Chapter 2

Gene Orthology Assessment with *OrthologID*

Mary Egan, Ernest K. Lee, Joanna C. Chiu, Gloria Coruzzi, and Rob DeSalle

Abstract

OrthologID (http://nypg.bio.nyu.edu/orthologid/) allows for the rapid and accurate identification of gene orthology within a character-based phylogenetic framework. The Web application has two functions – an orthologous group search and a query orthology classification. The former determines orthologous gene sets for complete genomes and identifies diagnostic characters that define each orthologous gene set; and the latter allows for the classification of unknown query sequences to orthology groups. The first module of the Web application, the gene family generator, uses an E-value based approach to sort genes into gene families. An alignment constructor then aligns members of gene families and the resulting gene family alignments are submitted to the tree builder to obtain gene family guide trees. Finally, the diagnostics generator extracts diagnostic characters from guide trees and these diagnostics are used to determine gene orthology for query sequences.

Key words: Single linkage cluster, orthology, phylogeny, alignment, diagnosis genomics.

Homology is either the cornerstone of biology or a term ripe for burning.
John Maynard Smith

1. Introduction

Sub-genomic studies (that analyze hundreds to thousands of gene regions for perhaps hundreds of terminal taxa) face new computational challenges. Responding to these challenges is resulting in the development of new methodologies and tools for tree building and analysis. Sub-genomic studies, however, follow the same ground plan as most molecular systematic studies, albeit on a grand scale. It may be that the next stimulus to intellectual debate in the field of systematics will come as a result of responding to

David Posada (ed.), *Bioinformatics for DNA Sequence Analysis,* Methods in Molecular Biology 537

DOI 10.1007/978-1-59745-251-9_2

challenges faced by true phylogenomic studies in which entire genomes are used as the basis for comparison. The presence of gene families and horizontal gene transfer pose challenges for both character coding as well as for identifying which are the appropriate terminals in the analysis. These challenges to be faced in phylogenomic studies hearken back to those faced in morphological studies. There is likely to be a shift in the direction of intellectual debate in the phylogenetic analysis equation, the direction of this shift being toward data matrix assembly and homology assessment. Several types of genomic studies are already implicitly or explicitly addressing the problem of homology.

Homology is often considered one of Darwin's most impressive contributions to evolutionary thinking. Darwin was one of the first to discern the differences between homology and analogy. Whereas homology refers to traits that are the same due to common ancestry, analogy refers to traits that are similar due to evolutionary convergence. Homology then becomes a term of absoluteness and must be discovered via hypothesis testing, and similarity becomes a term of measurement and is calculated from observation of two entities. The literature on homology is rich with debate, hence the quote that starts off this chapter. While phylogenomic studies may stimulate intellectual debate and growth in systematic theory, the converse may also be true – that systematics may provide an aspect of the intellectual underpinning necessary for the continued development of genomic studies.

In the 1980s Fitch and several colleagues published an important clarification of terms then being used in the earliest of comparisons of molecular sequences. Their note suggested that scientists take care in using the word homology. Essentially, Fitch et al. pointed out that most molecular biologists at the time were misusing and hence overusing the term homology. For instance, a typical sequence comparison paper from that period would claim that two sequences of 100 amino acids long had 70% homology if 70 out of 100 residues in the two proteins were the same. Fitch and colleagues pointed out that this was a misuse of the term homology, which indicates common ancestry. When two sequences are compared there is necessarily a lack of investigation of common ancestry because it takes at least three target sequences and an outgroup sequence to discover common ancestry. Those readers who have heard Fitch speak are probably familiar with his famous punch line about homology – "Homology is like pregnancy. Someone is either pregnant or not. A person cannot be 70% pregnant." In this context, two sequences can be homologous, but cannot be 70% homologous. Rather two sequences are 70% similar (for further discussion of similarity versus common ancestry in orthology assessment, *see* **Note 1**).

In addition to establishing this important distinction concerning homology, Fitch and colleagues also established a framework for how we should examine genes in multigene families, by proposing

that the term orthology refer to genes that are identical by descent as a result of speciation of two entities. The term paralogy then refers to a pair or set of genes related to each other but not through a speciation event. In this case, paralogy refers to members of a gene family that have arisen through duplication and not followed by a speciation event. While both terms refer to kinds of homologous relationships, orthology is what an anatomist would refer to as homology and paralogy would be akin to serial homology. A third term coined xenology refers to the similarity of entities as a result of horizontal transfer. It should be obvious from the terminology that any departure from orthology for members of gene families complicates and even negates sound evolutionary or biological analysis. The old adage of comparing apples to oranges also applies to genes in gene families.

To automate the rapid assessment of orthology of genes in gene families we have developed a Web based program called *OrthologID (1)*. This program uses the concept of common ancestry to establish homologous relationships of genes obtained from whole genome sequencing, EST studies, or other genome analyses. There are other methods that exist that have been used to establish orthology, such as simple BLAST, BLAT, and COG approaches. Because BLAST best hits have been shown not to identify the closest phylogenetic neighbor *(2)*, a problem exists with relying solely on this approach (and indeed others that rely on BLAST or are similar to BLAST). However, BLAST, BLAT, and other techniques such as COGs and other distance measures (*see* **Note 1**) can be informative first steps in topographical assessment. We consider these approaches to be valid generators of *hypotheses* when determining orthology and as we point out below they are incorporated into the algorithm we have developed. Our description of the program will first describe the rationale for the approach we have devised, then describe in detail the algorithm and its component parts, and finally some worked examples are presented.

An automated approximation of the phylogenetic gene tree approach to orthology determination would need to be developed in order to be able to use this approach on a genomic scale. For small gene families or for limited numbers of taxa, it is possible to use this approach with currently available analytical tools. These would involve initial similarity searches to identify putative gene families (for multiple taxa simultaneously), alignment, tree building, and screening trees for diagnostic characters to identify orthogous gene family members of each taxon, These analyses would be repeated for each new sequence to be placed among its orthologous group. To use this approach on a genomic scale, new tools have been developed. The main difference between the manual phylogenetic gene tree approach and the automated approach implemented in *OrthologID* is the exclusive use of completely sequenced genomes for constructing gene family trees. These

trees (termed guide trees) are screened for the presence of characters diagnostic of orthologs using the CAOS algorithm *(3)* and the identification of unknown queries is made through comparison of the guide tree diagnostics and the query sequence. In this way trees are not required to be constructed each time a new query's orthology is identified.

Table 2.1
Description of homology approaches showing the methods used to establish homology and focus of application

Approach and steps	Description of step	Purpose or method	Applied to
dePinna (dP)			
1. Primary homology	Establish character coding	Interpret anatomy (similar to character assignment)	Anatomy
2. Secondary homology	Phylogenetic analysis	Discover shared and derivedness to establish homology (identical to step 3 in BS)	
Brower Schawaroch (BS)			
1. Topographical similarity	Sequence alignment	New aspect not in dP	Molecular sequences
2. Character assignment	Assess character transformations	Simple for DNA and proteins to establish homology	
3. Phylogenetic analysis	Discover shared and derivedness	(Identical to step 2 in dP)	
Gene family homology (GFH)			
1. Topographical similarity	BLAST/BLAT	Establishes hypothesis of gene family inclusion	Gene presence absence studies; gene family studies
2. Character assignment	Alignment of gene	Establishes character assignment for sequences	
3. Phylogenetic analysis	Because target now is organismal phylogenetic analysis accomplishes gene homology assessment	Establishes gene family homologies by demonstrating shared derived origin for family members	

Establishing homology or orthology of genes in gene families can be viewed as slightly different from the way that DNA sequence characters are treated during alignment (for a discussion of the phylogenetic basis of homology assessment as applied to sequence alignment, *see* **Note 2**).

The *OrthologID* approach to orthology assessment is similar in many ways to the Brower and Schawaroch *(4)* scheme for homology assessment of phylogenetic characters via alignment. The differences between that approach and orthology determination is as follows: First, the potential members of gene family are identified using topographical similarity. Topographical similarity for orthology studies is very similar to the Brower and Schawaroch *(4)* scheme, except that the determination of topographical similarity is complicated by potential paralogy problems. Instead of going straight to alignment to determine topographical similarity a preliminary step of determining group membership using similarity as a means to assess group membership is implemented. The inclusion of genes into a particular gene family is often accomplished by setting a similarity cutoff (usually using similarity comparisons like E-values of sequences like BLAST *(5)*, BLAT, or COG *(6)*; *see* **Note 3**) and including all genes in an ortholog group that conform to the predetermined cutoff. Once this first step is accomplished the alignment step can be undertaken. Once the alignment step is accomplished the establishment of character state identity is straightforward and can be followed by the test of the hypothesis using phylogenetic approaches. **Table 2.1** summarizes the differences in the three ways of looking at homology.

2. Program Usage

The Web based *OrthologID* server allows the user to input a sequence or sequences of unknown orthology and receives the ortholog identification along with critical diagnostics for the ortholog groups. The *OrthologID* approach was developed specifically to handle the burgeoning amount of EST data from plant genomics. An overview of the *OrthologID (1)* approach is shown in **Fig. 2.1**. *OrthologID* uses the three step approach of orthology identification that are outlined in **Table 2.1** and discussed above. The approach is similar in some ways to PhiG developed by Dehal and Boore, (*7*; *see* **Note 4**).

The program is based on the structure and maintenance of a database that we call the *OrthologID* database (**Fig. 2.1**). To date the Plant *OrthologID* database is composed of five fully sequenced genomes – *Arabidopsis thaliana, Oryza sativa, Populus trichocarpa, Physcomitrella patens*, and *Chlamydomonas reinhardtii*,

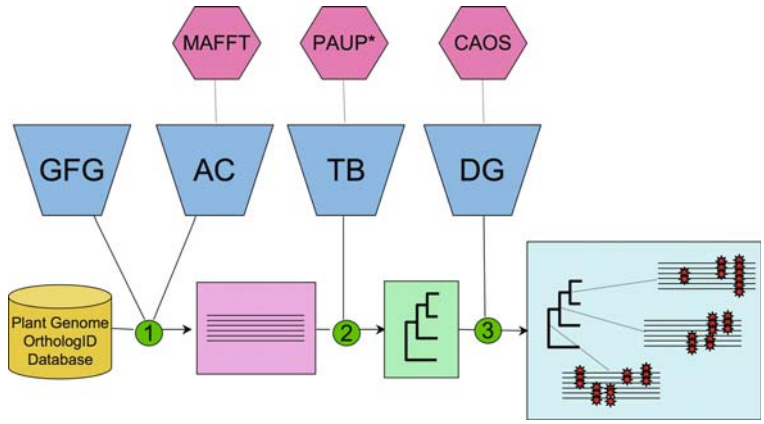

Fig. 2.1. Schematic diagram of the *OrthologID* pipeline. Existing programs are depicted as hexagons (MAFFT, PAUP, and CAOS). Trapezoids indicate *OrthologID* operations implemented by existing software. The rectangles at the bottom represent the product of each step in the pipeline; Step 1: the alignment product, Step 2: the phylogenetic tree product, and Step 3: the generation of diagnostics. Abbreviations: GFG = gene family generator; AC = alignment constructor; TB = tree builder; DG = diagnostics generator.

with a total of 179,005 genes. This database can be continually updated when new fully sequenced and annotated genomes come online. Only sequences from organisms with well-annotated whole genome sequences are used in the construction of guide trees that train the CAOS algorithm to find "diagnostic" amino acid or nucleic acid sites and uses those diagnostics to place an unknown from a less densely sampled genome into an ortholog group. We prefer using genes from only fully sequenced genomes to produce guide trees for two reasons. First, the fully sequenced genomes are also annotated to a better degree than incomplete genomes and EST projects. Second, the absence of a particular gene in a gene family cannot be determined from a partial genome or from EST sequences of the transcriptome of a genome; we consider guide trees constructed from such genomes to be potentially incomplete.

Emanating from the database are four subprograms that perform the following four tasks leading eventually to the construction of a guide tree for a particular gene family.

(1) A gene family generator (GFG; **Fig. 2.1**) that utilizes an e-value based approach to sorting genes into "gene families." Unlike the single linkage cluster algorithm mentioned in **Note 4** or the Coginator that is used to generate COG *(6)* families, we use only raw E-values to sort through the genes to place them in families.

(2) Next an alignment constructor (AC; **Fig. 2.1**) takes the genes placed into a gene family and aligns them with default parameters of the MAFFT *(8)* alignment program.

(3) The gene family alignments are then analyzed in the tree builder (TB; **Fig. 2.1**) using parsimony in PAUP* *(9)*.

(4) The trees are then processed by a diagnostics generator (DG; **Fig. 2.1**) where they are used as guide trees. Diagnostics are extracted from the guide trees using the CAOS algorithm *(3, 10)*.

In this way for each gene family a set of diagnostic rules is generated that can then be used in the Web interface to facilitate the identification of unknown query sequences supplied by the user. Because the diagnostic rules are used to generate the identification, the identifications can also include the diagnostics for inclusion of a query into an ortholog group.

The *OrthologID* program is structured so that other phylogenetic, alignment and DGs can be interchanged. For instance, the basic program can accommodate other alignment programs than MAFFT *(8)*, or other phylogenetic tree building programs than PAUP* *(9)*. In addition, while we prefer parsimony as the method for generating trees, the general program is built to also accommodate likelihood, distance, or Bayesian methods.

To date *OrthologID* has been constructed to facilitate identification of plant orthologs. We are in the process of extending the approach to accommodate other databases, and these new databases will operate the same as the Plant *OrthologID* database. The *OrthologID*website (**Fig. 2.2**) can be accessed at http://nypg.bio.nyu.edu/*orthologID/*. There are three "hot" buttons on this page: an "orthologous group search" button, a "query orthology classification" button, and an "about *OrthologID*" button. These hot buttons gain the user access to the two main ways to interact with *OrthologID*. The first, the "orthologous group search" button, allows the user to access

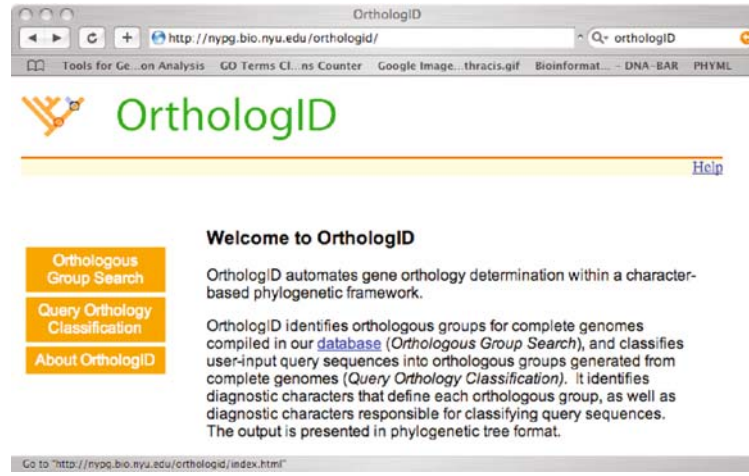

Fig. 2.2. Screenshot of *OrthologID* homepage (http://nypg.bio.nyu.edu/orthologid/).

specific gene family trees and to observe the diagnostics determined by *OrthologID*. This option requires a TAIR gene number for a gene or gene family annotated by The Arabidopsis Information Resource (http://www.arabidopsis.org/) for input. The second way of interacting with *OrthologID* is through the "query orthology classification" button. This option takes the query sequence and attaches it to the gene family tree and shows diagnostics for the whole gene family tree. This option requires a FASTA file of a gene sequence either of known gene family membership or an unknown sequence as input. The third button on the page gains access to the online introduction to *OrthologID*. Below we show worked examples for both options.

3. Examples

3.1. Worked Example # 1. Orthologous Group Search

Press the "orthologous group search" button on the *OrthologID*-homepage. Then, simply paste a query TAIR Gene Family Number into the provided query box (the red circle in **Fig. 2.3**). In this case we have pasted the TAIR number of the malate dehydrogenase gene (AT3G47520) family into the query box.

Press the "Search" button to the right of the query box. If the query is in the Plant *OrthologID* database, the *OrthologID* web server will return the identification of the gene family at the top of the screen (circled in red), a phylogenetic tree on the left and alignment of sequences from gene family members on the right (**Fig. 2.4**). The buttons on the top right of the screen allow the user to scroll across the alignment as indicated.

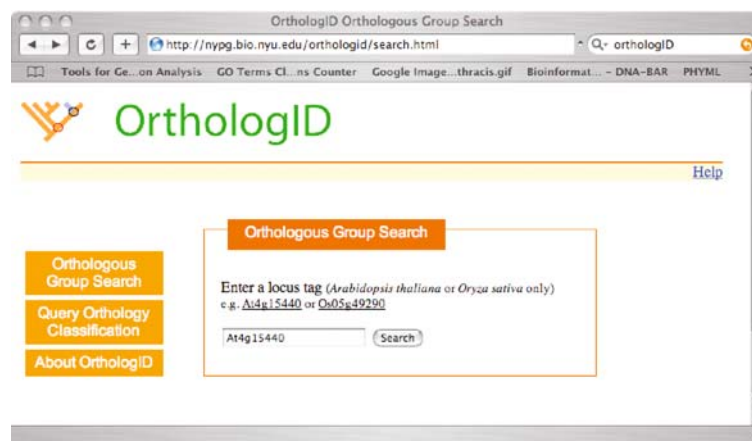

Fig. 2.3. Screenshot of web Group Search in *OrthologID* (http://nypg.bio.nyu.edu/orthologid/search.html).

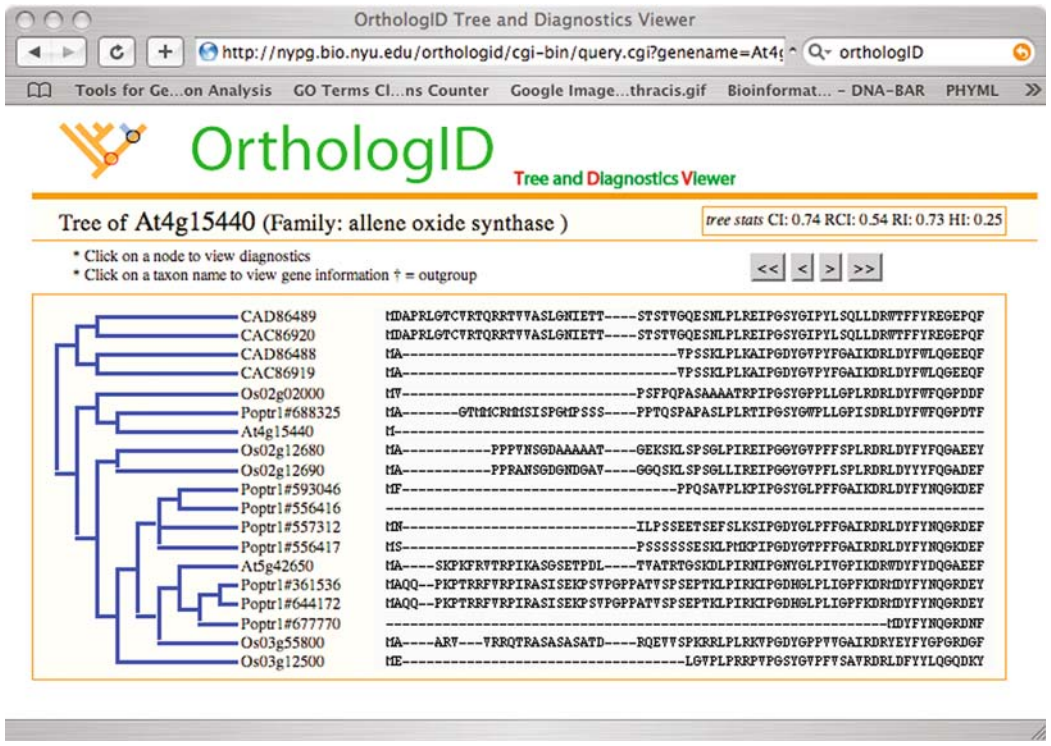

Fig. 2.4. Screenshot of results of querying the Group Search page with AT3G47520. The *arrow* indicates where the diagnostic query for **Fig. 2.5** is clicked.

This screen has two further interactive tools. First, any of the nodes can be clicked on to show the diagnostic amino acid sites and their states for the various ortholog groups displayed (by scrolling over the phylogenetic tree's nodes and clicking on any node). In this example we have clicked on the node near the arrow. When this node is clicked, all of the diagnostic sites in the guide tree are highlighted in red in the accompanying alignment and the group under examination is boxed off in light blue (**Fig. 2.5**).

For any gene in the tree to the right, by simply clicking on the gene number, full information on that gene can be obtained. In this case we have clicked on the *Populus* gene family member 686130 and this links out to the JGI *Populus trichocarpa* website with all of the annotation information for this specific protein.

3.2. Worked Example # 2. Query Orthology Classification

Press the "query orthology classification" button on the *OrthologID*-homepage. Paste a query FASTA sequence from an unknown protein sequence. In this case we have pasted a *Solanum* malate dehydrogenase gene family member into the query box (**Fig. 2.6**). The search button starts the search procedure. If the gene family that the query sequence belongs to is not in the Plant *OrthologID* database, then a "No Hits" response will be returned.

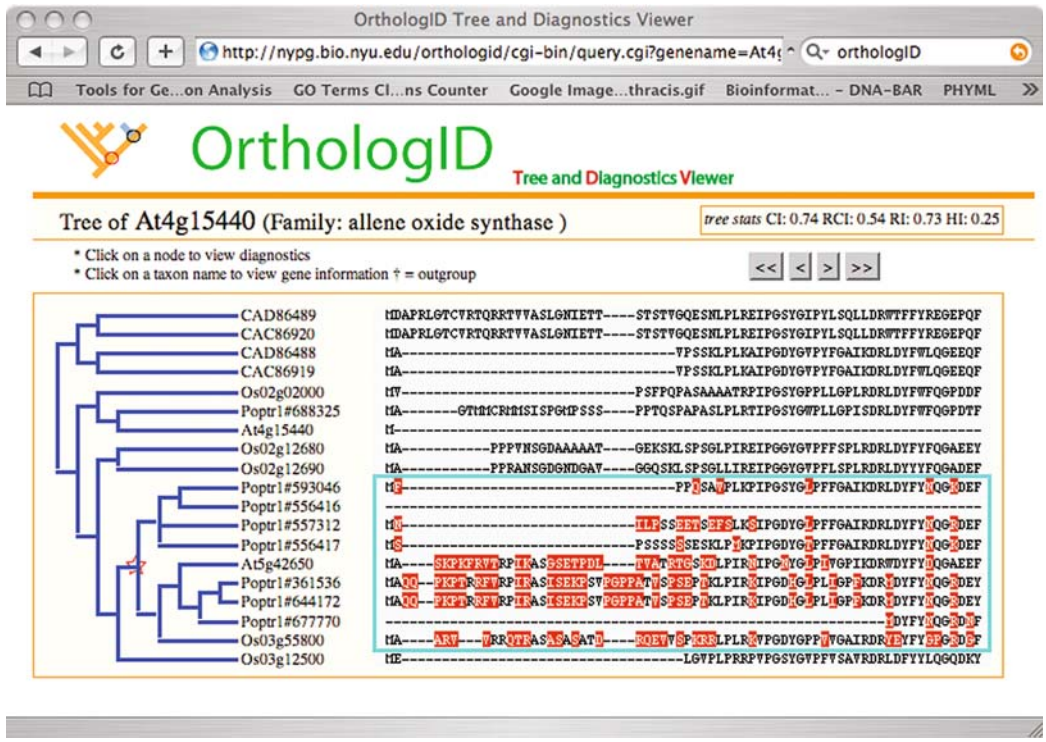

Fig. 2.5. Screenshot of querying the tree at the position marked by an *arrow* in **Fig. 2.4**. A *star* marks the node of interest and the characters that are involved in diagnosis of the group of genes in the box are highlighted.

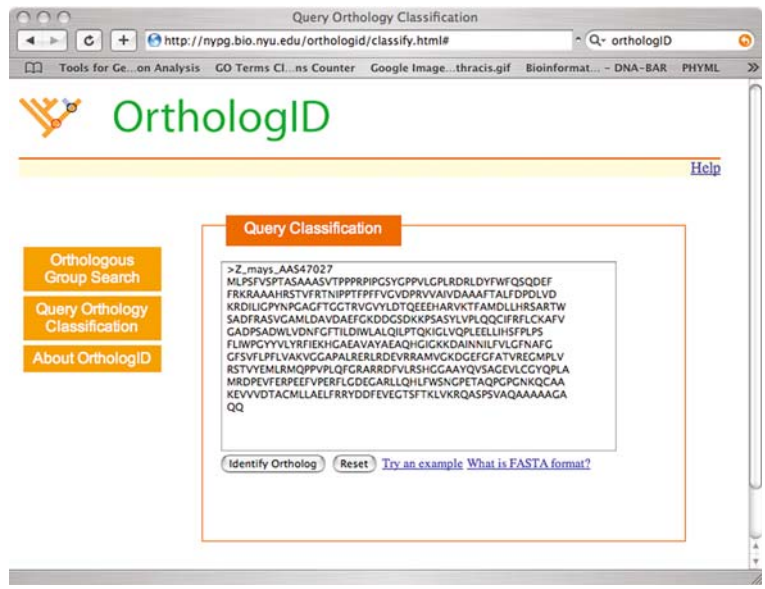

Fig. 2.6. Screenshot of the query ortholog classification in *OrthologID* (http://nypg.bio.nyu.edu/orthologid/classify.html).

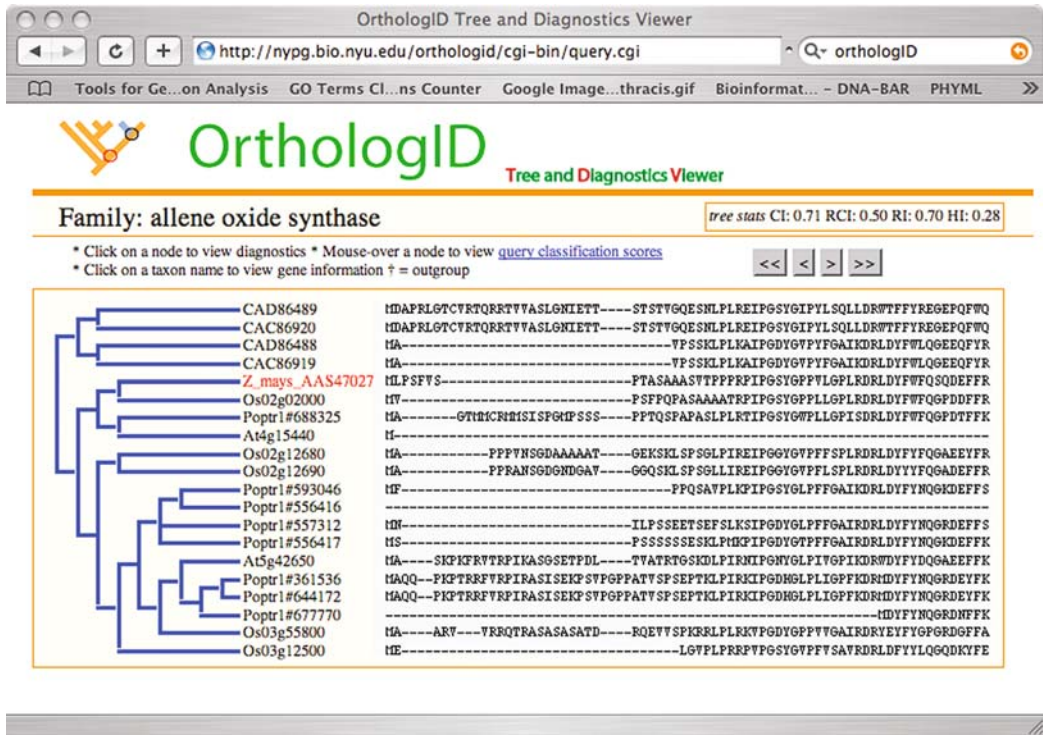

Fig. 2.7. Screenshot of the results of a query sequence placed in the query ortholog classification in *OrthologID*. The *oval* indicates the gene family the query sequence is orthologous to. The query sequence is shown attached to a node in the tree.

If the query sequence home gene family is in the Plant *OrthologID* database, the *OrthologID* web server will return the identification of the gene family at the top of the screen (circled in red), a phylogenetic tree on the left with the query sequence attached to the tree (in red), and an alignment of sequences from gene family members on the right (**Fig. 2.7**).

As with the results from the orthologous group search function, the tree can also return annotated gene information for each gene in the tree, by clicking on the gene name or number, and the diagnostic sites for each of the potential groups in the tree by clicking on the nodes in the tree (**Fig. 2.8**).

Another interactive aspect of the "query orthology classification" page is also shown in **Fig. 2.8**. When the node that contains the original query sequence is clicked, two numbers appear. The first gives the number of characters that are part of the diagnostics and the second number gives a score for the placement of the query sequence into the group it is placed (*see* **Note 6**).

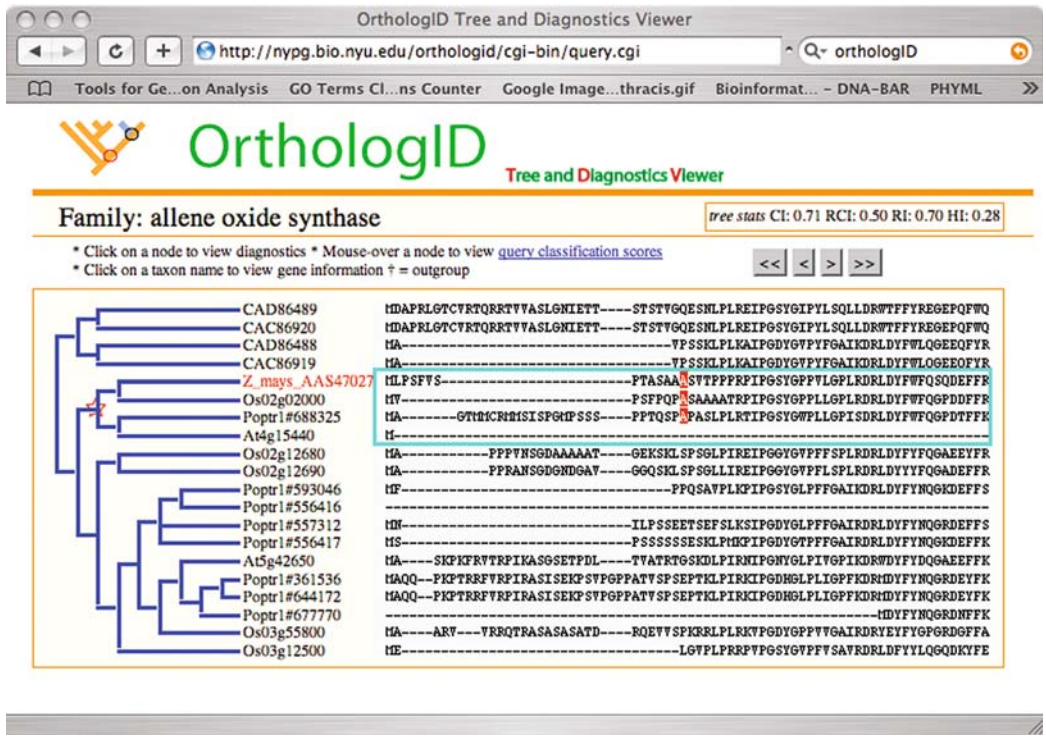

Fig. 2.8. Screenshot of an examination of the placement of the query sequence from **Fig. 2.7**. The box shows the statistics for diagnosing the query sequence to its position in the tree. The diagnostics used to place the query sequence are indicated in the *box* and highlighted.

4. Notes

1. Distances and orthology. Some authors have argued that distance measures are adequate indicators of homology or of significant evolutionary or biological relationships. On the surface this assumption makes some sense. When comparing several protein sequences it is assumed that the two with the greatest similarity are each other's closest relatives and hence, if they originate in different species, they are assumed to be orthologs. Eisen *(11)* and Thornton and DeSalle *(12)* have discussed the ins and outs of distance-based homology studies. They cite four assumptions and problems of this approach: (1) divergence rates are identical in all lineages. This assumption is often not the case in gene family evolution; (2) the pairwise similarity scores include not only phylogenetically informative synapomorphies but also shared ancestral characters (symplesiomorphies) and unique derived ones (autapomorphies); (3) multiple hits problems often violate the additivity of distances; and (4) accurate estimates of divergence are needed.

2. Phylogenetics and orthology assessment via sequence alignment and *OrthologID*. DePinna *(13)* suggested that homology assessment consists of two steps. The first requires that a hypothesis of homology be posed based on some measure of sameness of entities he termed primary homology. This concept should sound familiar and often times this primary homology is established by similarity measures or assessing similarity of two entities, be they genes or organisms. The hypothesis of homology or primary homology statement is then tested using phylogenetic analysis. If the primary homology statement is not rejected, it becomes a secondary homology statement. In this case, the secondary homology statement is a shared derived character or a synapomorphy. Brower and Schawaroch *(4)* refined this homology scheme by adding a step. They begin the assessment of homology with an initial step they call topographic similarity assessment. Once topographic similarity is established, a step involving establishing character state identity follows. The final step is the testing of the hypotheses established in the first two steps. Brower and Schawaroch *(4)* introduced their scheme because they realized a basic difference between molecular and morphological approaches to systematics.
3. The importance of E-values. In addition to information about functionally annotated genes and structure, one of the most commonly used tools for identifying genes is the use of an E-value cutoff. The choice of E-value is somewhat arbitrary though and some studies have noticed a "big genome attraction" effect. This phenomenon results from the idea that organisms with large genomes will include remnants of a large number of gene families. Some researchers have resorted to a process called "conditioning" in order to overcome this problem. In recent attempts to explore "E-value space," researchers examine phylogenies constructed using matrices assembled using a range of E-value cutoffs. Lienau et al. *(14)* examined the impact of E-value choice on the gene presence absence tree of life. When very stringent E-values (very small E-values) are used as a cutoff, the trees become less resolved, but what resolution exists agrees well with what we know about organismal history. The reason for the lack of resolution is that when very small E-values (such as e-300) are used, most of the information about presence/absence of genes is eliminated from a matrix *(14)*.
4. Comparison of *OrthologID* with PhiG. The PhiG approach is a straightforward application of Fitch's *(15)* original solution to the orthology paralogy problem. The approach involves five steps: "(1) an all against all BLASTP of the complete proteomes; (2) global alignment and distance calculation of the gene pairs identified by BLAST; (3) iterative, hierarchical clustering; (4) multiple sequence alignment (MSA) creation

and editing; and (5) gene tree reconstruction" *(7)*. Note that these steps are also very similar to the steps outlined in **Table 2.1**, with steps 1, 2, and 3 approximating the topographical similarity step in the Gfh approach, step 4 coinciding with character state assessment, and step 5 coinciding with the last step of all three approaches discussed in **Table 2.1**. The PhiG website adds an interesting aspect to gene family identification by putting the orthology statements derived from their five-step procedure into a chromosomal location context. In addition to this aspect of PhiG that relates to fully sequenced genomes, the package also uses a Hidden Markov Model (HMM) approach to facilitate the classification of putative proteins from less densely sampled genomes such as those generated by the EST approach.

5. The CAOS algorithm *(3, 10)*. The CAOS approach is based on population aggregation analysis as articulated by Davis and Nixon *(16)* and is used to discover the diagnostics in the *OrthologID* approach. In essence the guide tree allows for a phylogenetic grouping of protein sequences that can be used by the CAOS algorithm to find diagnostics. Diagnostics can be "single pure," "compound private pure," and even "compound nonprivate (polymorphic) pure" (see OrthologID website – http://nypg.bio.nyu.edu/orthologid/diagnostics.html). The compound classes of diagnostics take advantage of combining columns that in and of themselves are not diagnostic. The compound private pure diagnostics take advantage of private character states in particular ortholog groups and the compound polymorphic pure diagnostics utilize complex combinations of character states in positions to discover diagnostics in information that at first seems simply polymorphic.
6. Interpreting *OrthologID* scores. **Figure 2.9** shows an example of how to interpret the two numbers that are given when scrolling over the query sequence node on the "query orthology classification" page. In **Fig. 2.9**, there are the two examples we used to explain these numbers. In both, the first number indicates the number of diagnostic characters shared between the query and the gene(s) belonging to the clade it is placed into. In both the examples shown, this number is 50. Note that the absolute number of diagnostic characters may not be a reliable indication of how decisively the query is being placed into a clade. The second number, represented as a percentage score, is a better indication of the strength of query classification. In example 1, query sequence *Q1* shares 50 characters ($a = 50$) with the gene(s) in clade A, and zero characters ($b = 0$) with the gene(s) in clade B. As a result, *OrthologID*places *Q1* into clade A with a percentage score of

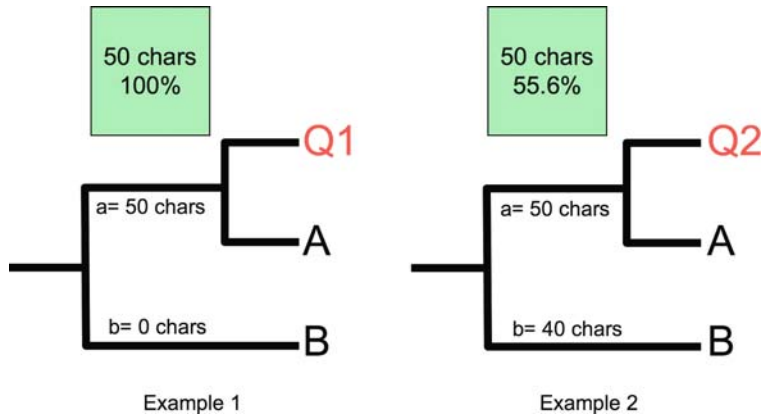

Fig. 2.9. An example of how to interpret the two numbers that are given when scrolling over the query sequence node on the "query orthology classification" page. The first number in both examples indicates the number of diagnostic characters shared between the query and the gene(s) belonging to the clade it is placed into. The second number, represented as a percentage score, is the better indication of the strength of query classification. In example 1, query sequence *Q1* shares 50 characters ($a = 50$) with the gene(s) in clade A, and zero characters ($b = 0$) with the gene(s) in clade B. As a result, *OrthologID* places *Q1* into clade A with a percentage score of 100%. The percentage score is calculated using the formula $(a/(a+b)) \times 100\%$. In example 2, $a = 50$ and $b = 40$; as a result, the percentage score is lower (55.6%), indicating that the strength of query placement for *Q2* is weaker than that of *Q1* in example 1.

100%. The percentage score is calculated using the formula $(a/(a+b)) \times 100\%$. In example 2, $a = 50$ and $b = 40$; as a result, the percentage score is lower (55.6%), indicating that the strength of query placement for *Q2* is weaker than that of *Q1* in example 1.

References

1. Chiu, J. C., Lee, E. K., Egan, M. G., Sarkar, I. N., Coruzzi, G. M., and DeSalle, R. (2006) *OrthologID*: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* **22**, 699–707.
2. Koski, L. B., and Golding, G. B. (2001) The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* **52**, 540–42.
3. Sarkar, I. N., Thornton, J. W., Planet, P. J., Figurski, D. H., Schierwater, B., and DeSalle, R. (2002) An automated phylogenetic key for classifying homeoboxes. *Mol Phylogenet Evol* **24**, 388–99.
4. Brower, A. V. Z., and Schawaroch, V. (1996) Three steps of homology assessment. *Cladistics* **12**, 265–72.
5. Altschul, S. F., Gish, W., Miller, W., Meyers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403–10.
6. Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33–6.
7. Dehal, P. S., and Boore, J. L. (2006) A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) Database. *BMC Bioinformatics* **7**, 201.
8. Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511–18.

9. Swofford, D. L. (2003) *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer Associates, Sunderland, Massachusetts.
10. Sarkar, I. N., Planet, P. J., Bael, T. E., Stanley, S. E., Siddall, M., DeSalle, R., and Figurski, D. H. (2002) Characteristic attributes in cancer microarrays. *J Biomed Inform* Apr/May; **35**(2), 111–22
11. Eisen, J. A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* **8**, 163–67.
12. Thornton, J. W., and DeSalle, R. (2000) Phylogenetics meets genomics: homology and evolution in gene families. *Annu Rev Genomics Hum Gene* **1**, 43–72.
13. DePinna, M. C. C. (1991) Concepts and tests of homology in the cladistic paradigm. *Cladistics* **7**, 367–94.
14. Lienau, E. K., DeSalle, R., Rosenfeld, J. A., and Planet, P. J. (2006) Reciprocal illumination in the gene content tree of life. *Syst Biol* **55**, 441–53.
15. Fitch, W. M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool* **19**, 99–113.
16. Davis, J. J., and Nixon, K.C. (1992) Populations, genetic variation and the delimitation of phylogenetic species. *Syst Biol* **41**, 121–35.