T-Tree, a new tool for taxonomy-based phylogenetic co-evolution analysis

Term Project for Biochemistry 218 Stanford University Fall 2007
Theodore C. Goldstein

ABSTRACT

Cladograms are dendrograms (tree-shaped diagrams) of proteins that are easily constructed by automated tools using phylogenetic techniques such as the tool ClustalW. Cladograms correspond to evolutionary relationships, but they are not easily projected onto actual consensus taxonomies. This paper describes T-Tree, an original application that maps bioinformatic-tool generated phylogenetic cladograms (such as those generated by ClustalW) onto taxonomic trees. T-Tree also provides a structure for performing additional phylogenetic analysis including filtering, data conditioning, and utilizing additional new statistical tools for detecting congruence between trees. T-Tree results provide a cognitive framework for hypothesizing whether candidate sets of molecules are ***co-evolutionary***, an important test for many investigations. If the cardinality of the tree is sufficient, T-Tree utilizes DeVienne's new *Icong* tree congruence tool for analyzing co-evolutionary relationships. This paper presents the results from applying T-Tree to data sets of nuclear hormone receptor *PPAR Gamma* and its ligand, *Insulin*. This work demonstrates how the T-Tree tool could be used to explore general co-evolutionary relationships between sets of evolving molecules including proteins and nucleic acids.

1. BACKGROUND

In the current post-genomic era, the scientific community is literally rapidly documenting the entire tree of life on the web.[1] Using new bioinformatics tools, we are shifting our focus and advancing from a protein-by-protein analysis approach to whole genome analysis. We can perform analysis on evolutionary theories that previously could not be proven or disproven for lack of data. One such theory proposed by Marc Kirschner and John Gerhart in *The Plausibility of Life*[2], makes the case that the Darwin's explanation for the origin of diversity is incomplete and that the results of recent discoveries in cell and developmental biology can be used to remedy this defect. Among the interesting theories that Kirschner and Gerhart propose is "weak linkage." They hold that weak linkage permits signal and response components to be combined in different contexts, allowing a novel outcome of development to be produced without the invention of new individual components. However, Kirschner and Gerhart do not provide any detailed mechanism for how this is accomplished.

Newer work done by [Bridgham 2006] [3] begins to answer the question of what is the mechanism behind weak linkage. Very impressively, they use phylogenetic analysis of nuclear hormones to posit a theoretical ancestral hormone. They then validate the study by synthesizing the ancestral hormone receptor and measuring its affinity binding across a number of ligands other than the known target ligands for the receptors. This sheds light on the way one complex systems of ligands and receptors evolved together. Bridgham and Thornton propose a new theory of molecular exploitation whereby a molecule can be recruited into a new role and hence into a new functional system complex.

THE PROBLEM

The theory of molecular exploitation supports the theory of weak linkage proposed by Kirschner and Gerhart. Bridgham presents an excellent explanation of the reconstruction of the ancestral nuclear hormone receptor. The theory of molecular exploitation is a great specific mechanism that helps describe how Darwinian evolution can advance new function before an entire system is in place. They ask that before a hormone is present, what is the source of the selection pressure for the receptor's affinity for it? They argue that, without the receptor, there is no selection pressure that could guide the evolution of the ligand. The difficulty with Bridgham's work is that it is very complicated to identify the relationship between closely paired receptors and ligands. They explore the relationships of the steroid hormone mineralocorticoid receptor MR and the glucocorticoid receptor (GR) and their respective ligands, aldosterone and cortisol, two well known closely related nuclear hormone receptors. This is an exciting area and one where bioinformatics tools could be useful to explore relationships of any set of co-evolving molecular systems, including protein, DNA and RNA molecules. All of these molecules can be viewed as co-evolving systems of information that are mutually responding to each other. The problem is how can we identify a set of co-evolving leading characters amongst a rapidly changing cast?

A SOLUTION

This paper describes the development of a new tool called T-Tree. The purpose of T-Tree is to provide guidance as to whether molecules are co-evolutionary. T-Tree maps phylogenetic cladograms onto the consensus taxonomy and creates a unified graph. T-Tree then performs a series of analyses that ask the question*, are these proteins co-evolutionary?* Proteins which are co-evolutionary could be subjects for further analysis. The taxonomic approach has many advantages over running similar statistical tests without the taxonomic tree:

- It is easier to visualize the evolutionary relationships implied by the cladogram.
- Like any hierarchy ontology (such as MeSH)[4], it becomes possible to normalize and assume that closely related species can be treated the same (or not) for purposes of analysis.
- When the cladogram contradicts the taxonomic tree, it is required to assess the reliability of the cladogram.
- By using metadata such as branch length from the cladogram, we can choose to normalize and exclude outliers from the data within certain scopes.

Co-speciation is the mutual evolutionary influence between two species, such as predator-prey or symbiosis. Each party in a co-evolutionary relationship exerts selective pressures on the other, thereby affecting each other's evolution. If the theory of molecular exploitation is correct, we may see the same similar patterns between mutually interacting molecules.

Can we isolate whole systems of biological interaction based solely on phylogenetic relationships? If we can find based solely on the degree of randomness of their tree patterns, we have a powerful new tool to focus research. In this study, we begin with the nuclear receptor PPAR Gamma and its ligand Insulin. However, the methods are purely graph based and should apply to a diverse range of investigations.

Borrowing from techniques developed by congruence testing for studies of host-parasite associations, we can investigate whether two families of molecules co-evolve, much as a host

and a parasite co-speciate. [DeVienne 2007] [5] has developed a congruence tool called *Icong* for comparing whether two trees are congruent by random chance or by a possible co-evolutionary relationship. DeVienne's theory is that congruence of the graph implies a high probability of co-speciation.

SUMMARY OF THE T-TREE ALGORITHM

A series of candidate receptor and ligand proteins are chosen by the user. The user retrieves the proteins from the Uniprot[6] web database and deposits them into a series of XML datasets. For each dataset, a cladogram is computed using ClustalW[7], producing a branch length annotated phylogram.

T-Tree then constructs a unified directed acyclic graph[8] of the organisms from the Uniprot datasets. T-Tree then parses the cladograms, and annotates each organism in the taxonomy with the proteins with a Binding Node. A more complete description of the algorithm is given below. T-Tree essentially does a bottom up tree match. The datasets are enriched by matching closely related organisms such as *Macaca mulatta* and *Macaca fascicularis,* which are cousins visible in the bottom-most clades of the taxonomic tree. These close cousin nodes are merged.

Distant cousins are excluded from the tree matching. Data must be normalized because of limits of phylogenetic reconstruction methods such as long branch attractors and the lack of unifying model of differences in evolutionary rates. This is done by a filter on the branch length described in [Lartillot 2007] [9]. Unlike classical computer algorithms for tree matching, the branch length is a significant factor and may cause nodes to be excluded from the analysis. T-Tree uses a branch length threshold parameter provided by the user. Further investigation and future versions may yield automatic determination of the branch length threshold parameter.

Once a matching subset of nodes is found, the collection of all sub-graphs with cardinality of seven or above are subjected to [DeVienne 2005] new *Icong* tree congruence tool for comparing whether two trees are congruent by random chance or by a possible co-evolutionary relationship.

METHOD DETAILS

The following proteins were formed as two data sets to see if there was co-evolution.

Scientific Name	Common name	PPARG Protein	Insulin Protein
Bos taurus	Domestic Cow	PPARG_BOVIN	INS_BOVIN
Canis familiaris	Dog	PPARG_CANFA	INS_CANFA
Cricetulus griseus	Hamster	PPARG_CRIGR	INS_CRILO
Homo sapiens	Human	PPARG_HUMAN	INS_HUMAN
Macaca mulatta	Macque	PPARG_MACMU	INS_MACFA
Mus musculus	Mouse	PPARG_MOUSE	INS1_MOUSE
Sus scrofa	Pig	PPARG_PIG	INS_PIG
Oryctolagus cuniculus	Rabbit	PPARG_RABIT	INS_RABIT
Rattus norvegicus	Rat	PPARG_RAT	INS1_RAT
Xenopus laevis	Frog	PPARG_XENLA	INS1_XENLA

Table 1 Proteins used in this analysis

The list of proteins was determined by doing a search on the Uniprot database http://beta.uniprot.org/. The result was downloaded in Fasta and XML format. Each dataset was then run through TimeLogic DeCypher's ClustalW general purpose multiple sequence alignment

program for proteins on http://decypher.stanford.edu. Of course, other ClustalW implementations would have been suitable as well. The cladogram was extracted and normalized for common nomenclature. The resulting four files (PPARG.xml INS.xml PPARG.ph INS.ph) were run through T-Tree constructing the raw annotated phylogram output in the appendix. T-Tree was also selected with the option to perform DeVienne's Icong analysis. This analysis was repeated at multiple branch length thresholds.

DETAILS OF THE T-Tree ALGORITHM

The T-Tree algorithm is a recursive, bottom up algorithm. The goal of the algorithm is to find the taxon of the taxonomic tree in common with a given binding node of the phylogram. It is similar to other Tree matching algorithms except that it skips clades. The O(n) of the algorithm is 1, as it efficiently uses the count of the found nodes to determine when the least common sub-tree has been found.

INPUT: A Taxonomic Tree and, a set of partially bound cladograms where each leaf represents a protein. Each protein has been bound to its organism given by the data provided by Uniprot in the XML description.

OUTPUT: A unified graph with interior nodes bound to the least common Taxon.

```
Taxon::propagateBindingsUpTaxon()
    numKids = len(self.kids)
    unionBindings ← null set
    if numKids == 0          Leaf Taxon Node
        return self.binding
    elif numKids == 1          Solitary Taxon Node
        return self.kids[0].propagateBindingsUpTaxon()

    foreach kid in self.kids   N-Ary Taxon Node
        cladogramBindings ← kid.propagateBindingsUpTaxon()
                intersection ← intersection ∩ kidBindings Calculate intersection from below

        foreach binding in cladogramBindings
            if binding in unionBindings and binding.foundBindingCount == binding.cardinality
                Found the originating common Taxon of all bindings
                self.originBindings ← self.originBindings ∪ binding
                    if binding.cardinality >= DeVienneAnalysis.CardMinium
                        DeVienneAnalysis.candidateTaxons.add(self)
            else
                if binding in unionBindings
                        binding.foundBindingCount ← binding.foundBindingCount + 1
                unionBindings ← unionBindings ∪ e
    return unionBindings
```

Each protein XML entry identifies its various standard information including name, accession number and lineage. The cladogram of the protein set is in Newick Format, with branch lengths. This is easily obtained from tools Protein ClustalW and HMMs from tools such as Decypher and many other implementations. Figure 1 shows the graphical version of the cladograms generated by the ClustalW. T-Tree parses the raw Newick format.

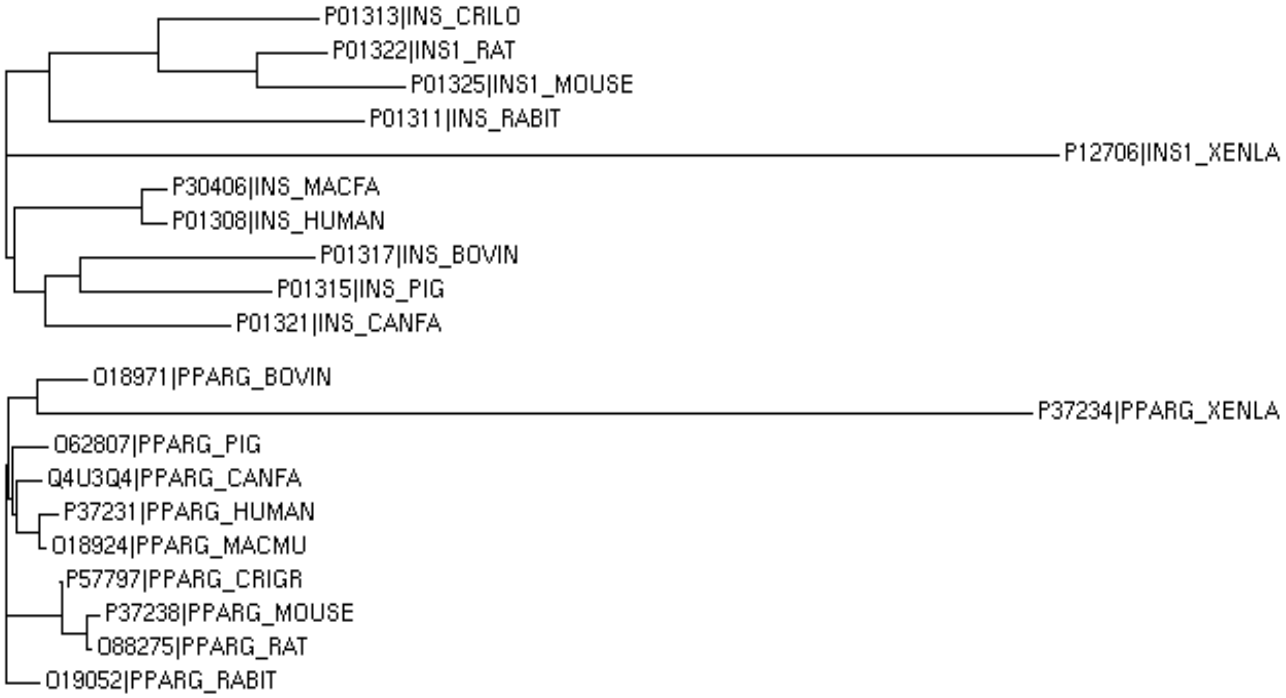


Figure 1 Cladograms generated by ClustalW

2. IMPLEMENTATION

T-Tree is implemented as a Python program with only a UNIX command line interface. However, it is straightforward to convert this implementation to a web-based implementation.

OBJECT MODEL

The object model for T-Tree (Figure 2 T-Tree Object Model) makes use of Python's excellent libraries. The XML parser utilizes the standard SAX parser. The SAX (Simple API for XML) is a serial access parser API for XML. SAX provides a mechanism for reading data from an XML document. It is a popular alternative to the Document Object Model (DOM). The SAX parser reads a series of XML entities and calls the functions startElement(), endElement() and characters() that are defined on the TaxonTreeBuilder (is a subclass of the XMLContentHandler).

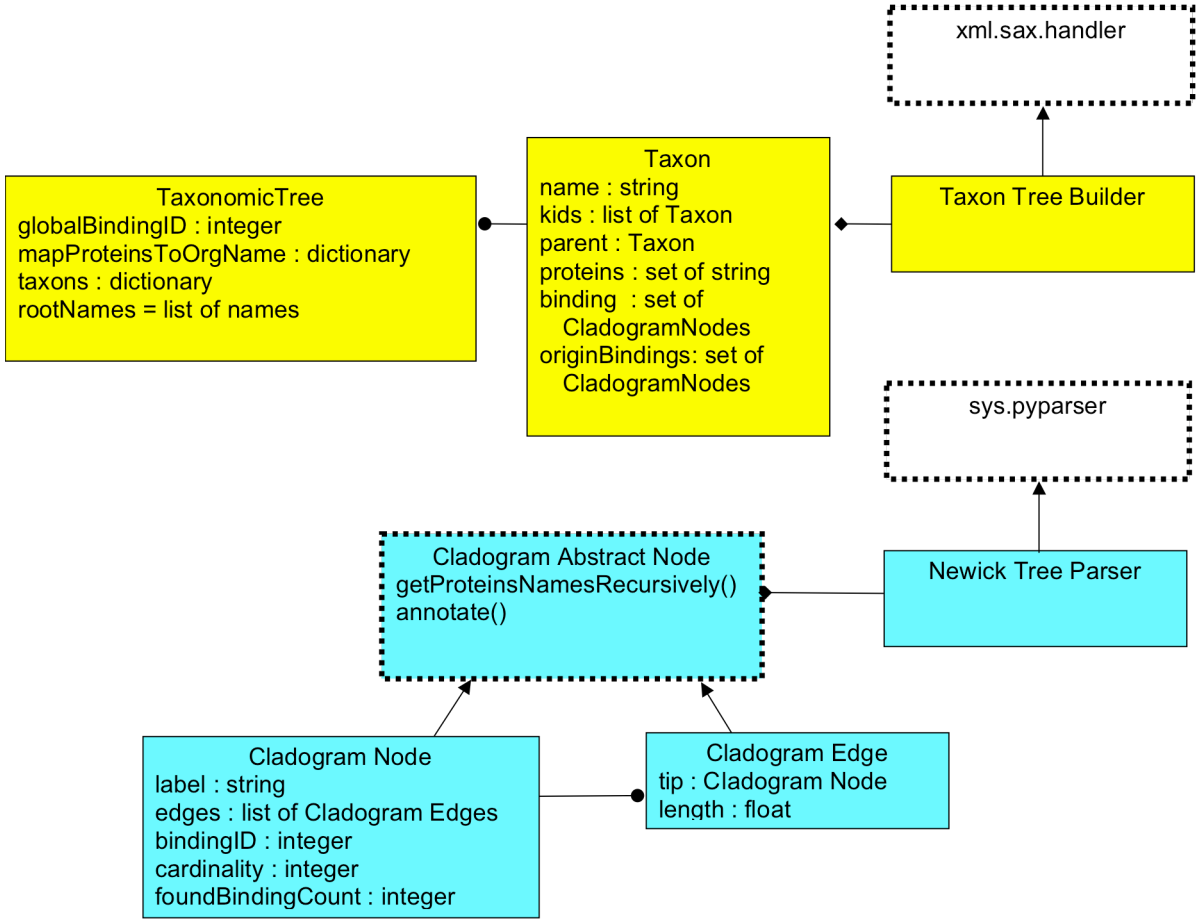


Figure 2 T-Tree Object Model

The NewickTree[10] parser is derived from sources by Rosengren,[11] itself a subclass of the PyParsing package. It builds a graph of the cladogram with separate objects representing the Nodes and Edges. The Taxonomic Tree object maintains a *Facade*[12] design pattern for access to internal representations.

3. RESULTS

The result of running T-Tree maps the biologically relevant taxon to the matching clades. The specific mapping is summarized in Figure 3 and Table 2. The lines are the taxonomic relationships generated by T-Tree that are mappable from the cladogram. Hamster species are treated as biologically similar, as are Macaca monkeys. The use of the DeVienne tool adds significant power to T-Tree. Further studies with other receptors and ligands are necessary to validate the T-Tree approach. Only when the branch length was below 0.25 (which eliminated *Xenopus laevis*) did we detect congruence.

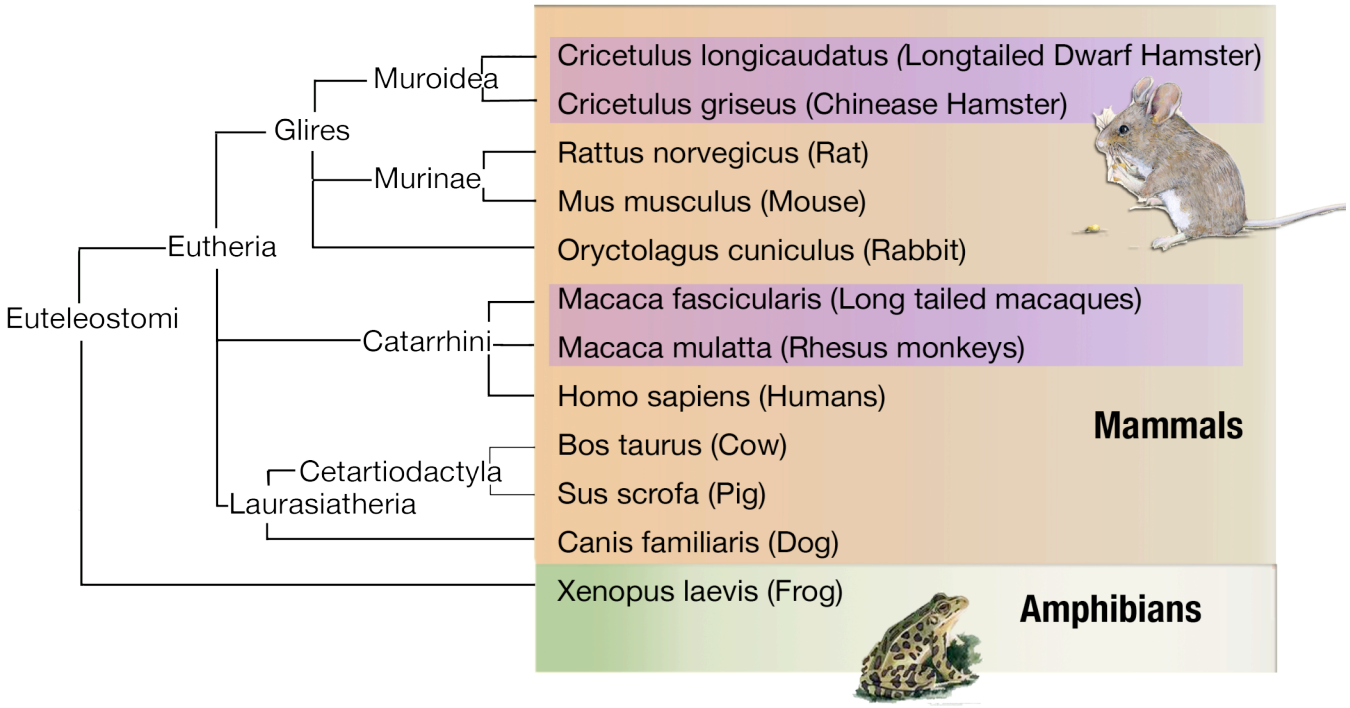


Figure 3 Taxonomy of Mammals and Amphibians[13]

Branch Length	Species	Icong	P Value	Interpretation
=> 0.25	All Ten species	0.14	0.026	Not more congruent than by chance
< 0.25	Nine species, excluding Xenopus	1.39	0.058	More congruent than by chance

Table 2 Summary of results: Branch length and effect on Icong and P Value

EVALUATION

A MeSH Pubmed search and a web search found many tools that perform *phylogenetic analysis*, *taxonomic analysis* or *dendrogram annotations*. Figure 4 depicts the tools that were closest to the spirit of T-Tree. GeneTree[14] and MacClade[15] are interactive tools that read tree formats from multiple data-sources and provide methods to redact, filter and annotate the trees.

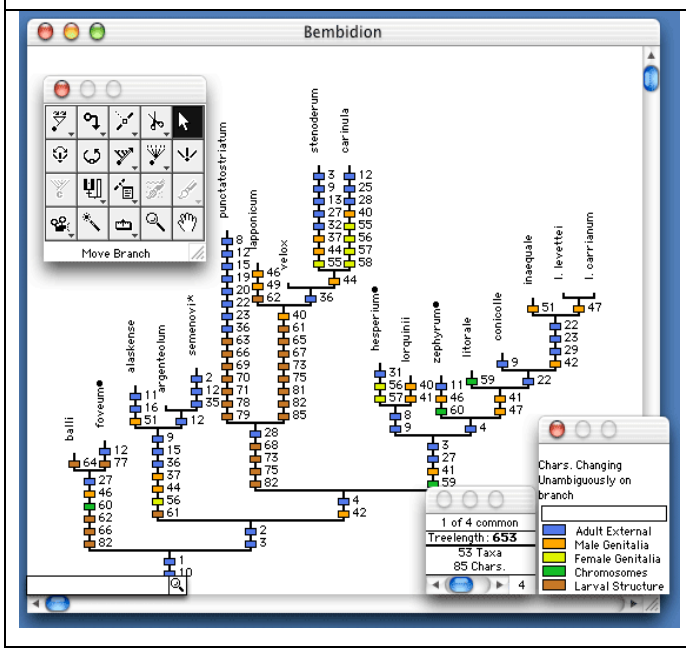
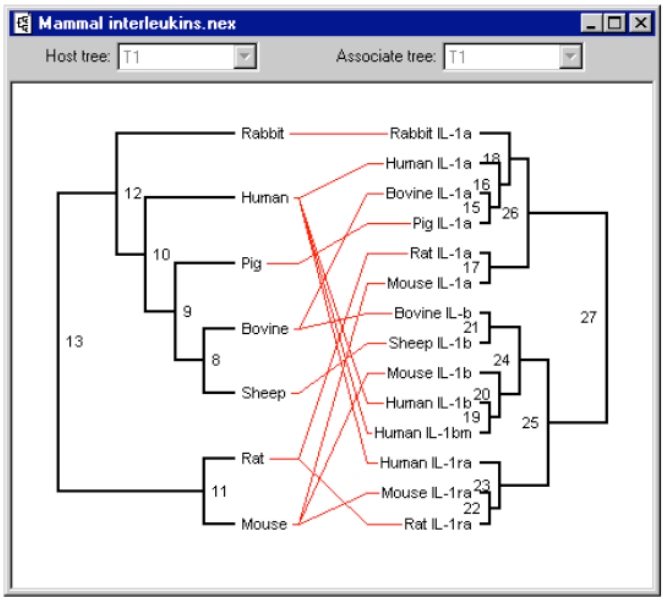
MacClade	GENETREE
	

Figure 4 Screenshots of MacClade and GeneTree

MacClade and GeneTree are elegant tools, but all of the annotations on the dendrograms are numeric data. None of these tools perform annotation on the taxonomy. In the DISCUSSION section below, it is conjectured that there are epistemological reasons why there are no tools for annotating taxonomies.

Therefore, T-Tree needs to be compared to hand annotation methods. A survey of various use sites, including the large web-based Tree of Life project (see APPENDIX section below for a full page screenshot) [16] provide no published algorithms about how these annotations are made. It would seem that it is done casually by eye. T-Tree's advantage over manual methods is that it creates a mathematically formal treatment of the data. The Taxon identified can then be used for further investigation, such as to identify other species for further evidence of coevolution or differentiation.

COMPARISON TO SIMILAR COMPUTATION METHODS

The central algorithm of T-Tree is a recursive bottom-up tree matching algorithm. It is similar to other well known tree matching algorithms [Hoffman 1982][17]. But T-Tree is only looking for the minimal intersection of the two trees, and not the identity match of any interior nodes. Therefore, only the cardinality of the sub-tree and whether a node spans the complete set of protein to species that are enclosed in its map matters. T-Tree's algorithm is much simpler than the MAST [AMIR 1997][18] , a very common algorithm used in matching cladograms for phylogenetic

analysis. In MAST, the input is a set of leaf-labeled trees and the goal is to compute a tree contained in all of the input trees with as many labeled leaves as possible. T-Tree is only concerned with nodes in common. A special case is that T-Tree can merge nodes at the periphery. As well, the constraint on matching the interior nodes is simply that the spanning nodes of the sub-tree match. Again, this is a much easier constraint than MAST, which penalizes the match (called the maximal weighted sub-tree) for levels of the taxonomy skipped. In T-Tree, a sub-tree is considered matched exactly when all leaves have been encountered.

4. DISCUSSION

T-Tree is an early effort and has significant limitations. While T-Tree is able to match trees whose species are closely related, there has to be at least one match in every class of molecule. T-Tree's ability to match trees with non-overlapping species could be improved by having some sort of fuzzy logic matching algorithm.

Controversy

Why are so few cladograms annotated with taxonomies and why are so few taxonomies annotated with phylogenetic information? Fifty years ago, there was a huge controversy between the *Cladists* and *Taxonomists*. The Cladists hold that Linnaean seven level (Kingdom, Phylum, Class, Order, and Family. Genus, Species) are overly simplistic and lack meaning in the face of evolutionary data. Modern Taxonomists have responded by inventing new level-names including superorder, suborder, infraorder, parvorder, magnorder. But many such as [Wolf 2005] [19] believe that the entire idea of taxonomies of non-Eukaratic species is obsolete because of the extent of horizontal gene transfer.

Taxonomists hold that taxa reflect phylogenies. They base taxa definitions on tangible characteristics that provide a testable hypothesis to determine if a given species is in a taxon or not. This applies whether the organism evolved by inherited traits or by gene-conversion, gene-sharing or other cross-organism transfer.

Possible Improvements to T-Tree

T-Tree focuses on paralogous evolution of two molecules within closely related species. Gene duplication events and radical molecular evolution where a new molecule displaces an existing molecule will end molecular co-evolution. This could be visible as a sub-tree no long being congruent.

It is possible that T-Tree can be used to find these events by their absence of congruence. There are many analogues in co-speciation that are worthy of investigation including vicariance (the division of a group of organisms by a geographic barrier, such as a mountain or a body of water, resulting in differentiation of the original group into new varieties or species).

T-Tree uses DeViene's Icong index and inherits many of Icong's issues. Icong was designed for host/parasite co-evolution and it is an open question whether it truly applies to receptor and ligands. The range of Icong does fits only a subrange of the standard distribution. However, this is a broad standard distribution range and is more than sufficient for our purposes.

DeVienne's index requires trees with at least seven nodes. It is frequently difficult to find at least seven nodes in each molecule class (ligand or receptor). Each molecule must be complete enough for a ClustalW analysis. The Icong index does not work where leaves on one tree are associated with multiple leaves on the other tree. This is an issue in host/parasite relationships.

But protein co-evolution studies match across organisms and therefore multiple interactions are less likely.

User Model Improvements

Receptor and ligands are well known and selecting them is easy. But selecting an appropriate set of input proteins for other co-evolutionary studies will frequently be difficult and likely impossible for Bacteria where horizontal gene transfer is likely to confound the principal. However, techniques such as [Marshall 2005][20] may provide a means to select appropriate proteins by using distance matrices and other maximal concordant genes and proteins.

T-Tree's input currently requires manual retrieval of sequences from Uniprot. This could be automated and an easy-to-use web-wrapper would make it much easier to use. NCBI ASN.1 format should also be supported. T-Tree's raw output is unattractive. The diagram above is hand built from the T-Tree's output, but could be automatically generated.

5. CONCLUSION

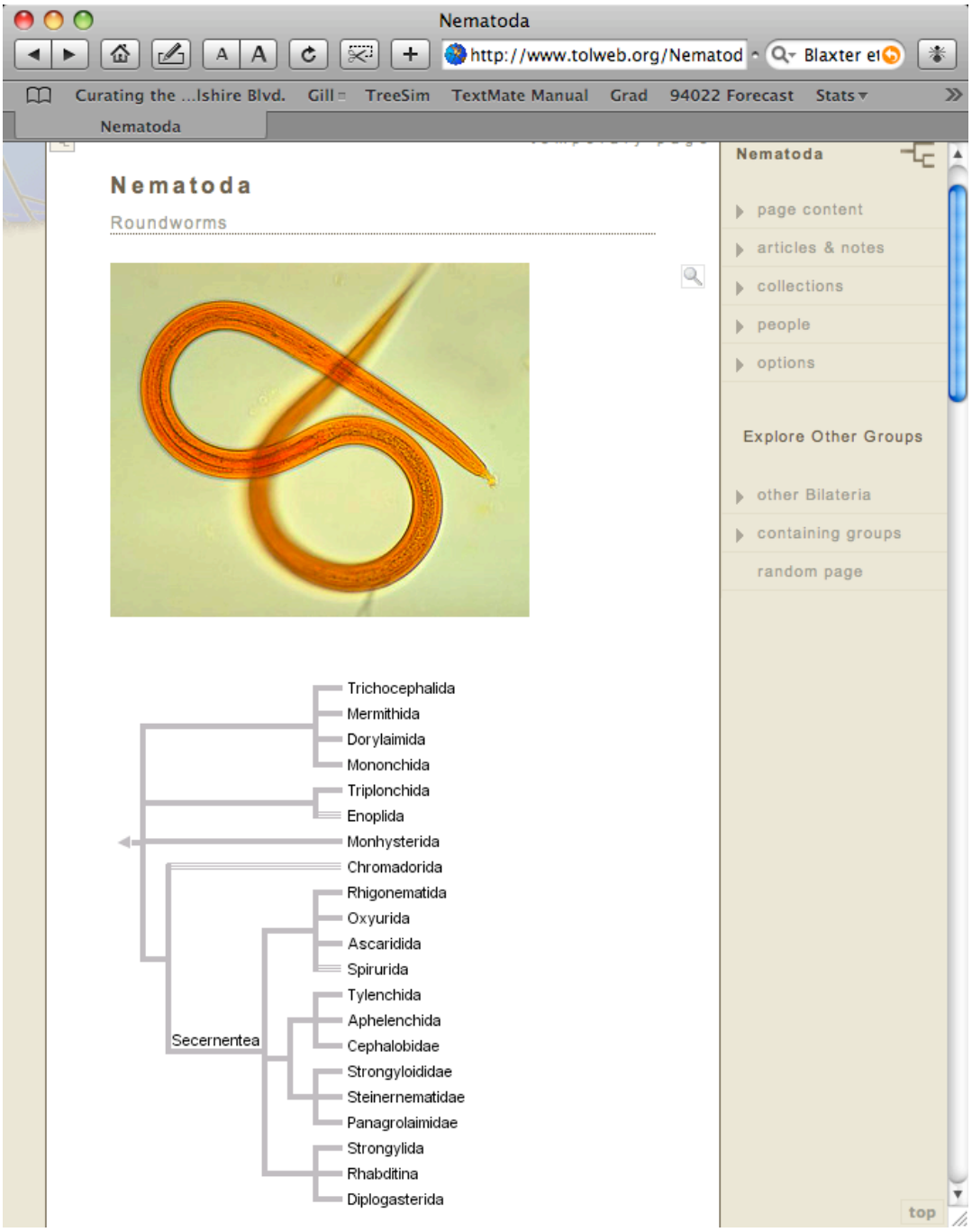
The results are consistent with the biology. The branch length excluding Xenopus is justified since Xenopus is a very different organism from the Mammals. According to [Hedges 2002][21], the last common ancestor taxa is Euteleostomi, having diverged approximately 360 million years ago (mya). This level of divergence is sufficient to explain the lack of convergence. The Icong ratio with and without is an order of magnitude more convergent. This is exciting support for use of Icong as a measure of co-evolution.

Recent trends in bioinformatic analysis exemplified by Koonin [Wolf 2005][19] and others of non-eukaryotic organisms argue that horizontal gene transfer is so prevalent as to make taxonomies useless. However, we should not discard the genome with the bathwater. Taxonomies are still informative for most eukaryotic species, T-Tree shows that there may yet be a role for taxonomies in bioinformatics.

T-Tree is a useful new tool that begins to answer the question: *did two molecules co-evolve*? Early studies with nuclear hormone receptors and ligands look very promising. Further investigation is necessary to validate whether this approach generalizes beyond nuclear hormone receptors and ligands and across a range of many molecular systems. T-Tree's use of taxonomies assists with managing and normalizing the cladograms. As well, the new relationships of taxonomies provide interesting and informative information in their own right.

6. APPENDIX

- **TREE OF LIFE SCREENSHOT**



XML SAMPLE

Here is an example XML entry from Uniprot. Information not relevant to T-Tree has been redacted including the extensive cross reference information to articles.


```
<entry dataset="Swiss-Prot" created="1994-10-01" modified="2007-11-13"
version="113">
    <accession>Q96J12</accession>
    <name>PPARG_HUMAN</name>
      <protein>
          <name>Peroxisome proliferator-activated receptor Gamma</name>
          <name>PPAR-Gamma</name>
      <name>Nuclear receptor subfamily 1 group C member 3</name>
    </protein>
    <gene>
      <name type="primary">PPARG</name>
    </gene>
    <organism key="1">
      <name type="scientific">Homo sapiens</name>
      <name type="common">Human</name>
      <dbReference type="NCBI Taxonomy" id="9606" key="2" />
      <lineage>
          <taxon>Eukaryota</taxon>
          <taxon>Metazoa</taxon>
          <taxon>Chordata</taxon>
          <taxon>Craniata</taxon>
          <taxon>Vertebrata</taxon>
          <taxon>Euteleostomi</taxon>
          <taxon>Mammalia</taxon>
          <taxon>Eutheria</taxon>
          <taxon>Euarchontoglires</taxon>
          <taxon>Primates</taxon>
          <taxon>Haplorrhini</taxon>
          <taxon>Catarrhini</taxon>
          <taxon>Hominidae</taxon>
          <taxon>Homo</taxon>
      </lineage>
    </organism>
    <sequence length="505" mass="57567" checksum="F16E5CAB122EBB32"
modified="1999-08-01" version="2">
      MGETLGDPPVDPEHGAFADALPMSTSQEITMVDTEMPFWPTNFGISSVDLSVMDDHSHSF
      DIKPFTTVDFSSISAPHYEDIPFTRADPMVADYKYDLKLQEYQSAIKVEPASPPYYSEKT
      QLYNRPHEEPSNSLMAIECRVCGDKASGFHYGVHACEGCKGFFRRTIRLKLIYDRCDLNC
      RIHKKSRNKCQYCRFQKCLAVGMSHNAIRFGRMPQAEKEKLLAEISSDIDQLNPESADLR
      ALAKHLYDSYIKSFPLTKAKARAILTGKTTDKSPFVIYDMNSLMMGEDKIKFKHITPLQE
      QSKEVAIRIFQGCQFRSVEAVQEITEYAKNIPGFINLDLNDQVTLLKYGVHEIIYTMLAS
      LMNKDGVLISEGQGFMTREFLKSLRKPFGDFMEPKFEFAVKFNALELDDSDLAIFIAVII
      LSGDRPGLLNVKPIEDIQDNLLQALELQLKLNHPESSQLFAKVLQKMTDLRQIVTEHVQL
      LHVIKKTETDMSLHPLLQEIYKDLY
    </sequence>
</entry>
```

T-TREE OUTPUT

```
BINDING 1 ['INS1_RAT'] ( RAT , MOU )
BINDING 1 ['INS1_MOUSE'] ( RAT , MOU )
BINDING 2 ['INS_CRILO'] ( CRI , ( RAT , MOU ) )
BINDING 2 ['INS1_RAT', 'INS1_MOUSE'] ( CRI , ( RAT , MOU ) )
BINDING 3 ['INS_CRILO', 'INS1_RAT', 'INS1_MOUSE'] ( ( CRI , ( RAT , MOU ) ) ,
RAB )
BINDING 3 ['INS_RABIT'] ( ( CRI , ( RAT , MOU ) ) , RAB )
BINDING 4 ['INS_MACFA'] ( MAC , HUM )
BINDING 4 ['INS_HUMAN'] ( MAC , HUM )
BINDING 5 ['INS_BOVIN'] ( BOV , PIG )
BINDING 5 ['INS_PIG'] ( BOV , PIG )
BINDING 6 ['INS_BOVIN', 'INS_PIG'] ( ( BOV , PIG ) , CAN )
BINDING 6 ['INS_CANFA'] ( ( BOV , PIG ) , CAN )
BINDING 7 ['INS_MACFA', 'INS_HUMAN'] ( ( MAC , HUM ) , ( ( BOV , PIG ) , CAN
) )
BINDING 7 ['INS_BOVIN', 'INS_PIG', 'INS_CANFA'] ( ( MAC , HUM ) , ( ( BOV ,
PIG ) , CAN ) )
BINDING 8 ['INS_CRILO', 'INS1_RAT', 'INS1_MOUSE', 'INS_RABIT'] ( ( ( CRI , (
RAT , MOU ) ) , RAB ) , XEN , ( ( MAC , HUM ) , ( ( BOV , PIG ) , CAN ) ) )
BINDING 8 ['INS1_XENLA'] ( ( ( CRI , ( RAT , MOU ) ) , RAB ) , XEN , ( ( MAC
, HUM ) , ( ( BOV , PIG ) , CAN ) ) )
BINDING 8 ['INS_MACFA', 'INS_HUMAN', 'INS_BOVIN', 'INS_PIG', 'INS_CANFA'] ( (
( CRI , ( RAT , MOU ) ) , RAB ) , XEN , ( ( MAC , HUM ) , ( ( BOV , PIG ) ,
CAN ) ) )
( ( ( CRI , ( RAT , MOU ) ) , RAB ) , XEN , ( ( MAC , HUM ) , ( ( BOV , PIG )
, CAN ) ) ) BINDING 9 ['PPARG_BOVIN'] ( BOV , XEN )
BINDING 9 ['PPARG_XENLA'] ( BOV , XEN )
BINDING 10 ['PPARG_HUMAN'] ( HUM , MAC )
BINDING 10 ['PPARG_MACMU'] ( HUM , MAC )
BINDING 11 ['PPARG_CANFA'] ( CAN , ( HUM , MAC ) )
BINDING 11 ['PPARG_HUMAN', 'PPARG_MACMU'] ( CAN , ( HUM , MAC ) )
BINDING 12 ['PPARG_PIG'] ( PIG , ( CAN , ( HUM , MAC ) ) )
BINDING 12 ['PPARG_CANFA', 'PPARG_HUMAN', 'PPARG_MACMU'] ( PIG , ( CAN , (
HUM , MAC ) ) )
BINDING 13 ['PPARG_BOVIN', 'PPARG_XENLA'] ( ( BOV , XEN ) , ( PIG , ( CAN , (
HUM , MAC ) ) ) )
BINDING 13 ['PPARG_PIG', 'PPARG_CANFA', 'PPARG_HUMAN', 'PPARG_MACMU'] ( ( BOV
, XEN ) , ( PIG , ( CAN , ( HUM , MAC ) ) ) )
BINDING 14 ['PPARG_MOUSE'] ( MOU , RAT )
BINDING 14 ['PPARG_RAT'] ( MOU , RAT )
BINDING 15 ['PPARG_CRIGR'] ( CRI , ( MOU , RAT ) )
BINDING 15 ['PPARG_MOUSE', 'PPARG_RAT'] ( CRI , ( MOU , RAT ) )
BINDING 16 ['PPARG_BOVIN', 'PPARG_XENLA', 'PPARG_PIG', 'PPARG_CANFA',
'PPARG_HUMAN', 'PPARG_MACMU'] ( ( ( BOV , XEN ) , ( PIG , ( CAN , ( HUM , MAC
) ) ) ) , ( CRI , ( MOU , RAT ) ) , RAB )
BINDING 16 ['PPARG_CRIGR', 'PPARG_MOUSE', 'PPARG_RAT'] ( ( ( BOV , XEN ) , (
PIG , ( CAN , ( HUM , MAC ) ) ) ) , ( CRI , ( MOU , RAT ) ) , RAB )
BINDING 16 ['PPARG_RABIT'] ( ( ( BOV , XEN ) , ( PIG , ( CAN , ( HUM , MAC )
) ) ) , ( CRI , ( MOU , RAT ) ) , RAB )

Eukaryota
  Metazoa
    Chordata
      Craniata
        Vertebrata
```

```
          Euteleostomi originBindings= 8 9 16 13
            Mammalia
              Eutheria originBindings= 11 12 7
                Euarchontoglires
                  Glires originBindings= 3
                    Rodentia
                      Sciurognathi
                        Muroidea originBindings= 2 15
                          Cricetidae
                            Cricetinae
                              Cricetulus
                                Cricetulus longicaudatus bindings= 2 3 8
set([u'INS_CRILO'])
                                Cricetulus griseus bindings= 15 16
set([u'PPARG_CRIGR'])\
                          Muridae
                            Murinae originBindings= 14 1
                              Rattus
                                Rattus norvegicus bindings= 2 8 3 14 1 16 15
set([u'INS1_RAT', u'PPARG_RAT'])
                              Mus
                                Mus musculus bindings= 2 8 3 14 1 16 15
set([u'INS1_MOUSE', u'PPARG_MOUSE'])
                    Lagomorpha
                      Leporidae
                        Oryctolagus
                          Oryctolagus cuniculus bindings= 3 8 16
set([u'INS_RABIT', u'PPARG_RABIT'])
                  Primates
                    Haplorrhini
                      Catarrhini originBindings= 10 4
                        Cercopithecidae
                          Cercopithecinae
                            Macaca
                              Macaca fascicularis bindings= 4 7 8
set([u'INS_MACFA'])
                              Macaca mulatta bindings= 10 11 16 12 13
set([u'PPARG_MACMU'])
                        Hominidae
                          Homo
                            Homo sapiens bindings= 10 7 8 4 16 12 11 13
set([u'PPARG_HUMAN', u'INS_HUMAN'])
                Laurasiatheria originBindings= 6
                  Cetartiodactyla originBindings= 5
                    Ruminantia
                      Pecora
                        Bovidae
                          Bovinae
                            Bos
                              Bos taurus bindings= 6 8 9 5 16 7 13
set([u'INS_BOVIN', u'PPARG_BOVIN'])
                    Suina
                      Suidae
                        Sus
                          Sus scrofa bindings= 6 8 5 16 12 7 13
set([u'PPARG_PIG', u'INS_PIG'])
                  Carnivora
```

```
                    Caniformia
                      Canidae
                        Canis
                          Canis familiaris bindings= 7 8 6 16 12 11 13
set([u'INS_CANFA', u'PPARG_CANFA'])
            Amphibia
              Batrachia
                Anura
                  Mesobatrachia
                    Pipoidea
                      Pipidae
                        Xenopodinae
                          Xenopus
                            Xenopus laevis bindings= 8 9 16 13
set([u'PPARG_XENLA', u'INS1_XENLA'])

Branch Length Threshold 0.23
DeVienne P Value = 0.0255466906219714
Icong = 1.38568129330254
More Congruent than by chance.

Branch Length Threshold 0.35
DeVienne P Value = 0.0584042198701816
Icong = 1.31286117798342
NOT more Congruent than by chance.
```

REFERENCES

[1] See http://www.tolweb.org/tree/

[2] Kirschner M.W. , Gerhart J.C.- *The Plausibility of Life: Resolving Darwin's Dilemma*, Yale University Press, New Haven (2005).

[3] Bridgham JT, Carroll SM, Thornton JW. - Evolution of hormone-receptor complexity by molecular exploitation. Science. 2006 Apr 7;312(5770):61-3.

[4] http://www.ncbi.nlm.nih.gov/sites/entrez?db=MeSH

[5] de Vienne M, Giraud T, Martin OC - Bioinformatics, 2007 , Advanced Acces Publication Oxford Univ Press A congruence index for testing topological similarity between trees, http://bioinformatics.oxfordjournals.org/cgi/reprint/btm500v1.pdf

[6] Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider M., The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365-370(2003).

[7] 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 Nov 11;22(22):4673-80.

[8] Skiena, S. *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica.* Reading, MA: Addison-Wesley, p. 190, 1990.

[9] Lartillot N., Brinkmann H, Philippe H. - Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model BMC Evol Biol. 2007; 7(Suppl 1): S4. Published online 2007 February 8. doi: 10.1186/1471-2148-7-S1-S4.

[10] Description, history and grammar at http://evolution.genetics.washington.edu/phylip/newicktree.html

[11] Newick Parser Grammar available at http://www.koders.com/python/fidDC964FCB0B0F1E23AC6BB6CE28DBC1CF553E57AD.aspx?s=cdef%3Atree

[12] Gamma, E., Helm, R., Johnson, R., and Vlissides,- *J. Design Patterns: Elements of Reusuble Object-Oriented Software.* Addison-Wesley, 1995.

[13] Diagram inspired by [Hedges 2002] (see 21 below).

[14] Page, R.D.M. and Cotton, J.C. (2000) - GeneTree: a tool for exploring gene family evolution. In Sankoff, Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families. Kluwer Academic Publishers, Dordrecht, pp. 525-536.

[15]see http://david.bembidion.org/macclade.html and Maddison, W. P. and D. R. Maddison. 1992. MacClade: Analysis of Phylogeny and Character Evolution, version 3.0. Sinauer Associates, Sunderland, Massachusetts.

[16] Blaxter M, Caenorhabditis elegans Is a Nematode, Science 11 December 1998: Vol. 282. no. 5396, pp. 2041 - 2046.

[17] Hoffman C. M., O'Donnell M. J. , "Pattern Matching in Trees," JACM 29, pp. 68-95, Jan, 1982.

[18] Amir A., Keselman D., Maximum agreement sub-tree in a set of evolutionary trees: Metrics and efficient algorithms, SIAM J. Comput. 26 (6) (1997) 1656–1669.

[19] Wolf Y, Rogozina I, Grishinb N and Koonin E, Genome trees and the tree of life, Trends in Genetics Volume 18, Issue 9, 1 September 2002, Pages 472-479

[20] Marshall B., Goldberg D. Automatic selection of representative proteins for bacterial phylogeny, BMC Evolutionary Biology 2005; 5: 34.

[21] Hedges, S. B. - The origin and evolution of model organism, Nature Rev. Genet. 3, 838−849 (2002).