Mona G. Mahmoud, Enas H. Ghallab, Magdi G. Shehata
[1]Department of Entomology,
Faculty of Science,
Ain Shams University,
Cairo, Egypt

Yasser M. Abd El-Latif
[2]Department of Mathematics,
Faculty of Science,
Ain Shams University,
Cairo, Egypt

Abstract—**Bioinformatics has been regarded as one of the rapidly-evolving fields with enormous impact on the history of life and biomedical sciences. It is an interdisciplinary science that integrates life sciences, mathematics and computer science in order to extract meaningful biological insights from large data sets of raw DNA and protein sequences. Our objective was the development of an entomogenomics database (provisionally named EntomDB) for education and research in entomology (entomology is the science of insects). This DB includes DNA/protein sequence data selected from genomes of major insect models of importance in biology and biomedical research. EntomDB will represent a customized easy, interactive and self-learning resource tool for beginner users in poor-resource settings. This will enable the users to learn basic skills in bioinformatics and genomics, needlessly to search through the numerous databases currently available on the World-Wide Web with their complex interfaces and contents. EntomDB will help students and young researchers in studying the primary structure, splicing, and translation and predict function of different genes by using simple simulation methods. It is also designed to be adaptable to work off-line, in case no internet connection is available. EntomDB is primarily designed for entomology discipline; however, it can easily be adapted for other disciplines in life and biomedical sciences. EntomDB will have important educational and developmental outcomes in promoting bioinformatics learning in the developing world and provide affordable first-level training for advanced degree and research levels.**

Keywords— Bioinformatics; local database; entomology

I. Introduction

Bioinformatics and genomics are two of the modern life sciences that are rapidly evolving and changing the history of science. Bioinformatics is defined in the broadest sense as the application of computation in biology to handle, manage and analyze large data sets in short time with high speed and accuracy. It is an interdisciplinary science that strongly links life sciences, molecular biology, mathematics and computer science.

It has been a powerful tool in the development of genomics and related sciences and a major tool for gaining new insights of biological data. Bioinformatics has become a crucial area of development and investment for major interest groups in academia, industry and governments, with multi-billion US dollars investments worldwide in Europe, USA and Asia [1], [2], [3], [4], [5]. Bioinformatics, as quoted from Hwa Lim [6]:

"Bioinformatics, which makes biotechnology predictive, is the discipline that concerns with the study of information content and information flow in biological systems. Understandably, it first became viable only in the second half of the 1980s, after the beginning of the genome initiative and advances in computer technologies. Any earlier would have been rather premature, except for the study of living systems as information systems and modest efforts. Due in part to the avalanche of genetic data from genome and proteome projects and from the capability to collect other medical and healthcare data, bioinformatics deals primarily with these data... it is thus not surprising that many people define bioinformatics as collection, analysis, management and dissemination of biological (medical and healthcare) data"…. The process can be repeated, but the point is clear. This is one reason why creation of databases is one of the most active activities in bioinformatics."

Based on these assertions, each part of Bioinformatics definitions will be perceived differently by different users [6].

This genomics revolution has been sparked by the availability of huge data of raw genome sequences and the outcome of various analyses, which necessitated/led to parallel advancements in biotechnology platforms and tools, the most important of which is bioinformatics. This intimate and mutual interrelationship between genomics and bioinformatics is so obvious, that we cannot separate the two from each other.

Genomics is the study and analysis of genomes (genome is the entire genetic material of an organism, which acts as the blueprint for this organism's functioning and life). With the sequencing of the human genome in 2001 [7], [8] (one of the most important scientific achievements in the history of science) there has been an explosion of genome sequencing projects and related sciences. The sequenced genomes include mouse, fruit flies, worms and yeast as well as microbes. Estimates show that the GenBank and molecular databases contain gene/protein sequences and related biological information of thousands of organisms from all biological taxa. These huge repositories of data are categorized into 14 categories with a total over 1000 databases (http://www.oxfordjournals.org/nar/database/a). All these data and databases are easily and freely accessible from the GenBank (http://www.ncbi.nlm.nih.gov) or as downloadable FTP files. There are 80-190 billion nucleotide bases (the measuring unit of gene sequence) in the GenBank, which doubles almost every 2 months. According to recent information at NCBI (National Center for Biotechnology Information) (last accessed March 22, 2011), the genome sequencing projects are reaching the 10000 limit at different stages: completed, in-assembly, or in-progress. There are 731 eukaryotic nuclear genomes including 71 insect genome sequencing projects, of which two are completed, 44 in assembly and 25 in-progress. The NCBI Entrez Gene contains data of >5.4 million genes of over 6200 organisms, which are accumulated and maintained through collaboration of many international partners [9].

Currently there are over 1200 bioinformatics links collected in a specific directory. This directory is a compilation of molecular biology servers, bioinformatics tools and online resources. The availability of such tools has enormously shifted science, from the laboratory-based to the information-based with fast and easy communication between different laboratories in the world [10], [11], [12].

Early practitioners in bioinformatics were self-taught to solve specific problems in their research. With the current revolution in biological research and the key role played by bioinformatics, it has become crucial to have new generations of graduates, which are equipped with a broad range of skills in basic and molecular biology, computational biology and mathematics to carry out the huge tasks set out by modern biology. A major consequence to this urgent need was that many universities have adopted educational and research programmes in bioinformatics and genomics that include a range of theoretical and practical skills to link experimental biology to computational biology [13]. Many of these training programmes can be conducted online, to overcome the obstacles of distance and resources.

According to the level of use and expertise, users of bioinformatics can be classified into three categories from the beginner to the expert [13]:

Super-users, with good knowledge and understanding of basic tools, but have no programming or (DB)-developing skills,

Power-users, with good knowledge and understanding of basic tools, underlying algorithms, with good programming and DB-developing skills,

Bioinformaticians, those that can develop, implement algorithms, develop new tools and software, simulate and model and use IT systems to manage and extract biological information from raw sequence data.

The British QAAA (Quality Assurance and Accreditation Agency) review identified the skills needed for biology graduates, which include solving problems using computers and apply electronic means as source of information and communication as well as learning quantitative and qualitative methods. Based on the level of learning and mastering these skills, graduate students are classified into threshold or minimum standard and good standard students [13].

For the developing world, as in Egypt, for example, the need to apply such approaches of *in silico* biology (computational and bioinformatics) that applies a great deal of computer work and predictive analysis of online data is rather crucial than for the developed world. This because of the serious lack of expertise and resources to carry out basic and advanced research in genomics and bioinformatics that are essential tools in modern biology, in particular the use of biotechnology in all fields of life science and society affairs. Such approach of online and computer-based learning will produce new generations with basic understanding of genomics and bioinformatics methods and principles applied in various life science fields and services. Through such first-level learning, they will be able to move from under super-users to high levels of power-users and experts. These future graduates will be the main source of skilled experts and users that are urgently needed for academia and industry.

The urgency of the critical situation of bioinformatics in Egypt is well reflected in a specialist document of the National Specialized Councils, an offshoot from the Presidential Department in 2002/2003. This document clearly highlighted the serious gaps in Egypt in bioinformatics teaching and research, and the crucial need to develop a focused educational and research programmes in bioinformatics. This is in the light of its global importance in biotechnological applications and investments. Of the recommendations of this expert committee are:

1) *Development of bioinformatics education system,*

2) *Increasing research programmes in bioinformatics both in the laboratory and online,*

3) *Linking Egyptian programmes to international resources, instead of replicating them.*

This urgent need for teaching and research in bioinformatics as expressed at the highest political level in the country, the proposed project to develop a system that is easy and accessible to first-level users in Egypt was initiated to be the first in its kind in Egypt, with the potential probability to be further distributed in the Arabic region.

Therefore, constructing a local customized entomological database able to collect, organize and deal with the volume and complexity of data relevant to the important model insect species will be a great challenge in entomology field. Building a local specialized database including DNA/protein sequences relevant to the selected insect species will significantly facilitate learning basic skills in bioinformatics without the need to search within over than 1000 molecular databases currently available on the World Wide Web (www).

II. Materials and Methods

A. Approach:

Insects are considered as the largest and the most widely-distributed group of animals on earth. At present, there are over 71 sequenced insect genomes that are freely accessible from the GenBank (http://www.ncbi.nlm.nih.gov) including the fruit fly *Drosophila melanogaster* (a major model for scientific research), the mosquitoes *Anopheles gambiae* and *Culex pipiens* (as disease carriers for both animal and human), the locust *Locusta migratoria* and the red flour beetle *Tribolium castaneum* (as harmful and destructive pests of many economic plants) and the honey bee *Apis mellifera* and the silk worm moth *Bombyx mori* (as beneficial insects important in agriculture and industry).

The proposed database is a collection of gene sequence data obtained from major genome databases as the GenBank (http://www.ncbi.nlm.nih.gov) and specific insect genome databases such as the Flybase (http://www.flybase.org) of the fruit fly *D. melanogaster* and AnoBase (http://www.anobase.org) (or vectorbase.org) of the mosquito *An. gambiae*. The gene sequence data and relevant information were fed into Microsoft SQL Server 2005. These data include: gene nucleotide sequence (genomic, cDNA, mRNA) in a FASTA file format; sequence identifiers (GenBank identifier, accession number, sequence length etc.); amino acid sequence of the protein encoded by this gene or a translation of the sequence in all possible reading frames and other relevant data.

B. *Architecture:*

Gene sequences of small size, e.g. 200 bp (bp= base pairs; it is the measuring unit of gene nucleotide sequence) were used to test the system. To train the system, basic procedures of searching, retrieving and analyzing a given gene sequence were performed. The type of information obtained, results, speed and accuracy of the system were assessed and compared to similar functions and results when using the parent global system to perform the same procedures and queries.

C. *Identification of the intron splice junctions and mRNA translation:*

Almost all eukaryotic genes contain intervening non-protein coding sequences or introns interrupting the coding regions or exons. After removal or splicing of introns, exons will be linked together in processed messenger RNAs (mRNAs) ready for translation to proteins. There are at least eight different kinds of intron splice sites or junctions that have been found in eukaryotic genes. The GU-AG-intron is the predominant type associated with eukaryotic protein-coding genes. This GU-AG type gets its name from the fact that the intron 5'-splice site or donor starts with a GU (GT) and 3'-splice site or acceptor ends with an AG. In addition, the intron sequences between the donor and acceptor sites bases have specific consensus sequence.

The splice sites or junctions of introns were identified using the method described [14]. The probability of a "GU" as a splice junction was computed from the flanking (preceding and succeeding) nucleotides. The potential splice junctions were printed in order from most to least likely, using the first character index of value 1. This method is commonly used instead of Perl indices using the first character index of value 0. Further, the programme used makes use of Perl treatment of strings, in that stepping off the left or right end of a string is allowed; the result is simply the null string.

After identification and removal of introns, and conjoining of exons, the resultant contiguous string of nucleotides or the open reading frame (ORF) of the gene will be translated to amino acid sequence of a given protein or a polypeptide. Therefore, translation is also known as the process of converting the information from the nucleotide sequences in the mRNA to the amino acid sequence encoding for a protein or a polypeptide.

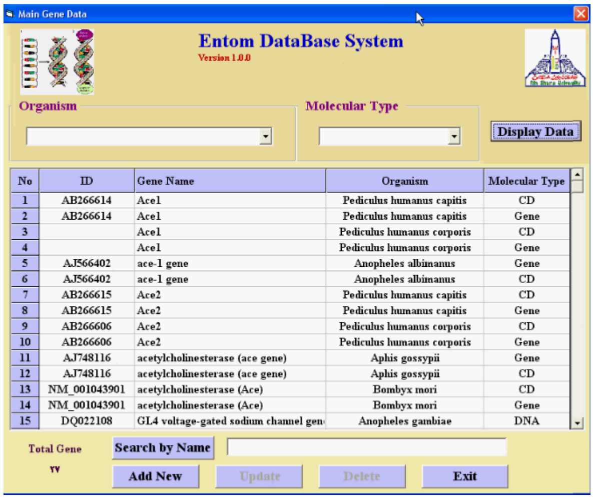
The programme outlined [14] is used for the conceptual translation of an ORF string of nucleotides to amino acids. It prints all three reading frames in the forward or sense direction, skipping nucleotides at the beginning to produce the additional frames and ignoring the nucleotides at the end that do not form a group of 3 or codon (a codon; every 3 nucleotides correspond to an amino acid). It should also be noted that the genetic code is degenerate; there are some amino acids that are encoded by more than one codon.

III. Results & Discussion:

A. *The training gene set and basic processes:*

Our local database, EntomDB includes information on DNA/protein sequences of different genes selected from major insect models. We have imported some biological operations to study the structure and function of each gene. At present, the local customized database includes a training set of eighteen gene sequences with all its operations successfully achieved (Fig. 1).

According to the central dogma of molecular biology (Fig. 2), the genetic material (DNA) present within the cell nucleus is firstly transcribed into mRNA (an intermediate state less stable than DNA), which is then transported into the cytoplasm to be translated into amino acids, the building unit of the proteins that will perform different cellular functions. Both biological processes of DNA transcription and RNA translation are reported by our database for each inserted gene. Splicing is an essential step of RNA transcript processing prior to translation, which leads to excision of introns, which are then removed and exons are joined together to form a contiguous gene transcript. This step takes place for most of eukaryotic mRNAs before they can be used to produce correct proteins through translation. Nearly all eukaryotic nuclear introns begin with the nucleotide sequence GT (donor site) and end with AG (acceptor site) (the GT-AG rule) (Fig. 3). At the time of splicing, the intron is composed of RNA not DNA, so the beginning of the sequence is GU not GT (Fig. 4). For mRNA translation, there are 6 possible reading frames (3 frames for each strand of the two DNA strands of a gene). The reading frame or ORF that will be translated into a contiguous string of amino acids was manually selected. This ORF starts with ATG codon of the amino acid methionine (Met or M for short) and ends with one of 3 stop codons, UAG or UAA or UGA.



Main Gene Data

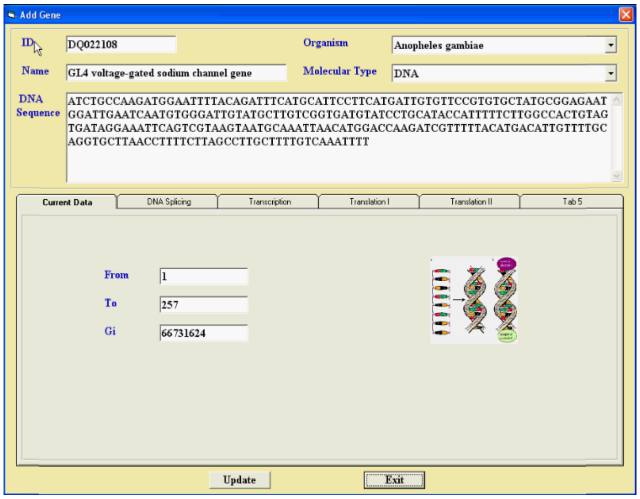
Entom DataBase System
Version 1.0.0

Organism □ Molecular Type □ [Display Data]

No	ID	Gene Name	Organism	Molecular Type
1	AB266614	Ace1	Pediculus humanus capitis	CD
2	AB266614	Ace1	Pediculus humanus capitis	Gene
3		Ace1	Pediculus humanus corporis	CD
4		Ace1	Pediculus humanus corporis	Gene
5	AJ566402	ace-1 gene	Anopheles albimanus	Gene
6	AJ566402	ace-1 gene	Anopheles albimanus	CD
7	AB266615	Ace2	Pediculus humanus capitis	CD
8	AB266615	Ace2	Pediculus humanus capitis	Gene
9	AB266606	Ace2	Pediculus humanus corporis	CD
10	AB266606	Ace2	Pediculus humanus corporis	Gene
11	AJ748116	acetylcholinesterase (ace gene)	Aphis gossypii	Gene
12	AJ748116	acetylcholinesterase (ace gene)	Aphis gossypii	CD
13	NM_001043901	acetylcholinesterase (Ace)	Bombyx mori	CD
14	NM_001043901	acetylcholinesterase (Ace)	Bombyx mori	Gene
15	DQ022108	GL4 voltage-gated sodium channel gen	Anopheles gambiae	DNA

Total Gene ٢٧

Fig. 1. A screen shot of the database interface including a number of genes from different organisms.



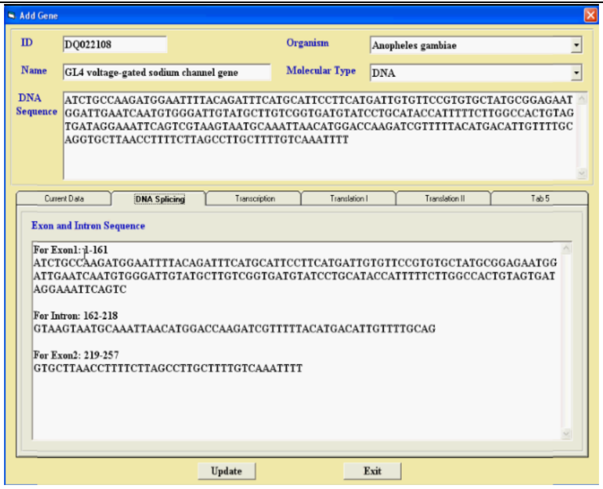
Add Gene

ID: DQ022108
Organism: Anopheles gambiae
Name: GL4 voltage-gated sodium channel gene
Molecular Type: DNA
DNA Sequence:

```
ATCTGCCAAGATGGAATTTTACAGATTTCATGCATTCCTTCATGATTGTGTTCCGTGTGCTATGCGGAGAAT
GGATTGAATCAATGTGGGATTGTATGCTTGTCGGTGATGTATCCTGCATACCATTTTTCTTGGCCACTGTAG
TGATAGGAAATTCAGTCGTAAGTAATGCAAATTAACATGGACCAAGATCGTTTTTACATGACATTGTTTTGC
AGGTGCTTAACCTTTTCTTAGCCTTGCTTTTGTCAAATTTT
```

From: 1
To: 257
Gi: 66731624

Fig. 2. A screen shot each gene showing its sequence information, Genbank ID, the source organism, molecular type.



Add Gene

ID: DQ022108
Organism: Anopheles gambiae
Name: GL4 voltage-gated sodium channel gene
Molecular Type: DNA
DNA Sequence:

```
ATCTGCCAAGATGGAATTTTACAGATTTCATGCATTCCTTCATGATTGTGTTCCGTGTGCTATGCGGAGAAT
GGATTGAATCAATGTGGGATTGTATGCTTGTCGGTGATGTATCCTGCATACCATTTTTCTTGGCCACTGTAG
TGATAGGAAATTCAGTCGTAAGTAATGCAAATTAACATGGACCAAGATCGTTTTTACATGACATTGTTTTGC
AGGTGCTTAACCTTTTCTTAGCCTTGCTTTTGTCAAATTTT
```

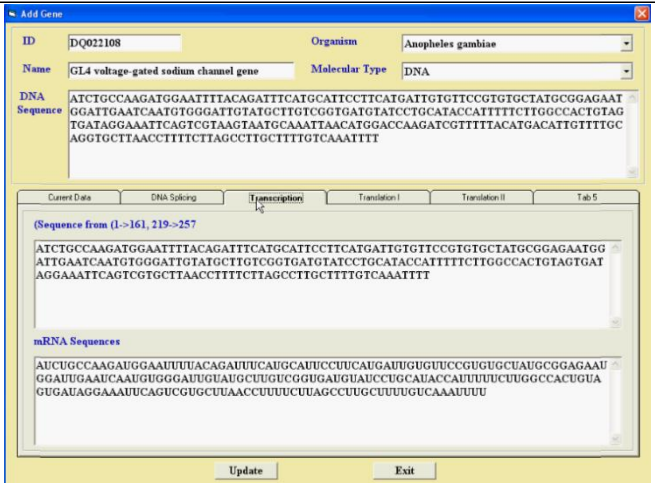
Exon and Intron Sequence

```
For Exon1: 1-161
ATCTGCCAAGATGGAATTTTACAGATTTCATGCATTCCTTCATGATTGTGTTCCGTGTGCTATGCGGAGAATGG
ATTGAATCAATGTGGGATTGTATGCTTGTCGGTGATGTATCCTGCATACCATTTTTCTTGGCCACTGTAGTGAT
AGGAAATTCAGTC

For Intron: 162-218
GTAAGTAATGCAAATTAACATGGACCAAGATCGTTTTTACATGACATTGTTTTGCAG

For Exon2: 219-257
GTGCTTAACCTTTTCTTAGCCTTGCTTTTGTCAAATTTT
```

Fig. 3. A screen shot of mRNA splicing: represents the exon/intron regions within each gene.



Add Gene

ID: DQ022108
Organism: Anopheles gambiae
Name: GL4 voltage-gated sodium channel gene
Molecular Type: DNA
DNA Sequence:

```
ATCTGCCAAGATGGAATTTTACAGATTTCATGCATTCCTTCATGATTGTGTTCCGTGTGCTATGCGGAGAAT
GGATTGAATCAATGTGGGATTGTATGCTTGTCGGTGATGTATCCTGCATACCATTTTTCTTGGCCACTGTAG
TGATAGGAAATTCAGTCGTAAGTAATGCAAATTAACATGGACCAAGATCGTTTTTACATGACATTGTTTTGC
AGGTGCTTAACCTTTTCTTAGCCTTGCTTTTGTCAAATTTT
```

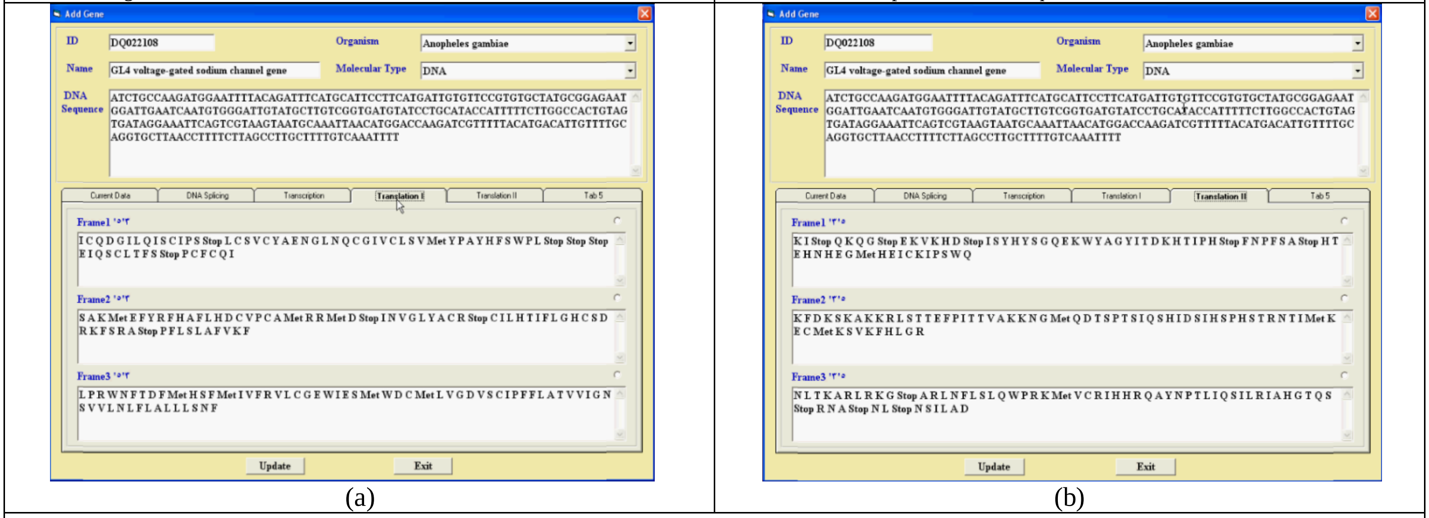
(Sequence from (1->161, 219->257

```
ATCTGCCAAGATGGAATTTTACAGATTTCATGCATTCCTTCATGATTGTGTTCCGTGTGCTATGCGGAGAATGG
ATTGAATCAATGTGGGATTGTATGCTTGTCGGTGATGTATCCTGCATACCATTTTTCTTGGCCACTGTAGTGAT
AGGAAATTCAGTCGTGCTTAACCTTTTCTTAGCCTTGCTTTTGTCAAATTTT
```

mRNA Sequences

```
AUCUGCCAAGAUGGAAUUUUACAGAUUUCAUGCAUUCCUUCAUGAUUGUGUUCCGUGUGCUAUGCGGAGAAU
GGAUUGAAUCAAUGUGGGAUUGUAUGCUUGUCGGUGAUGUAUCCUGCAUACCAUUUUUCUUGGCCACUGUA
GUGAUAGGAAAUUCAGUCGUGCUUAACCUUUUCUUAGCCUUGCUUUUGUCAAAUUUU
```

Fig. 4. A screen shot of transcription, the gene sequence is transcribed to obtain the mature spliced mRNA sequence.



Add Gene

ID: DQ022108
Organism: Anopheles gambiae
Name: GL4 voltage-gated sodium channel gene
Molecular Type: DNA
DNA Sequence:

```
ATCTGCCAAGATGGAATTTTACAGATTTCATGCATTCCTTCATGATTGTGTTCCGTGTGCTATGCGGAGAAT
GGATTGAATCAATGTGGGATTGTATGCTTGTCGGTGATGTATCCTGCATACCATTTTTCTTGGCCACTGTAG
TGATAGGAAATTCAGTCGTAAGTAATGCAAATTAACATGGACCAAGATCGTTTTTACATGACATTGTTTTGC
AGGTGCTTAACCTTTTCTTAGCCTTGCTTTTGTCAAATTTT
```

Frame1 '٥'٣

```
I C Q D G I L Q I S C I P S Stop L C S V C Y A E N G L N Q C G I V C L S V Met Y P A Y H F S W P L Stop Stop Stop
E I Q S C L T F S Stop P C F C Q I
```

Frame2 '٥'٣

```
S A K Met E F Y R F H A F L H D C V P C A Met R R Met D Stop I N V G L Y A C R Stop C I L H T I F L G H C S D
R K F S R A Stop P F L S L A F V K F
```

Frame3 '٥'٣

```
L P R W N F T D F Met H S F Met I V F R V L C G E W I E S Met W D C Met L V G D V S C I P F F L A T V V I G N
S V V L N L F L A L L L S N F
```

(a)

Add Gene

ID: DQ022108
Organism: Anopheles gambiae
Name: GL4 voltage-gated sodium channel gene
Molecular Type: DNA
DNA Sequence:

```
ATCTGCCAAGATGGAATTTTACAGATTTCATGCATTCCTTCATGATTGTGTTCCGTGTGCTATGCGGAGAAT
GGATTGAATCAATGTGGGATTGTATGCTTGTCGGTGATGTATCCTGCATACCATTTTTCTTGGCCACTGTAG
TGATAGGAAATTCAGTCGTAAGTAATGCAAATTAACATGGACCAAGATCGTTTTTACATGACATTGTTTTGC
AGGTGCTTAACCTTTTCTTAGCCTTGCTTTTGTCAAATTTT
```

Frame1 '٣'٥

```
K I Stop Q K Q G Stop E K V K H D Stop I S Y H Y S G Q E K W Y A G Y I T D K H T I P H Stop F N P F S A Stop H T
E H N H E G Met H E I C K I P S W Q
```

Frame2 '٣'٥

```
K F D K S K A K K R L S T T E F P I T T V A K K N G Met Q D T S P T S I Q S H I D S I H S P H S T R N T I Met K
E C Met K S V K F H L G R
```

Frame3 '٣'٥

```
N L T K A R L R K G Stop A R L N F L S L Q W P R K Met V C R I H H R Q A Y N P T L I Q S I L R I A H G T Q S
Stop R N A Stop N L Stop N S I L A D
```

(b)

Fig. 5. A screen shot of translation, the mRNA sequence is then translated into 6 frames and only one frame, the open reading frame (ORF) will be translated into a contiguous amino acid chain of a protein or a polypeptide.

D. Determination of the intron splice junctions:

We were able to define the intron splice sites using the GU-AG rule as shown above. No strictly followed rules appear to govern the distribution of these introns though they are generally less common in simpler eukaryotes. Introns are a very common feature in the genes of higher eukaryotes such as vertebrates and at least one intron exists in a gene. Aside from the sequences required for splicing, the length and nucleotide sequences of introns appear to be highly variable. In contrast, the position of introns within any given gene does seem to be evolutionarily-conserved, i.e., they are often in identical positions in alignments of the sequences of homologous genes.

All intron 5'-splice and 3'-splice junctions appear to be functionally equivalent to the splicing apparatus or spliceosome. In normal circumstances, splicing occurs only between the junctions of the same intron (cis-splicing). The molecular basis for differentiating between splice junctions of the same and different introns appears to be more complex and remains an important question for molecular biologists [14]. The majority of eukaryotic genes appear to be processed into a single type of spliced mRNA (monocistronic). However, some genes give rise to more than one type of mRNA sequence (ploycistronic) due to alternative splicing. In one extreme example of alternative splicing, one single human gene has been shown to generate up to 64 different mRNAs from the same primary transcript.

The nuclear membrane of eukaryotic cells provides a physical barrier that separates the process of transcription and translation in a way that never occurs in prokaryotes; where translation by ribosomes typically begins as soon as an RNA polymerase enzyme has begun to make an RNA copy of a gene coding region.

After the completion of the human genome project, a huge amount of DNA sequence information has been generated. The dealing with this large amount of data requires a clear need to gather, filter and evaluate this mass of information so that it can be used with greater efficiency. Bioinformatics is the science that has emerged in response to the development of genomics so it could handle, manage and analyze large data sets in short time with high speed and accuracy. Due to its importance, teaching bioinformatics at the undergraduate level in a biology laboratory program using customized, interactive and problem-based approaches has become an essential requirement in life and biomedical science curricula in the last two decades [5], [15], [16], [17], [18], [19], [20]. They all established guidelines for the contents and format of undergraduate bioinformatics courses and training programmes using customized software and tools with hyperlinks to web applications. These approaches will help increase cooperation between scientists of various departments, such as biology, mathematics and computer science, to conduct a given course. This first-level training will form the basis for advanced postgraduate and degree programs [13].

Due to the importance of incorporating bioinformatics within introductory undergraduate biology courses (curricula), many strategies have evolved. One of the applied strategies is the use of customized tools and databases to facilitate teaching, time saving and preventing students from being overwhelmed by complex graphical user interface and outputs of international publicly available bioinformatics databases such as NCBI, ExPASy, nps@, etc. [19]. The complexity of web applications could be distracting and discouraging to many beginner students to learn bioinformatics; as they might feel "lost in cyber space" and loose insight of the main purpose of using a given tool. Further, the delay in obtaining results due to busy servers could cause frustration of both students and tutors and lose confidence in the exercise [19]. Additionally, using computers as essential tools in scientific education and research will strengthen the skills of graduates in using computers to access databases, manipulate and store specific information and communicate effectively. All these skills are essential requirements for modern graduates who are seeking good career opportunities [18].

There is a continuous and enormous development in the fields of bioinformatics and genomics and the fields employing them in the academia and business sectors. In the time being and in the future, there will be an increasing demand for graduates with good skills in bioinformatics and genomics.

For future plans and with the success of this EntomDB prototype, we will test the possibility of publishing a version with Arabic interface, which will significantly increase the base of users and attract more audience. Gene sequences of larger size will be tested with adding more functions and queries. This will further train and test the rigor of the system to handle larger body of input data and perform more complex functions. The entire genome sequence of the other selected insect models will be tested and more data and functions will be added to the system and the efficiency of access and retrieval of specific data will be tested. Hyperlinks with global tools and databases will also be inserted in the EntomDB for advanced users. The system will be tested for the possibility of off-line usage and a prototype CD of the system will be developed.

In conclusion, development of a local customized entomological database for research and education will increase the ability of undergraduate students and young scientists to master broad and sophisticated bioinformatics tools and apply them for better understanding of biological processes and systems. This database is an easy-to-use, self-learning tool to prepare the users for more specialized data analysis. Increasing the use of *in silico* “dry laboratory” bioinformatics of genome sequences will help obtain preliminary results to guide “wet laboratory” experiments, thus minimizing the time and costs.

References

[1] D. W. Mount (2001). Bioinformatics: Sequence and genome analysis, Cold Spring Harbor Laboratory Press, New York.

[2] C. Ouzounis (2002). Biology by other means. Editorial, Bioinformatics, 18(12),1551-1552.

[3] A. Valencia (2002). Bioinformatics and the theoretical foundation of molecular biology. Editorial, Bioinformatics, 18(13),377-378.

[4] A. Campbell, L. Heyer (2003). Discovering genomics, proteomics and bioinformatics. Pearson Education, Benjamin Cummings, San Francisco, pp. 352.

[5] Y. Ai, Jermiin, N. Firth (2003). Teaching bioinformatics: A student-centered and problem based approach. In: CAL-laborate, a collaborative publication on the use of Information Technology in tertiary teaching and learning for the life sciences, published by UniServe Science, The

University of Sydney, Australia, http://science.uniserve.edu.au/, June, 2003, pp. 25-30.

[6] H. A. Lim (2005). Informatics, Bionformatics and Binformatics, a plenary lecture at the International Meeting on Frontiers of Physics, the Year of Physics, IMFP2005-Binformatics, pp.:1-18, http://www.dtrends.com/Bioinformatics/bioinfoIMFP05.html.

[7] J. C. Venter, M. D. Adams, E. W. Myers,P. W. Li, R. J. Mural, G. G. Sutton, et al. (2001). The sequence of the human genome. Science, 291(5507), 1304-51.

[8] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, et al. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860-921.

[9] E. W. Sayers, T. Barret, D. Benson. E.Bolton, et al., (2010). Database resources of the National Center for Biotechnology Information, Nucleic Acids Research, Vol. 38, (Database issue), D5-D16.

[10] M. Galperin (2006). The molecular biology database collection: 2007 update. Nucleic Acids Research, 35, (Database issue), D3-D4.

[11] D. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D. J. Wheeler (2007). Genbank. Nucleic Acids Research, 35 (Database issue), D3-D4.

[12] J. A. Fox, S. McMillan, B. F. Francis Ouellette (2007). Conducting research on the web: 2007 update for the bioinformatics links directory. Nucleic Acids Research, 35 (Web Server issue), W3-W5.

[13] C. Hack, G. Kendall (2005). Bioinformatics: Current practice and future challenges for life sciences education. Biochem. Mol. Biol. Edu, 33(2), 82-85.

[14] D. E. Krane, M. L. Raymer (2005). Fundamental concepts of bioinformatics, Pearson Education, Benjamin Cummings, San Francisco, pp. 249-276.

[15] J. Tillotson (2002). Strategies for introducing computer technologies into a biology laboratory program, Biochem. Mol. Biol. Edu., 30(4), 232-234.

[16] N. B. Centeno, J. Villa-freixa, B. Oliva (2003). Teaching structural bioinformatics at the undergraduate level, Biochem. Mol. Biol. Edu, 31(6), 386-391.

[17] J.Cohen (2003). Guidelines for establishing undergraduate bioinformatics courses. Journal of Science Education and Technology, 12(4), 449-456.

[18] J. Boyle (2004). Bioinformatics in undergraduate education, Biochem. Mol. Biol.Edu, 32 (4), 236- 238.

[19] M. Neumann, N. Provart (2006). Using customized tools and databases for teaching bioinformatics in introductory biology courses. Assoc. Biol. Lab. Edu. (ABLE), 27, 321-328.

[20] S. Cattley, J. Arthur (2007). BioManager: the use of a bioinformatics web application as a teaching tool in undergraduate bioinformatics training, Briefings in Bioinformatics, 8(6), 457-465.