Biome Distributions

Linas Vepstas

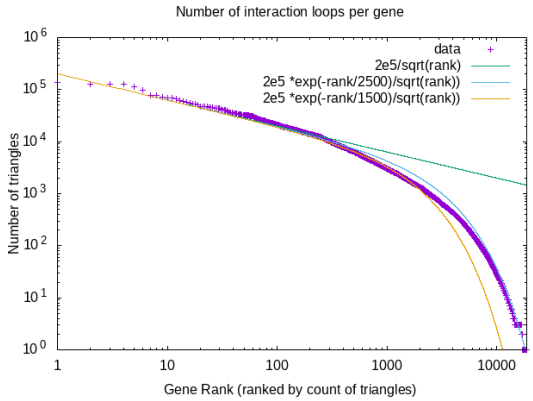
23 January 2020

Abstract

An exploration of the statistical distribution of interactions between genes, proteins and pathways, extracted from public genome, proteome and rectome datasets. This is a work-in-progress. It is incomplete.

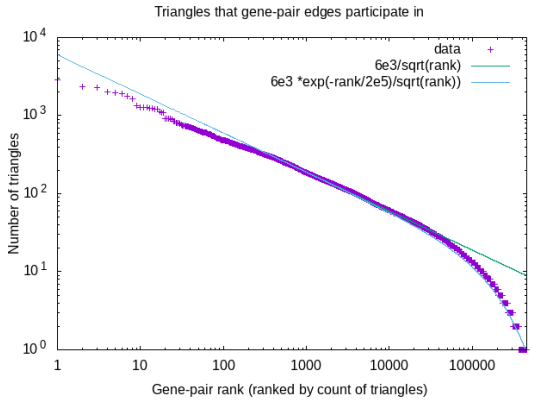
Zipf graphs

See the email chain titled "Genome distribution!?" and also the opencog benchmark "query-loop" for graphs of distributions as well as explanation of what this is. Here's a graph:



This shows 18766 genes in three-way triangular relationships. The parameter fits are just numerology.

There were 455572 distinct edges in the triangles. Here is thier distribution:



Both graphs show a square-root fall-off. Why?

MI graphs

How does one compute MI, again?

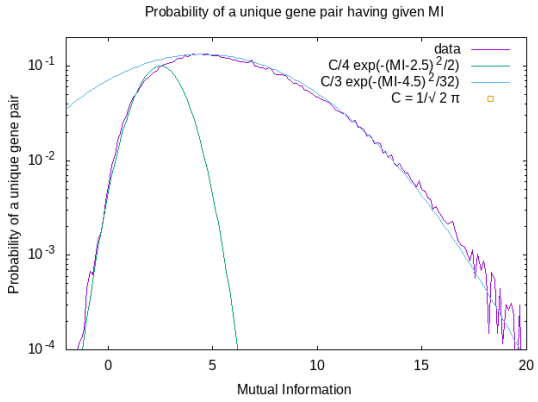
The starting point is interaction triangles, again. So, again, each triangle has three corners, the genes, and three edges, the gene pairs. By counting all possible triangles, one obtains a count of how often each edge is seen -- that is, how often an edge appears in some triangle. That is, one has a count N(ga, gb) which is the number of triangles that the pair (ga,gb) participated in. The rest is downhill. MI is as usual, Yuret's lexical attraction. etc.

When one has counts $N(x,y)$ one can readily compute the lexical attraction between the elements (x,y) as

$$MI(x,y) = \log_2 \frac{N(x,y)N(*,*)}{N(x,*)N(*,y)}$$

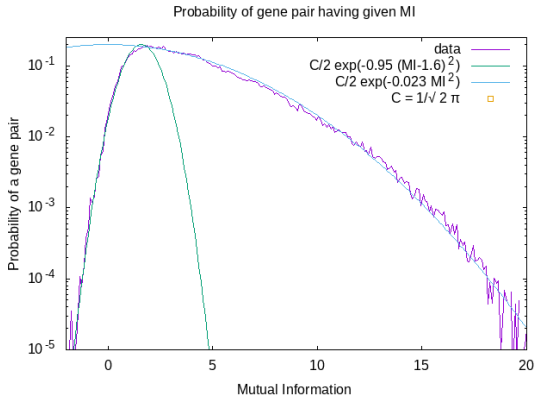
How is the MI distributed as a function of pair frequency? There are two such distributions that one can graph. Both distributions are probabilities of gene-pairs, as a function of MI. The first considers only unique gene-pairs, the second are weighted gene-pairs. That is, in the first, if two gene-pairs were found to have the same MI, then they are equiprobable. In the second, if two gene-pairs were found to have the same MI, then the relative probabilities of each is the same as the number of times they were observed. (Equivalently, one may also say that these are "graphs without and with degeneracy", or "unweighted and weighted graphs").

So, first a graph without degeneracy (all gene pairs are treated as equi-probable)



The normalization is such that the area under the data curve will integrate to 1.0. The data is fitted to two gaussain distributions. Recall that gaussians appear to be parabolas on a semi-log plot. The aparamters of the fit are as shown. They appear to be very nice integers, but this may be just a coincidence. The fits are hand-built, or "eyeballed", and *not* the result of some numerical least-squares curve-fit.

The second graph is very similar, but this time shows the weighted distribution. Again, it is easily fit with two gaussians, but this time, the centers of the gaussians are quite different, as are the widths.



Pentagons

The pentagon has five vrtexes: a pathway (as a single, named vertex), two proteins on that pathway, two genes, each of which express the coresponding protein, and finally, the two genes interacting with each other. Here, there are five edges, and three edge types: (protein in pathway), (gene expresses protein) and (gene-pair interaction). Thus, characterizing this statisitically requires six Zipfian graphs: three for the different vertex types, and three for the different edge types. Each edge type will also have an MI distribution, and so six MI graphs: one for each type, weighted and unweighted. So, twelve graphs in total.

(Based on earlier work, I beleive that that they will all look very similar to one-another... and explanation for that similarity is unknown to me ... the similarity is both reassuring and un-nerving: reassuring, in that there is a certain regularity to the giant mess of biology; un-nerving in that, perhaps, we've over-generalized and lost too many important details...)