Phenotype prediction

Even for genes whose molecular function and cellular roles are known, understanding their role in affecting a certain phenotype remains a challenge. Apart from the Mendelian single gene traits, a substantial portion of the phenotypes we observe in nature are an effect of complex interplays between numerous genes in addition to various environmental factors. Such 'complex traits' are hard to predict and the development of methods for uncovering genotype-phenotype relationships has been identified as one of the major post-genomic challenges [1] .

Comparative genomics has been proposed for uncovering such gene-trait relationships [1, 2]. This approach begins by constructing phenotypic profiles, which indicate which organism exhibits a particular phenotype – this is similar to the concept of phylogenetic profiles[3]. Then causal relationships between genes and traits can be deduced from the co-occurrence of genes and phenotypes across a large number of genomes. The underlying principle is that species sharing a phenotype are likely to utilize orthologous genes in the involved biological process. These ideas were applied to predict genes involved in well characterised traits such as hyperthermophily [4] and flagellar motility [5]. Several approaches have been developed for this comparative analysis. For example, Tamura et al.[6] proposed a rule based data mining algorithm to associate Clusters of Orthologues Groups of proteins (COGs) with phenotypes; Slonim et al. [7] proposed an information-theoretic approach to extract preferentially co-inherited clusters of genes having significant association with an observed phenotype. Paccanaro and Gerstein have recently developed a correlation-based method [8] that was able to discover genotype-phenotype associations combining phenotypic information from a biomedical informatics database, GIDEON, with the molecular information contained in Clusters of Orthologous Groups of proteins (COGs) [9].

1. Bork, P., et al., *Predicting function: from genes to genomes and back.* J Mol Biol, 1998. **283**(4): p. 707-25.
2. Huynen, M., T. Dandekar, and P. Bork, *Differential genome analysis applied to the species-specific features of Helicobacter pylori.* FEBS Lett, 1998. **426**(1): p. 1-5.
3. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4285-8.
4. Makarova, K.S., Y.I. Wolf, and E.V. Koonin, *Potential genomic determinants of hyperthermophily.* Trends Genet, 2003. **19**(4): p. 172-6.
5. Levesque, M., et al., *Trait-to-gene: a computational method for predicting the function of uncharacterized genes.* Curr Biol, 2003. **13**(2): p. 129-33.
6. Tamura, M. and P. D'Haeseleer, *Microbial genotype-phenotype mapping by class association rule mining.* Bioinformatics, 2008. **24**(13): p. 1523-9.
7. Slonim, N., O. Elemento, and S. Tavazoie, *Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks.* Mol Syst Biol, 2006. **2**: p. 2006 0005.
8. Goh, C.S., et al., *Integration of curated databases to identify genotype-phenotype associations.* BMC Genomics, 2006. **7**: p. 257.
9. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.