Biochemical features associated to cancer mutations

E. Porta[1], AM. Rojas[1], and I. Cases[1]

[1]Computational Cell Biology Group, Institute for Predictive and Personalized Medicine of Cancer, Badalona, Spain

The development and popularization of genome-wide technologies to perform experiments has provided the scientific community tools capable of taking a broad picture of the events happening in a cell. In this scenario, computational approaches that rely on computer-friendly annotations, such as controlled vocabularies or ontologies, to analyze the data have proven useful. The most widely used ontology to perform these studies is the Gene Ontology (GO). This approach has been particularly useful in the case of complex phenotypes, where focusing on single genes can be misleading since there are many genes involved. In these cases, analyzing the data from a broader perspective by aggregating the information using the ontology structure can be useful in order to extract new insights in the biology of the disease and learn about the differences in the genes underlying different phenotypes.

One particular disease that has benefited from these approaches is cancer, where several GO terms (such as "Apoptosis", "DNA repair", "Intracellular signaling" etc) have been identified to be over or under-represented in its associated genes in different enrichment analysis[1]. These associations have been used to make predictions on new cancer-related genes or to infer new functional annotations in known cancer-associated genes [1].

Enrichment analysis is not restricted to the Gene Ontology, however there have been few attempts to use this approach with other biological ontologies. We think that given the increasing interest in next-generation sequencing techniques and the exponential growth in the number of genomes and mutations coming from cancer samples it would be interesting to extend the enrichment analysis to this type of data because, not only different genes cause different diseases, also different mutations in the same gene can be associated to different phenotypes [2].

We have used the Sequence Ontology[3] (SO) and the Disease Ontology (DO) to perform an enrichment analysis in a dataset of human disease-associated missense mutations coming from Online Mendelian Inheritance in Man (OMIM), the Genetic Association Database (GAD) and the Catalogue Of Somatic Mutations In Cancer (COSMIC) to identify under or over-represented SO terms in cancer-related mutations.

Using this approach we have identified 3 SO terms enriched in cancer-associated mutations ("serine-rich regions", "aminoacid-biased regions" and "intrinsically unstructured regions") and 3 other SO terms depleted in cancer-related mutations ("disulphide bridges", "peptide localization signals" and "transmembrane regions") that highlight different relevant features of the disease.

1. Hu, P., Bader, G., Wigle, D. a & Emili, A. Computational prediction of cancer-gene function. *Nature reviews. Cancer* **7**, 23-34 (2007).
2. Zhong, Q. et al. Edgetic perturbation models of human inherited disorders. *Molecular systems biology* **5**, 321 (2009).
3. Eilbeck, K. et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* **6**, R44 (2005).