Uncovering microbial duality within human microbiomes: a novel algorithm for the analysis of host-pathogen interactions

Edgar D. Coelho, Joel P. Arrais, and José Luís Oliveira

Abstract— **Microbial species thrive within human hosts by establishing complex associations between themselves and the host. Even though species diversity can be measured (alpha- and beta-diversity), a methodology to estimate the impact of microorganisms in human pathways is still lacking. In this work we propose a computational approach to estimate which human pathways are targeted the most by microorganisms, while also identifying which microorganisms are prominent in this targeting. Our results were consistent with literature evidence, and thus we propose this methodology as a new prospective approach to be used for screening potentially impacted pathways.**

I. Introduction

The human organism is an ecosystem harboring about ten times more microbial cells than its own somatic and germ cells. This was the *motto* behind The Human Microbiome Project (HMP), which was created in attempt to establish new parameters for health and disease [1]. The development and introduction of next generation sequencing techniques propelled the identification of the human microbiome (the sum of our microbial symbionts), which in turn allowed the characterization of microbial species by their gene products.

Microorganisms must establish protein-protein interactions (PPIs) with the host to be able to thrive [2]. Since each PPI complex has a defined output in physiological context, one can hypothesize that if a human protein is interacting with a microbial protein, the former will not be available to establish a complex with its physiological interacting counterpart [reviewed in 3]. This could negatively impact pathways of the host (*e.g.*, [4]), but positive impact is also a possibility (*e.g.,* [5]).

Even though the diversity of microbial communities, within and across samples (alpha-diversity and beta diversity, respectively) for various body sites, has already been described [6], we did not find any approach to measure the potential positive or negative impact of microorganisms in human pathways.

*Research supported by Fundação para a Ciência e Tecnologia under Grant SFRH/BD/86343/2012 (to E.D.C.) and by the project NeuroPath (CENTRO-07-ST24-FEDER-002034), co-funded by QREN, "Mais Centro" program and EU.

E. D. Coelho and J. L. Oliveira are with the Department of Electronics, Telecommunications and Informatics, Institute of Electronics and Telematics Engineering of Aveiro, University of Aveiro, Portugal (e-mails: eduarte@ua.pt; jlo@ua.pt).

J. P. Arrais is with the Department of Informatics Engineering, Centre for Informatics and Systems, University of Coimbra, Portugal (phone: (+351)2397900; e-mail: jpa@dei.uc.pt).

In this work we propose an algorithm to estimate the impact of microorganisms in human pathways. We tested our approach in ten different microbiomes, obtaining results consistent with the literature.

II. Methods

A. Data Collection

Human gene data were obtained from the Human Protein Atlas [7], and filtered according to its expression level – only genes considered as being highly expressed in a certain body site were taken into account. Microbial gene data were collected from HMP, including abundance scores for each sample.

Sample extraction site data were also collected to establish the baseline microbiomes for each body site. The UniProt accession IDs of each possible gene product from the obtained gene pool. Each microbial and human protein acquired this way was sorted into the body site where their respective precursor genes were originally found. A summary of the final ten microbiomes is shown in Table 1.

After successful identification of the protein pool possibly present in each microbiome, we predicted the protein interactions most likely to occur in those sites (*i.e.,* the interactome for each microbiome). This task was fulfilled by using an adaptation of a previously described methodology [8] by the same authors.

Human pathway data was retrieved from the Reactome database [9]. This includes the name of the pathway and of all the proteins present in the respective pathway.

TABLE I. Number of proteins in each body site

	Protein Origin	
Body site	***Human***	***Microbial***
Appendix	16,952	174,293
Cervix	12,568	159,056
Colon	26,505	174,293
Esophagus	12,962	158,126
Nasopharynx	18,272	159,799
Oral mucosa	9,003	180,033
Rectum	22,916	174,293
Skin	10,858	176,186
Small Intestine	24,082	174,293
Vagina	8,372	96,561

All data from the Human Protein Atlas, the HMP, and the Reactome database used in this work were downloaded in June 2014.

B. PPI Prediction Within Microbiomes

The PPI prediction approach used in this work was based in [8]. This methodology focused on five feature clusters constructed from: literature; primary protein sequence information; orthologous profiles; biological process similarity, and; enriched conserved domain pairs. These clusters were optimized for the prediction of PPIs in proteins with a high degree of annotation. The most general and descriptive feature cluster was based on protein sequence data, and since the sequences of all proteins in our datasets are known, this means our feature cluster has 100% data coverage. For this reason, we opted to optimize the original algorithm for sequence -based PPI predictions.

The final classifier was tested using 5-fold cross validation against a dataset of high-quality, experimentally identified, and manually curated protein interactions (approximately 20,000 PPIs). These data were obtained from the BioGRID database [10] in March, 2014. PPI predictions in this work were obtained with an average accuracy of 0.9.

B. Data clustering

The final pre-processing step involved clustering all organisms present by their respective phyla, so we can understand which phyla impacts each human pathway the most. We opted for a phylum-level analysis since we are analyzing ten different microbiomes. Indeed, this level of taxonomic resolution has been used in a recent related work [6]. Since avian, bovine, and porcine proteins were identified in our protein pool, we removed all proteins under the *Animalia* kingdom, as we believe many of these proteins could be considered sample contaminants present by the time of sample extraction.

Even though all the proteins present in our data set could be clustered in 24 different phyla, we chose to only analyze the impact of *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, *Firmicutes*, *Fusobacteria*, *Proteobacteria*, *Spirochaetes*, and *Tenericutes*, as in these phyla are included the most notable human bacterial pathogens.

In addition, from a list of 201 human pathways that were targeted for interaction by at least one microbial protein, we chose to only analyze the ten pathways that were most targeted for PPIs by microbial proteins. The list of human pathways analyzed in this work is shown in Table 2.

All the proteins that integrated any of the 201 pathways were cross-linked with the human proteins present in our dataset, for each body site. This resulted in a list that linked each event to the human proteins that were expressed in each region, together with the total number of proteins that participated in each event. In addition, sub-pathways were grouped by their respective parent pathways. As such, our analysis will fall on the apoptosis, cell cycle, DNA repair, gene expression, and metabolism pathways.

C. Calculating the Microbial Impact in Human Pathways

The main hypothesis behind the development of our algorithm was simple: if a microbial protein interacts with a human protein that is part of a pathway, one can expect a decrease in pathway activity downstream of the targeted protein, since the targeted protein is transiently occupied. Of course, this will depend on the abundance of both targeting (microbial) and target (human) proteins. With this in mind, we developed the following equation:

$$\theta_e = \frac{1}{H_e}\sum_{i=1}^{h_e}\frac{1}{m_i}\sum_{j=1}^{m_i}\omega_j \, . \qquad (1)$$

In brief, for the analysis of the level of impact (θ) induced to each event (e), we averaged the sum of the abundance (ω_j) of each microbial protein (j) by the total number of microbial proteins (m_i), which interact positively with each human protein (i). The value of m_i represents the number of the most abundant microbial proteins which interact with each human protein of the total number of affected human proteins (h_e). This value is then weighted by the total number of human proteins (H_e) that participated in the event.

TABLE II. PATHWAYS TARGETED THE MOST BY MICROBIAL PROTEINS

Pathway Name	Parent Pathway	Reactome ID
Translesion synthesis by POLH	DNA repair	110320
Insulin effects increased synthesis of D-Xylulose-5-P	Metabolism	163754
PAOs oxidise polyamines to amines	Metabolism	141334
Resolution of D-loop structures through Holliday junction intermediates	DNA repair	75228
Resolution of D-loop structures	DNA repair	75149
TET 1, 2, 3 and TDG demethylate DNA	Gene express.	5221030
Processing of DNA double-strand break ends	DNA repair	83626
Transport of the SLBP independent mature mRNA	Gene express.	159227
Progressive synthesis on the C-strand of the telomere	Cell cycle	174414
Activation and oligomerization of BAK protein	Apoptosis	111452

For instance, when the total number of human proteins in any pathway p increases, one can expect a low impact θ_p. This can be caused by possible compensatory mechanisms in pathway p, where an isoform of the targeted protein plays the role of the latter in the pathway, or even due to up-regulation of gene expression as part of a feedback mechanism [11, 12]. Finally, we proceeded to the quantitative estimation of the impact of each phylum in human pathways.

III. RESULTS AND DISCUSSION

A. Impact mapping

Fig. 1 presents the results of our proposed approach. Regions with higher microbial abundance are easily spotted, as they have the highest impact values (the appendix, colon, oral mucosa, rectum, and small intestine). The other five body sites not only present lower impact values, but also seem to have dominant phyla.

In comparison with the appendix, colon, oral mucosa, rectum, and small intestine, these sites seem to possess reduced species diversity, as more than one phyla can be observed having an impact in the pathways of the former sites.

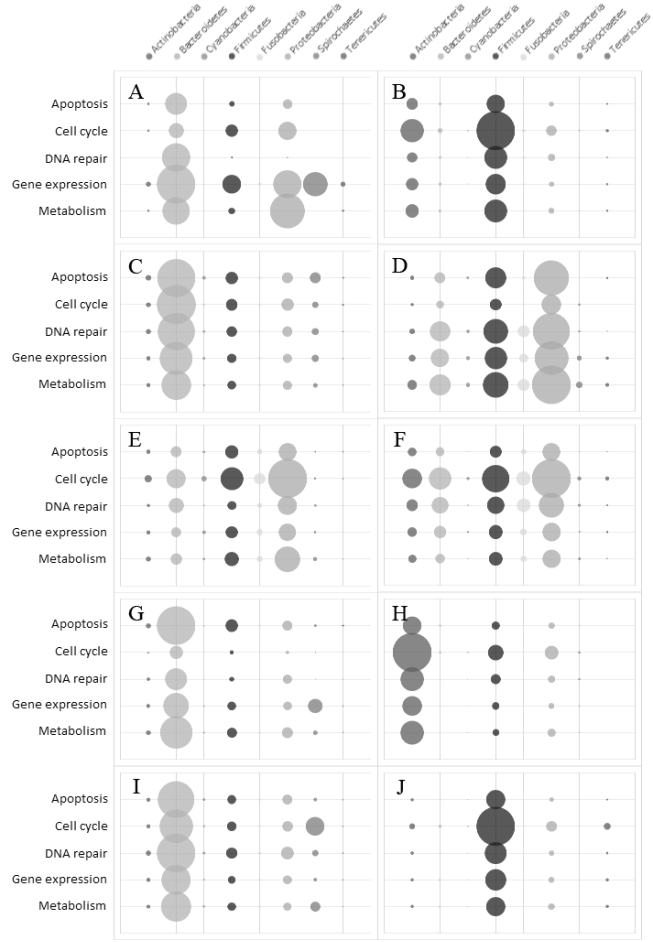

Figure 1. Representation of microbial impact in the pathways of ten different microbiomes. Circle radius represents impact value. *(A –Appendix; B – Cervix/Uterus; C – Colon; D – Esophagus; E – Nasopharynx; F – Oral cavity; G – Rectum; H – Skin; I – Small intestine; J - Vagina).*

Nonetheless, these results depend heavily on the extent of the knowledge of microbial genomes (and proteomes). That is, fully uncovered microbial proteomes will allow to optimize the prediction of protein interactions to a whole new level, enhancing all computational approaches based on such predictions.

B. Characterization of impact maps

We found that the phyla impacting the appendix, colon, rectum and small intestine (Fig. 1, Panels A, C, G, and I) the most were *Bacteroidetes*, *Firmicutes,* and *Proteobacteria*. In addition, when compared with the other six microbiomes, these regions suffered the greatest microbial impact. Literature evidence shows that these phyla make up for almost all of the species in the gut microbiota [13]. Our results also show that the main human pathways being targeted by microbial proteins are apoptosis, gene expression, and metabolism. Indeed, we found evidence that the gut microbiota has been implicated in metabolic disturbances, such as obesity, outflow during energy production, and even the regulation of host genes that control metabolic processes [14, 15].

Vaginal and cervical/uterine (Fig. 1, Panels B and J) regions also show a very similar pattern. Both regions were dominated by *Firmicutes*, consistently with literature evidence [16]. In this microbiome, we identified cell cycle, metabolism and DNA repair as the pathways being impacted the most. Such results also have literature support, as some *Lactobacilli* species were shown to decelerate cell cycle progression in the cervix, while a single *Lactobacillus* species was found to transiently accelerate cellular division in the human host [17].

The esophagus (Fig. 1, Panel D) possessed great microbial diversity, being mostly impacted by species under the *Proteobacteria*, *Firmicutes*, and *Bacteroidetes* phyla. The most affected pathways in the esophagus were metabolism and DNA repair, which according to existing literature, can occur once bacterial species activate their damage repair mechanisms. This will in turn induce DNA damage in the host, thus activating its DNA repair mechanisms [18, 19].

The oral cavity and nasopharynx (Fig. 1, Panels E and F) presented similar patterns, being impacted the most by Proteobacteria, *Firmicutes*, and *Bacteroidetes*. In this microbiome cell cycle was the pathway being targeted the most by microbial proteins. This may be explained by the ability of species like *Aggregatibacter actinomycetemcomitans* (under the *Proteobacteria* phylum) and *Porphyromonas gingivalis* (*Firmicutes* phylum), to induce cell cycle arrest, leading to the downregulation of host genes involved in cell cycle regulation [20, 21].

Lastly, our findings showed that *Actinobacteria*, *Firmicutes* and *Proteobacteria* were the phyla with most impact on the skin cellular events (Fig. 1, Panel H), which is consistent with the literature [22, 23]. We found that the pathways being targeted the most in this microbiome were cell cycle, metabolism and DNA repair. These are most likely explained by the presence of bacterial DNA in the host dermis, even when this DNA originates from dead bacteria [24, 25]. However, these studies were highly suggestive that bacterial cells can indeed colonize the dermis [26].

IV. Conclusion

We believe one cannot consider the human being as an individual organism, but rather as a multitude. Indeed, it is clear that not only human cells and proteins should be considered when it comes to systems biology. Humans as an ecosystem are constantly evolving, and various intrinsic (*e.g.*, ethnicity, age) and extrinsic factors (*e.g.*, habits, physical activity, diet) contribute to the development and establishment of the microbiomes of each individual. In addition, the microbiomes of healthy individuals for the same body sites can be dissimilar, rendering the task of identifying baseline microbiomes extremely challenging.

We should also emphasize that knowing which proteins in human pathways are being targeted by microbial proteins may provide new grounds for the research of specific microbial pathogenesis mechanisms. This knowledge would prove of utmost interest to the pharmacological and clinical

research areas, as it could be used for the development of new drugs and diagnostic methods.

The proposed algorithm consists of a first approach to systematically analyze the impact of microbial communities in humans. Based on the high degree of consistency between the obtained results and the existing literature, we believe the herein proposed algorithm has been successful in identifying events that actually occur *in vivo*.

ACKNOWLEDGMENT

We thank the IEETA Bioinformatics group for all the support during the designing and production of this work.

REFERENCES

[1] P. J. Turnbaugh, R. E. Ley, M. Hamady *et al.*, "The human microbiome project: exploring the microbial part of ourselves in a changing world," *Nature,* vol. 449, no. 7164, pp. 804, 2007.

[2] B. Henderson, "Cell stress proteins as modulators of bacteria-host interactions," *The Biology of Extracellular Molecular Chaperones: Novartis Foundation Symposium, No. 291*, J. G. Derek J. Chadwick, ed., pp. 141-154: Wiley, 2008.

[3] E. D. Coelho, J. P. Arrais, and J. L. Oliveira, "From protein-protein interactions to rational drug design: are computational methods up to the challenge?," *Curr Top Med Chem,* vol. 13, no. 5, pp. 602-18, 2013.

[4] A. M. D. Machado, C. Figueiredo, R. Seruca *et al.*, "Helicobacter pylori infection generates genetic instability in gastric cells," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer,* vol. 1806, no. 1, pp. 58-65, 8//, 2010.

[5] Y. Aso, and H. Akazan, "Prophylactic effect of a Lactobacillus casei preparation on the recurrence of superficial bladder cancer. BLP Study Group," *Urol Int,* vol. 49, no. 3, pp. 125-9, 1992.

[6] C. Huttenhower, D. Gevers, R. Knight *et al.*, "Structure, function and diversity of the healthy human microbiome," *Nature,* vol. 486, no. 7402, pp. 207-214, 06/14/print, 2012.

[7] M. Uhlen, P. Oksvold, L. Fagerberg *et al.*, "Towards a knowledge-based Human Protein Atlas," *Nat Biotech,* vol. 28, no. 12, pp. 1248-1250, 12//print, 2010.

[8] E. Coelho, J. Arrais, S. Matos *et al.*, "Computational prediction of the human-microbial oral interactome," *BMC Systems Biology,* vol. 8, no. 1, pp. 24, 2014.

[9] D. Croft, A. F. Mundo, R. Haw *et al.*, "The Reactome pathway knowledgebase," *Nucleic Acids Research*, November 15, 2013, 2013.

[10] C. Stark, B. J. Breitkreutz, T. Reguly *et al.*, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res,* vol. 34, no. Database issue, pp. D535-9, Jan 1, 2006.

[11] W. K. Chan, G. Yao, Y.-Z. Gu *et al.*, "Cross-talk between the Aryl Hydrocarbon Receptor and Hypoxia Inducible Factor Signaling Pathways: Demonstration of Competition and Compensation," *Journal of Biological Chemistry,* vol. 274, no. 17, pp. 12115-12123, April 23, 1999, 1999.

[12] T. Nguyen, P. J. Sherratt, and C. B. Pickett, "Regulatory Mechanisms Controlling Gene Expression Mediated By The Antioxidant Response Element," *Annual Review of Pharmacology and Toxicology,* vol. 43, no. 1, pp. 233-260, 2003.

[13] M. Hattori, and T. D. Taylor, "The human intestinal microbiome: a new frontier of human biology," *DNA Res,* vol. 16, no. 1, pp. 1-12, Feb, 2009.

[14] V. Tremaroli, and F. Backhed, "Functional interactions between the gut microbiota and host metabolism," *Nature,* vol. 489, no. 7415, pp. 242-249, 09/13/print, 2012.

[15] J. K. Nicholson, E. Holmes, J. Kinross *et al.*, "Host-Gut Microbiota Metabolic Interactions," *Science,* vol. 336, no. 6086, pp. 1262-1267, June 8, 2012, 2012.

[16] I. Cho, and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nat Rev Genet,* vol. 13, no. 4, pp. 260-270, 04//print, 2012.

[17] K. Vielfort, L. Weyler, N. Söderholm *et al.*, "Lactobacillus Decelerates Cervical Epithelial Cell Cycle Progression," *PLoS ONE,* vol. 8, no. 5, pp. e63592, 2013.

[18] D. Žgur-Bertok, "DNA damage repair and bacterial pathogens," *PLoS pathogens,* vol. 9, no. 11, pp. e1003711, 2013.

[19] A. M. Machado, C. Figueiredo, R. Seruca *et al.*, "Helicobacter pylori infection generates genetic instability in gastric cells," *Biochim Biophys Acta,* vol. 1806, no. 1, pp. 58-65, Aug, 2010.

[20] J. J. Mans, K. von Lackum, C. Dorsey *et al.*, "The degree of microbiome complexity influences the epithelial response to infection," *BMC Genomics,* vol. 10, pp. 380, 2009.

[21] M. Handfield, J. J. Mans, G. Zheng *et al.*, "Distinct transcriptional profiles characterize oral epithelium-microbiota interactions," *Cell Microbiol,* vol. 7, no. 6, pp. 811-23, Jun, 2005.

[22] I. Cho, and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nat Rev Genet,* vol. 13, no. 4, pp. 260-70, Apr, 2012.

[23] E. A. Grice, H. H. Kong, G. Renaud *et al.*, "A diversity profile of the human skin microbiota," *Genome Research,* vol. 18, no. 7, pp. 1043-1050, July 1, 2008, 2008.

[24] T. Yuki, H. Yoshida, Y. Akazawa *et al.*, "Activation of TLR2 enhances tight junction barrier in epidermal keratinocytes," *J Immunol,* vol. 187, no. 6, pp. 3230-7, Sep 15, 2011.

[25] Y. Lai, A. Di Nardo, T. Nakatsuji *et al.*, "Commensal bacteria regulate Toll-like receptor 3-dependent inflammation after skin injury," *Nat Med,* vol. 15, no. 12, pp. 1377-82, Dec, 2009.

[26] Y. Lai, A. L. Cogen, K. A. Radek *et al.*, "Activation of TLR2 by a small molecule produced by Staphylococcus epidermidis increases antimicrobial defense against bacterial skin infections," *J Invest Dermatol,* vol. 130, no. 9, pp. 2211-21, Sep, 2010.