Using Multiple Sources of Orthology Information to Improve Sensitivity and Specificity
John Calley, Xueqian Gong, and Cara Ruble

Different sources of orthology information differ in their intended uses (phylogenetic reconstruction or functional prediction), scope of coverage (different organisms), level of organization (gene or protein), name space (NCBI, Ensembl, etc.), and in their methods of compilation. In practice, scientists often choose a single resource for annotation of comparative genomic data, resulting in an incomplete or non-specific data analysis package. Many orthology resources are underutilized due to practical difficulties involved in using them, particularly differences in the names used versus the names needed (I.E., resource is expressed in Ensembl protein IDs, but EntrezGene names are needed). Additionally, there is much to be gained by comparing predictions derived from different resources, but this can be difficult to perform in a manner that provides more meaningful data than that obtained by using orthology resources individually. These issues are particularly problematic for analyzing high-throughput data requiring automated decisions based on accurate orthology information.

We present a process for mapping multiple orthology resources into one common form, taking into consideration the information lost in this process and presenting a set of tools that use this common form to compare and combine resource content. Use of multiple resources allows us to increase sensitivity by covering genes that do not exist in all resources and to increase specificity by noting areas of agreement between different resources.