Title
Electromicrobiology: a systems perspective

Permalink
https://escholarship.org/uc/item/2r80r0zp

Author
Nagarajan, Harish

Publication Date
2012

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA, SAN DIEGO

Electromicrobiology: A Systems Perspective

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in
Bioinformatics & Systems Biology

by

Harish Nagarajan

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Milton H. Saier Jr., Co-Chair
Professor Vineet Bafna
Professor Theresa Gaasterland
Professor Karsten Zengler
Professor Kun Zhang

2012
The dissertation of Harish Nagarajan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2012
DEDICATION

To the loving memory of my father.

To my mother, my pillar of support.
EPIGRAPH

I rarely plan my research,
   it plans me.
—Max F. Perutz
# TABLE OF CONTENTS

Signature Page ................................................................. iii
Dedication ................................................................. iv
Epigraph ................................................................. v
Table of Contents ........................................................ vi
List of Figures ........................................................ x
List of Tables ........................................................ xxi
Acknowledgements ......................................................... xxii
Vita ................................................................. xxv
Abstract of the Dissertation .............................................. xxvii

<table>
<thead>
<tr>
<th>Chapter 1</th>
<th>Introduction ......................................................... 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>What is Electromicrobiology? ........................................ 1</td>
</tr>
<tr>
<td>1.1.1</td>
<td>Microbial Fuel Cells ........................................ 2</td>
</tr>
<tr>
<td>1.1.2</td>
<td>Microbial Electrosynthesis ......................................... 3</td>
</tr>
<tr>
<td>1.1.3</td>
<td>Direct Interspecies Electron Transfer ................................ 4</td>
</tr>
<tr>
<td>1.2</td>
<td>Understanding the microbial genotype - phenotype relationship using genome-scale systems biology approaches 5</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Application of COBRA methods to characterize microbial community interactions ........................................ 10</td>
</tr>
<tr>
<td>1.3</td>
<td>Extending systems biology to characterize electromicrobiological applications ........................................ 14</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Characterizing microbial fuel cells .................................... 15</td>
</tr>
<tr>
<td>1.3.2</td>
<td>Characterizing microbial electrosynthesis .......................... 16</td>
</tr>
<tr>
<td>1.3.3</td>
<td>Investigating direct interspecies electron transfer in syntrophic associations ........................................ 16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 2</th>
<th>De novo assembly of the complete genome of an enhanced electricity producing variant of <em>G. sulfurreducens</em> 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Abstract ......................................................... 18</td>
</tr>
<tr>
<td>2.2</td>
<td>Introduction ......................................................... 19</td>
</tr>
<tr>
<td>2.3</td>
<td>Results ......................................................... 21</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Meta-Assembly strategy ........................................ 21</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Assembly validation ........................................ 25</td>
</tr>
</tbody>
</table>
2.3.3 KN400 is the first microbial genome that has been completely assembled de novo using only next-generation sequencing technologies.

2.3.4 Genomic characteristics of KN400 are typical of microbial genomes at large.

2.4 Discussion

2.5 Materials and Methods

2.5.1 Data

2.5.2 Meta-Assembly

2.5.3 Validation by Sanger Sequencing

2.5.4 Gene Prediction and Comparative Analysis

2.6 Acknowledgments

Chapter 3 Characterizing the interplay between multiple levels of organization of bacterial sigma factor regulatory networks

3.1 Abstract

3.2 Introduction

3.3 Results

3.3.1 The topology of the sigma factor network in *Geobacter sulfurreducens*

3.3.2 The operational state of the sigma factor network

3.3.3 Functional effect of the sigma factor network operation

3.4 Discussion

3.5 Methods

3.5.1 Bacterial strains, medium, and growth conditions

3.5.2 ChIP-chip and ChIP-seq

3.5.3 Identification of sigma factor binding regions

3.5.4 Binding motifs of the sigma factors in *G. sulfurreducens*

3.5.5 Western blot

3.5.6 Functional enrichment analysis of the regulons

3.5.7 *In silico* gene essentiality simulations

3.6 Acknowledgements

Chapter 4 Genome-scale metabolic characterization of electrosynthetic bacteria

4.1 Introduction

4.2 Metabolic Reconstruction

4.2.1 Characterization of the *Geobacter metallireducens* GS-15 genome-scale reconstruction

4.2.2 Genome-scale reconstruction of the homoacetogen, *Clostridium ljungdahlii*
Figure 1.1: **Aspects of Electromicrobiology:** A. Microbial Fuel Cell where the organism respires on organics and uses extracellular electron acceptors such as electrodes or metals to derive energy; B. Microbial Electrosynthesis: The scenario where, microbes use electrode as the electron donor to fix CO$_2$ to organic compounds like butanol; C. Direct Interspecies Electron Transfer: Syntrophic associations in which microorganisms establish electrical connections for the transfer of electrons from one partner to another . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

Figure 1.2: **Fundamentals of the genome-scale metabolic genotype-phenotype relationship:** The COBRA approach is based on three primary fundamental concepts: network constraints (a-d), objective functions (e), and the association of reactions with the genome. (a) A complex mixture of molecules (red) can react to yield end products (blue). (b) The stoichiometry of this reaction network is described mathematically in a stoichiometric matrix, with each column representing the stoichiometry of a reaction. Negative and positive values represent reactants and products, respectively. Reaction flux is limited by thermodynamics and catalytic capacities (Vm=Vmax), described by upper and lower bounds on flux for each reaction (green). (c) Reaction constraints result in a solution space that contains all feasible flux distributions. Additional constraints (e.g., mass balance, the steady-state assumption, and measured metabolite consumption rates) reduce the space of feasible flux distributions, as shown by the pink line. (d) In vivo biochemical networks involve additional complexity. Gene regulation can change the abundance of catalysts (e.g., the transformation of D to E). Often components are also localized in different organelles (e.g., E and F), thereby blocking reactions. (e) The biomass objective function describes an evolutionary pressure for microbial growth, and describes the metabolic demands to make basic metabolite building blocks for all cellular components (e.g., membranes, macromolecules, ATP, etc.). (f) The association of metabolism with the genome is done by mathematically linking the genome to transcripts, proteins, and chemical reactions. The gene-protein-reaction schema is used to describe gene association in the models, and provide an interface for the integration of high-throughput data. . . . . . . . . . 8
Figure 1.3: Introducing COBRA methods to study community interactions: COBRA methods are providing insight into the metabolic interactions in various types of microbial communities. (a) To study the mutualistic behavior of co-dependent mutant *E. coli*, researchers used MOMA [1] to simulate synergistic growth of pairs of auxotrophic *E. coli*. (b) Shadow prices from FBA simulations of these pairs were used to compute cooperation efficiencies between strains, which were subsequently compared with measured fitness improvements. (c) Competition in communities was modeled using DMMM [2] to understand how communities of *Geobacter* and *Rhodoferax* compete for resources, and how the demographics vary under different nutrient ratios, thereby affecting the efficiency of bioremediation efforts. Host-pathogen interactions between *M. tuberculosis* and a human macrophage were studied using COBRA. (d) While transcriptomic data were employed to build host-pathogen models at different stages of infection, the cellular objective of internalized *M. tuberculosis* is not known, so refinements to the objective function were predicted from transcriptomic data to account for changes in required amounts of compounds like lipids and amino acids (AAs). (e) This information was used to compute flux states of internalized *M. tuberculosis* with MCMC sampling [3]. This demonstrated a suppression of central metabolism and activation of the glyoxylate shunt, represented here by enolase and isocitrate lyase, respectively. The role of communities in evolution has been studied using Reductive evolutionary simulation [4]. In particular, this method predicted the minimal set of genes needed to for *Buchnera* to grow in the rich inards of the aphid. The predicted minimal gene sets (f) and temporal order of gene loss (g) were consistent with the gene content and phylogenetic structure of several *Buchnera* species.

Figure 2.1: Assembly Strategy: A) Hybrid Assembly Phase; B) Scaffold Bridging and Finishing Phase; C) Scaffold Ordering Phase: The left branch of the decision tree consists of permutations that can be confirmed by the PCR performed while the right branch consists of those permutations that cannot be confirmed by the particular PCR. The faded permutations are those which have been eliminated by the PCRs while those in bold are those that are remaining. (Gel Inset: Showing PCR products for all the 9 PCRs performed in the search strategy to confirm the correct orientation of the scaffolds); D) Genome Finishing Phase.
Figure 2.2: Custom program used in the Scaffold Bridging and Finishing Phase of the assembly process.

Figure 2.3: Assembly Validation approaches: A) Sanger sequencing approach; B) Comparative Genomics approach.

Figure 2.4: Genome-Level comparison of KN400 and PCA. Shown in this figure is a dot-plot of the genome-wide alignment of KN400 and PCA. Along the X-Axis is the KN400 genome and the PCA genome is shown along the Y axis.

Figure 2.5: Gel Picture confirming the 79 kb deletion (Region2) in KN400. PCR was performed with primer sets in order to amplify over the break (shown in panel B). The expected product size is 207bp. Panel A shows that we can amplify over the break only in KN400 and not in PCA, confirming the deletion of region 2 in KN400.

Figure 2.6: Density of five different genomic properties in the space of microbial genomes. A) GC Content B) Genome Size C) Number of rRNAs D) Number of Replicons E) Number of tRNAs. Shown in red circle, is the value of KN400’s genomic property.

Figure 2.7: Distribution of fragment sizes from 454 paired end sequencing data used for de novo assembly.

Figure 3.1: Sigma factor network topology: (a) The topology of the interaction network of the four major sigma factors in G. sulfurreducens. (b) The topology of the interaction network of the four major sigma factors in E. coli. The sigma factor network topology of G. sulfurreducens suggests a robust network with all of the alternative sigma factors ($\sigma^H$, $\sigma^N$, and $\sigma^S$) regulating the housekeeping $\sigma^D$. Auto-regulation of three sigma factors ($\sigma^D$, $\sigma^H$, and $\sigma^N$) in G. sulfurreducens further highlights critical topological differences compared to the E. coli sigma factor network.

Figure 3.2: Operational state of the G. sulfurreducens sigma factor network: The interaction network of the four major sigma factors with the metabolic genes in G. sulfurreducens is shown. An in silico gene essentiality screen was carried out under four different conditions to assess the effect of genetic perturbations on the sigma factor network operation. The genes determined to be essential under all four conditions are shown by red nodes; other non-essential genes are shown in gold.
Figure 3.3: **Functional state of the *G. sulfurreducens* sigma factor network:** (a) Comparative transcriptomic profile between *G. sulfurreducens* grown as planktonic cells reducing fumarate and as an electrogenic biofilm on an anode. About 10% of the transcriptome showed differential expression between the two conditions. (b) Venn diagram showing the comparative binding profiles of the three sigma factors $\sigma^D$, $\sigma^N$, and $\sigma^S$ for planktonic cells and electrogenic biofilm. The size of each sigma factor regulon from electrogenic biofilm cells is shown in parentheses. The invariant binding profiles between the two conditions indicate a near-static functional state for the sigma factor network in *G. sulfurreducens.*

Figure 3.4: **Analysis of RpoS protein concentrations during growth of *G. sulfurreducens*.** (A) Samples were collected from batch grown, wild type *G. sulfurreducens* on fresh water acetate media with fumarate as an electron acceptor. (B) Proteins were separated using electrophoresis on a 10% polyacrylamide gel. RpoS protein concentrations during exponential phase and stationary phase were determined using Western blot analysis. RpoS (42 kDa in size) concentrations remained at fairly constant levels across all the five time points (P1-5).

Figure 3.5: **Coordinated regulation of energy metabolism by the alternative sigma factors:** Analysis of the differences in expression and sigma factor binding profiles between the two divergent growth conditions (planktonic cells and electrogenic biofilm) in the context of the genome-scale metabolic network. A snapshot of the key steps involved in the energy metabolism during this shift is shown. $\sigma^N$ primarily regulated energy metabolism during this shift by controlling the terminal steps in the respective respiratory processes (DcuB, and the outer membrane cytochromes, OmcO/P and OmcZ). $\sigma^S$ regulated the NADH dehydrogenase (nuo operon). The genes upregulated and downregulated in this shift are shown in green and red, respectively. OmcZ and NuoIJKL were also found to be co-regulated by $\sigma^D$. The alternative sigma factors thus play a major role in the tight regulation of energy metabolism during this shift.
Figure 3.6: Interplay between the multiple levels of sigma factor network organization. The robust topology of the sigma factor network in *G. sulfurreducens* suggests a broader role for alternative sigma factors. This resulted in a modular operational state with $\sigma^N$ primarily regulating genes involved in energy metabolism and maintaining redox homeostasis, while the housekeeping $\sigma^D$ regulated biosynthetic processes. The sigma factor network has a near-static functional state with invariant binding profiles between the two divergent growth conditions (planktonic cells and electrogenic biofilm). Analysis of the transcriptomic changes and sigma factor binding profiles during this shift indicated that the necessitated tight regulation of energy metabolism during this shift is mirrored in the modularity of the operational state. This coordinated regulation in the operational state can further be attributed to the robust topology of the sigma factor network where the major sigma factors auto-regulate each other providing feedback mechanisms to fine tune expression.

Figure 3.7: Binding motif of the four major sigma factors in *Geobacter sulfurreducens*.

Figure 4.1: Reconstruction Workflow: (a) The workflow describing the reconstruction of the *G. metallireducens* reconstruction. The reconstruction process was initiated by comparing the updated genome annotation for *G. metallireducens* to the existing reconstruction to create a list of discrepancies that was manually reviewed and curated. Content that was in agreement between the updated annotation and reconstruction was used to generate a draft set of intracellular reactions. Lipid, membrane, murein, and LPS content was removed from this list as a periplasm compartment was added to the reconstruction. Manual curation (12) was aided by the KEGG [5], MODEL SEED [6], and MetaCyc [7] databases. Further, numerous publications and literature sources (i.e., the bibliome) were used to refine the network content. The manual review process resulted in a draft reconstruction that was used in conjunction with a formulated biomass objective function in simulations to validate the content of the reconstruction and generate a final version. B Venn Diagram showing the comparative analysis of network content.

Figure 4.2: Reconstruction Workflow: Iterative reconciliatory reconstruction workflow adopted for generating the metabolic model for *C. ljungdahlii*. 

---

xv
Figure 4.3: Example of new content in the reconstruction Nitrate reduction pathway in C. ljungdahlii. 78

Figure 4.4: Carbon fixation pathways in G. metallireducens A. A map of the two carbon fixation pathways included in the iAF987 reconstruction. The figure shows the two carbon fixation pathways encoded by genes annotated in the updated genome annotation. 81

Figure 4.5: Carbon fixation pathways: A map of the Wood-Ljungdahl pathway of carbon fixation and the genes encoding the CODH ACS complex. 83

Figure 4.6: Protocol adopted for reconstructing electron transfer pathways 84

Figure 4.7: Modified ETS of G. metallireducens A: Generation of a proton gradient due to an electrogenic fumarate reductase while respiring on fumarate. B: Dissipation of membrane potential due to succinate oxidation while respiring on Fe(III). Electron transfer to Fe(III) via a chain of cytochromes. 86

Figure 4.8: Phenotypic analysis of respiration on intracellular electron acceptors Phenotypic Phase Plane (PhPP) analysis of the G. metallireducens GS-15 dcuB strain on three different carbon sources. The reconstructed model of G. metallireducens GS-15 was utilized to analyze the growth of the dcuB strain on acetate, ethanol, and butanol as electron donors with fumarate as the electron acceptor. Inputs to the simulations were taken from the ranges observed in the experimental growth screens of each. Each panel shows the line of optimality on the diagonal, the donor limiting region in purple, and the acceptor limiting region in green, and blue regions are regions that outside the functionality of the network. Growth on acetate revealed an optimal ratio of donor and acceptor and a growth rate near the optimal growth rate. Growth on ethanol and butanol revealed that the cells were observed to be functioning in the suboptimal donor-limiting regions and the growth rates are far below the optimal growth rate for the observed acceptor (fumarate) uptake rates. 88
Figure 4.9: Phenotypic analysis of respiration on extracellular electron acceptors (Fe(III)) Phenotypic Phase Plane (PhPP) analysis of *G. metallireducens* GS-15 to examine the cost of extracellular electron transfer. The reconstructed model of *G. metallireducens* GS-15 was utilized to analyze the growth of wild type *G. metallireducens* GS-15 on acetate, ethanol, and butanol as electron donors with Fe(III) as the electron acceptor. The far left column shows the calculated PhPPs and the observed experimentally measured ratio between substrate uptake and Fe(III) conversion rate (dashed line). The hypothesis of a directly proportional energetic cost to the Fe+3 conversion rate was examined and a cost of 0.4 ATP hydrolyzed per electron transferred to Fe(III) resulted agreement between the line of optimality (i.e., the intersection of Fe+3 conversion limited and substrate uptake limited regions) and the experimentally observed ratio for all three cases.

Figure 4.10: Energy conservation mechanisms in *C. ljungdahlii* A: Electron bifurcation in the Wood-Ljungdahl pathway at the Methylenetetrahydrofolate dehydrogenase small subunit to generate reduced ferredoxin. B: Energy conservation through proton gradient established at the Rnf complex. C: Identification of a gene encoding the electron bifurcating transhydrogenase for the generation of NADPH during autotrophic growth (Nfn complex). D: Electron bifurcating hydrogenase for the reduction of ferredoxin from hydrogen during autotrophic growth.

Figure 4.11: Framework for modeling microbial electrosynthesis A: Schematic for modeling microbial electrosynthesis. The current drawn by the microbe at the cathode is used as the input constraint to the model which contains the necessary electron transfer pathways and carbon fixation pathways. This is then used to simulate electrosynthetic conditions by maximizing for energy generation and analyze the product formation scenarios. B: Simulation of electrosynthetic production of acetate and formate in *G. metallireducens*.
Figure 5.1: **Integrated framework for modeling syntrophy and direct electron transfer.** A: Protocol adopted for reconstructing extracellular electron transfer pathways. B: Properties of the updated and expanded genome-scale metabolic reconstructions used. C: Formalism for reconstructing syntrophy. Using an extension of the compartmentalized framework, a shared metabolite pool (SMP) is included to facilitate exchange of metabolites between the species as well as with the environment. D: Modeling Direct Interspecies Electron Transfer (DIET). The maximum flux through DIET is calculated based on the mid-point potential of OmcS.

Figure 5.2: **Physiological growth screens performed on the WT and CS K.O strains.** The uptake rates calculated from this data were used to constrain the model. While constraining the model with uptake rates, the fumarate uptake rate was calculated as the net fumarate and malate rates.

Figure 5.3: **Optimal fraction of community constituents.** To estimate the relative abundance of the community, the growth rate was sampled over a range of fixed relative abundance of *G. metallireducens*. This resulted in a predicted optimal fraction of 27-30% of *G. metallireducens*.

Figure 5.4: **Carbon and electron flow in the consortium.** Flux distribution in the wild type syntrophic interaction between *G. metallireducens* and *G. sulfurreducens*. All flux values are shown in mmol/gDWh.

Figure 5.5: **Impact of nitrogen limiting conditions on the syntrophic association.** To assess the impact of nitrogen fixation on the syntrophic association, simulations were carried out in ammonia free media. In a scenario where *G. sulfurreducens* is the sole organism fixing nitrogen, the maximum relative abundance *G. sulfurreducens* could achieve was around 45%. No feasible scenarios were computed for syntrophic growth of both organisms, if *G. metallireducens* had to be the sole organism fixing N₂.
Figure 5.6: **Insights into mechanisms of electron transfer from transcriptomic data.** A: Fold changes observed for the differentially expressed genes in the hyb K.O strain and CS K.O strain. B: Metabolic interpretation of the transcriptomic differences in the hyb K.O strain relative to the WT strain. The main genes unregulated in the hyb K.O strain include the outer membrane cytochromes that are associated with the pathway for DIET in the consortium. C: Metabolic adjustments in the CS K.O strain based on the transcriptomic data. The elimination of interspecies electron transfer via acetate oxidation confers a suboptimal direct electron transfer phenotype to the consortium. This is mirrored by the significant suppression of the electron transfer pathway involved in DIET in *G. metallireducens*. D: Interaction motif describing the synergistic and antagonistic relationship between DIET, acetate transfer and interspecies hydrogen transfer.

Figure 5.7: **Dynamics of interaction between the partners as they respond to selection pressure of syntrophic growth.** Phenotypic phase plane analysis reveals the presence of two distinct phases with maximal and suboptimal direct interspecies electron transfer.
Figure 5.8: **Comparative analysis of interspecies electron transfer mechanisms.** A. Relative efficiency of the different modes of interspecies electron transfer. DIET+Acetate case represents the WT strain, DIET case represents the CS K.O strain and the hydrogen transfer case is simulated using the coculture *P. carbinolicus* and *G. sulfurreducens*. The maximum energy yield per mol of electron donor is calculated for each mode of interspecies electron transfer. The theoretical maximum for the net reaction in this consortium is 8.27. The pie charts represent the relative abundance of the interacting partners during each mode of interspecies electron transfer. B. Relationship between community dynamics and mode of interspecies electron transfer. Simulations were carried out with a community model capable of doing hydrogen transfer and DIET. Simulations were conducted over the entire range of optimality levels of both the interacting organisms and whether DIET or hydrogen transfer occurred was monitored. The blue area represents the sole occurrence of DIET during the interaction and the orange area shows the region where DIET and hydrogen transfer occur. Interspecies hydrogen transfer occurred in addition to DIET, in regions where the optimality levels of the electron accepting organism was at least 1.8 times that of the electron donating organism.

Figure 5.9: **Comparison of energy conservation strategies adopted**
A: Energy conservation by *G. metallireducens* during DIET. ATP is primarily generated due to the proton gradient established at the inner membrane when electrons are transferred from NADH to menaquinone and subsequently to the periplasmic cytochromes. B: Energy conservation by *P. carbinolicus* during interspecies hydrogen transfer. ATP is primarily generated by substrate level phosphorylation during the formation of acetate from ethanol. Hydrogen is further generated through electron confurcation at the Nfn complex and the subsequent action of the NADP-dependent hydrogenase.

Figure 5.10: **Effect of proton availability on energy efficiency of the different modes of electron transfer.** Robustness analysis was performed on the influx of protons into *G. sulfurreducens* to assess its impact on energy yield of DIET, DIET+acetate transfer and interspecies hydrogen transfer. While the proton flux had a linear relationship on the efficiency of DIET and DIET+acetate transfer, there was no impact observed in the case of interspecies hydrogen transfer.
Figure 5.11: **Detailed schematic of the expansion of the compartmentalized framework adopted for integrating two different genome-scale metabolic models.** The outer membrane transporters in each of the metabolic models were replaced by transporters that facilitated transfer from the respective periplasmic compartments to the SMP. Exchange reactions were added for all metabolites present in the SMP.
LIST OF TABLES

Table 2.1: Summary and Statistics of different stages of Meta-Assembly. Phase A is the Hybrid Assembly that comprises of three distinct steps. ................................................................. 24
Table 2.2: Changes made due to alignment of Illumina reads in the Genome-Finishing Phase ................................................. 25
Table 2.3: Comparison of the approach Meta-Assembly to other commonly used assembly programs .............................................. 34
Table 2.4: Details of the next generation sequencing data used for de novo assembly of KN400 .............................................. 37
Table 2.5: Sequence and location of primers. (Lower case denotes reverse complement) .................................................. 39
Table 2.6: Primer pairs and product analysis of each PCR during the Scaffold Ordering Phase ............................................... 39
Table 3.1: Properties of regulons of the four major sigma factors in Geobacter sulfurreducens ................................... 46
Table 4.1: Subsystem distribution of reactions unique to G. metallireducens reconstruction ........................................... 72
Table 4.2: Known pathways for autotrophic carbon fixation and their energy requirements .............................................. 79
Table 5.1: Details of the different mutant strains modeled to obtain insights into mechanisms of electron transfer ................. 113
ACKNOWLEDGEMENTS

This thesis would not have been possible without the many sacrifices, support, and guidance of countless people.

I would like to thank my advisor, Bernhard Palsson, for providing me with this great opportunity to work in his research group. Specifically, I would like to thank him for his guidance and mentorship on various aspects of scientific and professional development. In particular, I would like to acknowledge him for moulding my perspective on scientific research. The skills I have learnt from him will have a lifelong impact on my career. I would like to thank Karsten Zengler for being a great mentor guiding me through various aspects of this thesis. He has always been around to lend a patient ear to all my questions and hypothesis. I learnt a lot from our interactions on various microbiological aspects and will cherish them forever.

I must definitely thank the numerous researchers and collaborators at the Systems Biology Research Group at UCSD, for providing an intellectually stimulating environment over the past five years. I would like to thank Christian Barrett, Adam Feist, Pep Charusanti, Neema Jamshidi, Vasily Portnoy for their collaboration and being senior mentors in the lab. I have thoroughly enjoyed the fruitful collaborations and discussions I have had with Nathan E. Lewis, Mallory Embree, Haythem Latif, Joshua Lerman, Merve Sahin, Aarash Bordbar, Ali Ebrahim, Teddy O’Brien, Yu Qiu, Juan Nogales Enrique, Hojung Nam, Niko Sonnenschein, and Roger Chang. I would also like to thank my office mates during various stages of my PhD Jeff Orth, Ronan Fleming, Kenyon Applebee, DaeHee Lee, for their discussions. I would also like to thank all the other researchers of the Systems Biology Research Group for their interactions over the past five years. Finally, I would like to thank Marc Abrams and Kathy Andrews for providing the administrative support.

I would like to thank my classmates Nitin Udpa, Roger Chang, and Nisha
Rajagopal for all the support during these five years. The time during our first year will always be cherished. I would also like to thank other friends at the Bioinformatics & Systems Biology graduate program, Josue Perez, Vipul Bhargava, Jan Schellenberger, and Mary Pacold for their support and discussions. Special thanks to all the graduate coordinators at the Bioinformatics program - Jan Lenington, Laura Gracia and Kathleen Kazules for their support.

My time at UCSD would not have been possible without the sense of balance provided by my friends Arun Manohar, Balaji Sriram, Nitin Udpa, Joshua Lerman, Josue Perez, and Nisha Rajagopal.

I would also like to thank Derek Lovley and his group at University of Massachusetts for a fruitful collaboration. I am grateful to Prof. Milton Saier, Prof Kun Zhang, Prof. Vineet Bafna, and Prof. Terry Gaasterland for serving on my doctoral committee. In particular I would like to thank Prof. Zhang for allowing access to his Illumina sequencer to generate datasets that were instrumental in the completion of this thesis.

Lastly, this thesis would definitely not have been possible without the support of my family. I am greatly indebted to my mother for all the sacrifices she has made to raise me and provide me with this opportunity. She has been a huge pillar of support and nothing would have been possible without her.


Chapter 2 in full is a reprint of the material Nagarajan, H., Butler JE, Klimes A, Qiu Y, Zengler K, Ward J, Young ND, Methe BA, Palsson BO, Lovley DR, Barrett CL. De Novo assembly of the complete genome of an enhanced

Chapter 3 in full is a reprint of the manuscript under review Qiu Y, Nagarajan H, Embree M, Shieu W, Abate E, Juarez K, Elkins J, Cho B-K, Nevin KP, Barrett CL, Lovley DR, Palsson BO, Zengler K. Characterizing the interplay between multiple levels of organization of bacterial sigma factor regulatory networks. The dissertation author was the second author of this paper, responsible for all the computational analysis and systems-level interpretation of the results.

Chapter 4 is in part, adapted from a paper that is being prepared for publication Feist AM, Nagarajan H, Aklujkar M, Tremblay PL, Rotaru AE, Lovley DR, Zengler K. Integrated model-driven and experimental characterization of the growth capabilities of Geobacter metallireducens GS-15. The dissertation author was one of the primary researchers responsible for the reconstruction and analysis of carbon fixation and energy conservation.

Chapter 4 is in part, adapted from a paper that is being prepared for publication Nagarajan H, Nogales J, Sahin M, Embree M, Ebrahim A, Feist AM, Lovley DR, Palsson BO, Zengler K. Using genome-scale models to aid biological discovery of acetogenic metabolism. The dissertation author was the primary author of this paper responsible for the research.

Chapter 5 in full has been submitted for publication of the material Nagarajan H, Embree M, Rotaru AE, Shrestha PM, Feist AM, Palsson BO, Lovley DR, Zengler K. Multi-omic characterization and modeling of interspecies electron transfer mechanisms and microbial community dynamics of syntrophic associations. The dissertation author was the primary author of this paper responsible for the research.
VITA

2007 B.Tech. in Biotechnology, Indian Institute of Technology, Madras

2012 Ph. D. in Bioinformatics & Systems Biology, University of California, San Diego

PUBLICATIONS


4. Lewis NE*, Liu X*, Li Y*, **Nagarajan H**.*, et al. The genomic divergence of CHO cell lines is revealed by the genomic sequence of the Chinese Hamster, *Cricetulus griseus*. *Under Revision*. * Equal contribution


ABSTRACT OF THE DISSERTATION

Electromicrobiology: A Systems Perspective

by

Harish Nagarajan

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California, San Diego, 2012

Professor Bernhard Ø. Palsson, Chair
Professor Milton H. Saier Jr., Co-Chair

The ability of microorganisms to exchange electrons directly with their environment has large implications for our knowledge of industrial as well as for environmental processes. For decades, it has been known that microbes can use electrodes as electron acceptors in microbial fuel cell settings. Recently, it has been shown that organisms are also capable to accept electrons directly from an electrode for fixation of carbon dioxide into multi-carbon molecules (Microbial Electrosynthesis). The origin of these industrially relevant processes probably lies in the ability of microorganisms to transfer electrons directly between each other. Such interactions between microorganisms play a major role in the functioning of global biogeochemical cycles. Hence, there is a great need to gain mechanistic
insights into the various factors governing microbial extracellular electron transfer. In this thesis, a genome-scale systems biology approach is applied to characterize these three aspects of electromicrobiology (microbial fuel cell, microbial electrosynthesis, and interspecies electron transfer). First, we apply next-generation sequencing (NGS) technologies to de novo assemble the complete genome sequence of an enhanced electricity producing variant of an organism in a microbial fuel cell. This was further extended to the characterization of regulatory networks governing electrogenic biofilm growth using multi-omic data sets. In the second part, a genome-scale characterization of the metabolic capabilities of two electrosynthetic bacteria is presented. Finally, we demonstrate how analysis of multi-omic data in the context of genome-scale models of microbial consortia enables us to decipher the underlying mechanism and cellular requirements for direct electron transfer in microbial associations.
Chapter 1

Introduction

1.1 What is Electromicrobiology?

The ability of microorganisms to exchange electrons directly with their environment has large implications for our knowledge of industrial as well as for environmental processes. Electron transfer reactions are known to be fundamental to energy conservation and metabolism of bacteria. The study of electronic properties of microorganisms that govern these interactions with the environment is called electromicrobiology [8]. Microbial extracellular electron transfer is the underlying driver of electromicrobiology. Advances in this field have been largely dominated by the study of dissimilatory metal reduction and microbial fuel cells (Fig. 1.1A).
**Figure 1.1: Aspects of Electromicrobiology:** A. Microbial Fuel Cell where the organism respires on organics and uses extracellular electron acceptors such as electrodes or metals to derive energy; B. Microbial Electrosynthesis: The scenario where, microbes use electrode as the electron donor to fix $CO_2$ to organic compounds like butanol; C. Direct Interspecies Electron Transfer: Syntrophic associations in which microorganisms establish electrical connections for the transfer of electrons from one partner to another

### 1.1.1 Microbial Fuel Cells

Microbial fuel cells are devices that typically utilize bacterial metabolism to harvest electrical current from a wide range of organic substrates. *Geobacteraceae* are one such class of organisms that can use electrodes as electron acceptors for anaerobic respiration and thus making it possible to harvest electricity from waste organic matter.[9]. The other major class of bacteria that have been shown to be capable of generating electricity in a microbial fuel cell is *Shewanella*. However, these two classes of bacteria adopt very different mechanisms for electron exchange with extracellular electron acceptors such as the electrode. While *Shewanella oneidensis* interacts with electrodes primarily through soluble electron shuttles (flavin molecules), *Geobacteraceae* establish direct contact with the electrode surface via outer-surface $c$-type cytochromes [8]. One of the members of this class, *Geobacter sulfurreducens* produces the highest current densities known for pure cultures, in microbial fuel cells. The availability of the full genome sequence for *G. sulfurreducens* and the fact that it is amenable to genetic manipulations
have made it the model organism for studies related to microbial fuel cells. *G. sulfurreducens* is also referred to as electricigens, since it conserves energy by oxidizing organic compounds like acetate to carbon dioxide and transfer electrons to anodes. Genome-scale studies and electrochemical analyses have indicated that the *Geobacter* cells form microbial nanowires that establish long range contact with anodes and facilitate direct electron transfer.[10]. It has further been shown that the pili in *Geobacter* have tunable metallic like conductivity and aid in the formation of conductive biofilms[11].

### 1.1.2 Microbial Electrosynthesis

Apart from the application of microbial fuel cells, two other potential applications of electromicrobiology have been demonstrated recently. One of these applications exploits the possibility of electron flow in the reverse direction compared to microbial fuel cells. One such application is called the process of microbial electrosynthesis, where electrons are fed to microorganisms at the cathode to enable reduction of carbon dioxide (Fig. 1.1B) [12]. In this process, the bacteria on the cathode uses electrons as the only source of energy for directly converting carbon dioxide into multi-carbon organic compounds. This is the scenario where microbes use cathodes as the electron donor to reduce carbon dioxide to multi-carbon organic compounds. This process can be leveraged to produce liquid transportation fuels and other useful commodity chemicals. Microbial electrosynthesis can be thought of as an artificial form of photosynthesis when driven by solar power. However, microbial electrosynthesis differs significantly from photosynthesis in that carbon and electron flow is primarily directed to the formation of extracellular products, rather than biomass. So far, only a few acetogenic microorganisms have been shown to be capable of accepting electrons from the cathode to reduce carbon dioxide to limited organic compounds such as acetate and 2-oxobutyrate [13].
1.1.3 Direct Interspecies Electron Transfer

Insights from investigation into microbe-electrode interactions have also impacted understanding of functioning of anaerobic ecosystems. The nutritional dependencies between microorganisms govern the functioning of these ecosystems and thereby play a major role in regulating the global biogeochemical cycles. Syntrophy is one such microbial association that is central to the functioning of a diversity of methanogenic environments such as wetlands, aquatic sediments, animal intestinal systems, and oil reservoirs, as well as in anaerobic digesters converting organic wastes to methane. As recently reviewed in detail [14], the metabolism of many organic compounds in methanogenic systems, such as fatty acids and aromatic compounds, requires cooperation between the microorganisms metabolizing those compounds and methanogens. Furthermore, the metabolism of compounds that can be fermented, such as sugars and amino acids, is also often improved by syntrophic interactions between fermentative microorganisms and methanogens. This is because although the overall conversion of fatty acids, aromatics, and fermentative compounds to methane and carbon dioxide is thermodynamically favourable, methanogens cannot effectively utilize any of these multi-carbon compounds, with the exception of the simplest organic acid, acetate. Therefore, other microorganisms must metabolize the organics that methanogens cannot utilize to acetate. This requires a partial anaerobic oxidation of the organic compounds, typically with the production of carbon dioxide and reducing equivalents that the cell must dispose in order to regenerate intracellular electron acceptors, such as NAD, FAD, and ferredoxin, that are required for continued metabolism. The classic, best understood method for the regeneration of reduced electron carriers is the reduction of protons to produce hydrogen ([15], [16], [14]). The oxidation of most intracellular electron acceptors with the reduction of protons is energetically unfavourable at standard conditions and thus hydrogen can only be produced if methanogens rapidly consume the hydrogen produced and maintain hydrogen at concentrations low enough to make their production energetically favourable. The concept of interspecies hydrogen transfer was first elucidated over 40 years ago and many studies have provided evidence consistent with interspecies hydrogen transfer or
the functionally equivalent interspecies formate transfer ([15], [16], [14]).
However, recent studies have described the possibility of direct interspecies electron
transfer (DIET) as an alternative to the long-standing paradigm of interspecies hy-
drogen transfer in some methanogenic environments [17, 18, 8]. Defined co-cultures
of *Geobacter metallireducens* and *Geobacter sulfurreducens* grow syntrophically via
DIET in medium with ethanol as the electron donor and fumarate as the electron
acceptor [19, 20, 21].

The exact mechanism of electron transfer during DIET or during micro-
bial electrosynthesis is not fully known yet. However, it has been shown that in
*Geobacter* species, different outer membrane cytochromes and pili genes are over-
expressed under conditions of electron transfer to and from electrodes [22]. This
suggests that the mechanisms for electron transfer from electrodes to the microbe
could differ significantly from the mechanisms for electron transfer to electrodes. It
also remains to be seen which of these mechanisms would *Geobacter* species adopt
in the case of electron transfer to syntrophic partners.

This thesis aims to adopt a genome-scale systems biology approach to char-
acterize various metabolic and regulatory aspects of these three applications of
electromicrobiology.

### 1.2 Understanding the microbial genotype - phenotype relationship using genome-scale systems biology approaches

One of the central goals of systems biology is to achieve a fundamental
understanding of relationship between genotype and phenotype. The advent of
genome-sequencing proved to be a turning point in enabling our understanding
of this fundamental relationship. Full genome sequences provide comprehensive,
albeit not yet complete, information about the genetic elements that create an
organism. The comprehensive understanding for some cellular processes, such as metabolism, has resulted in structured knowledge-bases that can be mathematically represented [23, 24, 25]. This mathematical representation enables the computation of phenotypic states [26, 27, 28, 29] based on genetic and environmental parameters. Remarkably, this provides a mechanistic representation of the microbial metabolic genotype-phenotype relationship.

Constraint-based models of genome-scale metabolic networks capture the genotype-phenotype relationship by simultaneously accounting for constraints on phenotype imposed by physicochemical laws and genetics. The realization that these quantitative genotype-phenotype relationships could be constructed from a genome has driven the emergence of this area of research, and the flood of increasingly rich high-throughput data has accelerated the evolution of constraint-based reconstruction and analysis (COBRA) methods from a set of basic tools for metabolic network analysis into a powerful analytical framework that is increasingly used.

The COBRA approach is based on a few fundamental concepts. These concepts include the imposition of physicochemical constraints that limit computable phenotypes (Fig. 1.2a-d), the identification and mathematical description of evolutionary selective pressures (Fig. 1.2e), and a genome-scale perspective of cell metabolism that accounts of all metabolic gene products in a cell (Fig. 1.2d,f). These fundamental concepts are briefly described here.

Metabolism is a complex network of biochemical reactions. The reaction occurrence is limited by three primary constraints: reaction substrate and enzyme availability, mass and charge conservation, and thermodynamics. For metabolism, reaction substrates must be present in a cell’s microenvironment or produced from other reactions, and enzymes must be available. Mass conservation further limits the possible reaction products and their stoichiometry, while thermodynamics constrain reaction directionality. For a given organism, this information can be obtained from careful biochemical and genetic studies or inferred from related organisms, and then catalogued in metabolic reconstruction knowledgebases [23, 25].
In the COBRA framework, a metabolic reconstruction is converted into an *in silico* model by mathematically describing the reactions and adding network inputs and outputs (e.g., uptake and secretion products). Much like a cell has one genome and many transcriptional states, an organism has one metabolic reconstruction from which context-specific models can be derived, each representing cellular functions under different conditions.
Figure 1.2: Fundamentals of the genome-scale metabolic genotype-phenotype relationship: The COBRA approach is based on three primary fundamental concepts: network constraints (a-d), objective functions (e), and the association of reactions with the genome. (a) A complex mixture of molecules (red) can react to yield end products (blue). (b) The stoichiometry of this reaction network is described mathematically in a stoichiometric matrix, with each column representing the stoichiometry of a reaction. Negative and positive values represent reactants and products, respectively. Reaction flux is limited by thermodynamics and catalytic capacities ($V_m=V_{max}$), described by upper and lower bounds on flux for each reaction (green). (c) Reaction constraints result in a solution space that contains all feasible flux distributions. Additional constraints (e.g., mass balance, the steady-state assumption, and measured metabolite consumption rates) reduce the space of feasible flux distributions, as shown by the pink line. (d) In vivo biochemical networks involve additional complexity. Gene regulation can change the abundance of catalysts (e.g., the transformation of D to E). Often components are also localized in different organelles (e.g., E and F), thereby blocking reactions. (e) The biomass objective function describes an evolutionary pressure for microbial growth, and describes the metabolic demands to make basic metabolite building blocks for all cellular components (e.g., membranes, macromolecules, ATP, etc.). (f) The association of metabolism with the genome is done by mathematically linking the genome to transcripts, proteins, and chemical reactions. The gene-protein-reaction schema is used to describe gene association in the models, and provide an interface for the integration of high-throughput data.
Physicochemical constraints on the metabolic network are mathematically described by a matrix representing the stoichiometric coefficients of each reaction ((Fig. 1.2a-b). Known upper and lower bounds on each reaction flux are imposed as additional constraints. Mathematically, these constraints define a multidimensional solution space of allowable reaction flux distributions, and the actual expressed flux state resides in this solution space. Additional constraints can further shrink the solution space to focus in on the actual flux state of the network ((Fig. 1.2c). These additional constraints may include enzyme capacity, spatial localization, metabolite sequestration, and multiple levels of gene, transcript, and protein regulation ((Fig. 1.2d).

Constraint-based modeling has rapidly developed since the advent of whole-genome sequencing [30, 31]. A genome provides the genetic basis for an organism’s metabolic network, and genome annotation defines the relationships between genes, enzymes, and the reactions they catalyze ((Fig. 1.2f) [32]. Annotated genomes and their associated biochemical and genetic data have facilitated the development of carefully curated metabolic network reconstructions containing thousands of reactions. When a reconstruction knowledgebase for an organism is converted into a genome-scale model (GEM), the mathematical representation provides constraints, and the objective function can be used to represent the optimal biological functions the organism strives to achieve. Thus, simulation of an organism’s phenotypes can be performed using its GEM.

The genome-scale view of metabolism of these models has two primary implications. First, in principle, they account for all known metabolic genes in a cell. Thus, when used in genome-scale dataset analysis (e.g., proteomics, metabolomics, etc.) [33], they provide novel insight since they account for real chemical connections between components (Fig. 1.2f). Second, since metabolic genes are associated with the biochemical functions of their gene products, simulations of metabolite flow through the network can provide mechanistic predictions of how each gene product affects the metabolic network function. Thus, cell phenotypes can be computed and data can be interpreted with GEMs, thereby providing mech-
anistic insight into how the cell genotype may contribute to the cell phenotype. Furthermore, the recent advances in genomic techniques have provided unprecedented access to data spanning multiple levels of cellular organization such as the transcriptome and the proteome. Since reconstructed networks leverage genomic data for insight and phenotype prediction, the development of COBRA methods has accelerated, and evolved into over hundreds of methods with various applications including biological discovery, metabolic engineering, studying bacterial evolution.

1.2.1 Application of COBRA methods to characterize microbial community interactions

Individual COBRA methods can answer numerous scientific questions. However, multiple methods can be deployed in parallel to obtain additional insights into a question of interest. Moreover, different models can be easily swapped or combined to test hypotheses relevant to different species. Thus by using a community of methods and several data types, deeper insights into larger questions may be attained. For example, COBRA methods have complemented each other and provided insight into microbial community interactions.

The community structure in an organisms microenvironment can shape metabolic pathways usage. Organisms compete for scarce resources or depend on the metabolic capabilities of their cohabitants. Evolution often selects for cells that leverage this community structure [4]. COBRA methods are now characterizing metabolisms role in microbial community structure [34, 35, 36]. These studies are providing insight into mutualism [37], competition [2], parasitism [38, 39], and community evolution [4, 40].
**Figure 1.3: Integrating COBRA methods to study community interactions** COBRA methods are providing insight into the metabolic interactions in various types of microbial communities. (a) To study the mutualistic behavior of co-dependent mutant *E. coli*, researchers used MOMA [1] to simulate synergistic growth of pairs of auxotrophic *E. coli*. (b) Shadow prices from FBA simulations of these pairs were used to compute cooperation efficiencies between strains, which were subsequently compared with measured fitness improvements. (c) Competition in communities was modeled using DMMM [2] to understand how communities of *Geobacter* and *Rhodoferax* compete for resources, and how the demographics vary under different nutrient ratios, thereby affecting the efficiency of bioremediation efforts. Host-pathogen interactions between *M. tuberculosis* and a human macrophage were studied using COBRA. (d) While transcriptomic data were employed to build host-pathogen models at different stages of infection, the cellular objective of internalized *M. tuberculosis* is not known, so refinements to the objective function were predicted from transcriptomic data to account for changes in required amounts of compounds like lipids and amino acids (AAs). (e) This information was used to compute flux states of internalized *M. tuberculosis* with MCMC sampling [3]. This demonstrated a suppression of central metabolism and activation of the glyoxylate shunt, represented here by enolase and isocitrate lyase, respectively. The role of communities in evolution has been studied using Reductive evolutionary simulation [4]. In particular, this method predicted the minimal set of genes needed to for *Buchnera* to grow in the rich innards of the aphid. The predicted minimal gene sets (f) and temporal order of gene loss (g) were consistent with the gene content and phylogenetic structure of several *Buchnera* species.
**Mutualism**

Synthetic mutualism between auxotrophic *E. coli* mutants was recently studied using COBRA methods [37]. The authors grew pairs of auxotrophic mutants and then modeled their coupled metabolism using MOMA to identify mutant pairs that exchange essential metabolites to improve growth (Fig. 1.3a). FBA shadow prices demonstrated the balance between the cost (from metabolite loss) and the benefit (from receiving missing essential metabolites) to each rescued auxotroph. The cooperative efficiency (i.e., the ratio of uptake benefit to production cost) recapitulated the observed growth of the co-cultures. Substantial increases in growth (Fig. 1.3b) were witnessed in co-cultures that exchanged beneficial, but less costly metabolites (i.e., higher cooperative efficiency). Although it is difficult to directly measure metabolite exchange between the auxotrophs, the computed cooperation efficiency provides an indirect quantitative assessment of the metabolite cross-feeding in this mutualistic system.

**Competition**

Metabolic competition for scarce nutrients has also been assessed with COBRA methods. Dynamic multi-species metabolic modeling (DMMM) characterized the competition for acetate, Fe(III), and ammonia between *Geobacter sulfurreducens* and *Rhodoferax ferrireducens* in subsurface anoxic environments (Fig. 1.3c) [2]. DMMM simulates the growth rate of both organisms and the rates of change of external metabolites, to dynamically predict population changes in the community. Using DMMM, the community composition was predicted under geochemically distinct conditions of low, medium, and high acetate flux. Under low acetate flux, DMMM predicts *Rhodoferax* dominates the community when sufficient ammonia is available, whereas *Geobacter* dominates under low ammonia and high acetate flux. This difference was attributed to the nitrogen fixation abilities of *Geobacter*, as well as its higher acetate uptake rate compared to *Rhodoferax*. Moreover, it was also predicted that under nitrogen fixing conditions, *Geobacter* increases its respiration at the expense of biomass production, thus showing how balancing community structure can impact the efficacy of uranium bioremediation.
in low ammonium zones.

**Parasitism**

Host-pathogen interactions have been studied with COBRA methods [38]. A recent study modeled the metabolic interactions between a human alveolar macrophage and *M. tuberculosis* [39]. Context-specific models of infection were built with GIMME [41] and Shlomi-NBT-08 [42] using transcriptomic data from three types of *M. tuberculosis* infections. Next, the *M. tuberculosis* objective function was revised using infection-specific gene expression data to better represent the metabolic activity of the internalized pathogen (Fig. 1.3d). Gene deletion analysis was compared with in vivo gene essentiality data, and MCMC sampling was also used to demonstrate a substantial alteration in metabolic pathway usage in *M. tuberculosis* during macrophage infection, including a suppression of glycolysis and an increased dependency on glyoxylate metabolism (Fig. 1.3e). This constraint of central metabolism during *M. tuberculosis* infection was also suggested by DCP, another method related to FBA [43]. This suppression of certain metabolic pathways with an increased dependency on normally latent pathways may provide novel antibiotic targets.

**Community evolution**

In evolution, genetic drift and selective pressures cause organisms to optimize their cellular machinery for a particular niche [44]. This assumption of cellular optimization has made COBRA methods useful tools to investigate hypotheses concerning organismal evolution, as reviewed by Papp, et al. [29]. In nature, the optimization of microbial metabolism is a multi-species affair, as demonstrated by the aphid endosymbiont *Buchnera aphidicola*. This descendant of the *Enterobacteriaceae* family has suffered drastic loss of genomic material as it evolved in its hosts nutrient-rich innards. Since *B. aphidicola*is related to *E. coli*, reductive evolutionary simulation (a gene deletion analysis derivative) [4] on the *E. coli* model provided minimal metabolic gene set predictions. These predicted minimal sets are highly consistent with the metabolic gene content of *B. aphidicola* (Fig. 1.3f).
In addition, the predicted temporal order of gene loss was significantly consistent with the phylogenetically reconstructed gene loss timing among the genomes of five *Buchnera species* (Fig. 1.3g) [40], thus suggesting that the bacterium optimized its pathway usage for its new rich habitat. Interestingly, metabolic pathways retained in the computed minimal gene sets highlight the bacterium's role in symbiotic evolution. Retained pathways contained reactions needed for producing riboflavin and essential amino acids lacking from the aphid diet, thereby highlighting their role in the symbiotic relationship [4]. Thus, COBRA methods are helping to describe how the community shapes gene content in evolving symbiotic communities [29].

### 1.3 Extending systems biology to characterize electromicrobiological applications

Genome-scale metabolic reconstructions have been used for a wide range of applications [28] mainly facilitated by constraint-based modeling (COBRA) [45]. The COBRA process consists of a network reconstruction workflow followed by its conversion into a mathematical format. The reconstruction and modeling techniques have been well developed and documented over the last ten years ([46], [25], [47]). This technique was applied first on *Geobacteraceae* to build the genome-scale metabolic model of *Geobacter sulfurreducens* [48] in an initial attempt to determine if genome-scale models could be used to predict bioremediation. This model has been used to describe growth under a diversity of conditions ([49]); predict the phenotype of gene deletions with 90% success ([50]); and was a valuable tool for metabolic engineering [51]. More recently, the metabolic model for *Geobacter metallireducens* was also developed [52]. Additional modeling has provided important insights into the factors leading to effective uranium bioremediation by Geobacter species ([53]). The application of genome-scale models has greatly improved understanding of bioremediation at the Rifle Site CO. However, there has not been much progress in terms of modeling microbe electrode interactions with *Geobacteraceae*. Therefore, the primary goal of the research described here is to extend the scope of COBRA approaches and genome-scale systems biol-
ogy methods to characterize electromicrobiological applications.

1.3.1 Characterizing microbial fuel cells

As mentioned earlier, while a large number of cytochromes (111) that facilitate the electron transfer to electrodes in microbial fuel cells, little is known about the explicit mechanism of all of these outer membrane cytochromes and their role in electron transfer. Attempts to improve current production by genetic manipulation and overproduction of cytochromes or nanowires were met with little success.[51]. This indicates that current production by bacteria in a microbial fuel cell is a highly regulated process that requires more changes than in a few genes related to bacterial respiration.

On the other hand, adaptive evolution and selection pressures have met with greater success. By applying a selection pressure of -400mV at anodes, a novel strain KN400 was isolated from the biofilm after 5 months. This strain was found to be capable of enhanced electricity production (8 times higher power production) and also possess clear phenotypic differences in the outer surface that interacts with the electrode [54]. In order to completely characterize KN400 and understand the basis for the observed phenotype it is extremely important to have the complete genome sequence.

Chapter 2 of this thesis describes the application of next-generation sequencing (NGS) technologies to de novo assemble the complete genome sequence of an enhanced electricity producing variant of *G. sulfurreducens*. Chapter 3 of this thesis pertains to the characterization of regulatory networks of *G. sulfurreducens* at multiple levels of complexity and at diverse growth conditions. Specifically, this chapter deals with the properties of the sigma factor regulatory network at a genome-scale for *G. sulfurreducens* growing as an electrogenic biofilm.
1.3.2 Characterizing microbial electrosynthesis

In order to completely understand the process of microbial electrosynthesis and successfully apply it for metabolic engineering efforts to produce liquid transportation fuels and commodity chemicals, it is imperative to characterize the metabolic and energetic constraints on carbon fixation and electron transfer pathways. This is of prime importance especially because, the metabolism of the bacteria acting as the catalyst is the effective rate limiting step for the process. Constraint-based metabolic modeling and analysis has been a valuable tool for discovering and understanding new capabilities and content in bacteria, as well as in guiding synthetic biology efforts for targeted production. Chapter 4 presents the genome-scale reconstructions of two such electrosynthetic bacteria and also the description of a modeling framework to characterize microbial electrosynthesis.

1.3.3 Investigating direct interspecies electron transfer in syntrophic associations

While several biochemical approaches have been able to describe the overall physiological activity and mechanisms of many syntrophic associations [55, 56], there has been a lack of a mechanistic understanding of the relationship between the complex nutritional and energetic dependencies and their functioning. This calls for the development of genome-scale approaches that can provide a link between the genotypes of the microbes, their reaction mechanisms and eventually their effect on the functioning of the microbial community [57, 58]. One of the major goals of microbial ecology and systems biology is to comprehensively understand, characterize and obtain meaningful insights into the various modes of interactions in communities and their functional effect on the ecosystem [59]. However, the full realization of the potential of CoSy biology has been impeded by the lack of a comprehensive characterization of the various mechanisms of energy transfer and their effect on the functioning of the microbial associations. In Chapter 5, a multi-omics modeling workflow that combines genomic, transcriptomic, and phenotypic data with constraint-based modeling using genome-scale metabolic models
to investigate metabolic dynamics and mechanisms of electron flow in syntrophic co-cultures is presented. This workflow also includes the modeling of direct interspecies electron transfer.

Chapter 2

*De novo* assembly of the complete genome of an enhanced electricity producing variant of *G. sulfurreducens*

2.1 Abstract

State-of-the-art DNA sequencing technologies are transforming the life sciences due to their ability to generate nucleotide sequence information with a speed and quantity that is unapproachable with traditional Sanger sequencing. Genome sequencing is a principal application of this technology, where the ultimate goal is the full and complete sequence of the organism of interest. Due to the nature of the raw data produced by these technologies, a full genomic sequence attained without the aid of Sanger sequencing has yet to be demonstrated.

We have successfully developed a four-phase strategy for using only next-generation sequencing technologies (Illumina and 454) to assemble a complete microbial genome *de novo*. We applied this approach to completely assemble the 3.7 Mb genome of a rare *Geobacter* variant (KN400) that is capable of unprecedented current production at an electrode. Two key components of our strategy enabled
us to achieve this result. First, we integrated the two data types early in the process to maximally leverage their complementary characteristics. And second, we used the output of different short read assembly programs in such a way so as to leverage the complementary nature of their different underlying algorithms or of their different implementations of the same underlying algorithm.

The significance of our result is that it demonstrates a general approach for maximizing the efficiency and success of genome assembly projects as new sequencing technologies and new assembly algorithms are introduced. The general approach is a meta strategy, wherein sequencing data are integrated as early as possible and in particular ways and wherein multiple assembly algorithms are judiciously applied such that the deficiencies in one are complemented by another.

2.2 Introduction

The sequencing of the first bacterial genome in 1995 has left a lasting impact on the field of prokaryotic genomics. The next revolution in the field of genomics has been the development and progress of high-throughput sequencing technologies. These next-generation sequencers, mainly those from Illumina and 454 Life Sciences (454), generate millions of short reads that are more error-prone than the traditional Sanger sequencing. However, these technologies have greatly reduced the cost of sequencing per base and thus have opened up a wide range of applications. The major applications include resequencing of closely related individuals for personalized genomics and \textit{de novo} sequencing of new microbial genomes.

\textit{De novo} sequencing using next-generation technologies has necessitated the development of new algorithms for assembling the short and more error-prone reads that they generate. Several \textit{de novo} assembly algorithms based on de-Bruijn graphs (EULER-SR [60] and Velvet [61]), hash-extension (VCAKE) [62], overlap layout (EDENA) [63] and for paired-end reads (ALLPATHS) [64] have been recently developed. These algorithms are capable of assembling millions of short-reads from next-generation sequencing technologies into thousands of contigs with varying de-
degrees of efficiency.

While 454 reads are longer than Illumina reads (~250-450 bp compared to ~36-100 bp), they have a higher indel error rate when compared to Illumina reads. The longer 454 reads, though, inherently offer advantages over the shorter Illumina reads for *de novo* assembly. Illumina reads, despite being much shorter, provide a higher depth of coverage than 454 reads. This complementary nature of Illumina and 454 reads has been exploited by some recent methods that have produced an assembly of *P. syringae pathovar oryzae*, consisting of 126 scaffolds, 2002 unincorporated contigs, and an N50 of 91.5 kb [65]. Another report integrated these two data types using a different approach to assemble an *Acinetobacter baylyi* strain into 10 scaffolds with an N50 of 1Mb [66]. Salzberg and colleagues assembled a virulent strain of *P. aeruginosa* PA01 into a large 6.3 Mb scaffold [67] using a mixed comparative and *de novo* approach that included a gene-boosted strategy. However, this approach heavily relied on comparative information and thus cannot be classified as *de novo*.

Despite these recent reports that indicate significant progress by integrating Illumina and 454 technologies, complete *de novo* assembly of microbial genomes from only short reads and without aid from Sanger sequencing still remains an unsolved challenge. This challenge is critically important [68], for a single, circular nucleotide sequence of the complete chromosome is a necessary prerequisite for confident and complete research based on a genome. As an answer to this challenge, we have developed a strategy (Meta-Assembly) for complete, whole-genome *de novo* assembly and applied it to a novel *Geobacter* variant (KN400) that is capable of unprecedented current production at an electrode [54]. Our Meta-Assembly strategy adopts a bi-level integrative approach that leverages the different and complementary results provided by multiple assembly programs over and above the integration of complementary data types to obtain a complete, whole-genome assembly. We have applied this strategy using 50X Illumina GA1 singleton reads and 16X 454 GS-FLX paired-end sequencing reads and assessed our finished as-
assembly by sequencing nearly 1% of the genome by Sanger sequencing and by a comparison to the genome of a highly related strain.

2.3 Results

2.3.1 Meta-Assembly strategy

Our Meta-Assembly approach (Figure 2.1) consists of four distinct phases: Hybrid Assembly, Scaffold Bridging and Finishing, Scaffold Ordering and Genome Finishing.

Figure 2.1: Assembly Strategy: A) Hybrid Assembly Phase; B) Scaffold Bridging and Finishing Phase; C) Scaffold Ordering Phase: The left branch of the decision tree consists of permutations that can be confirmed by the PCR performed while the right branch consists of those permutations that cannot be confirmed by the particular PCR. The faded permutations are those which have been eliminated by the PCRs while those in bold are those that are remaining. (Gel Inset: Showing PCR products for all the 9 PCRs performed in the search strategy to confirm the correct orientation of the scaffolds); D) Genome Finishing Phase.
In the Hybrid Assembly phase (Figure 2.1A), we first filtered and assembled the Illumina reads alone using the de-Bruijn graph based algorithm EULER-SR [60]. This assembly consisted of 4233 contigs with an N50 of 1.48 Kb. We then assembled these contigs, along with all of the reads not assembled by EULER-SR [60] and all of the 454 reads (i.e. neglecting the pair information) using Roche’s algorithm Newbler. This combining step was a critical aspect in maximizing the complementary information in Illumina and 454 reads, as shown by the resulting assembly of 270 hybrid contigs with an N50 of 92.67 kb (Table 2.1). We then leveraged the mate pair information by combining these 270 “hybrid” contigs with the paired 454 reads using Newbler’s scaffolder. The contigs that did not form part of one of the output scaffolds (unscaffolded contigs) were utilized later in the final Finishing phase. This scaffolding step resulted in a greatly improved assembly, giving three de novo scaffolds of lengths 3.18 Mb, 5.7 kb and 524 kb (respectively scaffolds A, B, and C in Fig. 2.1A) with a total length of 3.7 Mb.

Exploiting the complementary nature of assembly algorithms significantly improves the quality of the de novo scaffolds

The de novo scaffolds A and C from the Hybrid Assembly phase contained numerous stretches of degenerate nucleotides, and to resolve them we applied a post-processing step that exploited the coverage provided by short-reads. We developed a Scaffold Bridging and Finishing phase for the purpose of linking the de novo scaffolds and for resolving the intra-scaffold degenerate nucleotide positions that were introduced by the scaffolder (Fig. 2.1B). In this phase, we leveraged the complementary nature of the assemblies generated by programs like EULER-SR [60], Velvet [61] and Newbler. Since EULER-SR [60] and Newbler generate slightly different sets of contigs, we created a second set of hybrid contigs from Illumina and 454 reads using EULER-SR. We aligned these hybrid contigs against the de novo scaffolds using NUCMER and analyzed the alignment for the three scenarios that could potentially bridge the scaffolds and resolve the degenerate nucleotides (Methods and Fig. 2.2). We found that none of the hybrid EULER-SR [60] contigs aligned in such a way that they bridged any pair of de novo scaffolds (Fig.
2.2A). We were able to resolve all intra-scaffold degenerate nucleotide positions by either substituting the corresponding bases from hybrid EULER-SR [60] contigs that overlapped with flanking regions of Ns in the \textit{de novo} scaffolds (Fig. 2.2B), or by removing the degenerate bases when the regions flanking them aligned to contiguous regions on the hybrid EULER-SR contigs (Fig. 2.2C). (We implemented the same approach using hybrid contigs generated by Velvet [61] as the complementary set instead of EULER-SR [60] and obtained a similar result.) At this stage, our assembly could contain small indels due to 454 sequencing or due to our custom program. To correct these, we aligned the Illumina reads using the Smith-Waterman capabilities of MosaikAligner (Stromberg and Marth in preparation). We also analyzed these scaffolds for potential repeats/duplications by examining the read coverage and also the multiplicity of the vertices in the repeat graph that is part of EULER-SR’s output. This analysis revealed that scaffold B was indeed duplicated and a BLAST [69] search identified it as an rRNA gene. At the end of this phase, then, our assembly consisted of four scaffolds, identified in Fig. 2.1B as A (3.15 Mb), B (5.7 Kb) occurring twice, and C (518 Kb).

**Figure 2.2:** Custom program used in the Scaffold Bridging and Finishing Phase of the assembly process.
Table 2.1: Summary and Statistics of different stages of Meta-Assembly. Phase A is the Hybrid Assembly that comprises of three distinct steps.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Assembler</th>
<th>No. of Scaffolds</th>
<th>N50(kb)</th>
<th>Degenerate Positions</th>
<th>Assembly Size(Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>EULER-SR(Illumina Alone)</td>
<td>4233</td>
<td>1.487</td>
<td>0</td>
<td>3.51</td>
</tr>
<tr>
<td>A</td>
<td>Newbler(454 Reads + Illumina)</td>
<td>270</td>
<td>92.67</td>
<td>0</td>
<td>3.72</td>
</tr>
<tr>
<td>A</td>
<td>Newbler Scaffolder (Mate Pairs)</td>
<td>3</td>
<td>3184.3</td>
<td>41421</td>
<td>3.71</td>
</tr>
<tr>
<td>B</td>
<td>Scaffold Bridger/Finisher</td>
<td>4</td>
<td>3184.3</td>
<td>0</td>
<td>3.71</td>
</tr>
<tr>
<td>C</td>
<td>Scaffold Ordering</td>
<td>1</td>
<td>3714.2</td>
<td>0</td>
<td>3.71</td>
</tr>
<tr>
<td>D</td>
<td>Finisher</td>
<td>1</td>
<td>3714.2</td>
<td>0</td>
<td>3.71</td>
</tr>
</tbody>
</table>

An efficient PCR-based search strategy results in the correct orientation of the scaffold and a circular genome

In the Scaffold Ordering phase, we considered the KN400 genome to be a signed circular permutation of the four scaffolds—giving 24 unique possible permutations (Fig. 2.1C). To determine the correct relative orientation of the scaffolds, we employed a polymerase chain reaction (PCR)-based search strategy. Our approach consisted of nine PCRs, six of which serially eliminated 23 possible permutations. The orientation ABcB was confirmed by the remaining three PCRs (Figure 1C). This PCR-based ordering approach enabled us to link the four scaffolds into one circular chromosome of 3.71 Mb. This approach must not be confused with the standard gap-closing approaches adopted, because we do not use PCRs to fill gaps but only to confirm the relative orientation of the scaffolds and just link them up. That is, there were no intervening nucleotides between the four scaffolds. In fact, our approach goes a step further in validating the sequence by accounting for the reverse complement of the scaffolds, thereby eliminating any potential false rearrangements that might be introduced in the assembly.
Depth of coverage offered by Illumina reads corrects the indels introduced by 454 and scaffold finishing and ordering:

We corrected for indels and any errors introduced during our scaffold finishing and scaffold ordering phase by aligning the Illumina reads to the ordered scaffold ABcB using MosaikAligner (Stromberg and Marth in preparation). The result was a complete circular genome consisting of 3,714,272 bp. The statistics of changes made by this alignment are provided in Table 2.2. The genome can be accessed from GenBank under the accession number CP002031.

Table 2.2: Changes made due to alignment of Illumina reads in the Genome-Finishing Phase

<table>
<thead>
<tr>
<th>Changes</th>
<th>Number of Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs</td>
<td>101</td>
</tr>
<tr>
<td>Deletions</td>
<td>18</td>
</tr>
<tr>
<td>Insertions</td>
<td>7</td>
</tr>
</tbody>
</table>

2.3.2 Assembly validation

To validate the de novo assembly approach adopted here and to estimate the accuracy of the obtained genome sequence, we amplified 32kb (~1% of the genome) of KN400 and performed Sanger sequencing on both the forward and reverse strands (Fig. 2.3A). We compared these sequenced regions to the corresponding genomic region obtained from Meta-Assembly using megaBLAST [70] alignment algorithm. We found that out of the 32680 bp sequenced by Sanger sequencing, there were 32675 perfect matches, four SNPs and one 1 bp insertion with respect to the assembled KN400 genome sequence. All of these differences occur in a 30 bp region of the genome that is covered by just one 454 read. This means that the accuracy of this short region is a direct reflection of the quality of the single overlapping read. For this work, we utilized first generation short read technologies, and read quality and quantity have dramatically improved thereafter. Such low coverage regions are rare even in our assembly, and with contemporary data output they would likely be nonexistent. That is, investigators using our approach with the improved read data would almost certainly not have such assembly er-
rors because 1) there would likely be no such regions since data output is so high and 2) the accuracy of individual reads is so much higher. It is worth pointing out that any approach based on short reads would be limited by the single 454 reads spanning this region, but what sets our approach apart is that we were able to actually place this single read in the context of assembling a circularly-closed bacterial genome.

**Figure 2.3: Assembly Validation approaches:** A) Sanger sequencing approach; B) Comparative Genomics approach.
2.3.3 **KN400 is the first microbial genome that has been completely assembled de novo using only next-generation sequencing technologies**

We predicted 3356 ORFs and 54 RNAs in the KN400 genome using the RAST pipeline [71] and manual curation. We found the completed KN400 genome to be collinear over its entire length with no major rearrangements (Fig. 2.4)–and approximately 97% identical at the sequence level to *Geobacter sulfurreducens* PCA [72]. Because the genomes were so similar, we evaluated the correctness of our *de novo* assembly by assessing the commonalities and the differences between PCA and KN400 genomes. We adopted a comparative genomics approach at the ORF level to assess the differences and similarities between KN400 and PCA (Fig. 2.3B).

![Figure 2.4: Genome-Level comparison of KN400 and PCA. Shown in this figure is a dot-plot of the genome-wide alignment of KN400 and PCA. Along the X-Axis is the KN400 genome and the PCA genome is shown along the Y axis.](image)
Meta-Assembly approach facilitates accurate prediction of genomic regions unique to PCA

We identified three primary genomic regions that were specific to the PCA strain and had no correspondence with the KN400 genome: a 32kb region between GSU0039 and GSU0064 (in KN400) (Region 1), a 78kb region between GSU2105 and GSU2183 (Region 2), and a 16 kb region between GSU2588 and GSU2601 (Region 3). The largest region, GSU2105-GSU2183, had 79 genes in PCA between the orthologs to ORFs KN400.2161 and KN400.2163. Forty-six of these genes were hypothetical, and 11 were transposases or other integrative genetic elements. The remaining genes were predicted to encode several sensors and regulators and a DNA-binding repair protein. The deletion of this region was confirmed by performing a PCR over the break (Fig. 2.5). The second largest region, GSU0039-GSU0064, had 26 genes in PCA between the orthologs to KN400.0039 and KN400.0040. Fifteen of these were hypothetical and 4 were transposases. In addition, there were 3 CRISPR-associated proteins. The third region, GSU2588-GSU2601, had 11 genes in PCA between the orthologs to KN400.2568 and KN400.2571. Four of these were transposases and seven hypothetical proteins. The abundance of transposons in these PCA-strain-specific regions suggests that they resulted from the activity of mobile genetic elements within the *Geobacter sulfurreducens* genome since the time the two strains diverged. None of these regions encoded any protein predicted to be required for growth in the PCA strain [48].
Figure 2.5: Gel Picture confirming the 79 kb deletion (Region2) in KN400. PCR was performed with primer sets in order to amplify over the break (shown in panel B). The expected product size is 207bp. Panel A shows that we can amplify over the break only in KN400 and not in PCA, confirming the deletion of region 2 in KN400.
Genomic Regions unique to KN400 relative to PCA provides insights into possible evolutionary paths for KN400

Of the 3356 ORFs predicted in KN400, we found that 3088 (92%) had complete reciprocal orthologs with the same relative ordering in PCA. Conservation of the orthologs between the two genomes was quantified by the ratio of the bit scores from a KN400-PCA to a KN400-KN400 BLAST alignment. The orthologs had an average bit score ratio of 93.0%.

We compared the remaining 268 proteins in KN400 that did not have reciprocal orthology in PCA to all proteins in the NCBI RefSeq database [73] to determine the organism which encoded the protein with the highest sequence similarity. Fifty-six proteins had no significant match in the database, indicating they were specific to the KN400 genome, and were annotated as hypothetical proteins. Thirty-six proteins were most similar to PCA, and 90 were most similar to other Geobacteraceae, indicating that they were vertically inherited. This left 86 proteins that were found in the KN400 but were most similar to a non-Geobacteraceae species. These 86 genes were found primarily in 12 small regions in KN400 that aligned poorly with the PCA genome. These regions ranged in size from 3.3 to 21.2kb, and seven of them included genes for at least one transposase or integrase. In particular, copies of the two-subunit transposase ISGsu7 were associated with strain-specific islands in both PCA and KN400. This supports the idea that these regions may also have been produced by the activity of mobile genetic elements since the two strains diverged.

C-type cytochromes play a key role in the transfer of electrons from central metabolism to external electron acceptors like Fe(III) and electrodes [74]. Comparative analysis of KN400 and PCA showed that several genes encoding cytochromes contain single nucleotide polymorphisms between the strains, including the gene for the outer-membrane cytochrome OmcS (GSU2504), which has been shown to be required for electron transfer to insoluble Fe(III)[75]. In addition, analysis of the genes specific to the KN400 strain also shows that there are at least three transport proteins that are not found in the PCA strain. Further analysis of the link between these types of genetic differences and the phenotype of the KN400
strain during electricity production is underway (Butler J.E. et al, in prep).

2.3.4 Genomic characteristics of KN400 are typical of microbial genomes at large

To assess whether KN400 was a particularly fortuitous choice for attempting a de novo assembly using only short reads, we assessed its genome complexity by performing a comparative analysis using five different genomic properties that have been characterized using 895 microbial genomes. We obtained data relating to the GC content, genome size, number of replicons, number of rRNAs and number of tRNAs for a sample of 895 microbial genomes from the Genome Atlas Database (http://www.cbs.dtu.dk/services/GenomeAtlas-3.0/). We computed the percentile ranks for KN400 genome to evaluate its relative complexity in the space of all microbial genomes across these five dimensions (Fig. 2.6). Based on size and GC content, the KN400 genome has percentile ranks of 53 (3.71 Mb) and 75 (61% G+C). Furthermore, about 60% of the microbial genomes consisted of a single replicon – as is the case with KN400. KN400 has two ribosomal RNA operons, which is the most frequent number of rRNA operons among the 895 microbial genomes. Apart from large repeats like the rRNA operons, local inverted repeats like transposases also characterize genomic complexity and thus have an effect on the assembly process. To evaluate the transposase content of KN400, we relied on a recent survey [76] of the transposases present in all available microbial genomic and metagenomic databases. This survey found that the average genome contains 11 transposases per 1kb. Based on this result, a genome the size of KN400 would be expected to contain 38 transposases. Our analysis indicated that KN400 contains 30 transposases. Based on these comparisons, we concluded that KN400 is a typical microbial genome and not an outlying “simple” genome.
Figure 2.6: Density of five different genomic properties in the space of microbial genomes. A) GC Content B) Genome Size C) Number of rRNAs D) Number of Replicons E) Number of tRNAs. Shown in red circle, is the value of KN400’s genomic property.
2.4 Discussion

De novo assembly of complete microbial genomes using new DNA sequencing technologies and without the aid of Sanger sequencing has been an unsolved challenge. We have developed a bi-level integrative approach, the Meta-Assembly, that answers this challenge.

Our Meta-Assembly strategy is composed of four key phases. In phase one we integrated Illumina and 454 reads at the very beginning of our assembly process to generate hybrid contigs, instead of using Illumina reads only for error correction of an assembly generated from just 454 reads [66]. This early integration step was very important for reducing the number of degenerate nucleotide positions (Table 2.1 & Table 2.3) and thus for the overall quality of the assembly. Incorporating the Illumina reads early in the assembly process significantly reduced the number of degenerate nucleotides in the assembly (~41000 N’s) compared to when they are used for just error correction of the assembly generated by Newbler (~90,000 N’s). In addition, we used EULER-SR [60] instead of VCAKE [62] as the short read assembler—in distinction from an earlier report [65]. The fact that de-Bruijn graph based algorithms like EULER-SR [60] and Velvet [61] outperform VCAKE [62] has been documented in an earlier study [77], and we found the same trend with our data as well. Since assembly of Illumina reads is the first step of the hybrid assembly phase, the quality of the initial assembly has the greatest impact on the outcome of the entire process. Moreover, EULER-SR [60] is also capable of performing a de novo assembly with a mixture of Illumina and 454 reads, but its performance does not degrade with increasing read length ([77]). This proved to be a significant advantage, for we were able to exploit the complementary nature of EULER-SR [60] and Newbler to develop the Scaffold Bridging and Finishing Phase—enabling us to resolve all of the degenerate nucleotides.

In the second phase, we maximized the complementary information provided by different assembly algorithms. This component of our strategy is a key distinguishing aspect of our approach. Although Newbler alone was able to assemble the reads into five scaffolds, the resulting assembly had a considerable number of degenerate positions which could not be resolved just from an error correction
Table 2.3: Comparison of the approach Meta-Assembly to other commonly used assembly programs

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Number of Scaffolds</th>
<th>N50(kb)</th>
<th>Degenerate Positions</th>
<th>Assembly Size(Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-Assembly</td>
<td>1</td>
<td>3714.2</td>
<td>0</td>
<td>3.71</td>
</tr>
<tr>
<td>EULER-SR</td>
<td>150</td>
<td>58</td>
<td>0</td>
<td>3.70</td>
</tr>
<tr>
<td>Velvet</td>
<td>329</td>
<td>45</td>
<td>486532</td>
<td>3.89</td>
</tr>
<tr>
<td>Newbler alone</td>
<td>5</td>
<td>3184.3</td>
<td>90449</td>
<td>3.72</td>
</tr>
</tbody>
</table>

In the third phase, the simple PCR-based search strategy allowed us to quickly order and orient the scaffolds into a circular genome. This is another unique aspect of our approach in that we address the problem of relative orientation of the scaffolds as well as their ordering with just a few PCRs. While we use the PCRs to order the scaffolds into a circular genome, we did not fill any gaps as no sequence information is obtained from the PCRs. We note that as technology improvements allow paired-end sequencing reads with longer inserts, the necessity of this PCR step will decrease.

In the fourth and final phase, we aligned Illumina reads against the ordered scaffold to account for indels and errors induced during the scaffold finishing phase.

To our knowledge this is the first reported de novo assembly of a complete genome using next generation sequencing technologies. Furthermore, our comprehensive comparative analysis of genomic characteristics of 895 microbial genomes reveals that KN400 is a characteristic microbial genome and is not an outlier in the space of all microbial genomes. We view our result as the demonstration of general a strategy for assembling genomes, wherein multiple data types are integrated at specific steps in the process to maximize the potential of their complementary nature and wherein multiple assembly programs are utilized such that deficiencies
in one algorithmic approach are compensated by the strengths of another algorithmic approach. As new sequencing technologies and new assembly programs become available, they can be readily incorporated in this framework. Genome assembly will remain challenging for the foreseeable future, and we view the idea of such a readily extensible meta approach as one of the most promising ways to meet this challenge.

2.5 Materials and Methods

2.5.1 Data

We utilized an electrode isolated strain of a rare variant of *Geobacter sulfurreducens* KN400 that exhibits enhanced current production relative to *Geobacter sulfurreducens* PCA [54]. For the purpose of a *de novo* assembly, we carried out both Illumina sequencing and 454 Life Sciences pyrosequencing (shotgun and paired end) for KN400. The details of the sequence data used are shown in the Table 2.4. The paired end reads had an average insert size of 3020 bp and the distribution of the fragment size is given in Fig. 2.7.
Figure 2.7: Distribution of fragment sizes from 454 paired end sequencing data used for *de novo* assembly.
Table 2.4: Details of the next generation sequencing data used for de novo assembly of KN400

<table>
<thead>
<tr>
<th>Technology</th>
<th>Number of Reads</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina GA1</td>
<td>5048331</td>
<td>50X</td>
</tr>
<tr>
<td>Unpaired 454</td>
<td>125625</td>
<td>8X</td>
</tr>
<tr>
<td>Paired End 454 (GS FLX)</td>
<td>199523</td>
<td>8X</td>
</tr>
</tbody>
</table>

2.5.2 Meta-Assembly

Hybrid Assembly Phase

We assembled Illumina GA1 reads using the de-Bruijn graph based short-read assembler EULER-SR [60] with the vertex size parameter set to 25. Prior to assembling the short-reads, we filtered them for obvious failure modes using the filterIlluminaReads script as part of the EULER-SR package [60]. We took all of the 454 reads as singletons (i.e. neglecting the paired end information) and assembled them using Roche’s assembler (Newbler) with default parameters (minimum overlap length 40, minimum overlap identity 90%). In order to integrate the Illumina and 454 technologies, we combined the contigs generated from the EULER-SR [60] and Newbler assemblies using the incremental assembly option in Newbler to generate a set of hybrid contigs. At this stage, we leveraged the mate-pair information provided by the 454 reads in order to build scaffolds from these hybrid contigs.

Scaffold Bridging & Finishing Phase

In order to maximize the complementary nature of the assembly algorithms, we developed a meta-approach that reconciles the assemblies produced by either EULER-SR [60] or Velvet [61] and Newbler. The de novo scaffolds generated from phase A are largely due to Newbler and contain a lot of degenerate nucleotides. We created a second set of hybrid contigs from both the Illumina and 454 reads using EULER-SR with a vertex-size of 25. Using the NUCMER alignment tool of MUMMER package [78], we aligned these hybrid EULER-SR [60] contigs against
the de novo scaffolds with the break length parameter of 10,000. We further implemented a custom program (Fig. 2.2) which accounted for the following three scenarios, in order to link up the de novo scaffolds and resolve the degenerate nucleotides.

- If any of these contigs aligned in such a way that they were bridging any two of the de novo scaffolds, those scaffolds were linked up (Fig. 2.2A).

- In the event that these contigs overlapped with the flanking regions of degenerate nucleotides (N’s), we replaced the N’s with the corresponding region of the contig.(Fig. 2.2B)

- If the regions flanking the degenerate positions aligned to contiguous regions in the contig set, we removed that stretch of degenerate nucleotides (Fig. 2.2C).

We further augmented this custom program with an alignment of Illumina reads against these scaffolds in order to account for the indel errors due to 454 sequencing. We used the MosaikAligner in the “all” alignment mode with a hash size of 13 and 20 bp as alignment candidate threshold, allowing a maximum of 4 mismatches.

**Scaffold Ordering Phase**

We considered the KN400 genome as a signed circular permutation of the finished scaffolds and adopted a PCR based scaffold ordering approach in order to orient them into a circular genome. We designed a search strategy comprised of nine PCRs in order to determine the correct orientation. For performing the PCRs, total G. sulfurreducens (KN400) genomic DNA was prepared using the MasterPure Complete DNA Purification kit (Epicentre Biotechnologies, Madison, WI) according to manufacturer’s directions. Taq DNA polymerase (QIAGEN Inc., Valencia, CA) was used for all PCR amplifications. The sequence and the location details of the primer sets used for the PCRs are provided in Table 2.5, while the combinations of the primer pairs for each PCR and the expected amplicon size are given in Table 2.6.
Table 2.5: Sequence and location of primers. (Lower case denotes reverse complement)

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Location, Scaffold</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2F(b1R)</td>
<td>TCAGCAACTCCCCCTACTCCCATCA</td>
<td>end,B (beginning,b)</td>
</tr>
<tr>
<td>B1R(b2F)</td>
<td>CATCATCTCGCCTGCGGTGTCA</td>
<td>beginning B,(end b)</td>
</tr>
<tr>
<td>A2F</td>
<td>CTGGCCGAGGGCGAAGCTGAAAGAC</td>
<td>end,A</td>
</tr>
<tr>
<td>A1R</td>
<td>TGACGCGACTCCTGTAGGACCTC</td>
<td>beginning,A</td>
</tr>
<tr>
<td>C2F(c1R)</td>
<td>CGTGAAGCGATGTGCGAATGCGAAC</td>
<td>end,C (beginning,c)</td>
</tr>
<tr>
<td>C1R(c2F)</td>
<td>GGGGCTCGTGAAGTCACGAGTGCA</td>
<td>beginning C,(end c)</td>
</tr>
</tbody>
</table>

Table 2.6: Primer pairs and product analysis of each PCR during the Scaffold Ordering Phase

<table>
<thead>
<tr>
<th>PCR</th>
<th>Primer pairs</th>
<th>Product Observed (Y/N)</th>
<th>Amplicon size(if observed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB/bb</td>
<td>B2F,B2R</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>Bb</td>
<td>B2F,b1R</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>bB</td>
<td>b2F,b1R</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>Ab</td>
<td>A2F,b1R</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>BC</td>
<td>B2F,C1R</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>cB</td>
<td>c2F,B1R</td>
<td>Yes</td>
<td>351bp</td>
</tr>
<tr>
<td>AB</td>
<td>A2F,B1R</td>
<td>Yes</td>
<td>313bp</td>
</tr>
<tr>
<td>Bc</td>
<td>B2F,c1R</td>
<td>Yes</td>
<td>387bp</td>
</tr>
<tr>
<td>BA</td>
<td>B2F,A1R</td>
<td>Yes</td>
<td>404bp</td>
</tr>
</tbody>
</table>

Genome Finishing Phase

In order to account for any possible indel errors due to the 454 sequencing as well as our Meta-Assembly approach, we re-aligned all of the Illumina reads to the ordered scaffold using MosaikAligner (Stromberg and Marth in preparation) with the same parameters as described in Phase B of the Meta-Assembly approach.

2.5.3 Validation by Sanger Sequencing

We validated our assembly approach and computed error rates by performing Sanger sequencing of about 1% of the KN400 genome.
2.5.4 Gene Prediction and Comparative Analysis

At each step of the finishing process, we predicted open reading frames (ORFs) in the KN400 draft genome using the Rapid Annotation by Subsystem Technology (RAST) pipeline [71] in order to check if the standard properties like gene density were characteristic of a bacterial genome. We performed a whole genome alignment of KN400 and PCA using Mauve (Darling et al., 2004) with a seed-weight of 15, a minimum island size of 50, and the Muscle 3.6 algorithm in order to check for completeness of all the homologous gene sequence. We further confirmed the absence of any genes with respect to PCA by aligning the Illumina reads to the corresponding region.

Comparative Genomics Analysis

We obtained the PCA genome sequence, translated ORFs, and functional annotation from NCBI (RefSeq ID NC_002939) and the RefSeq database [73] and performed an alignment using BLAST [69]. We extracted the Identifiers, functional annotations, and organism names for the closest matches for all ORFs. We calculated bit score ratios as the bit score of the best match in a KN400-PCA or KN400-RefSeq BLASTp or tBLASTn comparison adjusted by the bit score of the KN400 protein aligned to itself or its own genome. Orthologs were defined as reciprocal best matches in whole-genome KN400-PCA BLASTp comparisons [79]. We also performed a bidirectional Smith-Waterman alignment of the ORFs of KN40 and PCA using the ssearch35 program of the FASTA v3.5 package [80].

2.6 Acknowledgments

We would like to thank Nicholas Webster for providing access to Newbler and the data analysis cluster and Mark Chaisson and Pavel Pevzner for discussions on genome assembly and EULER-SR. We would also like to acknowledge Nathan E. Lewis for his valuable inputs on representation of figures. This research was supported by the Office of Science (BER), U. S. Department of Energy, Cooperative Agreement No. DE-FC02-02ER63446.
Chapter 2 in full is a reprint of the material Nagarajan, H., Butler JE, Klimes A, Qiu Y, Zengler K, Ward J, Young ND, Methe BA, Palsson BO, Lovley DR, Barrett CL. De Novo assembly of the complete genome of an enhanced electricity-producing variant of Geobacter sulfurreducens using only short reads. PLoS One. (2010) Jun 8;5(6):e10922. The dissertation author was the primary author of this paper responsible for the research.
Chapter 3

Characterizing the interplay between multiple levels of organization of bacterial sigma factor regulatory networks

3.1 Abstract

Bacteria contain multiple sigma factors, each targeting diverse, but often overlapping sets of promoters, thereby forming a complex network. The layout and deployment of such a sigma factor network directly impacts global transcriptional regulation and ultimately dictates the phenotype. Here, we integrate multi-omic datasets to determine the topology, the operational, and functional states of the sigma factor network in Geobacter sulfurreducens. This revealed a robust network with a unique topology of interacting sigma factors. Analysis of the operational state of the sigma factor network showed a highly modular structure with $\sigma^N$ being the major regulator of energy metabolism. Surprisingly, the functional state of the network during the two most divergent growth conditions was nearly static, with sigma factor binding profiles almost invariant to environmental stimuli. This first comprehensive elucidation of the interplay between different levels of the sigma fac-
tor network organization is fundamental to characterize transcriptional regulatory mechanisms in bacteria.

3.2 Introduction

The core RNA polymerase in bacteria typically consists of five core subunits (α2ββω) and a dissociable sixth subunit, the sigma factor (σ), which is critical for transcriptional initiation. When a given sigma factor associates with the core RNA polymerase, the resulting holoenzyme acquires the ability to recognize promoter motifs and initiate transcription. All bacterial species contain a housekeeping sigma factor that is responsible for initiation of transcription from the majority of the promoters in the genome. Additionally, bacteria encode several to dozens of alternative sigma factors that control transcriptional initiation of a subset of the genes [81]. The number and the utilization of these alternative sigma factors varies widely among different bacterial species, and is related to the diversity of lifestyles [81, 82]. For example, while *Escherichia coli*, encodes seven sigma factors, the soil bacterium *Streptomyces coelicolor*, contains 60 alternative sigma factors [81]. The roles of these alternative sigma factors also differ substantially and can range from spore formation, to stress responses, such as those occurring while in stationary growth phase or during heat shock. Furthermore, these different sigma factors often interact with each other by targeting overlapping sets of promoters, thereby resulting in a complex regulatory network to modulate the variety of cellular process [83]. Several approaches, mostly computational, have been employed to determine binding sites for the elucidation of the regulatory network [84, 85, 86]. While computationally derived binding sites have been used to reconstruct sigma factor networks, these methods do not account for condition specific binding information, and thus cannot determine the functional state of the network. Therefore, experimental methods, such as chromatin immunoprecipitation coupled to microarray (ChIP-chip) or to sequencing (ChIP-seq), have recently been applied to determine condition specific binding of sigma factors to DNA directly [87, 88]. Recently, we demonstrated the utility of a systems approach that inte-
grates a diversity of multi-omics data to characterize the structural, operational, and functional organizations of the genomes of the generalist \textit{E. coli} and the specialist \textit{Geobacter sulfurreducens} [89, 90]. Here, we elucidated the comprehensive sigma factor network in \textit{G. sulfurreducens}, a bacterium that has been studied extensively for its impact on the natural environment and its capability of electricity production from organic waste [91]. The network of the four major sigma factors in \textit{G. sulfurreducens} ($\sigma^D$, $\sigma^H$, $\sigma^N$, and $\sigma^S$) was resolved by combining experimentally determined sigma factor binding with computational approaches in the context of the previously described transcriptional unit architecture [90]. In addition to determining the topology of the \textit{G. sulfurreducens} sigma factor network, we characterized its operational state and its effect on the physiological functional state of the cell. This comprehensive multi-level characterization of sigma factor regulons and their complex regulatory network serves as the scaffold necessary to build the entire transcriptional regulatory network.

3.3 Results

3.3.1 The topology of the sigma factor network in \textit{Geobacter sulfurreducens}

\textit{G. sulfurreducens} encodes six different sigma factors. In addition to the housekeeping $\sigma^D$, it also contains the extracytoplasmic stress $\sigma^E$, the flagellar $\sigma^F$, the heat stress $\sigma^H$, the nitrogen limitation related $\sigma^N$, and the stress and starvation induced $\sigma^S$. 
Figure 3.1: Sigma factor network topology: (a) The topology of the interaction network of the four major sigma factors in *G. sulfurreducens*. (b) The topology of the interaction network of the four major sigma factors in *E. coli*. The sigma factor network topology of *G. sulfurreducens* suggests a robust network with all of the alternative sigma factors (σ^H^, σ^N^, and σ^S^) regulating the housekeeping σ^D^). Auto-regulation of three sigma factors (σ^D^, σ^H^, and σ^N^) in *G. sulfurreducens* further highlights critical topological differences compared to the *E. coli* sigma factor network.
The network topology of the sigma factor network was determined using a ChIP-chip approach to obtain the genome-wide binding profiles for the four major sigma factors $\sigma^D$, $\sigma^H$, $\sigma^N$, and $\sigma^S$ under various growth conditions (Supplementary Table S1). Binding profiles of $\sigma^D$, $\sigma^N$ and $\sigma^S$ were obtained from cells grown with acetate as electron donor either planktonically with fumarate or as a biofilm on an electrode serving as the terminal electron acceptor. Furthermore, binding profiles of $\sigma^H$ and $\sigma^N$ were obtained from cells under heat shock stress and nitrogen limitation, respectively. $\sigma^E$ was excluded since the 32 flagella genes controlled by this sigma factor were found to be usually silent [54, 92]. Although several conditions were assayed to induce $\sigma^E$, no suitable condition for $\sigma^E$ could be obtained. Binding peaks for $\sigma^D$, $\sigma^H$, $\sigma^N$, and $\sigma^S$ under various conditions were identified by applying two peak calling algorithms (NimbleScan and MA2C [93]) to the ChIP-chip data sets. Peaks that were identified by both algorithms with at least a two-fold enrichment were considered binding regions (see Methods). These bona fide binding regions were subsequently mapped to the previously elucidated transcription unit architecture of *G. sulfurreducens* [90] to determine the regulon of each sigma factor. Overall, a total of 1,522 binding regions were identified in the *G. sulfurreducens* genome (Table 3.1 and Supplementary Table S1). These binding regions were mapped onto the promoter regions of 1,339 transcription units, covering greater than 80% (2,620 genes) of all genes. A total of 652 binding regions were identified for $\sigma^D$, which controlled transcription of about 60% of the genome (2,050 genes), confirming the role of $\sigma^D$ as housekeeping sigma factor.

**Table 3.1:** Properties of regulons of the four major sigma factors in *Geobacter sulfurreducens*

<table>
<thead>
<tr>
<th>Sigma Factor</th>
<th>Binding events</th>
<th>Regulated TUs</th>
<th>Regulated genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RpoD</td>
<td>652</td>
<td>875</td>
<td>2050</td>
</tr>
<tr>
<td>RpoH</td>
<td>275</td>
<td>337</td>
<td>791</td>
</tr>
<tr>
<td>RpoN</td>
<td>349</td>
<td>383</td>
<td>895</td>
</tr>
<tr>
<td>RpoS</td>
<td>246</td>
<td>357</td>
<td>838</td>
</tr>
</tbody>
</table>

Multiple genome-wide binding sites for $\sigma^H$ (275), $\sigma^N$ (349), and $\sigma^S$ (246) were identified. These binding sites controlled transcription of 802 ($\sigma^H$), 909 ($\sigma^N$),
and 863 ($\sigma^S$) genes (Table 3.1). We confirmed many genes previously described to be regulated by $\sigma^H$, such as heat-inducible chaperones (e.g., hspA) [94], or by $\sigma^N$, such as the nitrogenase molybdenum-iron cofactor biosynthesis protein (nifEN) [95]. In addition to nifEN being transcribed from a $\sigma^N$-specific promoter, $\sigma^N$ was the only sigma factor controlling this essential gene for nitrogen fixation. These results are conform to previously identified regulons of these sigma factors1. Each alternative sigma factor in *G. sulfurreducens* directly regulates 25% of the ORFs in the genome (Table 3.1), suggesting a broader role for these alternative sigma factors in *G. sulfurreducens*. The topology of the *G. sulfurreducens* sigma factor network indicated a robust network with all the alternative sigma factors ($\sigma^H$, $\sigma^N$, and $\sigma^S$) regulating the housekeeping $\sigma^D$ (Fig. 3.1a). Specifically, $\sigma^N$ is found to auto-regulate itself and regulate $\sigma^D$, thereby suggesting a pronounced role for $\sigma^N$ in addition to $\sigma^D$. In contrast, $\sigma^H$, $\sigma^N$, and $\sigma^S$ in *E. coli* exhibit a different topology (Fig. 3.1b). Neither $\sigma^H$ nor $\sigma^N$ auto-regulate themselves in *E. coli* and $\sigma^N$ does not regulate $\sigma^D$. These topological differences enable *G. sulfurreducens* to tightly regulate the expression of its sigma factors through feedback mechanisms and thus impacting the operation of the sigma factor network.

### 3.3.2 The operational state of the sigma factor network

To determine the operational state of the sigma factor network in *G. sulfurreducens*, we performed a functional enrichment analysis for the regulons of all four major sigma factors ($\sigma^D$, $\sigma^H$, $\sigma^N$, $\sigma^S$). Genes with known functions were assigned to 24 subsystems based on the SEED database [96], and a hypergeometric test was used to determine if a particular functional category is enriched in the regulon. The regulon of $\sigma^D$ was enriched in all major biosynthetic processes (nucleotide and amino acid metabolism), and in cell wall and capsule synthesis. The regulon of the heat stress $\sigma^H$ was also enriched for nucleotide and amino acid biosynthetic processes. The robust topology of the sigma factor network was evident from the enrichment for genes involved in membrane transport and protein metabolism in regulons of all four major sigma factors. Expectedly, $\sigma^N$ was the main sigma factor involved in regulating nitrogen metabolism. However, as hinted at by the topology
of the network, the enrichment analysis revealed an expanded role for $\sigma^N$ in the operational state of the network. In addition to nitrogen metabolism, the $\sigma^N$ regulon was enriched for genes involved in other cellular processes, such as cell wall and capsule synthesis, and membrane transport. Most importantly, $\sigma^N$ was found to be the primary sigma factor that regulates energy metabolism and respiration. Also, 18 genes that are transcribed from $\sigma^N$-dependent promoters encode two component system genes, indicating its significant role in cell signaling and transcription regulation. These results suggest a much more pronounced role for $\sigma^N$ in *G. sulfurreducens* than previously described. $\sigma^S$ was the only other sigma factor whose regulon was also enriched for genes involved in energy metabolism and respiration.

To assess the impact of the robust topology at the operational level, we investigated the effect of genetic perturbations on the sigma factor network operation. Since no validated essential gene list is available for *G. sulfurreducens*, we applied an *in silico* knock-out approach using a genome-scale metabolic model [48, 97]. Such approaches have been successfully applied to predict gene essentiality [98]. A total of 809 genes present in the *G. sulfurreducens* metabolic reconstruction were individually knocked-out in silico, and growth simulations representing four different conditions were performed. These conditions included planktonic growth with either Fe(III) or fumarate as electron acceptor, growth as an electrogenic biofilm on an electrode, and planktonic growth under nitrogen limitation. A gene was considered essential if the corresponding in silico knock-out resulted in growth-deficient simulations in all four growth conditions. Most of the essential metabolic genes were under the control of $\sigma^D$ (Fig. 3.2). Among the 229 essential genes identified, 75 of them were controlled solely by $\sigma^D$, whereas over half of them (123) were regulated by $\sigma^D$ and at least one other alternative sigma factor (Supplementary Table S2). Only a handful of essential genes were not transcribed from a $\sigma^D$ promoter, but from an alternative sigma factor promoter. For example, the gene GSU1745 that encodes the essential outer membrane porin OmpA [99] was only associated with a $\sigma^N$-dependent promoter. This observation is consistent with the previously failed attempts to generate a rpoN ($\sigma^N$) knock-out strain for *G. sulfurreducens* [95]. In addition to the genes essential across all four conditions, several genes
were determined to be essential only under certain specific conditions. Moreover, some of these genes were regulated solely by the housekeeping $\sigma^D$ and few of them were regulated by one of the alternative sigma factors. For example, 7 out of 17 genes that are involved in nitrogen metabolism had only $\sigma^N$-dependent promoters associated, and were found to be essential only under nitrogen limiting growth. These results clearly indicated that although most cellular processes essential under all conditions are regulated by $\sigma^D$, certain condition-specific essential processes are controlled by alternative sigma factors. Furthermore, 15 of the 229 essential genes had promoters associated with all four sigma factors. These included genes encoding for ATP synthase, as well as genes involved in amino acid and nucleoside biosynthesis. By utilizing multiple sigma factors for the most critical genes in the metabolic pathways, cells are able to fine tune expression levels under various conditions to easily adapt to environmental changes. A similar utilization of multiple sigma factor was also observed for genes related to other crucial cellular processes that are not part of the metabolic reconstruction, including genes involved in transcriptional machinery, such as infA, infB (transcription initiation factors), nusA (transcription elongation factor), rho (transcription termination factor) and rpoD ($\sigma^D$) itself. This broader role for alternative sigma factors in the operation of *G. sulfurreducens* sigma factor network is a likely reflection of its robust topology.
Figure 3.2: Operational state of the *G. sulfurreducens* sigma factor network: The interaction network of the four major sigma factors with the metabolic genes in *G. sulfurreducens* is shown. An in silico gene essentiality screen was carried out under four different conditions to assess the effect of genetic perturbations on the sigma factor network operation. The genes determined to be essential under all four conditions are shown by red nodes; other non-essential genes are shown in gold.
3.3.3 Functional effect of the sigma factor network operation

One of the important features of bacterial sigma factor usage is the selective preference for use of different sigma factors depending on the growth condition. For example, it is known in *E. coli* that the intracellular level of $\sigma^S$ is extremely low during exponential growth. Only when cells are in stationary phase, $\sigma^S$ accumulates in the cell, binds to promoters of its regulon and activates their expression [100]. To investigate the functional state of the sigma factor network in *G. sulfurreducens*, we compared binding profiles of $\sigma^D$, $\sigma^N$, and $\sigma^S$ under different growth conditions. The largest change in the transcriptome between all conditions examined was observed while cells were grown planktonically compared to an electrogenic biofilm on an electrode. The two divergent physiological states resulted in a difference of expression (at least two-fold) of >10% of the genome between these conditions (Fig. 3.3a). Comparison of sigma factor binding profiles revealed that all three sigma factors were active in both the planktonic cells as well as in the biofilm. $\sigma^S$, different to its counterpart in other bacteria, regulated over 800 genes in the exponential phase where cells had sufficient nutrients and carbon supply. Almost 30% of these genes were transcribed only from a $\sigma^S$-dependent promoter under this condition. The overall binding profiles of the three sigma factors during these two divergent physiological states were surprisingly similar (Fig. 3.3b), indicating that differential gene expression was not primarily regulated at the sigma factor level in *G. sulfurreducens*. The $\sigma^D$ regulon expanded from 1,974 genes in planktonic cells to 2,011 genes in electrogenic biofilm-forming cells with only 40 genes exhibiting differential $\sigma^D$ binding under these two conditions. The $\sigma^N$ regulon changed from 863 to 856 genes, with 50 genes differentially expressed. The functional regulon for $\sigma^S$ in exponentially grown planktonic cells (814 genes) was similar to the one for cells forming an electrogenic biofilm (807 genes), with 81 genes differentially expressed. To further elucidate the functional state of the sigma factor network, we assessed alternative usage of sigma factors under different growth conditions, and identified genes containing binding sites for multiple sigma factors in the promoter region. Although the transcriptome of the cell changed by
more than 10%, different sigma factor usage at the same promoter region was not observed.

Figure 3.3: Functional state of the *G. sulfurreducens* sigma factor network: (a) Comparative transcriptomic profile between *G. sulfurreducens* grown as planktonic cells reducing fumarate and as an electrogenic biofilm on an anode. About 10% of the transcriptome showed differential expression between the two conditions. (b) Venn diagram showing the comparative binding profiles of the three sigma factors $\sigma_D$, $\sigma_N$, and $\sigma_S$ for planktonic cells and electrogenic biofilm. The size of each sigma factor regulon from electrogenic biofilm cells is shown in parentheses. The invariant binding profiles between the two conditions indicate a near-static functional state for the sigma factor network in *G. sulfurreducens*. 
This near-static nature of the sigma factor binding profiles prompted us to analyze the network under additional growth conditions, early stationary phase (for $\sigma^S$) and nitrogen fixing condition (for $\sigma^N$). The binding profiles of $\sigma^D$, $\sigma^N$, and $\sigma^S$ were highly similar across these growth conditions as well. For example, the binding profiles of $\sigma^S$ in exponentially grown cells were similar to early stationary phase (both planktonic), confirming that $\sigma^S$ is present and active under both exponential and stationary phases of growth in *G. sulfurreducens* (Fig. 3.4). Only a few cases of conditional sigma factor usage were identified. Genes that are directly related to nitrogen fixation were not transcribed when ammonium was used as nitrogen source, but were specifically activated under nitrogen limitation with binding of $\sigma^N$ in their promoter regions, indicating conditional binding and activity of $\sigma^N$ at these promoters. This conditional usage of $\sigma^N$ is likely due to the difference in the modulation of $\sigma^N$ activity. Unlike the other sigma factors, the activity of $\sigma^N$ depends on the expression and binding of $\sigma^N$-dependent transcription factors to $\sigma^N$ targeted motifs.

![Figure 3.4: Analysis of RpoS protein concentrations during growth of *G. sulfurreducens*.](image)

(A) Samples were collected from batch grown, wild type *G. sulfurreducens* on fresh water acetate media with fumarate as an electron acceptor. (B) Proteins were separated using electrophoresis on a 10% polyacrylamide gel. RpoS protein concentrations during exponential phase and stationary phase were determined using Western blot analysis. RpoS (42 kDa in size) concentrations remained at fairly constant levels across all the five time points (P1-5).
While the comparative analysis of the sigma factor binding profiles suggests a near-static functional state of the sigma factor network, functional enrichment of the different regulons indicated modularity in its operation. To gain further insight into the effect of this modularity on the physiology of *G. sulfurreducens*, we analyzed the changes in expression between the two most divergent growth conditions (planktonic and electrogenic biofilm) in the context of the genome-scale metabolic network [48]. This analysis revealed that $\sigma^N$ primarily controls the energy metabolism during the shift from planktonic growth reducing fumarate to growth as an electrogenic biofilm, reiterating the trends observed from the functional enrichment analysis. Specifically, the gene encoding for DcuB, the terminal step in fumarate reduction, is regulated by $\sigma^N$ and its expression decreases by more than 4-fold during the shift to electrogenic growth. Moreover, OmcZ and other outer membrane cytochromes (OmcO, OmcN, and OmcP) that are involved in electron transfer to the anode [101] are also under $\sigma^N$ control and their expression levels are upregulated during this particular shift. Additionally, $\sigma^N$-controlled uptake hydrogenases that are not required for growth of *G. sulfurreducens* on the electrode were downregulated over 8-fold (Fig. 3.5). On the other hand, $\sigma^D$ regulated biosynthetic processes that are critical during this particular growth shift. This is manifested in the 2-fold increase in expression of a gene involved in iron-sulfur-cluster biosynthesis, recharge and transfer (GSU2570) during growth as an electrogenic biofilm. These results indicate that the near-static functional state of the sigma factor network mirrors the modularity of its operational state.
Figure 3.5: Coordinated regulation of energy metabolism by the alternative sigma factors: Analysis of the differences in expression and sigma factor binding profiles between the two divergent growth conditions (planktonic cells and electrogenic biofilm) in the context of the genome-scale metabolic network. A snapshot of the key steps involved in the energy metabolism during this shift is shown. $\sigma^N$ primarily regulated energy metabolism during this shift by controlling the terminal steps in the respective respiratory processes (DcuB, and the outer membrane cytochromes, OmcO/P and OmcZ). $\sigma^S$ regulated the NADH dehydrogenase (nuo operon). The genes upregulated and downregulated in this shift are shown in green and red, respectively. OmcZ and NuoIJKL were also found to be co-regulated by $\sigma^D$. The alternative sigma factors thus play a major role in the tight regulation of energy metabolism during this shift.
3.4 Discussion

In eubacteria, sigma factors play a critical role in transcriptional initiation by conferring promoter-binding specificity to the bacterial RNA polymerase. Understanding the complex network formed by the housekeeping and alternative sigma factors and their regulons is foundational to elucidate mechanisms of transcriptional regulation and to realize the genotype-phenotype relationship. A comprehensive study of bacterial transcriptional regulation relies on extensive genome-scale characterization of multiple cellular processes. Here, we used the multidimensional annotation of the *G. sulfurreducens* genome [90] as the pivotal scaffold to gain further insights into its transcriptional regulatory mechanisms. Specifically, we expanded our systems approach that integrates a wide range of omics data, to unravel the transcriptional regulation modulated by sigma factors in *G. sulfurreducens*. We determined the binding sites of sigma factors on a genome-wide basis and reconstructed the regulons for each of them under a variety of physiological states. In addition to determining the topology of the sigma factor network, we characterized the operational state of the network and analyzed its effect on the functional state (Fig. 3.6). This comprehensive sigma factor regulatory network covers more than 80% of the genome and provides a framework to reconstruct the complete transcriptional regulatory network in this organism.
Figure 3.6: Interplay between the multiple levels of sigma factor network organization. The robust topology of the sigma factor network in *G. sulfurreducens* suggests a broader role for alternative sigma factors. This resulted in a modular operational state with $\sigma^N$ primarily regulating genes involved in energy metabolism and maintaining redox homeostasis, while the housekeeping $\sigma^D$ regulated biosynthetic processes. The sigma factor network has a near-static functional state with invariant binding profiles between the two divergent growth conditions (planktonic cells and electrogenic biofilm). Analysis of the transcriptomic changes and sigma factor binding profiles during this shift indicated that the necessitated tight regulation of energy metabolism during this shift is mirrored in the modularity of the operational state. This coordinated regulation in the operational state can further be attributed to the robust topology of the sigma factor network where the major sigma factors auto-regulate each other providing feedback mechanisms to fine tune expression.
Using the transcription unit architecture, we determined that the house-keeping \( \sigma^D \) has over 2,000 genes in its regulon, and the other three alternative sigma factors each regulate the expression of 25% of the \( G. \ sulfurreducens \) genome. The experimentally determined transcription unit architecture and sigma factor regulons further enabled us to elucidate the binding motifs of these four major sigma factors. The binding motifs of the three \( \sigma^70 \) family sigma factors (\( \sigma^D, \sigma^H, \) and \( \sigma^S \)) closely resembled the classic sigma factor binding motifs in bacteria, but show slight differences in individual base pairs within the motifs (Fig. 3.7). The motif of \( \sigma^N \) is comparable to the known \( \sigma^N \) motif described in other bacteria (Fig. 3.7).

The reconstructed network further revealed that all of the alternative sigma factors in \( G. \ sulfurreducens \) regulate the housekeeping \( \sigma^D \), thereby forming a robust network. Another unexpected feature of the sigma factor network in \( G. \ sulfurreducens \) is the auto-regulation of its sigma factors. Both \( \sigma^N \) and \( \sigma^H \) auto-regulate their own expression, which is not the case in \( E. \ coli \), but known in other bacteria, such as in \( Rhizobium \ leguminosarum \) (\( \sigma^N \)) [102] and \( Caulobacter \ crescentus \) (\( \sigma^H \)) [103]. This feature could influence the operation of the sigma factor network in \( G. \ sulfurreducens \) by providing a direct feedback mechanism to fine-tune its sigma factor transcription levels. Additionally, the topology of the sigma factor network suggests a greater role for the alternative sigma factors than what has been reported for \( E. \ coli \).

The robust topology of the network resulted in a modular operational state with \( \sigma^N \) primarily regulating genes involved in energy metabolism and maintaining redox homeostasis, while the housekeeping \( \sigma^D \) regulated biosynthetic processes. Contrary to its typical role of controlling stress response [104], the regulon of \( \sigma^S \) in \( G. \ sulfurreducens \) is enriched for regulating energy metabolism. Consistent with this functional difference, the activity and protein levels of \( \sigma^S \) are almost unchanged between exponential and stationary growth phase (Fig. 3.4), a feature typically exhibited by housekeeping sigma factors. However, the size of RpoS (40 kDa) [105] and its binding motif are typical of bacterial \( \sigma^S \) (Fig. 3.4, Fig. 3.7). Taken together, the lack of an auto-regulatory feature and a capability to regulate other sigma factors further preclude the classification of this \( \sigma^S \) homolog in \( G. \).
sulfurreducens as an alternative housekeeping sigma factor. It could be speculated that another sigma factor, particularly $\sigma^E$, controls the main stress response in \textit{G. sulfurreducens}. However, induction of $\sigma^E$ was neither observed under stationary phase growth nor during heat or cold shock, suggesting a limited role of $\sigma^E$ under these conditions. This leads us to postulate that there is no designated sigma factor in \textit{G. sulfurreducens} that is predominantly active under stationary phase growth.

The effect of the robust topology on the operational state was further manifested in the shared regulation of 40% of the genes in the network by at least two sigma factors, providing a regulation redundancy that guarantees robust expression of these genes regardless of environmental stimuli. Indeed, most critical genes of cellular functions, such as ATP synthase, were under the regulation of $\sigma^D$ and at least one alternative sigma factors. Certain condition-specific essential genes, however, were regulated solely by alternate sigma factors. For instance, all genes encoding key enzymes for nitrogen assimilation (GDH, GS, and GOGAT) and the essential outer membrane porin (ompA) are transcribed by $\sigma^N$-dependent promoters. Therefore, while it has been possible to delete $\sigma^S$ in \textit{G. sulfurreducens} without affecting its viability [106], $\sigma^N$ deletion has not produced a viable phenotype [95].

Bacteria selectively regulate the relative abundance of different sigma factors in response to environmental stimuli. Typically, the usage of anti-sigma factors or sophisticated protease cascades [82, 107] can lead to a dynamic sigma factor regulatory network in which different portions of the network are switched on or off depending on the environment stimulus. However, our reconstructed sigma factor regulatory network in \textit{G. sulfurreducens} substantially differs from this model with a near-static network state between diverse physiological conditions. Systems-level analysis of transcriptomic changes and comparative binding profiles between the most divergent growth conditions attributed this difference in the functional state to topological and operational properties of the sigma factor network. In \textit{G. sulfurreducens}, $\sigma^D$ regulates biosynthetic processes and other housekeeping functions, while $\sigma^N$ is the primary sigma factor that regulates energy metabolism to maintain redox homeostasis. The two divergent growth conditions evaluated here represent very different modes of respiration, with electrogenic biofilm growth requiring ex-
tracellular electron transfer as opposed to intracellular fumarate reduction during planktonic growth. Both these growth modes necessitate a tight regulation of energy metabolism and maintenance of redox homeostasis, unlike the shift from respiration to fermentation in *E. coli*. Thus, the modular operational state is fundamental to the tight regulation of energy metabolism manifested in this near-static functional state of the sigma factor network. This is likely a consequence of the potential feedback mechanisms revealed by the robust topology of the network (Fig. 3.6).

Figure 3.7: Binding motif of the four major sigma factors in *Geobacter sulfurreducens*
The mechanisms controlling the levels of alternative sigma factors and how they compete with $\sigma^D$ for the core RNAP under different conditions needs further investigation. It is known for *E. coli* and some other bacteria that the $\sigma^D$-sequester Rsd plays an important role in modulating $\sigma^D$ availability and therefore facilitates the recruitment of alternative sigma factors. However, a homolog of this protein and other known anti-sigma factors could not be identified in *G. sulfurreducens*. Additionally, the *G. sulfurreducens* genome does not encode for the ClpXP protease cascade (including RssB and SprE) [108] to regulate the stability of $\sigma^S$, allowing the accumulation of stable $\sigma^S$ during exponential phase (Fig. 3.4). This further highlights the difference in the modulation of the sigma factor network of *G. sulfurreducens* and *E. coli*. It is possible that in a tightly controlled near-static sigma factor network, the dynamic response of the organism to environmental stimuli occurs predominantly at the level of the transcriptional regulatory network (Fig. 3.6). Also, it is known that ppGpp could target the RNA polymerase holoenzyme and modulate its activity on promoters [109]. Monitoring intracellular levels of ppGpp in *G. sulfurreducens* revealed that relGsu (RelA homolog) is involved in stress response such as stationary phase growth [110]. To evaluate the physiological role of sigma factors in response to such stress conditions, it would be informative to investigate the intracellular ppGpp levels in a double mutant of *G. sulfurreducens* (ΔrpoS::relGsu). Additionally, post-transcriptional regulation, or trans-acting transcription factors might be deployed instead to tune gene transcription in response to environmental changes in *G. sulfurreducens*. In summary, we have experimentally reconstructed a comprehensive sigma factor regulatory network in *G. sulfurreducens*. The robust network topology of this regulatory network is different from what has been described so far for other bacterial species. In addition to the topology, the detailed characterization of the operational state highlighted critical functional differences between the sigma factor network of a specialist like *G. sulfurreducens* and a generalist like *E. coli*. Analyzing the sigma factor regulatory network in the light of a comprehensive transcription unit architecture of a bacterium provides a framework to map binding activities of transcription factors as well as regulatory events of small signal molecules and will further our under-
standing of principles underlying transcription regulatory networks in bacteria.

3.5 Methods

3.5.1 Bacterial strains, medium, and growth conditions

*G. sulfurreducens* (ATCC 51573) was grown under strictly anoxic conditions at 30°C in mineral salt medium as previously described [111], with acetate as electron donor and fumarate or ferric citrate as electron acceptor. For growth in the absence of fixed inorganic nitrogen, ammonium chloride was omitted from the medium and N₂ served as the only nitrogen source. Cells in microbial fuel cells were grown as described previously [112]. For heat shock condition, cells were grown in mineral salt medium at 30°C until mid-exponential phase and then incubated in a 42°C water bath for 15 minutes.

3.5.2 ChIP-chip and ChIP-seq

ChIP-chip for RpoD was performed as described previously [90]. ChIP-chip for RpoH, RpoN and RpoS, was carried identical as described for RpoD but with polyclonal antibodies generated using rabbits as host animals. Genome wide sigma factor binding sites were determined for cells grown under various conditions. Prior to microarray hybridization, real-time quantitative PCR targeting previously known binding regions were carried out to verify enrichment of IP DNA fragments. qPCR and amplification of DNA was performed as previously described [87]. Microarray hybridization, wash and scan were performed in accordance with manufacturers instruction (Roche Nimblegen). ChIP-seq was performed with NEBNext ChIP-Seq Library Prep Master Mix Set for Illumina (NEB) with indexed adapters (Illumina). Sequence data were then aligned onto the *G. sulfurreducens* PCA genome (NC_002939) using Mosaik Aligner.
3.5.3 Identification of sigma factor binding regions

Binding regions of sigma factors were determined with both NimbleScan software as described before [90] and MA2C [93], and binding regions called in both software packages were used in downstream analysis.

3.5.4 Binding motifs of the sigma factors in *G. sulfurreducens*

Bacterial promoters can be grouped into two different families. The \( \sigma^D \) sigma factor family includes the housekeeping \( \sigma^D \) as well as most of the alternative sigma factors, such as \( \sigma^H \) and \( \sigma^S \). The binding motifs for this sigma factor family are generally composed of two consensus sequences centered at around -10 and -35 from the transcription start site [81]. The second family which contains \( \sigma^N \), contains -24/-12 type promoters. Promoters from both families are composed of multiple conserved elements, which are separated by variable length spacers, making them hard to predict computationally. Our experimentally identifying sigma factor binding regions in *G. sulfurreducens* consequently allowed elucidating the sigma factor binding motifs unambiguously. For each sigma factor, binding regions that were not shared with other sigma factors and additionally contained an experimentally determined transcription start site (TSS) were considered. Using BioProspector [113], we scanned 60 bp upstream of the TSS for potential two-part motifs. The results showed that all sigma factors belonging to the \( \sigma^D \) family (\( \sigma^D, \sigma^H, \sigma^S \)) contained a -35 element (TTGAC) that closely resemble the classic bacterial \( \sigma^D \) -35 element (TTGACA), with a less conserved -10 motif (TANNNT) (Fig. 3.7). The binding motif determined for \( \sigma^S \) had the least conserved -35 element within this family, but exhibited a relatively strong -10 and an extended -10 element (G at -14). On the other hand, \( \sigma^H \) binding sites had a strong -35 elements but the least conserved -10 element. It is known that many cis promoter elements can contribute to promoter selectivity of different sigma factors. For example, \( \sigma^S \) promoters in E. coli contain a hallmark C base at position -13 to make the promoter preferable to \( \sigma^S \) because of different charged amino acid located at
the beginning of the domain 3 alpha helix [114]. The subtle differences in the sigma factor binding motifs in *G. sulfurreducens* are very likely to contribute to the sigma factor selectivity at different $\sigma^D$ family promoters *in vivo*. The determined binding motif for $\sigma^N$ family promoters resembled the classic -24/-12 type motif found in other bacteria. It is known that the activities of $\sigma^N$ promoters are modulated by an enhancer dependent mechanism, and the selective expression of $\sigma^N$ promoters under different conditions in *G. sulfurreducens* could depend on the activity of the 30 predicted $\sigma^N$ dependent enhancers encoded in the *G. sulfurreducens* genome [115].

### 3.5.5 Western blot

Cells were harvested, pelleted via centrifugation and if not processed immediately, frozen at -80°C until use. Samples were resuspended in LDS sample buffer (Invitrogen) containing 10x reducing agent (Invitrogen) based on optical density in order to normalize protein content. Each sample was lysed at 95°C for five minutes and subjected to electrophoresis in a 10% polyacrylamide gel (Invitrogen). Resolved proteins were electrotransferred to a Hybond-ECL membrane (Amersham Biosciences). Nonspecific binding was prevented by incubating the ECL membrane in nonfat dried milk overnight. The ECL Western detection kit, 100 $\mu$g of *G. sulfurreducens* RpoS polyclonal antibody and horseradish-peroxidase-conjugated donkey anti-rabbit immunoglobulin G (Amersham Biosciences) were used to detect the presence of RpoS.

### 3.5.6 Functional enrichment analysis of the regulons

The functional categories for the *G. sulfurreducens* genes were obtained from the subsystems assigned in the SEED database [96]. To determine if the regulons of each sigma factor were significantly enriched for any particular functional category, a hypergeometric test was performed using the hygecdf function in MATLAB. A p-value cutoff of 0.05 was used to determine significance.
3.5.7 *In silico* gene essentiality simulations

Genes in the genome-scale metabolic model of *G. sulfurreducens* were individually knocked out and *in silico* growth simulations were performed. This was achieved using the singleGeneDeletion function of the COBRA toolbox [45].

3.6 Acknowledgements

We would like to thank Larisa Kagan for assisting with antibody purification and preparation. This research was supported by the Office of Science, Biological and Environmental Research (BER), United States Department of Energy, Cooperative Agreements No. DE-FC02-02ER63446 and DE-SC0004485.

Chapter 3 in full is a reprint of the manuscript under review Qiu Y, Nagarajan H, Embree M, Shieu W, Abate E, Juarez K, Elkins J, Cho B-K, Nevin KP, Barrett CL, Lovley DR, Palsson BO, Zengler K. Characterizing the interplay between multiple levels of organization of bacterial sigma factor regulatory networks. The dissertation author was the second author of this paper, responsible for all the computational analysis and systems-level interpretation of the results.
Chapter 4

Genome-scale metabolic characterization of electrosynthetic bacteria

A novel mechanism, known as microbial electrosynthesis, in which microorganisms directly use electric current to reduce carbon dioxide to multi-carbon organic compounds that are excreted from the cells into extracellular medium, has recently been discovered. Microbial electrosynthesis differs significantly from photosynthesis in that carbon and electron flow is primarily directed to the formation of extracellular products, rather than biomass. However, extensive knowledge about the metabolism of the organism as well as its extracellular electron transfer pathways is critical to realize the potential of this technology for the production of the desired fuel compound. So far, only a few microorganisms have been shown to be capable of accepting electrons from the cathode to reduce carbon dioxide to limited organic compounds such as acetate and 2-oxobutyrate. Constraint-based metabolic modeling and analysis has been useful for discovering and understanding new capabilities and content in bacteria, as well as in guiding metabolic engineering efforts for targeted production. Here, we present a detailed genome-scale characterization of metabolic capabilities of two different electrosynthetic bacteria *Geobacter metallireducens* and *Clostridium ljungdahlii*. 
4.1 Introduction

Electrical energy is increasingly becoming critical for the sustainable production of transportation fuels and biochemicals [116]. The recent discovery of microbial electrosynthesis provides an attractive and novel approach for the generation of valuable chemicals and fuels from electricity [13]. Microbial electrosynthesis represents the electricity-driven reduction of carbon dioxide into multi-carbon organic compounds, with the whole microorganism serving as the biocatalyst. When the electricity driving the process is generated from solar power or photovoltaic cells, microbial electrosynthesis can be thought of as an artificial form of photosynthesis. However, one of the critical factors distinguishing electrosynthesis from photosynthesis is the routing of carbon and electron flow towards extracellular products instead of biomass. Since the microorganism acts as the catalyst for this process, the choice of the appropriate microbe is critical for the successful application of microbial electrosynthesis.

There are two major metabolic constraints that govern the capabilities of a bacterium to perform electrosynthesis. First, it must possess the requisite metabolic machinery to fix CO$_2$. Second, the microbe should contain the appropriate electron transport system (ETS) that allows it to derive electrons directly from the electrode and pass it on to the appropriate reducing equivalents (ferredoxin, NADH, NADPH) for biosynthesis. Microbial electrosynthesis was first demonstrated by the production of acetate and 2-oxobutyrate by the gram-negative acetogen *Sporomusa ovata* at -600mV [117]. Later a few additional acetogens such as *Moorella thermoacetica* and *Clostridium ljungdahlii* have also been shown to be capable of performing microbial electrosynthesis [13].

Acetogenic microorganisms have unique metabolic pathways and energy conservation mechanisms that, if understood, could greatly increase options for the design of microorganisms for production of biofuels and enhance the range of biomass materials that could become economically viable feedstock. They were discovered for their capability to autotrophically reduce CO$_2$ to acetate and conserve
energy simultaneously using the Wood-Ljungdahl pathway [118]. This mode of carbon fixation is thought to be the most ancient aspect of metabolism [119]. Despite this, several physiological and biochemical aspects governing this metabolic capability have been poorly characterized with several new discoveries being made every year [120, 121]. One of the fundamental aspects of acetogenic metabolism discovered recently is the concept of flavin based electron bifurcation. In this mechanism, there is a concomitant coupling of an endergonic redox reaction with the oxidation of the same electron donor with higher potential electron acceptors. This aspect is believed to play a key role in the energy conservation mechanisms of acetogens [122]. A comprehensive characterization of these aspects of metabolism and energy conservation in acetogens is therefore fundamental to enhance the understanding of microbial electrosynthesis. To address this, we developed the genome-scale metabolic model for the acetogenic bacteria \( C.ljungdahlii \).

Apart from acetogens, a member of the class \( Geobacteraceae \) (\( Geobacter metallireducens \)) is also capable of performing electrosynthesis on the cathode [123]. The members of \( Geobacteraceae \) have unique extracellular electron transfer capabilities and have been one of the model organisms for studying microbe-electrode interactions [8]. Moreover, \( Geobacter \) species are capable of reduction of insoluble metal oxides coupled with the oxidation of a variety of organic compounds. To characterize microbial electrosynthesis in \( Geobacter \), it is therefore critical to understand the energy conservation mechanisms governing aspects of extracellular electron transfer and their interplay with metabolism. Constraints-based metabolic modeling was applied first on \( Geobacteraceae \) to build the genome-scale metabolic model of \( Geobacter sulfurreducens \) [48] in an initial attempt to determine if genome-scale models could be used to predict bioremediation. This model has been used to describe growth under a diversity of conditions ([49]); predict the phenotype of gene deletions with 90% success ([50]); and was a valuable tool for metabolic engineering [51]. More recently, the metabolic model for \( Geobacter metallireducens \) was also developed [124]. Additional modeling has provided important insights into the factors leading to effective uranium bioremediation by Geobacter species.
The application of genome-scale models has greatly improved understanding of bioremediation at the Rifle Site CO [27]. However, these models lack a detailed representation of the various extracellular electron transfer pathways and the energetics associated with them. To account for these explicitly and incorporate the knowledge of other metabolic capabilities such as CO$_2$ fixation, we have significantly expanded and updated the scope of the genome-scale metabolic reconstruction of *G. metallireducens*.

This chapter presents the genome-scale characterization of the metabolic and energetic aspects governing the two electrosynthetic bacteria *G. metallireducens* and *C. ljungdahlii*.

### 4.2 Metabolic Reconstruction

#### 4.2.1 Characterization of the *Geobacter metallireducens* GS-15 genome-scale reconstruction

The updated reconstruction of *G. metallireducens* GS-15 was generated by reconciling an existing genome-scale reconstruction [124] and an updated genome annotation, performing bottom-up reconstruction of additional metabolic pathways [125], and functionally testing the reconstruction for performance (Fig. 4.1A). Reactions that did not exactly match existing content were manually evaluated using standard manual curation procedures [125]. The final reconstruction contained 987 genes, 1284 reactions, and 1109 metabolites. The final reconstruction was compared to the previous version [124] and an automatically generated reconstruction from the Model SEED framework [6] (Fig. 4.1B) to identify newly reconstructed and unique content. The gene content in the *G. metallireducens* reconstruction was compared reconciled with the automated reconstruction obtained from ModelSEED [6]. The ModelSEED reconstruction was found to have 114 unique genes that were not present in the *G. metallireducens* reconstruction. Upon further analysis, it was determined that 86 of these genes were involved in
macromolecular synthesis, DNA replication and protein modifications which are beyond the scope of a metabolic network. Out of the remaining 28 genes, 8 of them did not have a specific reaction association in the ModelSEED (i.e. generic terms such as aminopeptidase, amidohydrolase). The ModelSEED annotations of the 20 metabolic genes were compared with the updated genome annotation presented in this work. It was found that these two sets of annotation were consistent for only two genes (Gmet_0988 and Gmet_2683). These two genes were added to the reconstruction by associating them to the appropriate reaction. In the case of the 18 genes, where a discrepancy existed between the updated genome annotation and ModelSEED annotation, the updated annotation was taken as gold standard. No reactions were added or removed from the reconstruction as a result of this analysis. The *G. metallireducens* reconstruction contains 227 genes not in either reconstruction, thus representing a significant advancement of coverage.
Figure 4.1: Reconstruction Workflow: (a) The workflow describing the reconstruction of the G. metallireducens reconstruction. The reconstruction process was initiated by comparing the updated genome annotation for G. metallireducens to the existing reconstruction to create a list of discrepancies that was manually reviewed and curated. Content that was in agreement between the updated annotation and reconstruction was used to generate a draft set of intracellular reactions. Lipid, membrane, murein, and LPS content was removed from this list as a periplasm compartment was added to the reconstruction. Manual curation (12) was aided by the KEGG [5], MODEL SEED [6], and MetaCyc [7] databases. Further, numerous publications and literature sources (i.e., the bibliome) were used to refine the network content. The manual review process resulted in a draft reconstruction that was used in conjunction with a formulated biomass objective function in simulations to validate the content of the reconstruction and generate a final version. B Venn Diagram showing the comparative analysis of network content. C: Validation of network content with omics data.
An analysis of the reaction content in the *G. metallireducens* reconstruction was performed to identify content unique to *G. metallireducens* included in the reconstruction that is not contained in the reconstructed networks for *E. coli* [126, 127], *S. cerevisiae* [128, 129], *M. barkeri* [130], *H. pylori* [131], *S. aureus* [132], *Y. pestis* [133], *K. pneumoniae* [134], *M. tuberculosis* [135], Human Recon 1 [136], *P. putida* [137] and *T. maritima* [138]. This *G. metallireducens* reconstruction contains 339 unique reactions compared to these networks (Table 4.1).

**Table 4.1:** Subsystem distribution of reactions unique to *G. metallireducens* reconstruction

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>Number of Reactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transport</td>
<td>80</td>
</tr>
<tr>
<td>Lipids and Glycan metabolism</td>
<td>64</td>
</tr>
<tr>
<td>Vitamins &amp; Cofactor Biosynthesis</td>
<td>50</td>
</tr>
<tr>
<td>Energy Metabolism</td>
<td>35</td>
</tr>
<tr>
<td>Aromatic Compound degradation</td>
<td>27</td>
</tr>
<tr>
<td>Alternate Carbon Metabolism</td>
<td>24</td>
</tr>
<tr>
<td>Amino Acid Metabolism</td>
<td>18</td>
</tr>
<tr>
<td>Central Metabolism</td>
<td>16</td>
</tr>
<tr>
<td>Nucleotide Metabolism</td>
<td>11</td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
</tr>
<tr>
<td>Metal Respiration</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>339</td>
</tr>
</tbody>
</table>

Furthermore, in comparison to the earlier reconstruction, it was revealed that there were a number of errors in the previous content which were fixed in the iAF987 reconstruction. Several unique pathways not previously appearing in any of the previously reconstructed networks used for comparison (Table 4.1) were included in the iAF987 reconstruction. Structurally, the periplasm was included as a distinct compartment in the reconstruction. The periplasm was determined to be important as *G. metallireducens* has the unique ability to transfer electrons extracellularly [8]. Thus characterizing the direct electron transfer pathways completely from the cytosol through the periplasm to the extracellular space will be key to understanding this unique capability. Additionally, the periplasm compartment addition allows for more accurate representation of metabolism, such as p-cresol and 4-hydroxy-benzyl alcohol degradation, which partially occurs in the
periplasm [139]. The iAF987 reconstruction has several unique metabolic features. Menaquinone biosynthesis via the futalosine pathway has been identified and reconstructed in iAF987 [140, 141]. Iron-sulfur clusters are essential cofactors for the function of many metalloproteins and several iron-sulfur cluster-binding proteins (e.g., nitrite reductase) are encoded in the *G. metallireducens* GS-15 genome. Similarly, molybdenum is an important cofactor for the function of proteins [142]. As such, these pathways were included in the reconstruction of *G. metallireducens* GS-15 iAF987 and also the recently reconstructed *E. coli* K-12 MG1655 reconstruction concurrently [143].

**Omics data validates network content**

To validate the content of the iAF987 network, omic data profiling a growth shift on aromatic electron donor from acetate was integrated with the metabolic model. Specifically, the metabolic adjustment predicted by the iAF987 model for the shift from growth on acetate to benzoate was compared with differentially expressed genes observed in the transcriptomic data. This analysis, performed using MADE [144] revealed a 97% consistency between model predictions for differentially regulated genes and omic data (Fig. 4.1C). Specifically, the genes encoding for benzoyl-CoA reductase and other reactions involved in benzoate utilization were upregulated.

It was determined that this key enzyme that links the degradation of aromatic substrates (electron donors) to central metabolism, is not ATP driven as previously thought [124], but rather is likely membrane bound and proton translocating [145]. Thus, a proton translocating reaction was added to the reconstruction for this step in metabolism. A translocation stoichiometry of 3 protons per electron was determined to be the likely extent of coupling through a thermodynamic analysis.

**Functional Testing**

The functional capabilities of the iAF987 reconstruction was evaluated by simulating growth using the biomass objective functions (Methods). The capabili-
ties of the model to predict growth on experimentally validated electron donors and acceptors was tested. The iAF987 reconstruction was able to grow on 19 different electron donors (3-Methylbutanoic acid, p-Cresol, 4-Hydroxy-benzyl alcohol, 4-Hydroxybenzaldehyde, 4-Hydroxybenzoate, Acetate, Butanol, Butyrate (n-C4:0), Benzoate, Benzaldehyde, Benzyl alcohol, Ethanol, Isobutyrate, Phenol, Propionate (n-C3:0), Propanol, Pentanoate, Pyruvate, Toluene) and 8 different electron acceptors (Fe(III), Manganese Mn(IV), Uranium U(VI), Technetium Tc(VII), Vanadium V(V), Nitrate, Nitrite, Anode Electrode).

4.2.2 Genome-scale reconstruction of the homoacetogen, 

*Clostridium ljungdahlii*

The genome-scale metabolic network for C. ljungdahlii was reconstructed using a four-step integrative reconciliatory workflow involving four published models of related *Clostridia* species and two draft models (Fig. 4.2). The first draft metabolic model was generated based on the *C. ljungdahlii* genome annotation [146] using the AutoModel functionality of SimPheny (Genomatica, San Diego), while the second draft model was generated using the ModelSEED database [6]. In addition to these two draft models, homologs to *C. ljungdahlii* genes were identified in published genome-scale reconstructions of related *Clostridia* species (*C. acetobutylicum*, *C. thermocellum*, and *C. beijerencii*) [147, 148, 149, 150] using Smith-Waterman alignment. A 60% amino acid sequence identity cutoff was used to identify *C. ljungdahlii* homologs in the other *Clostridia* genomes. The reactions corresponding to these genes in the respective *Clostridia* models were compiled and reconciled with the two draft models. The list of discrepancies was manually curated with the aid of biochemical literature and databases such as KEGG [5] and SEED [6]. Manual evaluation of new content from the annotation and existing genome-scale reconstruction consisted of gathering genetic, biochemical, sequence, and physiological data and reconciling this information to determine the likelihood of each reaction being present in the organism. The curated reconstruction was evaluated for functional performance with the aid of a biomass objective function that was formulated using an existing template (Methods). Using infer-
ence based on pathway function, as well as the SMILEY computational algorithm [151, 152] which predicts reactions which fills gaps in a metabolic network, reaction content was added to the network so that it could produce the necessary biomass components. This resulted in a final network consisting 633 genes, 767 reactions, and 693 metabolites (iHN633). This iHN633 reconstruction represents the first genome-scale metabolic model of an acetogen.
Figure 4.2: Reconstruction Workflow: Iterative reconciliatory reconstruction workflow adopted for generating the metabolic model for *C. ljungdahlii*.
**Functional Testing**

The functional capabilities of the iHN633 reconstruction was evaluated by simulating growth under different conditions using the developed biomass objective function (Methods). The iHN633 was successfully able to simulate heterotrophic growth on several substrates (fructose, gluconate, arabinose, ribose, xylose, pyruvate, and several other amino acids). Moreover, autotrophic growth with CO$_2$ and H$_2$ as well as just CO alone was also feasible with the iHN633 reconstruction. Furthermore, the iHN633 model was able to successfully predict growth rate consistent with experimental observations for heterotrophic growth on fructose. When the iHN633 model was constrained with experimentally measured fructose uptake rates (-1.8 mmol/gDWh), the growth rate predicted by the model (0.077 hr$^{-1}$) was in good agreement with experimentally measured growth rates (0.072 hr$^{-1}$). The acetate and ethanol production rates were also consistent with observed values.

**Identification of a new nitrate reduction pathway in *C. ljungdahlii***

A novel nitrate reduction pathway for *Clostridia* was reconstructed in the iHN633 model based on the *C. ljungdahlii* genome. This pathway could serve as the third possible nitrogen assimilation route in addition to the glutamine synthetase and glutamine:2-oxoglutarate aminotransferase or by fixing molecular nitrogen using a molybdenum-dependent nitrogenase. This pathway similar to the one identified in *Nautilia profundicola* [153]. The nitrate reduction pathway consists of a soluble nitrate reductase (NarABC), a nitrite reductase (ArsA1B1C), and an additional hydroxyl amine reductase (Hcp1 or Hcp2) (Fig. 4.3). Several *Clostridia* can use nitrate as final electron acceptor performing a primitive nitrate respiration. The identification of this pathway represents an alternative to investigate growth capabilities in ammonium free conditions.
4.3 Key carbon fixation pathways

Autotrophic carbon fixation is one of the key metabolic characteristics important for microbial electrosynthesis. The assimilation of inorganic carbon (CO$_2$) to cellular carbon requires an energetic input as well as reducing equivalents. These reducing equivalents have to be regenerated in a metabolic cycle that reduces inorganic carbon to cellular carbon. Apart from the Calvin cycle used by plants and photosynthetic organisms, there are five other carbon fixation pathways that have been characterized till date [154]. Key enzymes which enable the function of these pathways and the ATP requirement for each of these pathways had been identified (Table 4.2) [154].

4.3.1 Carbon fixation pathways in *G. metallireducens*

The genome of G. metallireducens GS-15 encodes for two out of the six known carbon fixation pathways. The pathways which are encoded in the *G. metallireducens* genome and reconstructed in iAF987 are the reductive citric acid (TCA) cycle [155] and the dicarboxylatehydroxybutyrate cycle [156] (Fig. 4.4A). Key enzymes for the reductive TCA cycle include the 2-oxoglutarate synthase (abbreviated OOR2r in the reconstruction) and ATP-citrate lyase (ACITL) which
Table 4.2: Known pathways for autotrophic carbon fixation and their energy requirements

<table>
<thead>
<tr>
<th>Pathway</th>
<th>ATP equivalent for 1 pyruvate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calvin Cycle (Reductive PPP)</td>
<td>7</td>
</tr>
<tr>
<td>Reductive TCA Cycle</td>
<td>2</td>
</tr>
<tr>
<td>Wood-Ljungdahl Pathway</td>
<td>1</td>
</tr>
<tr>
<td>3-hydroxypropionate bicycle</td>
<td>7</td>
</tr>
<tr>
<td>3-hydroxypropionate-4-hydroxybutrate cycle</td>
<td>9</td>
</tr>
<tr>
<td>Dicarboxylate-4-hydroxybutyrate cycle</td>
<td>5</td>
</tr>
</tbody>
</table>

enable the citric acid cycle to run in reverse. For the dicarboxylatehydroxybutyrate cycle, the key enzyme is 4-hydroxybutyryl-CoA dehydratase (4HBCOAH). The genes encoding for these key reactions in the reconstruction are shown in (Fig. 4.4B). The two pathways, the reductive citric acid (TCA) and the dicarboxylatehydroxybutyrate cycles, share four reactions in the citric acid cycle. The reconstruction of these carbon fixation pathways led to the novel prediction of a new growth condition for *G. metallireducens*. Previously, it was thought that the only condition for evaluating autotrophic growth was during electrosynthesis. However, with the expanded content including the carbon fixation pathways, it was computationally predicted that *G. metallireducens* can grow with formate as the electron donor and Fe(III) as the electron acceptor. Investigating flux distributions revealed that the CO\(_2\) derived from formate oxidation is reduced via the reductive TCA cycle to form acetyl-CoA which is subsequently assimilated into biomass. The electrons derived from formate oxidation are split between this process as well as for Fe(III) reduction. The energy gained by proton gradient formed during Fe(III) reduction was instrumental for providing the required ATP for carbon fixation. A detailed analysis of the energy conservation mechanism during Fe(III) reduction.
is provided in Section 4.4.1. This prediction of growth on formate and Fe(III) has been experimentally validated. The importance of CO$_2$ fixation for growth on formate is further highlighted by a study done on *G. sulfurreducens* performing Fe(III) reduction with formate as the electron donor. While *G. sulfurreducens* was able to reduce Fe(III) with the electrons from formate, it required the addition of 0.1 mM acetate to assimilate cell carbon [157]. This was attributed to the lack of a reductive TCA cycle in *G. sulfurreducens* due to the absence of the ATP-dependent citrate lyase.
Figure 4.4: Carbon fixation pathways in *G. metallireducens*. A. A map of the two carbon fixation pathways included in the iAF987 reconstruction. The figure shows the two carbon fixation pathways encoded by genes annotated in the updated genome annotation.
4.3.2 Carbon fixation pathways in *C. ljungdahlii*

The *C. ljungdahlii* genome encodes for all the genes involved in the Wood-Ljungdahl pathway of carbon fixation. Accordingly, this pathway was incorporated into the iHN633 reconstruction. The pathway consists of two different branches, one $CO_2$ contributing to the methyl methyl group of acetyl CoA via the eastern branch and the other $CO_2$ contributing to the carbonyl group of the acetyl CoA via the western branch (Fig. 4.5). The key enzyme in this pathway is the CODH/ACS complex which performs the dual activity of reducing $CO_2$ to $CO$ via the carbon monoxide dehydrogenase activity and the subsequent formation of acetyl CoA through the acetyl-CoA synthase. It has been suggested that the CODH/ACS complex functions in a manner that the $CO$ resulting from CODH activity is kept as a bound metabolite in the complex to diminish thermodynamic barriers involved in this energy intensive process [158]. Furthermore, the genome of *C. ljungdahlii* suggested that $CO$ oxidation to $CO_2$ is likely to be catalyzed by a different carbon monoxide dehydrogenase gene (CLJU_c09090-09110) other than the ones encoded by the CODH/ACS complex. Taking these into consideration, the iHN633 reconstruction incorporates the CODH/ACS reaction as the net reaction of carbon monoxide dehydrogenase and acetyl CoA synthase activity, and associates the gene cluster CLJU_c09090-09110 to the reaction representing $CO$ oxidation to $CO_2$. The CODH/ACS complex uses corrinoid iron-sulfur protein as a cofactor. The Wood-Ljungdahl pathway consumes one ATP equivalent for the production of acetyl CoA in the methyl branch for the activation of formate to formyltetrahydrofolate. The methylenetetrahydrofolate dehydrogenase reaction catalyzed by the bifunctional FolCD (CLJU_c37630) was assumed to be NADPH dependent based on sequence similarity with the corresponding enzyme in the acetogen *Moorella thermoacetica* [159, 160]. The energy conservation mechanisms associated with autotrophic growth on $CO_2$ and $H_2$ using the Wood-Ljungdahl pathway is discussed in detail in Section 4.4.2.
Figure 4.5: Carbon fixation pathways: A map of the Wood-Ljungdahl pathway of carbon fixation and the genes encoding the CODH ACS complex.
4.4 Energy conservation mechanisms and electron transport system (ETS) of electrosynthetic bacteria

Reconstructing Extracellular Electron Transfer Pathways

Identify all possible electron carriers (e.g. cytochromes, ferredoxin, quinone, NAD, FAD)

Enumerate all possible electron transport routes through these carriers.

Assign appropriate stoichiometry and gene associations to the reaction.

Figure 4.6: Protocol adopted for reconstructing electron transfer pathways

In order to accurately account for the various energy conservation mechanisms and ETS in the electrosynthetic bacteria *G. metallireducens* and *C. ljungdahlii*, all the possible electron carriers (cytochromes, quinones, ferredoxin, NADH, NADPH, FAD) were identified. A list of all possible redox reactions involving these carriers was enumerated and an extensive thermodynamic and energetic analysis
was performed on these key reactions to appropriately account for the proton translocation stoichiometries. (Fig. 4.6).

### 4.4.1 Analysis of energy conservation and ETS of *G. metallireducens*

The most common terminal electron acceptor for *Geobacter metallireducens* is Fe(III), which is reduced extracellularly. While wild-type *G. metallireducens* cannot reduce fumarate, it has been shown that engineering the dicarboxylate exchanger (dcuB) confers the capability to grow on fumarate as the terminal electron acceptor [161]. The analysis described in Fig. 4.6 enabled the identification of key energy conserving steps during respiration on both extracellular and intracellular electron acceptors (Fig. 4.7). An example of this included findings related to the fumarate reductase enzyme.

**Fumarate reductase in *Geobacter* is electrogenic**

*Geobacter* species derive 8 electrons from the oxidation of acetate through the TCA cycle during Fe(III) reduction. The electrons are transferred to extracellular Fe(III) via a chain of extra-cytoplasmic cytochromes and the membrane bound menaquinone pool. Oxidation of succinate to fumarate in the TCA cycle plays a critical role in transferring electrons to the menaquinone pool. In *Geobacter* sp., this metabolic step is catalyzed by a bifunctional fumarate reductase/succinate dehydrogenase (FrdCAB) enzyme. While the redox coupling for fumarate reduction using menaquinol is energetically favourable, the oxidation of succinate using menaquinone is endergonic. This is primarily due to the fact that succinate/fumarate redox couple is more electropositive than the electron acceptor menaquinone. It has also been proposed that this apparent energetic cost for succinate oxidation could provide an explanation to the decreased growth rates observed for *G. sulfurreducens* while respiring on Fe(III) compared to fumarate reduction [161]. Moreover, initial studies in *B. subtilis* indicated the possibility of
a transmembrane proton potential acting as the driving force for such a reaction [162]. The previous Geobacter models [48, 124] do not account for this based on biochemical evidence in organisms such as W. succinogenes. However, more recent evidence indicates that succinate dehydrogenase catalyzed by the bifunctional Frd-CAB does operate in a reverse redox loop in D. vulgaris [163]. This study shows that succinate oxidation by menaquinol requires the dissipation of membrane potential as a driving force. Furthermore, they attribute this coupling nature to the absence of the uncoupling residue in FrdCAB of W. succinogenes. The presence of this residue is shown to dissipate the proton potential generated by fumarate reductase. Geobacter species also lack this residue. Hence, we propose that the fumarate reductase of Geobacter species is also electrogenic. Thus, an electrogenic fumarate reductase was incorporated in the iAF987 reconstruction for G. metallireducens.

Figure 4.7: Modified ETS of G. metallireducens A: Generation of a proton gradient due to an electrogenic fumarate reductase while respiring on fumarate. B: Dissipation of membrane potential due to succinate oxidation while respiring on Fe(III). Electron transfer to Fe(III) via a chain of cytochromes.
To understand the energetics of intracellular electron transfer, physiological data for *G. metallireducens* dcuB was collected during fumarate reduction with different electron donors. Using the previously established maintenance energy values, it was found that setting the observed uptake rate for fumarate of the dcuB strain grown on fumarate and acetate (i.e., the input electron donor) and optimizing for optimal growth (as simulated through maximal flux through the biomass objective function) resulted in prediction of both the growth rate and acetate uptake rate consistent with what was measured in the growth screen experiment. Importantly, no significant byproducts besides CO2 were predicted by the model to be exported under this condition, which was consistent with the growth screen experiment. This result demonstrated that the previously determined values for the growth and non-growth associated maintenance are consistent for simulating optimal growth of the *G. metallireducens* GS-15 dcuB strain with acetate as an electron donor. Analysis of the *G. metallireducens* GS-15 dcuB strain grown with ethanol and butanol as electron donors revealed that growth on these substrates is suboptimal. Fig. 4.8 shows a Phenotypic Phase Plane (PhPP) analysis [164] of optimal growth for the dcuB strain with the three different electron donors, acetate, ethanol and butanol. Briefly, the experimental ranges of the electron donors (acetate, butanol, ethanol) and the electron acceptor (fumarate) were analyzed through modeling to understand optimal network behavior. A phase plane for each condition is given where the diagonal represents the optimal ratio of donor to acceptor when simulating optimal growth of the dcuB strain. Further, the experimentally observed values from the growth screen experiments are plotted on the phase planes as averaged triplicate value with error bars for comparison. From this analysis, growth with acetate as the electron donor revealed optimal behavior and the experimental data is very near to the line of optimality. Growth with ethanol and butanol as the electron donor revealed that these conditions are in the suboptimal regions of growth when analyzed with the model and when considering the experimentally-measured fumarate uptake rate for each condition, the amount of electron donor was far lower than an optimal condition at that acceptor rate. From this observation, one can hypothesize that growth with ethanol and butanol
is inefficient due to hindrance of either key enzymes involved in electron transfer, or through a modeling analysis of essential reactions that are different between the growth simulations on the optimally-utilized substrate acetate, a poor alcohol dehydrogenase in the network (the same alcohol dehydrogenases metabolize both butanol and ethanol - Gmet_1046, Gmet_1053).

**Phenotypic analysis of Intracellular Electron Transfer (Fumarate reduction)**

![Diagram showing phenotypic analysis](image)

**Figure 4.8: Phenotypic analysis of respiration on intracellular electron acceptors** Phenotypic Phase Plane (PhPP) analysis of the *G. metallireducens* GS-15 dcuB strain on three different carbon sources. The reconstructed model of *G. metallireducens* GS-15 was utilized to analyze the growth of the dcuB strain on acetate, ethanol, and butanol as electron donors with fumarate as the electron acceptor. Inputs to the simulations were taken from the ranges observed in the experimental growth screens of each. Each panel shows the line of optimality on the diagonal, the donor limiting region in purple, and the acceptor limiting region in green, and blue regions are regions that outside the functionality of the network. Growth on acetate revealed an optimal ratio of donor and acceptor and a growth rate near the optimal growth rate. Growth on ethanol and butanol revealed that the cells were observed to be functioning in the suboptimal donor-limiting regions and the growth rates are far below the optimal growth rate for the observed acceptor (fumarate) uptake rates.
Examining the cost of extracellular electron transfer

Utilizing experimental growth data of *G. metallireducens* respiring on iron, we were able to hypothesize the cost of transferring electrons to extracellular substrates. External electron transfer was examined using growth data of *G. metallireducens* GS-15 cells respiring with various electron donors and Fe(III) as the electron acceptor. The experimental data obtained for these studies were the concentrations of the electron acceptor and donor over time, and from these measurements, the calculation of the ratio of donor taken up by the cell to acceptor generated by the cell is possible. Given the ratio of this redox reaction, we were able to compare optimal performance of the network to what was observed. For the same three substrates examined when using the dcuB strain, we examined the observed donor to acceptor ratio and compared to modeling simulations. Given the network that was originally reconstructed where electrons are transferred from inner membrane cytochromes to the acceptor, Fe(III), without any energetic cost, the ratio of donor to acceptor was suboptimal for all conditions examined (left panel of Fig. 4.9). To approximate the actual cost of transferring electrons externally to Fe(III), we tested different hypothetical costs that may be associated to electron transfer. These costs were an ATP cost per electron transferred. The approach was to determine the ATP cost needed to shift the observed ratio to the line of optimality for all or some of the conditions. From this, it was determined that cost of 0.4 ATP per electron transferred resulted in shifting the observed ratio to the line of optimality for growth with acetate, ethanol, and butanol as the electron donor. This results shows that there is one consensus change to the extracellular electron transfer pathway to bring all three of the growth conditions tested to an optimal network state, including growth on acetate which was previously described as an optimal substrate.
Figure 4.9: Phenotypic analysis of respiration on extracellular electron acceptors (Fe(III)) Phenotypic Phase Plane (PhPP) analysis of *G. metallireducens* GS-15 to examine the cost of extracellular electron transfer. The reconstructed model of *G. metallireducens* GS-15 was utilized to analyze the growth of wild type *G. metallireducens* GS-15 on acetate, ethanol, and butanol as electron donors with Fe(III) as the electron acceptor. The far left column shows the calculated PhPPs and the observed experimentally measured ratio between substrate uptake and Fe(III) conversion rate (dashed line). The hypothesis of a directly proportional energetic cost to the Fe+3 conversion rate was examined and a cost of 0.4 ATP hydrolyzed per electron transferred to Fe(III) resulted agreement between the line of optimality (i.e., the intersection of Fe+3 conversion limited and substrate uptake limited regions) and the experimentally observed ratio for all three cases.
4.4.2 Analysis of energy conservation and ETS of *C. ljungdahlii* during autotrophic growth

Acetogens have long been thought to be living at the thermodynamic limit due to the energy requirements of the Wood-Ljungdahl pathway [120]. Given that the ATP generated from acetate production is required for the activation of formate, it has been proposed that acetogens should have additional energy conservation mechanisms. Acetogens have generally been classified into those containing respiratory cytochromes that establish a proton gradient (*M. thermoacetica*) and those that do not. Organisms such as *Acetobacterium woodii* have been shown to employ a sodium gradient as an energy conservation mechanism [120]. The genome sequence has revealed that *C. ljungdahlii* falls into a third class of acetogen that neither uses respiratory cytochromes nor a sodium gradient for energy conservation [146]. As reviewed recently, flavin based electron bifurcation is expected to serve as an alternative mechanism of energy conservation in acetogens [122].

Analysis of the iHN633 reconstruction of *C. ljungdahlii* reveals that electron bifurcation and proton translocating ferredoxin oxidation are critical mechanisms for energy conservation during autotrophic growth in *C. ljungdahlii* (Fig. 4.10). Proton translocation by the membrane-bound Rnf complex while oxidizing ferredoxin and reducing NAD was predicted by the iHN633 model to be essential for autotrophic growth (Fig. 4.10B). This prediction was further validated by the lack of autotrophic growth in a strain where the Rnf complex was knocked out (Tremblay et al. personal communication).

An electron-bifurcating transhydrogenase (Nfn complex) that couples exergonic reduction of NADP with ferredoxin to drive the endergonic reduction of NADP with NADH had been characterized in *C. kluyveri* [165]. A homolog to the NfnAB genes in *C. kluyveri* was identified in the *C. ljungdahlii* genome. This gene (CLJU_c37240) was annotated as a glutamate synthase. As part of iHN633, this gene has been assigned to the electron bifurcating transhydrogenase reaction (Fig. 4.10C). This reaction is essential for the generation of NADPH during autotrophic growth. Recently, a similar electron bifurcating NADPH transhydro-
genase activity was also reported in the acetogen *M. thermoacetica* [166].

It was speculated that the highly exergonic reduction of methylene-tetrahydrofolate by NADH can be a site for electron bifurcation to generate additional reduced ferredoxin from NADH [146]. Investigation of the *A. woodii* genome revealed that the small sub unit of the methylene-tetrahydrofolate reductase (MetV) can act as a flavoprotein that could aid in this proposed electron bifurcation, thereby indirectly establishing a proton gradient through the Rnf complex [120]. A homolog to the *A. woodii* MetV gene was identified next to the MetF gene in *C. ljungdahlii* and investigation of the genome revealed a possible co-transcription of the MetV and MetF genes (CLJU_c37610 and CLJU_c37620 respectively). Based on these evidences, an electron bifurcating methylene-tetrahydrofolate reductase reaction was included in the iHN633 reconstruction (Fig. 4.10A).

Another instance of electron bifurcation in the autotrophic growth of *C. ljungdahlii* was observed in the first step of hydrogen activation to reduce ferredoxin. A bifurcating hydrogenase (HydABC) which uses the exergonic reduction of NADH from hydrogen to drive the endergonic reduction of ferredoxin from hydrogen has been identified to perform this process in *A. woodii* [120, 167]. There were two sets of homologous genes identified in the *C. ljungdahlii* genome. However, only one of them (CLJU_c07070) was found to be essential for growth on hydrogen and carbon dioxide (Ueki et al, personal communication). Hence this gene cluster has been assigned to the bifurcating hydrogenase activity in the iHN633 reconstruction (Fig. 4.10D).

In summary, the genome-scale modeling of *C. ljungdahlii* has revealed that flavin based electron bifurcation is found to play a critical role in the essential reactions for autotrophic growth.
Figure 4.10: Energy conservation mechanisms in *C. ljungdahlii*


The genome sequence of *C. ljungdahlii* reported the absence of any quinone
biogenesis genes [146]. However, reannotation and inspection of KEGG [5] revealed partial coverage of the quinone biosynthesis pathway. Subsequent quinone extraction and LC-MS analysis (Methods) revealed the presence of menaquinone derivatives in \textit{C. ljungdahlii}. However, these did not resemble the known respiratory quinones. Hence, it is believed that this quinone-like molecule could play a role in accepting electrons during electrosynthesis. This provides a platform for the identification of a potential novel component involved in electrosynthesis in \textit{C. ljungdahlii}.

### 4.5 Framework for modeling electrosynthesis

A mathematical framework was developed to enable the modeling of microbial electrosynthesis using constraint-based techniques. This framework relied upon a genome-scale model that had the necessary energy conservation mechanisms and electron transfer pathways, carbon fixation pathways that would enable autotrophic growth (Fig. 4.11A). The inputs into the system during electrosynthesis are primarily \(\text{CO}_2\) and electrons from the cathode. Given that \(\text{CO}_2\) availability is unconstrained, the flux of electrons into the cell is the limiting factor in electrosynthesis. This is analogous to constraining a model with the substrate uptake rate during normal growth conditions. The current uptake by the cells on the cathode was transformed into the electron flux using conversion parameters representing Faraday’s constant and the number of cells that can be packed on the surface of the electrode. Electrode dimensions of 2.5 cm x 7.75 cm x 1.25 cm was used for the calculation and the area of a single cell was assumed to be \(2.75 \times 10^{-8}\) sq.cm and its dry weight to be 0.42 pg[168]. A 10% packing efficiency on the electrode was further assumed to calculate the conversion. This resulted in the following relationship between the current drawn and the electron flux (E mmol/gDWh):

\[
\frac{dQ}{dt} = N_{\text{cell}} \times \frac{gDW}{cell} \\
\frac{dQ}{dt} = 2.505 \times E \ \mu A
\]

Using this relationship, the electrosynthetic production of acetate and formate by \textit{G. metallireducens} was simulated by maximizing the energy generation.
(ATPM). Acetate was formed via the reductive TCA cycle and formate was formed as a result of the formate dehydrogenase activity (Fig. 4.11B)

**Figure 4.11: Framework for modeling microbial electrosynthesis** A: Schematic for modeling microbial electrosynthesis. The current drawn by the microbe at the cathode is used as the input constraint to the model which contains the necessary electron transfer pathways and carbon fixation pathways. This is then used to simulate electrosynthetic conditions by maximizing for energy generation and analyze the product formation scenarios. B: Simulation of electrosynthetic production of acetate and formate in *G. metallireducens.*
4.6 Methods

4.6.1 Reconstruction Process

The reconstructions for *G. metallireducens* and *C. ljungdahlii* were generated in a multi-step process. First, the genome annotations for the organism were entered into the UCSD SimPheny (Genomatica, San Diego, CA) database. The next step involved the reconciliation of the existing *G. metallireducens* reconstruction and for *C. ljungdahlii*, the other Clostridial models. If an exact match for the reaction did not exist in the UCSD SimPheny database on the level of metabolites participating in the reaction, they were manually evaluated for inclusion (see below). Next, a comparison of the metabolic content included in the genome annotation not in SimPheny was performed. Reactions that did not exactly match existing content were manually evaluated. Manual evaluation of new content from the annotation and existing genome-scale reconstruction consisted of gathering genetic, biochemical, sequence, and physiological data and reconciling this information to determine the likelihood of each reaction being present in the organisms. This manual curation process has been described and reviewed several times [23, 32, 125]. In the manual review process, the KEGG database [5], the MODEL SEED database [6], and primary literature were used extensively in the manual curation process. Confidence scores were given for each reaction along with noteworthy evidence used to justify inclusion of a given reaction.

4.6.2 Generation of the Biomass Objective Function

The Biomass Objective Functions for both *G. metallireducens* and *C. ljungdahlii* were formulated using a previous template [127]. For *G. metallireducens*, the biomass content previously determined for the close species *Geobacter sulfurreducens* was used to determine the breakdown of macromolecules [48] except that for total carbohydrate as the distribution in the murein, lipopolysaccharide, and cytosolic fractions [169] was not indicated. Further, the genome annotation [170] was used for the breakdown of chromosome bases. A study on lipid and LPS chain length was used for the breakdown of acyl chain length [171] and the remaining
content was approximated using the full profile presented for the gram-negative bacterium \textit{E. coli} [127].

For \textit{C. ljungdahlii}, a similar template was followed [127, 172]. The protein content was determined to be 43\% and for the rest of the macromolecular breakdown, the biomass content previously determined for a gram positive bacterium \textit{B. subtilis} and \textit{C. beijerenckii} was used [173, 148]. However, for modeling teichoic acid composition, the distribution based on \textit{S. aureus} was used due to similarity in terms of low G-C content [174].

It should be noted that prediction of growth rate and unmeasured uptake rates are relatively insensitive to realistic variations in biomass macromolecular weight fractions [127].

4.6.3 Flux Balance Analysis Simulations

The reconstructed metabolic networks were represented in a mathematical format in a stoichiometric matrix \textbf{S}, where the rows correspond to the metabolites and columns correspond to the reactions in the network. Flux Balance Analysis simulations were carried out as described in [175] using the COBRA Toolbox [45] and the SimPheny framework (Genomatica, Inc., San Diego, CA) were used for simulations. The objective used in the simulations were typically maximizing growth through the biomass objective function.

4.6.4 Bacterial growth conditions

\textit{Clostridium ljungdahlii} (ATCC 55383) was grown anaerobically in 100 ml of PETC medium (ATCC medium 1754) within 125 ml serum bottles at 37\degree C. Nitrogen was used as a carrier gas at a flow rate of 20 ml/min. Fructose, acetate and ethanol concentrations were measured by HPLC, 717 plus AutoSampler (Waters) using 5 mM sulfuric acid as mobile phase. Detection was done at 410 nm. \textit{G. metallireducens} GS-15 and dcuB strains were grown on fresh water media with 55mM Fe(III) citrate and 40mM fumarate as the electron acceptors respectively. Both the strains were grown with acetate, ethanol, and butanol as electron donors.
Acetate, ethanol, butanol, fumarate, and succinate concentrations were measured using the same HPLC as used for *C. ljungdahlii* growth screens. The growth rate during iron respiration was calculated from the Fe(II) production rate.

### 4.6.5 Quinone profiling

The quinone extraction was performed in *C. ljungdahlii* as in the protocol outlined previously [176]. The extraction procedure was also implemented on *E. coli* as a positive control. First, 2 ml of *E. coli* and 4 ml of *Clostridium ljungdahlii* were quenched with 6 ml of ice-cold methanol (Sigma Aldrich). 6 mL of petroleum ether was mixed with solution and the mixture was vortexed for 1 min. The mixture was then centrifuged at 900xg for 2 min and the top phase was transferred to a new tube. Another 3 mL of petroleum ether (Sigma Aldrich) was added to the residual mixture and the steps after addition of petroleum ether were repeated. The upper phase were mixed with the previous upper phase and the mixture was dried using SpeedVac 110-120 (Savant). Finally, the dried extracted quinones were suspended in 100 µL of ethanol (Sigma Aldrich). Quinone analysis was conducted using both HPLC system with Pursuit XRs (Varian, CA) C18 reverse phase column with methanol as a mobile phase and flow rate of 1.0 ml/min at ambient temperature, and Accurate-Mass TOFMS 6230 (Agilent) at the UCSD Chemistry Core. As standards, ubiquinone-10 and menaquinone-4 were used. A dual-wavelength UV detector with 290 nm for ubiquinone and 248 nm for menaquinone detection was used.

### 4.7 Conclusions

In this study, functional genome-scale metabolic networks have been generated for two electrosynthetic bacteria *G. metallireducens* and *C. ljungdahlii*. These networks represent a detailed description of the key carbon fixation and energy conservation pathways the two major metabolic constraints governing electrosynthesis. Computational modeling has revealed new relevant growth conditions critical for characterizing autotrophic metabolism. In addition, physiological experiments
have enabled the elucidation of novel components and key players of the electron transport system in these electrosynthetic bacteria. Constraint-based metabolic modeling has been successfully applied to various metabolic engineering projects. These functional models will enable strain design strategies for optimal electrofuel production. The advances made have, for the first time, enabled a genome-scale characterization of electrosynthesis.

4.8 Acknowledgements

We would like to thank Dr. Su at the UCSD Chemistry core for his help with the quinone extraction procedure. This work was supported by the ARPA-E Electrofuels program U.S. Department of Energy, under Award Number DE-AR0000087 and by Office of Science (BER), U.S. Department of Energy (DE-SC0004485).

Chapter 4 is in part, adapted from a paper that is being prepared for publication Feist AM, Nagarajan H, Aklujkar M, Tremblay PL, Rotaru AE, Lovley DR, Zengler K. Integrated model-driven and experimental characterization of the growth capabilities of Geobacter metallireducens GS-15. The dissertation author was one of the primary researchers responsible for the reconstruction and analysis of carbon fixation and energy conservation.

Chapter 4 is in part, adapted from a paper that is being prepared for publication Nagarajan H, Nogales J, Sahin M, Embree M, Ebrahim A, Feist AM, Lovley DR, Palsson BO, Zengler K. Using genome-scale models to aid biological discovery of acetogenic metabolism. The dissertation author was the primary author of this paper responsible for the research.
Chapter 5

Multi-omic characterization and modeling of interspecies electron transfer mechanisms and microbial community dynamics of syntrophic associations

5.1 Abstract

Syntrophic associations are central to methanogenic communities and thus play a fundamental role in governing major global biogeochemical cycles. While biochemical approaches have described the physiological activity of these communities, there has been a lack of a mechanistic understanding of the relationship between the complex nutritional and energetic dependencies and their functioning. Here, we present an integrated multi-omic workflow that combines genomic, transcriptomic, and physiological data with constraint-based modeling using genome-scale models to characterize interspecies electron transfer mechanisms and community dynamics of syntrophic associations. We apply this workflow to describe the syntrophic ethanol oxidation by Geobacter metallireducens and Geobacter sul-
furreducens at three levels of complexity. First, through a detailed reconstruction of energetics involved in extracellular electron transfer, we provided mechanistic insights into direct interspecies electron transfer (DIET) between the two organisms. We further showed that genomic characteristics such as energy conservation via respiration as opposed to substrate level phosphorylation, proton availability, and optimality levels of interacting organisms influences the preference of DIET over interspecies hydrogen transfer. Second, metatranscriptomic profiling in conjunction with genome-scale modeling uncovered a synergistic relationship between acetate transfer and DIET. Third, investigating the dynamics of the community revealed that, while G. sulfurreducens responded to the selection pressure of rapid syntrophic growth at the genomic and transcriptomic level, G. metallireducens responded only at the transcriptome level. This further suggests that G. sulfurreducens is under a greater selection pressure to adaptively evolve and streamline its metabolic capabilities in order to establish DIET and maintain syntrophy. This multi-omic approach will thus enhance our understanding of the dynamics and functioning of syntrophic communities.

5.2 Author Summary

The decomposition of organic matter in anoxic environments is a major part of the global carbon cycle. This degradation relies in large part on the transfer of metabolites involving different organisms. While biochemical approaches have been able to shed light on the overall physiological activity and mechanisms of such syntrophic associations, it has been challenging to gain comprehensive insights into the factors favoring different electron transfer mechanisms, genomic features of individual constituents, as well as the dynamics of these communities. We applied a multi-omic modeling workflow that combines genomic, transcriptomic, and phenomic data with constraint-based modeling using genome-scale models to investigate metabolic dynamics and mechanisms of electron flow in the syntrophic association of Geobacter metallireducens and Geobacter sulfurreducens. Through a detailed reconstruction of pathways involved in extracellular electron transfer,
mechanistic insights into direct interspecies electron transfer between the two organisms are provided. Furthermore, *G. sulfurreducens* adapted to the selection pressure of rapid syntrophic growth by changes on the genomic and transcriptomic level, while *G. metallireducens* responded only at the transcriptomic level. This could have implications for our understanding of adaptive responses and factors that shape the evolution of these syntrophic communities, thereby expanding our knowledge of the global carbon cycle.

5.3 Introduction

Microorganisms in nature do not exist as pure cultures, but rather are involved in a wide variety of interactions and nutritional interdependencies with their bacterial and eukaryotic neighbours in their ecosystem. Syntrophy is one such nutritional dependency where two or more microbes combine each other’s metabolic capabilities for the degradation of a particular substrate that neither is capable of performing individually [177]. Syntrophy in which electrons are shared between two partners are central to the functioning of a variety of methanogenic environments such as wetlands, aquatic sediments, animal intestinal systems, and oil reservoirs, as well as in anaerobic digesters converting organic wastes to methane[178, 179]. Furthermore, interspecies electron transfer is considered to play an important role in microbial aggregates that anaerobically oxidize methane with the reduction of sulphate [180].

Traditional biochemical approaches have been able to shed light on the overall physiological activity and mechanisms of electron transfer in many of these microbial communities [55, 56]. Despite these approaches, it has been challenging to gain comprehensive insights into the dynamics, the factors favouring one mechanism of interspecies electron transfer over the other, as well as genomics features for the metabolic activity of individual constituents. This has necessitated for the development of genome-scale approaches that can provide a link between the genotypes of the microbes, their reaction mechanisms, and eventually their effect on the functioning of the microbial community [57, 58]. One of the major
goals of microbial ecology and systems biology is to comprehensively understand, characterize, and obtain meaningful insights into the various modes of interactions in communities and their functional effect on the ecosystem [59].

Constraint-based modeling approaches (COBRA) have provided significant mechanistic insights into the genotype-phenotype relationship of single bacterial species [97]. Recently, several COBRA methods have been deployed to investigate the various microbial interactions such as mutualism, competition, and parasitism [97, 34, 36, 181, 37, 35]. One of the first such modeling studies applied flux balance analysis to investigate the mutualism between the sulfate-reducing bacterium *Desulfovibrio vulgaris* and the methanogenic archaeon *Methanococcus maripaludis* [35]. An extension of dynamic flux balance analysis has also been used to characterize the competitive relationship between *Geobacter sulfurreducens* and *Rhodoferax ferrireducens* in the subsurface [2]. A COBRA-based computational framework has been applied to determine the media compositions that induce putative symbiotic interactions between all possible combinations of seven different microorganisms [34]. While this provided the proof-of-concept for integration of multiple genome-scale metabolic models, the nature of the microbial partners evaluated in this study precluded a detailed mechanistic investigation into all possible modes of interactions. Most recently, a general multi-objective optimization framework has been developed for analyzing interspecies metabolite transfers during various types of multi-species interactions[181]. Extension of these community systems (CoSy) biology [59] approaches for studying syntrophic associations requires a comprehensive characterization of the various mechanisms of energy transfer and their effect on the functioning of the microbial community.

The long standing paradigm for interspecies electron transfer in anaerobic methanogenic consortia has been interspecies hydrogen transfer, in which the electron-donating partner disposes of electrons by reducing protons to produce hydrogen and the electron-accepting partner accepts electrons by oxidizing hydrogen with the reduction of carbon dioxide to methane [177]. However, direct interspecies electron transfer (DIET) appears to be an alternative to interspecies hydrogen transfer in some methanogenic environments [17, 18, 8]. Defined co-cultures of
Geobacter metallireducens and Geobacter sulfurreducens grow syntrophically via DIET in medium with ethanol as the electron donor and fumarate as the electron acceptor [19, 20, 21] and provide a model for syntrophic growth in which both partners are genetically tractable [182, 183]. Furthermore, a syntrophic co-culture in which G. sulfurreducens also functions as the electron-accepting partner, but which functions via interspecies hydrogen transfer, is available for comparison [20].

Here, we expanded the scope of CoSy biology by developing an integrated modeling framework that combines physiological, transcriptomic, and next generation sequencing (NGS) data with genome-scale metabolic models to simulate and investigate syntrophic interactions and their associated electron transfer mechanisms. Specifically, we applied this framework to characterize the mechanistic basis of the syntrophic association of G. metallireducens and G. sulfurreducens. By modeling different genetic perturbations of this system, we provide additional insights into the process of DIET, hypothesized to be an important mode of electron transfer in certain syntrophic associations. We further use transcriptomic and genomic data to describe the dynamics of the interaction between the two individual Geobacter species and elucidate their adaptive response to syntrophic growth.

5.4 Results

5.4.1 Integrated framework for modeling syntrophy and direct electron transfer

A four-step modeling workflow was developed for the genome-scale investigation of metabolic interactions and electron flow in the syntrophic association between G. metallireducens and G. sulfurreducens. The first step was to represent the extracellular electron transfer pathways through all the known electron carriers in the cell, including cytochromes, quinone, ferredoxin, NAD, and FAD (Fig. 5.1A). In addition to accurately assigning the stoichiometry associated with the respective redox reactions, these pathways also account for thermodynamic
consistency, appropriate gene association, and cellular localization of the different electron carriers. This protocol was implemented to significantly expand and update the previous genome-scale reconstructions of *G. metallireducens* and *G. sulfurreducens* [124, 184, 27], to explicitly account for a detailed representation of the energy metabolism involved in extracellular electron transfer. The reconstructions were also updated to included the most updated annotations, and a distinct periplasm compartment. This resulted in a *G. metallireducens* model (iAF987) consisting of 987 genes and 1284 reactions and a model for *G. sulfurreducens* (iHN837) consisting of 837 genes and 1085 reactions, respectively (Fig. 5.1B, Methods). In the third step, modeling of metabolic interactions between the two species was enabled by integrating the two expanded genome-scale metabolic models into a combined model using an extension of the compartmentalized modeling approach [36, 181, 34]. This integrated genome-scale model of syntrophy consists of five compartments, where each species was assumed to have independent cytosol and periplasm compartments. The fifth compartment, the shared metabolite pool (SMP), represents an extracellular environment common to both the constituents. This integrated model consisted of 1824 genes and 2333 reactions in total. The transport reactions in the respective genome-scale models were interfaced with the SMP to facilitate metabolic exchanges between the constituents of the consortium as well as with the environment (Fig. 5.1C). In the fourth step, direct interspecies electron transfer (DIET) between the two species was accounted for by introducing a reaction that represents the stoichiometric transfer of electrons between metabolites that correspond to their respective outer-membrane cytochromes. Outer-membrane cytochromes in *Geobacter* species have been shown to play a critical role in extracellular electron transfer to minerals as well as other external electron acceptors such as the anode of a microbe fuel cell[101, 185]. Specifically, the gene encoding OmcS of *G. sulfurreducens* was shown to be essential for DIET in this syntrophic consortium [19]. Hence, this gene has been associated with the reaction representing DIET. The maximum electron transfer flux through DIET ($V_{\text{max}}^{\text{DIET}}$) was calculated based on the midpoint potential of OmcS (Fig. 5.1D, Methods). This integrated workflow enabled a multi-level characterization of the metabolic
interactions and electron flow in the syntrophic association of *G. metallireducens* and *G. sulfurreducens*.

---

**Figure 5.1: Integrated framework for modeling syntrophy and direct electron transfer.**  A: Protocol adopted for reconstructing extracellular electron transfer pathways. B: Properties of the updated and expanded genome-scale metabolic reconstructions used. C: Formalism for reconstructing syntrophy. Using an extension of the compartmentalized framework, a shared metabolite pool (SMP) is included to facilitate exchange of metabolites between the species as well as with the environment. D: Modeling Direct Interspecies Electron Transfer (DIET). The maximum flux through DIET is calculated based on the midpoint potential of OmcS.
5.4.2 Analysis of carbon and electron flow governing the syntrophic association

Constraint-based modeling methods were applied to the integrated genome-scale metabolic model to characterize the carbon and electron flow defining the syntrophic association of *G. metallireducens* and *G. sulfurreducens* with ethanol and fumarate as the respective electron donor and acceptor. The community model was optimized for the objective of maximal growth of both constituents (*G. metallireducens* and *G. sulfurreducens*) with the addition of a constraint for simulating DIET. Specifically, the electron transfer flux through DIET ($V_{\text{DIET}}^\text{max}$) was constrained to its maximal value ($V_{\text{DIET}}^\text{max}$) calculated based on the midpoint potential of the cytochrome OmcS (Fig. 5.1D, Methods). Experimentally measured physiological uptake rates (Fig. 5.2) and product secretion rates were used as the input/output constraints for the simulation. Thermodynamically infeasible loops were eliminated from the simulations using the loopless COBRA (ll-COBRA) procedure [186]. The ratio of flux through the respective biomass objective functions was used to estimate the relative abundance of the constituents in the community. Model simulations revealed that *G. sulfurreducens* accounted for 73% of the consortium (Fig. 5.3). This prediction is in accordance with experimental observations of a *G. sulfurreducens* dominated coculture with a composition of around 75% *G. sulfurreducens* [19]. DIET was predicted to be the optimal mode of electron transfer in this consortium with electron transfer flux through DIET being at its maximal value ($V_{\text{DIET}}^\text{max}$). Further analysis of the flux distribution revealed that *G. metallireducens* secreted acetate into the SMP as a result of ethanol oxidation. The secreted acetate was predicted to be utilized by *G. sulfurreducens* as a carbon source for growth. Acetate was also predicted to serve as an electron donor for fumarate reduction via the TCA cycle in *G. sulfurreducens* (Fig. 5.4). To assess the accuracy of these model-predicted flux distributions, a context-specific model based on transcriptomic profiling of this coculture was constructed using the GIMME algorithm [41]. This algorithm implements an optimization framework as well as a scoring function to determine the model that best describes the gene expression levels and metabolic objectives of the cell under the given conditions. The flux
distribution from the resulting context-specific model consisting of 988 reactions and 993 metabolites was found to be 99% consistent with transcriptomic data. Reactions such as citrate synthase and aconitase of *G. metallireducens*, formate dehydrogenase and the uptake hydrogenase of *G. sulfurreducens*, were not active in this context-specific model. This prediction agrees with the demonstration of syntrophic growth between *G. metallireducens* and a mutant of *G. sulfurreducens* lacking both the uptake hydrogenase and formate dehydrogenase [20]. An *in silico* simulation where each reaction was knocked out individually to assess their effect on the growth revealed that apart from uptake of ions, the main reactions predicted to be essential for sustenance of syntrophic interaction in this consortium are those involved in the ethanol oxidation, fumarate reduction, and direct electron transfer between *G. metallireducens* and *G. sulfurreducens*. 
Figure 5.2: Physiological growth screens performed on the WT and CS K.O strains. The uptake rates calculated from this data were used to constrain the model. While constraining the model with uptake rates, the fumarate uptake rate was calculated as the net fumarate and malate rates.
Figure 5.3: Optimal fraction of community constituents. To estimate the relative abundance of the community, the growth rate was sampled over a range of fixed relative abundance of *G. metallireducens*. This resulted in a predicted optimal fraction of 27-30% of *G. metallireducens*. 
Since both \textit{G. metallireducens} and \textit{G. sulfurreducens} are capable of fixing nitrogen, the impact of this energy intensive process on the syntrophic interaction between them was investigated by simulating syntrophic growth in ammonia free media. This analysis predicted that if \textit{G. sulfurreducens} were to be the organism fixing nitrogen for the community, the composition of the community would have to be at least 45\% \textit{G. metallireducens}. On the other hand, no feasible interaction was predicted if \textit{G. metallireducens} had to be the organism solely responsible for fixing nitrogen for the community. This prediction is validated by the lack of syntrophic growth observed in a coculture of \textit{G. sulfurreducens nifD} mutant which is unable to fix nitrogen [182] and a \textit{G. metallireducens} strain pre-adapted to fix nitrogen in a medium containing $N_2$ gas as the sole nitrogen source (Fig. 5.5).
Figure 5.5: Impact of nitrogen limiting conditions on the syntrophic association. To assess the impact of nitrogen fixation on the syntrophic association, simulations were carried out in ammonia free media. In a scenario where *G. sulfurreducens* is the sole organism fixing nitrogen, the maximum relative abundance *G. sulfurreducens* could achieve was around 45%. No feasible scenarios were computed for syntrophic growth of both organisms, if *G. metallireducens* had to be the sole organism fixing $N_2$.

5.4.3 Transcriptomic profiling analyzed in the context of the integrated model provides insights into the different mechanisms of electron transfer

To gain further insights into the different mechanisms of electron transfer in the consortium, we modeled two other mutant strains, where the possibility of acetate transfer or hydrogen transfer was eliminated. This was achieved by the deletion of citrate synthase (CS K.O. strain) and the uptake hydrogenase of *G. sulfurreducens* (hyb K.O. strain) to eliminate acetate oxidation and hydrogen transfer, respectively (Table. 5.1) [21]. Transcriptomic profiling by RNA-seq
on these three strains provided an additional layer of context to the community
model to assess the functional effect of these perturbations to electron transfer
mechanisms.

**Table 5.1**: Details of the different mutant strains modeled to obtain insights into
mechanisms of electron transfer

<table>
<thead>
<tr>
<th>Strain</th>
<th>Gene deleted</th>
<th>Possible modes of electron transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>none</td>
<td>DIET + Acetate + Hydrogen</td>
</tr>
<tr>
<td>CS K.O.</td>
<td><strong>gltA</strong> (Citrate Synthase) of <em>G. sulfurreducens</em></td>
<td>DIET + Hydrogen</td>
</tr>
<tr>
<td>hyb K.O.</td>
<td><strong>hybA</strong> (Hydrogenase) of <em>G. sulfurreducens</em></td>
<td>DIET + Acetate</td>
</tr>
</tbody>
</table>

Eliminating the possibility of hydrogen transfer resulted in the differential
expression of 145 genes in *G. sulfurreducens* (Fig. 5.6A). Specifically, we observed
an increased expression (> 4 fold) of the outer membrane cytochromes (*omcS, omcT*)
that are believed to be involved in DIET and a significant downregulation
(> 15 fold) in expression of the other hydrogenase (*hya*) in addition to the deleted
*hyb*. Consistent with the lack of hydrogenase activity in the context-specific model
for the WT strain, the predicted flux distribution for the hyb K.O. strain was simi-
lar to that of the WT strain. This similarity was further supported by comparable
ethanol and fumarate uptake rates observed in both these strains [21] (Fig. 5.6B).

Eliminating the possibility of electron transfer through acetate oxidation by
the deletion of the citrate synthase gene of *G. sulfurreducens* (CS K.O. strain) re-
sulted in the differential expression of 485 genes in both *G. metallireducens* and G.
sulfurreducens (Fig. 5.6A). 145 genes out of the 485 differentially expressed genes
encoded for at least one reaction in the genome-scale model. A statistical analysis
on these 145 genes based on the metabolic subsystems revealed a significant en-
richment for genes involved in energy metabolism and electron transfer (p-value
0.01). Experimentally measured physiological uptake rates for the CS K.O. strain
were used to constrain the model and to simulate the effect of the elimination
of additional electron transfer via acetate. The optimal ratio of the constituents in this consortium was predicted to be 44% *G. metallireducens*, which compared well with the experimental observation of about 50%. As expected, 67% acetate secreted by *G. metallireducens* accumulated in the SMP, due to the inability of *G. sulfurreducens* to oxidize it. In accordance with this prediction, the expression of acetate transporter genes in *G. sulfurreducens* decreased significantly (>15 fold) compared to that of the WT strain. Similar to the WT strain, DIET was predicted to be preferred over interspecies hydrogen transfer in the CS K.O. strain. However, the electron transfer flux through DIET in the CS K.O. strain occurred at just 75% the maximal value ($V_{\text{DIET}}^{\text{max}}$). This suboptimal electron transfer scenario observed in the CS K.O. strain could potentially result in a lower electrical conductivity in this consortium compared to the WT strain.
Figure 5.6: Insights into mechanisms of electron transfer from transcriptomic data. A: Fold changes observed for the differentially expressed genes in the hyb K.O strain and CS K.O strain. B: Metabolic interpretation of the transcriptomic differences in the hyb K.O strain relative to the WT strain. The main genes unregulated in the hyb K.O strain include the outer membrane cytochromes that are associated with the pathway for DIET in the consortium. C: Metabolic adjustments in the CS K.O strain based on the transcriptomic data. The elimination of interspecies electron transfer via acetate oxidation confers a suboptimal direct electron transfer phenotype to the consortium. This is mirrored by the significant suppression of the electron transfer pathway involved in DIET in *G. metallireducens*. D: Interaction motif describing the synergistic and antagonistic relationship between DIET, acetate transfer and interspecies hydrogen transfer.

The metabolic adjustment to this functional state was investigated by integrating the fold changes in gene expression of the two strains with the genome-scale model using MADE [144]. This analysis revealed that the agreement between metabolic adjustment predicted by the model with the transcriptomic profiling was 87% accurate with a significant increase in flux for 17 reactions and decrease in flux for 91 reactions. The reactions showing significant decrease in fluxes corresponded to the electron transfer pathway of *G. metallireducens* involved in DIET (CYTMQOR3ppD, NADH17ppD, OMCPPCO_D) (Fig. 5.6C). The changes
in the flux distribution observed from expression data were also further corroborated by the use of randomized Monte-Carlo sampling on the genome-scale model (Methods). This significant suppression of the critical components of the electron transfer pathway out of \textit{G. metallireducens} mirrors the suboptimal electron transfer scenario observed upon eliminating the additional electron transfer via acetate (Figure 3C). Taken together, the integrated analysis of transcriptomic data with genome-scale model of syntrophy revealed a unique interaction motif between the different modes of electron transfer. Since the elimination of acetate transfer conferred a suboptimal electron transfer and the elimination of the hydrogenase gene enhanced the expression of genes involved in DIET, we postulate that acetate transfer is synergistic to DIET and interspecies hydrogen transfer has an antagonistic relationship with DIET (Fig. 5.6D).

5.4.4 Investigating the dynamics of adaptation to syntrophy

The obligate syntrophic association between the two \textit{Geobacter} species was achieved through adaptive laboratory evolution under a selection pressure to switch from their Fe(III)-reducing growth mode to rapid syntrophic growth in ethanol-fumarate medium. Other members of the \textit{Geobacteraceae} like \textit{Pelobacter carbinolicus}, have been previously shown to evolve from an Fe(III) reducer to a syntroph performing interspecies hydrogen transfer [187, 20, 188]. In the association described in this study, \textit{G. metallireducens} functions as an ethanol-oxidizing syntroph performing DIET[19]. On the other hand, the selection pressure dictates \textit{G. sulfurreducens} to depend on \textit{G. metallireducens} to derive its carbon source as well as electrons for growth while reducing fumarate. The genome-scale model of the consortium was used to investigate what effect the selection pressure had on how the partners adapted for syntrophic growth as well as their interaction dynamics. First, we assessed the entire spectrum of feasible interactions in this consortium by evaluating the effect of relative rates of ethanol oxidation and fumarate reduction. This phenotypic phase-plane analysis [164] indicated the presence of two distinct feasible phases (Phase A and Phase B) based on the presence or absence of a lim-
itation on fumarate reduction rate. The line of optimality represents the ethanol and fumarate uptake rates that correspond to optimal growth of the consortium (Fig. 5.7). Phase A is closer to the line of optimality and there is no limitation due to fumarate uptake rate. In addition, this phase is also characterized by maximal electron transfer flux through DIET as described by $V_{\text{DIET}}^{\text{max}}$. This state reflects the ideal scenario for both $G. \text{metallireducens}$ and $G. \text{sulfurreducens}$ in this consortium. Experimentally determined uptake rates of WT aggregates are representative of Phase A, with maximal electron transfer through DIET. Typically, in Phase A, $G. \text{sulfurreducens}$ dominates the consortium and is most abundant. Phase B reveals a scenario where a limitation due to fumarate uptake rate results in a state of suboptimal electron transfer flux. Specifically, the electron transfer flux from $G. \text{metallireducens}$ to $G. \text{sulfurreducens}$ is limited by the rate of electrons that $G. \text{sulfurreducens}$ can transfer to fumarate. While, $G. \text{sulfurreducens}$ still derives the requisite amount of electrons to reduce fumarate, $G. \text{metallireducens}$ is forced to find other ways to transfer the electrons it derives from ethanol oxidation. An example of the physiological characteristic exhibited in Phase B would be the CS K.O strain, with a relative abundance of $G. \text{metallireducens}$ closer to 50%, the inability to oxidize acetate, and suboptimal DIET. This dynamics suggests that $G. \text{sulfurreducens}$ is the partner likely to have actively responded to the selective pressure of rapid syntrophic growth. Consistent with this, a 1 bp deletion in GSU1495 encoding PilR of $G. \text{sulfurreducens}$ shown to promote syntrophic growth via direct exchange of electrons in cocultures after 100 generations, was preserved at the end of 780 generations. Strikingly, even after 780 generations, no genetic mutations were observed in $G. \text{metallireducens}$. While $G. \text{sulfurreducens}$ undergoes several genomic changes in addition to transcriptomic changes to actively respond to the pressure of rapid syntrophic growth, $G. \text{metallireducens}$ responds only at the level of the transcriptome by regulating its expression (Fig. 5.6C).
Figure 5.7: Dynamics of interaction between the partners as they respond to selection pressure of syntrophic growth. Phenotypic phase plane analysis reveals the presence of two distinct phases with maximal and suboptimal direct interspecies electron transfer.

5.4.5 Comparative analysis of interspecies electron transfer mechanisms

In contrast to electron exchange via DIET in the *G. metallireducens/G. sulfurreducens* co-cultures, co-cultures of *Pelobacter carbinolicus* and *G. sulfurreducens* growing in a similar ethanol/fumarate medium exchange electrons via interspecies hydrogen transfer [20]. This difference can be attributed in part to the inability of *P. cabinolicus* to make effective electrical connections with extracellular electron acceptors [189] as well as the inability of *G. metallireducens* to generate hydrogen gas [190]. Electron transfer mechanisms are fundamental for syntrophic interactions to exist, and therefore it is important to assess their influence on the energetic yields of the interacting organisms. Therefore, for an appropriate com-
parison of the energy yields of the respective partners during the different modes of interspecies electron transfer, the coculture of *P. carbinolicus* and *G. sulfurreducens* was used as the system of choice for the hydrogen transfer scenario.

The genome-scale metabolic model of *G. metallireducens* was modified to account for the major differences involved in ethanol and hydrogen metabolism of *P. carbinolicus*. This modification mainly included the production of hydrogen via electron confurcation due to the addition of a NADP-dependent hydrogenase and the Nfn complex, the addition of the Rnf complex, and the acetylating acetaldehyde dehydrogenase. The phenomenon of electron confurcation has been shown to play an important role in interspecies hydrogen transfer during syntrophic associations [191]. The relative efficiency of the different modes of electron transfer was evaluated using the genome-scale model of this coculture and the syntrophic consortium of the two *Geobacter* species to obtain insights into the potential differences in energy conservation mechanisms. Specifically, the maximum molar yield of ATP obtained per mole of electron donor (ethanol) for both the partners in each of the different scenarios was computed.

The predicted energy yields of the two partners during DIET mirrored the ratio of the relative abundance observed in the CS K.O. strain. The presence of additional electron transfer via acetate was found to enhance the efficiency of DIET by 24% (Fig. 5.8A). In this case, the predicted relative energy efficiencies of the two partners involved were consistent with the observed and predicted values of their relative abundances in the WT strain. The energy yield of *G. sulfurreducens* was predicted to be about 82% of the total energy yield during hydrogen transfer. This prediction is in good agreement with the reported relative abundance of 85% *G. sulfurreducens* in its syntrophic association with *P. carbinolicus* [20]. While the energy yield for the electron accepting organism (*G. sulfurreducens*) during interspecies hydrogen transfer was of similar order of magnitude compared to DIET (2.7 vs 3.5), the energy yield of the electron donating organism was about 4 times lower during the case of interspecies hydrogen transfer (Figure 5A). The metabolic characteristic governing this significant difference was found to be the
mode of energy conservation adopted by *P. carbinolicus* and *G. metallireducens* during interspecies hydrogen transfer and DIET respectively. *P. carbinolicus* was found to primarily generate ATP via substrate level phosphorylation, whereas *G. metallireducens* conserved energy via respiration aided by the establishment of a proton gradient during DIET (Fig. 5.9). This mechanistic difference potentially explains the higher relative abundance of the electron accepting organism during interspecies hydrogen transfer as observed in the coculture of *P. carbinolicus* and *G. sulfurreducens* [20].

**Figure 5.8: Comparative analysis of interspecies electron transfer mechanisms.** A. Relative efficiency of the different modes of interspecies electron transfer. DIET+Acetate case represents the WT strain, DIET case represents the CS K.O strain and the hydrogen transfer case is simulated using the coculture *P. carbinolicicus* and *G. sulfurreducens*. The maximum energy yield per mol of electron donor is calculated for each mode of interspecies electron transfer. The theoretical maximum for the net reaction in this consortium is 8.27. The pie charts represent the relative abundance of the interacting partners during each mode of interspecies electron transfer. B. Relationship between community dynamics and mode of interspecies electron transfer. Simulations were carried out with a community model capable of doing hydrogen transfer and DIET. Simulations were conducted over the entire range of optimality levels of both the interacting organisms and whether DIET or hydrogen transfer occurred was monitored. The blue area represents the sole occurrence of DIET during the interaction and the orange area shows the region where DIET and hydrogen transfer occur. Interspecies hydrogen transfer occurred in addition to DIET, in regions where the optimality levels of the electron accepting organism was at least 1.8 times that of the electron donating organism.
The effect of microbial community dynamics on the preferential occurrence of DIET or interspecies hydrogen transfer was analyzed using a genome-scale model of syntrophy between G. metallireducens with the added capabilities of interspecies hydrogen transfer and G. sulfurreducens. The biochemical constraints that govern microbial communities often result in constituents operating at varying levels of optimality that compromises their respective growth objectives [192]. Using the genome-scale model, the entire spectrum of optimality for both the electron donating and electron accepting organisms were scanned to assess the occurrence of DIET and interspecies hydrogen transfer. It is important to note that the level of optimality is not equal to relative abundance of cells, but rather reflects the strategy adopted to optimize biomass yield. It was observed that interspecies hydrogen transfer occurred in addition to DIET, whenever the optimality levels of
the electron accepting organism was at least 1.8 times that of the electron donating organism (Fig. 5.8B). In other words, when interspecies hydrogen transfer was modeled in silico to occur along with DIET, *G. sulfurreducens* tends to adopt a strategy that is relatively more optimal. If the community operates in a manner such that the relative optimality levels of the two organisms are of similar orders of magnitude, it is predicted that the energetic requirements of *G. metallireducens* preclude the possibility of interspecies hydrogen transfer occurring. The is largely due to the energetic advantage conferred to the electron donating organism due to the proton gradient established during DIET compared to interspecies hydrogen transfer (Fig. 5.9). To further explore other environmental factors that could determine the preference of DIET as the mode of electron exchange, a sensitivity analysis was performed. This revealed that the proton intake flux by the partner accepting electrons (*G. sulfurreducens*) was the critical factor that affected the efficiency of DIET (Fig. 5.10). While proton intake flux had no impact on the energy efficiency during interspecies hydrogen transfer, it had a linear effect on the energy yield in DIET and DIET+acetate transfer. Hence, a combination of genomic characteristics, environmental factors such as proton availability in the local microenvironment, and aspects of community dynamics like relative optimality of electron acceptor and donor organisms are expected to play a role in determining the preference of interspecies electron transfer mechanisms.
Figure 5.10: Effect of proton availability on energy efficiency of the different modes of electron transfer. Robustness analysis was performed on the influx of protons into *G. sulfurreducens* to assess its impact on energy yield of DIET, DIET+acetate transfer and interspecies hydrogen transfer. While the proton flux had a linear relationship on the efficiency of DIET and DIET+acetate transfer, there was no impact observed in the case of interspecies hydrogen transfer.

5.5 Discussion

Microorganisms engage in diverse partnerships and complex interactions in their natural environment. The extent of these microbial partnerships and the nature of their interactions play a crucial role in the cycling of nutrients and energy in the environment. Syntrophic associations between microorganisms are dominant mediators of all the major biogeochemical cycles in anoxic environments [178, 179, 180]. The obligate nature of most of these syntrophic associations necessitate a tight coupling of the metabolism and energetics of the interacting microorganisms, typically resulting in the formation of compact aggregates that operate at thermodynamic equilibrium [177]. A comprehensive characterization of
these complex nutritional and energetic dependencies is fundamental to the understanding of this crucial environmental process. Despite the rapid development of high-throughput sequencing and metagenomic technologies, almost all modeling approaches used to study these interspecies interactions have been limited to characterizing the metabolic exchanges between the constituents without accounting for the functional effect of the potentially different mechanisms of reducing equivalent transfer. Here, for the first time, we present a multi-omic modeling workflow to investigate metabolic dynamics and mechanisms of electron flow in the laboratory evolved syntrophic coculture of *G. metallireducens* and *G. sulfurreducens* at three distinct levels of complexity.

The detailed reconstruction of the energetics associated with extracellular electron transfer pathways enabled the mechanistic basis of investigation of the relative efficiencies of direct interspecies electron transfer (DIET) versus other alternatives. In the syntrophic association of *G. metallireducens* and *G. sulfurreducens*, the model predictions reaffirmed the prevalence of DIET as the mode of electron exchange. The reactions essential to the maintenance of this syntrophic association exclusively corresponded to those involved in DIET. Furthermore, the energetic efficiency of DIET was predicted to be directly impacted by the proton influx, unlike the case of interspecies hydrogen transfer. This study further presents a comparative analysis of the factors that determine the preference of DIET or interspecies hydrogen transfer. Specifically, this is demonstrated by analyzing a similar system of syntrophic ethanol oxidation involving *P. carbinolicus* and *G. sulfurreducens*. This revealed that metabolic features such as energy conservation via proton gradient or substrate-level phosphorylation, environmental factors such as proton availability in the local microenvironment, and microbial community dynamics could play a role in deploying DIET or interspecies hydrogen transfer mechanisms during syntrophic associations.

In addition to characterizing the syntrophy at the level of metabolites and electron flow, our integration of transcriptomic and physiological data with
genome-scale models enabled us to describe the interplay between the mechanisms of electron flow. By modeling genetic perturbations that eliminated interspecies hydrogen transfer and additional electron transfer via acetate, we were able to arrive at an interaction motif to resolve the synergistic effect of acetate transfer on DIET. The ability to extract such interaction motifs is of particular importance when studying complex microbial communities where multiple mechanisms of electron exchange are likely to be prevalent. It is possible that other syntrophic consortia will exploit a similar synergy of different electron flow mechanisms in a condition-dependent manner. The knowledge of such unique interaction motifs between modes of electron transfer will also have a significant impact on the design of synthetic syntrophic consortia for several biological engineering applications in the future [193].

Finally, we were able to obtain novel insights into the dynamics of interaction between the partners in the consortium and the mechanisms of how they adapted to the selective pressure of syntrophic growth. Two distinct phases of syntrophic growth characterized by maximal and suboptimal electron transfer were observed. As the electron transfer scenario transitioned from suboptimal to maximal (Phase B to A), the relative abundance of *G. sulfurreducens* increased. This led us to postulate that *G. sulfurreducens* actively drives the association by dominating the interaction dynamics. This inference was found to be consistent with a previous finding that the beneficial mutation that promoted syntrophic association was in *G. sulfurreducens*. In other words, the organism that relies on the electrons transferred (*G. sulfurreducens*) from its partner had the need to actively respond to the selection pressure by undergoing changes at the genomic level. However, the partner that provides electrons (*G. metallireducens*) seemed to respond only at the level of the transcriptome. This type of dynamics could have several evolutionary implications, suggesting that in syntrophic associations, the accepting partner streamlines its metabolic capabilities to enhance the most efficient mode of electron transfer. This could be of particular relevance when characterizing methanogenic consortia. Given that methanogens are typically involved in syntrophic relation-
ships, it is possible that they respond similarly to the availability of the electron donor by streamlining their metabolic capabilities for the most efficient utilization strategy. In fact, a recent study on the functional response of M. maripaludis to syntrophic growth in association with D. vulgaris revealed that, the methanogen uses paralogous genes to adapt to the changing substrate availability when growing syntrophically compared to growing as a monoculture [194]. An approach like the one presented here, that integrates genome resequencing, metatranscriptomics, and genome-scale modeling, would likely allow for a deeper mechanistic insight into the dynamics and adaptation of the syntrophic interaction of M. maripaludis and D. vulgaris.

In summary, this study represents a multi-omic approach to elucidate the electron transfer mechanism and to characterize the metabolic phenotype of a laboratory evolved syntrophic consortium at multiple levels of complexity. This integrated approach demonstrates that energetic limits due to obligate syntrophic associations are reflected in the genome and the transcriptome of the microorganisms, thereby providing a comprehensive alternative framework for detailed mechanistic characterization. Since the approach developed here is modular, it can be easily extended for the characterization of other complex communities. The models being developed in these studies will be important for understanding the functioning of anaerobic microbial communities and could aid in predicting the influence of environmental changes on methane emissions and other aspects of the carbon cycle.

5.6 Methods

5.6.1 Reconstructing a genome-scale model of syntrophy

In order to model the syntrophy between G. metallireducens and G. sulfurreducens, the existing genome-scale metabolic reconstructions of the two species were updated and expanded. A comparative reconciliatory process involving automated reconstructions from MODEL SEED [24], draft reconstruction based on
the updated genome annotation, existing reconstruction, and extensive manual curation was used to update and expand the metabolic model of *G. metallireducens* (Refer to Supporting Material - (GmetReconpaper)). This updated model was used to expand the genome-scale metabolic model of *G. sulfurreducens*. Homologs to *G. sulfurreducens* genes were identified in the *G. metallireducens* genome using a bi-directional Smith-Waterman alignment. A 60% ID cutoff was used to identify the homologs. The most recent updated annotation (AE017180 version 2) of *G. sulfurreducens* was then used to curate the homology matches. A distinct periplasmic compartment was also incorporated into the model. This procedure also included the explicit reconstruction of the extracellular electron transfer pathways as described in Fig. 5.1A.

The two genome-scale reconstructions were integrated into a combined model by extending the compartmentalized framework. A compartment called the shared metabolite pool (SMP) was introduced. This compartment was meant to facilitate the exchange of metabolites between the two genome-scale models as well as their interaction with the extracellular environment. To achieve this, the outer-membrane transporters in each of the genome-scale models were replaced by transporters that facilitated transfer from the respective periplasmic compartments to the SMP. Exchange reactions were added for all metabolites present in the SMP (Fig. 5.11).
Figure 5.11: Detailed schematic of the expansion of the compartmentalized framework adopted for integrating two different genome-scale metabolic models. The outer membrane transporters in each of the metabolic models were replaced by transporters that facilitated transfer from the respective periplasmic compartments to the SMP. Exchange reactions were added for all metabolites present in the SMP.

5.6.2 Modeling direct interspecies electron transfer

DIET was accounted for by introducing a reaction that represents the stoichiometric transfer of electrons between the outer-membrane cytochromes of the two *Geobacter* species. The flux through this reaction was constrained based on the midpoint potential of OmcS, an outer-membrane cytochrome of *G. sulfurreducens*. The maximum electron transfer flux through DIET was calculated by scaling the midpoint potential of OmcS ($E_{omcS}$) using the formula

$$V_{max}^{DIET} = \frac{E_{omcS}}{E_{Fe(III)citrate}} \times Y_{(Fe/EtOH)} \times V_{EtOH}$$

In this equation, $E_{Fe(III)citrate}$ represents the midpoint potential of Fe(III) citrate and $Y_{(Fe/EtOH)}$ represents the ratio of Fe(III) reduction rate to the oxidation rate of the electron donor (ethanol).

5.6.3 Physiological data measurements and growth conditions

All growth experiments of *G. metallireducens* and *G. sulfurreducens* (WT and CS K.O. strains) were performed using strict anaerobic culturing techniques
as previously described [195]. Cultures were incubated in 160 mL serum bottles sealed with butyl rubber stoppers and filled with 100 mL of media. The culture medium was provided with 40 mM fumarate and 20 mM ethanol as the sole electron donor as described [19]. For determination of substrate depletion and metabolite production, samples were withdrawn under strict anaerobic conditions with hypodermic needles and syringes. Samples were passed through 0.2 m Acrodisc filters and were diluted prior to analysis. Changes in ethanol concentration over time were monitored by gas chromatography as previously described [17]. Changes in concentration of volatile fatty acids, such as fumarate, succinate, malate, and acetate were determined by high performance liquid chromatography as previously described [196]. Cell biomass increase over time was determined in four replicate co-cultures (100 mL each) of *G. metallireducens* with *G. sulfurreducens* (WT and CS K.O. strains). Aggregates were harvested at different growth stages by centrifuging for 15 min at 4,000 rpm, washed in isotonic buffer, and pelleted at 4,000 rpm. The pelleted cell suspensions were freeze-dried for 48 hours on a Labconco lyophilizer. Dry mass was weighed prior to re-suspension in 0.5% SDS for determination of total protein content using the bicinchoninic acid method [197] with a protein standard.

### 5.6.4 Simulating syntrophic growth using the genome-scale model

The genome-scale metabolic reconstruction of the syntrophic association was represented in a mathematical format in the form a stoichiometric matrix (*S*). The rows of the matrix represent the metabolites in the network, while the columns represent the reactions. Flux balance analysis (FBA), where the flux through a particular objective reaction is optimized subject to a set of constraints, was used to characterize the system [175]. In this particular case, the objective of the optimization was maximizing the flux through the biomass of both *G. metallireducens* and *G. sulfurreducens*. Typically, the constraints for the optimization problem include upper and lower bounds (ub, lb) for reaction fluxes calculated from thermodynamic and enzyme kinetics. In this simulation, an additional con-
constraint on the flux through DIET was applied based on $V_{\text{max,DIET}}$. Constraints on ethanol, and fumarate uptake rates were applied based on experimentally measured physiological rates. All ions, and water were allowed to enter the network freely. Additionally, the loopless COBRA constraint was applied to eliminate thermodynamically infeasible loops from the solution [186]. Reaction deletion analysis and phenotype phase plane analysis were carried out using the singleRxnDeletion and phenotypePhaseplane functions in the COBRA toolbox [45].

5.6.5 Monte Carlo sampling and differential reaction activity

Monte Carlo sampling was applied on the integrated genome-scale model of the syntrophic consortium to generate a set of uniform, feasible flux distributions. A modified version of the artificially centered hit and run (ACHR) algorithm called gpSampler [45, 3] was implemented to sample both the WT and CS K.O strains. Differential reaction activity was computed based on the sampled flux distributions as done in a previous study [39, 198]. Significance of change was determined at a false discovery rate of 0.05 [199]. Using the gene-protein-reaction associations in the genome-scale model, the genes corresponding to the reactions showing differential activity were obtained. This was compared with the list of differentially expressed genes obtained from transcriptomic profiling using RNA-seq.

5.6.6 Transcriptomic profiling of the aggregates

Total RNA was isolated from wild type and mutant aggregates using the acid phenol method [200, 201]. Residual DNA was removed with a 30 minute DNase I digestion at 37°C (Qiagen) followed by purification with the RNeasy Mini kit (Qiagen). A total of 2.5 µg of total RNA was treated with the Gram-Negative RiboZero kit (Epicentre). Paired end, strand specific RNA sequencing was performed using a variation of the dUTP method outlined in [202, 203] with the following changes. 100 ng of rRNA subtracted RNA was fragmented with RNA fragmentation reagents (Ambion) for 3 minutes at 70°C. First strand synthesis was
primed using random hexamers (Invitrogen). Downstream library construction was performed as previously described [204]. Final libraries were quantified using qPCR (KAPA) and sequenced on an Illumina Genome Analyzer II.

5.6.7 Transcript quantification

The obtained RNA-seq reads were aligned to the concatenated genome sequences of \textit{G. metallireducens} and \textit{G. sulfurreducens} (RefSeq NC_002939) using the short-read aligner Bowtie [205] with two mismatches allowed per read alignment. To estimate transcript abundances, FPKM values were calculated using the tool Cufflinks [206] with appropriate parameters set for the strand-specific library type and upper-quartile normalization. Differential expression analysis was carried out using cuffdiff, with upper-quartile normalization and appropriate parameters set for strand-specific library type. A fold change of greater than 2-fold and false discovery rate cutoff of 0.05 was used to determine significant differential expression between two different conditions.

5.6.8 Functional enrichment analysis

To determine if the differentially expressed genes were enriched for any particular metabolic subsystem in the genome-scale model, a hypergeometric test was performed using the hygecdf function in MATLAB. A \( p \)-value cutoff of 0.05 was used to determine significance.

5.6.9 Resequencing and mutation identification

Nine parallel replicate co-cultures of \textit{G. metallireducens} and \textit{G. sulfurreducens} were sequenced after 780 generations using Illumina sequencing and the resulting mutations were compared to those identified after 100 generations [19]. The concatenated genomes of \textit{G. metallireducens} (NC_007517) and \textit{G. sulfurreducens} (AE017180 version 2) were used as the reference for mutation identification. Illumina reads from the nine different resequencing experiments were iteratively aligned to this reference using MosaikAligner (Stromberg and Marth). The number
of allowed mismatches in the alignment was increased from 0 to 5 in each iteration with the unaligned reads used as the input for the subsequent iteration. The alignments were processed using a custom script written in-house (described in [207]) to obtain the fold coverage at single base-pair resolution. A coverage cutoff of 10X was employed for SNP identification and lack of coverage was used to identify deletion. Furthermore, a particular location was deemed polymorphic in each of the nine experiments, if the observed nucleotide count was greater than twice the count of the actual nucleotide in the reference sequence at that position. Finally, a particular location was considered to be a mutation, if it was found in at least seven of the nine replicates.

5.7 Acknowledgments

The authors would like to acknowledge Muktak Aklujkar for his very valuable comments on the manuscript, Dan Carlo Flores and Mollie Murnane for laboratory assistance with culturing, substrate monitoring, dry weight, and protein determination. The authors also thank Marc Abrams for help with editing the manuscript. This work was funded by Office of Science (BER), U.S. Department of Energy (DE-SC0004485).

Chapter 5 in full has been submitted for publication of the material Nagarajan H, Embree M, Rotaru AE, Shrestha PM, Feist AM, Palsson BO, Lovley DR, Zengler K. Multi-omic characterization and modeling of interspecies electron transfer mechanisms and microbial community dynamics of syntrophic associations. The dissertation author was the primary author of this paper responsible for the research.
Chapter 6

Conclusions and Outlook

This thesis demonstrates how genome-scale systems biology approaches can be applied to characterize electromicrobiological applications such as microbial fuel cell, microbial electrosynthesis and interspecies electron transfer. In the first part of this thesis, next-generation sequencing and multi-omics data integration was applied to characterize the genomic and regulatory aspects of \textit{G. sulfurreducens} growing in a microbial fuel cell setting. Specifically, a \textit{de novo} assembly strategy was developed to elucidate the complete genome of an enhanced electricity producing variant. Further, multi-omics data integration enabled the characterization of sigma factor regulatory network governing electrogenic biofilm at a genome-scale. This revealed insights into how energy metabolism is regulated in \textit{G. sulfurreducens}. Further studies on key transcription factors in this bacterium will provide a comprehensive picture of the interplay between the regulation of energy metabolism and microbial fuel cell performance.

In the second part of this thesis, the metabolism of two electrosynthetic bacteria were characterized at the genome-scale. Specifically, key carbon fixation and energy conservation pathways were analyzed. This part provided key insights into potential novel components involved in electrosynthesis and serves as a framework for genome-scale characterization of the process. In addition, the detailed description of energy conservation mechanisms enabled us to elucidate potential energetic costs involved in extracellular electron transfer. The modeling framework
can be augmented as further biochemical discoveries on the various components of extracellular electron transport are made. This part of the thesis also presented the first genome-scale model of an acetogen. Acetogenic microorganisms have unique metabolic pathways and energy conservation mechanisms that, if understood, could greatly increase options for the design of microorganisms for production of biofuels and enhance the range of biomass materials that could become economically viable feedstock. Hence the models developed in this part are expected to guide strategies for strain design and engineering of autotrophic metabolism either using syngas or under electrosynthetic conditions. This effort could potentially be strengthened by the development of next generation genome-scale models that include biochemical processes such as transcription, translation, protein translocation in addition to metabolism [208]. These models will be of particular relevance for studying aspects of extracellular electron transfer since most of the key components are multi-heme proteins whose synthesis and translocation to the outer membrane are energy intensive. Metabolic models do not account for such costs and a model with expanded scope is expected to have an increased predictive power.

In the third part of this thesis a multi-omics modeling workflow that combines genomic, transcriptomic, and phenotypic data with constraint-based modeling using genome-scale metabolic models to investigate metabolic dynamics and mechanisms of electron flow in syntrophic co-cultures is presented. One of the key aspects of this workflow is the modeling of direct interspecies electron transfer. The models presented in this study will be important for understanding the functioning of anaerobic microbial communities and could aid in predicting the influence of environmental changes on methane emissions and other aspects of the carbon cycle. With this integrated approach, we demonstrate that energetic limits due to obligate syntrophic associations are reflected in the genome and the transcriptome of the microorganisms. This provides a complementary alternative framework that relies on omics data for detailed mechanistic characterization of microbial communities. As stated earlier, the study of microbial interactions will also benefit from the expansion in scope of metabolic models to include other biochemical processes.
The insights learnt from electron transfer capabilities of single species and simple microbial consortia can be extended to characterize complex microbial communities and microbiomes. Such an extension might require application of sequencing technologies to characterize the composition of the microbiome and application of transcriptomic technologies in conjunction with draft genome-scale metabolic models to obtain mechanistic insights into the functioning of the community.
Bibliography


of coupling the complete oxidation of organic compounds to the reduction of iron and other metals. Arch Microbiol 159: 336-44.


