Second Year Progress Report

INTERCELLULAR GENOMICS OF SUBSURFACE MICROBIAL COLONIES

Grant # DE-FG02-05ER25676

Prepared for the

DEPARTMENT OF ENERGY/OFFICE OF SCIENCE

For the Period

February 2006 through February 2007

Submitted by

PI P. Ortoleva
Center for Cell and Virus Theory
Indiana University, Bloomington IN 47405
Tel: (812) 855 2717 email: ortoleva@indiana.edu

Co-PI K. Tuncay
Center for Cell and Virus Theory
Indiana University, Bloomington IN 47405
Tel: (812) 855 2717 email: ktuncay@indiana.edu

Co-PI D. Gannon
Department of Computer Sciences
Indiana University, Bloomington IN 47405
Tel: (812) 855 5184 email: gannon@cs.indiana.edu

Co-PI C. Meile
Department of Marine Sciences
University of Georgia, Athens GA 30602
Tel: (706) 542 6549 email: cmeile@uga.edu

Table of Contents

I Summary

This report summarizes progress in the second year of this project. The objective is to develop methods and software to predict the spatial configuration, properties and temporal evolution of microbial colonies in the subsurface. To accomplish this, we integrate models of intracellular processes, cell-host medium exchange and reaction-transport dynamics on the colony scale. At the conclusion of the project, we aim to have the foundations of a predictive mathematical model and software that captures the three scales of these systems – the intracellular, pore, and colony wide spatial scales.

In the second year of the project, we refined our transcriptional regulatory network discovery (TRND) approach that utilizes gene expression data along with phylogenic similarity and gene ontology analyses and applied it successfully to *E.coli*, human B cells, and *Geobacter sulfurreducens.* We have developed a new Web interface, GeoGen, which is tailored to the reconstruction of microbial TRNs and solely focuses on Geobacter as one of DOE's high priority microbes. Our developments are designed such that the frameworks for the TRND and GeoGen can readily be used for other microbes of interest to the DOE.

In the context of modeling a single bacterium, we are actively pursuing both steady-state and kinetic approaches. The steady-state approach is based on a flux balance that uses maximizing biomass growth rate as its objective, subjected to various biochemical constraints, for the optimal values of reaction rates and uptake/release of metabolites. For the kinetic approach, we use Karyote, a rigorous cell model developed by us for an earlier DOE grant and the DARPA BioSPICE Project.

We are also investigating the interplay between bacterial colonies and environment at both pore and macroscopic scales. The pore scale models use detailed representations for realistic porous media accounting for the distribution of grain size whereas the macroscopic models employ the Darcy-type flow equations and up-scaled advective-diffusive transport equations for chemical species. We are rigorously testing the relationship between these two scales by evaluating macroscopic parameters using the volume averaging methodology applied to pore scale model results.

II TRN Discovery and Analysis

A Overview

A key component of this project is the development and application of tools for TRN discovery. The application of those tools aims at the optimization of the use of microbes in energy production and environmental remediation. In particular, our systems microbiology TRN discovery tools will provide approaches to predict microbial behavior and ultimately to enable the computer-aided design of mutants for prescribed functions that can be performed in an environmentally safe manner. The TRND system is designed to interpret, and ultimately guide, gene expression experiments (Fig. 1). We have developed a robust methodology to use known TRN information as a training set and augment it by discovering new transcription factor (TF)/gene regulatory interactions by integrating a variety of approaches via a Bayesian framework, as discussed in the paragraphs below, in the appendices, and in our publications cited herein.

The TRN we seek to discover is a list of genes for each of which a set of TFs with up/down regulation is provided. This approach also provides the gene-gene regulation network as the genes that encode the components of each TF are also included in our TRNs. We use multiple methodologies to suggest enhanced TRNs. The result of each methodology is weighed proportional to its success rate using the corresponding training set. This approach goes beyond studies that focus on gene-gene networks as it provides more detailed information (such as gene A is up regulated by TF B) that can be tested experimentally and used in medical and biotechnical applications. We demonstrate that methodologies such as gene ontology and phylogenic similarity provide better results when a preliminary set of TF/gene interactions is used instead of a training set of gene-gene data.

Our TRND Web-based system accepts gene expression microarray data on a microbe of interest as input and yields its TRN as output. This TRN can then be used as input to a second, integrated workflow that creates a Fortran-readable, mathematical cell reaction-transport model of transcription/translation/post-translational processes and analyzes the model to determine factors (e.g., extracellular conditions or mutations) that support distinct types of cell behaviors (e.g., intracellular levels of RNA, proteins, and other genomic and proteomic components). TRND, the database, and the methods it uses are described below.

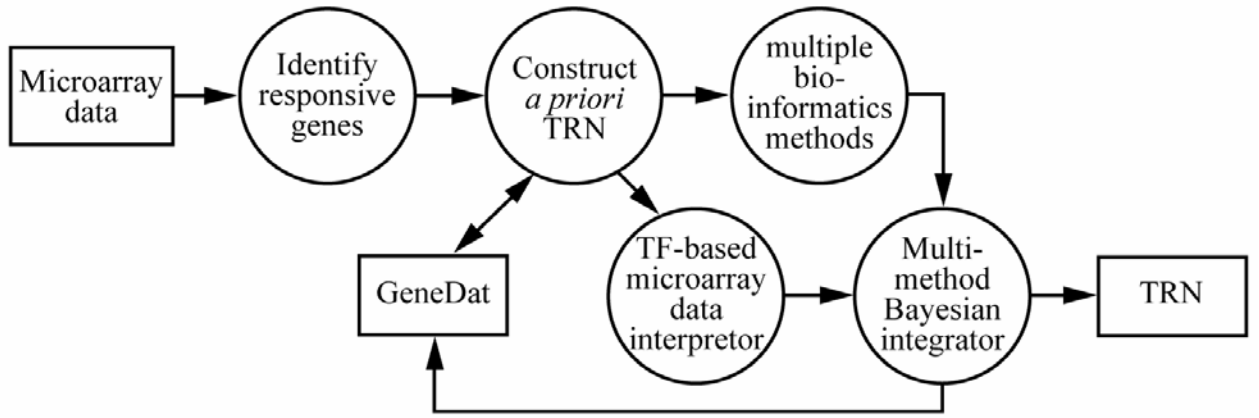


Fig. 1 Responsive genes from a gene-expression experiment initiate a query to extract an *a priori* TRN (training set) from our GenDat database. This preliminary TRN is used by our TF-based microarray interpreters and bioinformatics modules as a training set. The results of the individual modules are integrated via a Bayesian approach to discover TF/gene regulatory interactions. Results and the training set are also made available through our new database GeoGen which was developed to facilitate a higher level of collaboration between experimental and computational groups working on Geobacter species.

B FTF Module

Fast Transcription Factor (FTF) module (Sun et al. 2007; Tuncay et al. 2006; Qu et al. 2007) was initially developed by us for the DARPA BioSPICE project, and was refined, tested, and integrated into TRND last year. The FTF method is aimed at the construction of TF profiles from expression data. Considering the well-known noise in the data, the FTF algorithm uses a blending of the expression profiles of many genes to compute these TF profiles. Errors in a user-supplied preliminary TRN are corrected by our algorithm. In FTF, regulation of many genes by a given TF allows an integration of many gene responses to yield a given TF profile (TF activity versus time or across conditions). Use of the constructed TF profiles overcomes limitations of other methods wherein it is assumed that the profile of a TF's activity is represented by that of its encoding gene's RNA expression.

Network discovery requires many automated trials of possible TF/gene interactions; thus the FTF algorithm was designed to be extremely efficient. The essential equation on which FTF is based was arrived at empirically after extensive numerical experimentation with synthetic expression data and a known TRN, TF activities, and specified noise. The FTF equation takes the form

$$T_n^r - T_n^s = \sum_{i=1}^{N_g} H\left(m_i^r - m_i^s\right) b_{in} \Psi_{in} \quad ,$$

where T_n^r = activity of TF n at condition or time r, m_i^r = microarray response for gene i transcript at condition r, $b_{in} = +1/-1$ for gene i up/down regulated by TF n, $b_{in} = 0$ for no regulation, $H(x) = \pm 1$ for $x > 0$ or $x < 0$, = 0 for x = 0, and $\Psi_{in} = 2^{L_i} / (M_n (2^{L_i} - 1))$, for L_i = number of TFs controlling gene i, and M_n = number of genes TF n regulates as determined by the training set $\underline{\underline{b}}$. If there are N_{ge} experimental times or conditions, then one can write $N_{ge} \times (N_{ge} + 1)/2$ equations for the N_{ge} activities T_n^r, $r = 1,2,...N_{ge}$, for

each of the N_{TF} TFs. TF activities are obtained from the solution of the above equations via a least squares approach to integrate all the expression data in an objective fashion.

Given TF profiles constructed as above, a measure of the reliability of a given TF/gene regulatory interaction is determined by correlating the predicted TF profile with the observed gene's RNA expression microarray response. For example, if a TF upregulates a given gene, then the predicted TF profile and the observed microarray response for that gene are likely to be positively correlated. Such arguments are used to correct supplied TF/gene interactions and also to discover additional TF/gene interactions. Scores for FTF predicted interactions are calculated by taking the linear (Pearson) correlation between the predicted TF activities and gene expression data. A synthetic example that illustrates FTF is provided in Appendix A. Applications to *E.coli* and human B cell data (for which there are extensive gene expression data sets) are summarized in Appendices C and D, respectively.

C Gene Ontology Module

In this TRN construction approach, we use the biological process ontology developed by the Gene Ontology (GO) Consortium (*www.geneontology.com*) and hypothesize that the likelihood that a gene pair is regulated in the same manner increases with the similarity of their GO descriptions. GO analysis was proposed by Wu et al. (2005) who applied it to find functional modules in *E.coli*. Each GO term is placed in a directed acyclic graph. The GO similarity score between two genes is based on the number of shared ancestors. As a gene might be assigned multiple GO terms, we seek the maximum similarity score between all possible combinations.

To discover TF/gene interactions, our innovation is to reformulate the GO approach as follows. For a system of N_{gene} genes, there are $N_{gene} \times (N_{gene} - 1)/2$ gene-gene pairs. In order to score the gene A/TF B interaction, we first seek all genes regulated by TF B in the preliminary TRN or training set. Then we calculate the gene-gene similarity score for the gene of interest with each gene regulated by TF B. We assign the maximum of these scores to the gene A/TF B interaction. Although this appears to be a rough estimation of the TF/gene score, our computational experiments have shown that this score clearly distinguishes the probability distributions of the training and random sets of TF/gene interactions (Tuncay et al. 2006; Sun et al. 2006). We have also used this methodology for *G.sulfurreducens* as shown in Sect. IV.

D Phylogenic Similarity Module

Phylogenic similarity analysis, also proposed by Wu et al. (2005), is based on the hypothesis that two genes, from different but related organisms, with large phylogenic similarity score are likely to be in the same functional operon, regulon or pathway. Our *innovation* compared to previous approaches is the hypothesis that "two genes have high phylogenic similarity score, then they would be regulated in the same manner by the same set of TFs." Based on this hypothesis, we extend the preliminary TRN by calculating phylogenic similarity for gene-gene pairs following the methodology proposed by Wu et al. (2005) (referred to as "likelihood of neighboring profiles" in their work). We have extended the number of genomes used in the analysis from 134 to 229 and used the *E.coli* TRN as the training set, in contrast to the gene-gene pair training set used by Wu et al. (2005). Details of the methodology are provided in Appendix B.

E Multi-Method Integration

Given that the TRNs of interest involve many genes and TFs, they are vast and complex. Thus, any one of the methods cited above will not introduce sufficient information to reconstruct the TRN. To address this challenge, we have developed and tested an algorithm for integrating the information from many TRN construction methods. Our algorithm is as follows. TRN construction method *k* provides a score R^k for every possible TF/gene interaction. An experimentally-verified partial TRN from GenDat is used as the training set to determine $f_{tr}^{k}(R^k)$, the fraction of the known interactions in each of a number of intervals of R^k; similarly $f_{rand}^{k}(R^k)$ is obtained for randomly chosen TF/gene pairs. If the ratio $f_{tr}^{k}(R^k)/f_{rand}^{k}(R^k) >> 1$, interactions with a score R^k are highly likely to be correct. These Bayesian ratios are computed for each method and TF/gene pair. The sum of the $\log_{10}$ of these ratios is taken as a multi-method confidence measure S_{in} for gene *i* and TF *n*:

$$S_{in} = \sum_{k=1}^{N_{meth}} w_k \log_{10}\left(\frac{f_{tr}^{k}(R_{in}^{k})}{f_{rand}^{k}(R_{in}^{k})}\right).$$

Here, N_{meth} is the number of TRN construction methods and w_k is a weighting factor which we presently set to 1, but which could be optimized using a larger training set. Any TF/gene interaction with a sufficiently high log-sum confidence is accepted, resulting in an integrated predicted TRN. If a method fails to have a prediction for a TF/gene pair, it is excluded from the above calculation.

F The Nonlinear Dynamical Systems Analysis Module

Given the TRN for a microbe of interest, the question still remains regarding how one can derive its full biological implication. To address this challenge, we have developed a dynamical systems analysis module, NDS. NDS accepts a TRN, a list of simple or composite TFs, and the encoding genes of the TFs or TF components as input. NDS then automatically creates a Fortran-readable set of transcriptional/translational/post-translational reaction-transport, cell-model equations. In turn, the latter is analyzed to identify conditions (e.g., the extracellular medium or mutations) under which various distinct motifs of microbial behaviors are manifested (e.g., specialization in one versus another carbon or oxidation source). We have tested this module using data on human epithelial cells; it was shown that these cells can display dramatic transition as extracellular conditions pass through critical values. Also, it was shown that hysteretic behaviors can be manifest – i.e., over a given range of conditions a cell can display two or more distinct behaviors. Which state is occupied is determined by the initial state of the cell, described by levels of RNA and protein population. In this project, this will allow us to identify regions of extracellular conditions (e.g., temperature, O_2 concentration, nutrient availability, and solid phase composition) that optimize remediation or energy production.

G The TRND Network Discovery Web-Enabled System

Our objective in establishing TRND was to enable a semi-automated workflow that integrates the methods reviewed above in a Web-based format. The input to TRND is a user's gene expression data. Users can also add/edit TRN information and have access to a

database of available TRN information (see Sect. III for details). TRND interfaces allow for a range of microarray data formats. To start the computation/analysis, TRND extracts a preliminary TRN from our GenDat database (see Sect. III). Then the user is offered the choice of methods to use in reconstructing the TRN. Extensive editing functionalities are implemented to allow the user to upgrade the preliminary TRN. A visual tutorial is provided on TRND use that is downloadable from our *sysbio.indiana.edu* Web portal.

H TRND Testing and Validation

In the course of developing the individual modules and the TRND system described above, we have carried out studies to test and validate our techniques and software, gain experience with picking confidence cutoffs, reaction-transport and other parameters, and contribute scientific results. The choice of test/validation biological systems was dictated by (1) the availability of a large set of high-quality, gene-expression, microarray data and (2) the availability of an extensive training set of experimentally verified TF/gene regulatory and process-rate information. Given this validation of the TRND system (see the Appendices), we are applying it to *Geobacter* (see Sect. IV). The first test case was *E.coli* as it is believed to have the most well-understood TRN (Appendix C). The second test case was the human B cell as there is abundant high-quality data (336 data sets obtained on the same microarray platform) and adequate TF/gene regulatory information (9,500 TF/gene interactions) in GenDat (see Sect. III). These test cases helped us explore the advantages and weaknesses of the bioinformatics modules explained above.

III The GenDat, TRND Results, and GeoGen Databases

A Overview

Three databases have been created as part of our efforts in cellular regulatory network discovery. (1) GenDat is a database of TRN information for multiple species and cell lines. It is designed to provide training sets of experimentally verified TF/gene regulatory interaction information for use with our TRN construction system, TRND. GenDat enables TRND to provide a seamless, semi-automated workflow that yields a TRN given the gene expression data. (2) The TRND Results database provides TRND-predicted regulatory networks to the research community. (3) GeoGen is designed to integrate transcriptional regulatory information from across the *Geobacter* project; therefore, the objective is to provide a user-friendly interface to a database of TRN (and ultimately other cell regulatory) information on *Geobacter* species. GeoGen has a subset of the features in GenDat and TRND Results but is reorganized for greater ease of use, and it allows designated project collaborators to edit the database. GeoGen is designed to include computational results from multiple research groups and allows access to binding site information, a feature that doesn't exist in TRND Results and GenDat databases, and large datasets of predictions can be input by our systems manager. In summary, GeoGen provides an integration of efforts on *Geobacter* to facilitate the activities of both experimental and computational teams.

B GenDat and TRND Results

GenDat (*sysbio.indiana.edu/trnd/*) is our MySQL database of experimentally verified TRN information. This database holds gene, TF, and TF/gene interaction information. It

archives aliases and is drawn from a variety of sources. Associated tables contain sets of predicted TF/gene interactions. GenDat provides the training sets for our TRN discovery workflow, TRND (see Sect. II). In contrast to other databases, it (1) provides up/down regulatory interactions explicitly (i.e., the user is referred to the citations), (2) contains complete entries for specific pathways of interest, (3) can be downloaded to form a large training set, and (4) provides the TRN information in a format that can readily be used in an automated TRN discovery workflow, as in Fig. 1.

TRND Results is a database of transcriptional regulatory information predicted by of our TRND system. Users may access predictions on a microbe or cell type of interest. The information is organized according to the multiple methodologies used, for each TF/gene interaction and a confidence score is provided. An integrated multi-method confidence score is also provided so that users can choose the level of confidence they wish to adopt to screen out less reliable predictions. Individual methods for which predictions are provided include gene ontology (GO), phylogenic similarity, FTF and correlation analysis.

C GeoGen: Addressing Geobacter Project Integration

GeoGen is an interface and database designed to coordinate computational predictions and experimentally derived results on *Geobacter*. GeoGen facilitates the integration of results from diverse sources and allows the comparison between the results of the various approaches to network discovery. Predictions generated through computational methods include those from our TRND workflow. Additional experimentally verified information from the literature and computer-generated results will be added by authorized users on a continuous basis.

The entry and editing of data requires a deeper knowledge of the experimental data quality. For example, if incorrect information from GeoGen is used as part of a training set in a subsequent TRN construction, then results would be contaminated. To avoid the propagation of erroneous information from one source to another, not all collaborators should be able to edit all types of data. The present policy on data entry and editing is to allow selected collaborators to enter individual TF/gene interactions through the GeoGen Web interface. Large datasets (e.g., as generated computationally or from a survey of the literature) will be entered by us via our parsers, which will also help to avoid large-scale contamination of the database due to misunderstandings about data formats.

All *Geobacter* collaborators can extract data from GeoGen via a Web interface. Users may extract various types of information:

(1) all TFs regulating one gene,
(2) all genes regulated by one TF,
(3) the TRN as a downloadable Excel spreadsheet.
(4) TRN information selected by method, multiple methods, and information source.

GeoGen is designed to contain a spectrum of additional information:

- binding sites, including multiple sites on a given gene for a specific TF
- GSU numbers
- nature of the regulation (up/down) for a given TF/gene pair
- source of the information
- a measure of quality for the information

All these options have been fully implemented and we are in the process of adding a greater volume of data. As the needs of the *Geobacter* project arise, we shall add new types of information and ways of selectively harvesting the information.

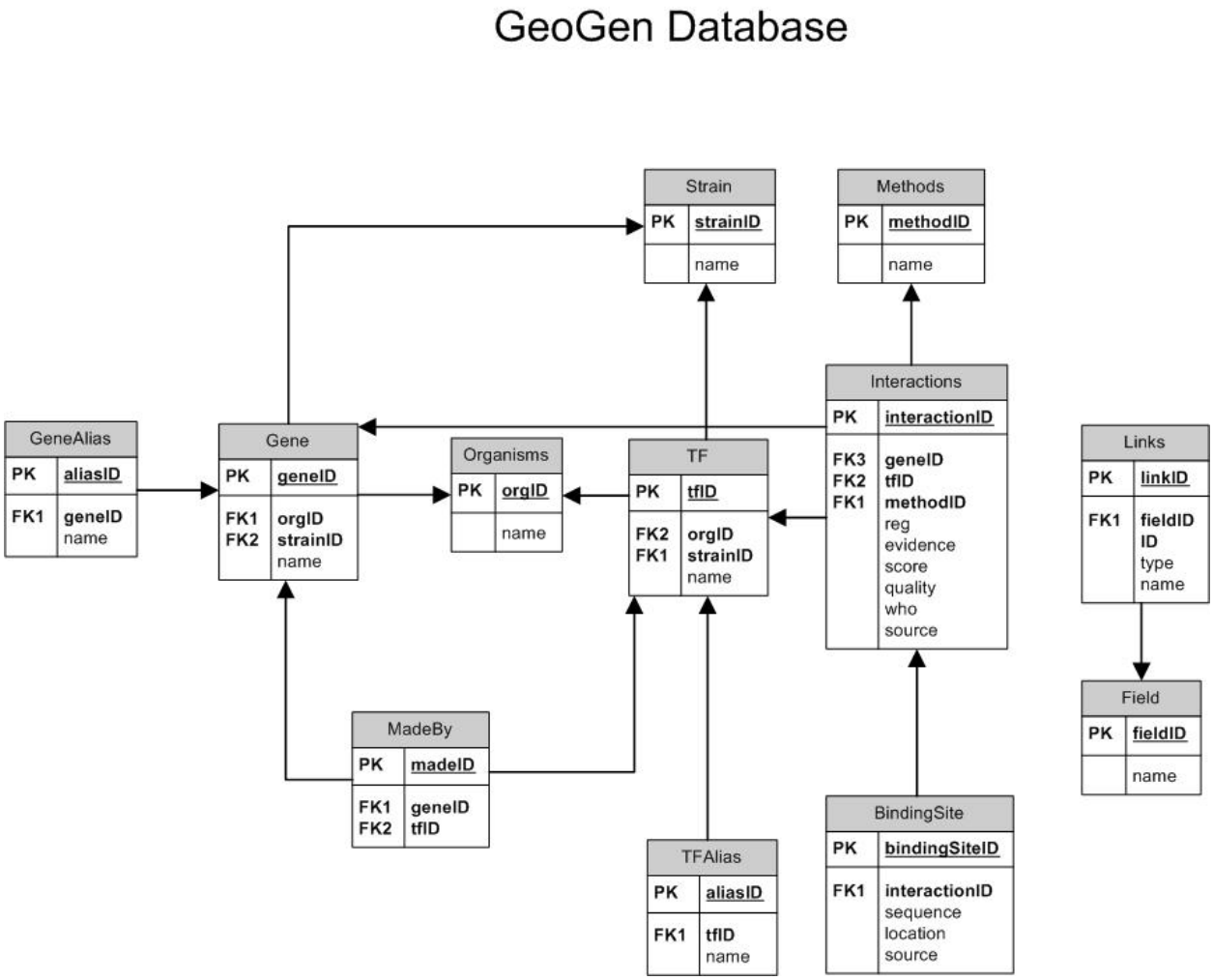


Fig. 2 GeoGen organization diagram. The database allows easy access to TF/gene interaction information from diverse sources (both experimental and computational).

GeoGen is available at *http://sysbio.indiana.edu/geogen*. We have tested GeoGen on Explorer 7 (Windows XP), Safari (OS X) and FireFox (all OSs). Since the objective of GeoGen is to provide an integrative function for the *Geobacter* project, we plan to evolve it in ways that are identified by a consensus of *Geobacter* project users.

The structure of GeoGen is shown schematically in Fig. 2. GeoGen is written in MySQL. Using database conventional definitions, symbols in the organization chart are as follows:

PK=primary key
FK1=foreign key one
FK2=foreign key two
FK3=foreign key three

GeoGen presently contains TF/gene interactions from sources indicated by (•) while those with (*) will be added once we receive feedback on this report:

- homology-based TRN information from our analysis
- TRND multi-method predictions
- Information from the literature[1,2,3,4,5].

* Computationally derived files of possible TF/gene interactions obtained by other research groups.

As shown in the screenshot in Fig. 3, access to GeoGen requires a user name and a password. Authorized users can add/revise TRN data through a Web interface. It is the user's responsibility to ensure that revisions and additions are correct. A user can only modify or delete information he/she entered. If data is owned by another user, they can request that it be deleted or changed. Such requests, like many other features of GeoGen, are initiated by a click of the mouse. These features are designed both for convenience and for the minimization of data-entry errors.

To protect the information in GenDat, several measures will be taken. We shall keep daily backups of the database in case of hardware failure, human error, or input of incorrect data. The backups are stored on our local RAID system and on Indiana University's mass storage facility, making loss of information essentially impossible.

---

[1] "Reconstruction of regulatory and metabolic pathways in metal-reducing δ-proteobacteria."
Dmitry A. Rodionov, Inna Dubchak, Adam Arkin, Eric Alm and Mikhail S. Gelfand
Genome Biology 2004, 5:R90 doi:10.1186/gb-2004-5-11-r90

[2] "Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family *Geobacter*aceae."
B. Yan, B.A Methe, D.R. Lovley, J. Krushkal
J. Theor. Biol. 230(1):133-44 (2004)

[3] "*Geobacter* sulfurreducens has two autoregulated lexA genes whose products do not bind the recA promoter: differing responses of lexA and recA to DNA damage."
M. Jara, C. Núñez, S. Campoy, A.R. Fernández de Henestrosa, D.R. Lovley, J. Barbé
J Bacteriol. 185:2493-2502. (2003)

[4] "Heat-Shock Sigma Factor RpoH from *Geobacter* sulfurreducens."
T. Ueki, D.R. Lovley
accepted for publication in *Microbiology* (2006)

[5] "DNA Microarray and Proteomic Analyses of the RpoS Regulon in *Geobacter* sulfurreducens."
C. Núñez, A. Esteve-Núñez, C. Giometti, S. Tollaksen, T. Khare, W. Lin, D.R. Lovley, B.A. Methé
J. Bacteriol. Apr;188(8):2792-800 (2006)

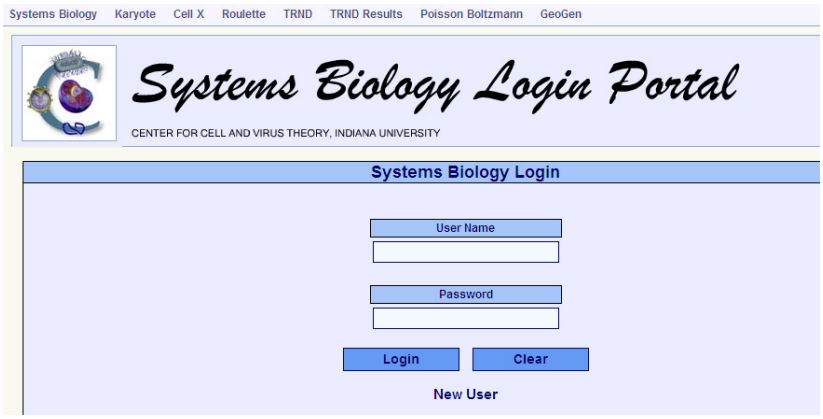


Fig. 3 GeoGen is a secure Web site. We create free accounts for interested researchers.

Once the user is logged on, four options are available as shown in the screenshot of Fig. 4. A user can search for a given TF or gene, view all TFs or genes, or add/revise data if authorized. As shown in Fig. 5, the search option leads to a menu where the user can decide to search for a gene, a TF, or, alternatively, view all regulatory information from a user-specified set of methods/sources. Users can enter a gene name or GSU number to search for a gene or a TF (Fig. 6) and the corresponding experimentally verified and computationally predicted regulatory network. New method/user categories can easily be added by our system administrator (contact M. Trelinsky at mtrelins@indiana.edu or email ortoleva@indiana.edu).

Fig. 4 Screenshot showing that a user can search by gene or a TF, view a list of TFs and genes, or add/revise data.

Fig. 5 Search option leads to a in which a user can choose to search by a gene or TF. Alternatively, all available regulatory information can be displayed.

Fig. 6 User can enter gene name or GSU number (recommended) to search for a gene or TF.

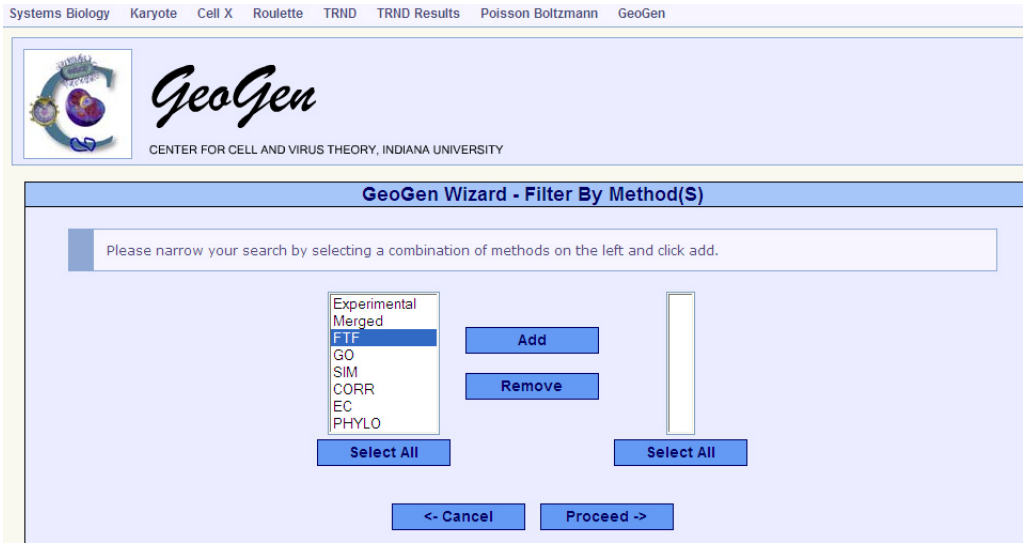


Fig. 7 Since GeoGen stores data from multiple sources (experimental or computational results of multiple research groups), a user can choose a combination of sources/methodologies available.

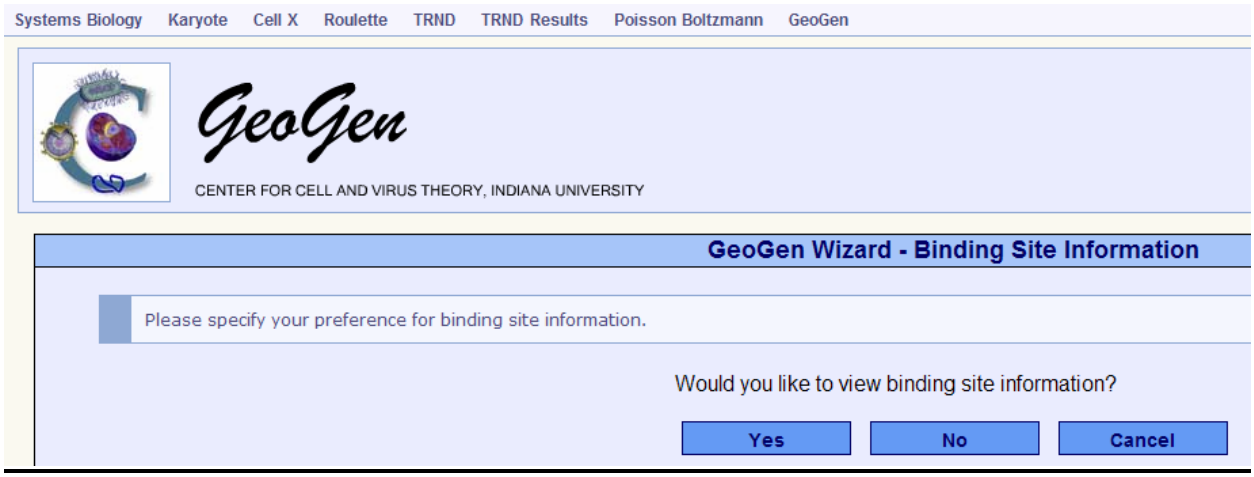


Fig. 8 Binding site information-viewing preference page

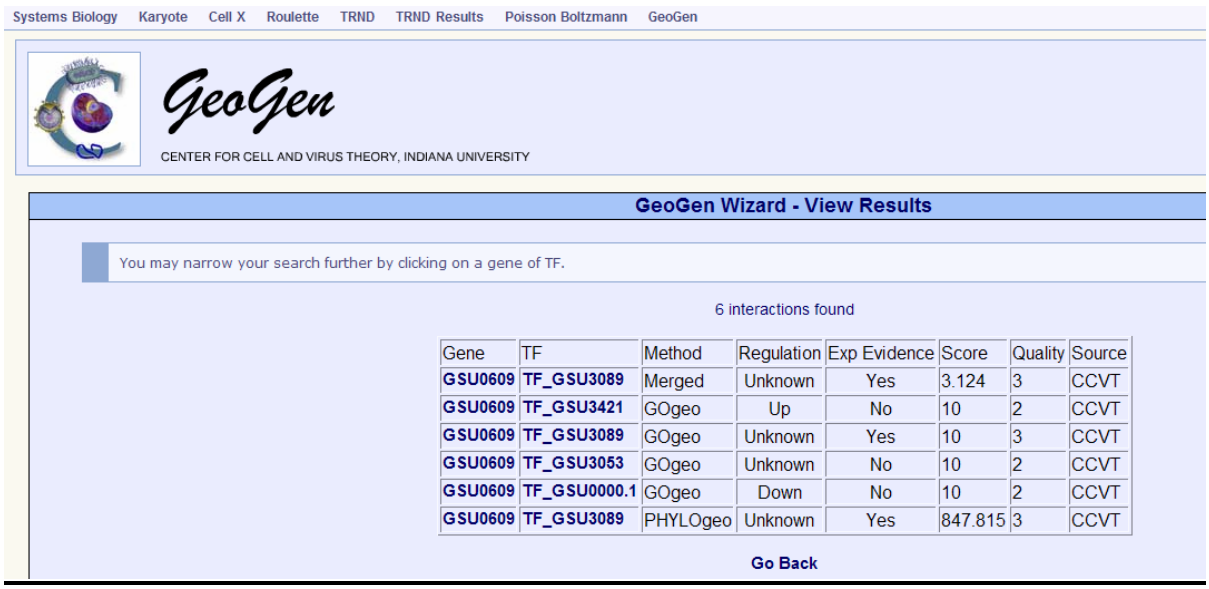


Fig. 9 TF/gene interactions resulting from a user's query

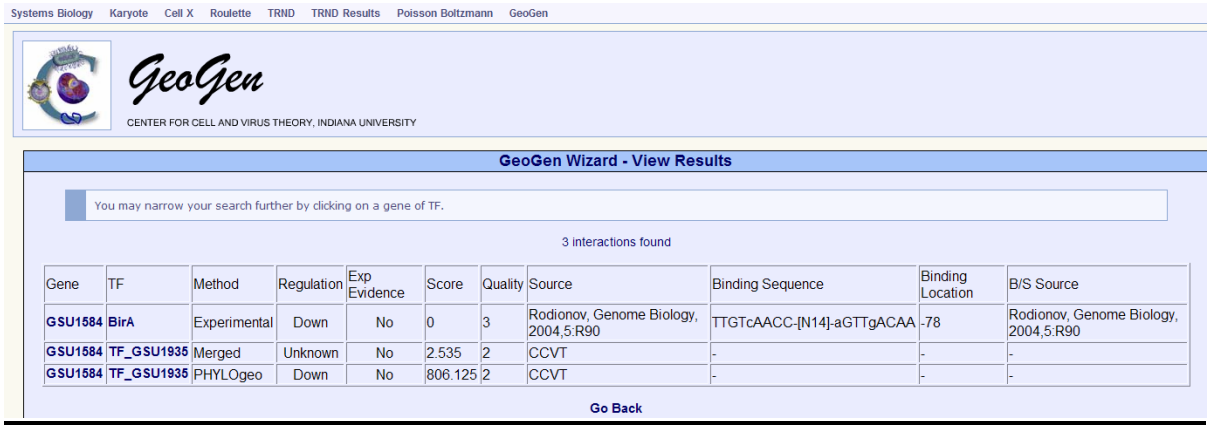


Fig. 10 View of TF-binding information for user-specified gene and TF

D Identified Caveats in the Discovery of TRNs

Two main issues hampering TRN discovery are data quality and translational and post-translational modifications of TF-gene expression profiles.

Data Quality: if a TF/gene interaction is predicted as a result of a single expression experiment comparing wild and mutated strains, the quality of this information may not be very high due to omnipresent microarray data uncertainty and the likelihood that the TRN of the wildtype itself is not well understood for a given gene. Thus, if TF T up-regulated gene G in the wildtype under the conditions of the reference experiment, a relatively lower expression level in the mutant could be misinterpreted as a down regulation of G by TF T in the mutant if only the expression ratio is used in the analysis. Similarly, the quality of computational predictions has a large degree of variation. Hence, most computationally-generated TF/gene interactions are accompanied by a score for the method. Its quality can be assessed via a Bayesian approach wherein the probability that an interaction with a given score in the training set is compared with that for a TF/gene pair chosen at random. Such Bayesian measures are provided with all TF/gene interactions predicted by our group. For this reason, we introduced an integer data field that indicates data quality. Hence, users can not only choose a combination of sources (experimental and computational), but also select a data quality cutoff they wish to impose. Quality factors we use presently are temporary and we seek feedback from collaborators to have a consensus for quality. For the computer-generated predictions, we suggest using a quality scale based on Bayesian ratios.

(Post-)Translational Modifications: It is a common misconception that TF activity can be represented by the expression of the TF-encoding gene. However, translational and post-translational processes can break the correlation between TF and gene expression profiles. Thus, the common notion of a gene-gene regulatory network should be viewed with caution. This leads to the potential misuse of data from the literature. A statement in the literature that gene A is regulated by TF X may be missing the fact that gene A is regulated by TF Y which was over-expressed due to an enhanced transcription of gene B that encodes TF X – i.e., the regulation of gene A by TF X is actually indirect. Mixing direct and indirect regulatory information in any TRN construction work must be done with great care and can lead to erroneous conclusions. We shall add a new field to discriminate between direct and indirect regulation if there is an expression of interest.

E Upcoming GeoGen Upgrades

A number of features that can readily be added to GeoGen will be included after receiving feedback on the present version. Examples for discussion include the following:

- options to include multiple *Geobacter* strains and to merge TRNs from different strains according to a user-controlled algorithm
- external links to provide further information for genes and TFs
- Excel headers showing the sources of the data
- greater filtering selection of data sources, e.g., by user method and confidence
- statistics and graphics on the TRN assembled according to the user-specified filter
- spider Web-like network images for the user-assembled TRN

- automated training set construction for TRND input to generate an augmented TRN based on user-supplied gene expression data
- advanced network analysis modules that can identify subnetworks that lead to dramatic behavioral transitions due to biochemical feedback (e.g., associated with the existence of multiple steady states for the same microenvironment, oscillatory states for time-independent microenvironment, and the tendency toward asymmetric division).

IV *G.sulfurreducens* TRN Construction

We are attempting to assemble a TRN for *G.sulfurreducens* of broad enough scope to enable computer-aided design of DOE-relevant systems. Progress to date and data sources are as follows.

A Homolog-Based TRN Training Set

A first-pass *G.sulfurreducens* TRN was created by identifying genes which are homologous to those in *E.coli* and *B.subtillis*. For example, suppose a TF up-regulates a given operon in *E.coli* in which genes g1, g2 and g3 are located. If there are homologs of the TF and these genes (in a single operon) in *G.sulfurreducens*, and all genes are transcribed in the same direction, then we assume that the same TF/gene interactions are likely to occur in *G.sulfurreducens*. In Table 1, we present the list of TFs that were common in *E.coli*, *B.subtilis*, and *G.sulfurreducens.* With this, we produced a preliminary TRN of 277 genes, 30 TFs, and 518 interactions (Table 2). Sources of the interaction information were EcoCyc for *E.coli* and http://dbtbs.hgc.jp for *B.subtilis*.

TFs from *E.coli*		**TFs from *B.subtilis***	
DnaA	GSU0000.1	DnaA	GSU0000.1
HyfR	GSU0359	HrcA	GSU0031
ZraR-P	GSU0372	Spo0A	GSU1037
IclR	GSU0514	PhoP	GSU1102
NtrC-P	GSU1003	PyrR	GSU1270
PhoB-P	GSU1102	AcoR	GSU1320
FhlA	GSU1129	PerR	GSU1379
NarL	GSU1293	SigF	GSU1525
Fur	GSU1379	SigL	GSU1887
LexA	GSU1617	BirA	GSU1935
NagC	GSU1702	BkdR	GSU2915
KdpE-P	GSU2484	SigD	GSU3053
IscR	GSU2571	SigA	GSU3089
CynR	GSU2787		
CusR	GSU2946		
ModE	GSU2964		
NikR	GSU2980		
ArcA	GSU3118		
NarP	GSU3229		
Fnr	GSU3421		

Table 1 TFs common in *G.sulfurreducens* and *E.coli* or *B.subtilis.*

TF_GSU0000.1	5	TF_GSU1702	2
TF_GSU0031	2	TF_GSU1887	7
TF_GSU0359	13	TF_GSU1935	2
TF_GSU0372	22	TF_GSU2484	9
TF_GSU0514	2	TF_GSU2571	4
TF_GSU1003	69	TF_GSU2787	5
TF_GSU1037	4	TF_GSU2915	3
TF_GSU1102	12	TF_GSU2946	8
TF_GSU1129	10	TF_GSU2964	22
TF_GSU1270	6	TF_GSU2980	5
TF_GSU1293	2	TF_GSU3053	39
TF_GSU1320	3	TF_GSU3089	89
TF_GSU1379	7	TF_GSU3118	51
TF_GSU1525	10	TF_GSU3229	7
TF_GSU1617	3	TF_GSU3421	95

Table 2 Number of genes regulated by each TF in the preliminary TRN.

All *G.sulfurreducens* genes have been entered into the GenDat and GeoGen databases, with annotations from TIGR. The TFs and interactions in the preliminary network obtained from homology have also been entered. In GenDat, the TFs are labeled as belonging to organism "*Geobacter homolog*" to differentiate them from experimentally verified TFs and interactions belonging to "*G.sulfurreducens*" (which are entered continuously as they become available).

B GO and Phylogenetic Similarity Analyses

A GO analysis (as described in Sect. II) was carried out, producing similarity scores for every possible gene-gene pair. This scoring indicates how closely genes are related in the gene ontology tree. These scores were then used with the preliminary homolog-based network (noted above) to obtain GO scores for all possible TF/gene pairs. A summary of the results is shown in Fig. 11 while the detailed TRN is in GeoGen. Application of our GO approach to *E.coli* shows that it provides TF/gene interactions with a high level of confidence.

We also performed a phylogenic similarity analysis based on the hypothesis that two genes with high phylogenic similarity score (they exist in a similar set of bacteria, in similar locations), then they would be regulated in the same manner by the same set of TFs. Our results for *E.coli* support this hypothesis so that, as with GO, we felt confident that it could be used to augment our preliminary *G.sulfurreducens* TRN. To calculate the phylogenic similarity scores, we first constructed a vector for each gene, the dimension of the vector being the number of genomes used in the analysis (229 as of December 2006). Then we use BLASTP to identify orthologous genes of a target genome in the reference genomes. If there is an orthologous gene in the i^{th} genome, then the i^{th} entry in this vector is assigned the order of the orthologous gene in the i^{th} genome. If an orthologous gene does not exist in the i^{th} genome, then this entry is taken to be zero. Once such a vector for each gene is constructed, we compute a phylogenic similarity measure for each gene pair using the expression provided in Appendix B. In Fig. 12, we show that TF/gene interactions in the preliminary network have significantly higher scores than a random TF/gene score. In

order to test whether GO and phylogenetic scores are correlated, for each GO score (between 2 and 13), we calculated the percentage of TF/gene pairs that scored higher than 500 in the phylogenetic similarity analysis. Fig. 13 shows that as GO scores increase, the probability of high phylogenetic scores increase as well.

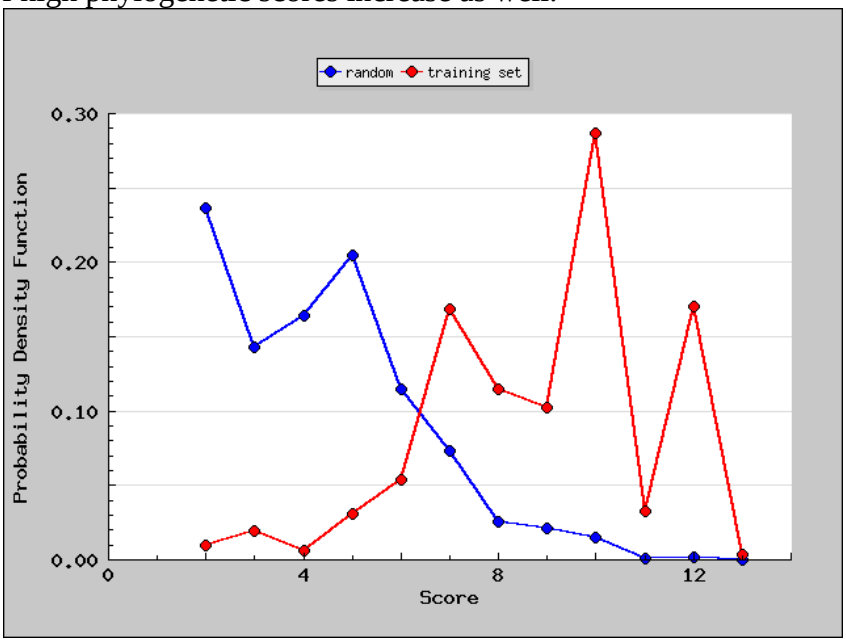


Fig. 11 Probability distributions for the GO scores for the TF/gene interactions in the preliminary TRN and all possible TF/gene pairs. These results suggest that, as also shown for *E.coli* and human B cell the hypothesis that genes that share more ancestors in the gene ontology tree are more likely to be regulated by the same set of TFs.

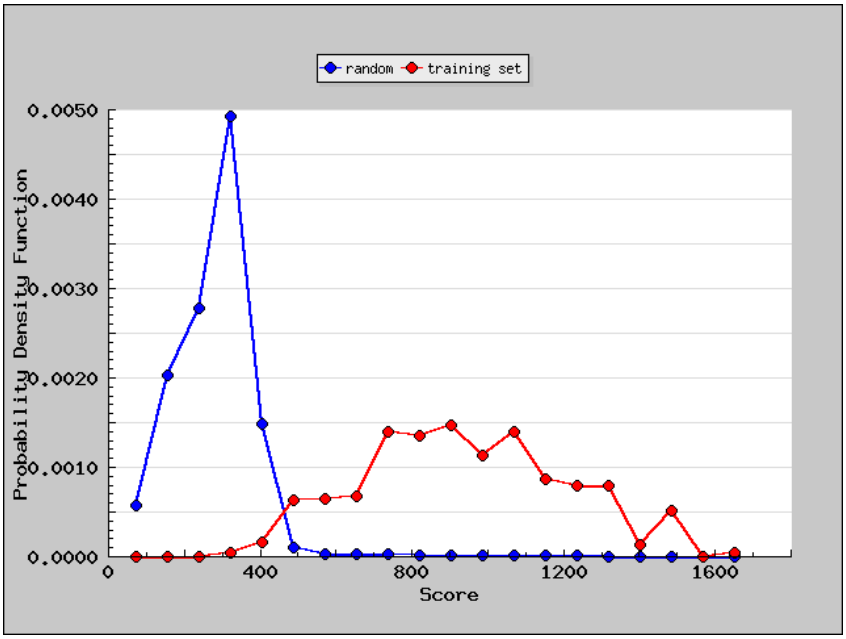


Fig. 12 Probability distributions for the phylogenic similarity scores for the TF/gene interactions in the preliminary TRN and all possible TF/gene pairs. As in Fig. 11, a statistically significant difference between the probability density functions are observed.

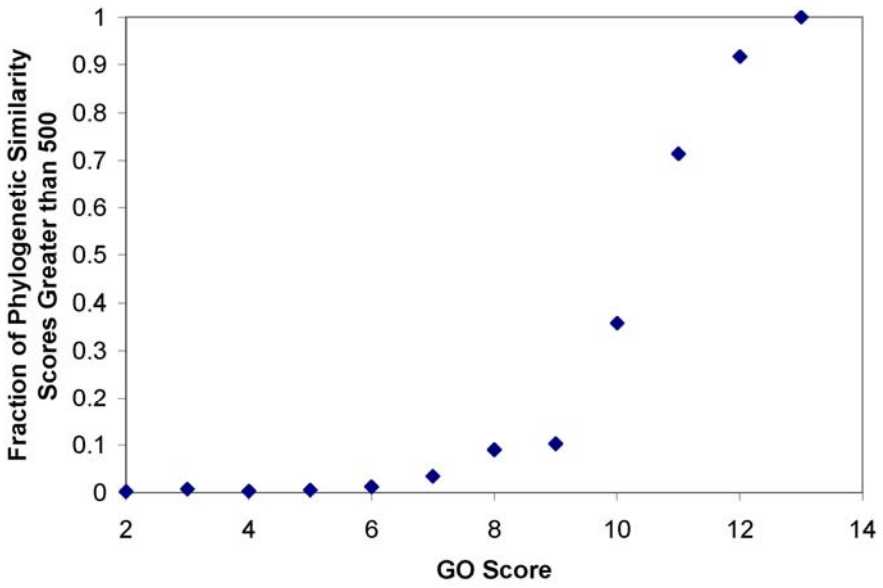


Fig. 13 Among the set of TF/gene interactions with a given GO score, we calculated the fraction of TF/gene interactions that scored higher than 500. The graph shows that high GO scores imply high phylogenetic scores.

C Assembling and Reformatting Available Expression Data for G.sulfurreducens

Information on the experiments listed on the *Geobacter* project Web site has been collected in an Excel spreadsheet and divided into various groups based on the control strain and experimental conditions. We have downloaded gene expression data for the experiments in the largest group (Group 1 with 20 experiments). The data has been reviewed critically and processed to prepare it for input to our TRND network construction system (*sysbio.indiana.edu*). This has involved writing scripts for a number of file formats to determine which genes had bad data versus those that were not differentially expressed; we also combined that information with the expression ratios for differentially expressed genes. Finally, a single file containing expression ratios for all *G.sulfurreducens* genes in the Group 1 experiments was constructed. 18 sets of expression data that were obtained using the same control cells were identified. Experiment numbers, source and condition are provided in Table 3.

If Significance Analysis of Microarrays (SAM) related files were available, we used the SAM analysis to create the list of differentially expressed genes and the associated expression ratios. We only kept genes with at least two reliable data in biological replicate experiments. If Linear Models of Microarray Data (Limma) files were available, we used the Limma analysis tool and set the cutoff p-value at 0.05 to calculate the differentially expressed genes. If neither SAM nor Limma files were available, we used the provided list of differentially expressed genes which may have been derived from a table in a paper. The final number of genes was 3,537. However, the majority of the genes were only expressed in a small fraction of experiments.

D Inconsistencies in the Microarray Data

Although descriptions for experiments 0035-GSUL and 0063-GSUL (Table 3 below) seem identical, our review of the expression data showed large variations between University of

Massachusetts and TIGR platforms. We also observed a significant variation between two experiments under identical conditions from the same laboratory (0056-GSUL and 0076-GSUL with amplified RNA). We removed 0076-GSUL from our analysis as suggested by University of Massachusetts researchers. The number of experiments was significantly fewer than that we used to demonstrate TRND on *E.coli* (65 experiments) and human B cells (336 experiments). First we made an attempt to include all genes, regardless of the number of experiments in which they were differentially expressed. The probability distributions of our FTF score for TF/gene interactions in the random and training sets were indistinguishable, in sharp contrast with our results using FTF on *E.coli* or human B cells. This suggests that either the data or this all-gene approach yields no TRN information. As our Bayesian multi-method integration approach uses the ratio of these probability distributions (see Sect. II: *Multi-Method Integration*), this result means that useful information cannot even be obtained from expression data in this case even when integrated with GO or other method.

E Microarray Information Content Improved by Gene Elimination

To address the aforementioned microarray data difficulties, we decided to only include genes with greater than 5 data points/conditions to avoid contamination of predictions with unreliable data correlations. Fig. 14 shows that 10% of the genes were not differentially expressed in any of the experiments, 18%, 24%, and 20% were expressed in one, two and three experiments, respectively. Analysis of the data showed that only 241 genes were differentially expressed in more than 5 experiments. We applied FTF to this limited ("significant") data. In this case, a clear difference between probability distributions for random and training sets was observed (Fig. 15). This result shows that our methodology for analyzing expression data applies to *G.sulfurreducens* and predictions for a greater number of genes and TFs will be made as the number of expression datasets increases.

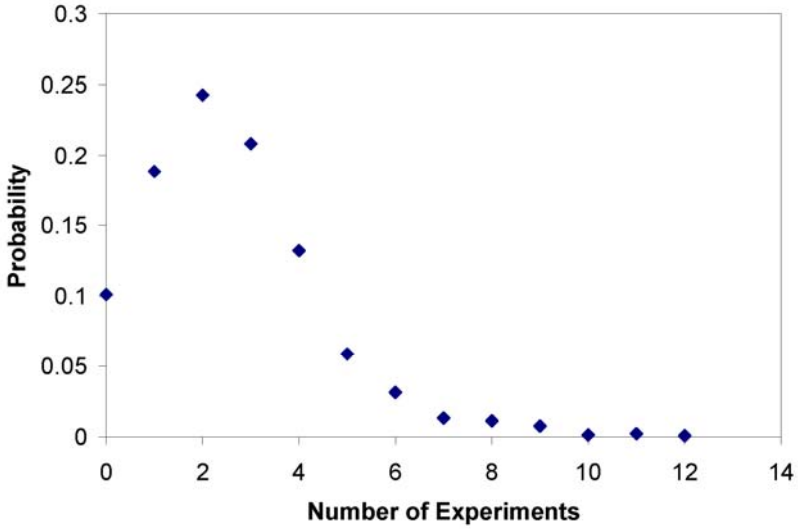


Fig. 14 Majority of genes (93%) were differentially expressed in less than 6 experiments. Therefore they had to be excluded from the expression analysis. However, they were included in GO and phylogenetic similarity analysis.

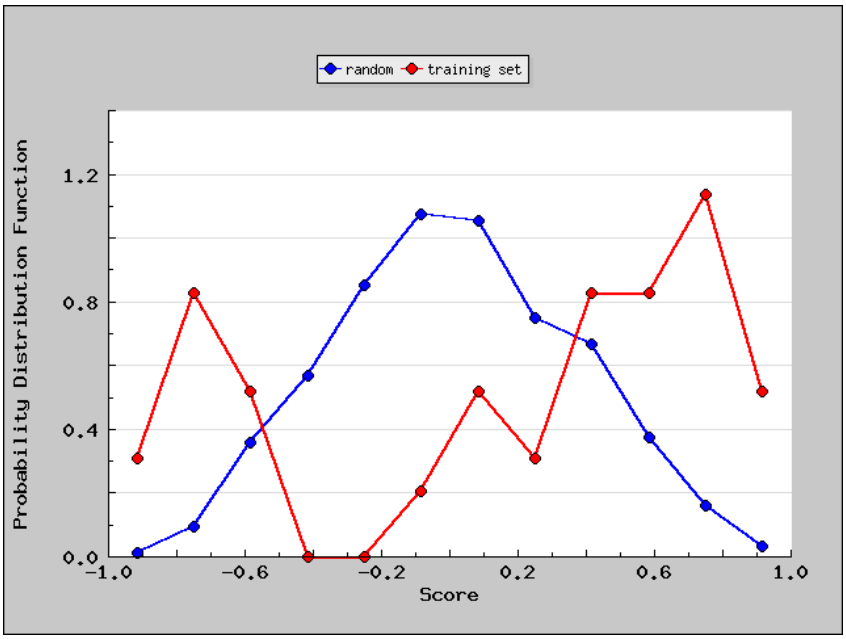


Fig. 15 Probability distributions for FTF scores for the TF/gene interactions in the preliminary TRN and all possible TF/gene pairs. This result was obtained using only 241 genes which were differentially expressed in 6 or more experiments. These results, although obtained with a very limited number of genes and TF/gene interactions, are similar to those obtained for *E.coli* and B cells (Appendices C and D). Therefore, we anticipate that as number of expression data sets increase, we will obtain more reliable results for a growing number of genes.

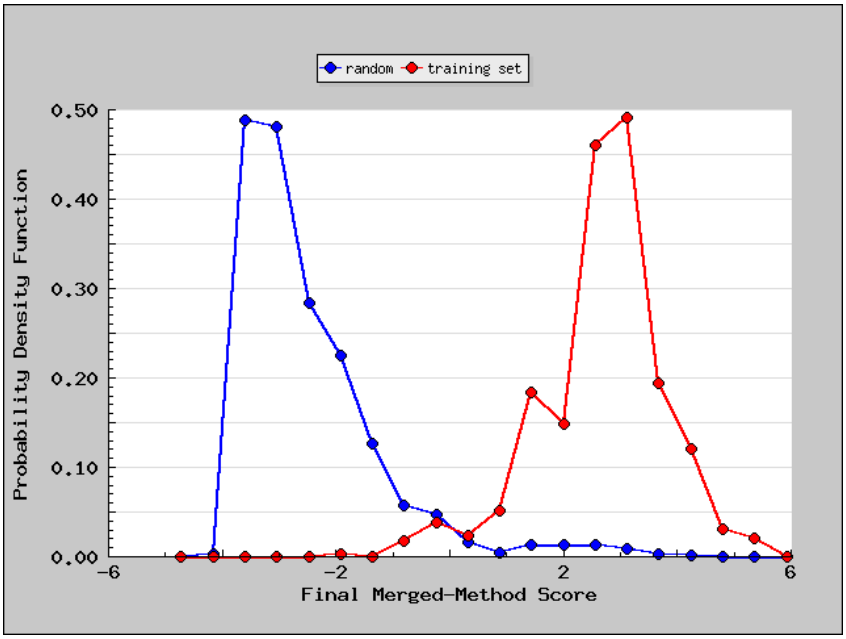


Fig. 16 Probability distributions for the final scores for the TF/gene interactions in the preliminary TRN and all possible TF/gene pairs. When we apply 2.0 as the threshold 785 TF/gene interactions are predicted. The results are available at *http://sysbio.indiana.edu/geogen.*

F Multi-Method Integration for G.sulfurreducens

The GO, phylogenic similarity and FTF results were used with the homology-based, first-pass training set to obtain a predicted TRN via our multi-method integration approach. This result, along with the individual scores that were used for the integration, is posted at our GeoGen Web page (*http://sysbio.indiana.edu/geogen*). Fig. 16 shows a clear distinction between the random and training sets. Scores for 84,435 TF/gene pairs were calculated. Only 2,548 TF/gene pairs were assigned scores for the three methodologies due to the limited availability of expression data. The highest scoring 169 TF/gene interactions are not in the training set provided in Appendix E. While a computational method can generate thousands of predictions, only by accompanying each of them with a score/quality measure can one know the subset that is credible. Thus, we conclude:

1. Our approach has great promise for delivering a *G.sulfurreducens* TRN, and
2. all TRN information in GeoGen should, we suggest, be accompanied with a score and a Bayesian ratio graph so that its credibility can be evaluated.

Experiment	Source	Experimental Strain/Condition	Notes
0006-GSUL	C.E. Nunez	rpoS (GSU1525) knockout	
0010-GSUL	L. DiDonato	relA (GSU2236) knockout	
0015-GSUL	G. Reguera	rpoE (GSU0721) knockout	
0016-GSUL	G. Reguera	Cells from a biofilm grown on Fe(III)oxide-coated surfaces	Control cells were planktonic (in solution)
0021-GSUL	B. Methe, K.P. Nevin	Medium lacking ammonia so cells have to fix nitrogen in order to grow	Done at TIGR.
0022-GSUL	B. Methe, K.P. Nevin	Fe(III) citrate used as the electron acceptor	
0023-GSUL	K. Juarez-Lopez	flp-1(GSU3421) knockout	
0024-GSUL	K. Juarez-Lopez	flp-2(GSU1992) knockout	
0025-GSUL*	A. Esteve-Nunez	Fumarate-limited growth	Original ratios were inverted to put the desired condition in the denominator
0027-GSUL	M.V. Coppi	sfrAB (GSU0509+GSU0510) knockout	
0028-GSUL	R.A. O'Neil	Fe(II) removed after steady state achieved	Cells were harvested as growth rate began to drop (at maxDensity*0.75)
0029-GSUL	K.P. Nevin, D. Holmes	Cells growing on the surface of an electrode within a Geobattery operating in potentiostat mode	
0030-GSUL	K.P. Nevin	Phosphate removed after steady state achieved	
0035-GSUL	G. Reguera	Temperature of 25C	Usual temperature is 30C. Done by TIGR.
0047-GSUL	B.C. Kim	omcF (GSU2432) knockout	
0056-GSUL*	R. DiDonato	Fumarate-limited growth	Different acetate and fumarate concentrations than in 0025-GSUL
0063-GSUL	B. Postier	Temperature of 25C	Usual temperature is 30C. Done at UMass.
0075-GSUL	B. Postier	Medium lacking ammonia so cells have to fix nitrogen in order to grow	Exp. title says the RNA was amplified. Done at UMass.

Table 3 Summary of available experimental data as of December 2006. The control condition for all experiments was the wild type strain with acetate as the electron donor, fumarate as the electron acceptor, and acetate as the limiting nutrient. Data sets marked with * were inverted (from a over b to b over a). The last three experiments were performed on University of Massachusetts arrays whereas the rest were performed on TIGR arrays.

V Cell–Environment Interaction

Three main avenues have been investigated to connect supra-pore scale environmental conditions to microbial cell metabolism: substrate abundance, energetic regulation, and differences reflected in intrinsic genetic potential at the modulon level.

An existing project examining the physiology of marine Roseobacters (Dr. M. A. Moran, UGA, lead PI) provided the opportunity for the Meile group to participate in a comparative genome analysis of Silicibacter. No significant variation in genetic potential related to redox sensing at the modulon level related was apparent between closely related organisms (Silicibacter pomeroyi, Silicibacter sp. M1040, Jannaschia Sp. CCS1) that inhabit distinct ecological niches. As a consequence, this avenue was not pursued further. A manuscript involving C. Meile has been submitted to Applied and Environmental Microbiology (Moran et al. in revision).

A second research avenue considered the role of free energy yield in determining the dominant metabolic pathways as it was suggested earlier that H_2 concentrations reflect the active microbial processes (e.g., Hoehler et al. 1998). In collaboration with Dr. S. B. Joye (University of Georgia at Athens, lead PI), we correlated energy yields based on substrate and product concentration measurements with measured process rates (acetogenesis, H_2 based methanogenesis, H_2 based sulfate reduction, acetate based methanogenesis and acetate based sulfate reduction) in a seafloor brine system. We found that the free energy of reaction and the corresponding measured process rates correlate poorly, illustrating potential limitations of this approach in the field. A manuscript is in preparation (Joye et al. in prep.).

As a consequence of the limitations to the above two approaches, we have focused on the simulation of a set of chemical substances that are known to be of importance in microbial metabolism. These substances can be tracked at the field scale, allowing us to compare model predictions and experimental measurements. We are currently developing models at three levels to accomplish this: metabolism of a single bacterium and reaction-transport modeling at the pore and field scales.

A Cellular metabolism formulation

We are pursuing two complementary avenues: A steady-state approach and a fully kinetic approach. For the steady-state approach, we have adapted the work by Mahadevan et al. (2006). In their metabolic network reconstruction that is based on genome analysis, they use a constraint-based modeling approach to estimate steady state intracellular fluxes and metabolite exchange with the environment. Mathematically, this approach can be formulated as a linear programming problem where the metabolic fluxes or reaction rates (f), in a network described by a stoichiometric matrix S, satisfy $S*f = 0$ (i.e., steady state) and adhere to a given set of physiological constraints on the magnitude of the fluxes. The solution (values for f) is then determined by optimizing for a specific biological function (for example maximum biomass growth). We have implemented the model of Mahadevan et al. (2006) in Matlab to estimate the metabolic fluxes under different conditions, e.g., acetate uptake fluxes. This allows us to estimate – for a given limiting substrate uptake flux – growth rates and uptake or release of metabolites accounted for in the model.

The complementary approach we have recently initiated uses Karyote and focuses on a kinetic description of the TCA cycle. Focusing on availability and use of acetate

(Table 4), Karyote predicts the concentrations of enzymes and metabolites within the cell. This approach has the benefit of resolving the dynamic nature of cell processes, but requires a great deal of knowledge about the metabolism of the organism of interest. While extensive studies have yet to be performed on the complete list of metabolic enzymes, limited data does exist to describe enzymes of importance in the uptake and use of acetate. In *G.sulfurreducens*, the utilization and fate of acetate can be described by three distinct and separate sections of metabolism (Table 4). The first steps involve the uptake and activation of acetate by the cell. The conversion of acetate into acetyl-CoA can occur by two different mechanisms. One is the direct conversion by phosphate transacetylase and acetate kinase. The alternative enzyme in acetate activation is acetyl-CoA transferase, which produces acetyl-CoA and succinate from succinyl-CoA and acetate. Next, the acetyl-CoA can be utilized by citrate synthase in the TCA cycle for ATP generation or it can be used by pyruvate-ferredoxin oxidoreductase to produce pyruvate for gluconeogenesis. Previous modeling studies have demonstrated that the fate of most acetate in the cell is through ATP generation (Mahadevan et al. 2006). The remaining acetate is used for cell growth. The flux of acetate from pyruvate through 2-phosphoglycerate, which is incorporated into biomass, can hence be used to determine cellular growth rates, which can be incorporated into larger (pore or macro-) scale models. We are currently in the process of implementing the reactions provided in Table 4.

Process/Enzyme	**Reaction**	**References for Kinetic Parameters**
Acetate Activation		
Acetate Kinase	ATP + Acetate → ADP + Acetyl-P	Galushko et al. 2000
Phosphate Transacetylase	Acetyl-P + CoASH → Pi + Acetyl-CoA	Galushko et al. 2000
Acetyl-CoA Transferase	Succinyl-CoA + Acetate → Acetyl-CoA + Succinate	Galushko et al. 2000
TCA Cycle		
Succinate Dehydrogenase	Succinate + NADP+ → Fumarate + NADPH	Galushko et al. 2000
Fumarase	Fumarate → Malate	Galushko et al. 2000
Malate Dehydrogenase	NAD+ + Malate → Oxaloacetate + NADH	Galushko et al. 2000
Citrate Synthase	Oxaloacetate + acetyl-CoA → CoASH + Citrate	Bond et al. 2005
Aconitase	Citrate → Isocitrate	Galushko et al. 2000
Isocitrate Dehydrogenase	Isocitrate + NADP+ → CO_2 + NADPH + 2-oxoglutarate	Galushko et al. 2000
Oxoglutarate oxidoreducatase	2-oxoglutarate + CoA → CO_2 + Succinyl-CoA	Galushko et al. 2000
Gluconeogenesis		
Pyruvate Ferredoxin Oxidoreductase	Acetyl-P + CO_2 → Pyruvate	Gebhardt et al. 1985
Pyruvate phosphate Dikinase	ATP + Pi + pyruvate → AMP + PP + PEP	Schwitzguebel et al. 1979
Enolase	PEP → 2PG	Weese, et al. 2005

Table 4 Enzyme and reactions relating to acetate modeled in Karyote. While a large portion of the kinetic parameters used came directly from Geobacter, some data is being obtained from other organisms.

B Reaction-Transport Model Developments

Pore Scale

To investigate the interplay between transport of chemicals and cellular functioning, we have implemented a 2-D representation of the pore scale using a finite element approach in COMSOL. We compute the flow field in a 6mm x 3mm domain, by imposing a pressure gradient and using periodic boundary conditions along the flow direction (Fig. 17). We then include expressions for the evolution of substrate concentrations and compute biomass distribution both in solution and attached to grain surfaces, subject to growth, death, sorption, and transport in the fluid phase. Growth and acetate uptake are computed either via a Monod-type dependency and growth efficiency, taking into account minimum acetate requirements, or using the results from the linear programming approach (see above). Preliminary results at this small scale indicate that under typical flow and production/consumption conditions, relatively small variation in substrate and biomass distribution at the pore scale is to be expected. Spatial heterogeneity of these distributions, however, can become significant in the presence of moving fronts (based on scenarios motivated by large scale model simulations; see below). This work has been part of a poster presentation at the Academy of the Environment meeting at the University of Georgia, and an invited talk at the American Society of Limnology and Oceanography (King and Meile 2006; Meile et al. 2007).

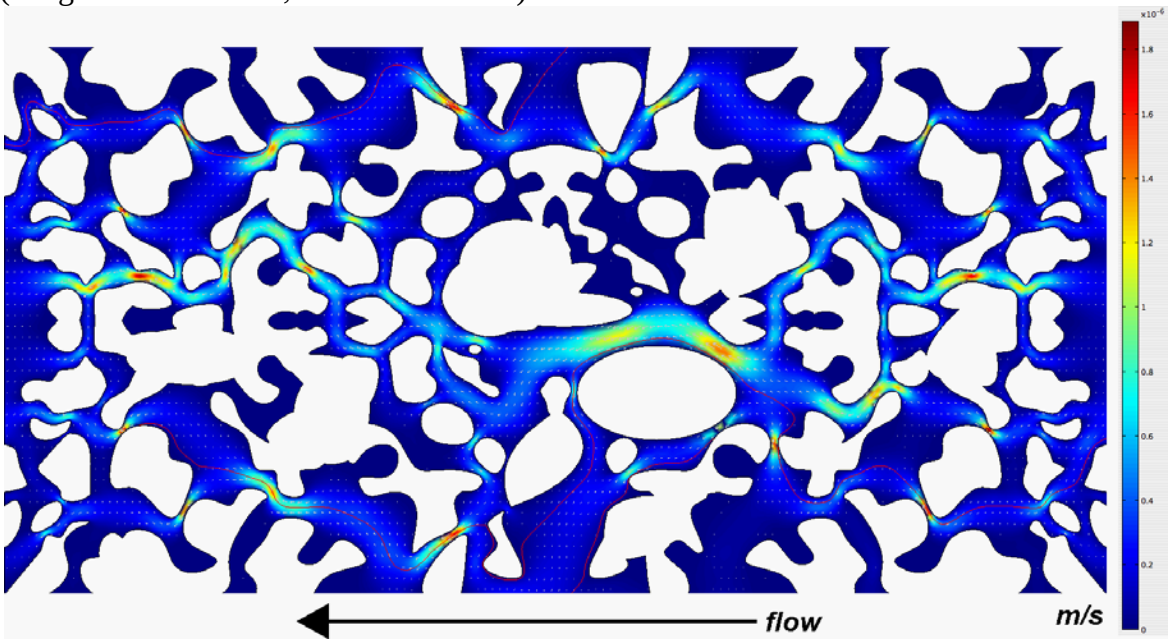


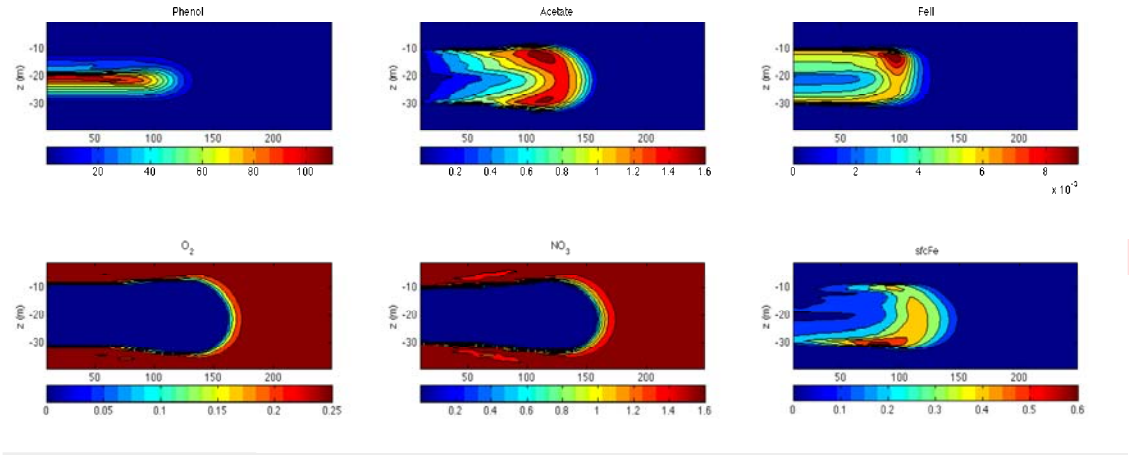
Fig. 17 Representation of flow velocities in a 2-D porous medium (~ 6mm by 3mm). Colors and streamlines depict fluid flow whereas white areas illustrate the grains within the system. The arrow represents the dominant flow direction along the imposed pressure gradient. The domain is periodic and represents a torus, i.e., outflow on the left and top is inflow at the right and bottom, and vice versa.

Large Scale

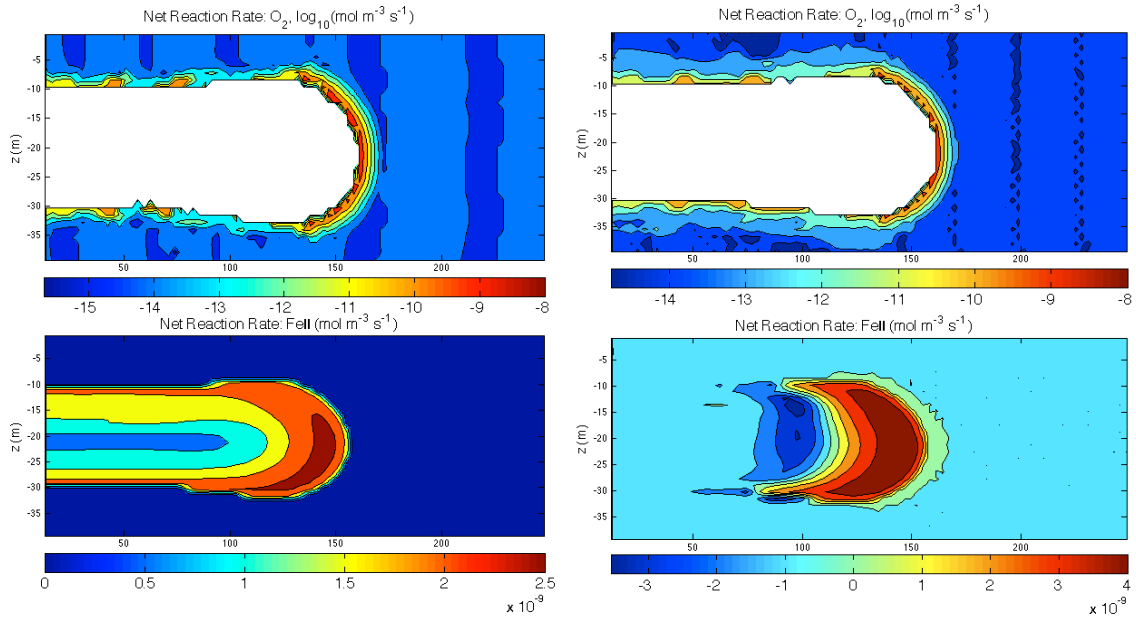
To provide pore scale simulations with a larger scale context, we are implementing reaction-transport models that are based on volume averaged properties. We compute concentration fields in porous media using a finite element approach, and use operator splitting method to achieve a modular code design (Appendix F). The governing equations

solved are $\phi \frac{\partial C_i}{\partial t} = \vec{\nabla} \cdot (D^* \vec{\nabla} C_i) - \vec{\nabla} \cdot (\phi \vec{v} C_i) + \phi R_i$ for solutes, and $\frac{dC_i}{dt} = R_i$ for solids, where C is expressed per volume of solute or solid phase, respectively, D^* is the dispersion tensor parameterized after Scheidegger (1961), $\vec{v}$ is the flow velocity and R is the net reaction rate per volume of a given phase, resulting from an arbitrary, user-defined set of reactions.

Building on a scenario of subsurface phenol contamination, we are currently investigating the impact of using a comprehensive reaction network. We are therefore expanding the set of processes taken into account in Watson et al. (2005) who considered primary reactions (i.e., reactions related to the breakdown of organic matter or contaminant derived electron donor) and sorption processes, by including secondary reactions – those that describe the interaction between reduced substances produced in the primary reactions (Fig. 18).



a.



b.

Fig. 18 Large scale (40m by 400m) simulations for a phenol contamination site following the scenario in Watson et al. (2005), after 13 yrs, including secondary reactions. (a) Concentration fields (mM) of select chemical using the full reaction network (clockwise: phenol, acetate, dissolved reduced iron, sorbed reduced iron, nitrate and oxygen). (b) Net rates (mol m^{-3} s^{-1}) of production and consumption for O_2 (top row; log scale) and Fe(II) (bottom row; linear scale) for primary reactions only (left) vs. a reaction scheme including primary, secondary and sorption reactions (right). The more comprehensive reaction network predicts zones of both net iron production and consumption.

VI Conclusions

- TRN results on *E.coli*, B cell and *G.sulfurreducens* show that our TRND approach is now mature and it can be applied to microbes of interest to the DOE.
- We now have a fairly broad metabolic and transcriptional regulatory network for *G.sulfurreducens*.
- We have made progress in modeling both pore and macroscopic scale reaction-transport models. We will expand on our initial pore scale model by investigating a more comprehensive reaction network, in particular the role of surface associated processes.
- We shall finalize our new pore scale model to resolve the three-dimensional nature of the porous media. We are currently working on the incorporation of a 3-D Stokes flow field with a comprehensive reaction network in Fortran.
- We shall complete the installation of our nonlinear dynamical systems analysis module as an additional site on our portal (*sysbio.indiana.edu*). Thus, with TRN information and associated transcription, translation and post-translational data as input, this workflow will allow one to discover conditions for which cell genomic/proteomic behavior will support one set of pathways, as well as conditions at which dramatic pathway switching will occur.

References

Basso, K., A.A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, A. Califano (2005). Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37, 382-390.

Bond, D. R., T. Mester, C. L. Nesbo, A. V. Izquierdo-Lopez, F. L. Collart, and D. R. Lovley. 2005. Characterization of citrate synthase from Geobacter sulfurreducens and evidence for a family of citrate synthases similar to those of eukaryotes throughout the Geobacteraceae. Appl. Environ. Microbiol. 71:3858-3865.

Cam, H., E. Balciunaite, A. Blais, A. Spektor, R.C. Scarpulla, R. Young, Y. Kluger, B.D. Dynlacht (2004). A common set of gene regulatory networks links metabolism and growth inhibition. *Molecular Cell* 16: 399-411.

Comsol. (2005). Comsol Multiphysics User's Guide, pp. 638.

Galushko, A. S. and B. Schink. 2000. Oxidation of acetate through reactions of the citric acid cycle by Geobacter sulfurreducens in pure culture and in syntrophic coculture. Arch. Microbiol. 174:314-321.

Hoehler T.M., M.J. Alperin, D.B. Albert, and C.S. Martens (1998). Thermodynamic control on hydrogen concentrations in anoxic sediments. *Geochim. Cosmochim. Acta* 62(10), 1745-1756.

Joye, S.B., V.A. Samarkin, B.N. Orcutt, K.-U. Hinrichs, I.R. MacDonald, C.D. Meile, M. Elvert, J.P. Montoya. Deeply sourced brines fuel microbial activity along the seafloor in the Gulf of Mexico (in preparation).

King, E. and C. Meile (2006). A computational study on pore scale substrate variability and microbial metabolism: Preliminary results. The Academy of the Environment meeting, October, Athens GA.

Mahadevan R., D.R. Bond, J.E. Butler, A. Esteve-Nunez, M.V. Coppi, B.O. Palsson, C.H. Schulling, and D.R. Lovley (2006). Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling. *Applied and Environmental Microbiology* 72: 1558-1568.

Meile, C., E. King, K. Tuncay, and P. Ortoleva (2007). Investigating the local microbial environment Pore scale modeling linked to Geobacter sulfurreducens metabolism. ASLO meeting February, Santa Fe.

Moran, M.A., R. Belas, M.A. Schell, J.M. Gonzalez, Sun, F., Sun, S. Binder, B.J., Edmonds, J., Ye, W., Orcutt, B., Howard, E.C., Meile, C. Palefsky, W., Goesmann, A., Ren, Q., Paulsen, I., Ulrich, L.E., Thompson, L.S., Saunders, E. and Buchan, A. Ecological Genomics of Marine Roseobacter (in revision, Applied and Environmental Microbiology).

Qu, K., A.E. Abi Haidar, J. Fan, D. Basu, G. Lin, L. Ensman, M. Jolly, P. Ortoleva (2007). Cancer Onset and Progression:A Genome-Wide, Nonlinear Dynamical Systems Perspective on Onconetworks. *Journal of Theoretical Biology* (in press).

Sayyed-Ahmad, A., K. Tuncay, P. Ortoleva (2007). Microarray analysis through transcription kinetic modeling and information theory, BMC Bioinformatics, 8, 20.

Scheidegger, A.E., 1961. General theory of dispersion in porous media. Journal of Geophysical Research, 66(10): 3273-3278.

Schwitzguebel, J.P. and L. Ettlinger. 1979. Phosphoenolpyruvate carboxylase from Acetobacter aceti. Arch. Microbiol. 122:109-115.

Sun, J., K. Tuncay, A.A. Haidar, F. Stanley, M. Trelinkski, P. Ortoleva (2006). Transcriptional Regulatory Network Discovery via Multiple Method Integration: Application to *E.coli* K12 (in revision).

Tuncay, K., L. Ensman, A.A. Haidar, F. Stanley, M. Trelinski, P. Ortoleva (2006). Transcriptional regulatory networks via gene ontology and expression data, *In Silico Biology*, 7, 3.

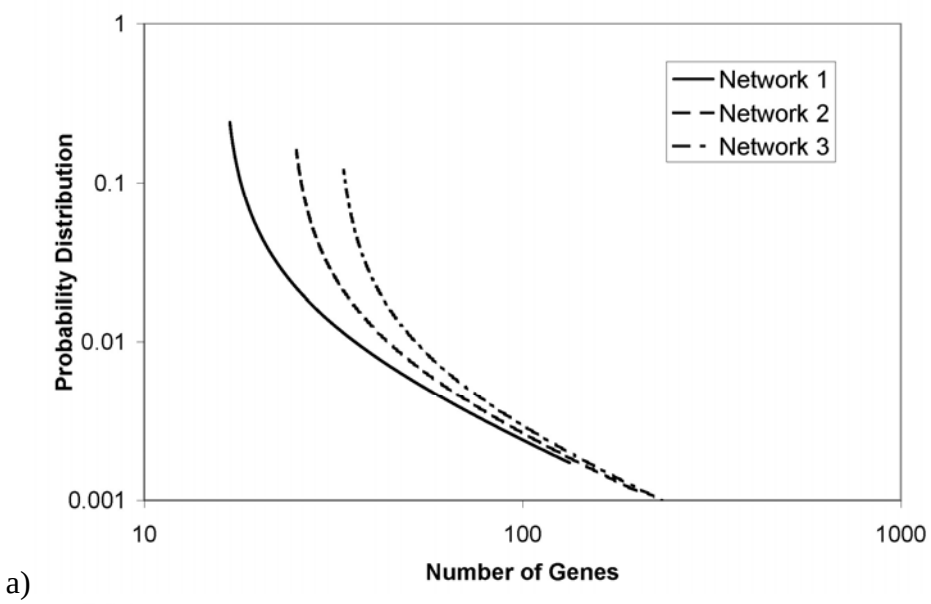
Watson I. A., S.E. Oswald, S.A. Banwart, R.S. Crouch, and S.F.Thornton (2005) Modeling the dynamics of fermentation and respiratory processes in a groundwater plume of phenolic contaminants interpreted from laboratory- to field-scale. *Environ. Sci. Technol.* 39: 8829-8839.

Weese, B. and W.C. Plaxton. 2005. Purification and characterization of a Homodimeric Enolase from Synechococcus. *J. Phycol.* 41: 515–522

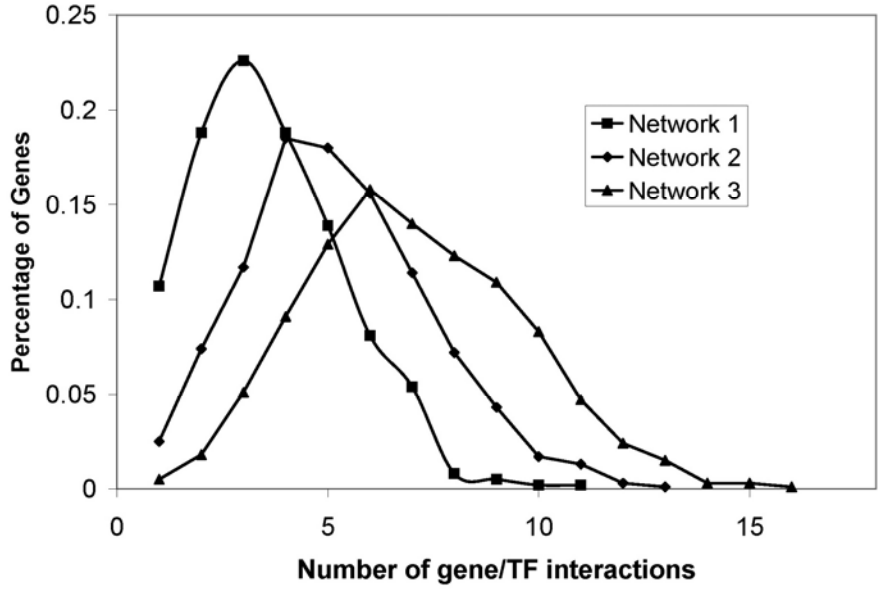
Wu, H., Z. Su, F. Mao, V. Olamn, Y. Xu (2005). Prediction of functional modules through comparative genome analysis and application of gene ontology. *Nucleic Acids Research* 33:2822-2837.

Appendix A: Construction of TF Activities

To test FTF we generated a TRN that consists of 1,000 genes and 100 TFs. The properties of the TRN are shown in Fig. A1. The synthetic expression data was generated by assumed random TF activities. Expression data for gene *i* was generated using $m_i^r = \sum_{n=1}^{N_{TF}} Q_{in} b_{in} T_n^r$. Here, m_i^r is the expression level of gene *i* in experiment *r*, T_n^r is the activity of TF *n* in experiment *r*, N_{TF} is the number of TFs, and Q_{in} is a measure of the binding affinity of TF *j* and gene *i*. Values of Q_{in} were allowed to change 20 fold and were generated randomly (in the logarithmic scale). Our synthetic examples with large TRNs show that, despite the simplicity of the FTF approach, the constructed TF activity profiles are reliable. For example, for a TRN that has the properties shown in Fig. A1, even when we eliminate 50% of the TRN to create a "preliminary TRN", 90% of the constructed TF activities have a correlation coefficient of at least 0.70 with the TF activities used to generate the synthetic expression data (when 20 or more microarray experimental conditions were used). Fig. A2 shows the dependence of the results on the number of experiments. This graph shows that, for practical reason, it is not feasible to recover the full network. Fig. A3a shows the effect of network structure on the results. As the network gets denser, the percentage of the network that can be recovered decreases. Fig. A3b illustrates the dependence of the percentage of recovery on the degree of incompleteness in the preliminary TRN. As anticipated, more complete preliminary TRNs allow a higher percentage of the unknown part of the network to be recovered using expression data. These results suggest that in a real world application such as *E.coli* (for which we have probably less than 40% of the TRN – based on the number of TF/gene interactions known and expected number of TFs), one can not expect to construct the full TRN using expression data alone, regardless of the number of expression datasets available.

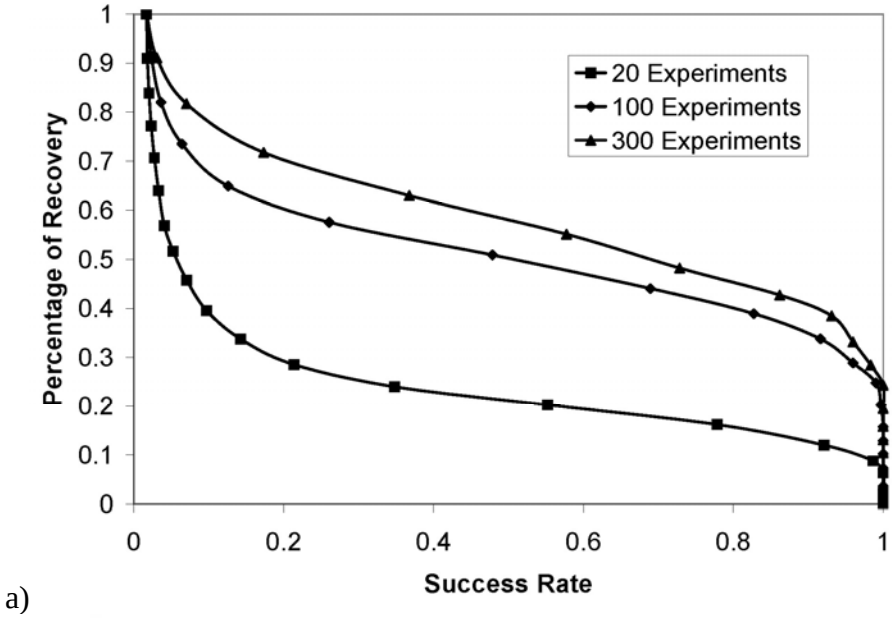


a)

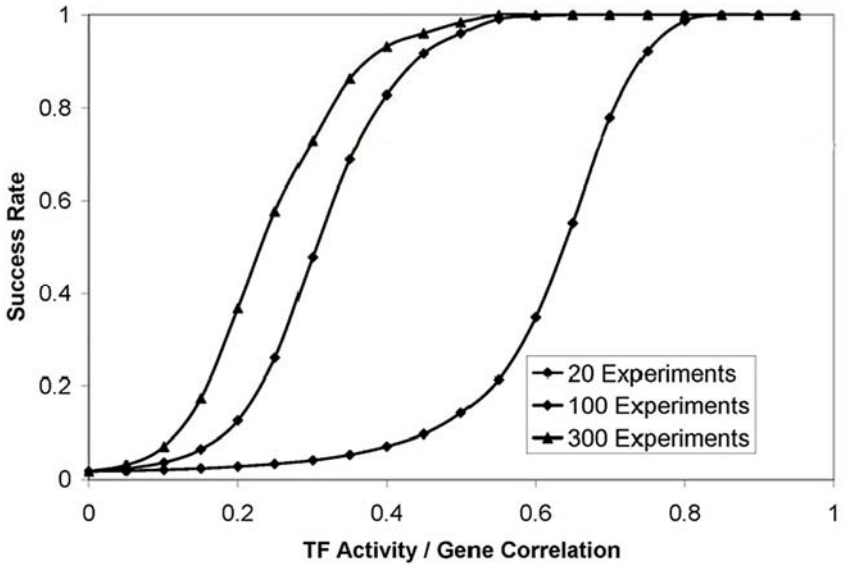


b)

Fig. A1 Properties of TRNs used in the synthetic examples. Networks that consist of 1,000 genes and 100 TFs are generated using the probability distribution for the number of genes regulated by a given TF shown in (a). The corresponding probability distribution for the number of regulators per gene is shown in (b). The average number of regulators per gene is 3.62, 5.22, and 7.02 for Networks 1, 2 and 3, respectively. Equal likelihood is chosen for up versus down regulation.

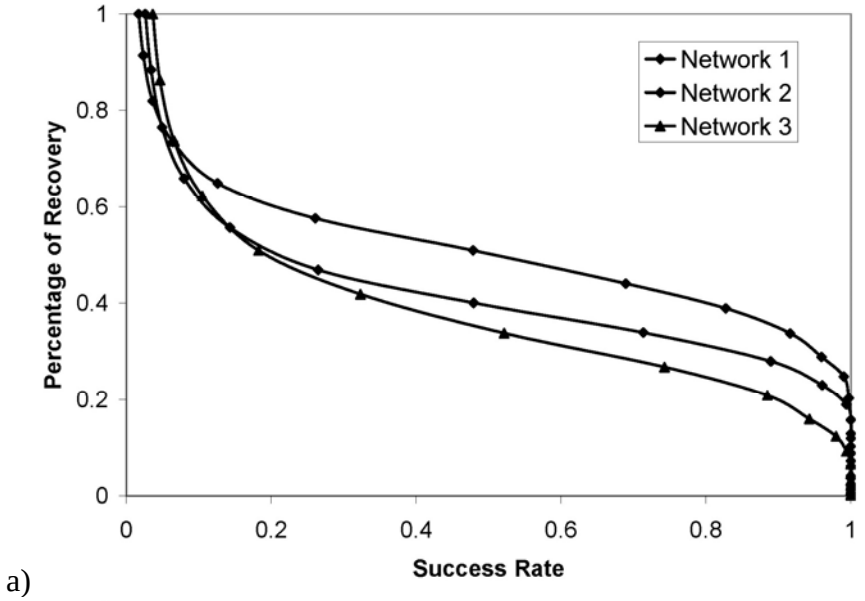


a)

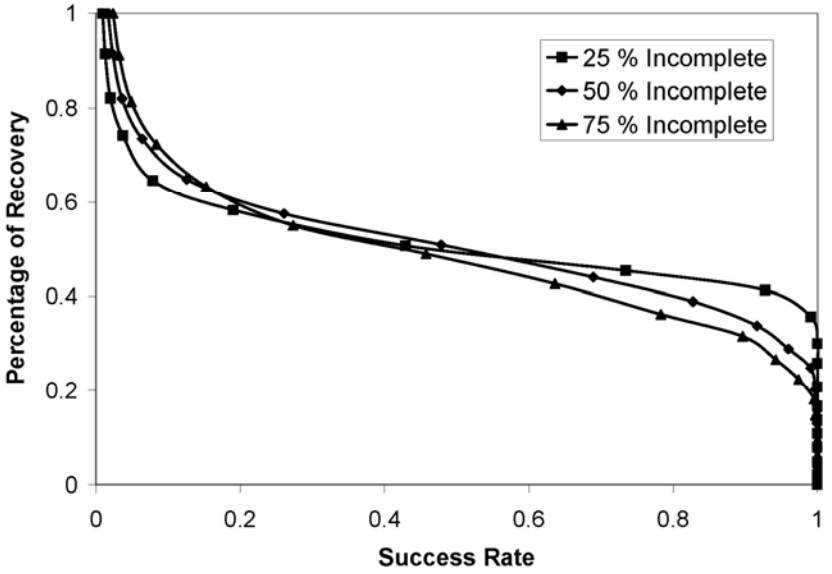


b)

Fig. A2 Reconstruction of TRNs. We have used the Network 1 of Fig. A1 and generated synthetic expression data. Then, we eliminated 50% of the network (randomly), and used FTF to reconstruct the deleted network. Fig. a) shows the percentage of the deleted network recovered as a function of success rate, a measure of the likelihood that an interaction is correct, as estimated from the training set (known interactions). As the number of microarray experiments increases, a higher percentage of the network can be reconstructed. However, full reconstruction requires too many experiments. Fig. b) shows success rate as a function of the absolute value of the linear correlation between the constructed TF activity profiles and gene expression data.



a)



b)

Fig. A3 Effect of TRN properties. We used Networks 1, 2 and 3 of Fig. 3 to generate 100 synthetic expression data sets, and eliminated 50% of the TF/gene interactions in the TRN. Shown is the percentage of the deleted network recovered as a function of success rate. As the number interactions increases, the percentage of the network that can be recovered decreases. b) Same as a) except we used Network 1 and eliminated 25%, 50%, and 75% of the network. As expected, higher percentage of the deleted network is recoverable when a more complete network is known.

Appendix B: Phylogenic Similarity Analysis

We first construct a vector for each gene in *E.coli*, the dimension of the vector being the number of genomes used in the analysis (in this study 229). We applied BLASTP to identify probable orthologous genes of a target genome in 229 reference genomes. The most significant BLASTP hit from each reference species was considered the true ortholog of the target species if the expectation value was less than 1.0e-10 (McCue et al. 2001). If there is an orthologous gene in the i^{th} genome, then the i^{th} entry in this vector is assigned the order (location) of the orthologous gene in the i^{th} genome. If an orthologous gene does not exist in the i^{th} genome, then this entry is taken to be *0*. Once such a vector for each *E.coli* gene is constructed, we compute a phylogenic similarity measure for each gene pair. Given two vectors $X_i = [x_{i1}, x_{i2}, \cdots, x_{i229}]$ for gene *i* and similarly X_j for gene *j*, we use the following phylogenic similarity measure for a gene pair:

$$S_{ij}^{PHY} = -\sum_{k=1}^{229} \log[P(x_{ik}, x_{jk})]. \qquad \text{(B1)}$$

Here $P(x_{ik}, x_{jk})$, the likelihood of genes *i* and *j*, is calculated from

$$P(x_{ik}, x_{jk}) \begin{cases} = (1-p_{ik})(1-p_{jk}) & \text{if } x_{ik}=0 \text{ and } x_{jk}=0 \\ = p_{ik}(1-p_{jk}) & \text{if } x_{ik}\neq 0 \text{ and } x_{jk}=0 \\ = (1-p_{ik})p_{jk} & \text{if } x_{ik}=0 \text{ and } x_{jk}\neq 0 \\ = p_{ik}p_{jk}\dfrac{d(x_{ik},x_{jk})(2N_k - d(x_{ik},x_{jk})-1)}{N_k(N_k-1)} & \text{if } x_{ik}\neq 0 \text{ and } x_{jk}\neq 0 \end{cases} \qquad \text{(B2)}$$

where

p_{ik} is the probability that gene *i* is present in genome *k*.

N_k is the total number of genes in reference genome *k*

$d(x_{ik}, x_{jk}) = abs(x_{ik} - x_{jk})$.

To calculate p_{ik}, we grouped 229 reference genomes into subgroups based on information gathered from *pathema.tigr.org* and *us.expasy.org/sprot/hamap/bacteria.html*. It is assumed that p_{ik} is identical within each subgroup for each gene. Then p_{ik} is taken to be the number of genomes that has an orthologous gene to the total number of genomes in the subgroup.

Appendix C: Application to *E.coli*

We used expression data obtained from NIH GEO (GSE7, GSE8, GSE9 - 65 datasets) and a training set of TF/gene interaction from EcoCyc (*www.ecocyc.org*). EcoCyc includes *E.coli* operons, promoters, TFs, and TF binding sites and describes the mechanisms of transcriptional regulation of *E.coli* genes. It contains the most complete description of the genetic network of any organism. EcoCyc and RegulonDB (Salgado et al. 2004) are curated to ensure that their data content is the same. The preliminary TRN used in this study included 984 genes, 144 TFs, and 2007 TF/gene interactions. Out of the 2,007 TF/gene interactions, 1,124 were up regulation, 766 were down regulation, 5 were uncertain, and 112 were dual regulation (both up/down). All methodologies provided in Sect. II.3 were used to calculate the final score (Sect. II.4) for all possible TF/gene interactions. All bacterial sequence information was downloaded from *ftp://ftp.ncbi.nih.gov/genomes/Bacteria*. The probability distributions of the integrated confidence score for the training and complete TF/gene sets are shown in Fig. C1. We applied a threshold of 1.3 to this score to find the most likely TF/gene interactions. The suggested TRN includes 3,694 new TF/gene interactions. After we performed the calculations we found 206 more TF/gene interactions in the RegulonDB (Salgado et al. 2004) and EcoCyc databases that were not included in the training set. 44 out of 206 regulatory interactions were predicted by our methodology. Out of 44 interactions, the nature of the regulation was correctly predicted for 33 of them. Regulation type couldn't be obtained for 7 interactions. The p-value for predicting at least 44 out of 206 TF/gene interactions to be less than 1.0e-50 (expected proportion=3.5e-04, number observed=44, sample size=3,694). We also used the gene expression data to further test the suggested TRN as follows. We obtained approximate TF activities for both the training and suggested TRNs. Then, for each gene we calculated the linear correlation coefficient between the expression data and the sum of TF activity profiles (all TFs affecting the gene, accounting separately up versus down regulation). Higher scores indicated better consistency between expression data and TRN. The average scores for the training TRN and the suggested TRN were 0.47 and 0.54, respectively, showing an improvement in the overall consistency of the TRN with gene expression profiles. When the same number of interactions is introduced randomly, average score drops to 0.43 (average of 1,000 Monte Carlo simulations).

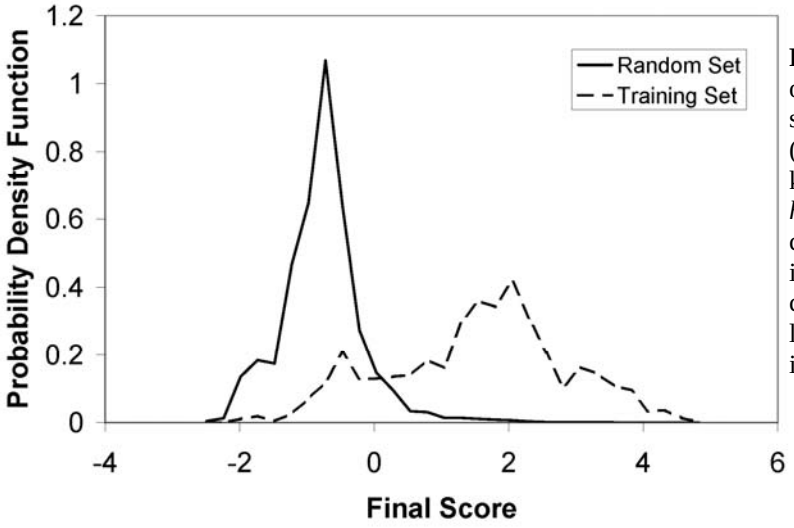


Fig. C1 Probability density function of combined scores for the training set (dashed) and the random set (solid). The training set is based on known TF/gene interactions from *http://ecocyc.org/*. The random set consists of all possible TF/gene interactions. It is seen that higher combined score implies higher likelihood of an actual TF/gene interaction.

Appendix D: Application to B Cell

336 sets of expression data on B cells, gathered by Basso et al. (2005), were obtained from the NIH Gene Expression Omnibus (GSE2350). The data includes normal purified cord blood (5 samples), germinal center (10 samples), memory (5 samples) and naive (5 samples) B cells, B cell chronic lymphocytic leukemia (34 samples), diffuse large B cell lymphomas (68 samples), Burkitt lymphoma (27 samples), follicular lymphoma (6 samples), primary effusion lymphoma (9 samples), mantle cell lymphoma (8 samples), hairy cell lines (16 samples), and 5 lymphoblastic cell lines. Detailed information on the experimental conditions is provided in Basso et al. (2005). 443 TFs and 4032 TF/gene interactions for 1,335 of the genes were found in GenDat.

First, the Gene Ontology method was used to predict and score interactions. Fig. D1 compares the probability distributions of GO scores for the random (all TF/gene pairs) and training set (TF/gene pairs from GenDat). The statistical significance of the difference was evaluated by the chi square test which resulted in a p-value much smaller than 0.0001 (using six bins). Therefore, the hypothesis "the likelihood that a gene pair is regulated in the same manner increases with the number of shared ancestors in the GO tree" is supported by our results.

We also applied the FTF method to the B cell data. The probability distributions for the correlation between the constructed TF activities and expression data are shown in Fig. D2 for training and random sets. As with the GO scores, the statistical significance of the results was evaluated by the chi square test, which resulted in a p-value much smaller than 0.0001 (using six bins).

Finally, we applied the correlation method to the B cell data. Fig. D3 shows the probability distributions for the random and training sets, confirming the hypothesis "higher gene-gene correlation implies greater likelihood of co-regulation."

These three methods were combined and the probability distributions of the integrated confidence scores for the training and random sets are shown in Fig. D4. We applied a threshold of 1.8 to this score to identify the most likely TF/gene interactions, i.e., to construct the final predicted TRN. To facilitate the use of our results by others, they are posted at *sysbio.indiana.edu/trndresults* (see also Tuncay et al. 2006). The preliminary TRN included 1,335 genes and 2,164 TF/gene interactions. In the final TRN, there were 14,616 TF/gene interactions that scored higher than the threshold. The number of genes with at least one TF/gene interaction was 2,164.

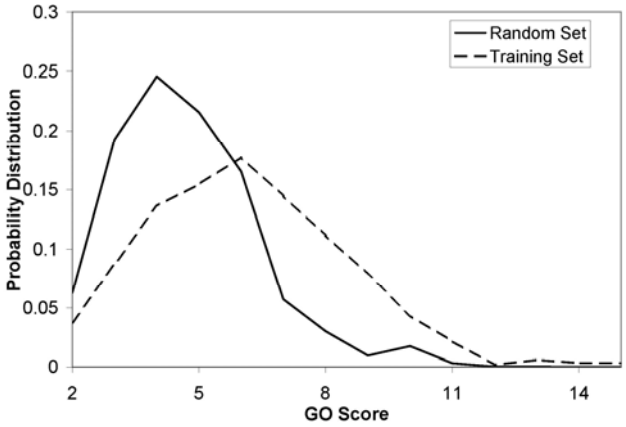


Fig. D1 Comparison of the probability distributions of GO similarity scores of the training set (dashed) and the random set (solid). The training set consists of all known TF/gene interactions for those genes with GO terms assigned. The random set consists of all possible TF/gene interactions for those genes with GO terms assigned. It is seen that higher GO similarity score implies higher likelihood of a TF/gene interaction, particularly when the GO similarity score is larger than 9.

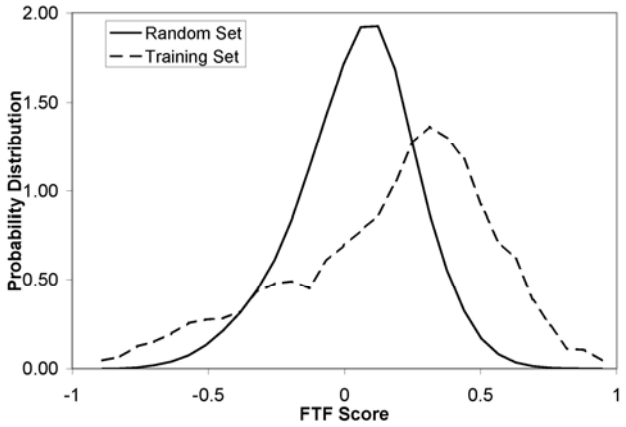


Fig. D2 Probability distribution of FTF scores for the training set (dashed) and the random set (solid). The x-axis is the interaction score while the y-axis here and in Figs 3 to 7, shows the probability density (i.e., probability per score interval) and not probability fraction, thus values can be greater than one but the integrated area is equal to one.

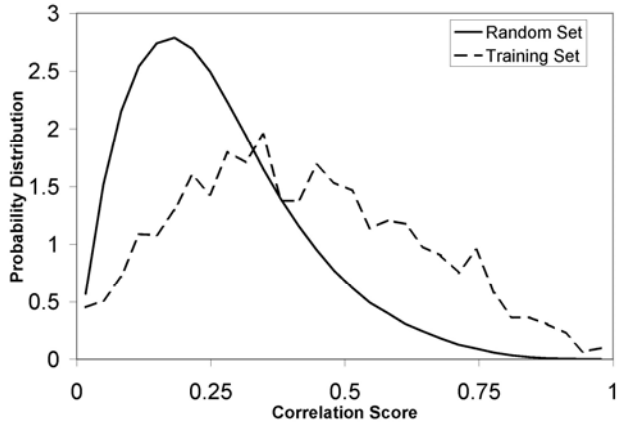


Fig. D3 Probability distribution of correlation scores of the training set (dashed) and the random set (solid) based on the gene/gene to TF/gene score transformation of C2.3. Although this method is based on linear correlation, it requires a preliminary TRN which clearly improves the results.

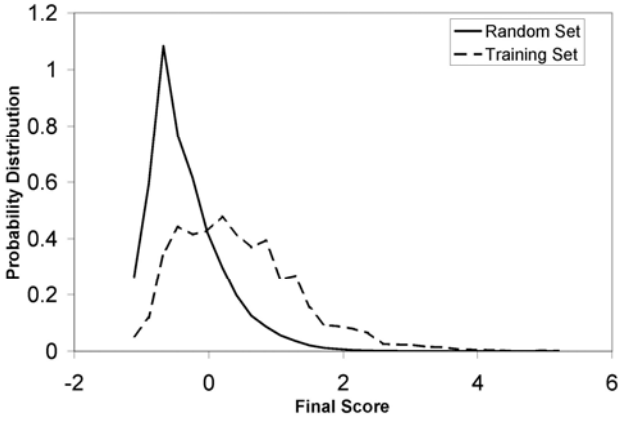
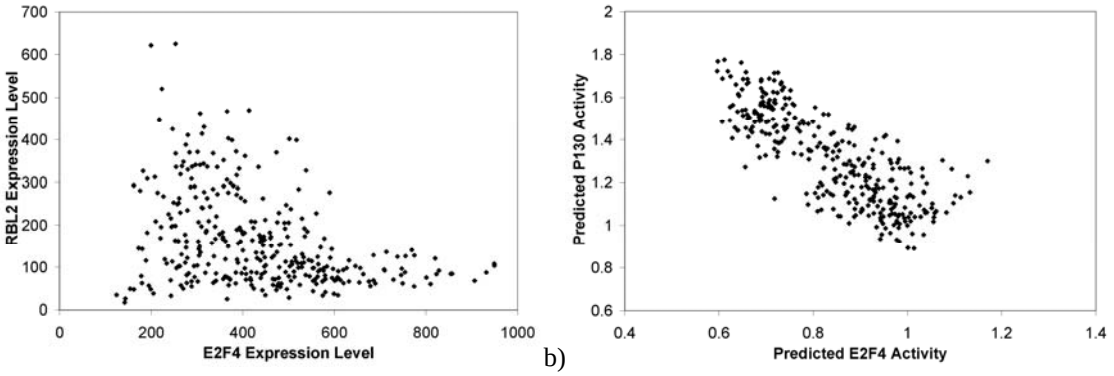


Fig. D4 Probability distribution of combined scores for the training set (dashed) and the random set (solid). The training set is based on all known TF/gene interactions. The random set consists of all possible TF/gene interactions. It is seen that higher combined score implies higher likelihood of a TF/gene interaction.

Validation 1: P130 and E2F4

After we prepared the preliminary TRN and obtained the enhanced TRN using the methodology described above, we located a manuscript by Cam et al. (2004) on transcription factors P130 and E2F4. In the following, we compare our predictions with their experimental results. The E2F family of TFs, which includes E2F1 to E2F7, regulates cell proliferation. P130 is a tumor repressor protein that falls into the pRB protein family, also known as pocket proteins. Pocket proteins directly inhibit E2F and recruit other factors to down regulate gene expression. E2F activity is also regulated through direct interactions with cyclin A, SP1 and P53 (Johnson 1998). All naturally occurring pocket family mutants isolated from human tumors lack the ability to bind and negatively regulate E2F. Cam et al. (2004) used genome-wide analysis of TF occupancy (via chromatin immunoprecipitation on microarrays - ChIP-on-Chip) for E2F4 and P130. Three arrest conditions were studied: quiescent and contact inhibited T98G cells and $P16^{INK4A}$ induced arrest of U2OS cells. 272 genes were found to be targeted by E2F4, P130, or both under any of the three conditions of growth arrest (Table 1, Cam et al. 2000). 171 of these target genes were found in the B cell expression data. At least 88% of the P130 and E2F4 targets were common to all 3 arrest conditions.

In the preliminary TRN, 12 genes were regulated by E2F4 (2 of them were found in Cam et al. 2004) and 43 genes were regulated by P130 (4 of them were found in Cam et al. 2004). 3 genes were coregulated by P130 and E2F4. Therefore, not only were the training sets for these TFs small, they also overlapped a very small set of those reported by Cam et al. (2000). TRND yielded 419 and 750 TF/gene interactions for E2F4 and P130, respectively. 50 (for E2F4) and 55 (P130) target genes that scored higher than the threshold were also reported in Cam et al. (2004). The p-value for this success rate (using a binary probability distribution) is much less than 1.0e-30. Co-regulation by P130 and E2F4 was an outcome of this study, despite the poor training sets. Fig. D5a is a scatter graph for expression of genes E2F4 and RBL2 (which codes for P130). The correlation coefficient is found to be -0.36. Fig. D5b shows the scatter graph for the activities of TFs E2F4 and P130. The correlation coefficient is calculated to be -0.80. Therefore, although E2F4 and RBL2 expression patterns were not highly correlated, due to post-translational modifications, the activities of these two TFs were found to be related, and a common set of targets were identified.



a) b)

Fig. D5 a) Scatter graph of E2F4 and RBL2 expression levels. The linear correlation coefficient is -0.36. Clearly, there is little relationship between the two sets of expression data. b) Scatter graph of the predicted E2F4 and P130 TF activities. The linear correlation coefficient is found to be -0.80. The training sets of E2F4 and P130 included 12 and 43 interactions, respectively. Only three of the genes were co-regulated by both TFs.

Validation 2: C-MYC

B cell expression data was also used by Basso et al. (2005) who made predictions for the TF C-MYC and compared the results with those available at *www.myccancergene.org*. Our training set for the C-MYC TF included 44 genes, 22 of them were identified as C-MYC targets at *www.myccancergene.org*. TRND provided 542 C-MYC targets, 190 of these predictions were identified as C-MYC targets at *www.myccancergene.org*. In this particular case, the correlation between MYC expression and our predicted TF C-MYC activity was fairly high, 0.49. Therefore, the assumption of representing C-MYC activity by the MYC expression pattern is justified. As a result, all three methods (TRND, Basso et al. 2005, and gene-gene correlation) yield similar results, though TRND shows a slight improvement over the others (Table 1). Comparison to E2F4 and P130 results illustrates that the success of a method in one subset of the TRN may not extrapolate to others, although TRND seems to be more generally applicable than other methods.

N	TRND	Basso et al.	Correlation
552	190	171	148
402	146	132	115
321	118	107	97
205	81	76	62
134	55	55	42
89	40	42	28

Table D1 For each method, the number of predicted C-MYC targets (out of the top N) that are identified at *mycancergene.org*.

Appendix E: Results for *G.Sulfurreducens*

Table E1 Top 169 TF/gene interactions in our TRND analysis that were not in the preliminary TRN. Final score was obtained as explained in *Multi-Method Integration* (Sect. II). As seen in Fig. 14, a very small fraction of TF/gene pairs scored higher than 3.0. If the nature of regulation could not be decided, a question mark is used. At least two methodologies out of three (GO, phylogenetic similarity, and FTF) were required to be considered as a potential TF/gene interaction. Since expression data was available for only 241 genes (with more than 6 usable data points), the weight of expression data in the predictions is very limited.

Transcription Factor	Gene	Final Score	Regulation
TF_GSU0359	GSU0338	3.939	up
TF_GSU1129	GSU0338	3.939	up
TF_GSU2964	GSU0338	3.939	up
TF_GSU0359	GSU0342	3.768	up
TF_GSU1129	GSU0342	3.768	up
TF_GSU2964	GSU0342	4.017	up
TF_GSU0359	GSU0343	3.36	up
TF_GSU1129	GSU0343	3.36	up
TF_GSU2964	GSU0343	3.36	up
TF_GSU0359	GSU0344	4.054	up
TF_GSU1129	GSU0344	4.215	up
TF_GSU2964	GSU0344	4.528	up
TF_GSU0359	GSU0345	3.921	up
TF_GSU1129	GSU0345	3.939	up
TF_GSU2964	GSU0345	3.939	up
TF_GSU0359	GSU0346	3.876	up
TF_GSU1129	GSU0350	4.02	up
TF_GSU2964	GSU0350	4.02	up
TF_GSU1129	GSU0351	5.047	up
TF_GSU2964	GSU0351	4.725	up
TF_GSU2484	GSU1102	3.262	up
TF_GSU1003	GSU1250	3.124	?
TF_GSU3421	GSU1250	3.124	?
TF_GSU1037	GSU1272	3.262	Down
TF_GSU3089	GSU1272	3.262	?
TF_GSU1037	GSU1273	3.313	down
TF_GSU3089	GSU1273	3.313	?
TF_GSU0372	GSU1989	3.262	up
TF_GSU0372	GSU1990	3.124	up
TF_GSU1887	GSU2448	3.313	?
TF_GSU1887	GSU2449	3.124	?
TF_GSU1102	GSU3118	3.262	up
TF_GSU0359	GSU3429	4.06	up
TF_GSU1129	GSU3429	3.939	up
TF_GSU2964	GSU3429	3.939	up
TF_GSU1129	GSU3430	3.993	up
TF_GSU2964	GSU3430	3.993	up
TF_GSU1129	GSU3431	4.06	up
TF_GSU2964	GSU3431	4.06	up
TF_GSU0359	GSU3432	3.939	up
TF_GSU1129	GSU3432	3.939	up

TF_GSU2964	GSU3432	3.939	up
TF_GSU0359	GSU3433	3.921	up
TF_GSU1129	GSU3433	3.921	up
TF_GSU2964	GSU3433	3.921	up
TF_GSU0359	GSU3434	3.876	up
TF_GSU1129	GSU3434	3.993	up
TF_GSU2964	GSU3434	3.993	up
TF_GSU0359	GSU3439	3.674	up
TF_GSU1129	GSU3439	3.738	up
TF_GSU2964	GSU3439	3.738	up
TF_GSU0359	GSU3441	3.674	up
TF_GSU1129	GSU3441	3.674	up
TF_GSU2964	GSU3441	3.674	up
TF_GSU0359	GSU3443	3.674	up
TF_GSU1129	GSU3443	3.674	up
TF_GSU2964	GSU3443	3.674	up
TF_GSU0359	GSU3444	3.993	up
TF_GSU1129	GSU3444	3.993	up
TF_GSU2964	GSU3444	3.993	up
TF_GSU0359	GSU3445	3.939	up
TF_GSU1129	GSU3445	3.939	up
TF_GSU2964	GSU3445	3.939	up
TF_GSU3053	GSU0111	3.124	?
TF_GSU3089	GSU0111	3.218	?
TF_GSU3089	GSU0112	3.09	?
TF_GSU3053	GSU0113	3.307	?
TF_GSU1102	GSU0149	3.124	up
TF_GSU1037	GSU0152	3.216	down
TF_GSU3089	GSU0152	3.216	?
TF_GSU0359	GSU0340	3.87	up
TF_GSU1003	GSU0340	3.87	up
TF_GSU1129	GSU0340	3.87	up
TF_GSU2964	GSU0340	3.87	up
TF_GSU3118	GSU0340	3.87	down
TF_GSU3421	GSU0340	3.87	?
TF_GSU0359	GSU0347	5.097	up
TF_GSU1003	GSU0347	4.781	up
TF_GSU1129	GSU0347	5.115	up
TF_GSU2964	GSU0347	4.794	up
TF_GSU3118	GSU0347	5.358	down
TF_GSU3421	GSU0347	4.02	down
TF_GSU0359	GSU0348	5.115	up
TF_GSU1003	GSU0348	4.781	up
TF_GSU1129	GSU0348	5.115	up
TF_GSU2964	GSU0348	5.115	up
TF_GSU3118	GSU0348	5.358	down
TF_GSU3421	GSU0348	4.02	down
TF_GSU0372	GSU0598	3.262	up
TF_GSU1003	GSU0598	3.262	?
TF_GSU3421	GSU0598	3.262	?

TF_GSU0372	GSU0599	3.124	up
TF_GSU1003	GSU0599	3.124	?
TF_GSU3421	GSU0599	3.124	?
TF_GSU0372	GSU0941	3.124	up
TF_GSU1003	GSU0941	3.124	?
TF_GSU3421	GSU0941	3.124	?
TF_GSU1102	GSU1099	3.325	up
TF_GSU0372	GSU1129	3.124	up
TF_GSU1003	GSU1129	3.124	?
TF_GSU3421	GSU1129	3.124	?
TF_GSU3089	GSU1178	3.154	?
TF_GSU1003	GSU1221	3.379	up
TF_GSU2964	GSU1293	3.257	up
TF_GSU3421	GSU1293	3.257	down
TF_GSU0372	GSU1296	3.262	up
TF_GSU1003	GSU1296	3.262	?
TF_GSU3421	GSU1296	3.262	?
TF_GSU2787	GSU1348	3.379	up
TF_GSU2964	GSU1348	3.379	down
TF_GSU0372	GSU1443	3.124	up
TF_GSU1003	GSU1443	3.124	?
TF_GSU3421	GSU1443	3.124	?
TF_GSU0372	GSU1653	3.262	up
TF_GSU1003	GSU1653	3.262	?
TF_GSU3421	GSU1653	3.262	?
TF_GSU0372	GSU1655	3.124	up
TF_GSU1003	GSU1655	3.124	?
TF_GSU3421	GSU1655	3.124	?
TF_GSU3089	GSU1828	3.36	?
TF_GSU1102	GSU1878	3.307	up
TF_GSU3089	GSU1906	3.593	?
TF_GSU0372	GSU1940	3.124	up
TF_GSU1003	GSU1940	3.262	?
TF_GSU3421	GSU1940	3.262	?
TF_GSU3089	GSU2025	3.199	?
TF_GSU0372	GSU2041	3.262	up
TF_GSU1003	GSU2041	3.262	?
TF_GSU3421	GSU2041	3.262	?
TF_GSU0372	GSU2042	3.124	up
TF_GSU1003	GSU2042	3.124	?
TF_GSU3421	GSU2042	3.124	?
TF_GSU1037	GSU2049	3.216	down
TF_GSU3089	GSU2049	3.216	?
TF_GSU3089	GSU2091	3.218	?
TF_GSU2484	GSU2145	3.262	up
TF_GSU3089	GSU2371	3.926	?
TF_GSU3089	GSU2445	3.262	?
TF_GSU1525	GSU2458	3.262	?
TF_GSU0372	GSU2492	3.124	up
TF_GSU1003	GSU2492	3.124	?

TF_GSU3421	GSU2492	3.124	?
TF_GSU0372	GSU2524	3.262	up
TF_GSU1003	GSU2524	3.262	?
TF_GSU3421	GSU2524	3.262	?
TF_GSU1887	GSU2588	3.379	?
TF_GSU1003	GSU2719	3.776	up
TF_GSU3118	GSU2719	3.776	down
TF_GSU3421	GSU2719	3.776	?
TF_GSU1037	GSU2874	3.732	down
TF_GSU3089	GSU2874	3.732	?
TF_GSU3089	GSU2879	3.795	?
TF_GSU2484	GSU2946	3.313	up
TF_GSU1270	GSU3058	3.124	down
TF_GSU1525	GSU3068	3.307	?
TF_GSU1525	GSU3074	3.325	?
TF_GSU1102	GSU3138	3.307	up
TF_GSU0372	GSU3217	3.262	up
TF_GSU1003	GSU3217	3.262	?
TF_GSU3421	GSU3217	3.262	?
TF_GSU0372	GSU3418	3.124	up
TF_GSU1003	GSU3418	3.124	?
TF_GSU3421	GSU3418	3.124	?
TF_GSU0359	GSU3436	3.921	up
TF_GSU1003	GSU3436	3.993	up
TF_GSU1129	GSU3436	3.939	up
TF_GSU2964	GSU3436	3.939	up
TF_GSU3118	GSU3436	3.993	Down
TF_GSU3421	GSU3436	3.993	Down

Appendix F: Large Scale Finite Element Model

The flow equation is given by:

$$\bar{\rho}\beta\phi\frac{\partial p}{\partial t}+\rho\frac{\partial \phi}{\partial t}=\vec{\nabla}\cdot(\frac{k\rho}{\mu}(\vec{\nabla}p-\rho\vec{g})) \tag{F.1}$$

where $\rho,\bar{\rho},\beta,\phi,\mu,t,k,p,g$ are fluid density [M L^{-3}], reference density, compressibility $\beta=-1/(V_f\partial V_f/\partial p)$ [(M L^{-1} T^{-2})$^{-1}$], porosity [-], dynamic viscosity [M L^{-1} T^{-1}], permeability [L^2], pressure [M L^{-1} T^{-2}], and gravitational acceleration [L T^{-2}], respectively and $\vec{\nabla}$ is the gradient operator [L^{-1}]. Using a Darcy approximation, fluid velocity $\vec{v}$ [L T^{-1}] is given by:

$$\phi\vec{v}=-\frac{k}{\mu}(\vec{\nabla}p-\rho\vec{g}) \tag{F.2}$$

The solute transport equations is:

$$\frac{\partial \phi C_i}{\partial t}=\vec{\nabla}\cdot(D^*\vec{\nabla}C_i)-\vec{\nabla}\cdot(\phi\vec{v}C_i)+\phi R_i \tag{F.3}$$

where *C* is the solute concentration in the fluid [M L^{-3}]. The diffusion tensor [L^2 T^{-1}] is defined by:

$$D_{ij}^*=\phi D^m\delta_{ij}+(\alpha_L-\alpha_T)\frac{v_iv_j}{|v|}+\alpha_T|v|\delta_{ij} \tag{F.4}$$

where $D^m,\delta_{ij},\alpha_L,\alpha_T$ are tortuosity corrected in situ molecular diffusion coefficient [L^2 T^{-1}], Kronecker symbol, and longitudinal and transverse dispersivities [L]. For solids, the governing equation is:

$$\frac{dC_i}{dt}=R_i \tag{F.5}$$

where *C* and *R* are the solid concentration and reaction rate in the solid matrix [M L^{-3}].

The governing equations are split into transport and reaction components (Fig F1). A Galerkin finite element approach is used to discretize the transport equations. As the transport of different chemical species is considered independent of each other, this allows sequential, uncoupled solution. This is done using an iterative conjugate gradient solver. The impact of reaction rates on the concentration fields is then calculated at each computational node using robust solvers for systems of coupled ordinary differential equations.

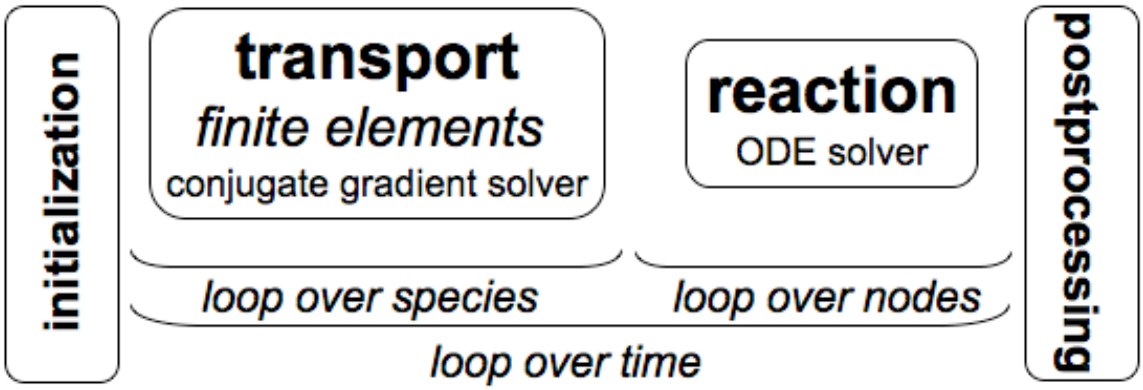


Fig. F1 Schematic of the large scale modeling approach. After model initialization, which includes a definition of the model scenario, the temporal evolution of the concentration field is computed using operator splitting. The transport of solutes is calculated on a finite element grid while the impact of kinetic reactions is implemented through solving a system of ordinary differential equations at each node.