**Sequence to Hierarchical Annotation Mapping**

The Sequence-to-Metacyc and Sequence-to-GO mappings follow similar steps, but the code base is not the same because of slight differences between the Metacyc and GO representations and their depths of redundancy. In both cases, the mapping of reads to Metacyc or GO can be made to various depths of granularity, depending on the investigator's depths of interest. Sequences can be assigned to multiple entities in both mappings, even at the same depth of classification. These entities have biologically meaningful classifications through their relationships that are tied together with parent/child relationships, interpreted as "is a". For example, "Thiamine-Biosynthesis" is a "Vitamin-Biosynthesis" pathway. The set of entities or categories and their parent/child relationship is the ontology.

Prior to the hierarchical annotation processing, reads or sequences from an assembly are passed through JCVI's annotation pipeline. Some of these sequences may be associated with Enzyme Commission (EC) IDs. These ECs are associated with reactions, and the reactions are associated with pathways. These pathways are hierarchically classified. For example, the "Degradation" entity may be divided into "Protein-Degradation", "Aldehyde-Degradation", "Carbohydrates-Degradation", etc… Then, "Carbohydrates-Degradation", may be broken down into "Polysaccharide-deg" and "Sugars-and-Polysaccharides". Since a particular sequence may fall into several or no subcategories, the counts for a subcategory will frequently not sum up and match the count for the parent category. However, no count for any individual child category should ever exceed the count for the parent category.

The database from which the ECs are associated with pathway is through MetaCyc. The database flatfiles need to be updated periodically. The key steps for processing the read-to-pathway associations is: 1.) extract EC to pathway mappings from MetaCyc database files. 2.) For each read that has been annotated with an EC number, associate the EC with a pathway ID. 3.) For each pathway a read is associated with, identify the parent category that the pathway is associated with, all the way up to the root. 4.) Finally, for each read, identify all the pathways that are associated with it, and generate a non-redundant set of pathways to assign counts to. This non-redundification of assignments is necessary to preserve the 1-count-1-read tally. Thus, a sequence can be associated with multiple children that have the same parents, yet have the sequence will not be multiply counted by the parent.

The output is a tree, as well as a table of counts. The tree is easier to peruse, since it allows the observer to see the richness of sub-classifications under each category. However, to compare the counts of two samples, the counts in a table format may be more useful. To generate table counts, a starting point category and depth are expected to be specified. This allows the user to identify which ontology of interest to focus on, and as well as how the deeply the classifications should be of concern.

The mapping of read-to-GO ontologies follows a similar process. The GO mappings are more numerous because the identified proteins are not limited to those associated with catalyzing reactions. The GO mappings are associated with the sequences through the Uniref100 assignments that are also assigned through JCVI's annotation pipeline. The GO mappings need to be processed differently because, unlike the MetaCyc ontology, a child category in GO, can be associated with multiple parents. Nonetheless, the final result can be assessed through a tree of counts and tables of counts, when the GO category and depth are specified. The GO ontology described at www.geneontology.org.