Biochem Anal Biochem 2017, Volume 7
DOI: 10.4172/2161-1009-C1-018

A shortcut from large-scale genomic data to protein structure and function

Michal Linial
Hebrew University of Jerusalem, Israel

Statement of the Problem: With 90 million proteins in the public databases and an unprecedented growth, only robust unsupervised and automated methods can realistically achieve the comprehensive functional annotations of the protein space. Importantly, in recent years, the accumulation of new sequences is mostly driven by deep sequencing efforts and a high-quality assembly phase. To address the pressing challenge of accurate annotation, we offer a classification resource that is created by an unsupervised analysis of full-length protein sequences. The charting of the protein space to families is model free with a quality that matches the state-of-the-art expert systems.

Methodology & Theoretical Basis: Machine learning (ML) methods are becoming fundamental in annotating unknown proteins and their heterogeneous properties. The quality of automatic charting the protein space is assessed by the structural conservation measures which provide an unavoidable bridge to function. The entire protein space is created using hierarchical clustering. We address the basis for detecting overlooked connections between families; the division of families to subfamilies and the requirement for a reliable inference.

Findings: The ProtoNet platform that includes analytical tools to explore the protein space will be presented included its version for genome scale analyses. An innovative structural biology approach focuses on determining residue-level protein properties, such as sites of post-translational modifications (PTMs). We formulated it in ML method for a residue-level property called ASAP. It extracts numerous features from raw sequences and supports easy integration of external features such as secondary structure, solvent accessibility, intrinsically disorder or PSSM profiles. Features are then used to train ML classifiers. ASAP can create new classifiers within minutes for a variety of tasks, including PTM prediction.

Conclusion & Significance: We present a set of methodologies for bridging protein function at different levels from single residue to a full proteins and genome. These tools and platforms match state-of-the-art results and shed light on unexplored evolutionary processed that cross all domains of life. We conclude that despite minimal sequence similarity fingerprints of evolutionary processes remain powerful for analyzing new proteomes at a genomic scale, for protein design tasks, mass spectrometry search engines and the discovery of new bioactive proteins.

michall@cc.huji.ac.il