Project specification

Implementing a Statistical Model for Clustering Metagenomic Contigs by Parallel Computing and Test Driven Development

Brynjar Smári Bjarnason
Computational and Systems Biology
School of Computer Science and Communication
KTH Royal Institute of Technology

CSC Supervisor: Jens Lagergren
School of Computer Science and Communication
KTH Royal Institute of Technology

Supervisor: Anders Andersson
School of Biotechnology
KTH Royal Institute of Technology

Co-supervisor: Christopher Quince
School of Engineering
University of Glasgow

February 26, 2013

The Problem

Metagenomics, or environmental genomics, is the exploration of microbial communities by sequencing environmental samples. For modern genomic research, the use of massive parallel sequencing has made the data processing step time consuming and difficult. This is especially the case for metagenomics where each sample contains a multitude of organisms and multiple samples are used to detect trends or differences over various conditions. To estimate the constitution and diversity of the microbial community studied, the raw sequence data needs to be divided into clusters, a procedure called clustering or binning. The binning procedure can take on either of two approaches, clustering reads or clustering contigs. Contigs are formed by assemblers that piece together overlapping short reads into longer sequences. The problem complexity decreases if reliable contigs are available to the clustering, since the assembly step also decreases data size significantly. Several assemblers dedicated for metagenomic data now exists, indicating the possibility to focus efforts on clustering of contigs instead of reads.

The information used by current clustering algorithm is mainly based on genome signatures and sequence homology. The steady increase in throughput and decrease in cost of sequencing have enabled the possibility to sequence multiple samples simultaneously. This provides additional information that could be used in the clustering process, namely the different abundances in different samples. Contigs belonging to the same cluster should under normal circumstances have the same kind of abundance-pattern over all samples.

Furthermore, most algorithms developed for binning are based on heuristics as opposed to a probabilistic modeling approach. Heuristic algorithms may very well offer some great performance but lack in rigidity and possibilities of interpreting the results. If a proper statistical model is the basis for the algorithm, this enables the use of a statistics toolbox developed by researchers for analyzing the properties of the algorithm.

The problem statement with respect to taxonomic levels might differ between settings, sometimes it is sufficient to cluster the data into bins corresponding to families, while at other settings it is desirable to go down to species or even strains. It is supposedly easier to separate higher taxonomic levels from each other.

Clustering metagenomic data is a problem hard to solve. The microbial communities are often large and diverse and have a high ratio between the highest and the lowest abundant organism. This results in a low signal to noise ratio for the lower abundance organisms making accurate predictions difficult. The small difference in genomic composition between strains of the same species further complicates the clustering down to this ultimate precision.

Goal

With this project the aim is to explore ways to efficiently and accurately cluster metagenomic contigs into taxonomy level based on the underlying mathematical models. Different clustering algorithms will be considered and compared such as Expectation-Maximization (EM), K-Means and others. Mathematical models, which are provided, will be implemented. The clustering and model implementations will be analyzed and optimized for parallel execution on high performance computing clusters. This is expected to be necessary given the size and complexity of the data and probability calculations respectively.

Some clustering algorithms do not work well with high dimensional data. The data for this project is expected to be high dimensional so the need for some dimension reduction or other ways to simplify the data will be investigated.

Reading List

The focus will be on clustering algorithms, specially on the EM. As the project progresses the aim is to look at other algorithms and compare efficiency and quality to the EM. Since all this process will be computationally expensive, parallel computing and optimization will be in focus as well.

High performance computing & Clustering methods

Parallel Clustering Algorithm for Large Data Sets with Applications in Bioinformatics	[Olman et al., 2009]
The Binning of Metagenomic Contigs for Microbial Physiology of Mixed Cultures	[Strous et al., 2012]
Removing Noise From Pyrosequenced Amplicons	[Quince et al., 2011]
MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample	[Wang et al., 2012b]
MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species	[Wang et al., 2012a]
A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio	[Leung et al., 2011]
A Parallel K-Means Clustering Algorithm with MPI	[Zhang et al., 2011]
High-Dimensional Clustering with Sparse Gaussian Mixture Models (Unpublished)	[Krishnamurthy,]
Fast Parallel Markov Clustering in Bioinformatics using Massively Parallel Computing on GPU with CUDA and ELLPACK-R Sparse Format	[Bustamam et al., 2011]
Unsupervised two-way clustering of metagenomic sequences	[Prabhakara and Acharya, 2012]
Clustering metagenomic sequences with interpolated Markov models	[Kelley and Salzberg, 2010]
Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics	[Weber et al., 2011]

Metagenomics

A human gut microbial gene catalogue established by metagenomic sequencing	[Qin et al., 2010]
A primer on metagenomics	[Wooley et al., 2010]
Microbiology: Metagenomics [Hugenholtz and Tyson, 2008]	
Application of tetranucleotide frequencies for the assignment of genomic fragments	[Teeling et al., 2004]
Community-wide analysis of microbial genome sequence signatures	[Dick et al., 2009]
Accurate phylogenetic classification of variable-length DNA fragments	[McHardy et al., 2007]
A novel abundance-based algorithm for binning metagenomic sequences using l-tuples	[Wu and Ye, 2011]

Methods

Test driven development (TDD) will be applied for the implementations. Each mathematical model and clustering algorithm will be implemented and analyzed for memory and execution efficiency. Each implementation will be examined for possibility of parallelization.

Delimitation

Development of the mathematical models will not be examined here.

Time Plan

Week	Task
3	Meet the group, get to know the problem
4	Administrative work, reading
5	Administrative work, reading
6	Reading
7	Vacation
8	Implement and analyze modeling composition
9	Read about clustering methods
10	Read amd design clustering algorithm
11	Implement clustering algorithm
12	Implement clustering algorithm
13	Vacation
14	Implement and analyze Covariance
15	Analyze and optimize
16	Analyze, optimize while working with real data
17	Analyze, optimize while working with real data
18	Write report
19	Write report
20	Write report
21	Write report
22	Write report
23	Write report
24	Write report
25	Defend thesis

References

[Bustamam et al., 2011] Bustamam, A., Burrage, K., and Hamilton, N. A. (2011). Fast Parallel Markov Clustering in Bioinformatics using Massively Parallel Computing on GPU with CUDA and ELLPACK-R Sparse Format. *IEEEACM transactions on computational biology and bioinformatics IEEE ACM*, 9(3):679–692.

[Dick et al., 2009] Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., and Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome biology*, 10(8):R85.

[Hugenholtz and Tyson, 2008] Hugenholtz, P. and Tyson, G. W. (2008). Microbiology: Metagenomics. *Nature*, 455(7212):481–483.

[Kelley and Salzberg, 2010] Kelley, D. R. and Salzberg, S. L. (2010). Clustering metagenomic sequences with interpolated Markov models. *BMC bioinformatics*, 11:544.

[Krishnamurthy,] Krishnamurthy, A. High-Dimensional Clustering with Sparse Gaussian Mixture Models. pages 1–8.

[Leung et al., 2011] Leung, H. C. M., Yiu, S. M., Yang, B., Peng, Y., Wang, Y., Liu, Z., Chen, J., Qin, J., Li, R., and Chin, F. Y. L. (2011). A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics (Oxford, England)*, 27(11):1489–95.

[McHardy et al., 2007] McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature methods*, 4(1):63–72.

[Olman et al., 2009] Olman, V., Mao, F. M. F., Wu, H. W. H., and Xu, Y. X. Y. (2009). Parallel Clustering Algorithm for Large Data Sets with Applications in Bioinformatics.

[Prabhakara and Acharya, 2012] Prabhakara, S. and Acharya, R. (2012). Unsupervised two-way clustering of metagenomic sequences. *Journal of biomedicine & biotechnology*, 2012:153647.

[Qin et al., 2010] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Paslier, D. L., Linneberg, A., Nielsen, H. B. r., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S. r., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Antolin, M., Artiguenave, F., Blottiere, H., Borruel, N., Bruls, T., Casellas, F., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Forte, M., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Jamet, A., Juste, C., Kaci, G., Kleerebezem, M., Knol, J., Kristensen, M., Layec, S., Roux, K. L., Leclerc, M., Maguin, E., Minardi, R. M., Oozeer, R., Rescigno, M., Sanchez, N., Tims, S., Torrejon, T., Varela, E., de Vos, W., Winogradsky, Y., Zoetendal, E., Bork, P., Ehrlich, S. D., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

[Quince et al., 2011] Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics*, 12:38.

[Strous et al., 2012] Strous, M., Kraft, B., Bisdorf, R., and Tegetmeyer, H. E. (2012). The Binning of Metagenomic Contigs for Microbial Physiology of Mixed Cultures. *Frontiers in Microbiology*, 3:410.

[Teeling et al., 2004] Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glöckner, F. O. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental microbiology*, 6(9):938–47.

[Wang et al., 2012a] Wang, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012a). MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. *Journal of computational biology : a journal of computational molecular cell biology*, 19(2):241–9.

[Wang et al., 2012b] Wang, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012b). MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics (Oxford, England)*, 28(18):i356–i362.

[Weber et al., 2011] Weber, M., Teeling, H., Huang, S., Waldmann, J., Kassabgy, M., Fuchs, B. M., Klindworth, A., Klockow, C., Wichels, A., Gerdts, G., Amann, R., and Glöckner, F. O. (2011). Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *The ISME journal*, 5(5):918–28.

[Wooley et al., 2010] Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS computational biology*, 6(2):e1000667.

[Wu and Ye, 2011] Wu, Y.-W. and Ye, Y. (2011). A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of computational biology : a journal of computational molecular cell biology*, 18(3):523–34.

[Zhang et al., 2011] Zhang, J., Wu, G., Hu, X., Li, S., and Hao, S. (2011). A Parallel K-Means Clustering Algorithm with MPI.