ESM 569 Essay

Hidden Geometry and Coarse-graining Metabolism Network

Shengzhu Yin

Department of Physics

University of Illinois Urbana-Champaign

E-mail: yin20@illinois.edu

Abstract: Metabolic Network Cartography is a key visualization of an organism's mechanism of processing. In a sense, life is an ultimate emergent phenomenon that can be found in nature. However, most cartography only contains topological information, i.e., a mathematical graph. In this essay, we will take a look into series of recent development that further incorporates geometrical information into these metabolic cartography. As a final result, we will see an emergent pattern by utilizing this additional information.

Contents

1 Introduction

1.1 Metabolic Network

Life is considered as one of the biggest emergent phenomena, yet it just seems impossible to understand its mechanism. Metabolism is about how an organism's body works, some of us might recall the Krebs cycle and Calvin cycle that we learned in grade schools. Aside from those "well-understood" metabolism pathways, there are countless pathways we do not understand. Not only that, interactions between these gargantuan pathways remain as mysteries. This collection of metabolic pathways are called a metabolic network. It is not hard to see the importance to decipher such unknown monsters since one of many reasons being having the solution will make us live longer. In fact, majority of pharmaceutical drugs' pathways are mostly not perfectly determined.

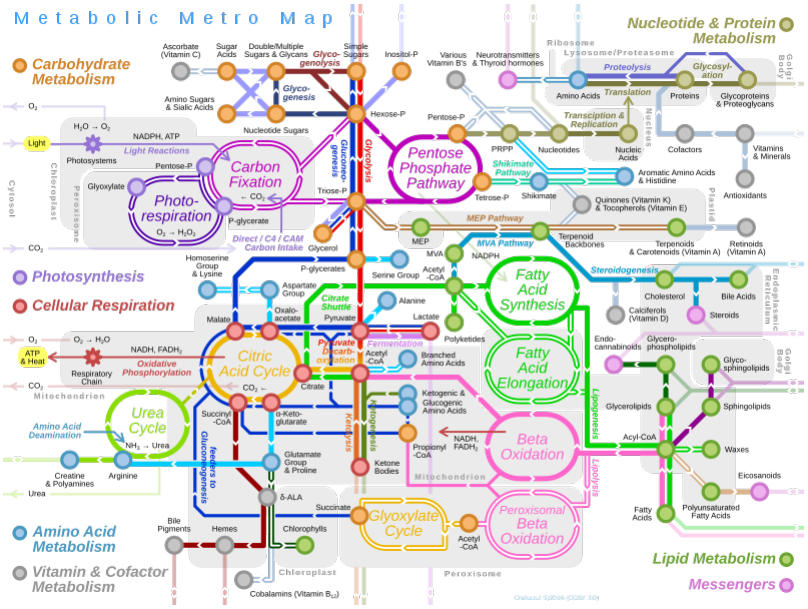

Figure 1. A relatively simple metabolic network. Usually things are a lot more complicated that the path overlaps become unbearable to see. [1]

1.2 Missing Information in Ordinary Metabolic Pathway Maps

As mentioned in the previous section, extracting additional information without more structural regularities is a tough problem. One problem with metabolic maps like figure 1 is that these maps are mathematically a graph; it is purely topological meaning that the length or shape of connections does not carry significance. M. Serrano, M. Boguna, and F.

Sagues had a great insight to give metric to such figures and infer a lot more information about the networks by using probabilistic methods with coarse-graining. Not only that, they found a universal expression that actually describes both E. Coli and humans. Given the complexity gap between two organisms, it is indeed astonishing. Note that this is not the first attempt for the scientific community to find criteria for clustering metabolic networks. Most network-based representation analyses before 2012 have failed. [3] In this essay, we will first discuss the mapping itself including the metric, then we will eventually move to a larger scale by a certain coarse-graining method that is analogous to Kadanoff's block spin. Along the way, there will be several data to show the validity of such mappings agrees with classical biochemical analysis.

2 Hidden Geometry

2.1 Summary of the Mapping

First, we shall consider a bipartite network representation that contains two kinds of vertices. One being the metabolites (ingredients or products that go into the reaction), and the second type of vertex being reactions. In this representation, any given metabolism map can be seen as a graph that is avoiding connections between the same types of nodes. See figure 2 a. Now, Serrano et. al. considers mapping to a more "organized space" to simplify the structure and give a simple metric. Luckily, one-dimensional circles were enough for this task. The answer is $\mathbb{S}^1 \times \mathbb{S}^1$. Although it is diffeomorphic to a two torus, it is more helpful to see this as two overlapping circles, Figure 2 b. To simply put, one can think one circle is for metabolite nodes to attach onto, and the other circle is for reaction nodes. Note that this is not the first time the authors put out such a model. In their previous paper [4], they proposed $\mathbb{S}^1$ model as well to study a completely different issue. We will address that a little bit at the end. Coming back to the original problem, let m be the metabolite and r the reaction. Then the natural metric is $d_{mr} := R\Delta\theta_{mr}$ on two circles with radius R. Where $\Delta\theta_{mr}$ is the angular separation between some metabolite node m and some reaction node r. We also define the probability measure

$$p\left(\frac{d_{mr}}{k_m k_r}\right) := Prob\,(m \text{ connecting to } r) \tag{2.1}$$

Here, the k_m and k_r are degrees of connection on some m node and some r node, i.e., the number of neighbors, which is a piece of topological information. This effective distance not only takes account of pure geometry but also the natural topology that resembles the reality– the interaction of a pathway is more active is has more legs stretched out to other reactions/metabolites. Furthermore, If one pays attention closely the metric given is analogous to inverted Newtonian gravity. The entire embedding process to $\mathbb{S}^1 \times \mathbb{S}^1$ is a rather complicated series of statistical detail (Maximum Likelihood Estimator, MLE, embedding) that disrupts most readers being focused; therefore, see appendix A. In theory, the explicit form of $p\big(\frac{d_{mr}}{k_m k_r}\big)$, which is the interaction strength, can be set to any sensible integrable function. However, the authors stick to the Fermi Dirac distribution in order to empirically fit parameters μ and β. Perhaps one surprising thing is that these fit parameters' values are shared among human and E. Coli as we will mention later. See Figure 2c.

$$p\left(\frac{d_{mr}}{k_m k_r}\right) = \frac{1}{1 + (\frac{d_{mr}}{k_m k_r})^{\beta}} \tag{2.2}$$

Another reason to set the function this way is not only due to a good fit to classical physicochemistry result, but it is also true that this particular choice gives maximal entropy interactions. It means that given a certain set of constraints to probabilistic calculations, the choice yields the most kind of randomness in interactions.[5][6]

Figure 2d is the Receiver Operating Characteristic (ROC) curve. If one defines a threshold in the connection probability, one can divide connections to true/false. The dotted line is the randomized 50/50 guess line. We see that the true positive rate (TPR) is much more than the false positive rate (FPR) by looking at the area under the curve. Also note that in figure 2 b, the image of two nodes in Figure 2a might be far apart even if the distance was close in the preimage, given that the metric exists. From here, we now see there is a certain meaning to the distance between two nodes! As one can see in the Fermi Dirac distribution (Fig. 2c), the connection probability $p\big(\frac{d_{mr}}{k_m k_r}\big)$ dies off if the effective distance grows larger. We call this pathway localization. This is a hint for a higher hierarchical structure thus where the coarse-graining comes in.

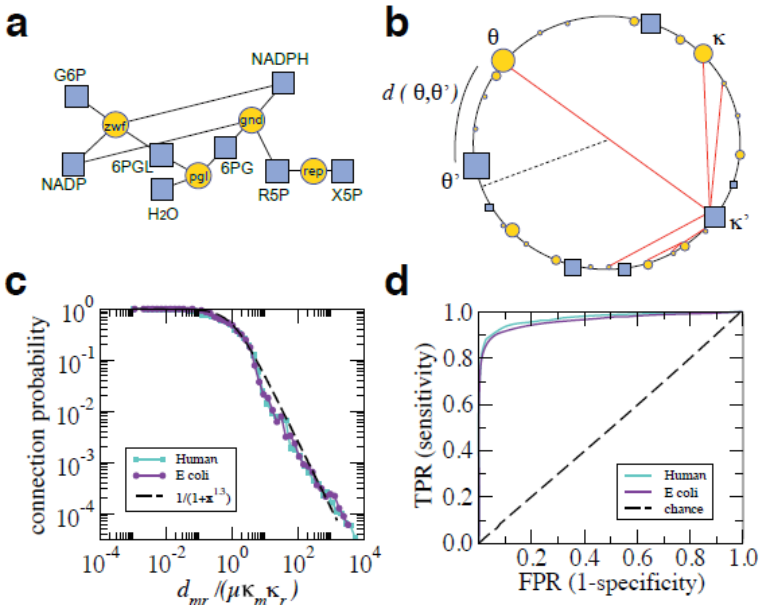

Figure 2. (a) bipartite metabolic path (b) metabolic pathway mapped to $\mathbb{S}^1 \times \mathbb{S}^1$ note that due to additional topological weights to the metric, two near neighbors, even if one defines a distance there, in a) can be far apart after the mapping (c) connection probability (Fermi Dirac)'s empirical fitting to classical physicochemical data. (d) The Receiver Operating Characteristic (ROC) curve. If one defines a threshold in the connection probability, one can divide connections to true/false. The dotted line is the randomized 50/50 guess line. We see that the true positive rate (TPR) is much more than false-positive rate (FPR) [2]

2.2 Validity of the Result

As one can see, Figure 3 is the global metabolic mapping $\mathbb{S}^1 \times \mathbb{S}^1$ of E. Coli. As we have discussed before, the yellow nodes' reactions and blue nodes are metabolites. Note the figure does not label all pathway names. To help in terms of visual comparison, there are multiple resizing and re-positioning of fonts and nodes in the figure. See the figure description. Now, to do a sanity check, the comparison between the angular decomposition of Figure 3 was made to compare with the already existing and trustworthy BiGG database (Figure 4). We see that the well-known strong signal of highly localized pathways is standing out. For example, the Oxidative Phosphorylation, Histidine, Glycolysis, Cofactor, and Prosthetic group pathways do match up with the classical physicochemical database (BiGG). Note the signal can appear at multiple different places on the ring with different probability amplitudes! The name tag is just the average angular position. For instance, each reaction $i = 1, ..., N_{path}$ in each pathway gets assigned an normalized vector r_i pointing outward. The average is just a vectorial arithmetic average of this. If the average carries a norm

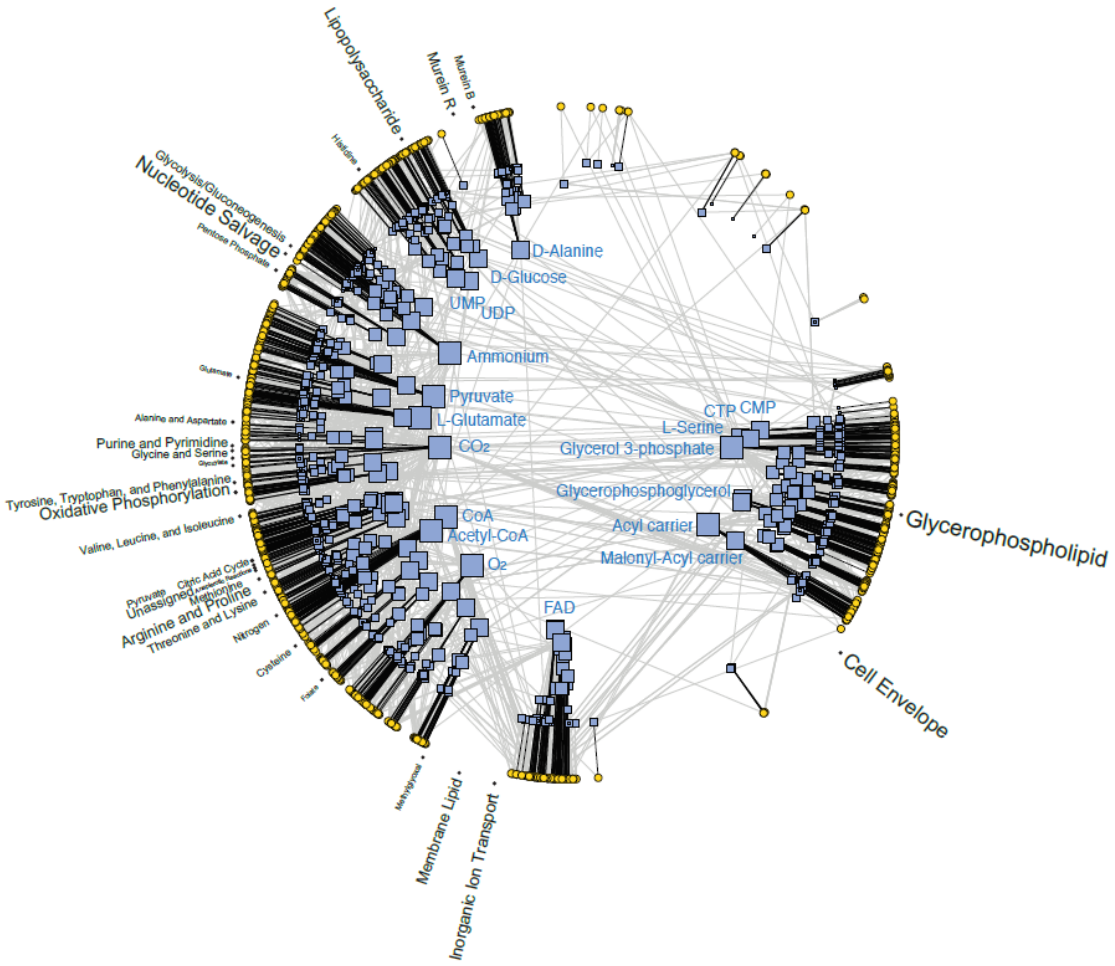

Figure 3. Global $\mathbb{S}^1 \times \mathbb{S}^1$ map of E. Coli's metabolism: Yellow nodes are reactions and blue nodes are metabolites. To help in terms of visual comparison, node size is proportional to the logarithm of the degree and placed in the radial direction according to formula $r = R - 2 \log k_m$. Grey lines have connection probability of less than 0.5, and black ones larger than 0.5. The pathway names are written in the averaged angular position of all reactions within the pathway, and the font size is proportional to log of number of reactions (indication of pathway "size") For human's see Appendix B [2]

of zero, it means that the probability distributions are evenly smeared out! (advantage of isotropic shape). Whereas the unity norm indicates that all reactions happen at one location. For human's case, see appendix B.

$$\langle \vec{r} \rangle \equiv \sum_{i=1}^{N_p} \vec{r}_i / N_p \tag{2.3}$$

Anyway, in figure 4, bars are color-coded according to their metabolic classes, written in figure descriptions. the y-axis is the pathway concentration. along with this data agreement

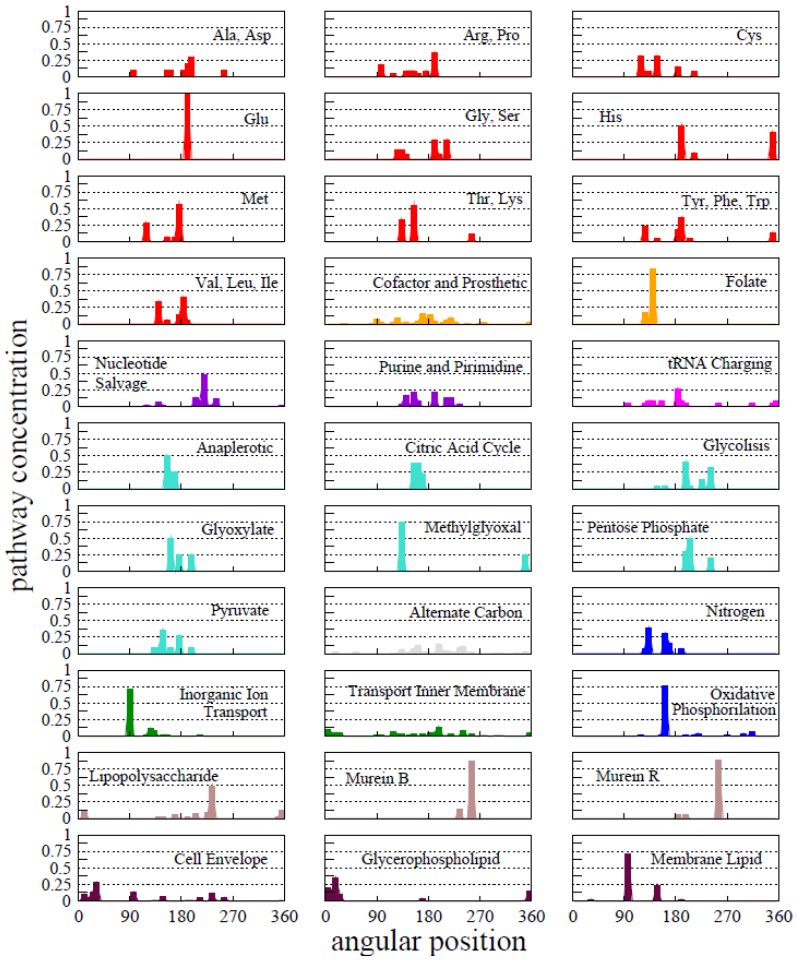

Figure 4. Graph of angular distribution of connection probability (dissected such that 2 degrees per bin). Different metabolic pathway classes are color coded. For example, red is for amino acids pathways. Orange: cofactors and vitamins. Violet for Nucleotide; turquoise for carbs; magenta for tRNA; grey for alternate carbon; blue for xTP, xDP; green for transports; brown for glycan; maroon for lipids.

with previous curve fittings, the data seems valid.

2.3 Coarse-graining and Emergence

In the previous two subsections, we saw that on the double rings, we saw some metabolic pathways are closer to the other, whereas some are distant. This indicates we could build a higher structure in terms of hierarchy. Recall that in the introduction, we have mentioned that clustering metabolic pathways were a hard problem. Consider Figure 3, the double circle map. Now slice it into 8 equal pieces like a pizza (Figure 5). Then the adjacencies are computed between pairs of pathways according to the list of reactions and metabolites that

are shared. However, before that, there is one problem: There are too many networks to extract meaningful information from this calculation. For example, 460 out of 561 potential pathways pairs overlap for E. Coli., and for humans, 1689 pairs out of 4278 pairs overlap (shares common metabolites). Hence, a disparity filter is used to wash out uniformly, randomly, distributed networks. The filter goes against the null hypothesis that states the local probability weights that are contained to a node are randomly distributed. Thus the p-value, probability that the null hypothesis is not rejected, between i and j node can be written as [7]

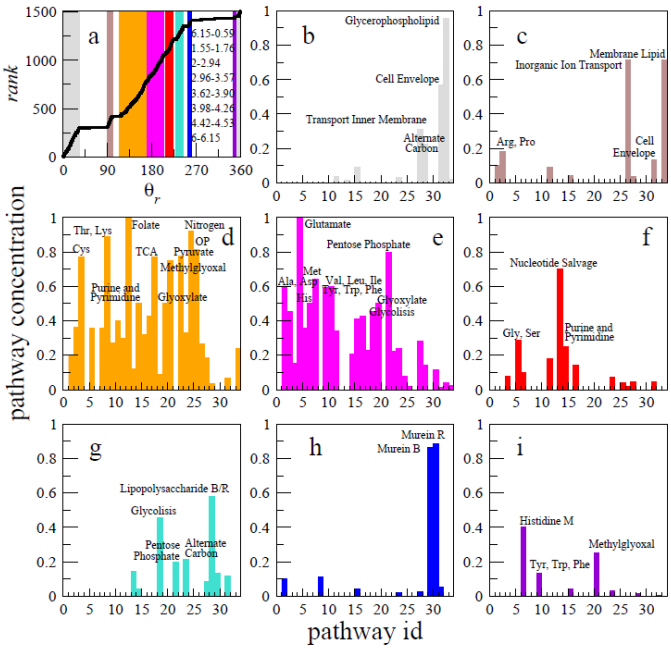

Figure 5. A different representation. 8 sectors are sliced and binned. We will compute adjacencies between pairs of pathways according to the list of reactions and metabolites that are shared.

$$p_{ij} = 1 - (k-1)\int_0^{w_{ij}/s} (1-x)^{k-2} dx < \alpha \tag{2.4}$$

Where α is the significance level, which is the choice of our hands. After this appropriate filtration, the cluster structure of emergent pathway maps reveals itself (Figure 6). Note that the measure here between the pathways are written as

$$Crosstalk_{P_a P_b} = \Sigma_{j\in P_a, k\in P_b, i\in v}\big(p(x_{ij}) + p(x_{ik})\big)|_{filter-survived-links} \tag{2.5}$$

Where $v \in M_{ab}$ is the set of metabolites shared by the same reaction that is present in both pathway P_a and P_b.

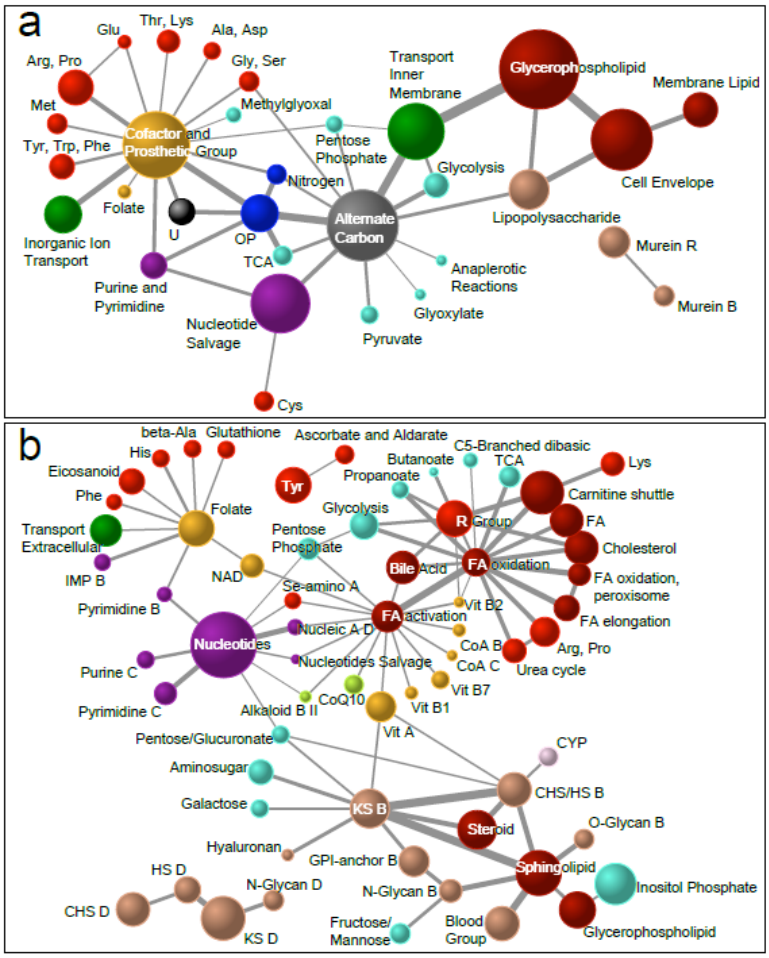

Figure 6. Emerged Metabolic backbone (pathway clusters) (a) E. Coli. (b) Human. The larger the ball is, the more reaction numbers it contains.

We see in figure 6 that both E. Coli. and Humans have star-shaped trees. This means that clustering was successfully done with "parents" and its "child". For example, the cofactor prosthetic group in E. Coli. plays a similar role as the human case: They supply amino acid "child" pathways. Furthermore, to recapitulate the filtering statement above, E. Coli. had 82 percent effective crosstalks (min 1.8 max 159.91), while humans only had 38.64 percent effective crosstalks (1689 out of 4278, as stated above, min 1.19 max 131.28). It means human pathways are more independent or carries more modularity. This entire procedure indeed resembles Kadanoff's block spin renormalization group procedure. The

filter acted as an averaging-out tool in this case.

3 Discussion and Conclusion

In this essay, we have reviewed [2] and multiple previous papers published by the author. Based on the data agreement with BiGG database and other classical physicochemistry facts, the results do look promising. Aside from that, the MLE method to find the $\mathbb{S}^1 \times \mathbb{S}^1$ embedding also looks pretty consistent as the parameter space was compact and theoretical and the database agreed with each other. One thing the author did not address much is the fitting parameter in the Fermi distribution looks universal, but it is just two cases. It would have been nice if the author extended some examples to show. Additionally, the existence of filtration. It is not too clear whether the result is stable against perturbation by the significance level (it would change the modularity of pathway networks). Lastly, it would also be nice to see whether this embedding still spits out a similar result if we swap to another symmetric compact manifold st. the likelihood function still is nicely behaved.

4 Appendix

4.1 Appendix A

See [2],[4],[7] for MLE search for θ_i and hidden degree parameter, filtration function, and finite size effect.

4.2 Appendix B (next page)

Acknowledgments

I would like to thank professor Nigel Goldenfeld for teaching us. The Youtube uploads really helped out me.

References

[1] en.wikipedia.org/wiki/File:MetabolicMetroMap.svg

[2] M. Serrano, M. Boguna, and F. Sagues, Molecular BioSystems (2012)

[3] S. Fortunato, Physics Reports 486, 75 (2010)

[4] M. A. Serrano, D. Krioukov, and M. Bogu na, Phys. Rev. Lett. 100, 078701 (2008).

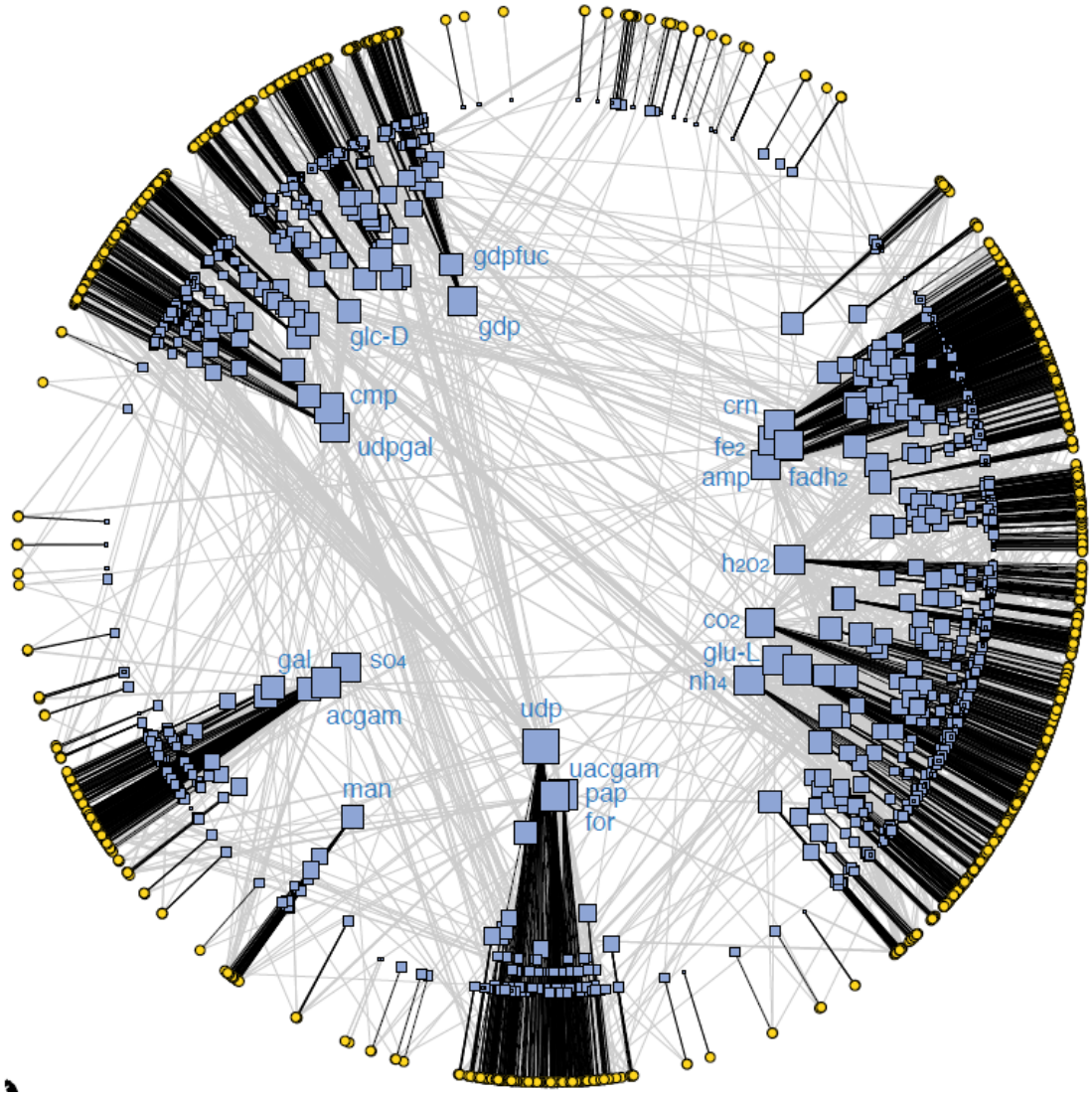

Figure 7. $\mathbb{S}^1 \times \mathbb{S}^1$ map of Human's metabolism: Yellow nodes are reactions and blue nodes are metabolites. To help in terms of visual comparison, node size is proportional to the logarithm of the degree and placed in the radial direction according to formula $r = R - 2 \log k_m$. Grey lines have connection probability of less than 0.5, and black ones larger than 0.5. The pathway names are written in the averaged angular position of all reactions within the pathway, and the font size is proportional to log of number of reactions (indication of pathway "size")

[5] D. Garlaschelli and M. I. Loredo, Phys. Rev. E 78, 015101 (2008).

[6] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Bogu na, Phys. Rev. E 82, 036106 (2010).

[7] M. A. Serrano, M. Bogu na, and A. Vespignani, Proc. Natl. Acad. Sci. USA 106, 6483 (2009).