

Mining Phenotypes for Protein Function Prediction

Ulf Leser

Humboldt-Universität zu Berlin

Joint work with Philip Groth, Bertram Weiss, and Hans-Dieter Pohlentz

Until very recently, phenotypes only very rarely were studied in a systematic manner. While ontologies for describing gene functions now have a 10 year long tradition, similar vocabularies for describing the phenotype of genes are only emerging now; similarly, the techniques for determining phenotypes on a large scale (especially RNAi) are available only for a few years, while genomic sequencing or gene expression studies are already established for a much longer time.

In this talk, we describe results from a study for exploiting phenotype descriptions for protein function prediction. We used the data from PhenomicDB, a phenotype database integrated from several publicly available data sources. Due to the lack of standardization, phenotypes in PhenomicDB can only be viewed as text (short statements, abstracts, singular terms ...). We clustered these texts and analyzed the corresponding gene clusters in terms of their coherence in functional annotation and their interconnectedness by protein-protein-interactions. We also devised a method for using the close similarity in their phenotype descriptions to predict the function of proteins. We show that this method yields a very good precision at acceptable coverage.