Comparative Analysis of Ortholog and Paralog Detection

Pankaj Bhambri[1] and Om Parkash Gupta[2]

[1](Ph.D. Research Scholar, Department of Computer Science and Engineering, I.K.G. Punjab Technical University, Kapurthala, Punjab, India)

[2](Associate Professor & Head, School of Elect. Engineering & Information Technology, Punjab Agriculture University, Ludhiana, Punjab, India)

Abstract: *Orthology and paralogy are central concept in evolutionary biology. Ortholog prediction (homologous genes diverged due to speciation) and paralog prediction (homologous genes diverged due to duplication) is an integral part of many comparative genomics methods. There are number of methods used to infer orthology and paralogy. Ortholog and paralog can be inferred using graph based and phylogenetic based methods. In this paper, both these methods are discussed along with the hybrid approaches. Different tools available these days are discussed and compared. In order to robustly use comparative genomics-based analysis to draw conclusions about gene functions, we need to understand the correspondence between genes and to do this we need a consistent framework to describe genes' evolutionary relationships. The term homologs refer to any genes that share a common origin but they do not necessarily have the same purpose. Homology between two genes needs not to be expressed in terms of percentage. For instance, we can say that mouse and human are homologous or not homologous, but we cannot say that mouse and human are 3 or 4 percent homologous to each other.*

Keywords*: ortholog, paralog, graph, phylogenetic*

I. INTRODUCTION

A tree depicts the graphical relationship between organisms, species or genomic sequence. To make any estimation about relationships between species, phylogenetic tree is constructed which link the organism [1]. Current evolution theory depicted that all species on the planet have descended from a common ancestor and this type of relationship is known as a phylogeny which can be demonstrated by the phylogenetic trees which further diagrammatically shows the evolutionary history of species of interest. A phylogenetic tree about existing organism is constructed on the basis of three characteristics such as morphological, physiological and molecular. The "tree of life" stands for a phylogenies of every organism, extinct and living [2]. There are two types of tree: rooted tree and un- rooted tree. Rooted tree are those in which all other nodes are derived from a single node and un-rooted tree are those that are not invented from a single node [1]. Root depicts the origin of evolution. There is standard graph notation to represent the phylogenetic tree. Leaves represent the current species or organism. Branches show the relationship between the organisms or species. Length of the branch shows the evolutionary time [3].

There are two types of homologs: orthologs and paralogs. Orthologs are the homologs genes that diverged due to speciation in last common ancestors. They must have diverged from same ancestral gene in the LCA (Last Common Ancestor). During cell division, species are divided into two then the copies of a single allele in the resulting two species are supposed to be orthologous to each other. Paralogs are the homologous genes that have diverged due to the duplication event. When a gene in an organism is considered as duplicated, it can be located at two distant positions in the identical genome and then these two copies are said to be paralogs. With time these genes raise new function which may or may not be related to the original one [4]. Paralogs are further sub classified as: out-paralogs and in-paralogs. Out-paralogs are those in which duplication event occur before speciation event and in-paralogs are those in which duplication occur after speciation event [5]. The term co-ortholog is sometimes used to describe these many-to-many relationships where there are more than one valid orthologs due to a recent gene duplication event [6]. Orthologs can have two properties: Reflexive or Transitive. a) Reflexive: If A1 is ortholog of B1 then B1 is ortholog of A1. b) Transitive: If A1 is ortholog of B1 and B1 is ortholog of C1, then A1 is not necessarily ortholog of C1. When assigning orthology to genes in more than two species, the relationships can become complicated as orthology is non-transitive in nature. Non-transitive nature of orthologs state that, it is not necessarily that two genes in different species that are both orthologs to a gene in the third species are orthologs to each other [7]. Accurate prediction of orthologs is central to comparative genomics, as they have the identical function through speciation and can be used in comparative genomics to assess the function of unknown similar genes [8]. Figure1.1 shows hypothetical phylogenetic gene tree which highlights all kinds of relationships of homologous genes. Genes are represented by the boxes and in the LCA of species 1, species 2 as well as species 3, there are two ancestral genes α and β.

Genes α1 and α2 in species1 and 2 have both diverged from a common ancestral gene α in the LCA. These genes that diverged when the species diverged are example of orthologs. Genes α1 and β3 are examples of paralogs because they arose from a gene duplication of the αβ gene ancestor in the LCA. They are further classified as out-paralogs because the gene duplication occurred prior to the divergence of the species 1 and 3. When the gene duplication occurs subsequent to species divergence, for example genes α3a and α3b, they are classified as in-paralogs. In-paralogs can also satisfy the criteria for orthologs as they diverged from a common ancestral gene in the LCA. The term co-ortholog is used to describe this type of many-to-many ortholog relationship (genes β1a, β1b, β1c and β2a, β2b are examples of co-orthologs).

II. ORTHOLOG AND PARALOG DETECTION METHODS

Detection Based on Phylogenetic Tree

Phylogenetic tree-based methods identify orthologs and paralogs with the building phylogenetic trees and then reconciling the phylogenetic gene trees with a supplied species tree. There are numerous methods for building and analyzing phylogenetic trees. Broadly these are categorized as either distance based or character based methods. Distance based methods are also known as phonetic methods. The most commonly used distance based methods are UPGMA (Un-weighted Paired Group Method with Arithmetic Mean), NJ (Neighbor-Joining), ME (Minimum Evolution) and FM (Fitch- Margoliash) method. In character based methods, characters are called as informative sites or non-informative sites. There are two major types of cladistic methods: MP (Maximum Parsimony) method and ML (Maximum Likelihood) method. To achieve high-throughput prediction, the analysis steps must be automated. The steps in a typical tree-based ortholog prediction pipeline are outlined in [9]–[11]. Gene sequences for the phylogenetic tree must be collected. Most programs use a sequence similarity search such as BLAST to select which genes from the genomes to use in the tree. After obtaining the gene sequences, a multiple sequence alignment must be constructed. Ideally, low quality regions in the alignment are filtered out. Then a phylogenetic tree is built from the multiple sequence alignment. Tree building methods fall into two types: distance based (e.g. UPGMA, neighbor-joining) and character-based (e.g. maximum parsimony). Distance based trees are computationally less intensive to build, but they are also less accurate in many situations [11]. Some ortholog prediction tools use methods such as bootstrapping [12] or build a consensus tree from multiple distinct tree building methods to improve confidence in the final tree [13]. Other methods examine features such as branch length to identify potential problems [9]. Then gene tree is reconciled with a supplied species tree. By inferring duplication and gene loss events along the branches, the gene tree is made to match the species tree. Once a reconciled gene tree is produced, orthologs and paralogs are inferred by examining the presence or absence of duplication events between the leaves of the tree.
A gene tree that is congruent with the species tree is inferred to contain orthologs. Generally, a phylogenetic approach is preferred for identifying orthologs [10], [11]. The methods can be less prone to error in presence of gene duplication and loss. However, tree-based methods have several limitations. There is usually a significant computational cost associated with phylogenetic tree building. Most tree-based methods do not scale well as the number of genomes and sequences increases [10]. The source Meta PhOrs is one of the largest repositories for orthologs prediction using phylogenetic tree-based methods. It contains ortholog predictions across 829 genomes. This scale of ortholog prediction is achieved by mining pre-computed phylogenetic gene trees from other sources [14]. Another limitation of the tree-base methods is the sensitivity of automated tree building to biases in the data or errors in the multiple sequence alignment. For example, a well-documented issue with tree building is long-branch attraction. Long branches are grouped together in a maximum parsimony approach even though they may not be taxa [15]. Finally, the requirement for a species tree is a significant limitation of tree-based methods. In many methods, the species tree must be a perfectly resolved, rooted, bi-furcated tree. A few tool shave developed methods for computing the root or working with multi-furcating trees [16].

Detection Based on Graph

Graph-based methods use pair-wise sequence similarities to predict orthologs (paralogs are not explicitly detected in this approach). The formative premise behind the graph-based methods is that because the orthologs diverged from a common ancestral gene, they should appear as the reciprocally most similar genes from the irrespective genomes. Because of its speed, BLAST is the most commonly used method for determining the best match. In the typical approach, BLAST will be run twice, using each genome as the query and subject. The top BLAST hits will be recorded for each gene in the genomes and orthologs are declared as the pair of genes that are the reciprocal of best BLAST hits of each other (this approach is referred to as RBB). Existing implementations can include different BLAST metrics to determine top hit E-value, percent identity or bit score, and how ties or multiple top hits are dealt with [10]. In practice, the RBB procedure works well and can outperform more complex ortholog predictions methods in some cases [17]. Most of the graph-based ortholog tools use RBB but then incorporate additional steps to improve accuracy, detect other types of homologous genes, or perform multi-species ortholog prediction. Graph-based approaches are computationally less intensive and more straightforward to automate, making them comparatively proficient at computing

orthologs for large datasets [10]. In Reciprocal Best Blast Hit (RBBH), queries are executed in both the directions for ortholog detection. This approach removes the poor BLAST matches but needs an entire set of genomes to work correctly. Some methods make the use of evolutionary distance to find the nearest gene instead of using similarities scores. Among any two genome, orthologs are the homologs with minimum divergence. Reciprocal Smallest Distance (RSD) confirms the two genes as orthologs considering the minimum evolutionary distance between them. RSD also includes the pairs which are excluded by RBBH. However, RSD is time consuming and more complex than RBBH. Its major drawback is the limited phylogenetic view (it only considers the top matches). RBB or other similar methods can fail in situations where there is differential gene loss or recombination of protein domains that can change the top hit to a non-orthologous gene [18].

Hybrid Based Methods

Ortholog prediction approaches have been developed that use elements of both the phylogenetic tree-based methods and the graph-based methods. In most tools that can be considered a hybrid approach, involves using a RBB procedure to identity candidates for phylogenetic gene tree building step [13], [19]. An example of a hybrid approach is the Phylogenetic Orthologous Groups (PHOGs) database, which uses a species tree to guide the creation of ortholog groups. Groups are constructed using RBB in a hierarchical fashion, starting with the most closely related species and moving up the species tree to include species with a larger phylogenetic range. Orthologs from the earlier groups are used as seeds in the later groups.

In protein network comparison method, interaction division conserved between two species is calculated. Massive false positives can be removed by relying on relations that is conserved between two species. Synteny is a locally conserved gene order that is identified in two or more genomes. Two species that have recently diverged from a common ancestor are expected to share similar set of genes. Synteny based methods explore the neighborhood of genes of interest for inferring orthologs. The portion of orthologs which have neighboring gene themselves forms the consistency marker [8]. Synteny analysis helps to examine the functions and structures of genes groups in the genome and their role in gene expression. Earlier, Synteny was used to illustrate co-localization of unlike genes in corresponding chromosomes of dissimilar species. But, now Synteny is also used to reveal the conservation of co-localization gene in the identical order within distinct genomes [20]. Synteny analysis is broadly used in plant comparative genomes. Plants genomes are highly complex and have variable size. Plant species enormously diverse in their growth habits, environment adaptation and nuclear genome structure [21]. Synteny blocks are genomic segments and it consist of a set of orthologous genes that share the same relative ordering on the chromosomes of two species [22]. Genes that are in Synteny block are mostly co-regulated and share identical functions. Due to gene's functional significance, there may be some selective forces that prohibit genes within it from escaping the blocks. A synteny block may become complex if it forms various types of functional clusters and topological arrangement [20], [23].

Furthermore, experimentation revealed the relationship between different plant species. Synteny between Arabidopsis (Athaliana) and tomato explored synteny between Arabidopsis and rice (O.Sativa) [24]. Whereas synteny between soya bean and Arabidopsis [25] explored the synteny between common bean and soya bean [26]. Synteny blocks are furthered classified in two blocks: conserved block and non conserved block. A conserved synteny block has block of genes that preserves the ordering and there is no mismatch within the block. A non conserved synteny block preserves the strandness of block of genes but has inequality within the block. Adhoc methods are used to find synteny blocks. Adhoc methods (comparatively slow and not fully reproducible) ignore strandness and are inappropriate for general application. Computational approaches are more effective with utilization of efficient algorithm. Orthocluster is one such method which is designed to identify synteny block of multiple genome. Here the orthologous relationships are available among the input genomes [20]. Orthocluster is a data mining tool for inferring synteny block, among multiple genomes. Firstly orthologous gene relationships between the interested species are identified. Thereafter, Orthocluster is used to infer synteny blocks. Inparanoid and Adhoc BLAST methods are used to identify orthologous genes, on the number, size and content of synteny blocks returned by Orthocluster using the oryzasavita and Arabidopsis thaliana genomes. In the identification of orthologous pair, three complementary methods are used to detect orthologous pairs. In the first method, BlastX is carried out to find mutually best hits as orthologs. In second method, orthologous which are missed by the first methods are identified using synteny anchors and synteny blocks. Lastly to detect orthologs, protein functional classification is used. These methods are able to identify orthologous pairs that were missed by mutually best hit approach [22].

III. TOOLS FOR DETECTION OF ORTHOLOG AND PARALOG

For the study of set of organisms with their gene families, detection & analysis work on orthology and paralogy is performed. This study yields the tremendous computational platforms for the identification of significant horizontal gene transfers. Analysis of orthologs becomes difficult among the protein families having large number of paralogs. Eukaryotic genome enhances some challenges to orthology analysis because the large

size of their genome, difficulty in defining accurate gene models, high number of gene duplication and complexity of protein domain architecture [27].

COCO-CL: Corelational Coefficient based Clustering method is used for the identification of genes orthologous groups and for hierarchical clustering of homology relations. Evolutionary distance to cluster is used by many hierarchical clustering methods but COCO-CL get advantages of the global topology of correlation network and search correlation of evolutionary historical data [28]. Gene tree is not used directly by this method, however it implies the use of evolutionary data, in the form of evolutionary distance matrix. The original COG cluster encloses one prokaryotic, one archaea and three eukaryotic organisms. Its concept is extended to eukaryotic genome TOGA [29], KOG [30], OrthoMcl [31] RIO: Resample inference of orthologs to detect orthologs using gene tree and species tree. To find the orthologs from phylogenetic tree, it adds bootstrapping [16]. There are the three concept super orthologs, ultra orthologs and sub-tree neighbors. Pfam (protein family) database was used as source of high quality multiple sequence alignment. Pfam alignments of domain in plants and worms [12]. RAP is another phylogenetic method for the detection of orthologs and paralogs. This algorithm improves upon traditional reconciliation method. With reference to n-array nodes which come across in species trees, it improves the congruence function. In RAP algorithm, single node is considered to be speciation until there is strong evidence that there is incongruence in gene tree and species tree. This algorithm takes into account topology, bootstraps and branch lengths. It is fast for reconciliation of large sets of phylogenetic trees [9]. LOFT: Levels of Orthology From Tree makes various hierarchical grouping that emphasizes different levels of relations between orthologs and paralogs. It is an alternative approach to gene/species reconciliation approach. This method introduces a technique named species-overlap for inferring speciation and duplication event from the gene tree. It does not need species tree. It considers a node to represent gene duplication, if its branches have overlapping sets of species. It is benchmarked against COG, reconstruction with trusted species tree and gene order conservation [32]. Tree Fam: it is a phylogenetic trees database for all the animal's gene families and for every TreeFam family. It provides homology prediction along with the evolutionary history. The sequences search is based on HMM, which place protein sequences into the TreeFam gene and results fast orthology identification. It uses the RAXML and Mafft tool for quick insertion into a tree and references alignment respectively. It contains twenty five sequenced animal genomes and four plant plus fungi relatives out group species [33]. It gives the prediction of orthology and tree for 109 species among 1536 families. It covers approximately two million sequences. A new version of TreeFam with added extra feature was released. The number of species was increased in TreeFam 9 from 79 to 109 [34]. HOPS: Hierarchical grouping of Orthologous and Paralogous Sequences is designed by storm and son hammer. By analyzing multi set of boots rapper trees, it assigns orthology. It uses a heurist based on sequences similarity. The sequences of tree are classified into four different categories ingroup1 and ingroup2 for two species of interest, out group and blank. The boots rapper method was typically applied to increase the accuracy of a statistical estimation. An advantage of this method is that all pair wise orthology assignments are involved to assign scores. Thus it does not allow orthologous relationships to be available in the original assessed tree. It has drawback that incomplete genome or partial gene loss may result in incorrect ortholog prediction [7]. It is applied to the prokaryotes and eukaryotes that appears in Pfam [35]. Metaphors: Meta phylogeny based orthologs is a public repository of phylogeny based on the paralogs orthologs. These orthologs are computed using phylogenetic tree available in twelve public repositories. In addition to it, maximum likelihood trees are constructed using protein families stored in OrthoMCL as well as from alignments, etc [36]. The pipeline of metaphors generally follow the procedure: all phylogenetic trees that contain any given pair of sequences are fetched, tree generated by suboptimal evolutionary models are removed by a filtering step. A subsequently species overlap algorithm [37] is implemented in ETE toolkit [38] and a consistency score is also calculated depending on orthology and paralogy tree. This method is applied to hundreds of genomes from eukaryotes and prokaryotes. PhylomeDB: It is a database for complete catalogs of phylogenies and was published in 2006. It is a repository of complete gene evolutionary histories encode in a genome [14], [39]. It gives trees and alignments enriched with appropriate observations along with prediction of homology relationships [13], [40]. It follows both gene-centric and genome- wide approaches [14]. Current release comprises 17 phylomes from different organism as human, bacteria and yeast. PhylomeDB is one of the biggest repositories of precomputed phylogenies and gives evolutionary computation for greater than 10 millions protein on approximately thousand species. In 42 of new phylomes, PhylomeDB assures the coverage of proteomes from the search for orthologs initiatives. Phylome database consist the large number of new phylomes. Version 4 of PhylomeDB provides reasoned sets of homology predictions using most up to date release of uniformity based prediction from the metaphorsDB [14]. PHOG: Berkely Phylofacts Orthology Group is method to detect orthologs. It does not use reconciliation approach for ortholog detection. It uses TreefamA, manually curate datasets as a benchmark [41]. Precision level and different taxonomic units are targeted via user set tree distance threshold in a web server. It uses pre computed trees [11]. PHOG-S is used for super orthologs and PHOG-O is used for standard orthologs. PHOG-T allows the user to control the tree distance threshold. PHOG method is applied to human mouse, zebra

fish and fruit fly sequence from TreeFam-A [41]. GreenPhyl: It is implemented for gene sequences having full genomes as input dataset. It was created from raw data using semi-automated gene family clustering implementation before the tree construction [11]. It is applied to all plant genomes [42]. COG: Clusters of Orthologous Groups discover three ways. BBH between orthologs sets co-orthologs in different species and these groups are extended until saturation, followed by manual ripping of large groups. They are inappropriately joined by multi domain proteins or difficult mixture of in and out paralogs. Later progress focuses on increasing the resources [43] and more proficient managing the large amount of genomes [44]. COG is applied to completely available sequenced genomes [45], recently comprises 631 genomes [46]. OrthoMCL: It is a clustering algorithm based on graph. It is developed to discover homologous protein based on sequences similarity. It differentiates relationship of orthologs from paralogs without computationally intensive phylogenetic analysis. This method makes groups of orthologs by a markov chain concerning iterative simulations, with cluster of desired rigidity identifies by trails and error [31]. The current OrthoMCL-database releases 4 have 138 genomes of mostly eukaryotes and prokaryotes. Protein sequence of 511797 out of 627098 were clustered into 70388 orthologs group [47]. Conventional approaches used to identification of orthologs, OrthoMCL group tend to be small. Sometime TribeMCL could be useful for finding ancient out paralogs for interested genes [48]. Inparanoid: It is an algorithm to generate groups of ortholog that contains all in-paralogs but no out-paralogs. It is graph based which begins with an extensive all vs. all (BLAST) alignment of protein sequences and then applies clustering rules to construct orthologs group [6], groups of genes between species based on phenotype identified by Inparanoid [49]. FSRD is a database of 1985 fungal stress response protein and uses Inparanoid to predict orthologs in 28 species along with human, pathogen and fungi [50]. Inparanoid orthologs is used to forecast either human genes are created by retro position [51] or not. Inparanoid is a process to pipeline so as to find Sno-RNA bearing host orthologous genes across eukaryote genomes [52]. Inparanoid 6 have 34 eukaryotic species and one prokaryotic out groups [53]. Inparanoid 8 include 213 species. These have 246 eukaryotes, 20 bacteria and 7 archaea [54]. OMA: Orthologous MAtrix is a database and method for the implication of ortholog among complete genome. OMA has wide scope, size and high quality of inferences. OMA handles splice variants. The longest variant is retained and the shortest variants are retrieved only when they differ at least 10% from the retained longer variant. This reduces the total sequences in the database. A heuristic was developed in which the variant with the highest number of genome matches in the all-against-all step is selected. It is difficult to evaluate orthology inference, however they find that new process is more inferred orthologous pairs to large OMA group [17]. OMA introduces a graph based technique for detecting hierarchical orthologous group with reconciliation. There are many major developments in OMA, first is new web interface, then gene orthology function prediction. Third is good support for genomes of plants and in particular homologs in the wheat genome. Next is a synteny analysis, lastly is statically computed hierarchal orthologous group subset in orhtoxml format [55]. Quartet S: It is another method that uses evolutionary evidences in a computationally capable manner. It correctly anticipates orthologs and paralogs by directing to use evolutionary evidences of duplication events in a quartet gene tree. This method is also useful for large scale detection of orthologs. Another variation of Quartet S was also defined and was named QuartetS-C that includes Quartet S along with clustering. Quartet S method is slightly superior to OMA 2008. Difference between Quartet S and Quartet S-C reveals that performance of ortholog detection is slightly improved after post processing clustering [56]. Round up: It is online database of orthology and their evolutionary distances. Round up 2.0 databases of orthologous genes for 1800 genomes, containing 226 eukaryotes, 1581 prokaryotes. Reciprocal Smallest Distance algorithm used to infer orthologous. Result of a noted query may be viewed in a different way. Genomic result may be downloaded in format suitable for functional as well as phylogenetic analysis [57]. OrthoInspector: This is the Software includes an algorithm for the fast detection of orthology and paralogy relation between species. In comparison with another method this improves sensitivity with a minimal loss to specificity. OrthoInspector software is used to study 59 eukaryotic species. It is easy and fast to use data management tool as an algorithm to generate fast sensitive anticipation of orthologyand inparalogy [58]. EchinoDB: It is a database that consist amino acid sequence ortho cluster from 42 Echinoderm transcriptome. Echinoderm is used to find orthologs suitable for phylogenetic analyses from next generation transcriptome data. RNA sequence is used to profile adult issue from echinoderm. EchinoDB is a repository of orthologous transcripts from echinoderm [59]. DODO: Domain based ortholog detection is a functional based new ortholog detection method to overcome the problems of identification of ortholog from a large number of genomes. It uses domain information to find orthology in remotely related genomes. DODO works into two step: firstly, on based of domain architecture it assigns protein in group secondly, within these groups, it identifies orthologs with much less complexity. Performance of DODO is directly affected by accuracy of domain identification [60]. FAT-CAT: Fast Approximation Tree Classification is a method of ortholog identification. It uses sub tree HMM scoring. Precision of FAT- CAT is enhanced with the usage of HMM at each node of each tree. Input for FAT-CAT is a gene sequence and output is a list of orthologs for that gene. According to the level of stringency, four different parameters are provided. It uses trees from the Phylofacts database. For checking whether the sub

tree identified is supported by different methods, additional third party orthology data is used by the method. Four stages are used for ortholog detection. Different parameters are used at different stages for the accuracy of ortholog detection. This method is able to provide accurate results when the query sequences have promiscuous domains. Disadvantage of this method is that it is computationally intensive and complex. FAT-CAT is slower than other ortholog web server. But, a fast variant of FAT-CAT that is FAST-CAT has also been designed. FAST-ACT is similar to FAT-CAT, except in the third stage where it avoid the computationally complex pair-wise alignments [61]. EggNOG: Evolutionary genealogy of gene Non-Supervised Orthologous Groups is database having orthologous groups predicted from Smith-Waterman alignment. The orthologous groups in EggNOG contain 1241751 genes [45]. This contains 2242035 proteins and gives a broad functional description for at least 88 percentages of them [62]. The third release of EggNOG includes non supervised orthologous group constructed from 1133 organism. In this release, search of homologous is based on SIMAP and group of orthologous is extended to 41 levels of selected taxonomic. EggNOG V3 contain 721801 orthologous group including a total of4396591 gene [63]. Fourth version of EggNOG database develops non supervised ortholog groups from full genome based on characterization and examine pipeline to resulting gene families. In comparison to previous release of EggNOG, V4 contains tripled the underlying species set to cover 3686 organisms. It also gives multiple sequences alignment and maximum likelihood. It provides orthologous group more precisely than previous [64]. Finally, EggNOG4.5 integrates a novel data set protein. It is most scalable and complete data base for prediction of ortholog [65]. OrthoAgogue: It is a tool for high speed prediction of homology relations within and between the species. OrthoAgogue perform much faster than OrthoMCL. It follows the identical in-paranoid algorithm as OrthoMCL, some variation are done to provide some flexibility. Only the best high score pairs in BLAST output for a given pair of sequences are used. Additional use of HSP helps to differentiate between same set of sequences. OrthoAgogue is an extremely efficient and flexible tool [66]. OrthoVenn: It is web platform to compare and annotate the ortholog gene cluster among multiple species. OrthoVenn include vertebrate, metazoan, plants, fungi and so many to identify orthologous gene cluster. It also provides Venn diagram for comparing two six species protein sequences. It allows for the identification orthologous cluster of single copy genes. It is efficient and user friendly, allow to examining and assigning the biological meaning of orthologous genes [67]. HCOP: HGNC Comparison of Orthology Prediction search tool provides an integrated database of orthology prediction. HCOP is a human centered ortholog prediction tool. It gives the ability to search orthologs between human and any other species. It also provides reciprocal search. The data has been taken from different methods including OMA, Inparanoid and TreeFam and only provides prediction relevant to human genes. YOGY: EukarYoticOrtholoGY is a web based integrated method to retrieve orthologs of eukaryotes. It merges orthology prediction from five different resources. Here, Query to web server shows the summary of genes as wella's orthologs prediction table from different methods. Gene Orthology term is also shown for functional inferences [68]. MetaPhors: Meta Phylogeny based Orthologs is a phylogenetic method that merges the prediction of different phylogenetic methods as raw data. Trees are obtained from different methods like TreeFam and PhylomeDB. In addition to it, maximum likelihood trees are constructed from multiple alignments or families of proteins stored in OrthoMCL, COG [36]. Species-overlap algorithm is used to predict orthology and paralogy. A consistency score is also calculated depending on the number of trees predicting orthology and paralogy [36]. DIOPT: Drosophila RNAi screening centre Integrative Ortholog Prediction combines mouse, fly, human, yeast and zebra fish ortholog prediction made by Inparanoid, OMA, Roundup, TreeFam and many other methods. It allows detecting orthologs of a gene in the selected output species. DIOPT indicates number of methods that support the anticipated orthologous gene pair and a weighted score based on functional assessment using high quality GO annotation and evaluate a simple score. It also shows protein and domain alignments along with percentage identity for anticipated ortholog pair. It helps to select the most appropriate matches among multiple possible orthologs [69].

IV. CONCLUSION

Detection of ortholog and paralog is a crucial task in comparative genomics. The methods that are used to infer pairs of orthologous and paralogous falls into two classes: tree based and graph based. Synteny can be used in orthologs identification. The tree based and graph based technique often have similar set of prediction orthologs, with different generally because of choice of speciation and duplication event used to express in paralogs as well as co-ortholog. Tree based methods are more specific and graph based methods are more sensitive. Tree based methods are computationally expensive and they do not perform well in horizontal gene transfer. However they employ explicit evolutionary models. Data set which is large in size mostly occurs in case of prokaryotes. Here, evolutions do not follow a simple tree method, and graph based methods are more suitable. To analyze orthology, several resources are available. In this paper, we have discussed the advantages and disadvantages of resources and also analyzed the quality of many tools. Table 1.1 elaborates the specification of different methods. The usage of tool depends upon the requirements of the user. The authors declare that there is no conflict of interest regarding the publication of this paper.

V. REFERENCES

[1] *J. Rizzo and E. C. Rouchka, "Review of Phylogenetic Tree Construction Review of Phylogenetic Tree Construction," Bioinforma. Rev., 2007.*

[2] *K. Dowell, "Molecular phylogenetics: an introduction to computational methods and tools for analyzing evolutionary relationships," Mol. Phylogenetics, pp. 1–19, 2008.*

[3] *B. G. Hall, "Phylogenetic Trees Made Easy: A How-to Manual," vol. 96, no. 4, pp. 469–470, 2011.*

[4] *S. Kim, "Clustering Methods for Finding Orthologs among Multiple Species Species," Thesis, no. August, 2007.*

[5] *M. D. Whiteside, "Computational Ortholog Prediction : Evaluating Use Cases and Improving High-Throughput Performance by," 2013.*

[6] *M. Remm, C. E. Storm, and E. L. Sonnhammer, "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.," J. Mol. Biol., vol. 314, no. 5, pp. 1041–52, 2001.*

[7] *C. E. V Storm and E. L. L. Sonnhammer, "Automated ortholog inference from phylogenetic trees and calculation of orthology reliability.," Bioinformatics, vol. 18, no. 1, pp. 92–99, 2002.*

[8] *J. Jun, I. I. Mandoiu, and C. E. Nelson, "Identification of mammalian orthologs using local synteny.," BMC Genomics, vol. 10, p. 630, 2009.*

[9] *J. F. Dufayard, L. Duret, S. Penel, M. Gouy, F. Rechenmann, and G. Perrière, "Tree pattern matching in phylogenetic trees: Automatic search for orthologs or paralogs in homologous gene sequence databases," Bioinformatics, vol. 21, no. 11, pp. 2596–2603, 2005.*

[10] *A. Kuzniar, R. C. H. J. van Ham, S. Pongor, and J. A. M. Leunissen, "The quest for orthologs: finding the corresponding gene across genomes," Trends Genet., vol. 24, no. 11, pp. 539–551, 2008.*

[11] *D. M. Kristensen, Y. I. Wolf, A. R. Mushegian, and E. V. Koonin, "Computational methods for Gene Orthology inference," Brief. Bioinform., vol. 12, no. 5, pp. 379–391, 2011.*

[12] *C. M. Zmasek and S. R. Eddy, "RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs.," BMC Bioinformatics, vol. 3, p. 14, 2002.*

[13] *A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney, "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates," Genome Res., vol. 19, no. 2, pp. 327–335, 2009.*

[14] *J. Huerta-Cepas, S. Capella-Gutierrez, L. P. Pryszcz, I. Denisov, D. Kormes, M. Marcet-Houben, and T. Gabaldon, "PhylomeDB v3.0: An expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions," Nucleic Acids Res., vol. 39, no. SUPPL. 1, pp. 556–560, 2011.*

[15] *T. O'Connor, K. Sundberg, H. Carroll, M. Clement, and Q. Snell, "Analysis of long branch extraction and long branch shortening.," BMC Genomics, vol. 11 Suppl 2, no. Suppl 2, p. S14, 2010.*

[16] *C. M. Zmasek and S. R. Eddy, "A simple algorithm to infer gene duplication and speciation events on a gene tree.," Bioinformatics, vol. 17, no. 9, pp. 821–828, 2001.*

[17] *A. M. Altenhoff, R. A. Studer, M. Robinson-Rechavi, and C. Dessimoz, "Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs," PLoS Comput. Biol., vol. 8, no. 5, 2012.*

[18] *D. L. Fulton, Y. Y. Li, M. R. Laird, B. G. S. Horsman, F. M. Roche, and F. S. L. Brinkman, "Improving the specificity of high-throughput ortholog prediction.," BMC Bioinformatics, vol. 7, no. 1, p. 270, 2006.*

[19] *H. Li, A. Coghlan, J. Ruan, L. J. Coin, J.-K. Hériché, L. Osmotherly, R. Li, T. Liu, Z. Zhang, L. Bolund, G. K. Wong, W. Zheng, P. Dehal, J. Wang, and R. Durbin, "TreeFam: a curated database of phylogenetic trees of animal gene families.," Nucleic Acids Res., vol. 34, no. Database issue, pp. D572–80, 2006.*

[20] *I. A. Vergara and N. Chen, "Large synteny blocks revealed between Caenorhabditis elegans and Caenorhabditis briggsae genomes using OrthoCluster.," BMC Genomics, vol. 11, p. 516, 2010.*

[21] *J. L. Bennetzen, "Comparative Sequence Analysis of Plant Nuclear Genomes: Microcolinearity and Its Many Exceptions," Plant Cell, vol. 12, no. July, pp. 1021–1030, 2000.*

[22] *X. Zeng, M. J. Nesbitt, J. Pei, K. Wang, I. a Vergara, and N. Chen, "OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics," EDBT '08 Proc. 11th Int. Conf. Extending database Technol., pp. 656–667, 2008.*

[23] *I. A. Vergara and N. Chen, "Using orthoCluster for the detection of synteny blocks among multiple genomes," Current Protocols in Bioinformatics, no. SUPPL. 27. pp. 1–18, 2009.*

[24] *J. Salse, B. Piégu, R. Cooke, and M. Delseny, "Synteny between Arabidopsis thaliana and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project.," Nucleic Acids Res., vol. 30, no. 11, pp. 2316–28, 2002.*

[25] J. L. Shultz, J. D. Ray, and D. A. Lightfoot, "A sequence based synteny map between soybean and Arabidopsis thaliana.," BMC Genomics, vol. 8, p. 8, 2007.

[26] P. E. McClean, S. Mamidi, M. McConnell, S. Chikara, and R. Lee, "Synteny mapping between common bean and soybean reveals extensive blocks of shared loci.," BMC Genomics, vol. 11, p. 184, 2010.

[27] F. Chen, A. J. Mackey, J. K. Vermunt, and D. S. Roos, "Assessing performance of orthology detection strategies applied to eukaryotic genomes," PLoS One, vol. 2, no. 4, 2007.

[28] A. Manuscript, "evolutionary correlations," Bioinformatics, vol. 22, no. 7, pp. 779–788, 2006.

[29] Y. Lee, R. Sultana, G. Pertea, J. Cho, S. Karamycheva, J. Tsai, B. Parvizi, F. Cheung, V. Antonescu, J. White, I. Holt, F. Liang, and J. Quackenbush, "Cross-referencing eukaryotic genomes: TIGR orthologous gene alignments (TOGA)," Genome Res., vol. 12, no. 3, pp. 493–502, 2002.

[30] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale, "The COG database: an updated version includes eukaryotes.," BMC Bioinformatics, vol. 4, p. 41, 2003.

[31] L. Li, C. J. Stoeckert, and D. S. Roos, "OrthoMCL: Identification of ortholog groups for eukaryotic genomes," Genome Res., vol. 13, no. 9, pp. 2178–2189, 2003.

[32] R. T. J. M. van der Heijden, B. Snel, V. van Noort, and M. A. Huynen, "Orthology prediction at scalable resolution by phylogenetic tree analysis.," BMC Bioinformatics, vol. 8, p. 83, 2007.

[33] J. Ruan, H. Li, Z. Chen, A. Coghlan, L. J. M. Coin, Y. Guo, J. K. Heacute;riche, Y. Hu, K. Kristiansen, R. Li, T. Liu, A. Moses, J. Qin, S. Vang, A. J. Vilella, A. Ureta-Vidal, L. Bolund, J. Wang, and R. Durbin, "TreeFam: 2008 Update," Nucleic Acids Res., vol. 36, no. SUPPL. 1, pp. 735–740, 2008.

[34] F. Schreiber, M. Patricio, M. Muffato, M. Pignatelli, and A. Bateman, "TreeFam v9: A new website, more species and orthology-on-the-fly," Nucleic Acids Res., vol. 42, no. D1, pp. 922–925, 2014.

[35] C. E. V Storm and E. L. L. Sonnhammer, "Comprehensive Analysis of Orthologous Protein Domains Using the HOPS Database Comprehensive Analysis of Orthologous Protein Domains Using the HOPS Database," Genome Res., vol. 13, no. 10, pp. 2353–2362, 2003.

[36] L. P. Pryszcz, J. Huerta-Cepas, and T. Gabaldón, "MetaPhOrs: Orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score," Nucleic Acids Res., vol. 39, no. 5, 2011.

[37] J. Huerta-Cepas, H. Dopazo, J. Dopazo, and T. Gabaldón, "The human phylome.," Genome Biol., vol. 8, no. 6, p. R109, 2007.

[38] J. Huerta-Cepas, J. Dopazo, and T. Gabaldón, "ETE: a python Environment for Tree Exploration.," BMC Bioinformatics, vol. 11, no. 1, p. 24, 2010.

[39] J. Huerta-Cepas, A. Bueno, J. Dopazo, and T. Gabaldón, "PhylomeDB: A database for genome-wide collections of gene phylogenies," Nucleic Acids Res., vol. 36, no. SUPPL. 1, pp. 491–496, 2008.

[40] S. Penel, A. Arigon, J. Dufayard, A.-S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière, "Databases of homologous gene families for comparative genomics.," BMC Bioinformatics, vol. 10 Suppl 6, p. S3, 2009.

[41] R. S. Datta, C. Meacham, B. Samad, C. Neyer, and K. Sjölander, "Berkeley PHOG: PhyloFacts orthology group prediction web server," Nucleic Acids Res., vol. 37, no. SUPPL. 2, pp. 84–89, 2009.

[42] M. G. Conte, S. Gaillard, G. Droc, and C. Perin, "Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants.," BMC Genomics, vol. 9, p. 183, 2008.

[43] R. L. Tatusov, E. V Koonin, and D. J. Lipman, "A genomic perspective on protein families.," Science, vol. 278, no. 5338, pp. 631–637, 1997.

[44] D. M. Kristensen, L. Kannan, M. K. Coleman, Y. I. Wolf, A. Sorokin, E. V Koonin, and A. Mushegian, "A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches," Bioinformatics, vol. 26, no. 12, pp. 1481–1487, 2010.

[45] L. J. Jensen, P. Julien, M. Kuhn, C. Von Mering, J. Muller, T. Doerks, and P. Bork, "eggNOG : automated construction and annotation of orthologous groups of genes," Nucleic Acids Res., vol. 36, no. October 2007, pp. 250–254, 2008.

[46] K. S. Makarova et al., "Comparative genomics of the lactic acid bacteria.," Proc. Natl. Acad. Sci. U. S. A., vol. 103, no. 42, pp. 15611–6, 2006.

[47] F. Chen, A. J. Mackey, C. J. Stoeckert, and D. S. Roos, "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.," Nucleic Acids Res., vol. 34, no. Database issue, pp. D363–8, 2006.

[48] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," Nucleic Acids Res., vol. 30, no. 7, pp. 1575–1584, 2002.

[49] J. O. Woods, U. M. Singh-Blom, J. M. Laurent, K. L. Mcgary, and E. M. Marcotte, "Prediction of gene–phenotype associations in

humans, mice, and plants using phenologs," BMC Bioinformatics, vol. 14, pp. 1–17, 2013.

[50] Z. Karanyi, I. Holb, I. Hornok, Laszlo Pocsi, and M. Miskei, "FSRD: Fungal stress response database," Database, vol. 2013. pp. 1–6, 2013.

[51] J. Ciomborowska, W. Rosikiewicz, D. Szklarczyk, W. Makalowski, and I. Makalowska, "'Orphan' retrogenes in the human genome," Mol. Biol. Evol., vol. 30, no. 2, pp. 384–396, 2013.

[52] M. P. Hoeppner and A. M. Poole, "Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility," BMC Evol. Biol., vol. 12, no. 1, p. 183, 2012.

[53] A. C. Berglund, E. Sjölund, G. Östlund, and E. L. L. Sonnhammer, "InParanoid 6: Eukaryotic ortholog clusters with inparalogs," Nucleic Acids Res., vol. 36, no. SUPPL. 1, pp. 263–266, 2008.

[54] E. L. L. Sonnhammer and G. Ostlund, "InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic," Nucleic Acids Res., vol. 43, no. D1, pp. D234–D239, 2015.

[55] A. M. Altenhoff, N. Šunca, N. Glover, C. M. Train, A. Sueki, I. Piližota, K. Gori, B. Tomiczek, S. Müller, H. Redestig, G. H. Gonnet, and C. Dessimoz, "The OMA orthology database in 2015: Function predictions, better plant support, synteny view and other improvements," Nucleic Acids Res., vol. 43, no. D1, pp. D240–D249, 2015.

[56] C. Yu, N. Zavaljevski, V. Desai, and J. Reifman, "QuartetS: A fast and accurate algorithm for large-scale orthology detection," Nucleic Acids Res., vol. 39, no. 13, pp. 1–10, 2011.

[57] T. F. Deluca, J. Cui, J. Y. Jung, K. C. St. Gabriel, and D. P. Wall, "Roundup 2.0: Enabling comparative genomics for over 1800 genomes," Bioinformatics, vol. 28, no. 5, pp. 715–716, 2012.

[58] F. Chester, "How to sell services more profitably," Harv. Bus. Rev., vol. 86, no. 12, p. 115, 2008.

[59] D. A. Janies, Z. Witter, G. V. Linchangco, D. W. Foltz, A. K. Miller, A. M. Kerr, J. Jay, R. W. Reid, and G. A. Wray, "EchinoDB, an application for comparative transcriptomics of deeply-sampled clades of echinoderms," BMC Bioinformatics, pp. 1–6, 2016.

[60] T. Chen, T. H. Wu, W. V Ng, and W. Lin, "DODO: an efficient orthologous genes assignment tool based on domain architectures. Domain based ortholog detection.," BMC Bioinformatics, vol. 11 Suppl 7, no. Suppl 7, p. S6, 2010.

[61] C. Afrasiabi, B. Samad, D. Dineen, C. Meacham, and K. Sjölander, "The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification.," Nucleic Acids Res., vol. 41, no. Web Server issue, pp. 242–248, 2013.

[62] J. Muller, D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. Von Mering, T. Doerks, L. J. Jensen, and P. Bork, "eggNOG v2.0: Extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations," Nucleic Acids Res., vol. 38, no. SUPPL.1, pp. 190–195, 2009.

[63] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. Von Mering, and P. Bork, "eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges," Nucleic Acids Res., vol. 40, no. D1, pp. 284–289, 2012.

[64] S. Powell, K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, T. Gabaldón, T. Rattei, C. Creevey, M. Kuhn, L. J. Jensen, C. Von Mering, and P. Bork, "EggNOG v4.0: Nested orthology inference across 3686 organisms," Nucleic Acids Res., vol. 42, no. D1, pp. 231–239, 2014.

[65] J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, and P. Bork, "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences," Nucleic Acids Res, vol. 44, no. D1, pp. D286–93, 2016.

[66] O. K. Ekseth, M. Kuiper, and V. Mironov, "OrthAgogue: An agile tool for the rapid prediction of orthology relations," Bioinformatics, vol. 30, no. 5, pp. 734–736, 2014.

[67] Y. Wang, D. Coleman-Derr, G. Chen, and Y. Q. Gu, "OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species.," Nucleic Acids Res., vol. 43, no. W1, pp. W78–84, 2015.

[68] C. J. Penkett, J. A. Morris, V. Wood, and J. Bähler, "YOGY: A web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms," Nucleic Acids Res., vol. 34, no. WEB. SERV. ISS., pp. 330–334, 2006.

[69] Y. Hu, I. Flockhart, A. Vinayagam, C. Bergwitz, B. Berger, N. Perrimon, and S. E. Mohr, "An integrative approach to ortholog prediction for disease-focused and other functional studies.," BMC Bioinformatics, vol. 12, p. 357, 2011.

Table 1.1 Specification of different methods

Method	Definition	Applied to	Pros	Cons
Corelational Coefficient based Clustering (COCO-CL)	Orthologous and paralogous sequences uses distance evolutionary histories. It takes an input of refined homologous proteins attained through less conservative clustering algorithm [28].	Several protein classification database, KOG [30], COG, OrthoMCl[31] and BLAST searches.	Semi independent method. Also used when species tree is unknown or uncertain.	It fails with ancient horizontal gene transfer, rampant gene losses. Incomplete dataset. It cannot quantify the quality of clusters [28].
Resample inference of orthologs (RIO)	Speciation duplication inference (SD) algorithm for completely resolved bootstrap resampled tree [16].	Protein families' alignment with domain in plants and worms.	Used to estimate reliability of orthologs. A procedure for automated phylogenetic.	Discrepancy between orthology bootstraps value leads to error. Output of RIO varying, if distance of query sequences to other seq. is unusually short or long [12].
Reconciliatured Arbres Phylogenetigyes (RAP)	Understanding the speciation and duplication events and then from a specified gene tree topology, it detects probable ortholog and paralogs.	Database of orthologous protein families HOVERGEN, HOBACGEN and HOGENOM	Graphical user interface. User set constraint, specify tree pattern and topology. Search for events in gene losses. Hidden paralogy also considered. Cope with uncertainties. Used for very large set of phylogenetic tree.	It does not weight gene losses. It assumes that gene transmission procedure is totally vertical [9].
Level of Orthology From Tree (LOFT)	Construct tremendous hierarchical grouping that brings the different level of relatedness between paralogs and orthologs.	Its benchmark against COG, reconciliation with gene order conservation and believed species tree.	Graphical user interface, stand alone user program. Offer high resolution. Reliable ortholog. Preserving the relationship between orthologous group. 77% classified as correct when bench mark is applied. Strong visualization inspection. Complete species tree reconciliation within 2 million seconds.	No freedom to manage predictions. It is suited for mainly large scale automated phylogenomics. Sometimes, it is less accurate as compare to other methods [32].
TreeFam	It uses numerous phylogenetic techniques to build sets of ortholog from a curate dataset of trees, expanded with supplementary generated tree [33].	It gives prediction of orthology and gene tree for 109 species in 1536 families (104 fully sequences animals genome plus 5 out-group species).	Manually crated based on tree. Orthologs pair can be downloaded. Uses both novel and known data from ensemble databases. Tree based result easily visualized.	It only includes taxa of animals with few exceptions of some plants and fungal species as out-groups [34].
Hierarchical grouping of orthologous and paralogous (HOPS)	It uses bootstraps tree to compute orthology support values for sequence pairs in a multiple sequences alignment.	Prokaryotes and eukaryotes domain appears in Pfam.	Integrated in Pfam server and domain integrated. Include partial genome sequenced species.	Only pair wise orthology between two into three eukaryotic clades. No prokaryotes. Not retrievable. Only run in web browser [7].
Metaphylogene based orthologs (MeraphOrs)	Species overlap algorithm is employed on phylogenetic tree obtained from wide variety of species to combine information [38].	It serves as global database of highly accurate phylogeny based ortholog and paralogy prediction. Several of genome from eukaryotes and prokaryotes.	It explored total number of tree for a given pair of sequences. Filters are used to give better results. It also uses evidence level. It provides freedom to control prediction.	It is not a standalone user program. Complexity is high [38].
PhylomeDB	It uses a phylogenetic pipeline that comprises alignment trimming and model testing.	Its releases contain 17 Phylome from eukaryotes and prokaryotes. Comprises 416093 tree and 165840 alignments.	Used as independent source of phylogenetic information. It hosts many different Phylome (novel data access). It also provide visualization	It contains many partially overlapping phylogenetic trees.

			feature and unique ID system [14].	
Berkeley Phylofacts orthology Group (PHOG)	Pre computed tree is used to target different taxonomic distance and precision levels in a prediction server.	Eukaryotic sequences from TreeFam database.	It uses novel phylogenetic approach. It provides average complexity and moderate execution speed. It provides 86 percentages recall.	Not stand alone user program. Orthology prediction is incomplete. PHOG has recalled 59 percentages which is less than Inparanoid. No synteny info. or gene neighborhood [41].
GreenPhyl	Semi automatic gene family clustering is used to create input repository from raw data before tree construction and stand-alone phylogenetic pipeline.	It has complete plant genomes.	Easily identifies ultra paralog relationship when no ortholog are detected. It detects orthologs even at lower threshold. It is standalone user.	It is more complex [42].
Clusters of Orthologous Groups (COG)	It identifies three ways BBH between co-orthologs or orthologs in three different species and these groups extended until saturation[43]. Further development focused on growing the resource and adding up automation [44].	Initially contains seven complete genome sequence available[43] with subsequent updates, expansion and several lineage specific derivative include eukaryotic phyla, archaea etc	It is manually crated. Standard for uniform protein function group. Easily addition of new.	It contains many out paralogs.
OrthoMCL	A markov clustering process involving iteration simulations forms group of orthologs and co-orthologs [31].	The first fully automated heuristic algorithm applied across multiple eukaryotic, OrthoMCL contains 138 genomes of mostly eukaryotic.	It is standalone user program and provides freedom to control prediction. Multiple specie comparison	It is command based. Some cluster contain out paralog. Include multiple splices variant of gene [47].
Orthologous matrix (OMA)	RSD evolutionary distance and accounting for different gene lag and gene fusion fission events are improvements upon conventional BBH.	Infer evolutionary relationship among currently 1706 complete proteomes [55].	It is standalone user program. It provides moderate execution speed and average complexity. Data is available in a wide range of formats and interface. It has also synteny view	It does not provide freedom to control prediction. Command interface. All verses all protein compare phase the most time consuming phase with more than 7 millions CPU hours logged to data [55].
Domain Based ortholog Detection (DODO)	Efficient BBH approach base on domain architecture. DODO work into two step firstly, on based of domain architecture it assign protein in group secondly, it further within these group it identifies orthologs with much less complexity.	Benchmarked against Inparanoid genomes.	It is standalone user program. Its complexity is low. Detect homologs group.	It cannot detect orthologs having different reported domain arch. If gene loss occurs in anchor genome, it could not detect orthologs relationship. Accuracy of domain identification affect performance [60].
Inparanoid	Detects BBH between a pair of organism and then uses additional statistical rules to add in paralogs a rising from duplication after speciation [6].	Current release comprises 213 species, 246 eukaryotes, 20 bacteria and 7 archaea [54].	It is standalone user program. It provides freedom to control prediction. Execution speed is moderate. Include genomes for all major eukaryotic clades.	It has command based interface. Orthology prediction is incomplete more complexity.
Evolutionary genealogy of gene Non supervised orthologous (EggNOG)	It provides OG of proteins at different taxonomic levels. OG constructed from smith waterman alignment [64].	Latest version of EggNOG cover more than tripled the underlying species to cover 3686 organism [65].	It provides gene ontology term and pair-wise orthology relationship. Complete redesign web interface	Inconsistency between levels prevents correct annotations across nested group. Online servers are often database oriented.

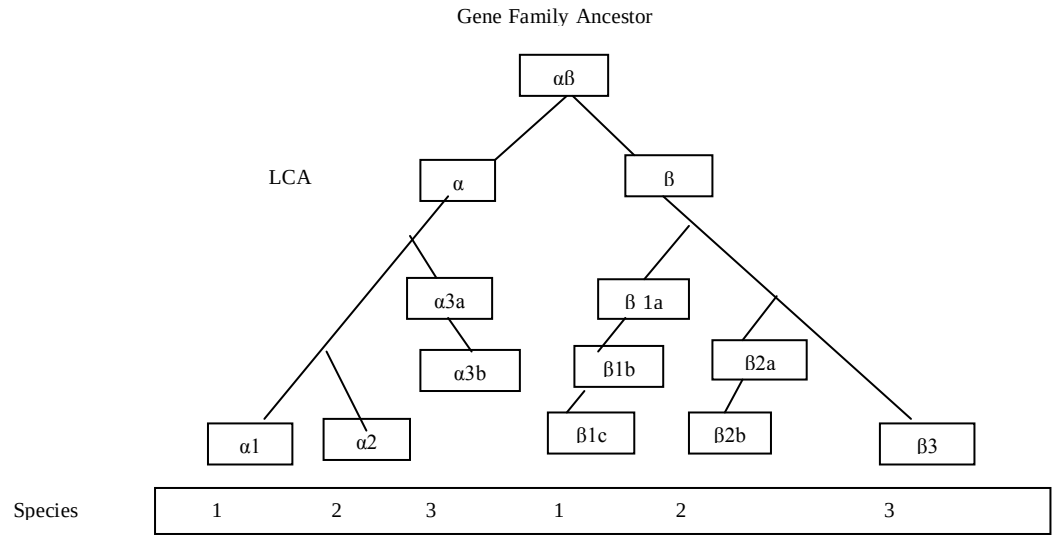

Figure1.1 Orthologous and Paralogous Genes