Validation of the Short Time-series Expression Miner (STEM) on Iron Cycling in a Shallow Alluvial Aquifer

Paul Soper
Dept. of Computer Science
University of Delaware
Newark, DE, USA
pdsoper@udel.edu

Stephen Herbein
Dept, of Computer Science
University of Delaware
Newark, DE, USA
sherbein@udel.edu

Michela Taufer
Dept. of Comptuer Science
University of Delaware
Newark, DE, USA
taufer@udel.edu

Clara Chan
Dept. of Geological Sciences
University of Delaware
Newark, DE, USA
cschan@udel.edu

***Abstract*— Zetaproteobacteria are aquatic iron-oxidizing bacteria. They play an important role in the rusting of underwater structures, but their metabolic mechanisms are not well understood. One vital set of data is the level of gene expression after initiating iron oxidation. Data are available for only four time points. We examined the possible use of specialized software, the Short Time-series Expression Miner (STEM), to analyze these data. To validate the method, a better-known system was studied. STEM correctly clustered relevant genes. This has encouraged us to apply STEM to Zetaproteobacteria.**

Keywords—clustering, gene expression, iron-oxidizing, bacteria

I. Motivation

Zetaproteobacteria oxidize iron structures in underwater environments from shallow aquifers to deep-sea hydrothermal vents (the source of our samples). Typical metabolic studies involve injecting a nutrient, sampling the microbial broth at several time points, destroying the cells to extract their RNA, separating the RNA iinto "bins", and analyzing the results. The results are thousands of genes measured at a handful of time points. The key problem is indentifying the relatively few genes of interest among them. This motivated the development of software specifically for such data – the Short Time-series Expression Minder (STEM)[1]. We applied to a simliar but better understood study of iron-oxidizing bacteria to determine its applicability to our data.

II. Methodology

A. Data Source and Processing

We chose to analyze data from *Metatranscriptomic evidence of pervasive and diverse chemolithoautotrophy relevant to C, S, N and Fe cycling in a shallow alluvial aquifer* by Jewell et al.[2]. In particular, we sought to reproduce two of their figures, shown here as Fig. 1 and 2. Their data was published in the Supplementary Information for the paper.

The data were far more extensive than we required. We used the data used in the figures: samples passing a 0.2 micron filter from bins 22.6 (2050 genes) and 22.9 (4994 genes). There are data for four time points: 0, 21, 35, and 47 days from the start of the experiment. The data were already in RPKM, so no normalization was necesary. We identified a number of the genes in the dat. They are listed in Tables 1 and 2.

B. The Short Time-series Expression Miner (STEM)

STEM was developed by Ziv Bar-Joseph and his associates at Carnegie Mellon, who published papers on its theory [3-5] and use [1]. Short term studies look for common patterns in the time series with the hope that they will refer to genes with related functions. Their main insight was that, with so few time points, one could generate all possible profiles, and then see which data matched them.

Data are filtered to remove any genes which have an expression level of 0 at any time point. The remaining values are converted to the log of the ratio of their value at he time they ere takento that at time t=0. (All profiles start at 0.) Candidate profiles are generated by limiting how much this ratio can change at each time point. The maximum change is a small integer – 2 by default. This limitation is what allows the generation of "all possible" profiles

The statistical significance of the time series is tested by comparing the observed sequence to all possible sequences of the same values (for each gene). Finally, profiles and their associated genes are clustered.

The software, written in Java, is available free for non-commericial use. The user interface is shown in Fig. 3. Note that sources for gene annotation, cross references, and locations can be specified, but were not available for our project at the time of this project.

III. Analysis

We ran STEM with the default number of profiles (50) and maximum profile change per time step (±2). Fig. 4 shows the results for bin 22.6. Each proile is numbered. Profiles with a color background are statistically significant, and those with the same color backgroud belong to the same cluster.

Identify applicable funding agency here. If none, delete this text box.

STEM's assignent of the identified genes are shown in Table 1. Note that each was assigned to a statistically significant profile, and those profiles all belong to the same cluster. The clustering shows a certain level of ambiguity. The overall cluster profile appers to be 1) level or down, 2) up, 3) level or down.

The assignment of genes to specific profiles is shown in Fig. 5. Their shapes appear similar to those in Fig. 1, given the different scales and different numbers of genes shown. The results for bin 22.9 are shown in Fig. 6 and tabulated in Table 2. Every identified gene is assigned to a statistically significant profile, and four of the five *dsr* genes are assigned to cluster 1.

IV. Results

The analysis of data from [2] validated the STEM method for the genes which had been identified. Application of the software to our own data also showed statistically significant profiles and clusters. STEM will be useful as we increase our understanding the iron oxidation metabolism of Zetaproteobacteria..

TABLE I. STEM Assignments for Genes in Bin 22.6

Bin 22.6	STEM Assignments for Bin 22.6		
	Gene	***Profile***	***Cluster***
1	dsrL	20	1
2	cyc2	29	1
3	mtoA	29	1
4	dsrO	29	1
5	mtoB	30	1

TABLE II. STEM ASSIGNMENTS FOR GENES IN BIN 22.9

Bin 22.9	STEM Assignments for Bin 22.9		
	Gene	***Profile***	***Cluster***
1	dsrO	5	1
2	dsrK	5	1
3	dsrP	5	1
4	dsrL	7	1
5	mtoB	21	3
6	dsrF	21	3
7	cyc2	42	4

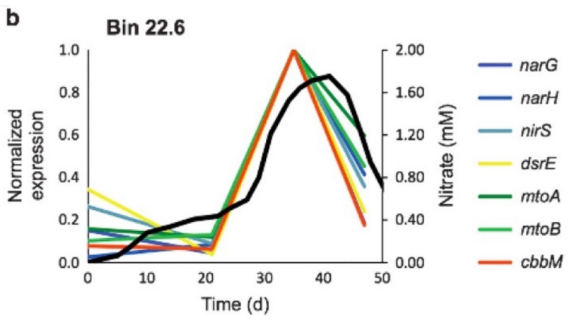

Fig. 1. Gene expression data from [2] for bin 22.6

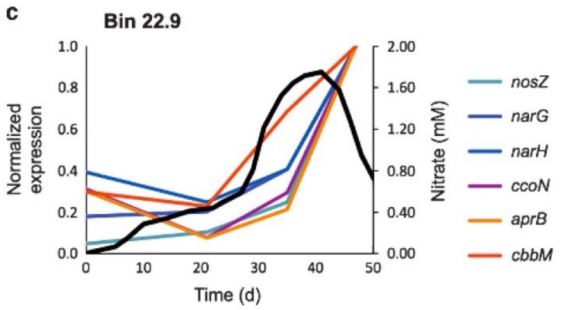

Fig. 2. Gene expression data from [2] for bin 22.9

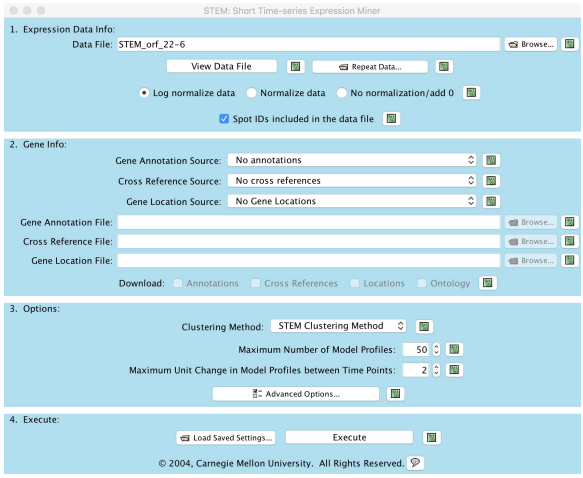

Fig. 3. The user interface for STEM

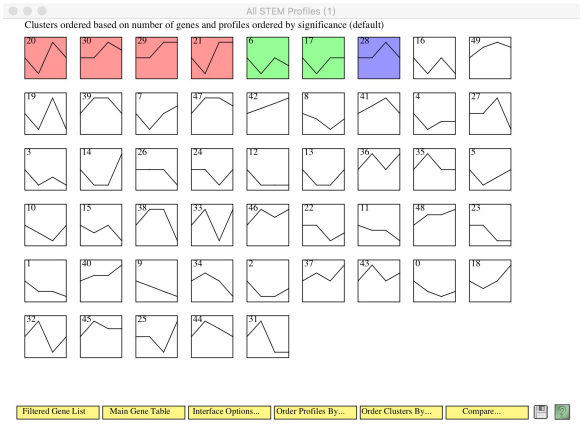

Fig. 4. STEM identification of statistically significant profiles for bin 22.6

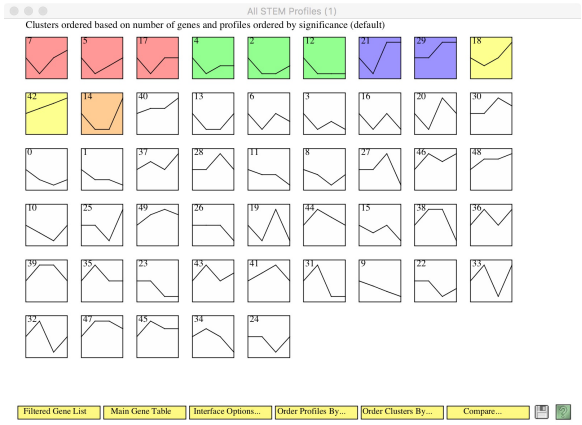

Fig. 6. STEM identificaiton of statistically significant profiles for bin 22.9

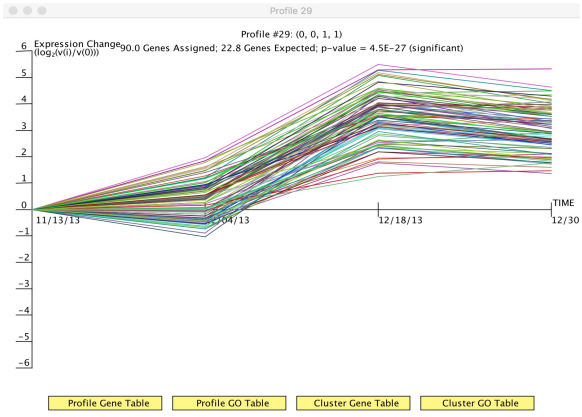

Fig. 5. Individual gene profiles in profile 29 from bin 22.6

REFERENCES

[1] J. Ernst and Z. Bar-Joseph, "STEM: a tool for the analysis of short time series gene expression data," BMC Bioinformatics, vol. 7, pp. 191, 2006.

[2] T.N.M. Jewell, U. Karaoz, E.L. Brodie, K.H. Williams and H.R. Beller, "Metatranscriptomic evidence of pervasive and diverse chemolithoautotrophy relevant to C, S, N and Fe cycling in a shallow alluvial aquifer," The ISME Journal, vol. 10, pp. 2106-2117, Sep. 2016.

[3] Z. Bar-Joseph, G. Gerber, D. Gifford, T. Jaakkola and I. Simon, "A new approach to analyzing gene expression time series data," pp. 39-48, Apr 18, 2002.

[4] Z. Bar-Joseph, G.K. Gerber, D.K. Gifford, T.S. Jaakkola and I. Simon, "Continuous representations of time-series gene expression data," Journal of Computational Biology : A Journal of Computational Molecular Cell Biology, vol. 10, pp. 341-356, 2003.

[5] Jason Ernst, Gerard J. Nau and Ziv Bar-Joseph, "Clustering short time series gene expression data," Bioinformatics, vol. 21, pp. i168, Jun 1,. 2005.