Poster C-40

Evaluating Orthology Detection Approaches in the Absence of a Gold Standard

Authors:
Feng Chen (*Biological Chemistry Graduate Program, University of Pennsylvania*)
Aaron J Mackey (*Department of Biology, and Genomics Institute, University of Pennsylvania*)
Jeroen K Vermunt (*Department of Methodology and Statistics, Tilburg University, Netherlands*)
David S Roos (*Department of Biology, and Genomics Institute, University of Pennsylvania*)

Short Abstract: Latent Class Analysis permits performance evaluation for multiple tests based on observed patterns of agreement. Applying this statistical methodology to various orthology identification methods reveals that INPARANOID, Orthostrapper, and OrthoMCL exhibit the best balance between sensitivity and specificity. OrthoMCL offers additional advantages of speed and applicability to multi-species datasets.

Long Abstract:

The rapid growth of genome sequence data, from an ever-increasing range of relatively obscure species, places a premium on the automated identification of orthologs to facilitate functional annotation, comparative genomics and evolutionary studies. The problem is particularly acute for eukaryotic genomes, because of their large size, the difficulty of defining accurate gene models, the complexity of protein domain architecture, and rampant gene duplications. Several strategies have been employed to distinguish probable orthologs from paralogs: phylogeny-based approaches include RIO (Resampled Inference of Orthology) and Orthostrapper/HOPS (Hierarchical grouping of Orthologous and Paralogous Sequences); strategies based on evolutionary distance metrics include RSD (Reciprocal Smallest Distance); BLAST-based approaches include Reciprocal Best Hits (RBH), COG/KOG (Cluster of Orthologous Groups), and INPARANOID. We have previously described the OrthoMCL algorithm, which improves on RBH by recognizing many-to-many co-ortholog relationships, using a normalization step to correct for systematic biases when comparing specific pairs of genomes, and using a Markov clustering algorithm to define ortholog groups. The OrthoMCL algorithm is fully automatable, requiring no manual curation.

Despite the large number of orthology identification approaches now available, no comprehensive comparison has yet been reported, in part because the lack of a genomic-scale error-free 'gold standard' data set makes it difficult to analyze performance. Functional genomics data are sometimes used as a surrogate for true orthology in performance assessments, but such data are known to include many false positives (FP) and false negatives (FN), and are difficult to apply across large evolutionary distances. Latent Class Analysis (LCA) is a statistical approach that has been widely employed for the analysis of multivariate categorical data in clinical diagnostics, marketing research, sociology, and other areas. LCA uses the agreement or disagreement data between methods to infer FP and FN rates, permitting quantitative comparisons in the absence of a gold standard dataset. Many biological questions have been addressed by multiple methods yielding binary (yes/no) outcomes but no clear definition of truth, making LCA an attractive method for applications in

Computational Biology. We have employed LCA to evaluate orthology identification methods, including all of the above algorithms, in addition to some homology detection algorithms, such as BLAST and TribeMCL. Whether a given pair of cross-species proteins is or is not orthologous, is an unobserved or 'latent' class, and all of the algorithms under consideration are used to make yes/no predictions as to orthology. Given the prediction results for a large set of homologous protein pairs, the likelihood function for this model can be expressed using the overall orthology probability and the FP and FN error rates for each algorithm. A maximum likelihood estimate of these model parameters is then used to represent benchmarking result. For this analysis, LCA was used to examine homologous protein pairs among 18,202 protein sequences from six complete eukaryotic genomes (Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Saccharomyces cerevisiae, and Schizosaccharomyces pombe), representing 1092 Pfam protein families.

From this analysis, most methods can be seen to balance sensitivity versus specificity. For example, homology detection methods display FP rates of 53-59% and FN rates of ~5%, while orthology detection methods based on phylogeny or evolutionary distance exhibit FP of 1-4% and FN of 21-67%. Between these two extremes are the diverse performances of BLAST-based orthology identification methods. As the first step for many BLAST-based approaches, RBH displays a low FP (9%), but its inability to recognize many-to-many co-ortholog relationships results in a high FN (33%). The consideration of co-orthologs by INPARANOID reduces FN to 21%, and FN rates are further decreased in methods that consider multiple genomes: 11% for OrthoMCL, and 4% for KOG. Ortholog clustering across multiple genomes inevitably bears a cost of increased FP rates, however: 18% for OrthoMCL, and 37% for KOG. Three algorithms exhibit both FP and FN error rate <25%, and can therefore be considered the best orthology identification methods: Orthostrapper, INPARANOID and OrthoMCL. In further studies, we have investigated the performance and optimization of these methods under different parameters, varying the orthology bootstrapping cutoff in phylogeny based approaches, the E-value cutoff in BLAST based approaches, and the MCL inflation value in Markov clustering algorithms (OrthoMCL, TribeMCL).

To compare OrthoMCL with the widely-used KOG algorithm, the stand-alone version of OrthoMCL was applied to the KOG reference dataset, including all proteins from seven eukaryotic genomes (the six species noted above, plus Encephalitozoon cuniculi). More than 50% of KOG groups were identically grouped by OrthoMCL, indicating that automated application of this method performs comparably to the manually-curated KOG database. ~35% of KOG groups were split into smaller groups by OrthoMCL, but comparison with Enzyme Commission (EC) annotations and protein domain architecture suggests that OrthoMCL groups exhibit a higher degree of consistency in the identification of putative protein function. Similar trends were observed in the application of OrthoMCL to the prokaryotic COG dataset. To facilitate applications of OrthoMCL's orthology prediction capabilities, proteome data from 55 complete genomes (all available eukaryotes, plus a representative selection of eubacteria and archaebacteria) was clustered into ortholog groups, and this data may be queried and downloaded from http://orthomcl.cbil.upenn.edu.