Poster M-1

CUPID: Core and Unique Protein IDentification system

Authors:
Raja Mazumder (*Protein Information Resource, Georgetown University*)
Darren A. Natale (*Protein Information Resource, Georgetown University*)
Sudhir Murthy (*DCWASA- DWT*)
Rathi Thiagarajan (*Protein Information Resource, Georgetown University*)
Cathy H. Wu (*Protein Information Resource, Georgetown University*)

Short Abstract: A pipeline using a combination of computational and manual analyses of BLAST results was developed to identify strain-, species-, and genus-specific proteins and to catalog the closest sequenced relative for each protein in a proteome. Such proteins can not only serve as taxon-specific diagnostic targets but also provide insight into the criteria that define an organism.

Long Abstract:
A pipeline using a combination of computational and manual analyses of BLAST results was developed to identify strain-, species-, and genus-specific proteins and to catalog the closest sequenced relative for each protein in a proteome. Proteins encoded by a given strain were preliminarily considered to be unique if BLAST, using a comprehensive protein database, failed to retrieve any protein not encoded by the query strain, species or genus (for strain-, species- and genus-specific proteins respectively), or if BLAST, using the best hit as the query (reverse BLAST), did not retrieve the initial query protein. Results were manually inspected for homology if the initial query was retrieved in the reverse BLAST but was not the best hit. Sequences unlikely to retrieve homologs using the default BLOSUM62 matrix (usually short sequences) were re-tested using the PAM30 matrix, thereby increasing the number of retrieved homologs and increasing the stringency of the search for unique proteins. The above protocol was used to examine several food- and water-borne bacterial pathogens. We found that the reverse BLAST step filters out about 22% of proteins with homologs that would otherwise be considered unique at the genus and species levels. Analysis of the annotations of unique proteins revealed that many are remnants of prophage proteins, or may be involved in virulence. The data generated from this study can be accessed and further evaluated from the CUPID (Core and Unique Protein Identification) system web site (updated semi-annually) at http://pir.georgetown.edu/cupid.

Over 300 pathogenic and non-pathogenic bacteria have been completely sequenced, including multiple strains from several species. The availability of sequence data from related genomes has facilitated comparative genomic analysis, which not only allows the study of major evolutionary processes, but also the determination of proteins conserved across—or unique to—different species. We have developed a general protocol to identify proteins unique to different taxa, and applied it to a set of food- and water-borne pathogens. Specifically, we: a) identify proteins that are unique to a particular strain, species, or genus; b) extract the set of proteins common to two or more strains or species; and c) determine the organism most closely related to a particular genus, species or strain.

It is evident from this study that a major reason that few unique proteins are found in some cases is the presence of sequence data for closely-related organisms and, by extension, the peculiarities of taxonomic designations. Therefore, only a general trend in the number of strain-, species-, or genus-specific proteins can be established. Strains with closely related sequenced strains tend to have relatively few unique proteins at that level while the converse is true for those without close relatives (compare Helicobacter pylori strains with Helicobacter hepaticus). The trend also holds true at the genus level (compare Escherichia or Helicobacter genus-specific proteins with Salmonella). We note that a proteome from a closely-related genus is represented in the protein database for both Escherichia (Shigella) and Helicobacter (Campylobacter). Shigella is so similar to E. coli that there are recommendations to consider them different species within the same genus, while Helicobacter pylori was once Campylobacter pylori.

Despite the conservative approach used here, one must be mindful of certain pitfalls in deriving lists of unique proteins. First, a protein might be labeled as unique only because homologs from other organisms were missed upon submission of the sequence, or because of some other conceptual translation problem. In all cases, the short list of potential unique proteins should be further screened computationally at the DNA level using tBLASTn. Second, many of the proteins identified here are remnants of prophage proteins. At the strain level, labeling of such proteins as unique may be more a reflection of a gap in whole-genome sequence information than of true specificity. Accordingly, discrimination between individual strains (isolates) may require laboratory comparison methods such as pulsed-field gel electrophoresis or whole-cell fatty acid analysis. In contrast, identifying species- or genus-specific proteins can be done with confidence when multiple representatives have been sequenced. In such cases, conservation within multiple strains of a species (for example) gives confidence in the “reality” of the uniqueness because that status has been conserved over time (core unique proteins: proteins that have related sequences in all selected organisms, but not in non-selected organisms).

Precise identification of pathogens is important so that adequate action can be taken to either eliminate or reduce the threat of infection. One use of the CUPID system is to help identify diagnostic targets specific to a particular clade of these pathogens. The unique proteins form a short list of diagnostic targets to be validated in the laboratory. Proteins predicted to be external to the bacterial cell – possibly involved in host interactions and virulence – may be used to develop protein-based detection systems. In addition, it should be possible to use the DNA encoding these proteins as the basis for diagnostics. In conclusion, the salient features of CUPID are: a) provides sets of proteins unique to a strain, species, and genus level; b) includes a check for additional homologs based on reciprocal hits; c) uses different parameters for short sequences; d) provides the identity of the nearest non-self neighbor; and e) allows retrieval of unique, core, and core unique proteins at different taxonomic levels.