

Integration of human metabolic pathway databases: a critical assessment

Miranda D. Stobbe^{1,3,*}, Perry D. Moerland^{1,3}, Antoine H.C. van Kampen^{1,2,3}

¹ Bioinformatics Laboratory, Academic Medical Center, The Netherlands

² Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands

³ Netherlands Bioinformatics Centre, Nijmegen, The Netherlands

* m.d.stobbe@amc.uva.nl

Metabolic pathway databases have proven their usefulness in various applications, ranging from analysis and interpretation of ‘omics’ data to phenotype prediction. As described by Green and Karp [1], the definition of a pathway given by a particular pathway database influences the outcome of certain analyses. We extend their comparison to different aspects of the metabolic network, namely EC numbers, genes, reactions, and the relations between them. We compared five, publicly available, human metabolic pathway databases: EHMN [2], H. sapiens Recon 1 [3], HumanCyc [4], KEGG [5] and Reactome [6].

The five databases combined contain 3442 genes and they agree on 395 genes (11%). We observed similar statistics for EC numbers, with an overlap of 207 (14%) of the 1542 contained in the union of the five databases. When we compare the metabolites based on their KEGG Compound ID the overlap is 11% of the 2341 in total. There are only 137 reactions (2%) on which all databases agree, when ignoring H⁺, e⁻ and H₂O. Pairwise comparisons of the databases showed that Reactome is largely a subset of the other databases and that EHMN is a superset in most respects.

There are various explanations for the lack of overlap we observed. For example, the use of a different definition of what is part of the metabolic network, the number of steps a process is described in, and the number of possible alternative substrates given for a specific enzyme. The lack of identifiers, particularly for metabolites, also influences the comparison. We illustrate these explanations for the differences we have found with a detailed comparison of the TCA cycle as represented in the five pathway databases.

Our results clearly show that the reconciliation of existing metabolic databases is desired and necessary. By combining the databases and resolving the differences we have found, coverage and quality of a human metabolic network could be improved. For the success of such a consensus human metabolic network, broad community support will be crucial

References

- [1] Green, M.L. and Karp, P.D, “The outcomes of pathway database computations depend on pathway ontology”, *Nucleic Acids Res.* (2006)
- [2] Ma, H. et al, “The Edinburgh human metabolic network reconstruction and its functional analysis”, *Molecular Systems Biology* (2007).
- [3] Duarte, N.C. et al, “Global reconstruction of the human metabolic network based on genomic and bibliomic data”, *PNAS* (2007).
- [4] Romero, P. et al, “Computational prediction of human metabolic pathways from the complete human genome”, *Genome Biology* (2004).
- [5] Kanehisa, M. et al, “KEGG for linking genomes to life and the environment”, *Nucleic Acids Res.* (2008).
- [6] Matthews, L. et al, “Reactome knowledgebase of human biological pathways and Processes”, *Nucleic Acids Res* (2009).