International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:8, No:8, 2014

Hypergraph Models of Metabolism

Nicole Pearcy, Jonathan J. Crofts, Nadia Chuzhanova

Abstract—In this paper, we employ a directed hypergraph model to investigate the extent to which environmental variability influences the set of available biochemical reactions within a living cell. Such an approach avoids the limitations of the usual complex network formalism by allowing for the multilateral relationships (i.e. connections involving more than two nodes) that naturally occur within many biological processes. More specifically, we extend the concept of network reciprocity to *complex hyper-networks*, thus enabling us to characterise a network in terms of the existence of mutual hyper-connections, which may be considered a proxy for metabolic network complexity. To demonstrate these ideas, we study 115 metabolic hyper-networks of bacteria, each of which can be classified into one of 6 increasingly varied habitats. In particular, we found that reciprocity increases significantly with increased environmental variability, supporting the view that organism adaptability leads to increased complexities in the resultant biochemical networks.

Keywords—Complexity, hypergraphs, reciprocity, metabolism.

I. Introduction

MANY biological systems can be described in terms of their interaction patterns [1]. These systems are typically modelled as simple or directed graphs and consist of a set of nodes representing the objects under investigation, e.g. metabolites, proteins or genes, and a set of edges representing relationships between node pairs. However, many biological processes have more complex relationships involving more than two nodes [2]. As an example, consider metabolic networks, which consist of many reactions of the form A + B $\longrightarrow$ C + D, involving four (or more) different metabolites. In this case, traditional network theory provides an inadequate description of the full chemical reaction system in at least two ways: (i) a chemical reaction system may be represented as a simple graph in a variety of different ways [3], [4], the choice of which accentuates different aspects of the metabolic process; and (ii) information is inevitably lost when reducing the full system to a simple graph [5], and so one risks oversimplifying the system of interest in a potentially significant way.

Hypergraphs provide an attractive alternative since they allow for the description of more general interactions consisting of more than two nodes. Metabolic networks are particularly amenable to such an approach, with nodes representing different metabolites and hyperedges, that is sets of nodes, representing chemical reactions. Moreover, for metabolic networks it is useful to distinguish between directed and undirected hypergraphs, as under normal physiological conditions many reactions can be considered as being irreversible.

N. Pearcy, N. Chuzhanova and J. J. Crofts are with the Department of Mathematics & Physics, Clifton Campus, Nottingham Trent University, Nottingham, NG11 8NS, UK, (e-mail:nicole.pearcy@ntu.ac.uk).

Whilst the topological characterisation of complex networks has received considerable attention over the past decade [6], [7], the theory of complex hyper-networks is far less developed, which, coupled with the increased algorithmic complexities that accompany such an approach, perhaps explain why this more natural framework has not been more widely adopted in the study of biological networks to date. Recently, however, a number of studies have attempted to extend complex network reasoning to this more complicated setting. For example, the commonly used clustering coefficient, a measure of the probability that any two neighbours of a given node are also neighbours, has been extended to hypergraphs [8], [9]. In [9] in particular, it was shown that the inverse scaling between network degree and clustering, typically reported in standard network analyses of metabolism and considered indicative of a hierarchical network structure [1], may actually be an artefact due to misrepresentation. Another important topological parameter that has been generalised to this more complicated setting is the subgraph centrality [8]. Centrality measures provide a measure of the relative importance of each node within a network, and the generalisation given in [8] provides such a characterisation for the nodes of a hyper-network. Other notable works include the extension of random graph models such as Erdös-Rényi and Barabási-Albert to hyper-networks [10]; the use of random walks to infer information flow and network architecture [11], [12]; and novel community detection algorithms for determining modular hyper-network structure [13], [14].

In this work we extend the concept of reciprocity to complex hyper-networks, and use it to study metabolism for a large cohort of bacterial species. Reciprocity measures the proportion of directed links to the total number of links in a network [15], and many real-world networks have been shown to display non-random reciprocity [16]. In the case of metabolic networks, reciprocity may be considered a proxy for how 'far-from-equilibrium' the biochemical reaction system is, thus enabling us to quantify the effect of environmental variability on the 'global reversibility' of reactions. More generally, the extension of complex network concepts, such as reciprocity, promises to provide further insight and understanding into the many complex biological systems for which a standard network representation provides an inadequate description.

The paper is organised as follows. In §II we start with a discussion of the necessary theoretical prerequisites regarding hypergraphs, before introducing the concept of reciprocity for complex hyper-networks. A description of the data and our experimental results is given in §III. The paper conclusion is given in §IV.

II. Complex Hyper-Networks

A. Preliminaries

A hypergraph is a pair of objects $H = (V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ is the set of n nodes and $E = \{E_1, E_2, ..., E_m\}$ is the set of m hyperedges. Each hyperedge is made up of subsets of V, such that $\bigcup_i E_i = V$ and $E_i \neq \emptyset$. For an undirected hypergraph, two nodes v_i and v_j are considered adjacent to one another if there exists a hyperedge k, such that $v_i \in E_k$ and $v_j \in E_k$. If, as we do in this work, one considers directed hypergraphs then each hyperedge is further subdivided into two sets - the tail set X and the head set Y, allowing us to distinguish between bidirectional and unidirectional relationships. Directed hyperedges are more precisely referred to as hyperarcs. For an illustration of a directed hypergraph see Fig. 1(a)-(b).

The number of hyperarcs containing a vertex, v_i say, in the tail set is called the out-degree of that vertex, $d^{\text{out}}(v_i)$, and the number of hyperarcs containing v_i in the head set is called the in-degree of that vertex, $d^{\text{in}}(v_i)$. More formally, we have

$$d^{\text{out}}(v_i) = |\{E_k \in E : v_i \in X_k\}|$$

and

$$d^{\text{in}}(v_i) = |\{E_k \in E : v_i \in Y_k\}|,$$

where $|x|$ denotes the cardinality of the set x.

A directed hypergraph can be represented by a variety of different matrices [17], the most popular of which is the *incidence matrix*, an $n \times m$ matrix $C(H)$ representing the relationships between the n nodes and m hyperarcs. The entries of the matrix $C(H)$ are given by

$$C_{ij} = \begin{cases} 1, & \text{if } v_i \in X_j, \\ -1, & \text{if } v_i \in Y_j, \\ 0, & \text{otherwise,} \end{cases}$$

that is, C_{ij} equals 1 or -1 depending upon whether v_i belongs to the head set or tail set of the jth hyperarc, respectively (see Fig. 1(c) for an example). Alternatively, a directed hypergraph can be represented by two incidence matrices - the negative (or outer) incidence matrix $C_-(H)$ and the positive (or inner) incidence matrix $C_+(H)$, representing the tail sets and head sets of the hyperarcs respectively. That is, the elements of $C_-(H)$ are equal to 1 if $v_i \in X_j$ and 0 otherwise. The elements of $C_+(H)$, on the other hand, are equal to 1 if $v_i \in Y_j$ and 0 otherwise [11].

Another important matrix representation is that of the *adjacency matrix*, $A(H)$, whose ijth element is given by the cardinality of the set of hyperarcs such that $v_i \in X_k$ and $v_j \in Y_k$. Note that the adjacency matrix can be derived from the inner and outer incidence matrices defined above as follows

$$A(H) = C_-(H)C_+(H)^T;$$

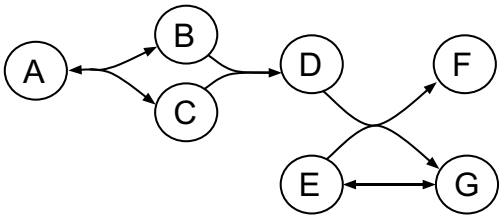
more formally, the elements of $A(H)$ are defined as

$$A_{ij} = |\{E_k \in E : \{v_i \in X_k, v_j \in Y_k\} \subset E_k\}|.$$

In this work we set the diagonal elements all equal to zero as hyper-loops are not allowed. An illustrative example is provided in Fig. 1(d).

Reaction 1: $\mathrm{A} \rightarrow \mathrm{B} + \mathrm{C}$
Reaction 2: $\mathrm{B} + \mathrm{C} \rightarrow \mathrm{D}$
Reaction 3: $\mathrm{D} + \mathrm{E} \rightarrow \mathrm{F} + \mathrm{G}$
Reaction 4: $\mathrm{E} \rightarrow \mathrm{G}$
Reaction 5: $\mathrm{B} + \mathrm{C} \rightarrow \mathrm{A}$
Reaction 6: $\mathrm{G} \rightarrow \mathrm{E}$

(a) Reaction network



(b) Directed hypergraph

$$C(H) = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} \begin{array}{c} \begin{array}{cccccc} R1 & R2 & R3 & R4 & R5 & R6 \end{array} \\ \begin{pmatrix} -1 & 0 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & -1 \end{pmatrix} \end{array}$$

(c) Incidence matrix

$$A(H) = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} \begin{array}{c} \begin{array}{ccccccc} A & B & C & D & E & F & G \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

(d) Adjacency matrix

Fig. 1. An example of a hypothetical reaction system and the accompanying hypergraph model and matrix representations

Another useful concept for our purposes is that of a *walk*. We define a walk of length l on a directed hypergraph as any sequence of vertices $v_1, v_2, ..., v_l, v_{l+1}$, such that $\{v_i \in X_k, v_{i+1} \in Y_k\} \subset E_k$ for some hyperarc k and for each $i = 1, 2, ..., l$. A *closed* walk (CW) is one for which the sequence starts and ends on the same vertex, i.e. $v_1 = v_l + 1$. Note that the vertices and hyperarcs involved in a walk are not necessarily distinct.

Finally, we require the notion of the *underlying directed*

hypergraph, $U(H)$, of a hypergraph, in which all hyperarcs of H are considered as being bidirectional, that is, for each hyperarc E_i there exists another hyperarc E_j whereby $X_i = Y_j$ and $Y_i = X_j$.

B. Reciprocity in Complex Hyper-networks

Reciprocity in standard graphs measures the proportion of mutual relationships within the network, or, in other words, the probability that an edge from B to A exists given that an edge from A to B exists. More formally, reciprocity is defined as

$$r = \frac{L^{\leftrightarrow}}{L}, \tag{1}$$

where $L^{\leftrightarrow}$ is the number of bidirectional edges and L is the total number of edges in the graph. Now, in order to extend the above measure of reciprocity to directed hyper-networks it is useful to recount the following theorem, which can be found, for example, in [18].

Theorem 1: The number of walks ($i \neq j$) or closed walks ($i = j$) of length k in a directed hypergraph are equal to the $(i, j)^{th}$ element of the matrix A^k.

From the above it is straightforward to rewrite (1) in terms of the adjacency matrix and thus to generalise to the case of directed hypergraphs.

$$r(H) = \frac{\mathrm{trace}(A^2)}{\mathrm{trace}(U^2)}. \tag{2}$$

Here A and U are the adjacency matrices of the directed hypergraph and underlying directed hypergraph, respectively, and the trace of a matrix is the sum of the diagonal elements. Note that in the case of a standard graph (1) and (2) are equivalent. Importantly, unlike the standard reciprocity measure, (2) includes information regarding the number of mutual hyper-connections that exist between a pair of vertices.

III. Results

A. Metabolic Data

The metabolic data employed in this study was acquired from the KEGG database on 12th October 2013 [19]. The organism specific reaction lists were derived using the reaction.lst file from the KEGG ftp site, which include full chemical reaction equations (i.e., with stoichiometric coefficients and currency metabolites). More specifically, KEGG XML files were used to extract the set of reactions for an organism, and then reaction.lst was used to obtain the full chemical equations for these reactions. These reaction lists are described with reaction ID's, metabolic map ID's and the chemical equation whose compounds are represented as the KEGG compound ID's. The following is an example from Butanoate metabolism:

R00212: 00620: C00024 + C00058 <=> C000010 + C00022

Note that these reaction lists have been derived by curating several chemical pathway maps from the KEGG database. Thus, since a reaction may be present within multiple metabolic maps there exist some reactions that are repeated within the list. Any repeat reactions with the same reaction ID are thus removed from the reaction list. However, due to errors within the KEGG database some of these repeats are not identical. This is due to the fact that chemical equations in different chemical pathways maps are catalysed by the same enzyme and thus have the same reaction ID, yet sometimes the reactions differ. These non-trivial cases, where we have differences between either the metabolites involved or the directionality of the reaction, were treated by taking the most comprehensive equation. For example, reaction R00212 is present within Butonoate metabolism (map ID: 00650) as reversible and Pyruvate metabolism (map ID: 000620) as irreversible. Here, we include the reaction as reversible, since this equation includes the most information and is thus regarded as the most comprehensive out of the two.

Another problem arises due to condensation and polymerization reactions. These type of reactions often involve compounds occurring as both a substrate and product in the reaction, which can be problematic when representing a reaction system in matrix form. The incidence matrix $C(H)$, for example, uses a -1 to represent the substrates and 1 to represent the products of a reaction in order to distinguish between the different sides of a reaction. Obviously, reactions involving overlapping substrates and products cannot be represented in this manner. For that reason, we exclude such reactions. Several other studies, such as those using flux based analysis [20], [21], have also excluded these reactions due to them being imbalanced.

B. Metabolic Networks as Directed Hypergraphs

We represent the metabolic networks as directed hypergraphs, that is, the n metabolites are represented by nodes and the set of chemical equations are represented as hyperarcs. Each hyperarc is subdivided into a tail set and head set to correspond to the substrates and products of the reactions, respectively. Reversible reactions are considered as two separate reactions in this set-up, that is, a reaction of the form $A + B \leftrightarrow C + D$ is treated as $A + B \rightarrow C + D$ and $C + D \rightarrow A + B$ (see Fig. 1 (a)-(b) for a hypothetical example). To simplify the analysis, we consider only the largest connected component for each network.

We considered 115 metabolic networks, each being categorised according to their environmental habitat (see Table I). The organisms can be found in a variety of conditions, ranging from highly specialised (e.g. symbiotic bacteria living within a host), to extremely heterogeneous conditions such as soil, and thus have evolved under very different selective pressures.

Fig. 2 shows a plot of the average hyper-network reciprocity, $r(H)$, versus environmental variability for the different bacterial networks. Note that the average here is taken over each of the 6 environmental classes: obligate, specialised, aquatic, facultative, multiple and terrestrial. Importantly, we found that the hyper-network reciprocity increased significantly with environmental variability. The lowest value

TABLE I
NETWORK STATISTICS FOR THE REACTION GRAPHS OF THE 115 BACTERIAL SPECIES STUDIED IN THIS WORK CLASSIFIED ACCORDING TO ENVIRONMENTAL VARIABILITY.

Environment*	Nodes			Hyperarcs		
	min	median	max	min	median	max
Obligate (34)	224	441	979	197	443	1156
Specialised (5)	643	695	743	707	805	841
Aquatic (4)	754	851	1014	819	944	1146
Facultative (41)	244	947	1308	204	1155	1510
Multiple (28)	631	900	1226	712	1078	1468
Terrestrial (3)	890	942	955	1086	1205	1219
Total (115)	224	748	1308	197	895	1510

*According to the NCBI, obligate bacteria have the most constant environment, followed by specialised and aquatic, and then facultative, multiple and terrestrial bacteria in that order. In the first column, numbers in brackets denote the number of networks in each class.

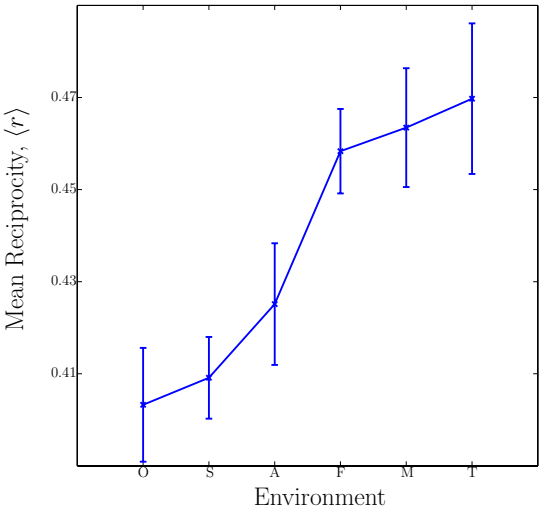


Fig. 2. Relationship between average hyper-network reciprocity, $\langle r \rangle$, and environmental variability. The six bacterial habitats along the x-axis are in order of environmental variability according to the NCBI classification scheme: **O**bligate, **S**pecialised, **A**quatic, **F**aculative, **M**ultiple and **T**errestrial. Here vertical bars denote the standard error of the mean

of reciprocity is found for the bacteria within the obligate class, followed by a slight increase for the specialised class, and then again slightly higher for the aquatic class, there is then a relatively steep increase to the facultative, multiple and terrestrial classes with a relatively small increase between each class. The group differences shown in Fig. 2 are significant by the Kruskal-Wallis (KW) test (p-value $< 10^{-4}$).

This result supports the idea that habitat lifestyle plays an important role in an organisms hyper-network topology, and is consistent with a number of other recent studies that found a relationship between genome size and variations in the environment [22]–[25]. In the current context, the relationship found in Fig. 2 can be viewed as an evolutionary adaptation caused by a larger amount of uncertainty present within a more varied environment, and thus the ease by which metabolites are returned is higher to enable greater adaptability to fluctuations within the environment.

IV. CONCLUSIONS

This paper extends the concept of network reciprocity to complex hyper-networks in order to study the role that directionality relationships play in shaping these more general, multi-faceted structures. These preliminary results on a large cohort of metabolic networks show that the new measure can be used to infer biologically relevant information. In particular, we found evidence for increased complexities in the metabolic hyper-networks of those organisms found in more hostile environments, in the sense that they displayed greater 'global reversibility'. In future work in this area we plan to expand the study performed here so as to include more detailed information regarding the bacterial species studied, for example, the effect of oxygen availability on hyper-network structure.

ACKNOWLEDGMENT

We thank Kazuhiro Takemoto for generously providing the metabolic data used in this study and for a number of useful discussions. NP is grateful to Nottingham Trent University for support via a 2008 RAE funded PhD scholarship.

REFERENCES

[1] M. Buchanan, G. Caldarelli, P. De Los Rios, F. Rao, M. Vendruscolo (Eds), *Networks in Cell Biology*, Cambridge University Press, 2010.
[2] S. Klamt, U. Haus, F. Theis, *Hypergraphs and cellular networks*, PLoS Computational Biology, **5**(5) e1000385 (2009).
[3] P. Holme, *Model validation of simple-graph representations of metabolism*, Journal of the Royal Society Interface, **6**(40) pp. 1027–1034 (2009).
[4] P. Holme, M. Huss, *Substance graphs are optimal simple-graph representations of metabolism*, Chinese Science Bulletin, **55**(27–28) pp. 3161–3168 (2010).
[5] R. Montaẽz, M.A. Medina, R.V. Solé, C. Rodríguez, *When metabolism meets topology: Reconciling metabolite and reaction networks*, Bioessays, **32**(3) pp. 246–256 (2010).
[6] Mark J. Newman, *Networks: An Introduction*, Oxford University Press, 2010.
[7] Ernesto Estrada, *The Structure of Complex Networks*, Oxford University Press, 2011.
[8] E. Estrada, J.A. Rodriguez-Velazquez, *Subgraph centrality and clustering in complex hyper-networks*, Physica A **364** pp. 581–594 (2006).
[9] W. Zhou, L. Nakhleh, *Properties of metabolic graphs: biological organization or representation artifacts?*, BMC Bioinformatics, **12**(132) (2011).
[10] J. Guillaume, M. Latapy, *Bipartite structure of all complex networks*, Information Processing Letters, **90** pp. 215–221 (2004).
[11] A. Ducournau, A. Bretto, *Random walks in directed hypergraphs and application to semi-supervised image segmentation*, Computer Vision and Image Understanding, **120** pp. 91–102 (2014).
[12] A. Bellaachia, M. Al-Dhelaan, *Random walks in hypergraph*, in Proceedings of the 2013 International Conference on Applied Mathematics and Computational Methods, Venice Italy, pp. 187–194 (2013).
[13] A. Vazquez, *Finding hypergraph communities: a Bayesian approach and variational solution*, Journal of Statistical Mechanics: Theory and Experiment, (2009) P07006.
[14] T. Michoel, B. Nachtergaele, *Alignment and integration of complex networks by hypergraph-based spectral clustering*, Physical Review E **86**(056111) (2012).
[15] S. Wasserman, K. Faust, *Social Network Analysis. Methods and Applications*, Cambridge University Press, 1994.
[16] D. Garlaschelli, M.I. Loffredo, *Patterns of link reciprocity in directed networks*, Physical Review Letters, **93**(268701) (2004).
[17] G. Gallo, G. Longo, S. Pallottino, S. Nguyen, *Directed hypergraphs and applications*, Discrete applied mathematics, **42**(2) (1993).
[18] J.A. Rodríguez, *On the Laplacian spectrum and walk-regular hypergraphs*, Linear and Multilinear Algebra, **51**(3) (2003).

[19] M. Kanehisa, *The KEGG database*, Silico Simulation of Biological Processes, **247**(91) (2002).
[20] A. Samal, A. Wagner, O.C.Martin, *Environmental versatility promotes modularity in genome-scale metabolic networks*, BMC Systems Biology, **5**(1) (2011).
[21] A. Samal, O.C. Martin, *Randomizing genome-scale metabolic networks*, PloS One, **6**(7) (2011).
[22] M. Parter, N. Kashtan, U. Alon, *Environmental variability and modularity of bacterial metabolic networks*, BMC Evolutionary Biology, **7**(1) (2007).
[23] A. Kreimer, E. Borenstein, U. Gophna, E. Ruppin, *The evolution of modularity in bacterial metabolic networks*, Proceedings of the National Academy of Sciences, **105**(19) (2008).
[24] S.C. Janga, M.M. Babu, *Network-based approaches for linking metabolism with environment*, Genome Biology, **9**(11), (2009).
[25] J.J. Crofts, E. Estrada, *A statistical mechanics description of environmental variability in metabolic networks*, Journal of Mathematical Chemistry, **52**(2) (2014).