CONSERVED HYPOTHETICAL GENES AND UNASSIGNED ENZYMES: HOW MUCH DO WE NOT UNDERSTAND

Michael Y. Galperin

NCBI, NLM, National Institutes of Health, Bethesda, Maryland, USA

The availability of complete genome sequences continues to change modern biology. With complete genomes, it is becoming possible to catalogue all proteins that are responsible for every essential cellular function (create a "genomic parts list"). This allows us to take an unbiased look at the live cell and evaluate how much do we really understand in its behavior and metabolism, signal transduction, cell division, and other processes. However, only a small fraction of proteins encoded in any given genome has ever been studied experimentally or will be studied in any detail any time soon. Therefore, analysis of complete genomes has to deal with the constantly growing numbers of "hypothetical" protein whose functions remain unknown. On the other hand, there are numerous enzymatic activities that have not been assigned to any protein sequences. This talk will discuss using comparative genomics to improve our understanding of metabolic and signaling pathways in microorganisms, including identification of "missing" enzymes and prediction of alternative enzyme variants from the lists of previously uncharacterized gene products. I will introduce our classification of "conserved hypothetical" proteins into "known unknowns" and "unknown unknowns", present several examples of the most interesting conserved hypothetical proteins, and discuss the approaches to their characterization. I will also try to show how comparative genomics can be used to prioritize future research in the least understood areas of molecular cell biology.