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.

It was naively hypothesized that genome/proteome reaction pathways form a scale-free network, and thus would have a Zipfian distribution. Much to our surprise, this is not the case! It seems like *everything* follows a square-root Zipfian distribution! I do not know of any network theory or biology theory that would explain this, so it is a surprise.

An exploration of the mutual information of interaction pathways is also performed. It appears that these are easily fit with a bimodal Gaussian distribution.

This is for human genome, reactome data. I don't doubt that the results are generic in biology.

The results here are preliminary; there are issues with ... my understanding of the pathway annotation may be flawed ...

Introduction

Publicly available genomics and proteomics databases describe a large number of interactions between genes and pathways. Taken together, these datasets describe a large graph, and the one may reasonably wonder about the properties and general structure of that graph. For example, one might as if the graph is scale-free, or if it has a hub-and-spoke structure, or make other graph-theoretic inquiries into it.

In this short monograph, a study is made of "triangles": three genes that mutually interact, and "pentagons": a pair of genes that express a pair of proteins that lie on a common pathway. If the genome (proteome, reactome) network were scale-free, then one might expect Zipfian distributions of triangles and pentagons.[1] This appears not to be the case. A further study is made of the mutual information ("lexical attraction") of pairs. It appears to be easy to describe this as a pair of Gaussians. A theoretical grounding for these results is unknown to the author. Clarifications are solicited.

[1] These questions originally arose during the characterization of a bioinformatics data-mining benchmark, found in `https://github.com/opencog/benchmark/query-loop`, and was elaborated in the "Genome distribution!?" email discussion.

Datasets

The graph networks explored here were constructed from public datasets from MCBI, ChEBI, PubMed, UniProt, SMPDB, Entrez and BioGrid. Two variant networks were constructed and explored. The two differ in how gene interactions were treated. In the first, genes that regulate one-another were treated as directed edges, in that gene A may regulate gene B, but not the other way around. This includes genes that may self-regulate. The second network was obtained from the first by symmetrizing all gene interactions (if A interacts with B, then B interacts with A) and removing all self-interacting genes.

Both graphs contain a total of 49050 genes. There are 540778 gene interactions in the first dataset, which are represented as directed edges. Of these, 347127 are symmetric (*i.e.* have a partner indicating the opposite direction). There were 2939 self-interacting genes. From this, we deduce a total of (347127 - 2939) / 2 = 172094 symmetric interactions, and 540778 - 347127 = 193651 non-symmetric interactions. The number of edges that have non-zero counts (i.e. appeared at least once in a triangle) is 455572.

For the second dataset, the self-interacting genes are removed, and matching symmetrized edges are created, for a total of 731490 directed edges, or half of that, 365745, when the relation is taken as symmetric.[2] The number of edges with non-zero counts (i.e. the number of edges that appear in at least one triangle) is 617530.

Pathways are from https://reactome.org and from http://smpdb.ca/

Triangles

A triangle is defined as a three-way interaction, of gene A regulates gene B, gene B regulates gene C and C regulates A. Such triangles are one of the most basic components of the topological structure of regulatory networks. It is worth understanding thier structure.

Any given gene may appear in multiple triangles; clearly, some genes might participate in many, and some in only a few. So, the first statistical question becomes: what is the distribution of these genes? This is answered via a Zipf-style graph: one makes a list of all of genes, ranking them according to the number of triangles they participated in, and then graphing, on a log-log graph, the number of triangles vs. the rank. Similarly, one can pose the same question, but for edges: how many triangles does a given gene-gene interaction participate in?

By counting the number of times an edge participates in some triangle, one obtains a count $N(g_a, g_b)$ that relates gene g_a to g_b. Given the counts of such pairs, this opens the door to a stable of standard statistical questions: what are the marginal disributions, the entropies, and the mutual information of such pairs? This is explored in a subsequent section.

Because these explorations are perfomed on curated datasets carried out by research labs, one has an open question of whether one is measuring "true biology", or whether

[2] Use the `(count-gene-interactions)` tool.

one is measuring the social network of the scientists, and thier intellectual interests. There does not appear to be any obvious way of answering this question.;

Characterization

The two datasets provide sometimes-similar and sometimes-different results. The differences are noted as appropriate.

In the first dataset, there are 3452807 pointed triangles. These are triangles with a distinguished point. By symmetry, ne would expect this number to be divisible by 3; it is not because many of the triangles are degenerate: There are 145571 degenerate triangles having the only two distinct vertexes, and 2939 "bouquets" - that is, triangles where all three corners are the same vertex. This leaves behind 1102412 distinct triangles having three distinct corners.

A total of 18766 genes participate in these interactions.[3] This differs from the number of genes in the dataset, as not all genes participate in triangles. There are 455572 distinct edges that appear in triangles. These edges appear a grand-total of 5050388 times in different triangles; thus, on average, any given gene interaction might appear in 11 different triangles. However, the concept of "average" is troublesome in Zipfian distributions; this and other marginals are expanded on in a subsequent section.

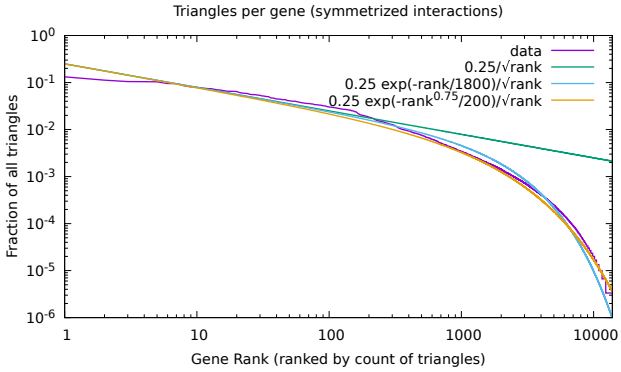
Because the first dataset contains self-interacting genes, not all triangles are "true" triangles; come may be degenerate. That is, one may have triangular relationships of the form "gene A regulates gene B, B regulates C and C regulates A," where B and C are the same gene. The desire to eliminate such degenerate triangles leads to contemplating the second dataset.

The second dataset has 10783686 pointed triangles; by symmetry, only 1/6 of these are distinct, leaving 1797281 distinct triangles. These triangles have a total of 731490 distinct edges. There is a total of 20123 genes participating in these triangles.

Zipf graphs

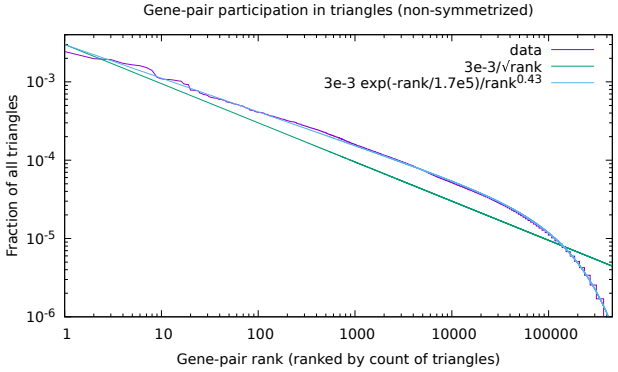
The following figure shows the number of triangles in which a gene participates in. This figure if for the symmetrized network; the corresponding figure for the unsymmetrized network is very similiar.

[3] Counted with `(length loop-participants)`.

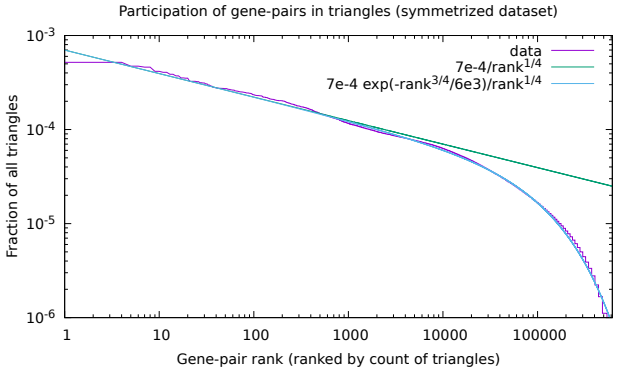


The first unexpected surprise is that this graph is not Zipfian. That is, the distribution is not of the form $1/(rank)^{\alpha}$ with $\alpha \approx 1$. Just eyeballing the general shape, as done above, one promptly arrives that $\alpha = 0.5$ for the leading part of the figure. The tail is certainly not Zipfian, either; it falls of sharply. A first guess, no better than voodoo numerology, suggests that the n-th ranked gene appears in $p(n) = 0.25e^{-\beta n}/\sqrt{n}$ of all triangles, for some constant β. Closer examination suggests $p(n) = 0.25\exp(-\beta n^{3/4})/\sqrt{n}$ provides a better fit. These forms are chosen to simply look appealing to the eye, rather than by minimizing some least-squares curve-fit. These seem like reasonable hypothesis, given the above graph. However, the situation gets a bit confused, when one looks at edges.

One may also ask how many triangles a given edge (a given gene-pair) participates in. A similar distribution results, although the suggestive $\sqrt{n}$ hypothesis does not seem to hold cleanly. As shown in the figure below, the exponent $\alpha \approx 0.43$ appears to offer a better fit. This figure is for the unsymmetrized dataset.



The symmetrization of the edges makes a dramatic change to the distribution exponent. The exponent now appears to be $\alpha \approx 0.25$, as shown in the figure below.



That such a technical change would have a rather profound effect on the structure of the graph is surprising.

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

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

The total MI for a dataset is defined as

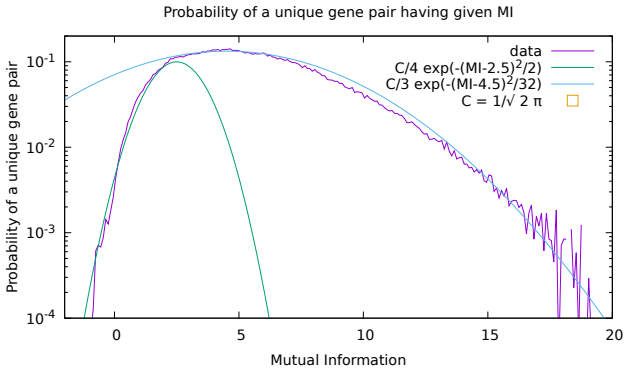
$$MI = \sum_{x,y} p(x,y)MI(x,y)$$

where $p(x,y) = N(x,y)/N(*,*)$ is the (frequentist) probability of observing the pair (x,y). For the symmetric dataset, it is 3.49. The total entropy for the dataset is 18.21. The sparsity is 9.36, where the sparsity is defined as $\log_2$ of the number of non-zero entries in the matrix. In this case, there are 20123 genes that participated in triangles, and 617530 non-zero entries (out of 404935129 possible; the $\log_2$ of this ratio provides the sparsity). Other statistics are possible: any given gene-pair participates in 400 triangles, on average; however, the concept of averages is rather flawed, when one has such Zipfian distributions.

So because the concept of average is flawed, we use MI instead... Hmm. what about the other stats (support, count, length!? I'm confused... OK: deconfuser: the matrix summary report shows *weighted* banach-space norms, so not raw, but weighted by frequency of row/column...)

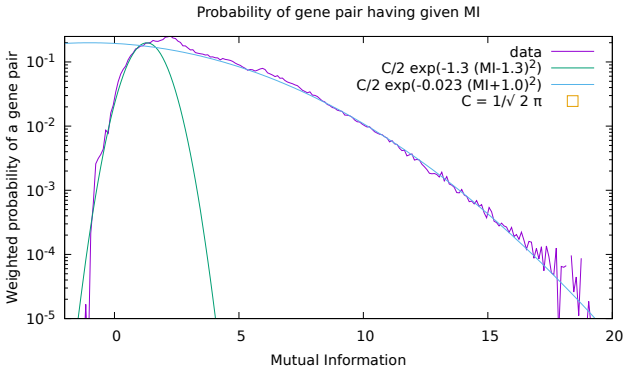
How is the MI distributed as a function of pair frequency? There are two such distributions that one can graph. Both distributions are probabilities of gene-pairs, as a function of MI. The first considers only unique gene-pairs, the second are weighted gene-pairs. That is, in the first, if two gene-pairs 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.



Again, the fit is eyeballed, with a preferrence of finding numerological simple-fraction fits, so as to expose simple laws, if these exist.

Pentagons

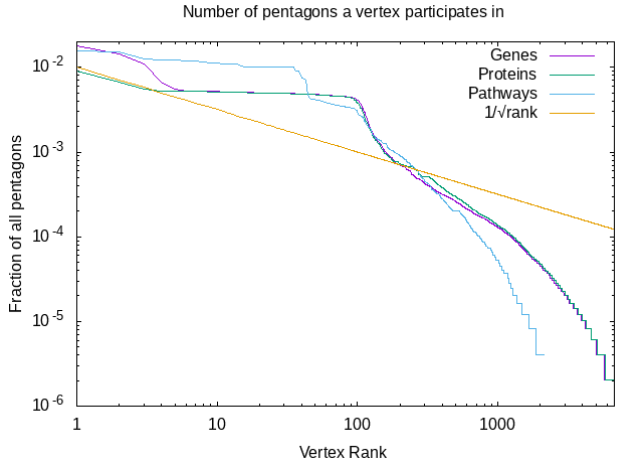
The pentagon has five vertexes: a pathway (as a single, named vertex), two proteins on that pathway, two genes, each of which express the coresponding protein, and finally, the two genes interacting with each other. Here, there are five edges, and three edge types: (protein in pathway), (gene expresses 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...)

Zipf

There are 50501 pathways in which some gene in the dataset is a member of. These are both SMP and R-HSA-tagged pathways (from the smpdb.ca and reactome.org, respectively).

There were 491558 distinct pentagons (or twice that many, if you count mirror symmetry.) There are 6694 unique genes that appeared in pentagonal relationships. There were 6735 proteins that appeared in pentagonal relations. There were 2129 distinct pathways that appeared in pentagonal relations. All have an unexpected distribution. All are graphed below, on the same graph.



The normalization is with respect to the total number of pentagons that a given vertex appears in. Thus, for example, the highest-ranked gene, and the highest-ranked pathway appears in approximately 1 out of 50 pentagons; the highest-ranked protien appears in approximately 1 out of 100, which can be read directly off the above graph.

There were 38843 distinct edges connecting pathways to proteins. There were

148040 distinct gene-protein expression edges. There were 77304 gene-gene interaction edges...