Title:
Blasting through Proteins

Authors.
Victoria Akintola[1] , and Jonathan E. Mohl [1,2]
[1]Bioinformatics Program, [2]Department of Mathematical Sciences, The University of Texas at El Paso.

Abstract.
Annotation of protein-coding genes is fundamental to interpreting genome sequences. However, these annotations reflect the biological knowledge and database content available at the time they are generated. As protein sequences from NCBI databases expand and previously uncharacterized proteins acquire experimentally or computationally supported functions, existing annotations may become incomplete or outdated. This work aims to re-evaluate previously annotated protein sequences in Integrated Omnicsbox from locally assembled genomes to determine whether new or more specific functional assignments are provided using bioinformatics resources. There are about one million potential protein sequences across 26 genomes (rotifers and ducks) and one transcriptome to re-analyze. Using the functional annotation package of OmicsBox, protein sequences were searched against an updated protein sequence database using BLASTP to identify homologous proteins and potential functional relationships. Significant matches were evaluated based on sequence identity, alignment coverage, e-value and functional description. InterProScan was performed subsequently through OmicsBox to identify protein families. The result from BLAST and InterProScan was integrated to update the gene ontology terms and assign Enzyme Commission numbers which provide a standardized classification of enzyme functions based on the biochemical reactions they catalyze, enabling accurate protein annotation and metabolic pathway identification. The updated functional information was compared with the original annotation files to classify proteins as having unchanged, refined, newly assigned or with unknown functions. This will help to assess the value of how frequently protein annotations should be revisited as biological databases evolve.