BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 4 2013, pages 511–512
doi:10.1093/bioinformatics/bts717

Systems biology

Advance Access publication December 24, 2012

Ultranet: efficient solver for the sparse inverse covariance selection problem in gene network modeling

Linnea Järvstråt[1], Mikael Johansson[2], Urban Gullberg[1] and Björn Nilsson[1,3,*]

[1]Department of Hematology and Transfusion Medicine, Lund University Hospital, SE-221 85 Lund, Sweden, [2]Department of Automatic Control, Royal Institute of Technology, SE-100 44 Stockholm, Sweden and [3]Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Graphical Gaussian models (GGMs) are a promising approach to identify gene regulatory networks. Such models can be robustly inferred by solving the sparse inverse covariance selection (SICS) problem. With the high dimensionality of genomics data, fast methods capable of solving large instances of SICS are needed.

We developed a novel network modeling tool, Ultranet, that solves the SICS problem with significantly improved efficiency. Ultranet combines a range of mathematical and programmatical techniques, exploits the structure of the SICS problem and enables computation of genome-scale GGMs without compromising analytic accuracy.

Availability and implementation: Ultranet is implemented in C++ and available at www.broadinstitute.org/ultranet.

Contact: bnilsson@broadinstitute.org or bjorn.nilsson@med.lu.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 3, 2012; revised on November 26, 2012; accepted on December 18, 2012

1 INTRODUCTION

Graphical Gaussian models (GGMs) are rapidly gaining traction as a promising approach to identify gene regulatory networks from genomics data (Markowetz and Spang, 2007; Schäfer and Strimmer, 2005b). GGMs are created by estimating the inverse Θ^* of the sample covariance matrix S, and then computing the partial correlation matrix $P^*_{ij} = \Theta^*_{ij}/\sqrt{\Theta^*_{ii}\Theta^*_{jj}}$. Assuming that the observations (typically gene expression values) have a multivariate Gaussian distribution, $P^*_{ij} = 0$ indicates that variables (genes) i and j are conditionally independent, given the other variables (i.e. all correlation between them can be explained by co-correlation with other variables); $P^*_{ij} \neq 0$ indicates that they are conditionally dependent (i.e., the correlation between them cannot be explained in full by co-correlation with other observed variables). The advantage of GGMs is that edges (non-zero P^*_{ij} s) are more likely to reflect direct dependencies between genes than edges in correlation networks (Markowetz and Spang, 2007; Schäfer and Strimmer, 2005a).

Alongside computing a relevant input matrix, the difficulty with GGM networks is to estimate Θ^*. A robust way to achieve this is to solve the penalized log-likelihood maximization problem

$$\Theta^* = \arg\max_{\Theta \succ 0} \log\det\Theta - \mathrm{tr}(S\Theta) - \lambda||\Theta||_1 \quad (1)$$

where $\lambda > 0$ controls the impact of the L^1 regularization term, and hence the sparsity of the network. This is called the sparse inverse covariance selection (SICS) problem. Solving SICS for high-dimensional genomics data is challenging because current standard methods, originally developed for lower-dimensional data, are associated with computational requirements that grow steeply with the number of genes. This leads to long wait times, increases the need for extraordinary computing resources and complicates analysis.

To address this issue, we developed an efficient network modeling tool, Ultranet, that solves the SICS problem significantly faster. As shown, Ultranet readily computes genome-scale GGMs with ~20 000 genes without compromising the analytic accuracy.

2 METHOD

As described in detail in Supplementary Information, Ultranet uses several different techniques to accelerate the computations.

First, as a preprocessing step, Ultranet identifies the connected components of the network by block-partitioning the adjacency matrix obtained by soft-thresholding S at λ. This decomposes the original problem into a set of smaller independent SICS problems. Because the time needed to solve SICS grows superlinearly with dimension, solving multiple small subproblems instead of one big problem is faster.

Second, to solve subproblems, Ultranet uses a first-order dual method based on accelerated projected-gradient descent (PGD). This method improves on earlier PGD approaches (Duchi *et al.*, 2008) by using simpler, yet effective, step-length selection rules and a heavy-ball momentum term to avoid zig-zagging. This eliminates most computations needed for step-length selection, and promotes smoother descent paths. The descent step is given by

$$W_{k+1} = \Pi[W_k + \alpha W_k^{-1} + \beta(W_k - W_{k-1})]_\lambda \quad (2)$$

where $W_k = \Theta_k^{-1}$, $k = 1, \ldots$ denotes the regularized covariance matrix at each iteration, and $\Pi[\,\cdot\,]_\lambda$ denotes projection onto the λ-box centered at S. The parameters α and β control the step length and impact of the momentum term, respectively.

Third, to initialize the solver, Ultranet uses a continuation approach where the problem is solved to low accuracy for decreasing λ values. This helps finding initial solutions near the optimum.

3 RESULTS AND DISCUSSION

We tested Ultranet on gene expression microarray data from various sources and of varying dimensionality. We made the

*To whom correspondence should be addressed.

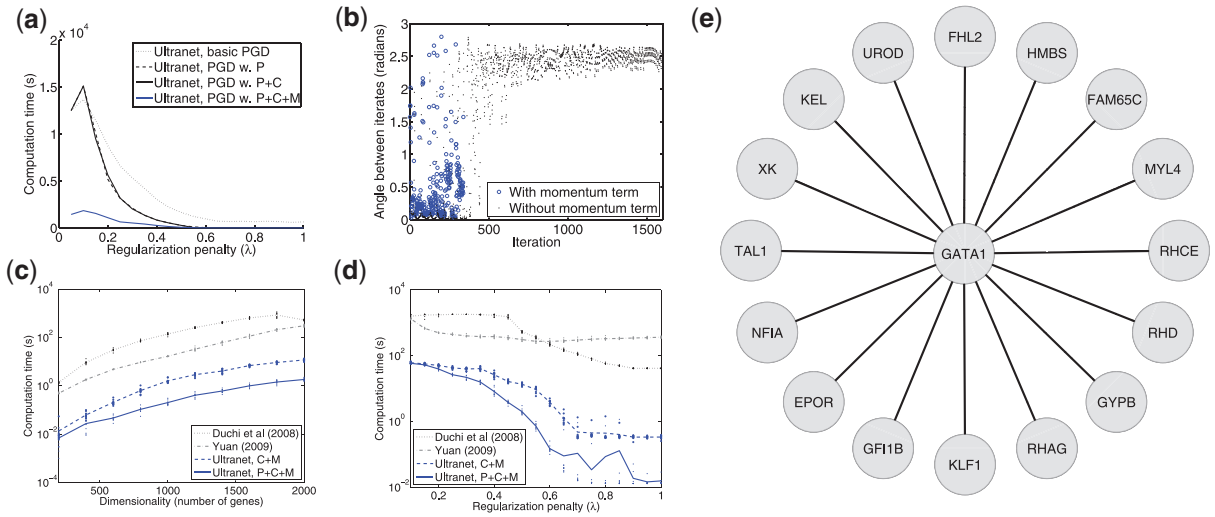

Fig. 1. Representative results for the Haferlach *et al.* (2010) dataset (2096 Affymetrix U133 Plus 2 arrays ×20 958 genes; Pearson correlations as S). (**a**) Computation time for basic PGD and PGD accelerated by block partitioning (P), continuation (C) and momentum term (M) (5000 random genes). (**b**) Angles between iterates ($\lambda = 0.2$; same genes). (**c**) Computation times for Ultranet versus control methods for varying λ (2000 random genes; average of 10 replicates). (**d**) Corresponding plot for varying S sizes ($\lambda = 0.5$). (**e**) Neighbors of *GATA1* for $\lambda = 0.8$ (computation time, 1.4 s with 20 958 genes). Normalized dual gap $< 10^{-3}$ used as stop criterion. Experiments performed on a dual Intel Xeon 5680 CPU w. 36 GB RAM. File access time not included

following recurrent observations (Fig. 1): first, adding block-partitioning to basic PGD (implemented in the same coding style) yielded substantial speed-ups for large λ; second, adding continuation yielded marginal speed-ups for intermediate λ in some datasets but did not yield significant speed-ups in general; third, in contrast, adding the momentum term accelerated convergence several fold for small λ because of dampened zig-zagging. Thus, we observed significant improvements over basic PGD across the full range of λ values.

We compared Ultranet with previous methods (Duchi *et al.*, 2008; Friedman *et al.*, 2008; Lu, 2009; Scheinberg *et al.*, 2010; Yuan, 2009). In our tests, the two fastest competitors were the methods by Duchi *et al.* (2008) (a PGD variant) and Yuan (2009) (an augmented Lagrangian method). Ultranet was 1–2 orders of magnitude faster than these methods without block partitioning, and 1–4 orders of magnitude faster with block partitioning (Fig. 1c and d).

We finally applied Ultranet to a large set of genome-wide gene expression profiles of bone marrow samples from a wide range of conditions (Haferlach *et al.*, 2010, NCBI Gene Expression Omnibus GSE13159) to estimate a draft regulatory network for human blood cell development. Examining the results, we focused on the neighborhood of *GATA1*, a transcription factor that drives the formation of red blood cells, a well-studied process. For all λ tested, the inferred *GATA1* neighborhood was clearly enriched for known red cell genes, including genes for transcription factors (*KLF1, GFI1B, TAL1*), blood groups (*RHD, RHAG, KEL, GYPB, XK*), heme synthesis enzymes (*UROD, HMBS*), and the erythropoietin receptor (*EPOR*) (Fig. 1e). Most neighbor genes contained *GATA1* ChIPseq (Chromatin Immuno-Precipitation followed by next-generation sequencing) peaks, supporting that they are true GATA-1 targets (Supplementary Table S1). The results illustrate the ability of GGMs to find biologically relevant networks, and the use of GGMs as a complement to motif searches and ChIP-seq to identify relevant transcription factor target genes. The results also illustrate Ultranet's ability to compute GGMs at a genome-wide scale.

In summary, Ultranet is a fast tool to infer gene networks based on GGMs. Ultranet is available for Microsoft Windows and Linux operating systems, exploits the multicore capacity of Intel x64 processors, and accepts data in tab-delimited text format.

Funding: This work was supported by the Swedish Foundation for Strategic Research (grant no. ICA08-0057), Marianne and Marcus Wallenberg's Foundation (grant no. 2010.0112) and the Swedish Children's Cancer Fund (grant no. 10/125).

Conflicts of Interest: none declared.

REFERENCES

Duchi,J.C. *et al.* (2008) Projected subgradient methods for learning sparse Gaussians. In: *Proceedings of the Conference Uncertainty in Artificial Intelligence*. Helsinki.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Haferlach,T. *et al.* (2010) Clinical utility of microarray-based gene expression profiling in the diagnosis and classification of leukemia. *J. Clin. Oncol.*, **28**, 2529–2537.

Lu,Z. (2009) Smooth optimization approach for sparse covariance selection. *SIAM J. Optim.*, **19**, 1807–1827.

Markowetz,F. and Spang,R. (2007) Inferring cellular networks—a review. *BMC Bioinformatics*, **8** (**Suppl. 6**), S5.

Scheinberg,K. *et al.* (2010) Sparse inverse covariance selection via alternating linearization methods. In: *Proceedings of Neural Information Processing Systems*. Vancouver.

Schäfer,J. and Strimmer,K. (2005a) An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–64.

Schäfer,J. and Strimmer,K. (2005b) A shrinkage approach to large-scale covariance matrix estimation and implications for genomics. *Stat. Appl. Genet. Mol. Biol.*, **4,** Article 32.

Yuan,X. (2009) Alternating direction methods for sparse covariance selection. *Optimization Online*, http://www.optimization-online.org/DB_FILE/2009/09/2390.pdf (17 January 2013, date last accessed).