OrthologID: automation of genome-scale ortholog identification within a parsimony framework

Chiu et al., *Bioinformatics,* 2006

seminar of bioinformatics

Triinu Kõressaar

27.03.2006

TARTU 2006

Two types of gene homology

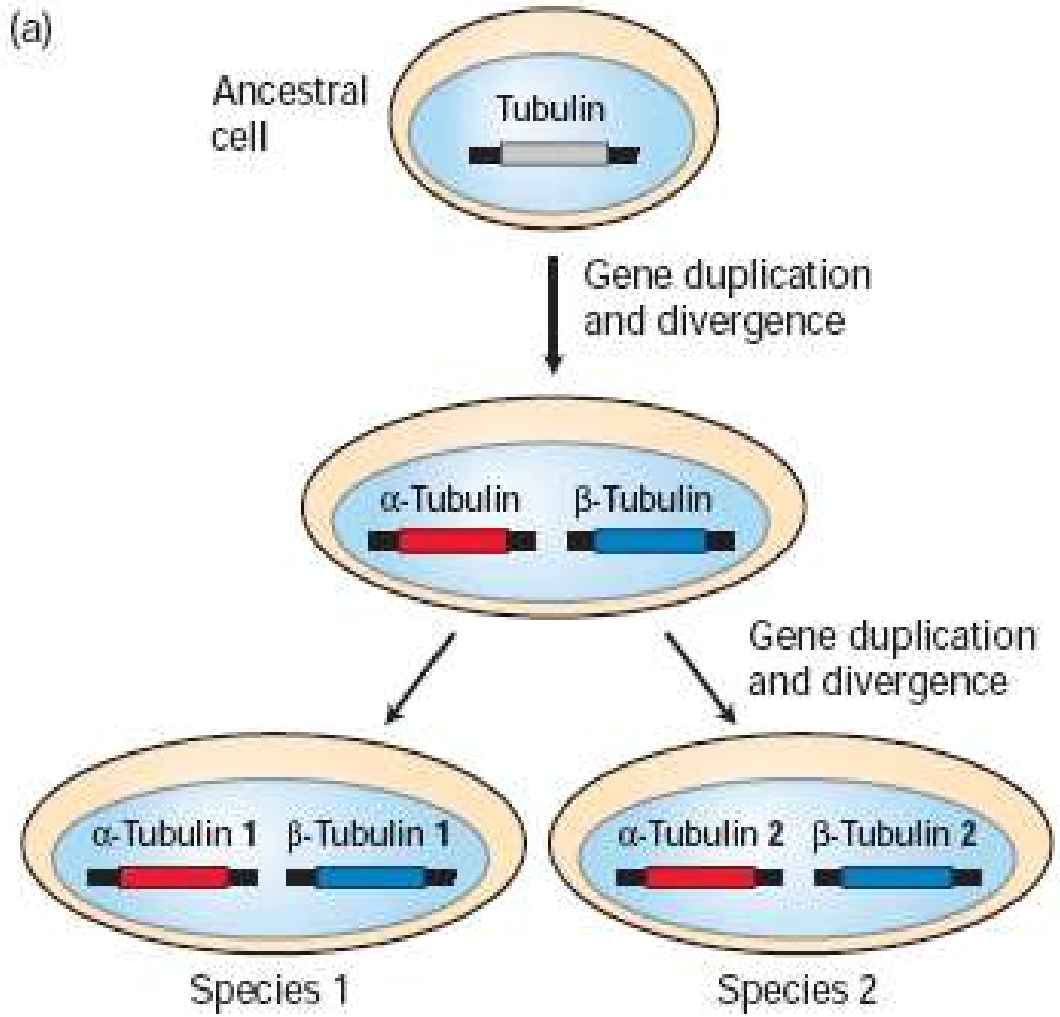
Paralogous - sequences that diverged as a result of **gene duplication** (e.g. α- and β- tubulin). Paralogous don't have to carry the same or similar function: due to lack of the original selective pressure upon one copy of the duplicated gene, this copy is free to mutate and acquire new functions.

Orthologous - sequences that arose because of **speciation** (e.g α-tubulin genes in different species). Orthologoues are genes thought to have evolved by vertical descent from a common ancestor. Orthologs will typically have the same or similar biochemical function.

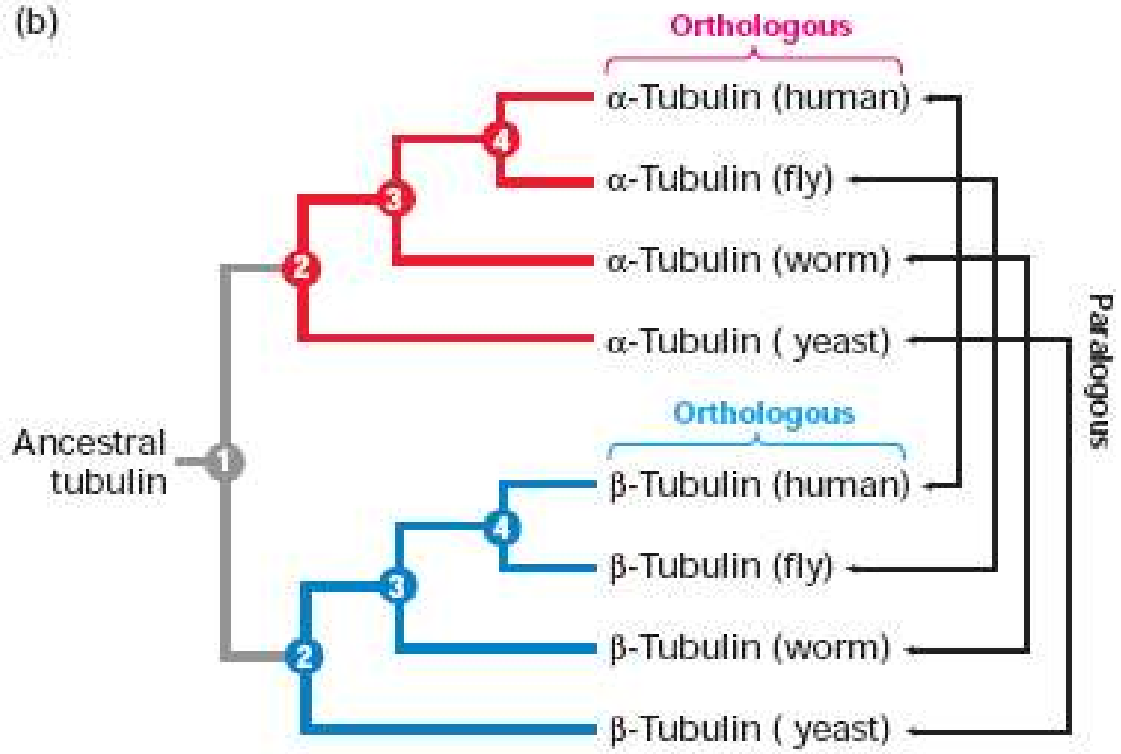
Identification of orthologues is :

- critical for reliable prediction of gene function(s) in newly sequenced genomes (comparative genomics).

- important in phylogenetics. In order to generate meaningful phylogenetic hypothesis for species evolution through character-based or distance-based analyses, it is essential that only orthologous gene sets are aligned and analyzed.



(a) Probable mechanism giving rise to the tubulin genes found in existing species. It is possible to deduce that a gene duplication event occured before speciation because the α-tubulin sequences from different species are more alike than are the α-tubulin and β-tubulin sequences within a species.



(b) A phylogenetic tree representing the relationship between the tubulin sequences. The branch points, indicated by small numbers, represent common ancestral genes at the time that two sequences diverged.

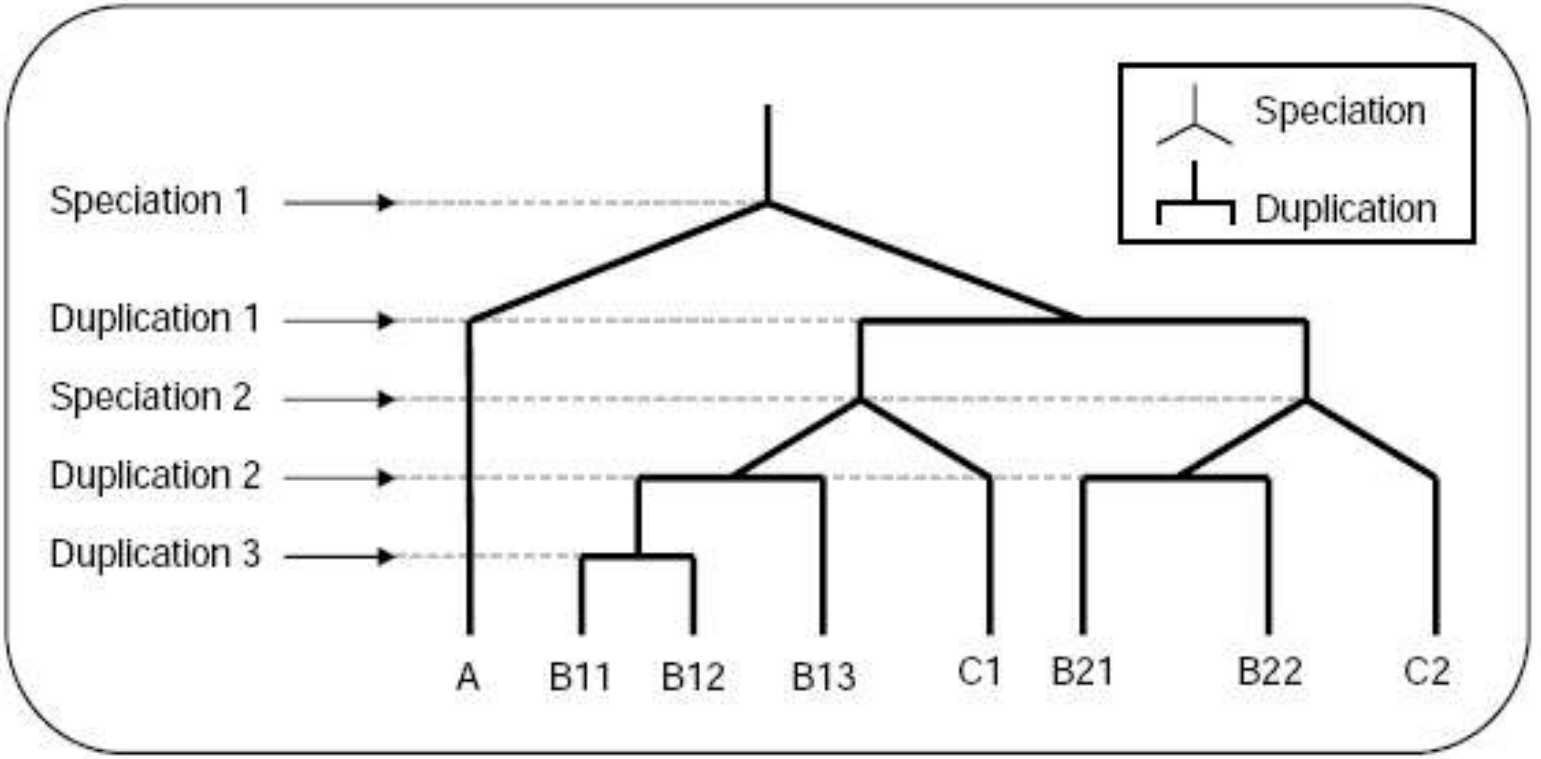
Orthologs and inparalogs

Gene orthology between two organisms is not necessarily a one-to-one reltion – it could be a one-to-many and many-to-many relation.

'**inparalogs**' ('recent' paralogs, true orthologs) indicate paralogs that arose through a gene duplication event after speciation

'**outparalogs**' ('ancient' paralogs) arise following a gene duplication preceding speciation

Outparalogs can never be orthologs, while inparalogs can form a group of genes that together are orthologous to a gene in another species (one-to-many and many-to-many relationship). Clustering inparalogs together allows proper identification of both one-to-one and many-to-many orthology cases



Gene evolution. One gene descending to three organisms A, B and C.

- Two genes whose common ancestor is at a λ junction (speciation) are orthologous, e.g. A with B11, and B21 with C2.
- Two genes whose common ancestor is at a horizontal bar junction (duplication) are paralogs, e.g. B11 with B13, and B12 with C2.
- Genes B11, B12 and B13 are inparalogs to A (and C) because the speciation event 1 (speciation event 2, resp.) occurred before the duplication events that gave rise to B11, B12 and B13.
- Genes B11, B12, B13 and C1 are outparalogs to genes B21, B22 and C2, as the initial duplication occured before B-C speciation.

Methods for finding orthologous genes (based on sequence similarity)

1. Pairwise similiarities

This is all-versus-all sequence comparison between two genomes to detect orthologs.

The principle is that if the sequences are orthologs, they should score higher with each other than with any other sequence in the other genome.

This method does not use multiple alignments or phylogenetic trees and therefore avoids potential errors that might be introduced at these steps.

These methods do not attempt to preserve the non-transitivity and hierarchic nature of the orthology relation. Suits well for comparative genomic analyses for identifying functions of new genes.

Most common approaches following this idea are based on all-versus-all **BLAST** searches.

E.g **Inparanoid** – orthologs and inparalogs from two species (Remm et al., 2001), **OrthoMCL** (extended Inparanoid) -orthologs and inparalogs from two and multiple species (Li et al., 2003), **COG** (Tatusov et al., 1997, 2000, 2001, 2003)

Methods for finding orthologous genes

2. Phylogenetic approach (1/2)

Orthologs are related through evolutionary history; phylogenetic trees (gene family trees) are the most natural way to detect orthologs.

Complete genomes from multiple species can be included in the analysis.

The methods following this approach are constructing phylogenetic trees with some poorly automatable steps and these algorithms demand large resources of computing power.

Methods for finding orthologous genes

2. Phylogenetic approach (2/2)

Exerting this approach for all genes of two or more genomes would require:

- clustering of homologs (problems with separating inparalogs and outparalogs)
- generation of correct multiple alignment for each group of homologous domains
- construction of a phylogenetic tree for each group (the topology of the phylogenetic tree is strongly dependent on the choice of tree building method)
- finally extraction of orthologs from these trees.

E.g **DomClust** (clustering protein sequences at the domain level) (Uchiyama, 2006), **OrthologID** (Chiu et al., 2006)

OrthologID (Chiu et al., 2006)

➔ Developed as a collaborative project by the New York Plant Genomics Consortium (NYPG)

➔ For facilitating the identification of gymnosperm EST sequences that are orthologous to the sequences in the complete genomes of *A. thaliana, O. sativa, P. trichocarpa* and *C. reinhardtii*

Web application that automates the labor-intensive procedures of gene orthology determination within a character based phylogenetic framework.

OrthologID can identify diagnostic characters that:
- define each orthologous gene set
- are responsible for classifying query sequences from other genomes into specific orthologous groups

OrthologID database includes
- several complete plant genomes
- unicellular outgroup

Backend of OrthologID

Databases of sequences (137,641) and phylogenetic trees (8,314)

Four interconnected modules:

1. Gene Family Creator (GFC)
2. Alignment Constructor
3. Tree Builder
4. Diagnostic Generator

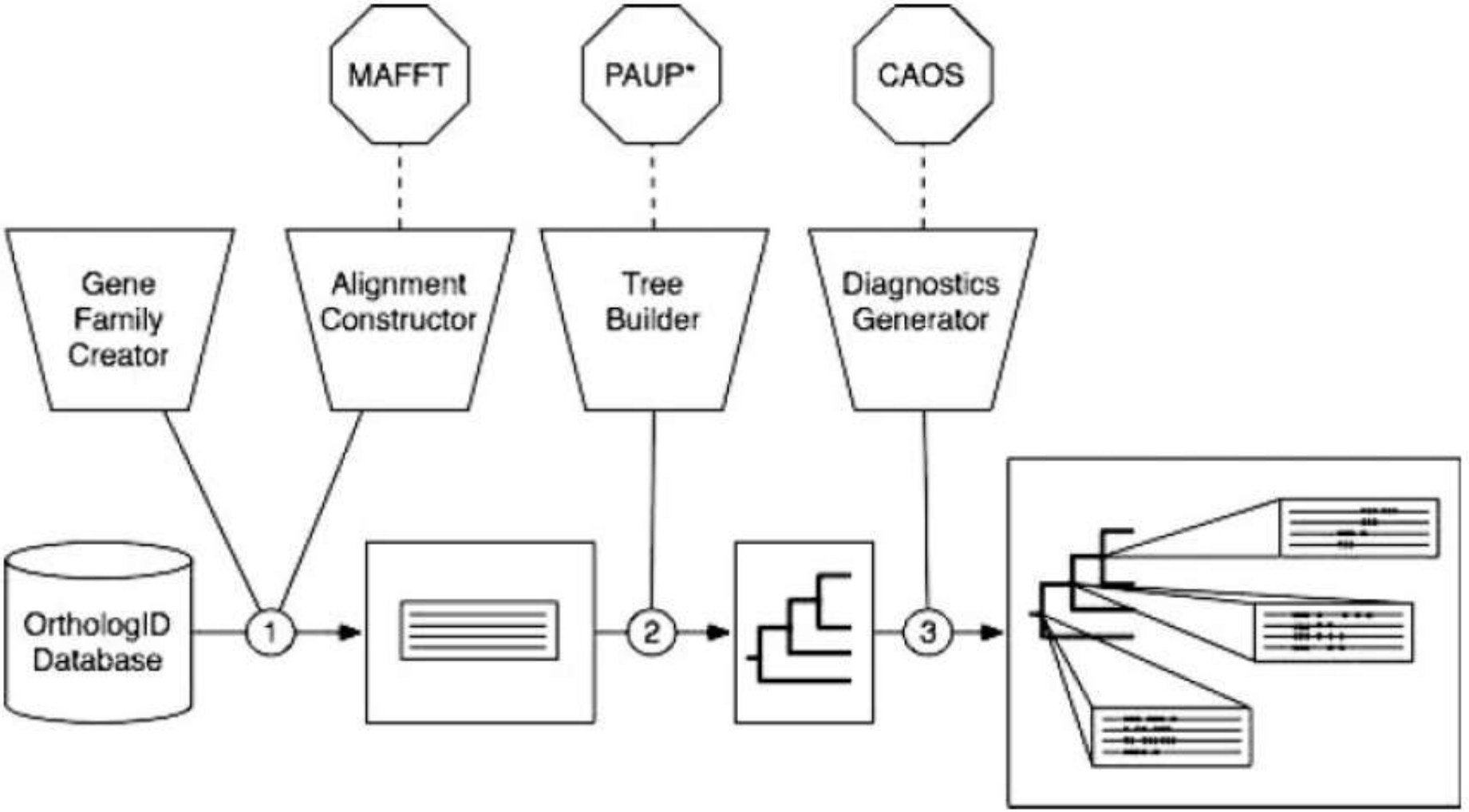


Figure. Overview of OrthologID. Maximum parsimony trees are generated and diagnostic characters are determined through an automated process:
(1) sequences are retrieved from OrthologID Database and clustered using the Gene Family Creator and aligned, using the Alignment Constructor (which interfaces with MAFFT)
(2) Phylogenetic trees are generated using the Tree Builder (which interfaces with PAUP*)
(3) diagnostic characters are ascertained using the Diagnostic Generator (which interfaces with CAOS).
Each OrthologID module, shown as trapezoids, are designed to function independently and allow the use of any processing tool (e.g. One could use ClustalW instead of MAFFT for sequence alignment)

Backend of OrthologID: module 1 – Gene Family Constructor, GFC

Clusters genes from complete genomes into gene families.

- Searches each ingroup gene against both ingroup and outgroup genomes using NCBI BLAST.
- Expectation value cutoff of 1e-20 is used

(For a pair of genes g1 and g2, g1 is defined as clusterable with g2 if the E-value in the BLAST of g1 against g2 is within the cutoff, and the alignable regions of the two genes are at least 80% of the longer sequence. A gene g is considered a member of the gene family F if at least one other gene in F is clusterable with g.)

- After all-against-all BLAST searches, GFC randomly picks a gene g from one of the ingroup genomes and looks for clusterable genes in the BLAST result of g.
- Each clusterable gene is added to the current family, and this gene's BLAST result is again searched for new members.
- Process is repeated until no more genes can be clustered to the current family.
- GFC then starts a new gene family, and the above steps are repeated.

Algorithmically this is realized with graphs.

Backend of OrthologID: module 2 – Alignment Constructor

Creates robust alignments for each gene family.

The multiple alignment program MAFFT version 5 is used for this purpose. MAFFT is considered one of the most efficient and reliable multiple alignment programs based on benchmark tests.

- The Alignment Constructor uses different sets of alignment parameters to create three different alignments for each gene family (three pairs of gap open penalty and offset values are used).
- Alignments are compared and alignment-ambiguous regions are culled.
- The resulting, culled alignment is then passed on to the Tree Builder.

Backend of OrthologID: module 3 – Tree Builder (1/3)

Generates gene family trees within a parsimony framework

Implements the parsimony ratchet (Nixon, 1999) using PAUP*(a pyhlogenetic tree building algorithm, Swofford, 2003).

For small gene families (with fewer than 13 sequences), exhaustive, branch and bound tree searches are performed (as implemented in PAUP) – finding the most parsimonious tree (branch rearrangement on trees are performed and only the most parsimonious trees or subset of suboptimal trees at each step are kept).

Backend of OrthologID: module 3 – Tree Builder (2/3)

For large gene families, tree space is rigorously explored using the parsimony ratchet:
1. An initial starting tree is generated (each iteration of a ratchet starts with a limited TBR ("tree bisection and reconnection") search to generate an initial tree)

2. The tree found in step 1 is used as a starting point for an iterative search strategy

3. A random subset of the characters (10-15%) is selected and perturbed.

4. The current tree is swapped using the perturbed weights to calculate length (typically TBR swapping will be used). Only one (or few) tree is kept during the search with perturbed matrix.

5. The weights are reset to the original weights. Using the current tree as a starting point (this is the final tree found in step 4) swapping proceeds (holding one or few trees) until an optimal tree is found for the unperturbed data

6. Go to 2 (3).

Backend of OrthologID: module 3 – Tree Builder (3/3)

-Each ratchet consists of 200 such iterations (e.g. if 500 taxons and 266 Mhz Pentium -> ca 6h).

-The Tree Builder computes 20 ratchets

-Performs a final TBR swap on the best trees, in order to visit multiple islands (suboptimal trees that are typically very close in both topology and length to the shortest trees) of tree space.

-Where more than one equally parsimonious tree results from the analysis, a strict consensus is computed.

-Consensus tree is used to identify orthology relationships in complete genomes, and used as a gene family guide tree for Query Orthology Classification.

Backend of OrthologID: module 4 – Diagnostic Generator (½)

Identifies diagnostic characters for orthologous groups using the **CAOS algorithm** and '**guide tree**' approach

· CAOS is a rapid algorithm for determining gene orthology based on derived traits shared between orthologous genes

· By the CAOS algorithm, OrthologID classifies new query sequences (full-length cDNA or EST) from genomes that are not completely sequenced, based on the phylogenetic and orthology relationships that are already determined through the analysis of complete genomes

Backend of OrthologID: module 4 – Diagnostic Generator (2/2)

- A complete parsimony gene family tree that is used to identify orthologous groups from complete genomes is used as a guide tree for classifying query sequences from other species.

- This guide tree (from module 3) is fed to the CAOS algorithm for the identification of characters that are diagnostic of each node and each orthologous gene set.

- In order to place query sequences into orthology groups assembled from complete genomes, CAOS screens the query sequence for the presence of characters that are diagnostic of nodes on the guide tree.

The CAOS algorithm and the use of guide trees are an improvement over traditional tree building approaches since the guide tree/CAOS approach eliminates the need to manually rebuild a gene family tree for each new query to be classified.

Frontend of OrthologID: web interface

http://nypg.bio.nyu.edu/orthologid/

Orthologous groups are presented through the **OrthologID Tree and Diagnostics Viewer** in an interacive phylogenetic tree format**.**

Allows users to:

1. *Orthologous group search* - search for orthologous gene sets in complete genomes that are available in the OrthologID database.

OrthologID

Orthologous Group Search

Query Orthology Classification

About OrthologID

Welcome to OrthologID

OrthologID automates gene orthology determination within a character-based phylogenetic framework.

OrthologID identifies orthologous groups for complete genomes compiled in our database (*Orthologous Group Search*), and classifies user-input query sequences into orthologous groups generated from complete genomes (*Query Orthology Classification*). It identifies diagnostic characters that define each orthologous group, as well as diagnostic characters responsible for classifying query sequences. The output is presented in phylogenetic tree format.

OrthologID

Orthologous Group Search

Query Orthology Classification

About OrthologID

Orthologous Group Search

Enter a locus tag (*Arabidopsis thaliana* or *Oryza sativa* only)
e.g. At4g15440 or Os05g49290

At4g15440 Search

OrthologID

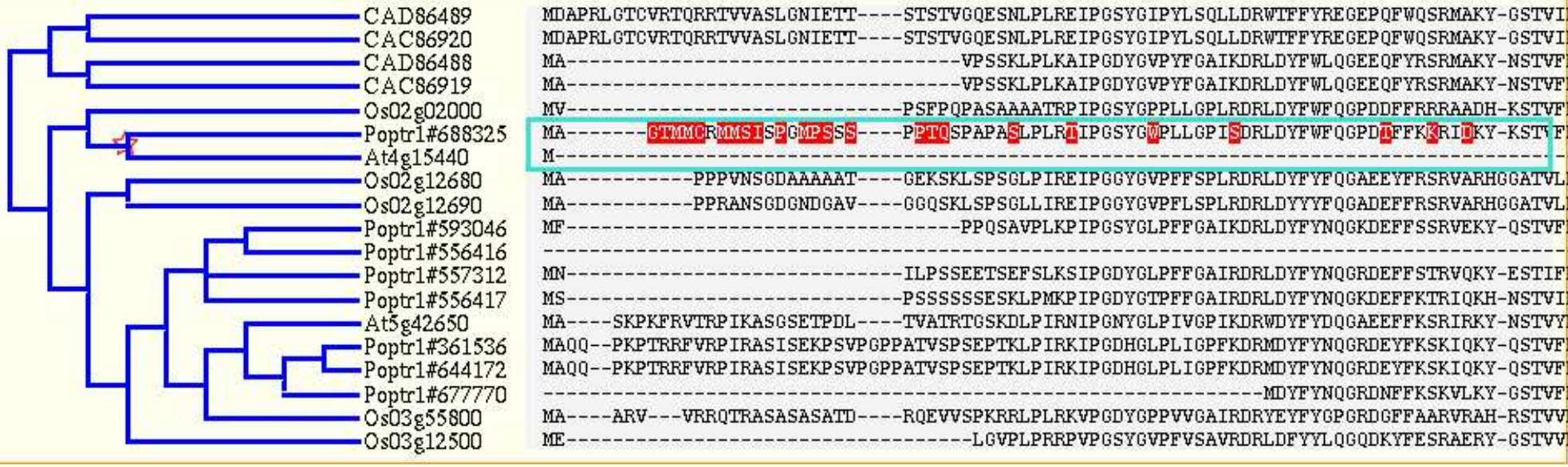
Tree and Diagnostics Viewer

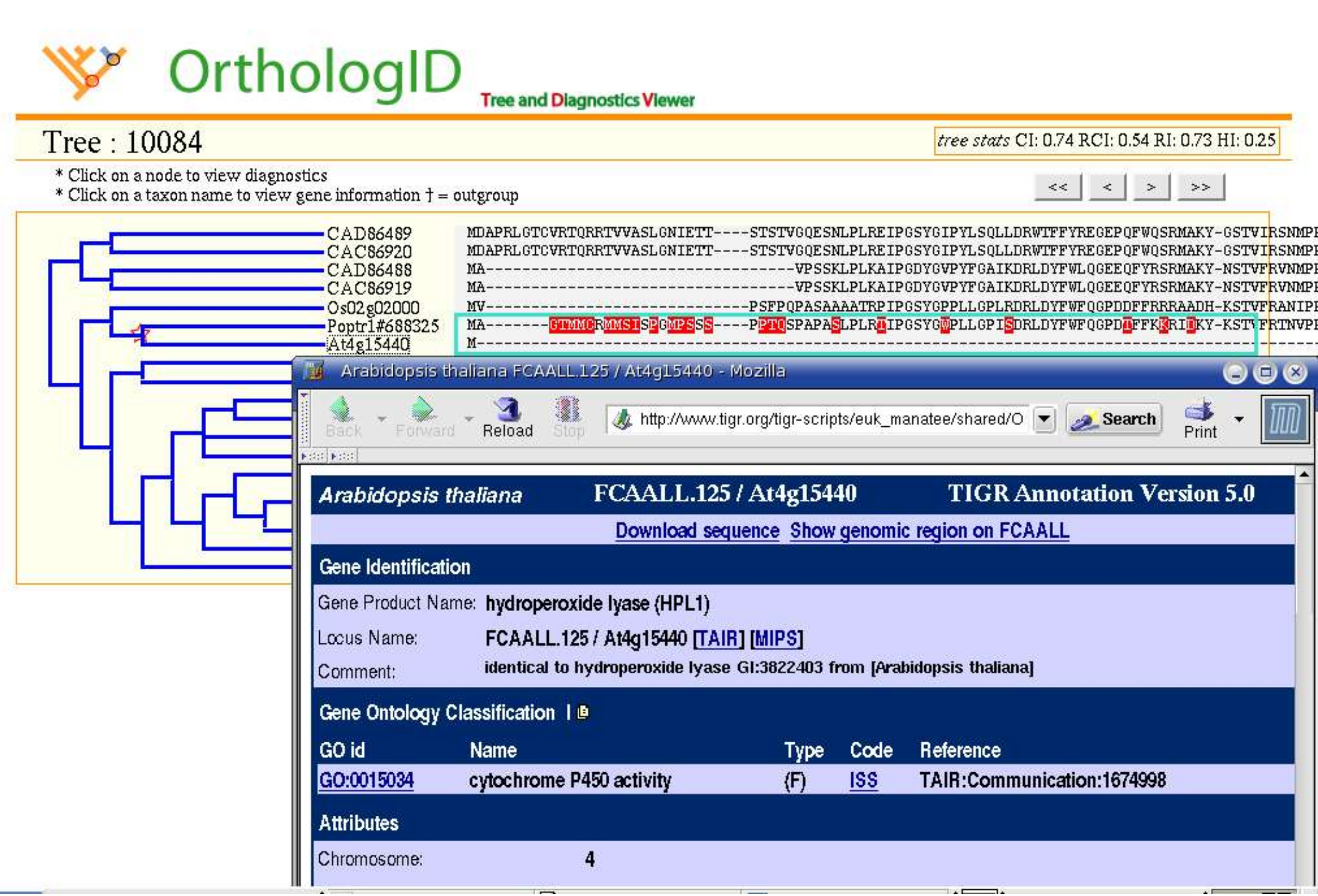
Tree : 10084

tree stats CI: 0.74 RCI: 0.54 RI: 0.73 HI: 0.25

* Click on a node to view diagnostics
* Click on a taxon name to view gene information † = outgroup

<< < > >>




OrthologID
Tree and Diagnostics Viewer
Tree : 10084
tree stats CI: 0.74 RCI: 0.54 RI: 0.73 HI: 0.25
* Click on a node to view diagnostics
* Click on a taxon name to view gene information † = outgroup
<< < > >>
CAD86489
CAC86920
CAD86488
CAC86919
Os02g02000
Poptr1#688325
At4g15440
MDAPRLGTCVRTQRRTVVASLGNIETT----STSTVGQESNLPLREIPGSYGIPYLSQLLDRWTFFYREGEPQFWQSRMAKY-GSTVIRSNMPE
MDAPRLGTCVRTQRRTVVASLGNIETT----STSTVGQESNLPLREIPGSYGIPYLSQLLDRWTFFYREGEPQFWQSRMAKY-GSTVIRSNMPE
MA----------VPSSKLPLKAIPGDYGVPYFGAIKDRLDYFWLQGEEQFYRSRMAKY-NSTVFRVNMPE
MA----------VPSSKLPLKAIPGDYGVPYFGAIKDRLDYFWLQGEEQFYRSRMAKY-NSTVFRVNMPE
MV----------PSFPQPASAAAATRPIPGSYGPPLLGPLRDRLDYFWFQGPDDFFRRRAADH-KSTVFRANIPE
MA-------GTMMCRMMSISPGMPSSS----PPTQSPAPASLPLRTIPGSYGWPLLGPISDRLDYFWFQGPDTFFKKRIDKY-KSTVFRTNVPE
M----------
Arabidopsis thaliana FCAALL.125 / At4g15440 - Mozilla
Back
Forward
Reload
Stop
http://www.tigr.org/tigr-scripts/euk_manatee/shared/O
Search
Print
Arabidopsis thaliana
FCAALL.125 / At4g15440
TIGR Annotation Version 5.0
Download sequence
Show genomic region on FCAALL
Gene Identification
Gene Product Name: hydroperoxide lyase (HPL1)
Locus Name: FCAALL.125 / At4g15440 [TAIR] [MIPS]
Comment: identical to hydroperoxide lyase GI:3822403 from [Arabidopsis thaliana]
Gene Ontology Classification
GO id
Name
Type
Code
Reference
GO:0015034
cytochrome P450 activity
(F)
ISS
TAIR:Communication:1674998
Attributes
Chromosome: 4

Frontend of OrthologID: web interface

2. *Query orthology classification* - classifies query sequences into existing orthology groups from complete genomes.

The users query sequence is obtained classification support index (CSI). This is the difference in the number of diagnostics in the descendent clades.

For every query the diagnostic characters responsible for placing the query into a particular clade are obtained. Additionally the query placement score is defined by the function:

$$S(a) = k_a / \left(k_a + k_{b1} + k_{b2} + \ldots + k_{bn} \right) * 100\ \%$$

where k_a is the number of diagnostic characters shared between the query and the sequence(s) in clade a and k_{bi} is the number of diagnostic characters shared between the query and the sequence(s) in one of the n sister clades, b_i of a.

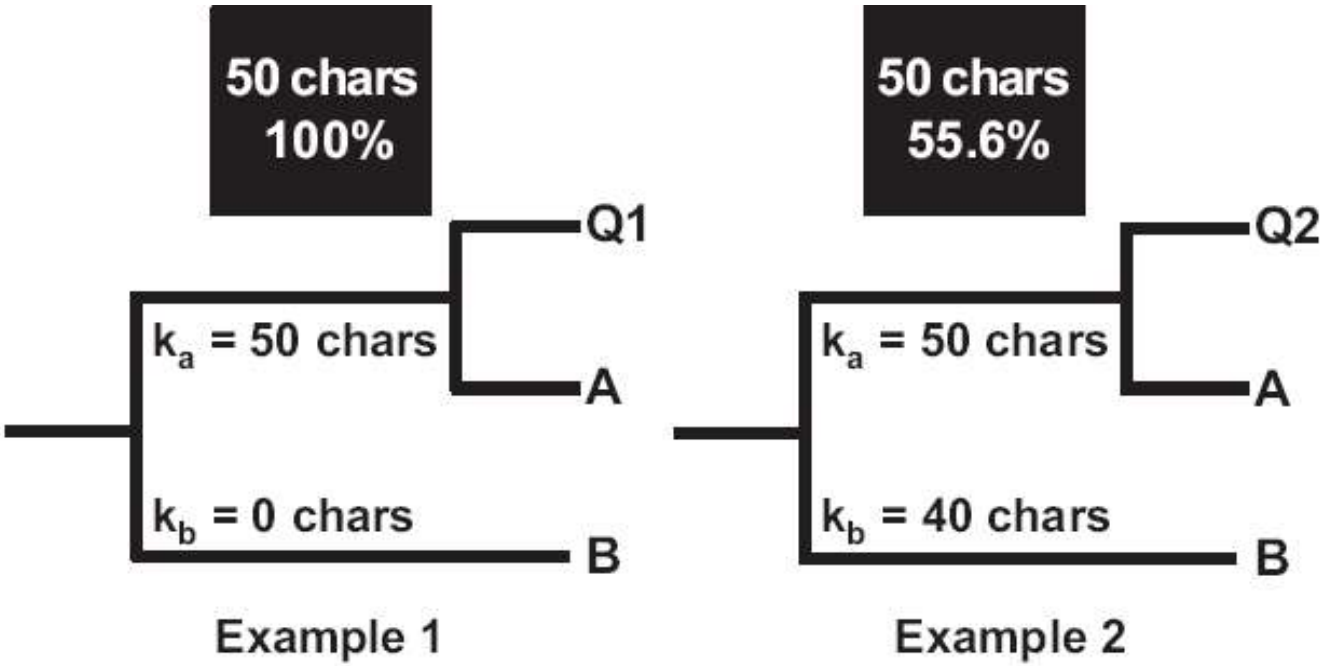


Fig. 3. Two examples illustrating the calculation of query placement score given in the pop-up boxes in the Tree and Diagnostics Viewer. In Example 1, query sequence Q1 shares 50 characters ($k_a = 50$) with the gene(s) in clade A and zero characters ($k_b = 0$) with the gene(s) in clade B. As a result, OrthologID places Q1 into clade A. The query placement score is expressed as $(k_a/k_a + k_b) \times 100\%$. The resulting score (100%) and the number of diagnostic characters responsible for placing Q1 (50) are shown in the pop-up box. In Example 2, $k_a = 50$ and $k_b = 40$; as a result, the strength of the placement of query Q2 is weaker than that of Q1 in Example 1. This is illustrated by the lower query placement score in Example 2 (55.6%).

OrthologID

Orthologous Group Search

Query Orthology Classification

About OrthologID

Query Classification

```
>Z_mays_AAS47027
MLPSFVSPTASAAASVTPPPRPIPGSYGPPVLGPLRDRLDYFWFQSQDEF
FRKRAAAHRSTVFRTNIPPTFPFFVGVDPRVVAIVDAAAFTALFDPDLVD
KRDILIGPYNPGAGFTGGTRVGVYLDTQEEEHARVKTFAMDLLHRSARTW
SADFRASVGAMLDAVDAEFGKDDGSDKKPSASYLVPLQQCIFRFLCKAFV
GADPSADWLVDNFGFTILDIWLALQILPTQKIGLVQPLEELLIHSFPLPS
FLIWPGYYVLYRFIEKHGAEAVAYAEAQHGIGKKDAINNILFVLGFNAFG
GFSVFLPFLVAKVGGAPALRERLRDEVRRAMVGKDGEFGFATVREGMPLV
RSTVYEMLRMQPPVPLQFGRARRDFVLRSHGGAAYQVSAGEVLCGYQPLA
MRDPEVFERPEEFVPERFLGDEGARLLQHLFWSNGPETAQPGPGNKQCAA
KEVVVDTACMLLAELFRRYDDFEVEGTSFTKLVKRQASPSVAQAAAAAGA
QQ
```

Identify Ortholog | Reset | Try an example | What is FASTA format?

OrthologID

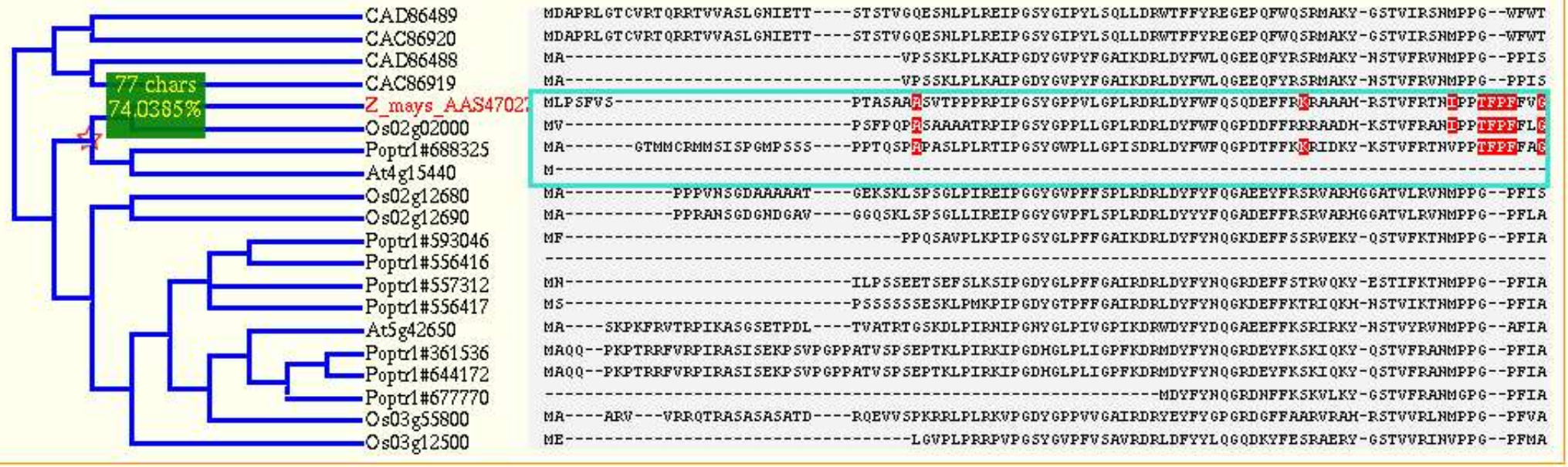
Tree and Diagnostics Viewer

Query : Z_mays_AAS47027

tree stats CI: 0.71 RCI: 0.50 RI: 0.70 HI: 0.28

* Click on a node to view diagnostics * Mouse-over a node to view query classification scores
* Click on a taxon name to view gene information † = outgroup

<< < > >>



Results of efficacy tests of OrthologID (1/2)

1. The placement of query sequences of OrthologID against the placement generated using full-scale parsimony analyses were examined:

- 36 plant sequences (other than the ones whose genomes are included in OrthologID database) were randomly selected from NCBI and New York Plant Genome (NYPG) databases.
- These genomes were submitted to OrthologID against the complete genomes of *A.thaliana, O. Sativa, P. Trichocarpa* and *C. Reinhardtii.*
- 77.8% of the 36 plant query sequences OrthologID and full-scale parsimony analyses resulted in the same orthology classification.

Results of efficacy tests of OrthologID (2/2)

2. The effectiveness of OrthologID for identifiyng orthologous gene sets (orthologs and inparalogs) for query sequences were examined:

- 36 plant query sequences from diverse range of plant species against the current OrthologID plant database

- 66.7% were placed into orthology groups with single orthologs or groups of inparalogs

Advantages of OrthologID:

- Parsimony phylogenetic analysis is a natural way to detect orthologs, thus using character-based parsimony framework for finding orthologous groups is a good approach
- OrthologID can differentiate between inparalogs and outparalogs
- Determines orthology relationsships between more than two complete genomes simultaneously
- OrthologID can screen query sequences (cDNA, EST) from new genomes for diagnostic characters and place them in orthology groups compiled using completely sequenced genomes (due to CAOS algorithm)
- OrthologID identifies diagnostic characters of orthologous gene sets

For future purposes:

To increase OrthologID scope and general utility the OrthologID ***database will be expanded*** to include complete genomes from other phylogenetic lineages, including ***prokaryotes and non-plant eukaryotes***.

References:

1. Remm M, Storm CEV, Sonnhammer ELL. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology.* 314(5), 1041-52.

2. Uchiyama I. (2006). Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. *Nucleic Acids Res.* 34(2), 647-58.

3. Li L, Stoeckert CJ, Roos Jr, Roos DS. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13(9), 2178-89.

4. Jothi R, Zotenko E, Tasneem A, Przytycka TM. (2006). COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics.* 22(7), 779-88.

5. Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, DeSalle R. (2006). OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics*. 22(6), 699-707.

6. Dutilh BE, Huynen MA, Snel B. (2006). A global definition of expression context is conserved orthologs, but does not correlate with sequence conservation. *BMC Genomics.* 7(10), 739-49.

7 Nixon, KC. (1999). The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis. *Cladistics.* 15, 407-414.