COMPUTER-BASED SCREENING OF PISCINE ENDOGENOUS RETROVIRUS FROM PUFFERFISH GENOME PROJECT DATABASE

MALATY SINNIAH

PROGRAM BIOTEKNOLOGI
SEKOLAH SAINS DAN TEKNOLOGI
UNIVERSITI MALAYSIA SABAH
2005

PUMS 99:1

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS@

JUDUL: Computer-based Screening of Piscine Endogenous Retrovirus from Pufferfish Genome Project Database

Ijazah: Sarjana Muda Sains

SESI PENGAJIAN: 2002/2005

Saya MALATY SINNIAH

(HURUF BESAR)

mengaku membenarkan tesis (LPS/Sarjana/Doktor Falsafah)* ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:

1. Tesis adalah hakmilik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. **Sila tandakan (/)

[✓]	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
[ ]	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
[ ]	TIDAK TERHAD	

Disahkan oleh

(TANDATANGAN PENULIS)

(TANDATANGAN PUSTAKAWAN)

Alamat Tetap: 515, Jln. Merak 5, Tmn. Thivy Jaya, 70300 Seremban, N.S.

DR. ROZIAH KAMBOL
Nama Penyelia

Tarikh: 28/3/05

Tarikh:

CATATAN: * Potong yang tidak berkenaan.

** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT dan TERHAD.

@ Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana secara penyelidikan, atau disertasi bagi pengajian secara kerja kursus dan penyelidikan, atau Laporan Projek Sarjana Muda (LPSM).

COMPUTER-BASED SCREENING OF PISCINE ENDOGENOUS RETROVIRUS FROM PUFFERFISH GENOME PROJECT DATABASE

MALATY SINNIAH

DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEHI IJAZAH SARJANA MUDA SAINS DENGAN KEPUJIAN

PROGRAM BIOTEKNOLOGI
SEKOLAH SAINS DAN TEKNOLOGI
UNIVERSITI MALAYSIA SABAH

2005

DECLARATION

I hereby declare that this dissertation is based on my original work except for certain citations, quotations and summaries, which have been duly acknowledged.

31st MARCH 2005

MALATY SINNIAH

HS2002-3063

AUTHENTICATION

Signature

1. **SUPERVISOR**

   (DR. ROZIAH KAMBOL) _____

2. **EXAMINER 1**

   (DR. VIJAY KUMAR) _____

3. **EXAMINER 2**

   (DR. WONG NYET KUI) _____

4. **DEAN**

   (ASSOC. PROF. DR. AMRAN AHMED) _____

ACKNOWLEDGEMENTS

I would like to take this great opportunity to show my appreciation and gratitude to those who have contributed in completion of this project. First of all, I would like to express my gratitude to my supervisor Dr. Roziah Kambol for her guidance throughout this project. Her helpful and supportive comments on my writing for this project helped me a lot throughout the finishing of this project. Without her, it is not possible for me to complete this project successfully.

In preparing for this study, I owe my gratitude to my friends and colleagues. I would particularly like to express my deepest thanks to my friends Saratha, Anusuya, and Priya who generously contributed their time and effort to help me to complete this project.

I would also like to thank my beloved family, especially my father Mr. Sinniah, both my brothers Manoharan and Thevaraj who have been a pillar of full strength and comfort me all the while.

Lastly, I would like to express my thanks to my dear friends especially Kanagam, Vellaseni and Suba for their continuous encouragement and support. Thanks for everyone whom I mentioned and to who indirectly involved in this project.

ABSTRACT

The main purpose of this study is to determine the existence of endogenous retrovirus in Pufferfish genome through screening the Pufferfish Genome Database. In this study, the reverse transcriptase portion of *pol* gene was examined as it exhibits the most sequences containing highly conserved motif. Four major steps that were carried out in order to fulfill the objectives were data mining of probe, screening, sequence alignment and phylogenetic tree construction. Probe sequences were screened in Pufferfish Genome Database through *Fugu* Genome Blast Search. Three sequences containing conserved motif of domains of reverse transcriptase gene in retrovirus were assembled as the preliminary result of this project. About 371 sequences were screened through the blast search. Continuous screening of all eight probe sequences enabled the assembly of 65 sequences. Out of that number, 18 sequences were selected for sequence alignment and phylogenetic tree construction. The sequences obtained through Blast Search were then aligned as it exhibits the highly conserved motif of all domains of reverse transcriptase gene in retroviruses. This was to detect the relationship between the established sequences within fish database and the newly gathered sequences. The Phylogenetic tree that was built to study the relationships between different species of organism. The phylogenetic tree revealed the grouping pattern of seven main genera of retroviruses and the relationships between the genera.

ABSTRAK

Tujuan utama kajian ini adalah untuk mengenalpasti kewujudan 'endogenus retrovirus' di dalam genom Pufferfish melalui penyaringan 'Pangkalan Data Genom Pufferfish'. Dalam kajian ini, bahagian gen *pol* dalam transkriptase berbalik retrovirus dikaji kerana ia mempamerkan motif terpelihara gen transkriptase berbalik retrovirus. Empat kaedah telah dijalankan dalam pada memenuhi objektif kajian ini. Ia adalah perangkaan jujukan '*probe*', penyaringan, penyusunan jujukan dan pembinaan pokok filogenetik. Jujukan retrovirus yang bertindak sebagai '*probe*' telah disaring didalam 'Pangkalan Data Genom Pufferfish' melalui Carian Blast Genom *Fugu*. Jujukan-jujukan yang mengandungi motif terpelihara gen transkripsi berbalik retrovirus telah dikumpulkan sebagai keputusan awal kajian ini. Penyaringan jujukan '*probe*' telah membolehkan sebanyak 371 jujukan disaring. Penyaringan berterusan kesemua jujukan-jujukan probe tersebut telah membolehkan sebanyaj 65 jujukan dipilih daripada 371 jujukan. Daripada bilangan tersebut, sebanyak 18 jujukan yang mengandungi motif terpelihara gen transkriptase berbalik dikumpulkan untuk penyusunan jujuan dan seterusnya untuk membina pokok filogenetik. Jujukan-jujukan yang didapati melalui carian blast kemudiannya disusun agar ia dapat mempamerkan motif terpelihara gen trankripsi berbalik. Kaedah ini dijalankan untuk mengenalpasti hubungan diantara jujukan yang telah sedia ada dalam pangkalan data ikan dengan jujukan yang baru ditemui. Kaedah penyusunan jujukan diikuti oleh pembinaan pokok filogenetik. Pokok filogenetik dibina untuk mempelajari hubungan diantara sepsis organisma yang berbeza. Pokok filogenetik yang dibina menerangkan corak pengkelasan 7 genera retrovirus endogenus dan hubungan di antara genera-genera tersebut.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The earlier science of gene screening includes methods such as proteins and peptides identification and quantification of amino acids in the lab. As time passes, those methods were adopted with the latest technology. Computer-based screening is one of the methods widely used to determine and identify the sequence of certain protein or nucleotides in a genome research. The most precise way to identify a protein is through its sequence. Protein identification can be performed by computer translation of DNA sequences.

The *Fugu rubripes* is one of at least 100 species of Pufferfish. The fish is also known with various names such as the blowfish, swellfish or ballonfish. Pufferfish is in the family of *Tetraodontidae* (order *Tetraodontiformes*) and it comes from Far East. Pufferfish contains a poison known as Tetrodotoxin from a symbiotic bacterium. This makes the pufferfish very poisonous yet they are classified as a delicacy in Japan.

The pufferfish is carefully prepared by removing the poisonous skin and organs and served as sushi , sashimi or cooked in broth.

databases are the most important source of information on proteins. Because of our incomplete understanding of the protein folding, the attempts to further predict the protein structure directly from sequence is hindered. By using this bioinformatics tool, the amino acid sequence of a protein, called primary, secondary and tertiary structure can be easily determined from the sequence on the gene that codes for it.

2.2 Genome Projects

In biology, the genome of an organism is a complete DNA sequence of one set of chromosomes. For every organism, from simple bacteria to complex human beings, the genetic information necessary for life is contained in its genome. The genome consists of tightly coiled threads of deoxyribonucleic acid (DNA) and is found in every cell of an organism. A DNA molecule consists of two strands wrapped around each other that resemble a spiral staircase. Each strand is a linear arrangement of repeating similar units called nucleotides, which are each composed of one sugar, one phosphate, and a nitrogenous base.

Each DNA molecule contains many genes that code for functional units of heredity. The sequence of a gene carries the information required for encoding proteins. The protein-coding instructions from the genes are transmitted indirectly through a transient intermediary molecule messenger known as messenger ribonucleic acid or mRNA. For the information within a gene to be expressed, a complementary RNA strand is produced from the DNA that serves as the template for protein synthesis (Campbell, 1992).

Evolution has conserved many of the DNA sequences which are being used in genes to code for certain genes that regulate gene expression. This has led to the finding of new genes by genome sequences comparisons between species.

The international effort in constructing detailed genetic and physical maps of the human genome has led the path to similar researches to map and sequence the genomes of a variety of organisms. Small genomes of some viruses and bacteria have been sequenced. Larger organisms that have been sequenced completely are the bakers' yeast (*Saccharomyces cerevisiae*), the roundworm (*Caenorhabditis elegans*) and the fruit fly (*Drosophila melanogaster*), (Barner and Gary, 1999).

Apart from that, another organism which has been studied in the field of genome research is mouse and rat. The study on mouse and rat has been started since the early 20th century. The mouse, with its short generation time, small size and plethora of phenotypic variants excelled as a tool of genetic investigations, especially after the conceptualization and creation of inbred strains (Barner and Gary, 1999).

The study of human and mouse genome complements each other because both organisms shows common set of genes. The DNA sequences of both human and mouse are also related. This makes it possible for human genes and proteins to function in mouse.

Human Genome Project (HGP) was launched in 1999 in the United States. The focus of HGP was to sequence the complete sequence of three billion DNA bases that compose the human genome (Nelson *et al.*, 2003). In 1992, the second generation of

human genetic map was developed (Collins *et al.*, 2003). The three billion projects were expected to take fifteen years but due to upcoming technologies, a finished version of human genome sequence was completed in 2003.

According to Nelson *et al.,* (2003), the entire human genome, only a small region of about three percentage appear to constitute the estimated 24179-87720 genes that are translated into necessary proteins. Whereas, the non-coding DNA known as junk DNA is not very obvious, yet they may play important role in packaging and gene expression.

Human Endogenous retroviruses (HERVs) are transmitted vertically through the germline and the overall HERVs constitute only about one percentage of human genome (Nelson *et al.*, 2003). Over twenty HERV families have been identified and a few of them have the ability to produce viral-like particles (Nelson *et al.*, 2003).

2.3 Pufferfish

Fugu Rubripes, Japanese Pufferfish is the first vertebrate which genome was completed after human. The *Fugu* Genome Project was initiated in 1989 in Cambridge by Sydney Brenner and his colleagues including Greg Elgar, Sam Aparicio, and Byrappa Venkatesh. The Pufferfish genome was decoded by an international consortium including the Joint Genome Institute in the United States, the Institute for Molecular and Cell Biology in Singapore and the Human Genome Mapping Resource Centre in Cambridge, UK. According to the *Fugu* Genome Project,

Fugu is the first vertebrate genome to be publicly sequenced and assembled (Davidson, 2000).

As vertebrates, fish and human not only share the defining characteristic of a backbone but also many anatomical and physiological similarities. According to the international *Fugu* Genome Consortium led by Sydney Brenner, the consortium chose a whole-genome shotgun sequencing approach that involved assembling over 3.7 million sequencing reads into large scaffolds.

A scaffold is a portion of the genome sequence reconstructed from end-sequenced from whole-genome shotgun clones (Verseweyveld, 2002). Scaffolds are composed of contigs and gaps. A contig is a continuous length of genomic sequence in which the order of bases is known to a high confidence level. Gaps occur where reads from the two sequenced ends of at least one fragment overlap with other reads in two different contigs. Since the lengths of the fragments are roughly known, the number of bases between contigs can be estimated. The goal of whole-genome shotgun assembly is to represent each genomic sequence in one scaffold; however, this is not always possible. One chromosome may be represented by many scaffolds (e.g., chlamydomonas) or just a single scaffold (e.g., human chromosome 19), (Verseweyveld, 2002).

This approach successfully covered more than 80% of the 365 megabase of the *Fugu* genome. Around 31059 predicted gene loci were found similar to the number of certified human genes. Gene loci represents one third of the genome even though it has the same number of genes and regulatory regions as the human, these elements are

embedded in only 365 million bases as compared to 3 billion bases that make up human DNA (Davidson, 2000).

Not only similarities, but there are also some differences between *Fugu* and human genome. The repetitive sequences of the *Fugu* genome account for 2.7%. It also has at least 40 different families of transposable elements, more than in human genome. Other than that, about a quarter of proteins in human doesn't have homology match with the *Fugu* such as genes encoding inflammatory cytokines and immune system components. *Fugu* also has potassium channel subunits and kinases, whereas human contains more zinc-finger and novel receptor families. Although the compact *Fugu* genome contains the same basic blueprint as the human genome, but the length of the sequence is only about one eight. This differentiation is due to the scarcity of so-called junk DNA, repeating sequenced that are not involved in making proteins in *Fugu*.

The compactness of *Fugu* genome can also be seen clearly by the small size of introns. These introns quote for three quarters which are less than 425 nucleotides long. However, further analysis of the genome showed the existence of quite a number of 'giant genes'. These 'giant genes' consist of extended introns which are spread over a distance, which are greater than human homologs.

So, the relative compactness of the *Fugu* genome makes the detection and analysis of both gene sequences and gene regulatory elements easier. The *Fugu* genome is ongoing and it promises further insights into genome structure and evolution.

2.4 Retrovirus

Retroviruses are classified in the *retroviridae* family and are distinct from all other RNA viruses because their genome is transcribed into DNA and integrated into the host genome. The enzyme responsible for generating the initial DNA copy of RNA is RNA-dependent DNA polymerase, now usually called reverse transcriptase (Adolph, 1996). Although all retroviruses have similar structure and genomic organization, there are large differences in their pathology and infectivity characteristics. Retroviruses are the only group of viruses known to transform cells and at the same time produce virus. The first retrovirus to be identified was the Rous Sarcoma virus (RSV). The Mouse Mammary Tumor Virus (MMTV) and the Gross mouse leukemia virus were the first oncogenic viruses of mammals recognized. (Kambol, 2003).

At first, retroviruses were observed as infectious virus particle and capable of transmission between cells (Lewin, 2000). So, the intracellular cycle which involves duplex DNA is thought of as a mean of reproducing the RNA virus. All retroviruses have approximately 80-100nm diameter, containing nucleoprotein cores. They are enveloped by plasma membrane which consists of virally coded glycoprotein derived from host cell.

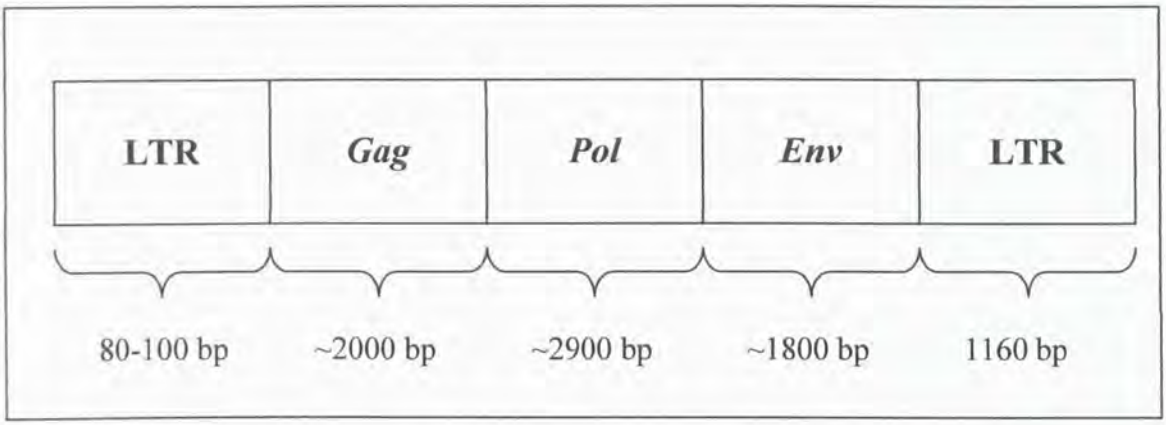

Figure 2.1 RNA form of virus

A typical retrovirus genome have three genes namely *gag-pol-env* as shown in Figure 2.1. The group specific antigen (*gag*) gene gives rise to the structure of internal protein components of the nucleoprotein core of the virion, such as matrix, capsid and nucleocapsid.

The *Gag* polyprotein of about 60,000 Daltons is encoded by a genome length of 38S mRNA. Later, the precursor protein is cleaved to form viral proteins which make up most of the virus particle. The reverse transcriptase (*pol)* gene codes for the function concerning the nucleic acid synthesis, recombination and integration to the host. Purified reverse transcriptase shows synthetic and degradative activities. The synthetic activity is aided by reverse transcriptase and degradative activities are characterized by ribonuclease H (RNase H). The *env* gene codes for the glycoprotein component of the envelope for virus infectivity. It is encoded by a sub genomic mRNA which has viral 5-` sequences spliced its 5-` end (Becker, 1979).

The full mRNA strand is translated to produce *Gag* and *Pol* polyproteins. The *Gag* is translated by reading the initial codon to the first termination codon. Whereas,

Pol region is expressed when the reading of the strand bypasses termination codon (Lewin, 2000).

2.4.1 Classification of Retrovirus

During the early stage, retroviruses were classified based on their morphological characteristics. In this study, retroviruses were observed as particles consist of a core known as nucleoid. This nucleoid is covered completely by an outer envelope. These particles have a spherical structure of about 100nm in diameter. Retroviruses were characterized in four groups using their morphological characteristics: Type A, B, C and D.

Type A retrovirus consist of a nucleoid surrounded by a double shell and these retroviruses lack the infectivity criteria. An example of this group is intracisternal particle A which can be found in rodent genome as retrovirus-like element. Type B particles are seen as nucleoid covered by an envelope. Type C particles are observed as a dense nucleoid in a sphere. These types are the major viruses which are being isolated from mammalian and avian taxon. Type D retroviruses are ring-shaped and usually found in primates. Examples of this group are Mason Pfizer Monkey Virus (MPMV) and simian AIDS virus (Kambol, 2003).

Later, retroviruses were characterized again based upon their pathogenecity and the way they infect the host. The three major groups are *oncovirinae, lentivirinae* and *spumavirinae* (Holland, 1992). *Oncovirinae* consist of viruses which are found in avian and mammalian species. Retroviruses in *lentivirinae* group are slow viruses and

they can be distinguished by the incubation time between infection and manifestation of disease. *Spumavirinae* are described as foamy viruses because of their ability to produce a spontaneous 'foamy' in tissue culture cells (Kambol, 2003).

However, the most recent approach was based on the sequence data and genome arrangement. Retroviruses are classified into seven genera; alfaretrovirus, beteretrovirus, gammaretrovirus, epsilonretroviruses, deltaretrovirus, lentivirus and spumaretrovirus. In this new approach, few new features were exhibited, for example type B and D retroviruses are combined in one genus. Not only that, a new genus for fish has been recognized and known as epsilonretrovirus.

Fish viruses are categorized in a new genus called epsilonretrovirus because severel accessory genes and a new tRNA primer have been identified in their complex genome structure. An example of virus in this group is known as Walleye dermal sarcoma virus (WDSV). This virus has a new unique $tRNA^{His}$ primer and three additional accessory genes (*orf-a, orf-b and orf-c*). Viruses in this particular genus are only found in piscine hosts and Walleye fish viruses are the major member in this genus (Kambol, 2003).

2.4.2 Retrovirus Life Cycle

The life cycle of a retrovirus begins by the binding of the viral envelope glycoprotein to a specific receptor on the cell surface of the host (Adolph, 1996). After binding, the virion penetrates the cell membrane by endocytosis or by direct fusion. In the

cytoplasm, the virion is partially degraded. The uncoated viral particle is then passes the cytoplasm and enters the nucleus.

During the migration of the viral particle to the nucleus, the single-stranded viral RNA is converted to double-stranded viral DNA. The converted DNA form is called a provirus and this provirus is the one which enters the nucleus. The provirus has complete long terminal repeats (LTR) on both end of DNA strand. The LTR have sequences necessary for integration and the regulation of transcription. The LTR is approximately 600bp in length and contains a duplicated transcription enhancer element followed by promoter region and a polyadenylation (pol A) signal (Adolph, 1996). Following process would be the integration of the provirus into the host genome. By integrating, retroviruses become part of the genomic information which will be passed to daughter cells (Adolph, 1996).

After undergo a few processes such as transcription and translation, new retroviral particles are produced. This particle is covered by capsid proteins. The envelope of the retroviral (virion) is formed at the plasma membrane and the virion is then released from the host cell by 'budding'.

Retroviruses are divided into two groups based upon the type of transmission and also categorized depending upon to their host range (Schupbach, 1981). As most retroviruses are infectious agents, they transmit in two ways. One is by horizontally, meaning that these retroviruses are transmitted from one individual to another from somatic cells. These retroviruses are known as exogenous. The other type of transmission is by vertically, from parent to offspring using germ-line transmission.

REFERENCES

Adolph, K. W., Ph.D., 1996. *Viral Genome Methods*. CRC Press, London.

Attwood, T. K. & Parry-Smith, D.J. 1999. *Cell and Molecular Biology: Introduction to Bioinformatics*. Prentice Hall, UK.

Barner, M. B. & Gray, I. C. 1999. *Bioinformatics for Geneticists*. Genetic Bioinfo, Glaxosmithklie Pharmaceuticals, UK.

Becker, Y., Hebrew U & Hadar, J. 1979. *Replication of Viral and Cellular Genomes*. Martimus Nijhoff Publishing, Boston.

Campbell, A. 1992. *Annual Review of Genetics, vol. 26.* Stanford University, Stanford, Palo AHo, California, USA.

Collins, F. s., Green, E. D., Guttmacher, A.E. & Guyer, M.S. 2003. A Vision for the Future of Genomics Research. *Naturel*. **(422)** : 835-847.

Dale, J. W. & von Schantz, M. 2002. *From Genes to Genomes: Concepts and Applications of DNA Technology*. University of Surrey, UK., John Wiley and Sous Ltd., UK.

Domingo, E., Webster, R. & Holland, J. 1999. *Origin and Evolution of Viruses. Academic Press.*

Davidson, H., 2000. *Genomic Sequence Analysis of Fugu rubripes CFTR and Flanking Genes in a 60 kb Region Conserving Synteny with 800 kb of Human Chromosome 7. Genome Research,* **(10):** 1194-1203. Medical Research Council Human Genetics Unit, Edinburgh, UK.

Gribskov, M. & Devereux, J. 1991. *Sequence Analysis Primer*. Stockton Press, UK.

Herniou, E., Martin, J.,Miller, K., Cook, J., Wilkinson, M and Tristem, M. 1998. Retroviral Diversity and Distribution in Vertebrates. *Journal Of Virology. (72):5955-5966,* Department of Biology, Imperial College, United Kingdom.

Holland, J. J., 1992. *Genetic Diversity of RNA Viruses: Current Topics in Microbiology 176 and Immunology.* Springer-Verlag publisher, USA.

Johnson, E. W. and Coffin, J. M., 1999. *Constructing Primate Phylogennies From Ancient Retrovirus Sequences, Inaugural Articles. (96): 10254-10260.* Department Of Molecular Microbiology, Boston.

Kambol, R., 2003. *Distribution and Evolution of Endogenous Retrovirus Within Amphibian and Piscine Hosts*. Unpublished Ph.D Thesis. Department of Biological Sciences,Imperial College, United Kingdom.

Lewin, B. 2000. *Genes VI* . Oxford University and Cell Press.

Neafsey, D. E. and Palumbi, S. R., 2003. Genome Size Evolution in Pufferfish: *A* Comparative Analysis of Diodontid and Tetraodontid Pufferfish Genomes. *Genome Research. (13): 821-830,* Department of Organismic and Evolutionary Biology, Cambridge, California, USA.

Nelson, P. N., Carnegie, P. R., Martin, J., Ejtehadi, H. D., Hooley, P., Roden, D., Rowland-Jones, S., Warren, P., Astley, J. & Murray, P. G. 2003. *Demystified -Human Endogenous Retrovirus. Molecular Pathology*. **(56)**: 11-18.

Schupbach, J. 1989. Current Topics in Microbiology 142 and Immunology: *Human Retrovirology Facts and Concepts*. Springer-Verlag publisher, USA.

Sensen, C. W., 2002. *Essentials of Genomics and Bioinformatics*. (E.D.). WILEY-VCH Verlag Gmbh, Weinheim, Germany.

Shen, C. H., and Steiner, L. A., 2004. Genomic Structure and Thymic Expression of an Endogenous Retrovirus in Zebrafish. Journal of Virology. (***78***): 899-911, Department of Biology, Cambridge

Verseweyveld L., 2002. *Fugu pufferfish DNA sequence offers short cut to Human Genome.* Virtual Medical World. Joint Genome Institute.