BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 13 2011, pages 1876–1877
doi:10.1093/bioinformatics/btr274

Systems biology

Advance Access publication April 29, 2011

parmigene—a parallel R package for mutual information estimation and gene network reconstruction

Gabriele Sales[1] and Chiara Romualdi[2,]*

[1]Department of Statistical Sciences and [2]Department of Biology, University of Padova, 35121 Padova, Italy

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Inferring large transcriptional networks using mutual information has been shown to be effective in several experimental setup. Unfortunately, this approach has two main drawbacks: (i) several mutual information estimators are prone to biases and (ii) available software still has large computational costs when processing thousand of genes.

Results: Here, we present *parmigene* (PARallel Mutual Information estimation for GEne NEtwork reconstruction), an R package that tries to fill the above gaps. It implements a mutual information estimator based on k-nearest neighbor distances that is minimally biased with respect to the other methods and uses a parallel computing paradigm to reconstruct gene regulatory networks. We test *parmigene* on *in silico* and real data. We show that *parmigene* gives more precise results than existing softwares with strikingly less computational costs.

Availability and Implementation: The *parmigene* package is available on the CRAN network at http://cran.r-project.org/web/packages/.

Contact: chiara.romualdi@unipd.it

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

Received and revised on March 17, 2011; accepted on April 22, 2011

1 INTRODUCTION

Inferring gene regulatory networks (GRNs) is currently one of the most challenging task in systems biology. Several approaches have been proposed to reconstruct GRNs using experimental data (paradigm of reverse engineering approach) such as Bayesian Networks (BN) (Friedman, 2004), Relevance Networks (RN), (Butte *et al.*, 2000) and Graphical Gaussian Models (GGM) (Schafer and Strimmer, 2005); see Werhli *et al.* (2006) for a comparative review. While BN and GGM should be able to distinguish between direct and indirect edges, RN does not. However, BN and GGM hardly work in the case of thousands of genes and with a small number of replicates, while RN have the ability to deal with them. The seminal paper of Basso *et al.* (2005) extends RN introducing an algorithm based on the Data Processing Inequality (DPI) for removing indirect edges. Their approach, called ARACNE, has been successfully applied to reconstruct the subnetwork of the MYC gene in human B cells. At the same time, other methodologies have been proposed to remove indirect edges: the CLR and the MRNET algorithms; for additional details, see Meyer *et al.* (2008). Although successfully applied in the reconstruction of GRNs, RN based on MI estimation are still affected by two major drawbacks: (i) some MI estimators are biased (see Supplementary Material for details) and (ii) even if RN computational cost is comparatively lower than that of BN and GGM, the time required to reconstruct the GRN of several thousand genes is still extremely large.

Here, we present *parmigene* (PARallel Mutual Information calculation for GEne NEtwork reconstruction) a novel fast and parallel R package that (i) performs network inference implementing a minimally biased MI estimator, following Kraskov's algorithm (hereafter *knnmi*) (Kraskov *et al.*, 2004) based on k-nearest neighbor distances, and (ii) uses OpenMP to implement a parallel GRN reconstruction with a strikingly fast run times. In the following, we report (i) the bias obtained by *knnmi* compared with the other MI estimators, (ii) the precision obtained by *parmigene* in the reconstruction of the MYC subnetwork using real data (Basso *et al.*, 2005) and (iii) the computational costs for GRN reconstruction.

2 RESULTS

Bias: MI bias has been estimated for variables with low and with perfect non-linear associations. We compare knnmi (from *parmigene*), MI based on kernel density estimator [implemented in KDEMI, Qiu *et al.* (2009)] and variable discretizations implemented in Bioconductor packages bioDist and *minet* (Meyer *et al.*, 2008) (the shrinkage entropy estimator, the Miller-Madow and the Schurmann-Grassberger estimators, hereafter called shrink, mm and sg). For additional details see Supplementary Material. Figure 1 shows the bias distributions, while Supplementary Table S1 reports the average biases. *parmigene* with knnmi shows the lowest bias. Lower ks seem to perform better in both scenarios; in the following, we will thus report only the results of *parmigene* with $k=3$.

Precision: The impact of the MI estimation on GRN reconstruction has been verified using real data on human B cells (GEO ID GSE2350) (Basso *et al.*, 2005). The precision has been estimated using the subnetwork generated by the hub gene MYC for which several true interactors are known. Margolin *et al.* (2006) proposed the ARACNE algorithm with the DPI implemented using a multiplicative tolerance (τ). *minet* implements only an additive tolerance (ε) that gives worse levels of precision (Table 2). *parmigene* implements both alternatives. Table 1 reports precision and rank score (normalized average rank of true positives); values closer to 1 represent the best results (see Supplementary Material for details).

The ARACNE inference method with the multiplicative tolerance gives the highest precision. Furthermore, *parmigene* always

*To whom correspondence should be addressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

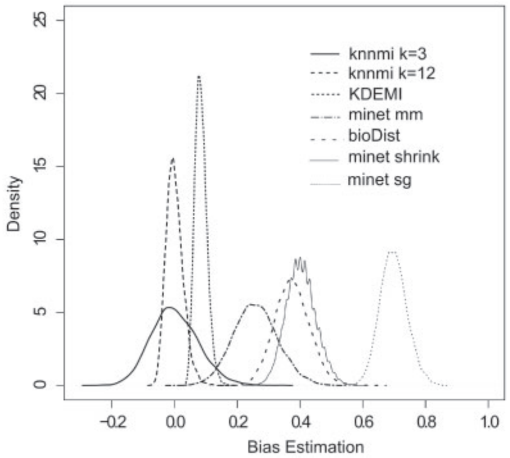

Fig. 1. Bias distribution of MI estimators in the case of Gaussian variables with low association levels.

Table 1. Precision estimation in the MYC subnetwork reconstruction

Algorithms	N	TP	Precision	Rank score
parmigene ARACNE $\tau = 0.15$	119	40	0.33	4
minet ARACNE[a] $\tau = 0.15$	1933	375	0.19	7
parmigene ARACNE $\varepsilon = 0.05$	56	16	0.29	4
minet ARACNE $\varepsilon = 0.05$	110	21	0.19	7
parmigene ARACNE $\varepsilon = 0.1$	308	66	0.21	6
minet ARACNE $\varepsilon = 0.1$	1079	191	0.18	7
parmigene CLR	6136	1083	0.17	7
minet CLR	5826	1029	0.17	7
parmigene MRNET	6462	1095	0.17	7
minet MRNET	6169	1047	0.17	7

See Supplementary Material for results using different combinations of ε and τ.
[a] MI matrix obtained with *minet* has been used as input for the ARACNE implementation in *parmigene*.
N, numebr of identified interactors;
TP, true positives.

outperforms the other estimators either for precision or for rank position.

Computational time: we tested the time required (i) to estimate MI and (ii) to infer GRNs with increasing number of genes. Table 2 shows the results. In the case of large-scale experiments (the entire gene expression matrix), *parmigene* computational time is strikingly better than those of exiting methods; it is 12 times faster in the MI estimation and 1.3 and 3.5 times faster, respectively, in the ARACNE and MRNET algorithms. Using parallel computations with a quad-core system, *parmigene* is 50 times faster than other methods for MI estimation.

Table 2. Computational time for MI estimation; knnmi uses $k = 3$

MI estimator	No. of genes	CPU time[a]	Wall clock time[a]
parmigene knnmi	100	0.005	0.006
minet mm	100	0.02	0.02
parmigene knnmi	1000	2.21	0.56
minet mm	1000	4.20	4.20
parmigene knnmi	8799	171.8	43
minet mm	8799	2120.7	2120.7

See Supplementary Material for computational times using different settings.
[a] Time calculated in minutes, test system: Intel(R) Core(TM)2 Quad Q9450 2.66 GHz, 8 Gb RAM, linux x86_64 2.6.36.

3 CONCLUSION

The MI estimator implemented in *parmigene* gives almost unbiased results and, more importantly, more precise results on real data. In addition, *parmigene* has strikingly fast execution times.

ACKNOWLEDGEMENTS

We acknowledge the CINECA Award N. HP10BDJ9X8, 2010 for the availability of high-performance computing resources and support.

Funding: University of Padova (CPDR070805 to G.S. and CPDR075919 to C.R.).

Conflict of Interest: none declared.

REFERENCES

Basso,K. *et al*. (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genetics*, **37**, 1061–4036.

Butte,A.J. *et al*. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. USA*, **97**, 12182–12186.

Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Kraskov,A. *et al*. (2004) Estimating mutual information. *Phys. Rev. E*, **69**, 066138.

Margolin,A. *et al*. (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7** (Suppl. 1), S7.

Meyer,P. *et al*. (2008) minet: a r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Qiu,P. *et al*. (2009) Fast calculation of pairwise mutual information for gene regulatory network reconstruction. *Comput. Methods Programs Biomed.*, **94**, 177–180.

Schafer,J. and Strimmer,K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.

Werhli,A.V. *et al*. (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, **22**, 2523–2531.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016