C-3-12

Correlation of annotation terms in heterogeneous databases and its application to Gene Set Enrichment Analysis (GSEA) and ontology construction

Katsuhiko Murakami[1] **Tadashi Imanishi**[1, 2]
k-murakami@aist.go.jp t.imanishi@aist.go.jp

[1] Molecular Profiling Research Center for Drug Discovery,
National Institute of Advanced Industrial Science and Technology,
2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] Tokai University School of Medicine, Tokai University,
143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan

Keywords: Gene Set Enrichment Analysis, Database, Annotation, Correlation, Ontology

Characterization of a given gene set, such as "gene set enrichment analysis (GSEA)" [1], has become an important task in omics era. In such analyses, annotations are used as if they were independent, despite that some annotations are correlated each other. To interpret complex multiple annotations, we comprehensively examined correlation among each annotation for human genes. We selected ten gene annotation (gene family, Gene Ontology, InterPro, KEGG pathway, protein-protein interaction, SCOP, SOSUI membrane protein prediction, OMIM, tissue specificity of gene expression, and subcellular localization) from the integrated human gene database, H-InvDB [2,3]. For all pairs of the terms, the correlations were evaluated using Fisher's exact (two-side) test with Bonferroni correction. As a result, we found 99,813 and 653 pairs with positive and negative correlation respectively. Many of the positive relationships were synonyms, such as "SCOP g.44.1.1 (RING finger)" and "IPR001841 (Zinc finger, RING-type)". We found other pairs with relevant but not apparent relationships, such as "GO:0006470, protein dephosphroylation" and "hsa04940: Type I diabetes mellitus". We also obtained negative relationships. Many of them seemed unlikely to co-occur in a gene and related to subcellular localization, such as extracellular and nuclear. Those information will help to refine predictive annotation, or perhaps include multifunctions of the genes. By integrating these annotation relationships together with other integrated databases, such as UniProt, we can re-evaluate complex annotations of GSEA results and produce a new summary report of the gene set, as well as a construction of an integrated ontology..

[1] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sc*i 102(43):15545-50, 2005

[2] Takeda J, Yamasaki C, Murakami K, Nagai Y, Sera M, Hara Y, Obi N, Habara T, Gojobori T, Imanishi T., H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. *Nucleic Acids Research* 41(Database issue): D915-9, 2013

[3] http://www.h-invitational.jp/