Power law correlation between gene connectivity and number of citations

Imparato, D.O.; Oliveira, R.A.; Andrade A.S.; Pasquali, M.A.; Dalmolin, R.J.

Departamento de Bioquímica, DBQ-UFRN, RN; Instituto de Medicina Tropical, IMT-UFRN, RN, Brasil

INTRODUCTION: An important part of systems biology relies on the usage of protein-protein interaction networks, although not much effort has been put into analyzing what role this information may play alongside different kinds of data. For instance, this lack of integration leaves room for possible predictions about how the depth of investigation put into a gene affects its number of annotated connections. **OBJECTIVES:** Explore a possible correlation between genes citations in biomedical literature and their number of interactions. **MATERIAL AND METHODS:** STRING-db PPI data was collected throughout versions 8.1 to 10. An R script was developed to automatically search for every PubMed reference of a gene and log its PubMed ID and publication date into an SQL database, given the unavailability of such retrieval. Text-mining evidence of protein links was removed, since it benefited greatly from PubMed itself and represented a biasing source. Confidence scores were recomputed and remaining data was filtered for medium confidence (score>=0.4). As ENSP IDs change over relatively short amounts of time, ID inconsistency across versions was taken into account and corrected for using ENSEMBL-Entrez ID mapping and protein synonyms data. **RESULTS AND DISCUSSION:** The obtained number of interactions and article references for each gene were plotted in a variety of ways using R. A power law correlation was observed when the number of interactions (k) was plotted against its mean number of references (c). Although fluctuations arise as k grows larger in response to lower sample sizes, a power regression line greatly described a large portion of the plot for every STRING-db version. **CONCLUSIONS:** The number of annotated gene interactions and its average number of references in biomedical literature follow a power law.

Keywords: systems biology, bioinformatics, ppi