Navigating the twilight zone: Exploring overlooked functions among cell modulators

Michal Linial, Dept of Biological Chemistry, The Sudarsky Center for Computational Biology, the Hebrew University of Jerusalem, Jerusalem 91904, Israel

Mapping of remote evolutionary links is a classic computational problem of much interest. Relating protein families allows for functional and structural inference on uncharacterized families. However, since many sequences have diverged beyond reliable alignment, these are too remote to identify by conventional methods. This situation can refer to as the twilight zone, a term that captures the difficulty of extracting the hidden evolutionary signal in the protein sequence space. The bright side of it comes from the enormous growth of genome sequences in recent years. This flood of information can be tamed to improve the evolutionary signal and to seek for hidden, already faded connections between protein families.

I will present a method to systematically identify such remote evolutionary relations between protein families, leveraging a novel evolutionary tree of all protein sequences and families. The tree is a hierarchical tree that includes millions of protein sequences [1]. It allows tracing even very faint links, owing to the robustness of considering the entire volume of pairwise sequence similarities at construction. I will present the power and limitation of navigating through such a tree and a method that systematically scans the tree for evolutionary breakpoint in putative ancient superfamilies [2,3].

I will then discuss the limitation of knowledge extractions tools that any experimentalist may face following a large-scale proteomics experiments. I will illustrate the benefit of using integrative tools and resources in exposing overlooked connections. Specifically, will illustrate how one can improve the ability to search for fast evolving sequences as in the case of animal toxins. The focus of the discovery tool will be towards toxins, viral proteins and cell modulators.

The research is supported by the SCCB, Sudarsky Center for Computational Biology and grants from the ISF and EU Prospects.

Background information:

