How the abundance in the universe of components determines the statistics of the shared components

A. Mazzolini[1], E. De Lazzari[2], M. Gherardi[3], M. Cosentino Lagomarsino[4], M. Caselle[1], M. Osella[1]

[1]Physics Department and INFN, University of Turin, Italy

[2]Quantitative biology department, UPMC, Paris, France

[3]Physics Department, University of Milan, Italy

[4]Istituto FIRC di Oncologia Molecolare, Milan, Italy

Several complex systems of diverse nature consist of realizations which can be broken into their elementary constitutive components, for example, books into words, genomes into genes, and men-made systems into building blocks. The statistics of the components (e.g., words) across realizations (e.g., books) shows several quantitative laws, such and the well-known example of the power-law distribution of component abundances, known as Zipf's law in the context of in the context of natural languages [1].

Central to the current debate in evolutionary genomics is a different law [2], the "distribution of shared genes", or "occurrence distribution", where a component occurrence is defined as the fraction of realizations in which the component is present. In genomes, the occurrence distribution shows a peculiar U-shape due to a large number of rare (i.e. belonging to very few species) and common genes (present in almost all the species), compared to genes at intermediate occurrences. While several possible theoretical explanations of the U-shaped gene occurrence distribution have been proposed, its causes are still under debate [3,4].

Here, we consider occurrence distributions in three datasets from genomics, linguistics (literary texts), and technology (LEGO toy constructions), showing that it is characterized by a general power law decay, and a dataset-specific size of the common component peak.

By means of a theoretical null model based on multinomial sampling we show that the characteristic U-shape can emerge as a statistical consequence of the abundance distribution (the Zipf's law) with some crucial small deviations.

This similarity between the empirical occurrence distribution and the null one allows also us to establish an analytical relationship between the law and the abundance statistics, identifying the crucial parameters affecting the power law decay and the size of the common component peak.

Our results suggest that several features of the occurrence distribution can be predicted by the knowledge of the abundance statistics, and therefore its global shape it is not so informative by itself. However, the distribution shows also small deviations from the null predictions, which are in fact its most interesting features, providing information specifically contained in the occurrence statistics.

[l] G.K. Zipf, **Human behaviour and**, (1949).

[2] Touchon et al, PLoS Genet **5**, e1000344 (2009).

[3] Pang et al, PNAS **110**, 6235 (2013).

[4] Haegeman et al, BMC genomics **13**, 196 (2012).