Current Challenges in Optimizing the Capture of all Publications Pertaining to a Protein Family and its Members

Colbie J. Reed, Rémi Denise, Geoffrey Hutinet, and Valérie de Crécy-Lagard
Microbiology & Cell Science Department, The University of Florida

Although conceptually simple, one of the major challenges remaining in the post-genomic era is the capture of all published experimental data available for members of a protein family. To date, two main methods are used to search the scientific corpus for works pertaining to a given sequence and its homologs: 1) text-based search using common family/homolog identifiers; 2) sequence- based search using a member sequence to query a single or several databases to identify existing links between published data and matching homologs. Here, select tools commonly used to capture the literature on a protein family are reviewed and current challenges discussed. Using DUF34 as an example protein family, a defined list of keywords and a subset of member sequences were determined for use in systematic text- and sequence-based queries, respectively. Tools and their performances were then evaluated and compared, revealing major differences in the number and quality of the identified publications. In addition to highlighting tool vulnerabilities as to better inform the common practices of researchers and their use of such resources, these results provided insights important for addressing key challenges of sequence-to-publication crosslinking, biological entity identification, and data publishing standards.