Extracting Sparse Canonical Correlations Between Microbial Communities and Deep Groundwater Geochemistry

Viivi Uurtio[1], Juho Rousu[1], Malin Bomberg[2], and Merja Itävaara[2]

[1] Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science, Aalto University,
Konemiehentie 2, FI-00076 Aalto, Finland
[2] VTT Technical Research Centre of Finland, Espoo, Finland

Abstract. Microorganisms are found in deep subsurface groundwater, upto kilometers deep. Microbial populations interact in communities with the geochemical resources and conditions of their habitat [1], in ways that researchers are only beginning to understand. Here we describe a study on deep subsurface microbial communities in Fennoscandian crystalline bedrock [2],[3]. The motivation for the particular study is risk assessment for long term disposal of nuclear waste [4], where the geobiochemical stability of the site and the potential chemical and physical effects of microbial activity need to be understood.

In order to model the complex network of microbial community interactions with the deep bedrock habitat, we analysed environmental samples obtained from the deep bedrock containing both microbial and geochemical variables. We focused on sulfate reducing bacteria that produce sulphide which may corrode the copper of the nuclear waste capsules. The bacteria were identified by their dissimilatory sulphite reductase marker genes (*dsr*B) that are present in all microorganisms performing dissimilatory sulfate reduction [5]. Computational analysis of this type of data requires a multivariate approach in order to extract correlations among the variables. Since the diverse bacterial interactions with the geochemistry of the habitat are complex and encompassed in high-dimensional data, the visualization and interpretation of the results of multivariate analysis is challenging.

We applied asymmetrical sparse canonical correlation analysis (SCCA) [6] in order to extract subsets of highly correlating sulfate reducing bacterial communities and geochemical measurements from two datasets obtained from deep bedrock drill holes in Finland [5],[7]. SCCA is a multivariate method that seeks semantic projections that use as few relevant features as possible to explain as much correlation as possible [6] by imposing *L1*-norm penalization on variable weights [8].

We imposed sparsity on either of the data views by penalizing dual variables related to the latent variables of the other view. In order to select an optimal level of sparsity for a data view, we performed 3-fold cross validation in which we computed the optimal feature weights at a range of levels of sparsity for a training set and assigned these to an unseen test set in order to obtain a predictive canonical correlation coefficient. The optimal level of sparsity resulted in highest predictive correlation coefficient of the projections. The optimal observed projection correlations were tested statistically by permutation tests in which the optimal level of sparsity was chosen for each permuted dataset.

We analysed the resulting projections by examining the correlation coefficients, i.e. cosine angles of the projections to the feature axes. In this way, we extracted the subsets of features contributing to the highly correlating projection direction. The method of computing linear correlation coefficients among the original and projected measurements was introduced by [9],[10] in the framework of two co-dependent datasets but has not yet been applied to SCCA. The highly correlating subsets of features were visualized by means of correlation plots [9],[10]. We also extended the correlation plot visualization to a clustergram in which the problem of overlapping features was overcome.

When sparsity was imposed on sulfate reducing bacterial data, we discovered a high positive correlation among the *Peptococcaceae* family and the geochemical measurements depth, electrical conductivity, total dissolved salts, total number of cells including the ionic chloride and calcium which points out a possible relation between salinity and sulfate reduction. Another correlation that contributes to this finding was seen when sparsity was imposed on the geochemical measurements, since the *Desulfobulbaceae* family was correlating negatively with ionic chloride. *Peptococcaceae* and *Desulfobacteraceae* families were correlating positively with pH measurements. This could be explained by the fact that sulfate reducers have an impact on it through

consumption of sulfate and production of sulfide which regulate the buffering capacity of the water [11] in their habitat.

Our approach finds biologically relevant correlations that can be used to unravel the complex interactions occurring in a microbial habitat. The optimization of the level of sparsity for each view prior to computation of the feature weights seems to be important in cases where the number of features in the two views differs greatly. Both visualization techniques, correlation plots and clustergrams, enabled biological interpretation of the results.

Asymmetrical SCCA algorithm together with optimization of the level of sparsity and statistical significance testing of the resulting projections provides a means to examine the relations among two sets of co-dependent variables in high-dimensional space. An alternative approach to improve the framework would be to include prior information about the relations among the variables before the computation of the projections.

References

1. Madigan M.T., Martinko J.M., and Parker J. *Brock biology of microorganisms*. Pearson Benjamin Cummings, 2009.
2. M. Itävaara, M. Nyyssönen, A. Kapanen, A. Nousiainen, L. Ahonen, and I. Kukkonen. Characterization of bacterial diversity to a depth of 1500 m in the outokumpu deep borehole, fennoscandian shield. *FEMS microbiology ecology*, 77(2):295–309, 2011.
3. L. Purkamo, M. Bomberg, M. Nyyssönen, I. Kukkonen, L. Ahonen, R. Kietävainen, and M. Itävaara. Dissecting the deep biosphere: retrieving authentic microbial communities from packer-isolated deep crystalline bedrock fracture zones. *FEMS Microbiology Ecology*, 85:324–337, 2013.
4. M. Nyyssönen, M. Bomberg, A. Kapanen, A. Nousiainen, P. Pitkänen, and M. Itävaara. Methanogenic and sulphate-reducing microbial communities in deep groundwater of crystalline rock fractures in olkiluoto, finland. *Geomicrobiology Journal*, 29(10):863–878, 2012.
5. M. Bomberg, M. Nyyssönen, and M. Itävaara. Quantitation and identification of methanogens and sulphate reducers in olkiluoto groundwater. Technical report, Posiva OY and VTT Technical Research Centre of Finland., 2010.
6. D.R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83, 2011.
7. M. Bomberg, M. Nyyssönen, and M. Itävaara. Characterization of olkiluoto bacterial and archaeal communities by 454 pyrosequencing. Technical report, Posiva OY and VTT Technical Research Centre of Finland., 2012.
8. J. Rousu, D. D. Agranoff, O. Sodeinde, J. Shawe-Taylor, and D. Fernandez-Reyes. Biomarker discovery by sparse canonical correlation analysis of complex clinical phenotypes of tuberculosis and malaria. *PLoS computational biology*, 9(4):e1003018, 2013.
9. B-H. Mevik and R. Wehrens. The pls package: Principal component and partial least squares regression in r. *Journal of Statistical Software*, 18, 2007.
10. I. González, K.-A. Lê Cao, M. J. Davis, and S. Déjean. Visualising associations between paired omics data sets. *BioData Mining*, 5, 2012.
11. G. Arp, V. Thiel, A. Reimer, W. Michaelis, and J. Reitner. Biofilm exopolymers control microbialite formation at thermal springs discharging into the alkaline pyramid lake, nevada, usa. *Sedimentary Geology*, 126(1):159–176, 1999.