Systems biology approach to discovery of phenotype-specific metabolic processes for engineering bacterial bio-hydrogen

Andrea M Rocha,[1] William Hendrix,[2,3] Matthew C Schmidt,[2,3] James R. Mihelcic,[1] Nagiza F Samatova[2, 3,*]

[1]Department of Civil & Environmental Engineering, University of South Florida, Tampa, FL 33620

[2]Department of Computer Science, North Carolina State University, Raleigh, NC 27695

[3]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

[*]E-mail: samatovan@ornl.gov

Abstract. Biological hydrogen is one of the most promising alternative energy and fuel resource today. Due to the ability of several naturally occurring microorganisms to generate hydrogen using different metabolic processes, engineered biohydrogen production is considered a feasible and sustainable technology. To improve upon current technological applications, biological and metabolic engineering approaches are underway to optimize metabolic processes involved in hydrogen production. In order to successfully modify metabolic pathways, full understanding of metabolic networks involved in expression of microbial traits in hydrogen producing organisms is necessary. In this paper, we demonstrate the application of two computational systems biology approaches to identify phenotype-related metabolic components related to hydrogen-producing bacteria. The first method utilizes prior knowledge to discover clusters of phenotype-related proteins in individual phenotype-expressing microorganisms. When applied to bacterial functional relationship data, key enzymes, such as [FeFe]-hydrogenase, [NiFe]-hydrogenase, and glutamate synthase, were identified. In the second method, our efficient graph search algorithm, called PRP-Finder, is used to identify key genes and metabolic pathways that are likely related to biohydrogen production. In addition, predictions on a potential interplay, or cross-talk, between metabolic pathways can be predicted using this approach. Information obtained from both methodologies provides valuable insights into metabolic networks and system controls involved in expression of microbial traits essential for advancing engineering approaches that result in more efficient biohydrogen production.

1. Introduction

Production of biological hydrogen (bio-hydrogen) is potentially a feasible technology for generation of alternative energy and fuels. The ability of several naturally occurring microorganisms to generate hydrogen using various metabolic processes, such as light and dark fermentation metabolic routes, makes bio-hydrogen production a favorable option for addressing bioenergy needs (Nath and Das 2004). Unfortunately, application of bio-hydrogen technologies is often limited due to overall low hydrogen production and yields generated by hydrogen-producing microorganisms (Nath and Das 2004; Jones 2008). This is due mainly to utilization of organic substrates by microorganisms (e.g.,

glucose) towards cellular growth rather than production of hydrogen (Nath and Das 2004). To help overcome this problem, one approach being considered is the genetic modification or bioengineering of microbial metabolic pathways away from biomass production and towards expression of phenotypic traits related to hydrogen production (Jones 2008).

Genetic modification methodologies have not been widely applied towards bio-hydrogen production because a fuller understanding of metabolic processes, networks, and their relationships involved in expression of desired phenotypic traits, such as hydrogen production, is necessary. Although numerous studies have identified hydrogen-producing organisms (Nath and Das 2004), this type of information remains unclear. To close this gap, we started with developing and employing two computational systems biology approaches to identify phenotype-related metabolic processes for fermentative hydrogen production. The former, or the "bottom-up" approach called **DENSE**, capitalizes on the availability of partial "prior knowledge" about the proteins involved in this process and enriches that knowledge with newly identified sets of functionally associated proteins present in individual phenotype-expressing microorganisms. The latter, or the "top-down" approach called **PRP-Finder**, capitalizes on high-throughput comparative analysis of multiple genome-scale metabolic networks to identify phenotype-related metabolic genes and pathways. Description and application of each approach towards bio-hydrogen production is described in the following sections.

2. Bottom-Up Discovery of Phenotype-Related Proteins Using Knowledge Priors

In fermentative hydrogen-producing organisms, such as *Clostridium acetobutylicum*, hydrogen yields are dependent on the presence and activation of hydrogen producing enzymes called hydrogenases (Vignais et al. 2001). Studies evaluating the role of hydrogenase in hydrogen production have shown that organisms can contain more than one type of hydrogenase that can each require sets of accessory proteins for activation. As such, the presence or absence of specific accessory proteins plays an important role in regulating the activity of hydrogenase and hydrogen production or uptake in microorganisms. In addition, many hydrogenases are thought to either directly or indirectly regulate other metabolic processes, such as nitrogen metabolism (Rey et al. 2007). To identify phenotype-related proteins required for activation and maturation of [FeFe]-hydrogenases and understand relationships between hydrogenase and other metabolic proteins, we applied our DENSE algorithm to identify proteins that are functionally associated with the hydrogenase, HydA, in *C. acetobutylicum.*

DENSE was able to identify three maturation proteins that are essential for expression of a [FeFe]-hydrogenase (Akhtar and Jones 2008). They are HydE (CAC1631), HydF (CAC1651), and HydG (CAC1356) (Figure 1). When these proteins are present and interact with HydA1, activation of the hydrogen producing [FeFe]-hydrogenase occurs. According to studies on hydrogenases, deletion of one the proteins will result in inactivation of the [FeFe]-hydrogenase (Akhtar and Jones 2008). In addition to identifying key protein clusters, the algorithm predicted an association between an uncharacterized protein (Figure 1; CAC0487) and the three maturation proteins. Identification of uncharacterized proteins may assist scientists and bioengineers in identifying potential functional roles of unknown proteins. For example, in our study (Hendrix et al. 2010), functional associations between HydA1 and proteins involved in activation of a hydrogen producing enzyme were identified. Since the uncharacterized protein is highly interconnected with the maturation proteins, it can be predicted that the protein is involved in development of the [FeFe]-hydrogenase (HydA1). Utilizing this information, the role of

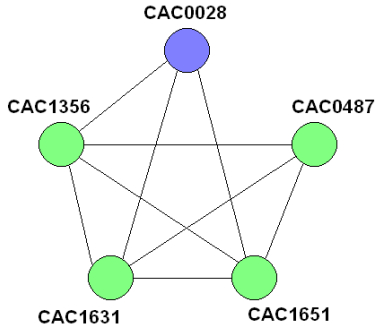

Figure 1. Dense and enriched sub-graphs identified by DENSE algorithm. Blue circle represents the knowledge prior protein HydA1. Image borrowed from Hendrix et al. (2010)

CAC0487 in relation to the three maturation proteins can be characterized through genetic studies and then applied to bioengineering hydrogen producers.

2.1. DENSE: Dense and ENnriched Sub-graph Enumeration

Given a phenotype-expressing organism, our DENSE algorithm tackles the problem of identifying genes that are functionally associated to a set of known phenotype-related proteins by enumerating the "dense and enriched" subgraphs in genome-scale networks of functionally associated or interacting proteins (Hendrix et al. 2010). We define a "dense" subgraph as one in which every vertex is adjacent to at least some γ percentage of the other vertices in the subgraph for some value γ above 50%, which correspond to a set of genes with many strong functional associations between them. In order to incorporate researchers' *prior* knowledge, though, we extend this definition by introducing an "enriched" dense subgraph in which at least μ percentage of the vertices are contained in the knowledge prior query set, for some parameter value μ. Genes contained in such dense and enriched subgraphs, or "μ-enriched, γ-dense quasi-cliques," have strong functional relationships with the previously identified genes, and so are likely to perform a related task.

DENSE uses an agglomerative approach to discover these μ-enriched, γ-dense quasi-cliques, starting with a single query vertex v_0, adding one vertex to the quasi-clique at a time, and backtracking as it finds maximal enriched quasi-cliques or subgraphs that cannot be contained in an enriched quasi-clique. Based on our theoretical results, we limit the vertices that can be added to the subgraph in order to reduce the necessary search space. A number of heuristic values are maintained, and when any of the heuristic values is reduced to zero, the related vertex becomes a "dead end" that is not added to the subgraph.

3. Top-Down High-Throughput Discovery of Phenotype-Related Metabolic Pathways

To discover phenotype-related enzymes and metabolic pathways related to bio-hydrogen production, we applied our PRP-Finder algorithm to evaluate 14 hydrogen-producing and 11 non-hydrogen producing bacteria. We selected microorganisms representative of the three different types of bio-hydrogen production—biophotolysis, light fermentation, and dark fermentation.

Our analysis of identified enzymes and reactions for hydrogen-producing organisms indicates that a large number of reactions involved in glycolysis are conserved across hydrogen producing organisms but not conserved across non-hydrogen producing organisms (Jones 2008). In addition, we identified key metabolic reactions and enzymes related to hydrogen production during mixed acid and dark fermentation (Figure 2). Specifically, we were able to identify formate hydrogen lyase (FHL), an enzyme responsible for the oxidation of formate and production of CO_2 (g) and H_2 (g) (Jones 2008).

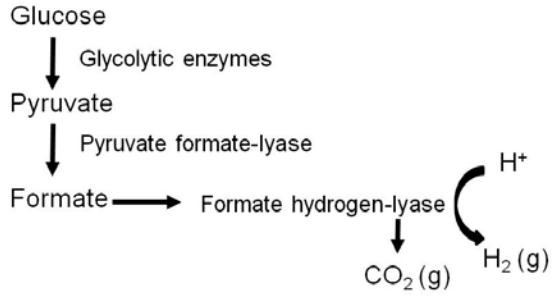

Figure 2. General overview of hydrogen production through mixed acid fermentation.

In addition to dark fermentation routes, the PRP-finder also identified metabolic reactions associated with nitrogen metabolism. These reactions include nitrogen fixation by nitrogenase, and assimilation of ammonia by glutamine synthetase and glutamate synthase. During nitrogen-fixation, N_2 (g) is converted to ammonia by the enzyme nitrogenase and H_2 (g) is produced (Rey et al. 2007). Ammonia produced by these reactions can then be assimilated by the organisms to produce the nitrogen-containing compounds glutamine and glutamate (White 2007). In hydrogen producing-organisms assimilation of ammonia is important since high ammonia concentrations may result in repression of nitrogenase and hydrogen production (McKinlay and Harwood 2010).

3.1. PRP-Finder algorithm

Comparative genome analysis has been successful in identifying some individual phenotype-related enzymes but has struggled to identify phenotype-related metabolic pathways. If a single enzyme is responsible for a given phenotype, then its presence in an organism will be strongly correlated to that of the phenotype (Pellegrini et al. 1999, Jim et al. 2004). However, an enzyme that is part of a metabolic pathway may be present in the organism yet the pathway is not. The presence of such enzymes may show little correlation to the presence of the phenotype.

To enable the prediction of phenotype-related metabolic pathways, we propose a phenotype-driven comparative analysis of genome-scale metabolic networks across phenotype exhibiting organisms and contrasting this comparison across organisms that do not exhibit the target phenotype. This way, we extend a comparative analysis approach to handle sets of enzymes—rather than individual enzymes—in the context of genome-scale metabolic networks. Since it is computationally intractable to consider all possible enzyme sets, some previous approaches have considered only those enzyme sets that correspond to a known metabolic pathway (Tamura and D'haeseleer 2008). However, such a methodology is limited because it cannot detect an interplay, or cross-talk, among phenotype-related metabolic pathways. This calls for approaches that could compare networks at the genome-scale rather than at the individual pathway level.

From a graph-theoretical perspective, we address a computational problem that aims to identify connected subgraphs that are both conserved and statistically enriched across metabolic networks of phenotype expressing organisms. We have developed a heuristic algorithm, called PRP-Finder that is the only known algorithm that can identify the desired subgraphs across tens and even hundreds of metabolic networks in practical time. PRP-Finder "grows" the desired subgraphs and limits the number of subgraphs grown (Schmidt and F.Samatova 2009). PRP-Finder runs in a matter of seconds, yet the subgraphs identified by the PRP-Finder are highly correlated to the subgraphs that are the most likely to represent phenotype-related metabolic pathways.

4. Conclusions

In summary, our hybrid top-down and bottom-up methodologies were able to identify phenotype-related metabolic processes and potentially important interactions among metabolic sub-networks (e.g., pathways and dense subgraphs). In both approaches, we were able to identify known enzymes and reactions associated with hydrogen production through fermentative reactions. In addition, we were able to identify potentially new interactions among the enzymes and the metabolic pathways. Overall, the information obtained from these methodologies, while preliminary, provides valuable insights about metabolic processes involved in expression of phenotypic traits important for engineering microbial organisms for enhanced bio-hydrogen production and improving upon current biotechnological applications through application of bioengineering tools.

Acknowledgements

The work of AMR was supported by the State of Florida 21st Century World Class Scholar Program, Alfred P. Sloan Foundation, and the Research Alliance in Math and Science (RAMS) program sponsored by the Office of Advanced Scientific Computing Research, U.S. Department of Energy. The work of MCS, WH, and NFS was supported by the by the "Ultrascale Computational Modeling of Phenotype-Specific Metabolic Processes in Microbial Communities" project from U.S. Department of Energy, Office of Science. The work was performed in part at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. De-AC05-00OR22725. This work has been authored by a contractor of the U.S. Government, accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

References

Akhtar, M.K., and P.R. Jones (2008). "Engineering of a Synthetic hydF=hydE=hydG=hydA operon for biohydrogen production." *Analytical Biochemistry* **373**: 170–172.

Hendrix, W., A.M. Rocha, M. Elmore, J. Trien, and N. F. Samatova (2010). *Discovery of Enriched Biological Motifs with Knowledge Priors*. BIOCOMP'10–11th International Conference on Bioinformatics and Computational Biology.

Jones, P.R. (2008). "Improving Fermentative Biomass-derived H_2 Production by Engineering Microbial Metabolism." *International Journal of Hydrogen Energy* **33**: 5122–5130.

McKinlay, J.B., and C.S. Harwood (2010). "Photobiological production of hydrogen gas as a biofuel." *Current Opinion in Biotechnology* **21**(3): 244–251.

Nath, K., and D. Das (2004). "Improvement of hydrogen production: Various approaches." *Appl. Microbiol. Biotechnol.* **65**: 520–529.

Rey, F.E., E.K. Heiniger, and C.S. Harwood (2007). "Redirection of metabolism for biological hydrogen production." *Applied and Environmental Microbiology* **73**(5): 1665–1671.

Schmidt, M.C., and N.F.Samatova (2009). *An Algorithm for Discovery of Phenotype Related Metabolic Pathways*. IEEE International Conference on Bioinformatics & Biomedicine Washington, D.C.

Tamura, M., and P. D'haeseleer (2008). "Microbial genotype-phenotype mapping by class association rule mining." *Bioinformatics* **24**(13): 1523–1529.

Vignais, P.M., B. Billoud, and J. Meyer (2001). "Classification and phylogeny of hydrogenases." *FEMS Microbiology Reviews* **25**(4): 455–501.

White, D. (2007). *The physiology and biochemistry of prokaryotes*. New York, Oxford University Press.