IITB-Monash Research Academy

An Indian-Australian research partnership

Indian Institute of Technology Bombay | MONASH University

Project Title:	**Computational Phylogenetics for Variant Manuscripts**	
Project Number	**IMURA0497**	
Monash Main Supervisor (Name, Email Id, Phone)	Dr. Reza Haffari, Gholamreza.Haffari@monash.edu,	*Full name, Email*
Monash Co-supervisor(s) (Name, Email Id, Phone)		
Monash Department:	Faculty of Information Technology, Monash University, Clayton.	
IITB Main Supervisor (Name, Email Id, Phone)	Prof. Pushpak Bhattacharyya, pushpakbh@gmail.com, pb@cse.iitb.ac.in,	*Full name, Email*
IITB Co-supervisor(s) (Name, Email Id, Phone)	Prof. Malhar Kulkarni, malhar@hss.iitb.ac.in, malharku@gmail.com,	
IITB Department:	Department of Computer Science and Engineering, IIT Bombay.	

Research Academy Themes:

Highlight which of the Academy's Theme(s) this project will address?
(*Feel free to nominate more than one. For more information, see www.iitbmonash.org)*

1. **Advanced computational engineering, simulation and manufacture**
2. Infrastructure Engineering
3. Clean Energy
4. Water
5. Nanotechnology
6. Biotechnology and Stem Cell Research
7. **Humanities and Social Sciences**

The research problem

In ancient times, text was written manually by scribes for the classical languages such as Sanskrit, Greek, Latin, etc. This ancient text is usually called as manuscript (Figure 1). In the case of Sanskrit, some examples of manuscripts are *Kāśikāvṛttī*, *CarakaSamhita*, *Mahabharata*, *Ramayana*, etc. The manuscripts were usually written on metal, leaves, wooden staves, etc.

Figure 1: Examples of Manuscripts

Earlier, the manuscripts used to be copied by scribes in many ways: by referring to the original document, by comparing various texts, by writing down the orally transmitted text, by using his/her expertise and understanding, etc. In copying a text, a scribe normally creates a new version of the text that usually differs from its original text. The next scribe reproduces variants of the previous copy. Moreover, he introduces new variants himself and possibly also eliminates some variants by correcting obvious mistakes. The process of copying and recopying produces a hierarchical pattern of variants, so that some variant readings can be identified. The detection of the hierarchical pattern of variants transmitted in the existing manuscripts provides the textual transmission. Over a period of time, manuscripts of the same text are created at different locations and at different time. As a result of this, multiple copies of a text having variant readings accumulated. Figure 2 illustrates the manuscripts having variant readings.

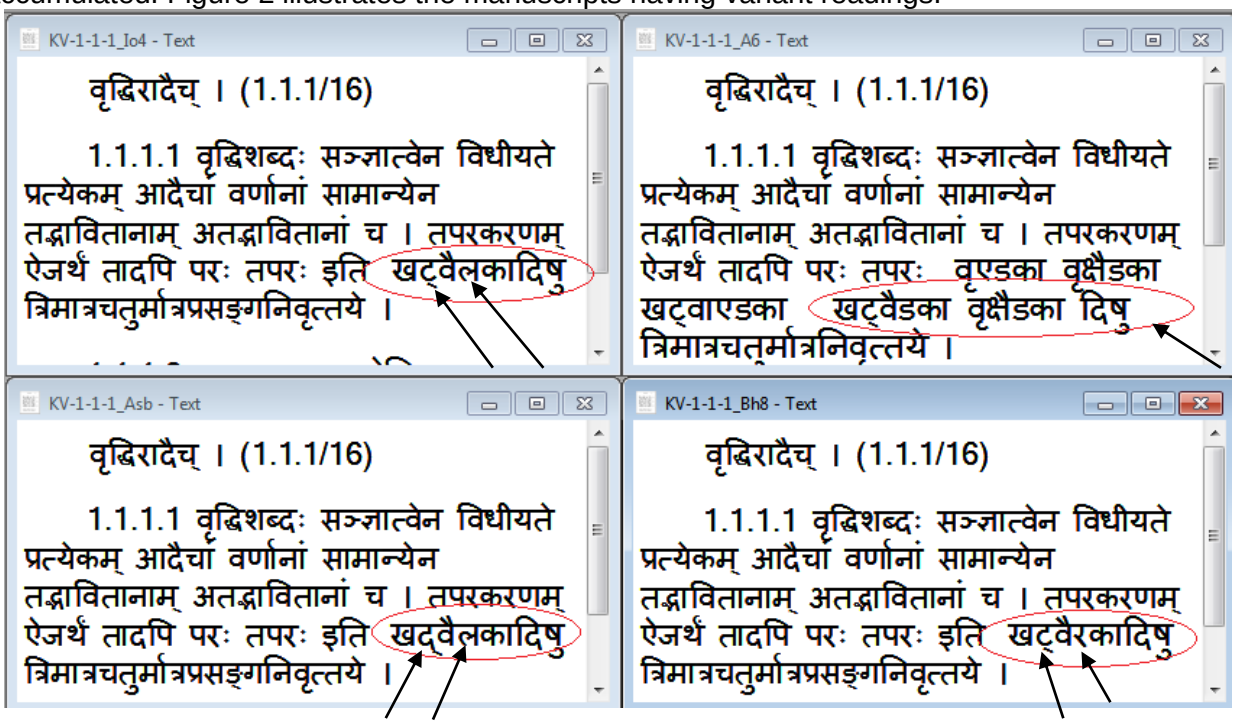

Figure 2: Variant Manuscripts of Same Text (arrows points to variants)

The research challenge here is to find the relationship between these variant manuscripts and find the source of the text.

We propose to use the method called as computational phylogenetics to construct

phylogenetic trees to find the relationship between variant manuscripts. A phylogenetic tree or an evolutionary tree is a branching diagram or tree showing the inferred evolutionary relationships among various biological species or other entities. Figure 3 shows an example of phylogenetic tree of life. An Analogy of phylogenetic tree of life, we can have phylogenetic tree of manuscripts.

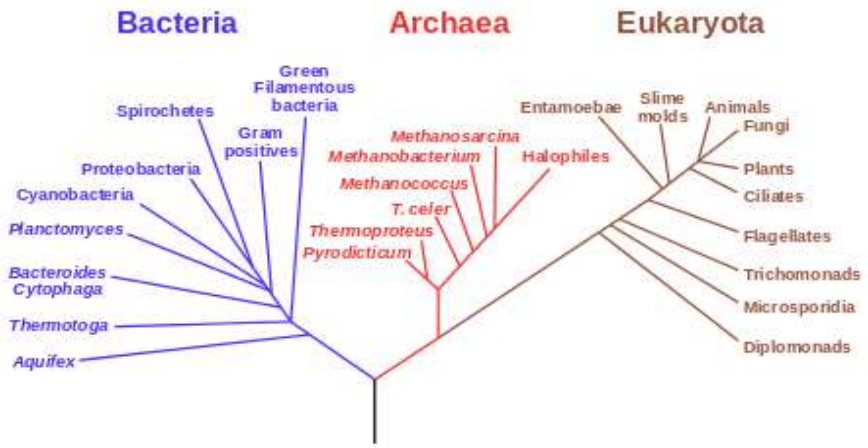

Figure 3: An Example of a Phylogenetic tree of Life (source: Wikipedia)

Computational phylogenetics is the application of computational algorithms, methods, and programs used for phylogenetic analyses. Here, the goal is to assemble a phylogenetic tree representing a hypothesis about the evolutionary ancestry of a set of genes, species, or other taxa (in our case variant manuscripts).

Figure 4 is an example of a manually drawn phylogenetic tree of Malayalam manuscripts *(Malayalam manuscripts of the Kāśikāvṛtti : A study by Malhar Kulkarni, 2006)*. Here, M is a source and Ma, Mb and Mc are its child. M is decided as a source based on the analysis made on the variant readings. In this process manuscripts have similar variants are grouped together and named as M1, M2,…, M11.

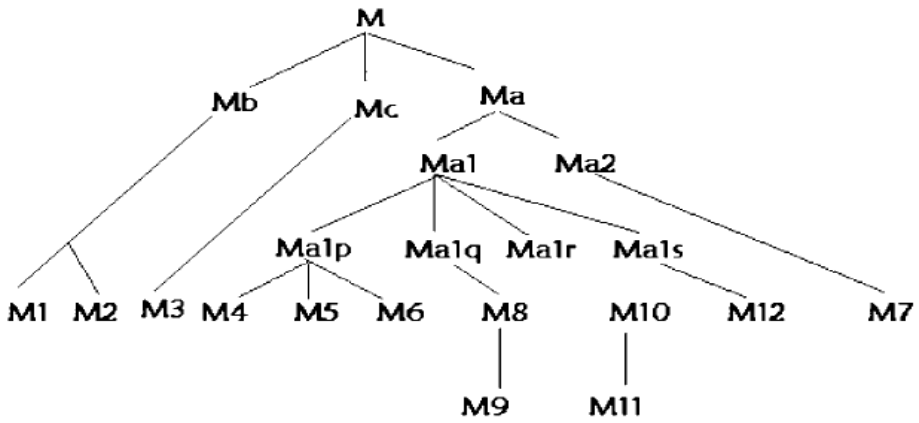

Figure 4: An Example of a Phylogenetic tree of manuscripts (source: Kulkarni(2006))

Figure 5 gives the block diagram of the process of computational phylogenetics for building phylogenetic trees for the variant manuscripts. The punched-tape shape in the diagram is a manuscript/text and rectangular shape is a process which in fact is a research problem to be explored and studied in detail under this project.

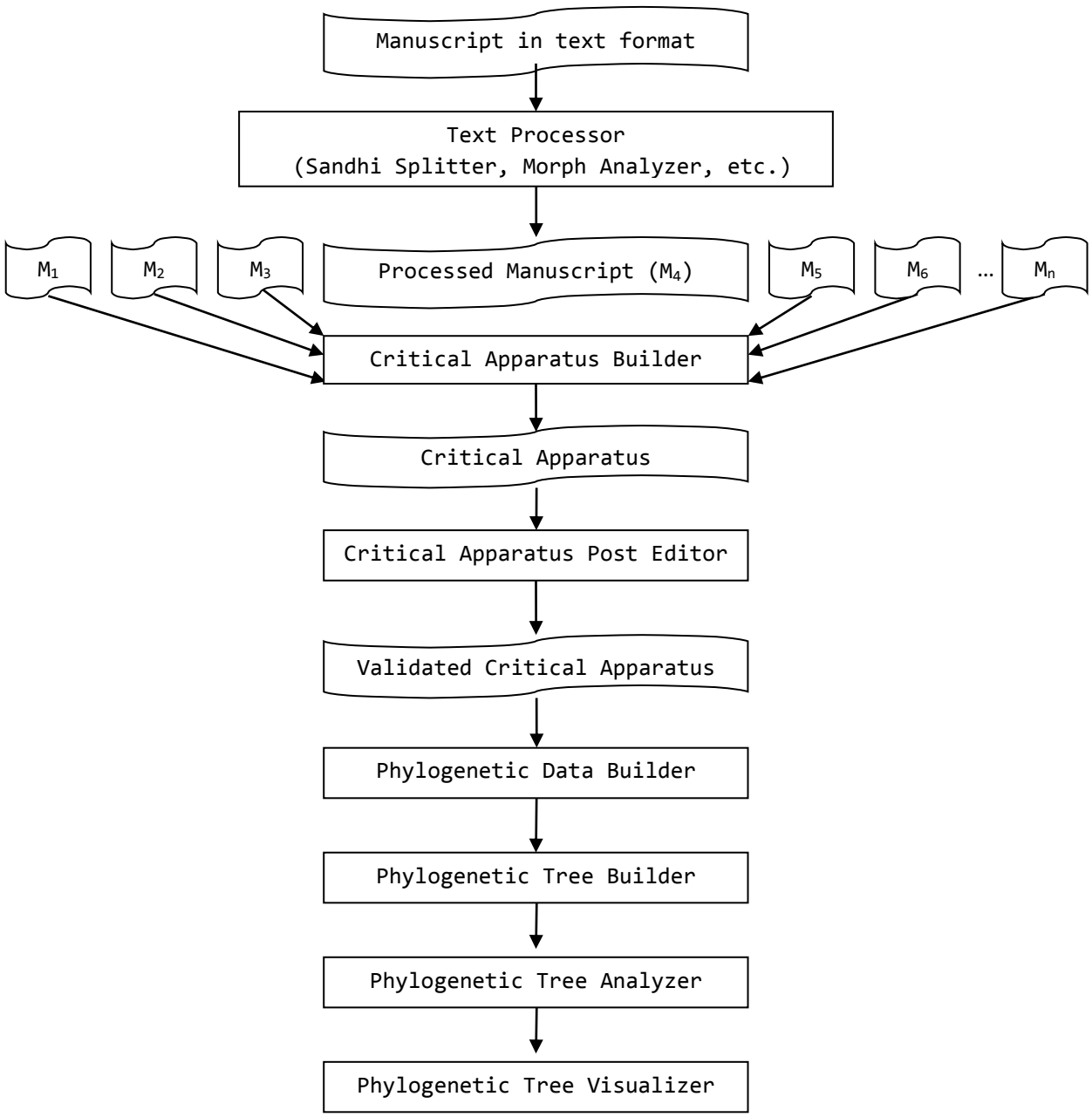

Fig. 5: Process of Computational Phylogenetics for Variant Manuscripts

The details of the modules of computational phylogenetics process is as follows:

1) **Text Processor:**
   The manuscript in text format shall be processed by passing through the morph analyser, Sandhi splitter, etc. so that each lemma (word) can be regenerated into its processable form. Here the input would be manuscript in text format and the output would be a simplified text.

2) **Critical Apparatus Builder:**
   Critical apparatus is the critical and primary source material that accompanies an edition of a text. It is often a by product of manuscript collation. Currently the critical apparatus is created manually by lexicographers. Critical apparatus shall be automatically built. Once we automate this process, it will reduce lot of manual work like going through each manuscript, cross checking, editing in the critical apparatus document, etc.

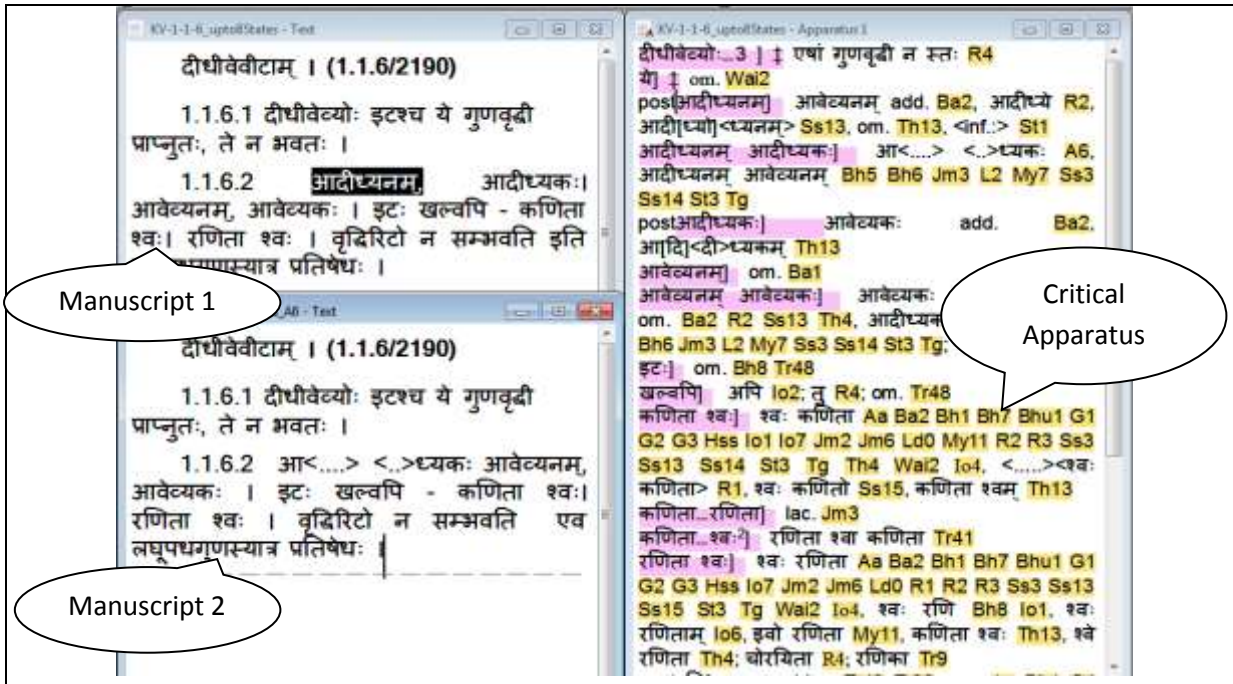

Figure 6: Samples of Manuscripts and Critical Apparatus documents

3) **Critical Apparatus Post Editor:**
   Critical apparatus documents shall be properly validated. This can be done manually or using the critical apparatus post editor.

4) **Phylogenetic Data Builder:**
   Once the validated critical apparatus is built, the apparatus data needs to be converted into a format which would help in building trees. There may be various formats the data can be exported to. The phylogenetic data builder will build the data into multiple input formats as per the requirement of the phylogenetic tree builder program.

5) **Phylogenetic Tree Builder:**
   The phylogenetic tree builder shall build various trees using phylogenetic methods such as maximum parsimony, maximum likelihood, distance methods, etc.

6) **Phylogenetic Tree Analyser:**
   Various trees built using phylogenetic tree builder will be analysed here using various parameters and the best tree is chosen which can give the best possible results.

7) **Phylogenetic Tree Visualiser:**
   Phylogenetic trees built shall be effectively presented here using phylogenetic tree visualiser. This will help in looking at the manuscript data in different perspective.

In this research project, we plan to develop Phylogenetic tree building modules such as Text Processor, Critical Apparatus Builder, Critical Apparatus Post Editor, Phylogenetic Data Builder, Phylogenetic Tree Builder, Phylogenetic Tree Analyser and Phylogenetic Tree Visualiser. This will help in finding efficiently and effectively the source and relationships of the manuscripts.

Project aims

The major aim of the project is to identify the most likely source of manuscripts. As we are unsure that which manuscript is the original among all the available manuscripts, using phylogenetic analysis we could be able to identify the most likely source of the text.

Other than the above objective there can be various other applications of computational phylogenetics. They are as follows:

- Collating manuscripts – compare and analyze the manuscripts and produce variations, this will help in the following tasks:
  - Identify the similar manuscripts, i.e. manuscripts having similar content.
  - Identify the variant readings, i.e. identifying the changes in manuscript. The changes could be omission, addition, replacement, etc
  - Grouping of manuscripts, i.e. the manuscripts which are repeatedly come together.
  - Find manuscripts which are never grouped together.
  - Find the dependency of grouped manuscripts.
  - Identify the relationships between manuscripts.
  - Identify the change of the word, change of the position of the word, etc.
  - Identify the semantic changes in the manuscripts, for example a small change in *hrasva dīrgha* may change the meaning of the word itself.
  - Identify the contaminated manuscripts, i.e. when two or more text versions being combined into one.
  - Identify the parallelism in manuscripts – Parallelism is the phenomenon that identical mistakes affect different lines of transmission.
  - Identify the originality of the text – Insertion and omission of certain words/phrases give rise to the spectacle about which would be the original text. E.g., In *Mahabharata's geeta* part, there are about 745 *shlokas*, while in *Bhagavad geeta* there are 700 *shlokas*. Here the question is which *shlokas* are original, the one from *Mahabharata* or the one from *Bhagvad geeta*?
  - Identify the *Ganapaath*. i.e., words of a particular time period. In manuscripts the words were included or excluded depending upon time period – e.g. some of the *vaidik* words which were used earlier is not being used these days.
- Automating variant reading identification, i.e. the process of Critical Apparatus Building can be automated by developing a tool/interface which can automatically identify the variant reading.
- Automatic phylogenetic tree builder. Manual tree reconstruction is still a challenge today; hence automating the most likely tree generation could be possible.
- Building phylogenetic tree visualizer to visualize resulting trees having thousands of species/variants**.**
- Building phylogenetic tree analyzer to analyze the variant readings and come up with the most likely source of the manuscript.
- Finding whether the phylogenetic analysis or tools used for DNA sequencing or similar techniques shall be used for variant reading analysis? That is, the variants used by the computer program to establish the branching of the tree really reveal the genealogical relationship of manuscripts.

Expected outcomes

The major outcomes of computational phylogenetics are as follows:

- Manuscript text processor
- Critical apparatus builder
- Phylogenetic data builder
- Phylogenetic tree builder
- Phylogenetic tree analyser
- Phylogenetic tree visualizer

How will the project address the Goals of the above Themes?

The computational phylogenetics falls under the theme '**Advanced computational engineering, simulation and manufacture'** and **'Humanities and Social Sciences'** among the 7 themes mentioned above by the IITB-Monash Academy. The computation phylogenetics involves research in the area of Artificial Intelligence, Machine Learning, Natural Language Processing along with Computational Linguistics along with the knowledge of classical languages like Sanskrit.

Capabilities and Degrees Required

List the ideal set of capabilities that a student should have for this project. Feel free to be as specific or as general as you like. These capabilities will be input into the online application form and students who opt for this project will be required to show that they can demonstrate these capabilities.

The minimum educational qualifications are:

Master's degree in Engineering/ Technology; or
Master's degree in Science; or
Master's degree in Computer Applications; or
Bachelor's degree in Engineering/Technology/Science and a valid GATE score

NOTE: Candidates without a valid GATE score or research fellowship can be considered for admission if they have a minimum of two years professional experience. Competition for places is high and a competitive selection process is applied to all applicants.

Potential Collaborators

Please visit the IITB website www.iitb.ac.in OR Monash Website www.monash.edu to highlight some potential collaborators that would be best suited for the area of research you are intending to float.

Please provide a few key words relating to this project to make it easier for the students to apply.

Computational Phylogenetics, Phylogenetics, Manuscripts, Variant Readings, Critical Apparatus, Phylogenetic Tree, Classical Languages, Natural Language Processing, Philology, Cladistics, Stemmatology, Phylogenetic Stemmatics.