

# **GENE CONTENT PHYLOGENY OF FLAVIVIRIDAE FAMILY**

**AMITABH GUPTA (040002)**

**KUNAL PUNJRATH (040088)**

**Supervisor  
Dr. SUJATA MOHANTY**



**May 2008**

**Submitted in partial fulfillment of the Degree of  
Bachelor of Technology**

**DEPARTMENT OF BIOTECHNOLOGY  
JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY  
UNIVERSITY, NOIDA**

## TABLE OF CONTENTS

| <b>Chapter<br/>No.</b> | <b>Topics</b>   | <b><i>Page<br/>No.</i></b> |
|------------------------|---|----------------------------|
|                        | Declaration   | 3                          |
|                        | Certificate from the Supervisor   | 4,5                        |
|                        | Acknowledgement   | 6                          |
|                        | Summary   | 7                          |
| <b>1.</b>              | <b>Introduction</b>   | 8                          |
| <b>2.</b>              | <b>Classification &amp; Phylogeny of Viruses</b>  | 12                         |
|                        | 2.1 Introduction  | 12                         |
|                        | 2.2 Classification of viruses   | 13                         |
|                        | 2.3 Phylogeny among viruses   | 15                         |
| <b>3.</b>              | <b>The Phylogenetic profiling technique</b>   | 16                         |
|                        | 3.1 Introduction  | 16                         |
|                        | 3.2 Materials & Methods   | 17                         |
|                        | 3.3 Result & Discussion   | 21                         |
| <b>4.</b>              | <b>The COG technique &amp; Phylogenomic method (Background)</b>                                   | 23                         |
| <b>5.</b>              | <b>Implementing the COG technique – Creating phylogenetic trees (Materials &amp; Methodology)</b> | 28                         |
|                        | 5.1 Materials   | 28                         |

|           |  |       |
|-----------|--|-------|
|           | 5.2 Methodology  | 30    |
|           | 5.2.1 COG Construction: COG Stringency Number                  | 30    |
|           | 5.2.2 The Gene Content Tree                                    | 31    |
|           | 5.2.3 Reciprocal Tree of COG Phylogenetic Profiles             | 34    |
| <b>6.</b> | <b>The COG technique – Outcomes (Results &amp; Discussion)</b> | 35    |
|           | 6.1 Results  | 35    |
|           | 6.1.1 Gene Content tree  | 38    |
|           | 6.1.2 Reciprocal tree  | 40    |
|           | 6.1.3 Polyprotein Neighbor Joining tree                        | 41    |
|           | 6.2 Discussion   | 42    |
|           | 6.2.1 Gene Content tree  | 43    |
|           | 6.2.2 Reciprocal tree  | 44    |
| <b>7.</b> | <b>Conclusion</b>  | 46    |
|           | References   | 47    |
|           | Appendices   |       |
|           | A. List of VOGs and Supplementary material                     | 51    |
|           | B. Phylogeny of Flaviviridae                                   | 52    |
|           | Publications   | 56    |
|           | Brief Bio-data of Students                                     | 57,60 |

## **DECLARATION**

We hereby declare that the project titled “**GENE CONTENT PHYLOGENY OF FLAVIVIRIDAE FAMILY**” submitted in partial fulfillment of B.Tech in Biotechnology has been carried out by us at JIITU,Noida under the guidance of **Dr. Sujata Mohanty**, Senior Lecturer,Department of Biotechnology, JIITU,Noida

We further declare that the project work or any part thereof has not been previously submitted for any degree or diploma in any university.

Date: 5<sup>th</sup> May, 2008

Amitabh Gupta  
Enrollment No.-040002

Kunal Punjrath  
Enrollment No.-040002

Place: JIITU,Noida

## **CERTIFICATE**

This is to certify that the work titled “**GENE CONTENT PHYLOGENY OF FLAVIVIRIDAE FAMILY**” submitted by “**Amitabh Gupta**” in partial fulfillment for the award of degree of Bachelor of Technology of Jaypee Institute of Information Technology University, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Dr. Sujata Mohanty

Senior Lecturer, Department of Biotechnology, JIITU

5<sup>th</sup> May 2008

## **CERTIFICATE**

This is to certify that the work titled “**GENE CONTENT PHYLOGENY OF FLAVIVIRIDAE FAMILY**” submitted by “**Kunal Punjrath**” in partial fulfillment for the award of degree of Bachelor of Technology of Jaypee Institute of Information Technology University, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Dr. Sujata Mohanty

Senior Lecturer, Department of Biotechnology, JIITU

5<sup>th</sup> May 2008

## ACKNOWLEDGEMENTS

We extend our warm and sincere thanks to our project supervisor **Dr. Sujata Mohanty** who was a staunch supporter and motivator throughout the project.

Right from the inception of this project work, our mentor guided us till the very end in the true sense of the word. She always came up with innovative ways and creative terms thus also helping me to instill and enhance the quality of creative thinking within us.

We would also like to express our gratitude to this alma mater JIITU, Noida for providing proper resources as and when required such as an all time internet facility and an excellent library.

We would also take this opportunity to thank Dr. Kamal Rawal for making the necessary computer laboratory arrangements for us.

One person who needs special mention is Professor Indira Gosh, Director Bioinformatics Centre, University of Pune, for sowing the initial idea for this project.

Also, we would like to thank Michael G. Montague and Clyde A. Hutchison from the University of North Carolina for providing us with the Perl scripts used in COG construction.

Hence without giving a warm thanks to all of them who made this project work a reality our work would be incomplete.

AMITABH GUPTA

KUNAL PUNJRATH

5<sup>TH</sup> MAY 2008

## SUMMARY

With complete sequencing of 2540 viral genomes, the value of genome sequencing projects will be determined by how the resulting mass of sequence data are analyzed. Whole genome phylogeny has been called one of the major open problems of comparative genomics. Here, we identified the cluster of orthologous group for 47 completely sequenced viral genomes of Flaviviridae family. Flaviviruses are responsible for causing many human encephalitic and hemorrhagic diseases such as tick-borne encephalitis, dengue fever, West Nile, and yellow fever.

We identified 35 clusters of orthologous groups related to viral proteins (VOG). Each VOG represented a family of gene products conserved across several Flaviviridae genomes. These families were defined without using an arbitrary threshold criterion based on sequence similarity.

The VOG data were used to construct whole-genome phylogenetic trees based on gene content. These parsimony trees agree well with trees based on neighbor-joining method and are robust when tested by bootstrap analysis. The VOG data also were used to construct a reciprocal tree that clustered genes with similar phylogenetic profiles. This clustering may give clues to genes with related functions or with related histories of acquisition and loss during evolution of viruses belongs to Flaviviridae family.

Amitabh Gupta

Kunal Punjraht

Dr. Sujata Mohanty



# CHAPTER 1

## INTRODUCTION

---

The term ‘phylogenomics’ indicates the construction of phylogeny based upon complete genome data. The initial step in every phylogenomic approach is to which genes are to be compared between species [1]. Orthology among species provide the raw material for this, three types of orthology definitions are available:

- 1- Pairwise orthology
- 2- Cluster orthology
- 3- Tree-based orthology

First two methods uses sequence similarity scores to define the orthologous groups of genes. Pairwise orthology is defined between only two species (ex- bi-directional best hits). Cluster orthology is the natural extension of pair wise orthology to more than two species (ex- Clusters of Orthologous Groups)

Tree-based orthology uses sequence similarity scores as well as analyze phylogenetic tree of homologous group of genes to obtain orthologous relationship.

There are four methods widely used for phylogenomics analysis –

- 1- Gene content
- 2- Superalignment
- 3- Superdistance
- 4- Supertree

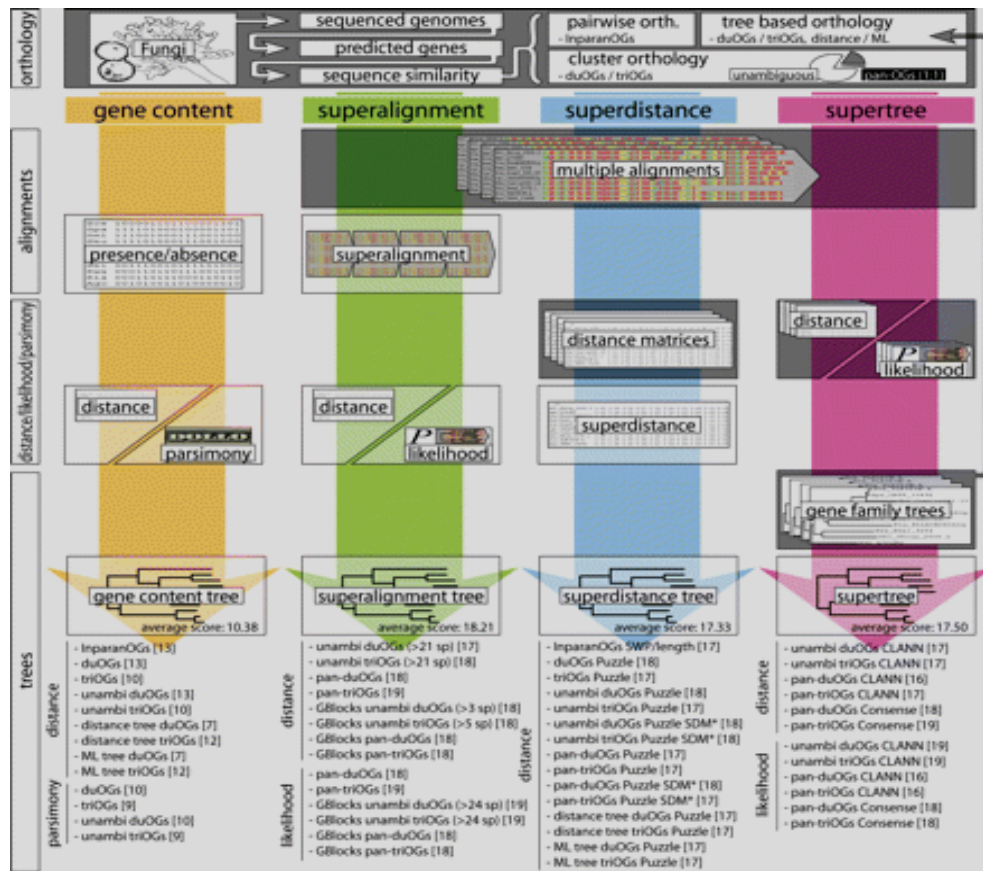
Last three methods are sequence based phylogenomics methods; the first step is to make multiple alignments for every orthologous group (OG).

In the superalignment approach, concatenating the multiple alignments for each orthologous group to form a superalignment combines the phylogenetic information. Subsequently, conventional phylogenetic inference methods can be used to transform the alignment into a phylogeny.

In the superdistance approach, first distance matrixes of all the gene families are calculated, and then phylogenomic distance between two species is then defined as the average distance between all the shared gene families.

In supertree approach, phylogenetic trees have been composed for all gene families then all the multiple gene trees combines into a single phylogenomic tree by an integration step.

In gene content approach, sequence information is only used for defining the orthologous groups (OGs) that treat as a gene content data. To infer a phylogenomic tree from gene content data, a binary character matrix indicating the presence or absence of the OGs in all the species can be treated in the same way as a multiple sequence alignment [1] [2].



**Fig. 1** Making phylogenomic trees[1]. Before starting tree inference, OGs are defined (top row). Phylogenomics follows the steps of phylogenetics, from multiple alignment through distance, likelihood or parsimony to the reconstruction of a phylogeny. Integrating separate phylogenetics for each gene family (gray boxes) to phylogenomics (white boxes) can be done at every one of these steps. This defines the phylogenomic approach: gene content (after OG definition), superalignment (after multiple alignment), superdistance (after distance calculation) or supertree (after reconstruction of gene family trees). The phylogenomic trees we reconstructed are listed at the bottom, the number between square brackets indicates the number of target nodes that the tree recovered correctly.

Here, we applied gene content phylogeny on Flaviviridae family presently comprise of genus Flavivirus, Pestivirus and Hepacivirus consists of 49 single-strand, positive-sense RNA viruses. A number of the flaviviruses are associated with human disease. For example, dengue virus, present as four serotypes (DENV-1 to DENV-4), is prevalent in over 100 countries and 2.5 million people live in dengue-endemic areas,

while yellow fever virus (YFV) affects 200,000 persons annually, with a case fatality rate of around 20 percent.

They are well studied with a large percentage of identified or partially identified genes. They are very divergent in many aspects, but, like cellular organisms, are known to have a conserved core of genes that encode many of the required functions of the virus life cycle including RNA polymerase, major capsid protein and core protein.

## CHAPTER 2

# CLASSIFICATION & PHYLOGENY OF VIRUSES

---

### 2.1 INTRODUCTION

We are well into the Age of Genomics, and genome sequences are providing a flood of valuable new information about (among other things) the evolutionary histories and relationships of contemporary organisms. This is nowhere more true, than among the viruses, where the economical sizes of the genomes mean that large numbers of genomic sequences can be determined and compared; as a consequence, there are now large databases cataloging the sequence relationships and inferred phylogenetic relationships among such notable groups as the Herpesviruses or strains of HIV. Viral genome sequences are also remarkably diverse, producing diverse proteins.

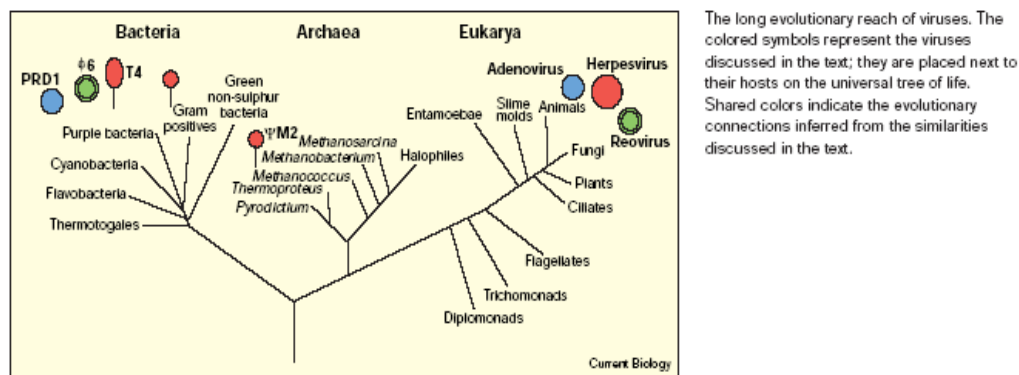


Fig 2.1 Tree of Life [14]

Our approach was confined to various taxonomic groups of viruses, and hence a need arises to understand the relevance and classification of viruses.

## 2.2 CLASSIFICATION OF VIRUSES

1. Host and organ affected – identify viruses by type of host infected or the type of tissue that was infected (ex. dermatropic if infect skin, neurotropic if infect nerve tissue)

2. **International committee on taxonomy of viruses (ICTV)** – establishes the rules for classifying viruses; established 1966; <http://www.ncbi.nlm.nih.gov/ICTV/>

a. **Virus family** – family is the highest taxonomic category; usually distinguished on basis of nucleic acid type, capsid shape, envelope, and virus size

b. **Virus genus** – genera also used but people are slow to accept

c. **Viral species** – a group of viruses that share the same genome and the same relationships with organisms

There are 2600 various properties which determine the groups.

3. **Baltimore classification** – The Baltimore classification is a classification system which groups viruses into families depending on their type of genome (DNA, RNA, single-stranded (ss), double-stranded (ds) etc.) and their method of replication (Fig. 2.2).

a) **Type I:** dsDNA viruses (Herpesviridae, Poxviridae, Adenoviridae and Papovaviridae)

b) **Type II:** ssDNA viruses (Circoviridae and Parvoviridae)

- c) **Type III:** dsRNA viruses (Reoviridae and Birnaviridae)
- d) **Type IV:** positive sense ssRNA viruses (Astroviridae, Caliciviridae, Coronaviridae, Flaviviridae, Picornaviridae, Arteriviridae and Togaviridae)
- e) **Type V:** negative sense ssRNA viruses (Arenaviridae, Orthomyxoviridae, Paramyxoviridae, Bunyaviridae, Filoviridae and Rhabdoviridae)
- f) **Type VI:** positive sense ssRNA viruses that replicate through a DNA intermediate (Reverse transcriptase) (Retroviridae)
- g) **Type VII:** dsDNA viruses with ssRNA intermediates (Reverse transcriptase)(Hepadnaviridae)

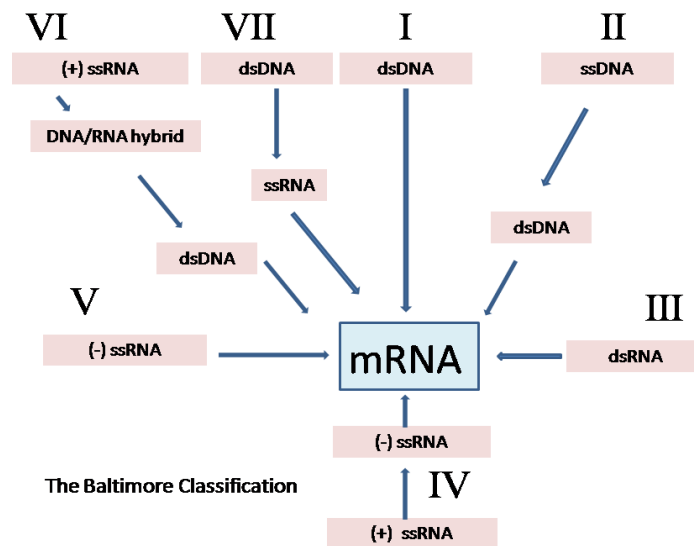


Fig 2.2 *The Baltimore Classification*

## 2.3 PHYLOGENY AMONG VIRUSES

Phylogeny is the description of biological relationships, usually expressed as a tree.

Phylogeny states as topology of the relationships based on the classification according to similarity of one or more sets of characters, or on a model of evolutionary process.

Molecular phylogenetics is the use of molecular sequences to construct evolutionary trees to study a family of related sequences that we know evolved from a common ancestor, and we want to know in which order these sequences diverged from one another. There are two approaches to deriving phylogenetic trees: *phenetic approach* is based on similarity; *cladistic approach* is based upon **genealogy** [1].

Due to selecting viruses of different taxonomical hierarchy, we were able to generate phylogenetic profile of complete protein set for closely as well as distantly related viruses and get an idea of which proteins were conserved during evolutionary process. Also this helped us in constructing the Gene content tree for flaviviridae family. The phylogeny of Flaviviridae family is depicted in Appendix B.



## CHAPTER 3

# THE PHYLOGENETIC PROFILING TECHNIQUE

---

### 3.1 INTRODUCTION

This method detects proteins that participate in a common structural complex or metabolic pathway. Proteins within these groups are defined as *functionally linked*. The Underlying hypothesis is that functionally linked proteins evolve in a correlated fashion, and, therefore, they have homologs in the same subset of organisms [11].

In short, if two proteins have homologs in the same subset of fully sequenced organisms, they are likely to be functionally linked. By exploiting this property we intended to do functional annotation of uncharacterized proteins.

We made phylogenetic profile of Picornavirale family using the proteomes of viruses in this family.

A Profile is a string with **n** entries: n corresponds to number of proteomes .In a profile the presence of a homolog (ORF) to a given protein in the n<sup>th</sup> proteome is represented with an entry of **unity** at nth position. If no homolog is found the entry is **zero**.

Proteins are **clustered** according to the similarity of their phylogenetic profiles. Similar profiles show a correlated pattern of inheritance and, by implication,

**functional linkage.** The method predicts that the functions of uncharacterized proteins are likely to be similar to characterized proteins within a cluster.

### 3.2 MATERIALS & METHODS

- **Proteome Data Sets:** Whole proteomes for 27 distinct members of the family *Picornaviradae* were downloaded from the NCBI database (<http://www.ncbi.nlm.nih.gov>). The virus list and its acronyms used in the analysis are given in Table 3.1

**Table 3.1 Viruses and abbreviations**

| <b>Virus name</b>                       | <b>Abbrv.</b> |
|---|---------------|
| Avian encephalomyelitis virus           | AvEnm         |
| Bovine enterovirus                      | BoEnt         |
| Duck hepatitis virus 1                  | DuH1          |
| Duck picornavirus TW90A                 | DuPTW         |
| Encephalomyocarditis virus              | Enmcd         |
| Equine rhinitis B virus 1               | EqRB1         |
| Foot-and-mouth disease virus type A     | F&MA          |
| Foot-and-mouth disease virus type Asia1 | F&MA1         |
| Foot-and-mouth disease virus type C     | F&MC          |
| Foot-and-mouth disease virus type O     | F&MO          |
| Foot-and-mouth disease virus type SAT1  | F&MSA1        |
| Hepatitis A virus                       | HA            |
| Human enterovirus 100                   | HuEnt1        |
| Human enterovirus A                     | HuEntA        |
| Human enterovirus B                     | HuEntB        |
| Human enterovirus C                     | HuEntC        |
| Human enterovirus D                     | HuEntD        |
| Human rhinovirus A                      | HuRA          |
| Human rhinovirus B                      | HuRB          |
| Poliovirus                              | Poli          |
| Porcine enterovirus A                   | PoEntA        |
| Porcine enterovirus B                   | PoEntB        |
| Porcine teschovirus                     | PoTe          |
| Saffold virus                           | Saff          |
| Simian enterovirus A                    | SiEntA        |
| Simian picornavirus 1                   | SiPi1         |
| Theilovirus                             | Thei          |

- **BLASTP**: Uses the BLAST algorithm to compare an amino acid query sequence against a protein sequence database. The BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) programs have been designed for speed to find high scoring local alignments. BLAST uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity (Altschul et al., 1990 [15]). Because of its design for speed, there may be a minimal loss of sensitivity to distant sequence relationships.
- **PSI-BLAST** (**P**osition-**S**pecific **I**terated **B**LAST) is a tool that produces a position-specific scoring matrix constructed from a multiple alignment of the top-scoring BLAST responses to a given query sequence [16] [17]. This scoring matrix produces a profile designed to identify the key positions of conserved amino acids within a motif. When a profile is used to search a database, it can often detect subtle relationships between proteins that are distant structural or functional homologues. These relationships are often not detected by a BLAST search with a sample sequence query.
- **Programme written in perl script:**
  - **formatdb.pl**: For creating blast output files for each virus's proteome with respect to reference organism's proteome.
  - **parser.pl**: For parsing the significant information from blast output files and creates parsing files for each virus except reference organism.
  - **hits.pl**: For creating result files for each virus from parsing files.
  - **db\_insert.pl**: Creates the phylogenic profile from the result files.

The following steps, enlist the procedure followed for creating phylogenetic profiles

- 1) First, we downloaded protein set of viruses of different taxonomical hierarchy from ENTREZ genome database [32] in FASTA format.
- 2) Stored these proteome as a .txt file specific for each virus in input files folder, made another .txt file consist a list of names for all viruses selected with there four letter specific code.
- 3) With the help of BLAST algorithm, we compared each virus' input file with reference organism's (*Equine rhinitis A virus*) input file, which are listed in virus list text file. This process was executed by the programme *formatdb.pl*.
- 4) After execution of programme we got blast output files for the listed viruses, with the help of these output files, we generate query files and result files for each virus. In query file, we parsing the query for each blast hit and its corresponding significant similar protein match, if significant match didn't found it parsing the only query pattern. We decided the **significant value** ( $E \leq 10e-5$ ) for parsing in *parser.pl*.
- 5) After parsing the value in query files we generated result files using *hits.pl* containing significant and non-significant hits in binary form (0/1).
- 6) The proteome database containing the phylogenetic profile was created by executing the *dbinsert.pl* script.
- 7) After creation of database we can identify the phylogenetic profile of each protein. We indicate the presence of protein in all viruses by value 1 or absent

by value 0. In this database we generated the phylogenetic profile of 100 against 10 proteins, which represent the string of 55 and 74 entries. The complete database can be accessed from the supplementary material CD.

Further, the same approach was repeated using PSI-BLAST search because:

- The initial blastp search is not very informative.
- There was a need to identify the distant members of some protein families.
- To create a model PSSM in one database and use that in another to find better matches.

We used blastpgp in PSI-BLAST mode to perform a standard blastp search against the target database. The sequences found in the first round are used to build a PSSM, which is then used in the next round of search. As explained in the paper by the Altschul et al and Schäffer et al, the traditional BLAST algorithm was implemented using an  $A \times A$  substitution matrix, where  $A$  is the alphabet size currently set at 28. PSI-BLAST instead uses a  $Q \times A$  matrix, where  $Q$  is the length of the query sequence. At each position, the score for a letter depends on the position with respect to the query and the letter in the subject sequences [17].

Blastpgp can also store and reuse the PSSM through `-C` and `-R` parameters, respectively. To reuse a stored checkpoint file, the exact same query has to be used. The motivation for the `-C` and `-R` parameters is to support changing databases, so that the PSSM model built by searching one database and can be reused to search another database. We used this to switch from a larger protein database to a smaller database as shown below.

```
blastpgp -i $query -d viral_db -h .001 -j 0 -C pssm_$qn
```

```
blastpgp -i $query -d $dbname -R pssm_$qn -o output/$qn/$db.out
```

The first run saves the PSSM that was used in the last round of search against the protein database *viral\_db*. The second run uses this saved PSSM to perform profiling against specific datasets.

### 3.3 RESULTS & DISCUSSION

Proteins involved in **related cellular mechanism** like **viral division**, which is coordinated with other cell cycle events such as **nucleic acid synthesis** that leads to **protein synthesis**, were found to be the most conserved.

| Table 3.2 Picornaviradae Vs Equine rhinitis A virus |                     |               |                   |
|---|---------------------|---------------|-------------------|
| Accession IDs                                       | Protein names       | BLAST results | PSI-BLAST results |
| NP_740379.1   | 2C                  | 27 (0 bit)    | 27 (0 bit)        |
| NP_740383.1   | 3D (RNA polymerase) | 27 (0 bit)    | 27 (0 bit)        |
| NP_740374.1   | VP2                 | 25 (2 bit)    | 27 (0 bit)        |
| NP_740375.1   | VP3                 | 24 (3 bit)    | 27 (0 bit)        |
| NP_653075.1   | Polyprotein         | 18 (9 bit)    | n/p               |
| NP_653076.1   | Polyprotein         | 18            | n/p               |
| NP_740382.1   | 3C (proteinase)     | 18            | n/p               |
| NP_740376.1   | VP1                 | 9             | 26 (1 bit)        |
| NP_740373.1   | VP4                 | 8             | 9                 |
| NP_740372.1   | Leader (proteinase) | 5             | 5                 |
| NP_740378.1   |                     | 5             | n/p               |
| NP_740377.1   |                     | 0             |                   |
| NP_740380.1   |                     | 0             |                   |
| NP_740381.1   |                     | 0             |                   |

From Table 3.2 it can be seen that 2C protein gave the maximum number of hits in the Picornaviradae family. This protein is significant because

- It is a 329 amino acid-protein that is essential for viral RNA synthesis and may perform multiple functions.

- It blocks transport of protein from ER to Golgi apparatus.

Apart from finding conservancy of few proteins in the proteomes of viruses, we were not able to establish functional co-relation among proteins.

The following factors prevented us from establishing functional linkage among proteins of picornaviradae family:

- 1) Lack of published material on the successful implementation of phylogenetic profiling in the same manner as done by us on viruses. This technique in its current form might be unsuitable for viruses due to the following reasons –
  - a. Smaller genome sizes of viruses compared to multi-cellular organisms. This made it impossible to find similar profile protein clusters.
  - b. High mutation rate in viruses makes it difficult to find homologs.
  - c. Difficulty in selecting a reference organism as evolution in viruses has been incomprehensible.
- 2) Lack of published material on viral translation cycle inside its host for all members of picornaviradae family.

## **CHAPTER 4**

# **THE COG TECHNIQUE AND PHYLOGENOMIC METHOD**

## **(BACKGROUND)**

---

Phylogenomics is correctly defined by tree construction through integrating vast amount of whole genome information. There are many approaches available for phylogenomics analysis like gene content, super alignment and super distance method. Gene conservancy among species provides the strong phylogenetic signal for evolutionary studies and tree building methods.

Phylogenomics tree based upon gene content are calculated from present or absent profiles using either distance or parsimony. In other phylogenomics tree construction methods are calculated from distance matrix using multiple sequence alignment.

Multiple sequence alignment is the frequently used for detection of conserved pattern in gene families shared by related species but the relationship between genes from different genomes are represented by a system of homologous families that include both orthologs and paralogs.

Here, we present a systematic comparison of two important factors of phylogenomic inference: the orthology approach and the level of integration of phylogenetic information to a genomic scale in case of viruses belongs to Flaviviridae family.

The relationships between genes from different genomes are naturally represented as a system of homologous families that include both orthologs and paralogs. The correct identification of orthologs (i.e. genes whose deepest relationship represents a



speciation event [1]) is crucial for reconstruction and interpretation of phylogenetic trees; and for addressing the following questions: How many common genes are shared by different species? What is the extent of the core of genes that shares a common history? Which genes underwent duplication, were lost, or horizontally transferred between different lineages? Most of the known methods used for detection of orthologs are based on sequence similarity and genome-specific best hits [2].

Today, a widely used method to identify sets of orthologs from a set of  $n$  species is the reciprocal best BLAST hit method (e.g., [2],[4],[7]). The method requires strong conservative relationships among the orthologs so that if a gene from species 1 selects a gene from species 2 as a best hit when performing a BLAST search with genome 1 against genome 2, then the gene 2 must in turn select gene 1 as the best hit when genome 2 is searched against genome 1. For a set of  $n$  species the reciprocal BLAST hit method requires the presence of all pairwise reciprocal connections between all species as depicted on Figure 4.1.

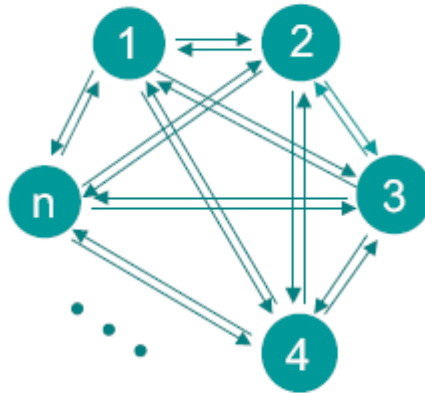


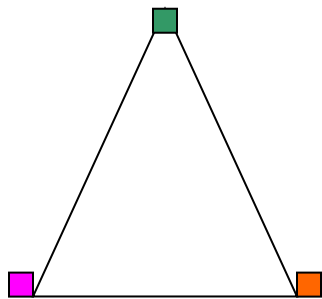
Fig 4.1 *Pairwise reciprocal connections*

**The reciprocal best BLAST hit method.** Circles represent genes from  $n$  different taxa; arrows signify best BLAST hit relationship.

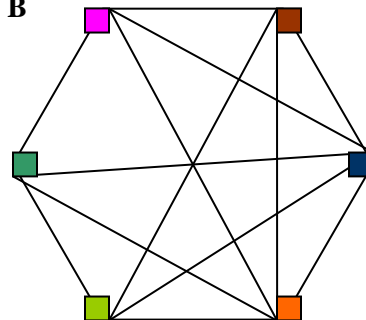
If gene duplications occurred in each of the given two clades subsequent to their divergence, only a many to many relationships will adequately describe orthologs, and accordingly, detection of the highest similarity will not result in identification of the complete set of orthologs. Given the existence of one-to-many and many-to-many orthologous relationships, *Tatusov et.al* [2]. redefined the task of identifying orthologs as the delineation of clusters of orthologous groups (COGs). Each COG consists of individual orthologous genes or orthologous groups of paralogs from three or more phylogenetic lineages. In other words, any two proteins from different lineages that belong to the same COG are orthologs. Each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events.

Here, we derived COG (clusters of orthologous group) for constructing the whole genome phylogenetic tree based on gene content data in case of viral genomes. In the Clusters of Orthologous Groups (COG) [13] strict reciprocity is replaced by a triangular best Blast hits relationship. First, triangles forming one-side circular best BLAST hits are constructed, and then triangles with common sides are merged together to form a cluster.

**A**

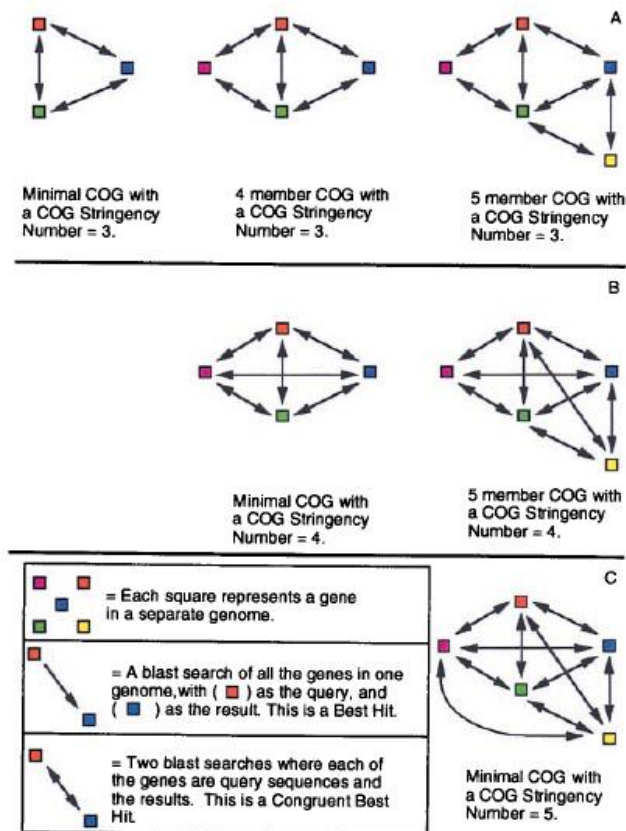


**B**



**Fig 4.2** Examples of COGs. Solid lines show symmetrical BeTs, Colored Square depict the genes in different genomes (A) Congruent BeTs form a triangle, the minimal COG. (B) Symmetrical BeTs arranged in triangle with common side.

One of the most important modifications in COG technique was the addition of variable CSN. In Tatusov *et al.* [2], triangles were constructed from congruent BeTs. This created COGs with a CSN of 3, based on the merging of triangles with common sides (Fig 4.2A). A tetrahedron consisting of four gene products from separate genomes with congruent BeTs to each other forms the basis for a COG with a CSN of 4. Larger COGs with CSN of 4 are formed by merging tetrahedrons with common faces (Fig 4.2B). Similarly, COGs of higher CSN result from the merging of higher-order constructs. The minimal COG for CSN = 5 is diagrammed in Fig 4.3 The CSN can be increased up to the number of genomes in the set.



**Fig 4.3** How the CSN functions to increase the rigor of the test for addition of new members to a COG as well as construction of minimal COGs. (A) COGs as constructed in ref. 4 by merging of triangles with common sides (CSN = 3). (B) A similar set of COGs (CSN = 4) as in A, but in each case one more congruent BeT is required for construction of each minimal COG and for addition of members (in this case by merging of tetrahedrons with common faces). (C) Shown is the minimal COG at CSN = 5.

Increasing the CSN increases the stringency of the test for adding a new member to a COG (Fig 4.3). Each member of a COG must have CSN-1 congruent BeTs with other members. Higher CSNs are especially important for use in small genomes, such as those of viruses, because the chance of an erroneous congruent BeT, resulting from chance sequence similarity rather than homology, increases dramatically if there are few genes in each genome. In case of viruses, clusters of orthologous groups for viral related proteins denoted as VOG.

### **IMPORTANT TERMS**

**Parsimony:** Possible trees are compared and each is given a score that is a reflection of the minimum number of character state changes (e.g., amino acid substitutions) that would be required over evolutionary time to fit the sequences into that tree. The optimal tree is considered to be the one requiring the fewest changes (the most parsimonious tree).

**Bootstrapping:** Alignment positions within the original multiple sequence alignment are resampled and new data sets are made. Each bootstrapped data set is used to generate a separate phylogenetic tree and the trees are compared. Each node of the tree can be given a bootstrap percentage indicating how frequently those species joined by that node group together in different trees. Bootstrap percentage does not correspond directly to a confidence limit.

## CHAPTER 5

# IMPLEMENTING THE COG TECHNIQUE – CREATING PHYLOGENETIC TREES (MATERIALS AND METHODOLOGY)

---

### 5.1 MATERIALS

**Viruses:** The 47 viral proteomes used in the analysis were obtained in FASTA format from NCBI.

(<http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/11050.html>) The viral proteomes were sorted manually into separate files for the sake of analysis.

**Viral Clusters of Orthologous Groups (VOGs):** 35 Clusters of Related *Viral* Proteins for flaviviradae family were identified using a modified COG approach mentioned by Montague and Hutchison [7], described later.

#### **Programme Used:**

- **BLASTP:** Uses the BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) algorithm [15] to compare an amino acid query sequence against a protein sequence database. The BLAST program used to find high scoring local alignments were obtained from the NCBI ftp server [35].

- **PHYLIP** (Version 3.67): PHYLIP, the Phylogeny Inference Package, is a package of programs for inferring phylogenies (evolutionary trees) [21]. PHYLIP is freely available from <http://evolution.genetics.washington.edu/phylip.html>. The following programme were used from the package:
  - **Seqboot** – Reads in a data set, and produces multiple data sets from it by bootstrap resampling.
  - **Mix** – Estimates phylogenies by some parsimony methods for discrete character data with two states (0 and 1). Allows use of the Wagner parsimony method, the Camin-Sokal parsimony method, or arbitrary mixtures of these.
  - **Consense** – Computes consensus trees by the majority-rule consensus tree method, which also allows one to easily find the strict consensus tree.
  - **Protdist** – For calculating the distance matrix needed for neighbor joining tree.
  - **Neighbor** – For constructing neighbor joining tree.
- **TreeView** (1.6.6): The outtree files obtained from PHYLIP were used to create trees shown in here.

## 5.2 METHODOLOGY

### 5.2.1 COG Construction: COG Stringency Number (CSN)

Every Flaviviridae Proteome in the system was assigned a number, and every protein in those proteomes was assigned a number. This was formatted with delimiter characters (M and X) – M1X4, M2X4, M3X4, for example, would be the 4th protein of proteome 1, 2, and 3 respectively. A best hit blast search result was formatted:

[Blast-query] f [Blast-Result], for example M2X14fM7X130 reads when the 14th protein of the 2nd proteome has blasted against the proteins of the 7<sup>th</sup> proteome; the 130th protein was the best hit. A congruent best hit (BeT), as described by Montague and Hutchison [7] and in the Tatusov et. al. original COG paper is a situation where Best Hits point back at each other. In this system, for the above example, that would mean the following are both true –M2X14fM7X130, and M7X130fM2X14.

For construction of COGs, triangles were constructed from congruent BeTs. This created COGs with a CSN of 3, based on the merging of triangles with common sides (Fig. 1A). A tetrahedron forms the basis for a COG with a CSN of 4. It consists of four gene products (proteins) from separate genomes/proteomes with congruent BeTs to each other. Larger COGs with CSN = 4 are formed by merging tetrahedrons with common faces (Fig. 1B). Similarly, COGs of higher CSN result from the merging of higher-order constructs.

This work was performed by invoking the following Perl scripts (provided by Montague and Hutchison [7]):

- **triagnew5.pl** takes a list of congruent best hits separated into 2 files "congruent" and "gruentcon" and makes cogs of various stringencies out of them. "congruent" and "gruentcon" should both be formatted, one best hit per line, like so:

M2X9fM3X29

M3X3fM6X1

M7X17fM12X19

"gruentcon" should have all the opposite best hits to the ones in "congruent". Only best hits (BeTs) that are congruent and thus are matched to their opposite should be in either file.

- **addtocoresh.pl** adds members to each cog that have the requisite number of best hits to the COG, but where those best hits are not congruent.
- **Cogcompare.pl** is intended for comparing COGs of one stringency to stringency. In this way it is possible, for example, to work out which COG at stringency 7 is a subset of the proteins in a COG at stringency 4.

### 5.2.2 The Gene Content Tree

The steps involved for the construction of tree by maximum parsimony using VOGs of CSN=3 are as follows:-



- 1) Creating Data Matrix: To each flaviviradae proteome, for each COG, a one or a zero was assigned. One signified that the proteome had at least one member in the relevant COG, and zero signified no members. The resulting COG membership data matrix is illustrated in Fig. Each row of this matrix is a binary sequence describing the COG content of one of the proteomes, with a length equal to the total number of COGs. This work was done by following perl scripts –
  - **formatdb\_cog.pl** performed the BLASTP comparisons for each protein in every viral proteome against each protein in all the COGs.
  - **parser\_cog.pl** parsed the significant hits ( $E = 1e - 5$ ) from blast output files into corresponding text files.
  - **hits\_cog.pl** converted significant and non-significant hits into binary form (0/1) and stored them in result files.
  - **Input\_phylip.pl** creates the binary data matrix considering all the result files. This matrix is further used in the discrete character methods in PHYLIP.
- 2) Resampling the input data: The data matrix or the discrete character matrix is bootstrapped to produce 500 replicates using **Seqboot** program provided in the PHYLIP package.

```

D:\Study Material\phytip 3.67\exe\seqboot.exe
Bootstrapping algorithm, version 3.67
Settings for this run:
D      Sequence, Morph, Rest., Gene Freqs?  Discrete Morphology
F      Use factors information?              No
J      Bootstrap, Jackknife, Permute, Rewrite? Bootstrap
%      Regular or altered sampling fraction? regular
B      Block size for block-bootstrapping?  1 <regular bootstrap>
R      How many replicates?                  500
W      Read weights of characters?           No
X      Read mixture file?                   No
N      Read ancestors file?                 No
S      Write out data sets or just weights?  Data sets
0      Terminal type <IBM PC, ANSI, none>?  IBM PC
1      Print out the data at start of run    No
2      Print indications of progress of run Yes

Y to accept these or type the letter for one to change

```

- 3) Creating Most Parsimonious trees: Wagner Parsimony is used to generate best trees using the 500 replicates. This is done with the help of the **Mix** program in PHYLIP.

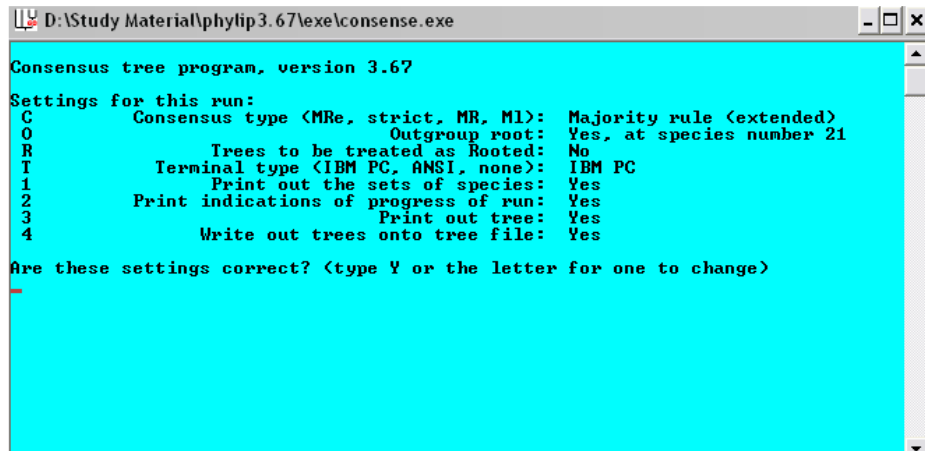
```

D:\Study Material\phytip 3.67\exe\mix.exe
Mixed parsimony algorithm, version 3.67
Settings for this run:
U      Search for best tree?                 Yes
X      Use Mixed method?                    No
P      Parsimony method?                    Wagner
J      Randomize input order of species?     Yes (seed = 673, 47 times)
O      Outgroup root?                       Yes, at species number 21
I      Use Threshold parsimony?              No, use ordinary parsimony
A      Use ancestral states in input file?   No
W      Sites weighted?                      No
M      Analyze multiple data sets?           Yes, 500 data sets
0      Terminal type <IBM PC, ANSI, none>?  IBM PC
1      Print out the data at start of run    No
2      Print indications of progress of run Yes
3      Print out tree                       Yes
4      Print out steps in each character     No
5      Print states at all nodes of tree    No
6      Write out trees onto tree file?       Yes

Are these settings correct? <type Y or the letter for one to change>

```

- 4) The Consensus Tree: The outfile from **Mix** program provides 50,000 trees, which are used to create the majority-rule consensus tree by invoking the **Consense** program in PHYLIP.

A screenshot of a Windows command prompt window titled "D:\Study Material\phytip3.67\exe\consense.exe". The window displays the "Consensus tree program, version 3.67" and its settings. The settings are listed as follows: Consensus type (MRe, strict, MR, M1) is set to "Majority rule (extended)"; Outgroup root is "Yes, at species number 21"; Trees to be treated as Rooted is "No"; Terminal type (IBM PC, ANSI, none) is "IBM PC"; Print out the sets of species is "Yes"; Print indications of progress of run is "Yes"; Print out tree is "Yes"; and Write out trees onto tree file is "Yes". The prompt "Are these settings correct? (type Y or the letter for one to change)" is shown at the bottom, with a red cursor on the line below it.

```
Consensus tree program, version 3.67
Settings for this run:
C      Consensus type (MRe, strict, MR, M1):  Majority rule (extended)
0      Outgroup root:                        Yes, at species number 21
R      Trees to be treated as Rooted:        No
I      Terminal type (IBM PC, ANSI, none):    IBM PC
1      Print out the sets of species:         Yes
2      Print indications of progress of run:  Yes
3      Print out tree:                       Yes
4      Write out trees onto tree file:        Yes
Are these settings correct? (type Y or the letter for one to change)
_
```

### 5.2.3 Reciprocal Tree of COG Phylogenetic Profiles

Each column in the COG data matrix (see Results, Fig 6.1) is a binary sequence that describes the distribution of members of a particular COG among the flaviviradae proteomes. These aligned sequences, termed phylogenetic profiles, were used to derive a reciprocal tree that clusters the COGs rather than the proteomes.

The reciprocal tree was constructed for of the 35 COGs at CSN = 3 by the Maximum Parsimony Method [7]. The steps involved in here were same as the steps (2) to (4) described earlier for construction of gene content tree. The only change made was that 100 boot replicates were used instead of 500.

## **CHAPTER 6**

### **THE COG TECHNIQUE – OUTCOMES (RESULTS AND DISCUSSION)**

---

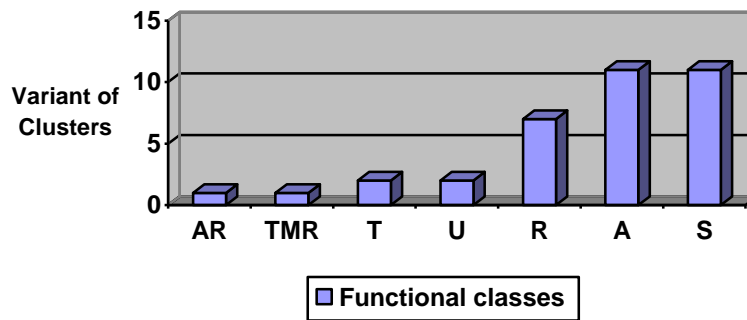
#### **6.1 RESULTS**

We applied COG (clusters of orthologous group) technique on viral genomes based upon orthologous relationship between viral species belonging to genus Flavivirus, Pestivirus and Hepacivirus of Flaviviridae family. This gave us VOG's (viral related clusters of orthologous group) which were utilized in establishing functional phylogenetic relationship by creating whole-genome phylogenomic trees based upon gene content data. We also produced a reciprocal tree that groups together VOGs that have a similar pattern of presence or absence in the studied viral genomes.

**VOG Construction:** We constructed VOGs by using reciprocal best-hit (BeT) approach or Congruent BeTs. These congruent BeTs further used in creation of triangle, which is an elementary minimal VOG. Each created VOG based on merging of triangle with common side. The created COG with a CSN = 3 (congruent stringency number), have congruent best hits found in three viral genomes.

For all 47 genomes of Flaviviridae family, 35 VOGs were identified at CSN = 3. (For complete list of VOGs created, refer Appendix A) These identified VOGs broadly classified into seven categories –

**Graph 6.1 VOGs of Flaviviridae**



In the graph 6.1, functional classes are delineated by following broad categories -  
A - Auxiliary proteins, S - Structural Proteins, R - Regulation of cellular mechanism  
U - Unknown proteins, T - RNA replication, transcription and modification,  
TMR - Replication, Movement and regulatory proteins.

These 35 clusters comprise of 355 proteins of viruses belonging to the Flaviviridae family.

**Table 6.1 Viruses and abbreviations**

| Virus name                       | Abbrv. |
|----------------------------------|--------|
| Alkhurma                         | Alkh   |
| Apoi                             | Apoi   |
| Border disease1                  | Bor1   |
| Bovine viral diarrhea1           | Bov1   |
| Bovine viral diarrhea2           | Bov2   |
| Bussuquara                       | Buss   |
| Cell fusing agent                | Cell   |
| Classical swine fever            | Clas   |
| Culex flavivirus                 | Cule   |
| Dengue1                          | Den1   |
| Dengue2                          | Den2   |
| Dengue3                          | Den3   |
| Dengue4                          | Den4   |
| Entebbe bat                      | Ente   |
| HepatitisC2                      | HeC2   |
| HepatitisC3                      | HeC3   |
| HepatitisC4                      | HeC4   |
| HepatitisC5                      | HeC5   |
| HepatitisC6                      | HeC6   |
| HepatitisC                       | HepC   |
| HepatitisG                       | HepG   |
| HepatitisGB A                    | HGBA   |
| HepatitisGB B                    | HGBB   |
| Iguape                           | Igua   |
| Ilheus                           | Ilhe   |
| Japanese encephalitis            | Japa   |
| Kamiti River                     | Kami   |
| Karshi                           | Kars   |
| Kokobera                         | Koko   |
| Langat                           | Lang   |
| Louping ill                      | Loup   |
| Modoc                            | Modo   |
| Montana myotis leukoencephalitis | MMLe   |
| Murray Valley encephalitis       | MuVE   |
| Omsk hemorrhagic fever           | Omsk   |
| Pestivirus Giraffe1              | PGi1   |
| Pestivirus Reindeer1             | PRe1   |
| Powassan                         | Powa   |
| Rio Bravo                        | RioB   |
| Sepik                            | Sepi   |
| St Louis encephalitis            | StLE   |
| Tamana bat                       | Tama   |
| Tick-borne encephalitis          | Tick   |
| Usutu                            | Usut   |
| West Nile                        | West   |
| Yellow_fever                     | Yell   |
| Yokose                           | Yoko   |

**Fig 6.1 Data Matrix**

| VOG No<br>Virus | 001 | 002 | 003 | 023 | 026 | 033 | 034 | .....265 |
|-----------------|-----|-----|-----|-----|-----|-----|-----|----------|
| Alkh            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Apoi            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Bor1            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Bov1            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Bov2            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Buss            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Cell            | 1   | 1   | 1   | 0   | 0   | 1   | 1   |          |
| Clas            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Cule            | 1   | 1   | 1   | 0   | 0   | 1   | 1   |          |
| Den1            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Den2            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Den3            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Den4            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Ente            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| HeC2            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| HeC3            | 1   | 1   | 1   | 0   | 0   | 0   | 0   |          |
| HeC4            | 1   | 1   | 1   | 0   | 0   | 0   | 0   |          |
| HeC5            | 1   | 1   | 1   | 0   | 0   | 0   | 0   |          |
| HeC6            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| HepC            | 1   | 1   | 1   | 0   | 0   | 0   | 0   |          |
| HepG            | 1   | 1   | 1   | 0   | 0   | 0   | 0   |          |
| HGBA            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| HGBB            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Igua            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Ilhe            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Japa            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Kami            | 1   | 1   | 1   | 1   | 0   | 1   | 1   |          |
| Kars            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Koko            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Lang            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Loup            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Modo            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| MMLe            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| MuVE            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Omsk            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| PGi1            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| PRe1            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Powa            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| RioB            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Sepi            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| StLE            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Tama            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Tick            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Usut            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| West            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Yell            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |
| Yoko            | 1   | 1   | 1   | 1   | 1   | 1   | 1   |          |

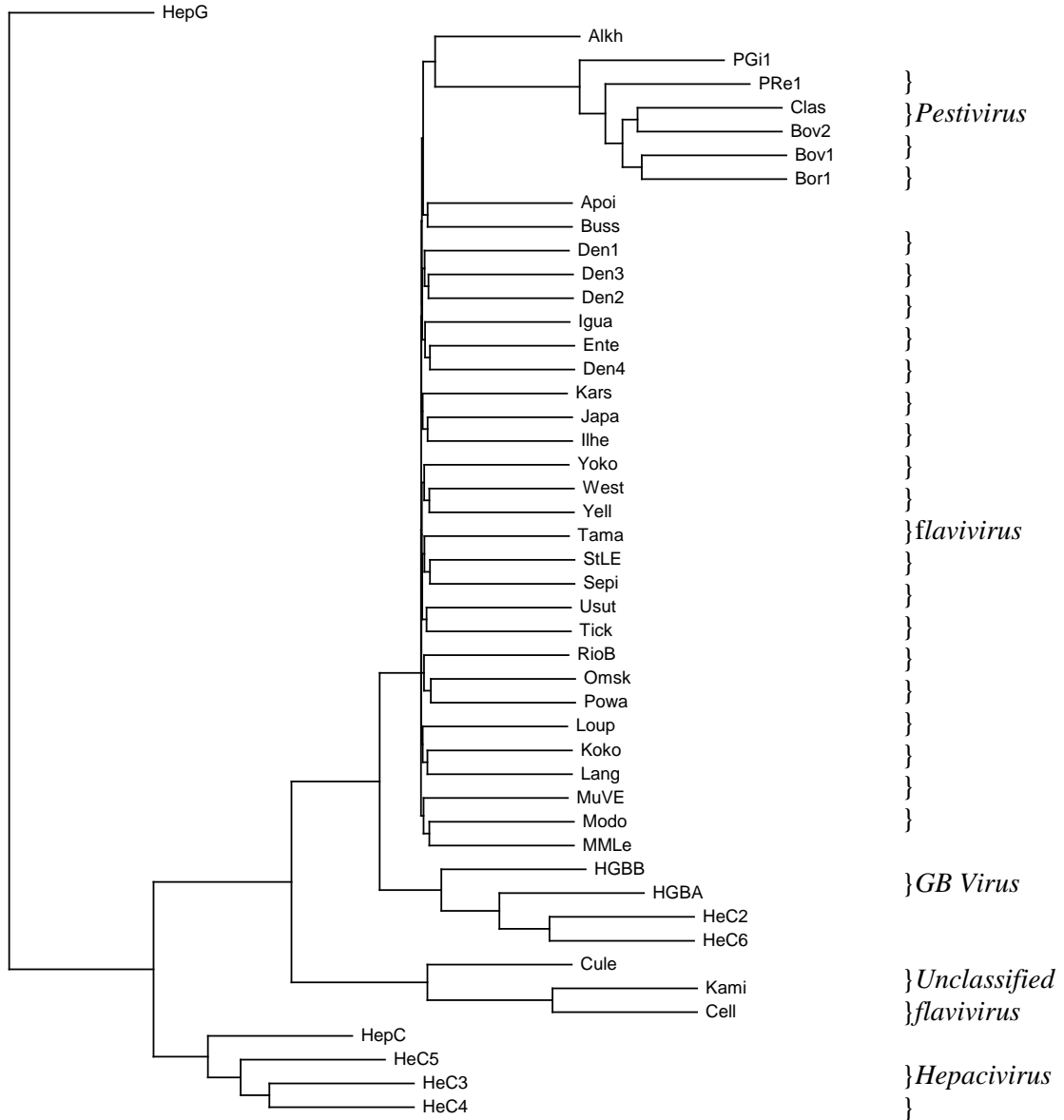
### **6.1.1 GENE CONTENT TREE**

VOGs clusters of viral related proteins represent the orthologous relationship between different viral genomes and give insight about evolutionary relatedness among them. By using VOG membership as a genetic trait, we constructed the whole genome phylogenetic tree for viruses belonging to Flaviviridae family. The Gene content tree is based upon gene content data delineated in the form of a data matrix (figure 6.1) that represents presence or absence of VOG membership among viral genome.

The data matrix analyzed with the help of maximum parsimony method (discrete character based) gave the Gene content tree (figure 6.2) with 500 pseudo-replicates.

For bootstrap values of each node please refer the trees' outfile in the supplementary material (Appendix B).

Fig 6.2 Gene content tree

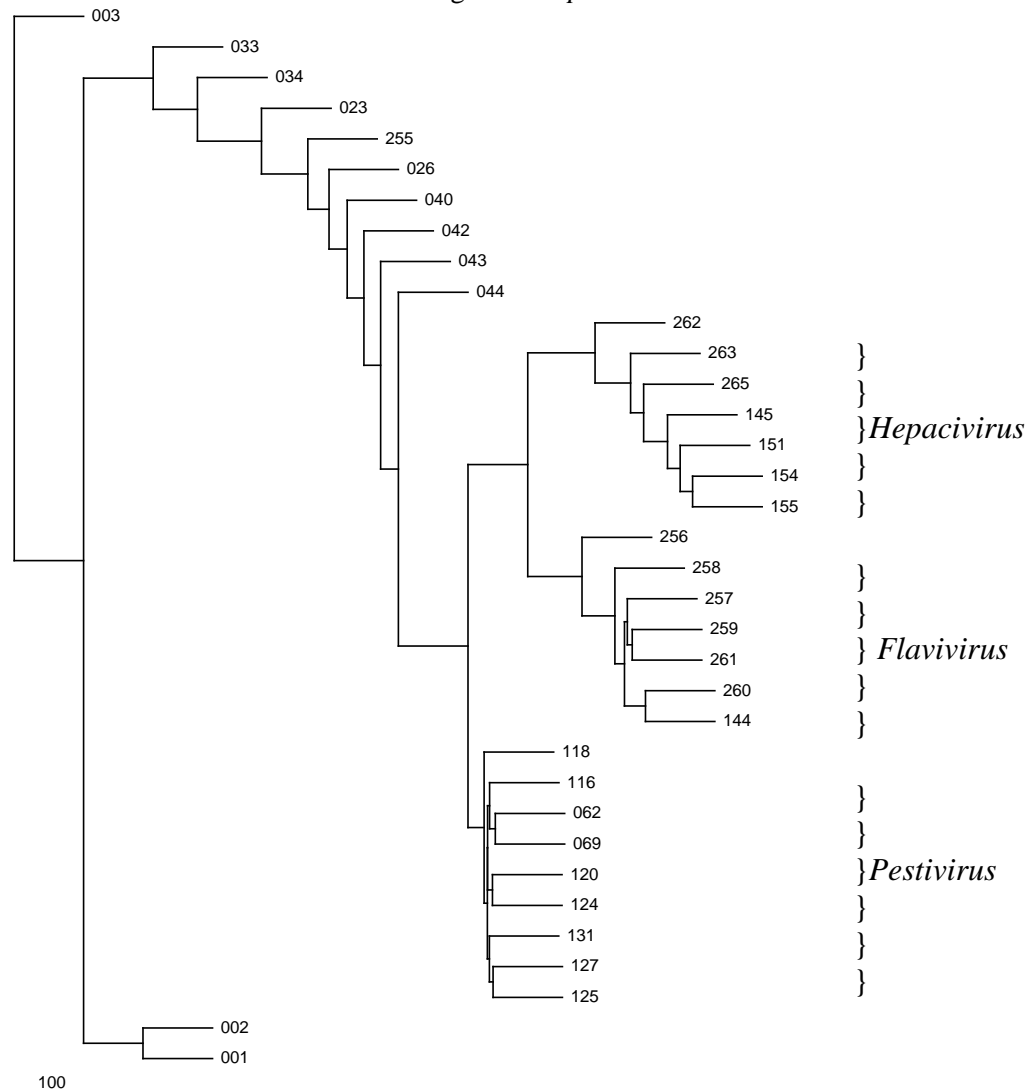




### 6.1.2 RECIPROCAL TREE

The reciprocal tree was constructed as explained earlier. We performed a clustering analysis of the phylogenetic profiles of the 35 VOGs at CSN = 3 by maximum parsimony method (discrete character based) to obtain the tree shown in figure 6.3

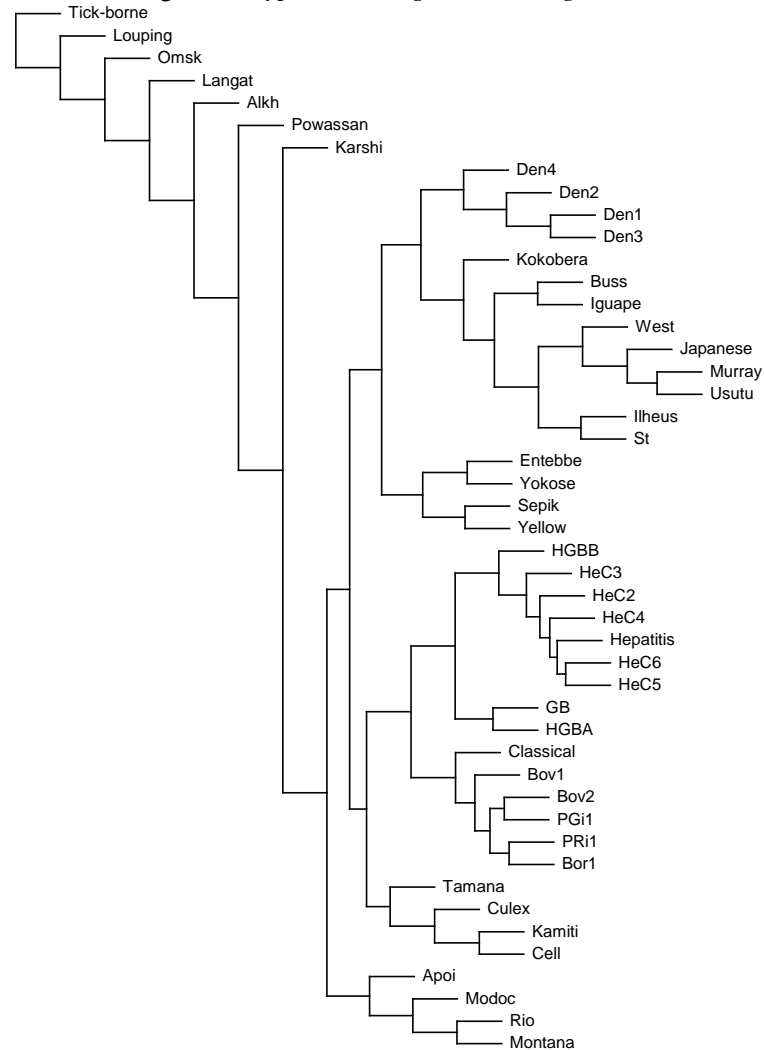
Fig 6.3 *Reciprocal tree*



### 6.1.3 POLYPROTEIN NEIGHBOR JOINING TREE

A neighbor-joining tree was constructed using the PHYLIP package based upon distance matrix created with the help of polyprotein alignment (from CLUSTALW [19]). This tree shown in figure 6.4 was created to compare and validate results obtained from the Gene content tree. Both trees share similar nodes are strongly supported by boot strap analysis.

Fig 6.4 *Polyprotein Neighbor Joining Tree*



## 6.2 DISCUSSION

Phylogenies are built upon the shared and contrasted traits of a group of organisms or their parts. A phylogeny represents paths of descent that depict the transmission of characters from ancestral lineage to descendent lineage. It consists of terminals (the species or gene sampled), internal nodes (ancestral state reconstructions of the associated characters), and branches (which define the relationships of the units and can represent the relative divergence among the terminals and nodes). Branch lengths may depict the time between speciation events according to the mutation rate or the number of mutations along a lineage, depending on the operational taxonomic units (OTUs).

The vast majority of genes in sequenced genomes are organized into gene families and superfamilies. In phylogenomic era, the relationships between different genes represented by homologous families consist of both orthologs and paralogs. Orthologs are the genes in different species evolved from an ancestral gene by speciation; by contrast paralogs are genes related by duplication within genome. The hierarchy of genes, like the nested organization of living organisms, has been produced primarily by a process lineage splitting and divergence –in this case, gene duplication followed by independent trajectories of sequence substitution.

COGs (clusters of orthologous group) provide a powerful tool for defining gene families in completely sequenced genomes. Here, we applied this technique in case of viral genomes and created viral specific COGs and proved that COG membership is a useful phylogenetic trait for understanding the evolutionary history of genomes. In previous studies, phylogenetic profiling used for functional characterization of proteins that are evolved in co-related fashion during evolutionary process. Here, phylogenetic profiles of VOGs membership represented by binary sequence, used for

understanding the evolutionary history of viral genomes. This was done by creating reciprocal tree based on phylogenetic profile of VOG membership in Flaviviridae genomes and gene content tree based on distribution of particular VOG in Flaviviridae genomes.

### **6.2.1 GENE CONTENT TREE**

Phylogeny reconstruction is based on aligned sequences of genes that conserve across species. In the gene content tree we utilized the strong phylogenetic signal that was present across whole genome especially those which are not universally conserved. Gene products that are conserved across all the genomes translate into a column of ones in the data matrix and thus are not informative. In this approach, differentiation is based upon partially conserved genes and presence or absence of VOG membership represented by binary sequence of 1 or 0 for each genome. It is possible to use all of the phylogenetic sequence analysis tools that are available, such as maximum parsimony (discrete character based) and bootstrap.

The distribution pattern of gene content data in the data matrix does not affect the construction of gene content tree. Our method of gene content tree segregates Flavivirus, Pestivirus and Hepacivirus into three groups but is not precise. A clear-cut tree could not be obtained due to two main reasons –

1. Presence of less number of genes in viral genomes of Flaviridae family, and
2. The VOGs were created at COG Stringency Number (CSN) = 3.

The gene content tree nodes (fig 6.2) contain three members of unclassified Flavivirus, six members of Pestivirus and four members of Hepacivirus which share a similar topology with the neighbor-joining tree (Fig 6.4) based on distance matrix of polyprotein comparison of viral genomes. These nodes were nearly present in all the gene content trees derived from rearranged input binary sequences for viral genomes.

This argues strongly for the ability of gene content trees in deducing evolutionary history from whole genome sequences.

In a phylogenetic tree based upon gene content, each node represents an ancestral organism that has gained or lost genes compared with its preceding ancestral node. It is useful to consider the fact of lateral gene transfer in whole genome phylogenetic tree based on gene content data. We assumed that speciation events happen after gene duplication or gene loss. Under this assumption, we can define the ancestor of an organism resulting from a gene acquisition or gene loss, which contributes to the majority of resulting genome. Finally we can conclude that based upon this assumption and definition, gene content tree approximates the actual history of speciation based on gene acquisition and loss.

### **6.2.2 RECIPROCAL TREE**

Reciprocal tree clusters the VOGs according to pattern similarity in viral genomes; VOGs with similar pattern cluster on the same node in a reciprocal tree. Topology of reciprocal (Fig 6.3) and neighbor-joining tree (Fig 6.4) is much similar than gene content tree (Fig 6.2). From the fact that phylogenetic VOGs consist of orthologs belonging to different viral species which derived from common ancestor, it can be assumed that the phylogenetic signal obtained from this methodology is stronger than gene content data.

In previous approach of phylogenetic profiling, it has been obtained that gene with similar phylogenetic profiles have share similar functions. The grouping of VOGs of unknown function in the reciprocal tree may be useful in identifying their functional role in evolutionary context. For example, we speculate that the Hepacivirus specific VOGs clustered in upper right of figure 6.4 may determine the envelope protein of Hepaciviruses.

The Obtained reciprocal tree, built from phylogenetic data has no direct phylogenetic interpretation. After analysis of obtained cluster in reciprocal tree, we can deduce that it arranges clusters according to phylogenetic relationship. Specific VOGs for Hepacivirus, unclassified Flavivirus and Pestivirus are arranged in different clusters. According to this observation, we can say that clustering of genes on the reciprocal tree, gives clues to groups of genes that have related evolutionary histories or related functions.

## **CHAPTER 7**

### **CONCLUSION**

---

The Phylogenetic Profiling technique, though a very elegant technique for prediction of protein functional co-relation in bacteria, does not apply flawlessly on viral genomes. In its original form the technique was difficult to implement and produced inconvincible results. Therefore, a modified approach was taken to establish evolutionary relationship among viruses based upon phylogenomics.

Whole genome phylogenetic tree based on gene content data provided relevant understanding about evolutionary history of viral genomes. Gene content tree shared similar nodes for three genres of Flaviviridae family with neighbor-joining tree. Further in this direction, we produced reciprocal tree based upon phylogenetic profiles of binary sequence that represented each VOG membership among viral genomes. In reciprocal tree, VOGs belongs to similar phylogenetic profiles share same clusters, these clusters also helpful for functional characterization of unknown proteins which belongs to clusters of known functional class.

Reciprocal tree also provides the information about phylogenetic relationship by distributing the VOGs according to their phylogenetic lineage. We obtained a similar evolutionary pattern for viral species that belonged to the same genus in both trees. This proved that gene content tree using VOG membership as a genetic trait is a potential approach for establishing whole genome phylogenetic tree construction.

---

## REFERENCES

### RESEARCH PAPERS

- 1) B. E. Dutilh <sup>1,\*</sup>, V. van Noort <sup>1</sup>, R. T. J. M. van der Heijden et al.; **Assessment of phylogenomic and orthology approaches for phylogenetic inference:** *Bioinformatics* 2007 23(7):815-824
- 2) Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278(5338)**:631-637.
- 3) Koonin EV: **Orthologs, paralogs, and evolutionary genomics.** *Annu Rev Genet* 2005, **39**:309-338.
- 4) Maria S Poptsova\* and J Peter Gogarte: **BranchClust: a phylogenetic algorithm for selecting gene families,** *BMC Bioinformatics* 2007, **8**:120  
[<http://www.bioinformatics.org/branchclust>]
- 5) Jonathan A. Eisen: **Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis.** *Genome Res.* 1998 **8**: 163-167.
- 6) S. Cook<sup>1</sup> and E. C. Holmes<sup>2</sup>; **A multigene analysis of the phylogenetic relationships among the flaviviruses (Family: *Flaviviridae*) and the evolution of vector transmission:** *Arch Virol* (2006) 151: 309–325
- 7) Montague MG, Hutchison CA 3rd: **Gene content phylogeny of herpesviruses.** *Proc Natl Acad Sci USA* 2000, **97(10)**:5334-5339.
- 8) Roger W. Hendrix, Jeffrey G. Lawrence, Graham F. Hatfull and Sherwood Casjens; **The origins and ongoing evolution of viruses .TRENDS IN MICROBIOLOGY** 504 VOL. 8 NO. 11.



- 9) M. Codoñer and S. F. Elena; **Evolutionary relationships among members of the *Bromoviridae* deduced from whole proteome analysis**. *Arch Virol* (2006) 151: 299–307.
- 10) Roger W. Hendrix; **Evolution: The long evolutionary reach of viruses.**
- 11) Yeates, Matteo Pellegrini, Edward M. Marcotte and Todd O.; **Assigning protein functions by comparative genome analysis: Protein phylogenetic Profiles**, *PNAS* 1999;96:4285-4288
- 12) David Eisenberg, Edward M. Marcotte, and Ioannis Xenarios & Todd O. Yeates; **Protein function in the post-genomic era: *Nature***, 823-826(2000).
- 13) Sorel T. Fitz-Gibbon\*, and Christopher H. House; **Whole genome-based phylogenetic analysis of free living micro-organisms: 4218-4222** *Nucleic Acids Research*, 1999, Vol. No. 21
- 14) Catarina Mota, Isabel Gordo; **Adaptive Mutations in Bacteria- High Rate and Small Effects** : *SCIENCE*, Vol 317, 813-816.
- 15) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool. *J Mol Biol*** 1990, **215(3)**:403-410.
- 16) Altschul SF, Koonin EV: **Iterated profile searches with PSIBLAST – a tool for discovery in protein databases. *Trends Biochem Sci*** 1998, **23(11)**:444-447.
- 17) Altschul et al.: **Gapped blast and PSI-blast: a new generation of protein database search programs. *Nucleic Acids Research***, pages 25(17):3389\_3402, 1997.
- 18) Felsenstein J: **Cases in which parsimony and compatibility methods will be positively misleading. *Syst Zool*** 1978, **27**:401-410.
- 19) Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*** 1994, **22**:4673 -44680.

- 20) **Don Gilbert:** ReadSeq, biosequence data translator, version 2,  
<http://iubio.bio.indiana.edu/soft/molbio/readseq/java/>.
- 21) Felsenstein, J. 2004. **PHYLIP (Phylogeny Inference Package) version 3.6**  
*Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*
- 22) Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4(4)**:406-425.
- 23) Zhaxybayeva O, Gogarten JP: **Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses.** *BMC Genomics* 2002, **3(1)**:4.
- 24) B.Néron<sup>1</sup>, P. Tufféry<sup>2</sup>, C.Letondal<sup>1</sup>: **Mobyle: a Web portal framework for bioinformatics analyses**, poster presented at NETTAB 2005

## **BOOKS**

- 25) Lynn Helena Caporale; Darwin in the genome – molecular strategies in biological evolution.
- 26) Paul G. Higgs and Teresa K. Attwood; Bioinformatics and Molecular Evolution: Chapter 4 : Models of sequence evolution, Chapter 8 : Phylogenetic Methods.
- 27) Charles W. Fox, Jason B. Wolf; Evolutionary Genetics; Chapter 11: New Genes, New Functions : Gene Family Evolution and Phylogenetics. Chapter 28 : Theory of Phylogenetic Estimation.
- 28) Daugherty, Projan; Microbial Genomics and Drug discovery.

## **WEBSITES**

- 29) [http://www.biochem.ucl.ac.uk/bsm/virus\\_database/VIDA.html](http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html)

- 30) [http://www.tulane.edu/~dmsander/Big\\_Virology/BVHomePage.html](http://www.tulane.edu/~dmsander/Big_Virology/BVHomePage.html)
- 31) <http://biowulf.nih.gov/apps/blast/doc/formatdb.html>
- 32) <http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/vifam.html>
- 33) [http://www.ncbi.nlm.nih.gov/projects/Gene/gentrez\\_stats.cgi](http://www.ncbi.nlm.nih.gov/projects/Gene/gentrez_stats.cgi)
- 34) <http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/>
- 35) <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/>
- 36) <http://www.ebi.ac.uk/cgi-bin/readseq.cgi>
- 37) <http://mobyle.pasteur.fr/cgi-bin/MobylePortal/>
- 38) [carnot.utmb.edu/flavitrack/index.php](http://carnot.utmb.edu/flavitrack/index.php)
- 39) [www.biovirus.org/](http://www.biovirus.org/)
- 40) <http://athena.bioc.uvic.ca/workbench.php?tool=vocs&db=flaviviridae>

## APPENDIX A

| VOG LIST  |         |
|-----------|---------|
| VOG name  | Abbrev. |
| VOGsr0001 | 001     |
| VOGsr0002 | 002     |
| VOGsr0003 | 003     |
| VOGsr0023 | 023     |
| VOGsr0026 | 026     |
| VOGsr0033 | 033     |
| VOGsr0034 | 034     |
| VOGsr0040 | 040     |
| VOGsr0042 | 042     |
| VOGsr0043 | 043     |
| VOGsr0044 | 044     |
| VOGsr0062 | 062     |
| VOGsr0069 | 069     |
| VOGsr0116 | 116     |
| VOGsr0118 | 118     |
| VOGsr0120 | 120     |
| VOGsr0124 | 124     |
| VOGsr0125 | 125     |
| VOGsr0127 | 127     |
| VOGsr0131 | 131     |
| VOGsr0144 | 144     |
| VOGsr0145 | 145     |
| VOGsr0151 | 151     |
| VOGsr0154 | 154     |
| VOGsr0155 | 155     |
| VOGsr0255 | 255     |
| VOGsr0256 | 256     |
| VOGsr0257 | 257     |
| VOGsr0258 | 258     |
| VOGsr0259 | 259     |
| VOGsr0260 | 260     |
| VOGsr0261 | 261     |
| VOGsr0262 | 262     |
| VOGsr0263 | 263     |
| VOGsr0265 | 265     |

| List of Supplementary Material<br>provided in the CD   |
|--|
| <ol style="list-style-type: none"> <li>1. Sequences of all 47 viruses in Flaviviridae family.</li> <li>2. Sequences of 27 viruses in Picornaviridae family.</li> <li>3. Perl Scripts</li> <li>4. Results <ul style="list-style-type: none"> <li>– Output files from Phylogenetic profiling work (BLAST, etc)</li> <li>– Phylogenetic profile (matrix)</li> <li>– VOG clusters and its sequences.</li> <li>– VOG Data Matrix</li> <li>– Phylogenetic trees input and Output files</li> <li>– Gene Content tree</li> <li>– Reciprocal Tree</li> <li>– Neighbor joining tree</li> </ul> </li> </ol> |

## APPENDIX B

### Phylogeny of Flaviviridae

- **Flavivirus** (arboviruses group B)
  - Dengue virus group
    - Dengue virus 1
    - Dengue virus 2
    - Dengue virus 3
    - Dengue virus 4
  - Japanese encephalitis virus group
    - Japanese encephalitis virus
    - Koutango virus
    - Murray Valley encephalitis virus
    - St. Louis encephalitis virus
    - Usutu virus
    - West Nile virus
  - Kokobera virus group
    - Kokobera virus
    - unclassified Kokobera virus group
  - Modoc virus group
    - Cowbone Ridge virus
    - Jutiapa virus
    - Modoc virus
    - Sal Vieja virus
    - San Perlita virus
  - mosquito-borne viruses
    - Ilheus virus
    - Sepik virus
  - Ntaya virus group Bagaza virus
    - Israel turkey meningoencephalomyelitis virus
    - Ntaya virus
    - Tembusu virus
    - Yokose virus
  - Rio Bravo virus group
    - Apoi virus
    - Bukalasa bat virus
    - Carey Island virus
    - Dakar bat virus
    - Entebbe bat virus
    - Rio Bravo virus
    - Saboya virus
  - Seaborne tick-borne virus group
    - Meaban virus
    - Saumarez Reef virus

- Tyuleny virus
- Spondweni virus group
  - Zika virus
- tick-borne encephalitis virus group
  - Kyasanur forest disease virus
  - Langat virus
  - Louping ill virus
  - Omsk hemorrhagic fever virus
  - Phnom Penh bat virus
  - Powassan virus
  - Royal Farm virus
  - Tick-borne encephalitis virus
- Yaounde virus
- Yellow fever virus group
  - Banzi virus
  - Bouboui virus
  - Edge Hill virus
  - Uganda S virus
  - Wesselsbron virus
  - Yellow fever virus
- unclassified Flavivirus
  - Aroa virus
  - Batu Cave virus
  - Bussuquara virus
  - Cacipacore virus
  - Cell fusing agent virus
  - Culex flavivirus
  - Flavivirus CbaAr4001
  - Flavivirus FSME
  - Gadgets Gully virus
  - Greek goat encephalitis virus
  - Iguape virus
  - Jugra virus
  - Kadam virus
  - Kamiti River virus
  - Kedougou virus
  - Montana myotis leukoencephalitis virus
  - Ngoye virus
  - Russian Spring-Summer encephalitis virus
  - Sokoluk virus
  - Spanish sheep encephalitis virus
  - Tai forest virus B3
  - Tamana bat virus
  - Tick-borne flavivirus
  - Wang Thong virus

- Flavivirus sp.
- **Hepacivirus**
  - Hepatitis C virus
    - Hepatitis C virus genotype 1
    - Hepatitis C virus genotype 2
    - Hepatitis C virus genotype 3
    - Hepatitis C virus genotype 4
    - Hepatitis C virus genotype 5
    - Hepatitis C virus genotype 6
    - unclassified Hepatitis C virus
- **Pestivirus**
  - Border disease virus
    - Border disease virus - BD31
    - Border disease virus - X818
    - Border disease virus 1
    - Border disease virus 2
    - Border disease virus 3
    - Border disease virus isolates
  - Bovine viral diarrhea virus 1
    - Bovine viral diarrhea virus 1-CP7
    - Bovine viral diarrhea virus 1-NADL
    - Bovine viral diarrhea virus 1-Osloss
    - Bovine viral diarrhea virus 1-SD1
    - Bovine viral diarrhea virus isolates and strains
    - Pestivirus isolate 97-360
    - Pestivirus isolate Hay 87/2210
    - Pestivirus strain mousedeer
    - Pestivirus type 1 isolates
  - Bovine viral diarrhea virus 2 (BVDV-2)
    - Bovine viral diarrhea virus 2-C413
    - Bovine viral diarrhea virus 2-New York'93
    - Bovine viral diarrhea virus-2 isolate 230/98-K1 (Gi-4)
    - Bovine viral diarrhea virus-2 isolate 230/98-K2 (Gi-5)
    - Bovine viral diarrhea virus-2 isolate 230/98-K3 (Gi-6)
    - Bovine viral diarrhea virus-2 isolate Giessen-3
    - Bovine viral diarrhea virus-2 isolate SCP
  - Classical swine fever virus
    - Classical swine fever virus - Alfort/187
    - Classical swine fever virus - Alfort/Tuebingen
    - Classical swine fever virus - Brescia
    - Classical swine fever virus - C
    - Classical swine fever virus isolates
    - Hog cholera virus strain Zoelen
  - unclassified Pestivirus
    - Bovine viral diarrhea virus 3

- Chamois pestivirus 1
  - Ovine pestivirus
  - Pestivirus Giraffe-1
  - Pestivirus HoBi
  - Porcine pestivirus
  - Pronghorn antelope pestivirus
  - Pestivirus sp.
- **unclassified Flaviviridae**
  - GB virus A
    - Douroucouli hepatitis GB virus A
    - GBV-A-like agents
  - GBV-C/HGV group
    - GB virus C
    - Hepatitis GB virus C-like virus
  - Hepatitis GB virus B
  - Marmoset hepatitis GB virus A
  - Turkey meningoencephalitis virus



## **PUBLICATIONS**

1. Amitabh Gupta, Department of Bioechnology, IIIT University, Noida, India.  
Poster titled: *Computer Aided Vaccine Design*. Conference: Bangalore Bio'2006
2. Amitabh Gupta, Department of Bioechnology, IIIT University, Noida, India.  
Poster titled: *Application of Artificial Neural Network and Immuno-Informatics for Vaccine Design*. Conference : IRIS National Fair- 2007
3. Amitabh Gupta, Department of Bioechnology, IIIT University, Noida, India.  
Poster Titled: *Application of Artificial Neural Network and Evolutionary Immuno-Informatics for Vaccine Deign*. Conference: British Council of India, Biotech Idea to Innovation Programme –2007 & Bio-Horizon- 2007.

|                 |
|-----------------|
| <b>BIO DATA</b> |
|-----------------|

Curriculum Vitae  
**Amitabh Gupta**

**Correspondence address:**

A-297/2, Rajendra Nagar,  
Bareilly- 2430001, U. P. India  
Mobile: 00919891320085

**Email:**

**amitabh\_gupta07@yahoo.co.in**

**Academic Qualification**  
**July 2004 – Present**

**Bachelor of Technology –  
Biotechnology**

Jaypee institute of Information  
Technology, Noida, India

CGPA – 7.0 (~ 75%)

**August 2002- March 2003**

12<sup>th</sup> Standard, First Division

**August 2000- March 2001**

10<sup>th</sup> Standard, First Division

**Final Year Research Project:** Gene content phylogeny of Flaviviridae family of single stranded RNA viruses. For establishing the whole genome phylogenetic relationship between members of this family, identification of Clusters of orthologous groups (COG) have done by best hit approach.

**Research Experience:**

Research Trainee (May 2007 –July 2007), Under Prof. Indira Ghosh, Director, Bioinformatics Centre, University of Pune, India

Functional co-relation between proteome set of microbes by phylogenetic profiling. A profile database of proteome set of 50 microbes was developed for functional characterization of hypothetical proteins of *Pseudomonas aeruginosa* and filter out novel drug targets.

Simultaneously pursued a project: Standardisation of Discotope B-cell epitope prediction server using 22 previously characterized virus antigen-antibody complexes under the guidance of Dr. Urmila Kulkarni, Lecturer, Bioinformatics Centre, University of Pune.

#### **Undergraduate Projects:**

- Development of a database management system in ORACLE to retrieve viral epitope and corresponding human immune cell receptor sequences for pathogenic viruses.
- Optimization of molecular descriptors using machine-learning approach to aid structure based drug designing for viruses.
- Designed a business plan ‘Online Consultancy Services for HIV patients’ as part of a project work for the B. Tech course titled Entrepreneurial Development.

#### ***Achievement and Awards***

1. Special Mention Jury Award, 2007, Biotech Idea to Innovation Programme 2007, British Council India. (<http://www.britishcouncil.org>).
2. Prof. William Webster scholarship, College level, 2004.
3. Represented College in Agilent Engineering & Technology Award 2008 (AETA).
4. Participated National level IRIS 2007 (Initiative for Research and Innovation in Science) organized by Department of Science & Technology, India, Intel and Confederation of Indian Industry for research proposal on vaccine design.
5. GATE 2007 [Graduate Aptitude Test in Engineering] – 93.7 percentile at all India level.
6. Won first prize, “FACE CLONNING”, college annual fest JIVE’07, event based upon painting competition on biotech theme.
7. Selected for top 5 teams out of 20 in ‘STRATGEMY’, college annual fest JIVE’07, event based upon the Ideas for ‘Restructuring of Indian Postal System’

#### **Conferences/Seminars/Workshops (Selected Only)**

1. Business plan, SymBio - Knowledge based consultancy venture, Karyon’2008, Department of biotechnology, Delhi college of Engineering.

2. Poster presentation – Application of Immuno-Informatics and Artificial Neural Network for Vaccine Design, Bio-horizon 2007, IIT-Delhi and Korean Biochemical Society.
3. Poster presentation – Computer-aided vaccine design, Bangalore BIO 2006, Vision -Biotechnology.
4. Participated International conference on computational biology, INCOB-2007, DBT (Department of biotechnology, India) and Jawaharlal Nehru University, New Delhi.
5. Participated, Bio-horizon 2006, Indian Institute of Technology, New Delhi.
6. Summer Training, Bioinformatics Centre, University of Pune, India

**Technical Skills:**

**Statistical packages:** SAS, **Programming Languages:** Perl / Bioperl, C and SQL, **Operating Systems:** Linux, Windows (XP/2000/Vista), **Database Systems:** Oracle, & MySql **Web:** Front page, HTML/DHTML, Visual Studio., **Image processing:** Adobe Photoshop & Illustrator, **Machine Learning Methods:** Neural Networks, Support Vector Machine.

**Bioinformatics software packages:** Hyperchem, PolMol, Rasmol, **Mummer 3.20**, **Phylip**, **VMD 1.8.6** & familiarize online tools for sequence alignment, protein interaction network, gene ontology, biochemical pathway analysis, epitope prediction, Immuno-informatics tools.

**Laboratory Techniques:** Microbial Culturing Methods, General molecular biology methods, PAGE, PCR, Site-directed mutagenesis, Enzyme assays, ELISA.

## BIO DATA

### KUNAL PUNJRATH

Contact address: 6/22, Aashirwad Enclave, Dehradun - 248178

Email: [kunalpunj Rath@rediffmail.com](mailto:kunalpunj Rath@rediffmail.com)

☎: +91 9891641975

Date of Birth: 16 April 1986

### ACADAMIC QUALIFICATION

| Year  | Degree/<br>Education board  | Chapter 5. Institute/ School                            | C.G.P.A/<br>Percentage |
|-------|-----------------------------|---|------------------------|
| 2004- | B. Tech in<br>Biotechnology | Jaypee Institute Of Information<br>Technology, Noida    | 6.4                    |
| 2003  | I.S.C (XII)                 | Carman Residential & Day School,<br>Dalanwala, Dehradun | 61.2%                  |
| 2001  | I.C.S.E (X)                 | Carman Residential & Day School,<br>Dalanwala, Dehradun | 77.83%                 |

### TECHNICAL SKILLS

|  |   |
|--|---|
| Familiar with<br>programming<br>languages  | <ul style="list-style-type: none"> <li>• Perl</li> <li>• C\C++</li> <li>• HTML</li> <li>• JavaScript.</li> </ul>  |
| Bioinformatics<br>tools  | <ul style="list-style-type: none"> <li>• PHYLIP</li> <li>• RasMol, MolMol</li> <li>• BLAST, ClustalW</li> <li>• Swiss-PdbViewer</li> <li>• HyperChem</li> </ul> |
| MySQL, MS Office, Adobe Photoshop, Adobe Illustrator, Macromedia<br>Dreamweaver. |   |

### ACADEMIC PROJECTS

|                    |   |
|--------------------|---|
| Final year project | <ul style="list-style-type: none"> <li>• <u>Gene content phylogeny of Flaviviridae family:</u><br/>Phylogentic profiling and Clusters of orthologous<br/>groups (COGs) approaches were implemented.<br/>Phylogenetic trees were created.</li> </ul> |
|--------------------|---|

|                                  |   |
|----------------------------------|---|
| Industrial Training              | <ul style="list-style-type: none"> <li>• <u>Microbial Profiling of deep Offshore Wells for Hydrocarbon Exploration</u> at Geo-chemistry division, KDMIPE ONGC, Dehradun. Methanococcus, Methanogens and Sulfate reducing Bacteria were profiled.</li> </ul>   |
| C programs for implementation of | <ul style="list-style-type: none"> <li>• <u>Needleman-Wunch algorithm</u>: Global alignment of two nucleotide or protein sequences, using 2D arrays.</li> <li>• <u>Smith-Waterman algorithm</u>: Local alignment of two nucleotide or protein sequences, using 2D arrays.</li> <li>• <u>UPGMA method</u>: To calculate the Phylogenetic distance between two or more species based on user specified data.</li> </ul> |
| Web Site                         | <ul style="list-style-type: none"> <li>• <u>Seven Wonders of the World</u>: Extensive details and photos of ancient, modern, and new seven wonders. Site developed using HTML, CSS, and Javascript.</li> </ul>  |

#### EXTRA-CURRICULARS

- Prefect for House in school. Organized inter-house competitions.
- Participated in Quiz competitions, and science fairs at school.
- Attended Bio-Horizon 2006, symposium held at IIT-D.