Phylogenetics & Evolutionary Biology

Thomas, J Phylogen Evolution Biol 2015, 3:1
http://dx.doi.org/10.4172/2329-9002.1000145

Review Article Open Access

Homo Sapiens: Epidermal Growth Factor Receptor (DNA Data Mining)

Evans Andrew Thomas*

Institute of Biology, Darwin's Research Institute of Biological Sciences, Manchester, UK

***Corresponding author:** Evans Andrew Thomas, Institute of biology, Darwin's Research Institute of Biological Sciences, Manchester, UK, Tel: 0040-492- 5783; E-mail: evansat1004@gmail.com

Rec date: Jan 15, 2015, **Acc date:** Feb 18, 2015, **Pub date:** Feb 27, 2015

Abstract

Epidermal growth factor plays an important role in the Regulation of the cell growth, proliferation and differentiation by binding to its receptor EGFR. IT has a receptor called EGFR it leads to the rapid internalization and releases the Lysozyme. It reduces the cell signalling, but it is possible only after the endocytosis only. Epidermal growth factor has the capability to transfer the epithelial cell. During the carcinomas in human epidermal growth factor and its ligand show high expression in that condition

Keywords: Data mining; Epidermal growth factor; Homo sapiens

Introduction

NCBI: NCBI means national centre for biotechnology information (NCBI). NCBI is the part in the United States of America. The National centre for Biotechnology information has advances the science and health by providing the access to biomedical and genetic information. Research in molecular and genetic process that control the diseases and health are done by National Library of Medicine (NLM). With the help of NCBI researchers select a tool that helps in finding the sequences. An important aspect that is science primer which helps in easy reading of introduction to many science topics like Bio-informatics, Molecular modelling, Genomic mapping and Molecular genetics. NCBI plays an important role in the sequencing of genome, protein structures and gap between sequences act. Sequencing of genome, gap between sequences and structural information stored in PUBMED is growing rapidly. Other than sequencing structural determination takes a lot of time and limited in its application [1].

Bioinformatics

The term Bioinformatics was first introduced in early 1990s because of Human Genome Project. Bioinformatics is the application of biology and information technology in the field of molecular biology. Due to the advances in the resources and technology Human Genome Project has impact on current research that is taking place in the world. Some of the important applications of Bioinformatics are

Sequence analysis

Genome annotation

Analysis of gene expression

Protein expression

Molecular medicine [2].

Databases

In the bioinformatics two types of databases are important they are

Protein databases

DNA databases

These two data bases analyse the biological databases and then format by functional and sequence information. Based on sequence analysis biological databases are of two types

Primary sequence database

Secondary sequence database

Re-engineering works are doing for making the biological data more easy and efficient to use. By this it is easy to get data from the different sources and increase the power and capability of biological resources [2].

Relational database management system plays an important role in the implementing of molecular biological sequences by using the ODBC and JDBC for data exchange. By using these methods many problems get disappeared [3].

Blast

BLAST means Basic Local Alignment Search Tool. Between the sequences the Blast find the region of similarities. It compares the sequence databases to protein or nucleotide sequence and calculates the matches. Blast also can be used to identify the phylogenetic Relation between the sequence and in gene families. Blast programme are mainly used to searching Protein and DNA database sequence similarities [4]. Same as in the Human Genome Project all the genes have been mapped and sequenced and the information is stored in gene bank. This information will be available by an accession number. Accession number contains the information regards an individual gene. The accession number BC037558 contains the information about the Homo Sapiens gene that is epidermal growth factor receptor pathway substrate 15 like-1, mRNA (cDNA clone) partial cds. It comes under fallowing taxonomical units – Eukaryota; Metazoa; Primates; Mammalia; Eutheria; Chordata; Craniata; Vertibreta; Eutheria; Euteleostomi; Hominidea; Homo [5].

Accession Number

Accession number is the unique number that is given to an individual for finding the nucleotide sequence in Bioinformatics. With the help of the accession number has to identify the DNA and protein sequence and data is to be recorded [6].

Epidermal Growth Factor

Epidermal growth factor plays an important role in the Regulation of the cell growth, proliferation and differentiation by binding to its receptor EGFR. IT has a receptor called EGFR it leads to the rapid internalization and releases the Lysozyme. It reduces the cell signalling, but it is possible only after the endocytosis only [7]. Epidermal growth factor has the capability to transfer the epithelial cell. During tha carcinomas in human epidermal growth factor and its ligand show high expression in that condition [8].

EPS15L1 epidermal growth factor receptor 15 like1:

Homo sapiens

Official Full name: Epidermal growth factor receptor.

Primary source: HGNC: 26634

Taxon ID: 9606

Organism: Homo sapien

Gene type: protein coding

Common name: man.

Lineage (full): Eukaryota; Metazoa; Primates; Mammalia; Eutheria; Chordata; Craniata; Vertibreta; Eutheria; euteleostomi; Hominidea; Homo

Also known as: EPS15R, EPS15L1.

Gene Structure

The gene is present in the chromosome 19. In the chromosome 19 it is present in the region of 19.13.11. EPS15L1 is the protein that encodes the gene (Figures 1 and 2) [9].

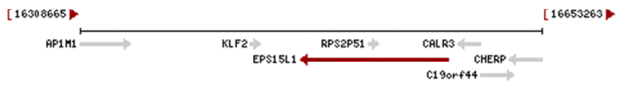

Figure 1: Gene structure: Chromosome 19 - NC_000019.9.

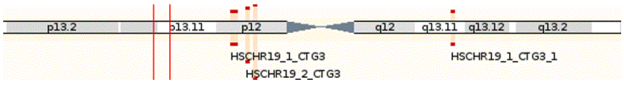

Figure 2: Location of epidermal growth factor in the chromosome.

Open Reading Frames (ORF)

In the DNA sequence ORF are used to identify the protein coding in DNA sequence. For example in the DNA sequence with equal percentage of each nucleotide a stop codon is expected and in gene prediction for prokaryotes look for stop codon fallowed by Open Reading Frame. In translation it is important to know that which nucleotide starts translation and when stops these are called open reading frame.

After the completion of the DNA sequence it is important to find the open reading frame (ORF). Each region of DNA contains six open reading frames three in one direction and three in other direction. Of these six only one frame with large length is used for translation. An open reading frame starts with an atg (Met) and ends with a stop codon (taa, tag, or tga).

To the accession number BC037558 totally there are Twenty five open reading frames, in these 12 with positive open reading frames and 13 with negative open reading frames. In the total 25 frames +3 frames is largest in length with 1806 base pairs and +2 is smallest with 129 base pairs. BY using the http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi we can find the open reading frames (Figures 3 and 4) [10].

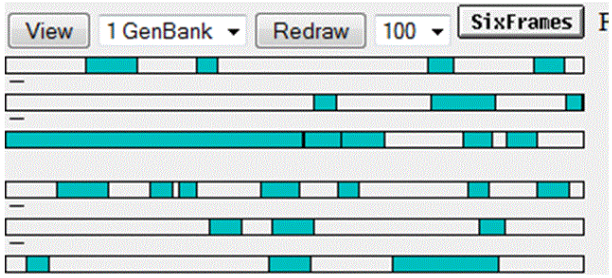

Figure 3: Open reading frames.

The above given is ORF frames and its length. From these ORF structure known that 25 open reading frames contain several ORFs [11,12].

Introns and Exons

Introns

Introns are the sequence that is present with in the gene. An intron is a nucleotide sequence that is located in the middle of the gene sequences.

Exons

Exons are the sequences that are present in DNA and are used to code amino acids in the protein. Exons are placed in to the mRNA to code for the amino acids.

ENSEMBL is the site that is used to find the information about the introns and exons with in a gene.

Frame	from	to	Length
+3	6	1811	1806
-3	2355	3005	651
+2	2591	2983	393
+1	490	807	318
-1	314	631	318
+3	2046	2306	261
-2	1624	1881	258
-3	1608	1859	252
-1	1562	1795	234
+3	1824	2045	222
-1	3233	3433	201
-2	1246	1440	195
+1	3214	3402	189
+3	3048	3236	189
+3	2787	2966	180
-2	2884	3042	159
+1	2575	2730	156
-1	881	1024	144
+2	1880	2017	138
-3	132	266	135
-1	2030	2158	129
-1	2816	2941	126
+1	1165	1290	126
-1	1058	1171	114
+2	3413	3513	102

Figure 4: ORFs length.

Gene: ENSG0000127527 of Homo sapiens epidermal growth factor receptor pathway substrate 15 like-1 has two transcripts (Figures 5 and 6).

Transcript ID: ENSR00000248070

Transcript ID: ENST00000455140

Discussion on transcripts ID:

Transcript ID ENST00000248070:

Exons: 23

Transcript Length: 2774 base pairs.

Translation length: 864 residues.

CCDS: CCDS32944

Transcript ID ENST00000455140:

Exons: 24

Transcript length: 3054 base pairs

Translation length: 910 residues

In both the transcripts Exons are present in the first transcript (ENST00000248070) there are 23 exons were as in the next transcript (ENST00000455140) there are 24 exons.

CCDS Report for consensus CDS:

Translation (864 aa):

MAAPLIPLSQQIPTGNSLYESYYKQVDPAYTGRVGASEAALFLK KSGLSDIILGKIWDLADPEGKGFLDK

QGFYVALRLVACAQSGHEVTLSNLNLSMPPPKFHDTSSPLMVT PPSAEAHWAVRVEEKAKFDGIFESLLP

INGLLSGDKVKPVLMNSKLPLDVLGRVWDLSDIDKDGHLDRD EFAVAMHLVYRALEKEPVPSALPPSLIP

PSKRKKTVFPGAVPVLPASPPPKDSLRSTPSHGSVSSLNSTGSLS PKHSLKQTQPTVNWVVPVADKMRFD

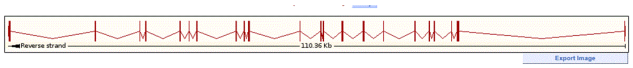

Figure 5: Introns.

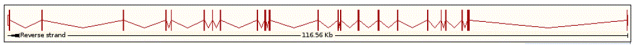

Figure 6: Exons.

EIFLKTDLDLDGYVSGQEVKEIFMHSGLTQNLLAHIWALADTR QTGKLSKDQFALAMYFIQQKVSKGIDP

PQVLSPDMVPPSERGTPGPDSSGSLGSGEFTGVKELDDISQEIA QLQREKYSLEQDIREKEEAIRQKTSE

VQELQNDLDRETSSLQELEAQKQDAQDRLDEMDQQKAKLRD MLSDVRQKCQDETQMISSLKTQIQSQESD

LKSQEDDLNRAKSELNRLQQEETQLEQSIQAGRVQLETIIKSLK STQDEINQARSKLSQLHESRQEAHRS

LEQYDQVLDGAHGASLTDLANLSEGVSLAERGSFGAMDDPFK NKALLFSNNTQELHPDPFQTEDPFKSDP

FKGADPFKGDPFQNDPFAEQQTTSTDPFGGDPFKESDPFRGSA TDDFFKKQTKNDPFTSDPFTKNPSLPS

KLDPFESSDPFSSSSVSSKGSDPFGTLDPFGSGSFNSAEGFADFSQ MSKPPPSGPFTSSLGGAGFSDDPF

KSKQDTPALPPKKPAPPRPKPPSGKSTPVSQLGSADFPEAPDPF QPLGADSGDPFQSKKGFGDPFSGKDP

FVPSSAAKPSKASASGFADFTSVS

Colour variation in the amino acids shows the variations in the regions.

There are 864 amino acids in the translation length of the transcript ID ENST0000248070 it has CCDS consensus sequence which is shown above. The variation in the colour shows the different in the regions [13].

Multiple Sequence Alignment

Multiple sequence alignment is done when it is need to compare homologous sequence it is important tool in bio informatics. For doing multiple sequences various programme are used like CLUSTAL W and T-coffee they will show various results of the alignments for the sequences [14].

The Protein FASTA sequence of my accession number BC037558 is taken from the BLAST results the sequence is copied and saved. Along with my accession number other four sequences are taken from four different accession numbers and saved. For doing multiple sequence

alignment we have used the website http://www.ebi.ac.uk in these website we will find the CLUSTAL W2 click on that and upload the sequences we will get the results in few minutes. The results will appeared as follows (Figure 7) [15].

SeqA	Name	Length	SeqB	Name	Length	Score
1	gi|10864047|ref|NP_067058.1|	864	2	gi|119604951|gb|EAW84545.1|	910	99.0
1	gi|10864047|ref|NP_067058.1|	864	3	gi|194383118|dbj|BAG59115.1|	910	99.0
1	gi|10864047|ref|NP_067058.1|	864	4	gi|297703989|ref|XP_002828907.1|	910	99.0
1	gi|10864047|ref|NP_067058.1|	864	5	gi|297276397|ref|XP_001113811.2|	910	98.0
2	gi|119604951|gb|EAW84545.1|	910	3	gi|194383118|dbj|BAG59115.1|	910	99.0
2	gi|119604951|gb|EAW84545.1|	910	4	gi|297703989|ref|XP_002828907.1|	910	99.0
2	gi|119604951|gb|EAW84545.1|	910	5	gi|297276397|ref|XP_001113811.2|	910	99.0
3	gi|194383118|dbj|BAG59115.1|	910	4	gi|297703989|ref|XP_002828907.1|	910	99.0
3	gi|194383118|dbj|BAG59115.1|	910	5	gi|297276397|ref|XP_001113811.2|	910	98.0
4	gi|297703989|ref|XP_002828907.1|	910	5	gi|297276397|ref|XP_001113811.2|	910	98.0

Figure 7: Score table. The above Figure shows the results of multiple sequence alignment of score table. Here five organisms sequence is taken in to the consideration.

CLUSTAL W2 2.1: Multiple sequence alignment.

Fallowed by these we can watch the cladogramor Phylogenetic tree which gives the results of evolutionary distance between the five organisms.

Phylogram

At the end length difference is given (Figure 8) [16].

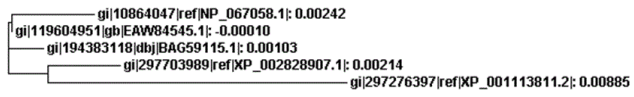

Figure 8: Phylogenetic tree.

Phylogenetics

Phylogenetics is process of placing the organisms in to groups or classes based on their similarity in evolutionary relationships. In case of molecular genetics the classification is based on the comparison on DNA or Amino acid sequences (Figure 9).

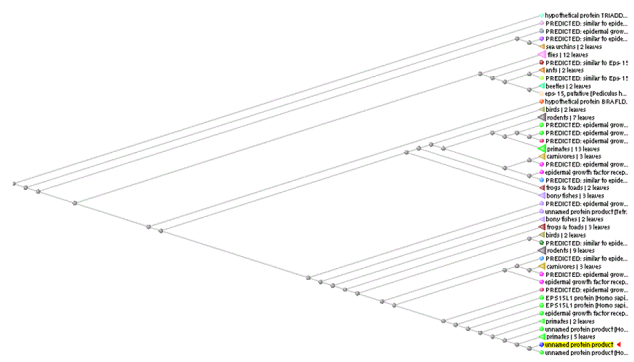

Figure 9: Blast tree view widget.

From the above tree we can see the similarity difference from these it is clear that query sequences has good similarity in between them. It can be explained as evolution and these are due to the mutations between the genes [17].

Conclusion

Epidermal growth factor plays an important role in the Regulation of the cell growth, proliferation and differentiation by binding to its receptor EGFR. EPS15L1 is the protein that encodes the gene.

Phylogenetic tree gives the results of evolutionary distance between the five organisms. From the phylogenetic tree we can see the similarity difference based on that it is clear that query sequences have good similarity in between them. It can be explained as evolution and these are due to the mutations between the genes.

References

1. Teixeira C, Gomes JRB, Gomes P, Maurel F (2011) Viral surface glycoprotein, gp120 and gp41, as potential drug targets against HIV-1: Brief overview one quarter of a century past the approval of zidovudine, the first anti-retroviral drug. Eur J Med Chem 46: 979-992.
2. Sanabani S, KleineNeto W, Kalmar EM, Diaz RS, Janini LM, et al. (2006) Analysis of the near full length genomes of HIV-1 subtypes B, F and BF recombinant from a cohort of 14 patients in Sao Paulo, Brazil. Infect Genet Evol 6: 368-377.
3. Castro NE, Perez LM, Burton GF, Crandall KA (2012) The evolution of HIV: inferences using phylogenetics. Mol Phylogenet Evol 62: 777-792.
4. Wainberg MA, Jeang KT (2008) 25 years of HIV-1 research - progress and perspectives. BMC Med 6: 31.
5. http://www.ebi.ac.uk/Tools/services/Web/toolresult.ebi
6. http://www.ncbi.nlm.nih.gov/blast/treeview/treeview.cgi
7. Pandhare J, Dash C (2011) A prospective on drug abuse-associated epigenetics and HIV-1 replication. Life Sci 88: 995-999.
8. Arhel N, Kirchhoff F (2010) Host proteins involved in HIV infection: new therapeutic targets. Biochim Biophys Acta 1802: 313-321.
9. Gelderblom H (1997) Fine Structure of HIV and SIV.
10. Lu K, Heng X, Summers MF (2011) Structural determinants and mechanism of HIV-1 genome packaging. J Mol Biol 410: 609-633.
11. Kypr J, Mrazek J, Reich J (1989) Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. Biochim Biophys Acta 1009: 280-282.
12. Khamsri B, Murao F, Yoshida A, Sakurai A, Uchiyam T, et al. (2006) Comparative study on the structure and cytopathogenic activity of HIV Vpr/Vpx proteins. Microbes and Infection 8: 10-15.
13. Palmisano L, Vella S (2011) A brief history of antiretroviral therapy of HIV infection: success and challenges. Ann Ist Super Sanita 47: 44-48.
14. Kitamura Y, Lee YM, Coffin JM (1992) Nonrandom integration of retroviral DNA in vitro: effect of CpG methylation. Proc Natl Acad Sci USA 89: 5532-5536.
15. Beer BE, Bailes E, Sharp PM, Hirsch VM (1999) Diversity and Evolution of Primate Lenti viruses. Human retroviruses and AIDS 460-474.
16. Wang X, Ragupathy V, Zhao J, Hewlett I (2011) Molecules from apoptotic pathways modulate HIV-1 replication in Jurkat cells. BiochemBiophys Res Commun 414: 20-24.
17. Chhatbar C, Mishra R, Kumar A, Singh SK (2011) HIV vaccine: hopes and hurdles. Drug Discov Today 16: 948-956.