What is the pig's order? Dealing with the ragged hierarchy of NCBI Taxonomy

Testsu Sakamoto[1], Lab Biodados[2],

1 Universidade Federal de Minas Gerais. Laboratório de Biodados.
2 Universidade Federal de Minas Gerais

Abstract

Any biological data are tightly linked to taxonomy and several bioinformatics tools use this information to answer biological questions regarding the species classification, diversity and evolution. One important taxonomic source is the NCBI Taxonomy. In this database, all sequence accessions deposited on INSDC are associated with their respective species which they belong; and the catalogued species are organized in a hierarchical structure that attempts to illustrate their evolutionary relationship. The hierarchical structure of NCBI Taxonomy is useful in retrieving the taxonomic lineage and classification since it uses taxonomic nomenclature and rank (class, order, family etc.) to name and classify the nodes comprising it. Though several bioinformatics analyses work with this taxonomic source, many of them face some difficulties because its hierarchy is of ragged type. As result, taxonomic lineages in the hierarchy are comprised of different number of nodes and some taxonomic ranks could be missing in some lineages. For instance, the taxonomic lineage of Sus scrofa (pig; taxid: 9823) does not have a node of subphylum, superclass, subclass or order ranks. Furthermore, the existence of nodes without a rank assigned (referred as "no rank") also contributes on generating ragged hierarchy. To address this issue, we developed an algorithm that takes the tree structure from NCBI Taxonomy and generates a balanced taxonomic tree. To create a balanced hierarchy, the algorithm firstly attempts to assign a taxonomic rank to a "no rank" nodes. Then, the algorithm creates/deletes nodes throughout the tree making it balanced. The algorithm also creates a name for the new nodes by borrowing the names from its ranked child or, if there is no child, from its ranked parent node. The new hierarchical structure was named Taxallnomy and it contains 29 hierarchical levels correspondent to the 29 taxonomic ranks used in the NCBI Taxonomy database. From Taxallnomy, user can obtain the complete taxonomic lineage with 29 nodes of all taxa available in the NCBI Taxonomy database. Taxallnomy is applicable to several bioinformatics analyses that depend on the data from NCBI Taxonomy. In this work, we demonstrated its applicability by embedding taxonomic information of a specified rank to a phylogenetic tree; and by making metagenomic profile according to a rank. The algorithm was written in PERL and all resource for Taxallnomy database can be accessed at biodados.icb.ufmg.br/taxallnomy.

Funding: FAPEMIG, CAPES