Bioinformatics, 31(14), 2015, 2332–2339
doi: 10.1093/bioinformatics/btv138
Advance Access Publication Date: 9 March 2015
Original Paper

OXFORD

Systems biology

Hi-Jack: a novel computational framework for pathway-based inference of host–pathogen interactions

Dimitrios Kleftogiannis[1,*], Limsoon Wong[2], John A.C. Archer[1,3] and Panos Kalnis[1]

[1]Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal, Jeddah 23955-6900, Saudi Arabia, [2]School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417, Singapore and [3]Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Jeddah 23955-6900, Saudi Arabia

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on November 19, 2014; revised on February 9, 2015; accepted on March 4, 2015

Abstract

Motivation: Pathogens infect their host and hijack the host machinery to produce more progeny pathogens. Obligate intracellular pathogens, in particular, require resources of the host to replicate. Therefore, infections by these pathogens lead to alterations in the metabolism of the host, shifting in favor of pathogen protein production. Some computational identification of mechanisms of host–pathogen interactions have been proposed, but it seems the problem has yet to be approached from the metabolite-hijacking angle.

Results: We propose a novel computational framework, Hi-Jack, for inferring pathway-based interactions between a host and a pathogen that relies on the idea of metabolite hijacking. Hi-Jack searches metabolic network data from hosts and pathogens, and identifies candidate reactions where hijacking occurs. A novel scoring function ranks candidate hijacked reactions and identifies pathways in the host that interact with pathways in the pathogen, as well as the associated frequent hijacked metabolites. We also describe host–pathogen interaction principles that can be used in the future for subsequent studies. Our case study on *Mycobacterium tuberculosis (Mtb)* revealed pathways in human—e.g. carbohydrate metabolism, lipids metabolism and pathways related to amino acids metabolism—that are likely to be hijacked by the pathogen. In addition, we report interesting potential pathway interconnections between human and *Mtb* such as linkage of human fatty acid biosynthesis with *Mtb* biosynthesis of unsaturated fatty acids, or linkage of human pentose phosphate pathway with lipopolysaccharide biosynthesis in *Mtb*.

Availability and implementation: Datasets and codes are available at http://cloud.kaust.edu.sa/Pages/Hi-Jack.aspx

Contact: Dimitrios.Kleftogiannis@kaust.edu.sa

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metabolism is an aggregation of chemical reactions utilized by cellular processes to generate vital metabolites for survival and growth. Metabolic networks are organized in pathways where reactions are connected to each other, meaning that substrates (input molecules) catalyzed by enzymes are transformed to products (output molecules; Croes *et al.*, 2006). These networks are represented as graphs. The available information consists of metabolites, enzymes, genes and relevant annotations, usually coming from experiments in model organisms (e.g. *E. coli*) and archived in repositories such as KEGG (Kanehisa and Goto, 2000), EcoCyc/MetaCyc (Karp *et al.*, 2000) and aMaze (Lemer *et al.*, 2004).

Data that harbours the relevant information enables the development of computational methods that identify interconnections between cellular processes, with a focus on metabolic pathway inference. One stream of methods targets synthetic biology and pathway reconstruction. Algorithms such as CMPF (Lim and Wong, 2012), MRSD (Xia *et al.*, 2011), Metabolic PathFinding (Croes *et al.*, 2005) and the FMM web-server (Chou *et al.*, 2009) take as input pairs of source and target metabolites, and search for routes in the metabolic network that satisfy specific constraints such as shortest distance between source and target or, minimal switching from one pathway to another. Another stream of methods targets inference of complex regulatory modules given sets of gene expression profiles (Chueh and Lu, 2012; Ourfali *et al.*, 2007).

The aforementioned methods opened new routes for investigating interactions between metabolic processes, simplifying functional mechanisms and deciphering relationships between genes, chemical reactions and metabolites using data from a single organism. However, studying the interplay between two different organisms that interact (host and pathogen) and inferring relationships between them is a more challenging problem. In addition, host–pathogen interactions are important for understanding infection mechanisms and developing better prevention and treatment strategies of infectious diseases. However, up to this point, state-of-the-art computational methods for studying host–pathogen interactions have aimed primarily at identifying host–pathogen protein–protein interactions (PPIs; Zhou *et al.*, 2013), rather than at the metabolic-pathway level.

Pathogens have the ability to survive and populate inside a host under extremely hostile conditions by invading the host's immune-response system (Pieters and Gatfield, 2002; Singh *et al.*, 2013). It is also well established that pathogens, upon infection, hijack some of the host's functional processes to produce nutrients required for survival (Niederweis, 2008). Thus, inferring pathways in the pathogen that are likely to interact with pathways in the host, based on the idea of hijacking metabolites or nutrients, is a novel and promising direction for studying host–pathogen interactions.

We propose Hi-Jack, a novel computational framework for inferring metabolic pathway interactions between hosts and pathogens based on the idea of metabolite hijacking. Hi-Jack searches for chemical reactions required for a pathogen's growth that satisfy certain metabolite/nutrient-acquisition criteria. Input to the algorithm is the metabolic network of a host organism and the metabolic network of a pathogen. Both metabolic networks are represented as directed graphs, and graph-theoretic metrics as well as graph statistics are collected to rank pathways according to their significance in hijacking interactions. The same ranking schema is used to identify chemical compounds that are candidate targets of hijacking by the pathogen, as well as chemical reactions of great interest, accompanied by their corresponding pathway details.

As a proof of concept, we studied *Mycobacterium tuberculosis* (*Mtb*), and we applied Hi-Jack to identify important metabolic pathway interactions with human. We focused on *Mtb* since it is an 'ancient' intracellular bacterium (meaning that tuberculosis is an old disease in terms of historical documentation) responsible for ~2.5 million deaths per year (Russell, 2001). *Mtb* is also known for its ability to manipulate phagosome to resist destruction by macrophages. It is a pathogen extremely resistant to drugs, and development of anti-tuberculosis drugs remains an important health problem worldwide. Hi-Jack analysis revealed that in human, pathways related to amino-acid metabolism, carbohydrate metabolism and metabolism of lipids are more likely to be hijacked by *Mtb*. Strong examples are fatty acid biosynthesis, pentose phosphate pathway, purine metabolism and arginine and proline metabolism. On the *Mtb* side, fatty acid biosynthesis, purine metabolism and biosynthesis of unsaturated fatty acids are ranked higher in our list of pathways. Moreover, we found chemical compounds that are frequent targets of hijacked chemical reactions in human. Hijacking of carbohydrates and lipids is in agreement with evidence coming from the literature, which indicates that *Mtb* under very hostile conditions in the host, switches to lipids as an alternative source of energy (Niederweis, 2008). The analysis also revealed that *Mtb* often hijacks compounds from human to produce metabolites linked to cell-wall construction.

The evidence for the Hi-Jack predictions mentioned above include the following principles: (i) proportion of pathways that contain computationally-predicted host–pathogen PPIs; (ii) proportion of pathways that contain genes required for bacterium optimal growth; (iii) tissue specificity of the hijacked pathways in the host and (iv) linkage of the predicted pathways with known drug targets. Our validation principles and some findings coming from literature confirmed the initial hypothesis and indicate that Hi-Jack has the potential to infer pathway-based host–pathogen interactions and identify metabolites likely to be hijacked by the pathogens. To our knowledge, this is the first study that utilizes metabolic pathway data from both host and pathogen metabolisms and ranks candidate interconnections at the pathway level based on their global significance in the phenomenon of hijacking vital metabolites. Also, the proposed validation principles have high potentials to be applied in future studies on host–pathogen interactions.

2 Methods

2.1 Metabolic network data

Although metabolic network data are available from many different sources (e.g. Rhea, BiGG, UniPathway, BioPath, BRENDA, SEED and Reactome), KEGG and MetaCyc are the largest curated databases of metabolic reactions and pathways containing significantly more reactions than all the other sources and having about the same level of chemical reactions in their pathways (Altman *et al.*, 2013). Hi-Jack was designed to work with metabolic-network data and annotations downloaded from KEGG database (Kanehisa and Goto 2000) using the KEGG API. We retrieved the list of human (hsa) metabolic pathways that contains 287 pathways. Similarly, we downloaded the list of *Mycobacterium tuberculosis* (mtu) pathways that contains 115 distinct pathways.

We retrieved KEGG chemical reactions that are mapped to metabolic pathways in the RPAIR format (Reactant Pairs). This data format decomposes chemical reactions into a set of single substrate-product pairs, together with the chemical structure transformation patterns characterized by the atom type changes at the reactant

center (Hattori *et al.*, 2003). A chemical reaction with multiple substrates/products is represented in the RPAIR format by multiple reactant pairs, each pair in the form of X → Y. As an example, chemical reaction beta-D-glucose → cellobiose + H_2O is decomposed into two RPAIRs: Beta-D-glucose → cellobiose and beta-D-glucose → H_2O. Those pairs that appear in KEGG metabolic pathway datasets are called the main pairs. Hi-Jack takes into account the main pairs that are involved in hijacked interactions between host and pathogen. In practice, within our deployed datasets, data are represented as triplets that contain pathway ids and pairs of compounds that interact in the corresponding pathways. RPAIR annotation has been successfully applied in several path-finding algorithms and metabolic-network-analysis frameworks (Faust *et al.*, 2009). Given the lists of RPAIR, we identified 5658 reactions in human and 3942 reactions in *Mtb*. Note that the raw human/*Mtb* datasets contain reversible reactions. However, in the data pre-processing step we filtered out RPAIR reversed directions that are not specified in any metabolic pathway.

A metabolite is considered not produced in the host (resp., pathogen) if either of the following two conditions is met: (i) The metabolite has no incoming edge in the host (resp., pathogen) metabolic pathway graphs. Obviously, such a metabolite can be assumed not produced in the host (resp., pathogen). (ii) The metabolite is in the beginning of a host (resp., pathogen) metabolic pathway, serving as 'starting point' of the associated biological process, and it is not produced by other host (resp., pathogen) pathways. Such a metabolite can start a cyclic process (e.g. the Calvin cycle). Even though such a metabolite typically has an incoming edge from the end of the cycle, some initial amount of it is needed to start the cycle. If it is not produced by any other pathways in the host (resp., pathogen), it has to be imported.

2.2 Hi-Jack implementation

Hi-Jack deploys a simple filtering algorithm that searches host and pathogen metabolic networks for RPAIRs that satisfy a metabolite-hijacking hypothesis. The default hijacking hypothesis is defined as follows: Suppose there is a reaction $C^1{}_p \rightarrow C^2{}_p$ in a pathogen, a reaction $C^1{}_h \rightarrow C^2{}_h$ in its host and the pathogen cannot generate the metabolite $C^1{}_p$; $C^1{}_p$ is not a frequent cofactor, specific inorganic molecules or basic energy carrier; then the pathogen will hijack reaction $C^1{}_h \rightarrow C^2{}_h$ in host, provided that $C^1{}_p$ equals $C^2{}_h$. In simple words, if the pathogen for various reasons—like hostile environment, host immune responses or drug effects—cannot create a substrate that is required by a reaction to generate a needed metabolite, it will obtain this substrate from its environment (host metabolism). This is a reasonable hypothesis for the following reason. Assuming $C^1{}_p$ is not made in the pathogen and is not imported, the reaction $C^1{}_p \rightarrow C^2{}_p$ is thus not useful in the pathogen. There is hence no evolutionary pressure to keep this reaction in the pathogen. It should eventually get 'optimized' out of the genome of the pathogen (especially in an ancient pathogen). That is, this reaction should not be present in the pathogen. Such an interconnected pair of reactions is called a 'hijacked reaction', and the compound that links these reactions is called 'point of hijacking' or 'target of hijacking'.

In a refined further analysis, a stricter hypothesis can be postulated that searches for reactions $C^1{}_h \rightarrow C^2{}_h$ in the host and $C^1{}_p \rightarrow C^2{}_p$ in the pathogen satisfying the additional condition that the $C^2{}_p$ compound is not produced in the host. This scenario simulates cases when the pathogen follows different underlying chemical processes from the host by utilizing non-homologous enzymes to produce essential metabolites for its own growth (e.g. *Mtb* has 55 enzymes non-homologous to human; Amir *et al.*, 2014).

A common problem when dealing with metabolic network data in the RPAIR format is the effect of 'currency compounds' (e.g. H_2O, ATP, NAD+, NADP+/NADPH, O_2, CO_2). These compounds, which are usually cofactors, small inorganic molecules or energy carriers, take part in many chemical processes and, thus, are characterized by much higher node degrees than other central metabolites in the metabolic network. We have plotted the node degree (viz. sum of in/out degree) of the compounds involved in our analysis in Supplementary Figures S1 and S2, illustrating the degree distribution of the compounds found in human and *Mtb*. The degree distribution is a power-law distribution: only a few compounds have very high degree. As a result, any path finding or filtering algorithm similar to Hi-Jack, applied on such metabolic network data, will use these highly connected compounds and thus many wrong or nonsense paths may be inferred.

To mitigate this problem, when filtering hijacked reactions and points of hijacking, Hi-Jack excludes paths involving 36 compounds that are highly connected. This strategy does not affect the mass balance of reactions and does not destroy the equilibrium of the metabolic pathways (Croes *et al.*, 2006). To illustrate this in a simple way, an example is provided in Figure 1. It becomes clear that without eliminating H_2O from the process, invalid connections are inferred that result in non-sense chemical procedures similar to the production of 5-aminolevulinate from beta-D-glucose.

After applying the hijacking hypothesis and identifying the points of hijacking, the algorithm returns a list of candidate interconnections. These candidate solutions are represented as RPAIRs in the host and pathogen, as well as the corresponding pathways in which these reactions occur. Note that one host pathway may have several hijacked RPAIRs that interconnect to one or more RPAIRs in the pathogen that belong to one or more pathways.

To rank the solutions, we used graph metrics and statistics collected from the metabolic network. In the simplest case, node in- and out-degrees are representative of the significance of each metabolite. Each hijacking match is scored using the formula:

$$HMS = \frac{\begin{array}{c}[in\,deg\,(C_h^1)+1]\times[in\,deg\,(C_h^2)+1]\\ \times[out\,deg\,(C_h^3)+1]\times[out\,deg\,(C_h^4)+1]\end{array}}{\begin{array}{c}[in\,deg\,(C_h^1)+1]\times[in\,deg\,(C_h^2)+1]\\ \times[out\,deg\,(C_h^3)+1]\times[out\,deg\,(C_h^4)+1]\end{array}}$$

Since all reactions in the deployed host and pathogen datasets are oriented and pathways are represented as directed graphs, in the *HijackingMatchingScore* (*HMS*) formula above, indeg(X) denotes the in-degree for the chemical compound X and outdeg(X) denotes the out-degree for X in the pathway.

In *HMS*, greater scores are assigned to a target of hijacking where the pathogen has fewer ways to produce it, the host has more ways to produce it, the pathogen has more ways to use it, and/or the host has fewer ways to use it. Variations of *HMS* utilize, instead of node degree, Page Rank or other metrics like centrality to weight compounds according to their significance in the flow of metabolites in the network; see the Supplementary Material and our web-repository. *HMS* is computed for every pair of interconnected reactions in host and pathogen using the hijacked metabolites. Then for every pathway, these scores are aggregated generating a global ranking score, *PathwayRank*, which reflects the likelihood of the underline pathway to interact with pathways in the other organism.

$$Pathway\,Rank = \frac{Hijacked\,\mathrm{Re}\,actions \times \sum_{k=1,m=1}^{\cup Hijacked\,\mathrm{Re}\,actions} HMS(Pk/Pm)}{\mathrm{Re}\,actions}$$

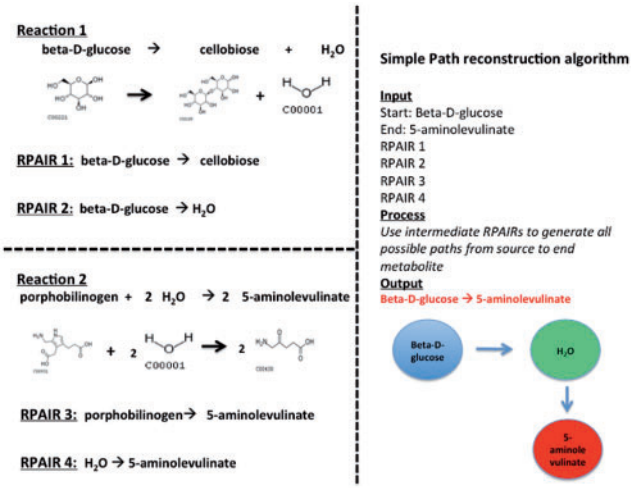
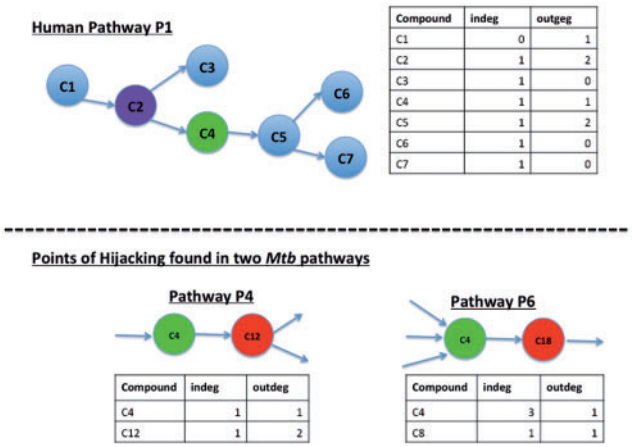


Fig. 1. The importance of eliminating currency compounds from metabolic network data applied on path finding algorithms. Without excluding H2O, the connection of RPAIR 2 with RPAIR 4 produces 5-aminolevulinate from beta-d-glucose using H2O as a linker. The inferred metabolic transformation is wrong since 5-aminolevulinate



Compound	indeg	outgeg
C1	0	1
C2	1	2
C3	1	0
C4	1	1
C5	1	2
C6	1	0
C7	1	0

Compound	indeg	outdeg
C4	1	1
C12	1	2

Compound	indeg	outdeg
C4	3	1
C8	1	1

Fig. 2. How Hi-Jack works. In the upper part, we have pathway P1 in human with 6 reactions. Given reaction C2→C4, we found in *Mtb* pathways two points of hijacking for compound C4 that are presented in the lower part with green color

Next we provide a concrete example that shows how Hi-Jack works: Let us assume pathway P1 in human contains 6 reactions as depicted in Figure 2. For simplicity we have also pre-computed in- and out-degrees for all of the compounds found in the pathway. Let us now focus on reaction C2 → C4 in P1. If we assume that for some reason the pathogen cannot produce compound C4, then C4 becomes a hijacking point for the pathogen. After searching in the pathogen's metabolism, we identify two reactions that require C4 to produce their metabolites. Specifically, in Figure 2, reaction C4 → C12 in the pathogen's pathway P4 and C4 → C8 in the pathogen's pathway P6 require substrate C4 to generate their products. In Figure 2, we have also pre-computed in and out degrees for all the compounds on the pathogen's side. We get $HMS(P1/P4) = 1$ and $HMS(P1/P6) = 0.33$. These scores represent the local 'strength' of interconnections between pathways P1/P4 and P1/P6. It follows that $PathwayRank(P1) = 0.22$, since it has 1 hijacked reaction out of 6 reactions that P1 contains.

The idea behind *PathwayRank* is relatively simple, indicating that longer pathways (in terms of number of reactions) with many hijacked reactions are more likely to interact with pathogen pathways. Following the previous example, let us assume pathway P2 in the host has exactly the same length 6 as pathway P1. Let us also assume that these pathways achieve the same sum of *HMS* (1.33 in our example) but P2 has 2 hijacked reactions instead of 1 in P1. The *PathwayRank* of P1 is 0.22 and P2 is 0.44, meaning that pathway P2—which has more targets for hijacking for a particular number of reactions—is more likely to interact with the pathogen.

Once we rank the pathways for both human and *Mtb*, it is easy to parse the results and identify interesting metabolites based on the frequency of occurrences in hijacked interactions. Note that, since *HMS* and *PathwayRank* are specific to host/pathogen data (e.g. in-/out-degree of a chemical compound depends on the organism's metabolism), Hi-Jack is capable of reflecting specific metabolic characteristics of different hosts and pathogens.

3 Results

3.1 Principles for inferring host–pathogen interactions

A reasonable way of validating hijacking interactions is to test whether the hijacked metabolites are capable of being transported out of the host cells and imported into the pathogen. However, studying pathogen membrane permeability is not straightforward and validation criteria based on nutrients transportation principles are not easily applicable. The reason is that principles that govern nutrients acquisition by pathogens, especially in mycobacteria, are not completely understood (Niederweis, 2008). In particular: (i) membranes and cell-wall systems of pathogens (e.g. cell envelope in *Mtb*) are very complex systems and experimental procedures not easily designed; (ii) identity and function of very few transporter proteins is known for *Mtb*; (iii) specificity of the transportation mechanisms (and the nutrients availability) depends on the location of *Mtb* inside the host (e.g. phagosome, macrophages or dendritic cells); (iv) uptaking processes often are dominated by direct diffusion or porin-mediated diffusion which are again not understood completely. Consequently, if a specific metabolite is not currently known to be transportable across membranes, there is insufficient evidence against this specific compound being hijacked and transported under some specific cellular conditions using specific mechanisms that are not yet known. That is, validation based on nutrient acquisition principles is overly conservative.

As there is no gold-standard baseline for validating pathway-based inference of host–pathogen interactions, some principles of pathway interconnections between host and pathogen are proposed in this section. This kind of analysis is essential for developing better prediction approaches and can be applied to subsequent studies. These principles go beyond classical performance evaluation metrics since they represent different levels of confidence on whether Hi-Jack's predictions are meaningful. The principles are:

1. *Proportion of host and pathogen pathways containing host–pathogen PPIs:* High-confidence computationally-predicted host–pathogen PPIs suggests interactions between host and pathogen. A metabolic pathway which has many proteins involved in such PPIs therefore implies metabolic interference between the two organisms, and further suggests co-localization of proteins and sub-cellular localization of host and pathogen. We obtained computationally-predicted human-*Mtb* PPIs from Zhou *et al.* (2013) and Zhou *et al.* (2014). This dataset is an integrative collection of 1097 PPIs for the two organisms generated by a homology-based approach and a novel accurate approach for predicting host–pathogen PPIs. For every predicted

host (resp., pathogen) pathway, we computed the number of proteins in that pathway that have PPIs with the pathogen (resp., host). Note that we performed this analysis twice, for human and *Mtb* pathways.

2. *Proportion of pathways that contain genes required for survival and growth in the pathogen*: Essential genes are critical for the survival of an organism. Providing a more precise definition of essential genes is not straightforward since different critical gene sets depend on the circumstances and the environmental conditions where the organism lives. Defining these gene sets becomes even more complicated when dealing with pathogens since pathogens have some cell-autonomous functions and, at the same time, they rely on host organisms. In *Mtb*, several experimental studies use mainly transposon site hybridization experiments (TraSH) to identify sets of essential genes. For example Rengarajan *et al.* (2005) reports three gene groups required by different stages of *Mtb* survival in macrophages. However, the fact that the experimental conditions described in these models are very specific is a limiting factor for generalization. To describe in a more general way a pathogen's life cycle, in this study we focus on a list of 614 'universal' essential genes required for optimal *Mtb* growth. These genes derived from TraSH experiments (Sassetti *et al.*, 2003) are archived in the DEG database (Luo *et al.*, 2014). To use this information, we compute the proportion of identified essential genes that are mapped to different pathways assuming that pathways with more essential genes are characterized as 'essential pathways' (Sassetti *et al*, 2003), which incorporate functions and mechanisms for producing or uptaking critical nutrients for growth.
3. *Tissue-specificity of the host pathways*: Obligate intracellular pathogens are solely dependent on the host cell. Therefore, a pathogen and its host–organism interaction targets are more likely to be co-localized in specific tissues. In the case of *Mtb*, we expect the predicted host pathways to be activated in tissues accessible by the bacterium, like lung tissue and lymph nodes. To perform this analysis, we retrieved from the TIGER database (Liu *et al.*, 2008) a list of expressed tissues for every gene. From them, we extracted the genes that are expressed in lungs and lymph nodes. Then for every predicted pathway, we collected the underlying genes and we mapped them to the list of genes that are expressed in lung tissues and lymph nodes. To decide whether a pathway is activated in a tissue we required a predefined number (e.g. 75%) of the mapped genes to be at least expressed in these tissues. Note that we searched for sufficient numbers of expressed genes and we did not decide based on the actual expression levels. If we had decided based solely on expression levels, a few overexpressed genes might dominate our estimations and we might conclude misleading results about pathways activation in these specific tissues. As information of expression in tissues is incomplete, the validation principle we applied is necessarily quite loose.
4. *Linkage of predicted pathways with known drug targets:* Development of anti-tuberculosis drugs is a cutting-edge research area. High-throughput screening of chemical compounds and computationally predicted drug targets are alternative approaches that complement each other for this difficult task. Supporting Hi-Jack's pathway predictions with known or predicted drug targets gives an indirect way to validate the results. Drugs are bioactive molecules that target critical functions for pathogen survival/growth. On the pathogen side, pathways that contain many drug targets are more likely to take part in hijacking processes. Being a drug target suggests that specific reactions and mechanisms related to 'molecular exchange' are present and activated in the pathogen. In other words, if a pathogen's pathway allows a drug to be transported inside membranes, it implies a direct host/pathogen interaction in that pathway. We applied this analysis on the outcome of a large-scale high-throughput screen that resulted in a catalog of 776 active chemical compound structures against tuberculosis (Martinez-Jimenez *et al.*, 2013). Alternatively, results can be obtained using predicted drug targets for *Mtb* from Anishetty *et al.* (2005).
5. *Manual annotation of genes (if available), chemical compounds and pathways:* Finally, knowledge coming from experimental procedures and evidence from the literature about genes with significant functional roles in tuberculosis further support Hi-Jack's predictions. Such examples are known *Mtb* virulent genes, known transporter proteins, and processes related to core cellular functions like cell-wall synthesis and mycolic-acid biogenesis.

3.2 Applying Hi-Jack to *Mtb* metabolism

Here, we studied *Mtb* metabolism and report results obtained by Hi-Jack. Overall, Hi-Jack identified hijacking phenomena between 36 human pathways and 46 *Mtb* pathways. We present in Supplementary Tables S1 and S2 the top-ranked pathways for both organisms. These human top-ranked pathways have on average 17.9 hijacked reactions per pathway with the pathogen. In contrast the rest lower-ranked pathways have 7.3 hijacked reactions per pathway. Next, we studied further the KEGG classes that categorize pathways according to their biological role; cf. histograms in Figures 3 and 4.

We found that, in human, pathways that belong to carbohydrate metabolism achieve the highest score, followed by pathways pertaining to amino-acid metabolism, and pathways related to metabolism of lipids. These findings are in agreement with the few known nutrient-acquisition principles for mycobacteria, which identify transporters of carbohydrates and lipids as well as up-taking mechanism for amino acids (Niederweis, 2008). On the *Mtb* side, the top-ranked pathways more often belong to carbohydrate metabolism and amino-acid metabolism, and the list of pathways is quite similar to human pathways but enhanced with pathways related to biosynthesis of unsaturated fatty acids and glycerophospholipids indicating that the pathogen requires specific compounds for its growth. Next, we retrieved compounds that are frequent hijacking points from host, and we report in Supplementary Table S3 the top-ten chemical compounds. Overall, we observe that the most-frequent targets from human are lipids and compounds or molecules that take part in fatty acid biosynthesis. Supplementary Table S4 illustrates the most frequently RPAIRS that are hijacked by the bacterium. Based on this set of interconnected RPAIRs we report candidate interactions between human and *Mtb* pathways. To visualize the interconnections between host/pathogen pathways the bipartite graph presented in Supplementary Figure S3 shows the *Mtb* pathways that are connected with the top-3 ranked pathways in human. Interesting is the fact that human glycine, serine and threonine metabolism is potentially linked with 17 *Mtb* pathways, human pentose phosphate metabolism with 13 pathways in *Mtb* and glycerophospolipid metabolism with 11 *Mtb* pathways.

Next, using principles presented in the previous section, we supported our results by computational evidence. Figure 5 presents the proportion of pathways in human that contain host–pathogen–PPIs. As a control study, we partitioned initially the retrieved pathways in two groups, one for those that hijacking occurs, and one for all the

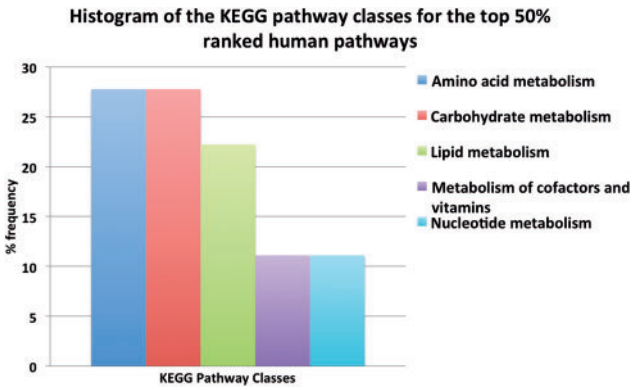


Fig. 3. KEGG classes for top 50% ranked pathways retrieved in Human

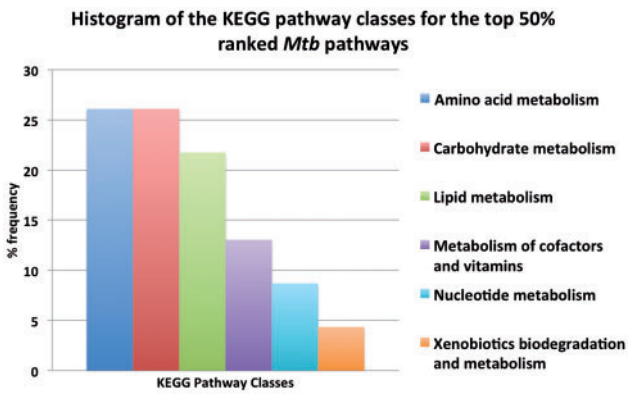


Fig. 4. KEGG classes for top 50% ranked pathways retrieved in *Mtb*

other pathways where there is no hijacking score and thus hijacking does not occur (Group C). Then we partitioned the group of hijacked pathways in two equal size subsets: the first 50% of pathways correspond to the top-ranked pathways (Group A) and the rest 50% to the lower-ranked pathways (Group B). For each group, we computed the proportion of human pathways in which human-*Mtb* PPIs occur. We observe that only 20.4% of pathways that belong to Group C contain human-pathogen PPIs. Group B, which is the group of lower rank, has lower numbers of PPIs compared to Group A of higher ranks. Supplementary Table 5 presents more details about pathways in group A, as well as the exact number of host–pathogen PPIs per pathway. Similarly, Figure 6 presents results for human-*Mtb*-PPI enrichment for the *Mtb* side. Following the same control study as before, we generated three distinct groups, one that corresponds to the top-ranked pathways for hijacking, one that corresponds to the lower-ranked pathways for hijacking and one for all the other pathways that there is no hijacking phenomena. In this case, it is again clear that the groups of hijacking (Group A and B) achieve much higher numbers of PPIs in the pathogen. Again the unranked pathways (Group C), which correspond to the 'no-hijacking' case, achieve a much lower 27.1% support from PPIs. Supplementary Table 6 presents more details about the group of top-ranked pathways in *Mtb*, as well as the exact number of host–pathogen PPIs per pathway. It becomes apparent that the top-ranked pathways based on the hijacking score have higher number of host–pathogen PPIs.

Regarding the set of essential genes required for optimal *Mtb* growth, Figure 7 presents the proportion of pathways that contain such genes in different groups according to their relative ranking. Group C, corresponding to the 'no-hijacking' case, achieves 32.1% support from essential genes, whereas Group A and B that correspond to hijacked pathways have much higher support by essential genes (86.9 and 78.2% resp.). Supplementary Table S7 provides more detailed results for *Mtb* pathways belonging to the top-ranked Group A.

As to the tissue specificity of the top-ranked pathways, we found that all of the reported pathways have many genes (but not all) expressed in lungs and lymph nodes. In lungs pentose phosphate pathway and fatty acid biosyntesis have the highest proportion of expressed genes (82.1 and 83.3% resp.). In lymph nodes arginine and proline metabolism has the highest number of expressed genes followed by pentose phosphate metabolism. Overall we observe that the number of expressed genes in the reported pathways in lymph nodes is much lower than that in lungs. We thus infer that some candidate hijacked pathways, such as alpha-linolenic acid metabolism, that has few expressed genes in both tissues appear not promising candidates for hijacking phenomena. Figure 8 presents the detailed results.

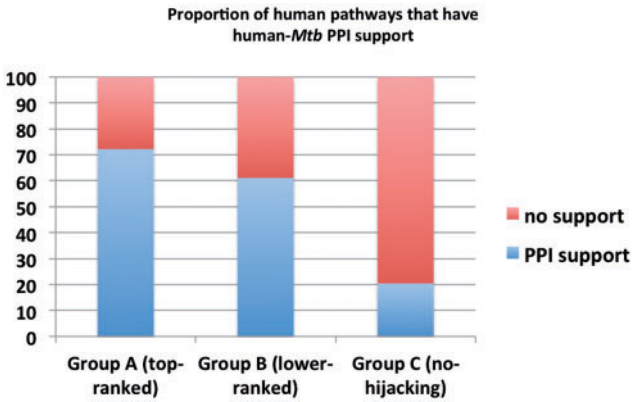


Fig. 5. Proportion (%) of pathways in human grouped in three categories according to Hi-Jack's scoring scheme that have support by host–pathogen PPIs

Next, we supported our results with predicted drug targets. Figure 9 presents the proportion of pathways containing known drug targets obtained from the TCAMSTP dataset (Martinez-Jimenez *et al.*, 2013) in Group A, B and C of human pathways. We found that pathways involved in hijacking interactions have greater number of drug targets compared to those that do not contain hijacked RPAIRS. This confirms our initial hypothesis that pathways involved in hijacking should have more mechanisms for molecular exchange activated. Supplementary Table S8 presents more details for our top-ranked predicted pathways that are enriched in drug targets in *Mtb*.

As some of our indirect validation principles appear lenient and the differences in the proportion of pathways supported by host–pathogen PPIs, essential genes and drug targets sometimes look relative small, we computed the per-pathway average number of host–pathogen PPIs, essential genes and drug targets that corresponds to groups of different rank. The results in Figure 10 convincingly show that groups of higher rank achieve higher average support by PPIs, essential genes for growth and drug targets.

We also computed an additional 'strict' performance metric, which estimates the proportion of pathways in the pathogen that are supported by essential genes for growth and drug targets and PPIs with the host. These results are presented in Figure 11.

4 Conclusion

We proposed Hi-Jack, a novel computational framework for inferring metabolic-pathway associations between host and pathogen.

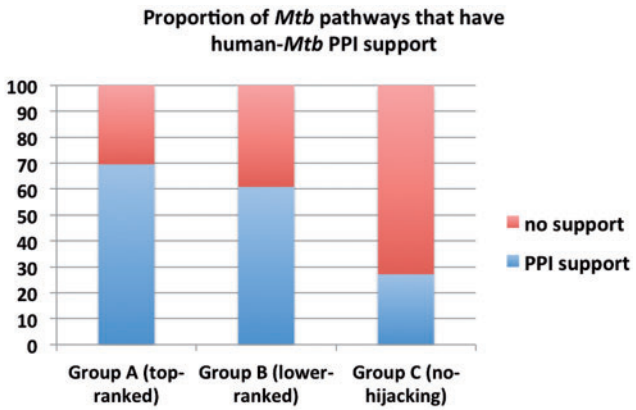


Fig. 6. Proportion (%) of pathways in *Mtb* grouped in three categories according to Hi-Jack's scoring scheme that have support by host–pathogen PPIs

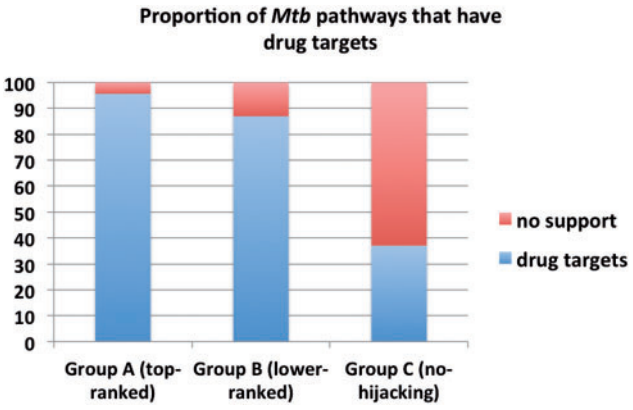


Fig. 9. Proportion (%) of pathways in *Mtb* grouped in three categories according to Hi-Jack's scoring scheme that are linked with known drug targets

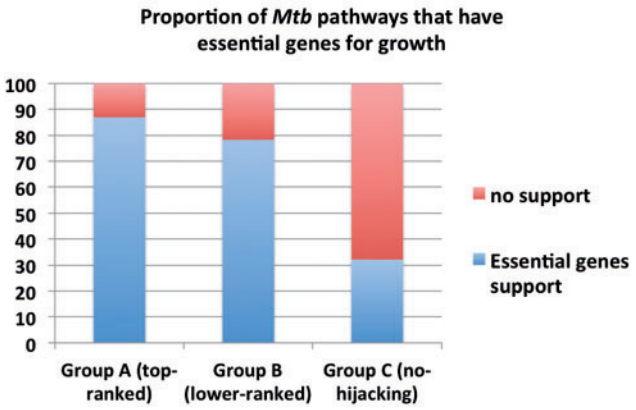


Fig. 7. Proportion (%) of pathways in *Mtb* grouped in three categories according to Hi-Jack's scoring scheme that have support by essential genes for optimal growth

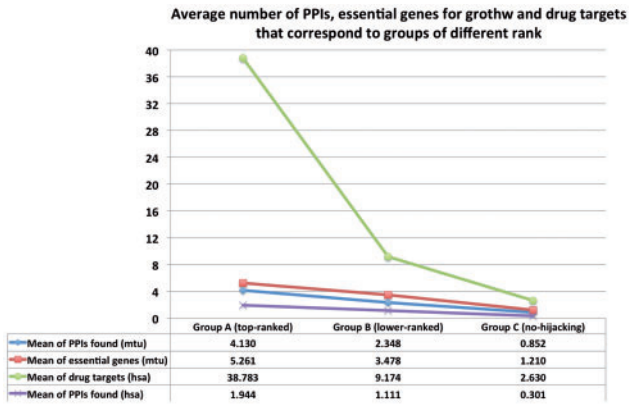


Fig. 10. Average number of PPIs, essential genes and drug targets that correspond to groups of different rank

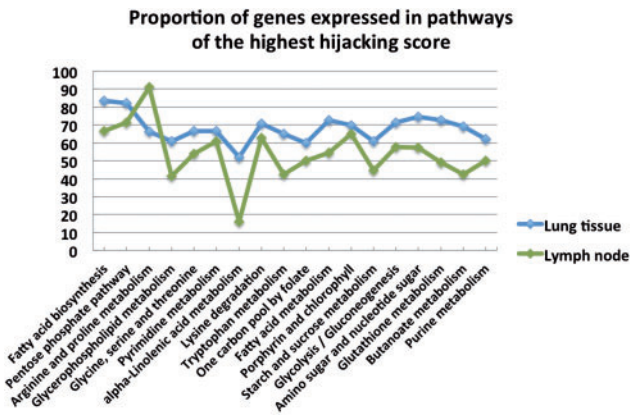


Fig. 8. Proportion (%) of expressed genes in lung tissue and lymph nodes for all the pathways that belong to the top-ranked group based on the hijacking score

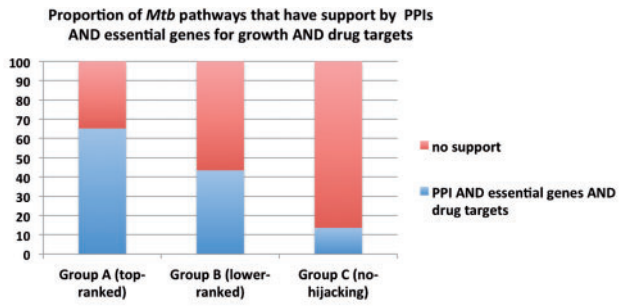


Fig. 11. Proportion of pathways in *Mtb* that have host–pathogen PPIs ANPPIs AND essential genes for growth and drug targets

The framework defines a metabolite-hijacking hypothesis and deploys a simple filtering algorithm to rank host–pathogen interaction pathways based on the idea of metabolite acquisition by the pathogen. We also proposed basic principles of hijacking interactions that support our results by different levels of confidence, applicable in future studies.

As a case study, we focused on *Mycobacterium tuberculosis*, and we revealed possible mechanisms that support pathogen growth and survival. We found that amino-acid metabolism, carbohydrate metabolism, and pathways related to metabolism of lipids are top-ranked pathways for both human and *Mtb*. For *Mtb*, frequent pathways are also related to lipid and glycerophospholipid metabolism, as well as biosynthesis of fatty acids. In addition, our predictions are supported by computational evidence, viz. host–pathogen–PPI support, significant number of essential genes for optimal growth and known drug targets. Surprisingly, 91% of all ranked pathways in *Mtb* are supported by at least one drug target. In addition,

tissue-specificity analysis for lung tissue revealed that the top-ranked human pathways have a lot of expressed genes and thus, hosting hijacking phenomena is possible.

Nonetheless, there is still space for improvements: (i) Integrating metabolic network data from both KEGG and MetaCyc may improve the quality of the results. (ii) We are planning to study more pathogens like *M. leprae*, *M. ulcerans* and *H. influenzae* which appear to have very different growth requirements, meaning that pathways required for optimal growth, enzymes and transportation mechanisms are different from those in *Mtb*. (iii) Refining the hijacking hypothesis by searching for compounds that satisfy a predefined similarity threshold as defined by a compound similarity metric like Tanimoto coefficient is very interesting extension. d/ From the software development perspective, the implementation of a more general algorithm that identifies hijacking points given an arbitrary number of steps away from the targeted reaction is an interesting idea for the future.

Acknowledgements

We thank Hufeng Zhou for providing the human-*Mtb* PPI datasets and Kevin Lim for his helpful comments.

Funding

This research has been funded by KAUST Base Research Funds to P.K.

Conflict of Interest: none declared.

References

Altman,T. *et al.* (2013) A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**, 112.

Amir,A. *et al.* (2014) Mycobacterium tuberculosis H37Rv: in silico drug targets identification by metabolic pathways analysis. *Int. J. Evol. Biol.*, **2014**, 284170.

Anishetty,S. *et al.* (2005) Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis. *Comput. Biol. Chem.*, **29**, 368–378.

Chou,C.H. *et al.* (2009) FMM: a web server for metabolic pathway reconstruction and comparative analysis. *Nucleic Acids Res.*, **37**, W129–W134.

Chueh,T.H. and Lu,H.H. (2012) Inference of biological pathway from gene expression profiles by time delay boolean networks. *PloS One*, **7**, e42095.

Croes,D. *et al.* (2005) Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res.*, **33**, W326–W330.

Croes,D. *et al.* (2006) Inferring meaningful pathways in weighted metabolic networks. *J. Mol. Biol.*, **356**, 222–236.

Faust,K. *et al.* (2009) Metabolic pathfinding using RPAIR annotation. *J. Mol. Biol.*, **388**, 390–414.

Hattori,M. *et al.* (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Karp,P.D. *et al.* (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res.*, **28**, 56–59.

Lemer,C., *et al.* (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res.*, **32**, D443–D448.

Lim,K. and Wong,L. (2012) CMPF: class-switching minimized pathfinding in metabolic networks. *BMC Bioinformatics*, **13**, S17.

Liu,X. *et al.* (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.

Luo,H. *et al.* (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.*, **42**, D574–D580.

Martinez-Jimenez,F. *et al.* (2013) Target prediction for an open access set of compounds active against *Mycobacterium tuberculosis*. *PLoS Comput. Biol.*, **9**, e1003253.

Niederweis,M. (2008) Nutrient acquisition by mycobacteria. *Microbiology (Reading, England)*, **154**, 679–692.

Ourfali,O. *et al.* (2007) SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics (Oxford, England)*, **23**, i359–i366.

Pieters,J. and Gatfield,J. (2002) Hijacking the host: survival of pathogenic mycobacteria inside macrophages. *Trends Microbiol.*, **10**, 142–146.

Rengarajan,J., Bloom,B.R. and Rubin,E.J. (2005) Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc. Natl. Acad. Sci. USA*, **102**, 8327–8332.

Russell,D.G. (2001) *Mycobacterium tuberculosis*: here today, and here tomorrow, Nature reviews. *Mol. Cell Biol.*, **2**, 569–577.

Sassetti,C.M., Boyd,D.H. and Rubin,E.J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.*, **48**, 77–84.

Singh,Y. *et al.* (2013) *Mycobacterium tuberculosis* controls microRNA-99b (miR-99b) expression in infected murine dendritic cells to modulate host immunity. *J. Biol. Chem.*, **288**, 5056–5061.

Xia,D. *et al.* (2011) MRSD: a web server for metabolic route search and design. *Bioinformatics (Oxford, England)*, **27**, 1581–1582.

Zhou,H., Jin,J. and Wong,L. (2013) Progress in computational studies of host–pathogen interactions. *J. Bioinf. Comput. Biol.*, **11**, 1230001.

Zhou,H. *et al.* (2013) Stringent DDI-based prediction of *H. sapiens–M. tuberculosis* H37Rv protein-protein interactions. *BMC Syst. Biol.*, **7**, S6.

Zhou,H. *et al.* (2014) Stringent homology-based prediction of *H. sapiens–M. tuberculosis* H37Rv protein–protein interactions. *Biol. Direct*, **9**, 5.