XI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 24-25-26 de outubro de 2013

ENRICHMENT ANALYSIS OF FUNCTIONAL TERMS IN DRUG TARGET GENES INVOLVED IN NEUROLOGICAL DISORDERS BASED ON THE HYPERGEOMETRIC TEST

P Cacheiro[1], MJ Sobrido[1,2,3], C Cadarso-Suárez[4], A Carracedo[1,2,3]

[1]Grupo de Medicina Xenómica, Universidade de Santiago de Compostela
[2]Fundación Pública Galega de Medicina Xenómica, Instituto de Investigaciones Sanitarias de Santiago de Compostela (IDIS)-SERGAS
[3]Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER)
[4]Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela

ABSTRACT

Enrichment analysis allows to study biological annotation terms overrepresented in a biomedical repository (e.g. a gene list). In this work we aimed at investigating functional characteristics of a set of genes involved in Mendelian neurological disorders, which are also validated therapeutic targets.Three different biological annotation resources were employed: the Gene Ontology, the KEGG pathways and the Pfam domains database. The test based on the hypergeometric distribution was used to assess statistical significance. After applying correction for multiple testing, enriched functional terms for the different categories were identified. The results obtained from this study might provide further insight into common attributes among successful neurological drug targets.

Keywords: Enrichment analysis, hypergeometric test, drug target, genetic disorders

1. INTRODUCTION

In the search for novel therapeutic molecules, knowledge extracted from successful drug targets is of great relevance. Information on functional aspects of genes and gene products can be gathered from different public repositories, including the Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Protein domain database Pfam.

The GO annotation database consists of three structured controlled vocabularies of terms describing biological processes (BP), cellular components (CC) and molecular functions (MF) associated to gene products. Information about canonical pathways is contained in the KEGG database, which comprises a set of networks of molecular interactions including metabolic and signaling routes and human diseases. The Pfam database contains information about conserved protein domain families based on protein regions of sequence similarity.

The aim of this study was to investigate functional aspects associated to drug target genes related to neurological phenotypes, by looking for functional categories overrepresented in a given gene list compared to a general background gene list.

2. METHODOLOGY

The genes to be tested were identified through the search for neurological records in OMIM (Online Mendelian Inheritance in Man), an encyclopedia of genes causing human diseases. The total collection of neurological genes obtained from OMIM was set as the reference list (N = 1067). A subset of that list - the test set - was comprised of the genes known to be drug targets according to DrugBank, a database of drugs and drug targets (N = 272).

This choice of reference set was made in order to avoid potential bias due to overrepresentation of terms linked to the nervous system.

Among the different statistical methods that can be used to estimate enrichment P-values, the hypergeometric test calculates the probability of enrichment of a given term in the target set against the reference set (or universe) by assuming that the list of elements are sampled from a hypergeometric distribution:

$$p(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

Where N are the number of genes in the reference set, n are the number of genes in the target set, M and k are the number of genes annotated to a specific term in the reference set and the target set, respectively.

Enrichment analysis of functional characteristics according to GO categories, KEGG pathways and Pfam protein families was performed by means of Bioconductor R package Category (version 2.26.0), which includes an implementation of the hypergeometric test. Given that several categories are tested at the same time, multiple testing correction was performed using the R package multtest. The Benjamini & Hochberg (1995) step-up FDR-controlling procedure was applied to calculated adjusted P-values.

3. RESULTS AND DISCUSSION

The total number of terms significantly enriched was of 27 for GO Molecular Function, 12 for GO Cellular Component, 29 for GO Biological Process, 20 for KEGG pathways and 2 for Pfam domains.

Results from the hypergeometric test regarding KEGG pathways and Pfam domains are shown in tables 1 and 2, respectively. Only results with adjusted P-value <0.05 are shown.

Table 1: Results from the enrichment analysis considering KEGG pathways

KEGG ID	KEGG Term	P-value (uncorrected)	Adjusted P-value (BH correction)
01100	Metabolic pathways	1.445e-10	2.615e-08
00330	Arginine and proline metabolism	5.829e-07	3.978e-05
00190	Oxidative phosphorylation	6.593e-07	3.978e-05
05012	Parkinson´s disease	2.160e-06	9.775e-05
05016	Huntington´s disease	1.498e-05	5.422e-04
05010	Alzheimer's disease	3.491e-05	1.053e-03
00640	Propanoate metabolism	9.139e-05	2.363e-03
00071	Fatty acid metabolism	1.813e-04	4.104e-03
00280	Valine, leucine and isoleucine degradation	2.456e-04	4.940e-03
00380	Tryptophan metabolism	5.105e-04	9.240e-03
00970	Aminoacyl-tRNA biosynthesis	1.283e-03	0.0172
00650	Butanoate metabolism	1.283e-03	0.0172
00260	Glycine, serine and threonine metabolism	1.283e-03	0.0172
00620	Pyruvate metabolism	1.329e-03	0.0172
05215	Prostate cancer	2.073e-03	0.0225
00250	Alanine, aspartate and glutamate metabolism	2.109e-03	0.0225
00020	Citrate cycle (TCA cycle)	2.109e-03	0.0225
04080	Neuroactive ligand-receptor interaction	2.352e-03	0.0237
04914	Progesterone-mediated oocyte maturation	3.030e-03	0.0289
00310	Lysine degradation	4.641e-03	0.0419

Table 2: Results from the enrichment analysis considering Pfam domains

PFAM ID	PFAM Term	P-value (uncorrected)	Adjusted P-value (BH correction)
PF07714	Protein tyrosine kinase	1.215e-06	4.362e-04
PF00001	Transmembrane receptor (rhodopsin family)	2.666e-04	0.0478

By using a simple statistical method and biological data available in public databases, we can obtain valuable information about a gene set of interest. In this particular case, for instance, we could identify two protein families (tyrosine kinases and rhodopsin-type receptors) overrepresented within the neurological drug targets compared to the whole set of neurological genes.

Special attention must be paid to the selection of the reference set, since the choice of alternative gene universes might have a potential impact on the results. Furthermore, due to the structure of the public databases used to retrieve the terms, the test results are likely to include closely related terms, regardless of multiple testing corrections. Another factor potentially introducing bias in the list of terms for comparison might be the fact that it is expected to have more abundant and more specific information in the databases on genes that are drug targets, as well as on diseases and pathways that have an available drug. Thus, the results should be interpreted cautiously.

ACKNOWLEDGMENTS

Funding: Proyecto InnoPharma. Ministerio de Economía y Competitividad. Fondos FEDER.

REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet,25,25-29.

Benjamini Y and Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B, 57, 289-300.

Gentleman R, Falcon S, Sarkar D (2013) Category: Category Analysis. R package version 2.26.0.

Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, 37, 1-13.

Kalathur RK, Hernández-Prieto MA, Futschik ME (2012) Huntington's disease and its therapeutic target genes: a global functional profile based on the HD Research Crossroads database. BMC Neurol, 28,12:47.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res,40 (Database issue), D109-D114 .

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.. Nucleic Acids Res, 39 (Database issue), D1035-1041.

Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), http://omim.org/.

Pollard KS, Gilbert HN, Ge Y, Taylorand S, Dudoit S (2013) multtest: Resampling-based multiple hypothesis testing. R package version 2.16.0.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res,40(Database issue), D290-301.