BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 15 2009, pages 1959–1960
doi:10.1093/bioinformatics/btp307

Genome analysis

RJaCGH: Bayesian analysis of aCGH arrays for detecting copy number changes and recurrent regions

Oscar M. Rueda*,† and Ramon Diaz-Uriarte*

Structural Biology and Biocomputing Programme, Spanish National Cancer Center (CNIO), Melchor Fernández Almagro 3, Madrid, 28029, Spain

Received on March 4, 2009; revised on April 20, 2009; accepted on April 30, 2009

Advance Access publication May 6, 2009

Associate Editor: John Quackenbush

ABSTRACT

Summary: Several methods have been proposed to detect copy number changes and recurrent regions of copy number variation from aCGH, but few methods return probabilities of alteration explicitly, which are the direct answer to the question 'is this probe/region altered?' RJaCGH fits a Non-Homogeneous Hidden Markov model to the aCGH data using Markov Chain Monte Carlo with Reversible Jump, and returns the probability that each probe is gained or lost. Using these probabilites, recurrent regions (over sets of individuals) of copy number alteration can be found.

Availability: RJaCGH is available as an R package from CRAN repositories (e.g. http://cran.r-project.org/web/packages).

Contact: rueda.om@gmail.com; rdiaz02@gmail.com

1 INTRODUCTION

Genomic DNA copy number alterations (CNAs) are associated with complex diseases (McCarroll and Altshuler, 2007), and are often studied using array-based comparative genomic hybridization (aCGH). To be immediately useful in both clinical and basic research scenarios, aCGH data analysis requires accurate methods that do not impose unrealistic biological assumptions and that provide direct answers to the key question, 'What is the probability that this gene/region has CNAs?' Estimates of the probabilities of alteration (instead of *P*-values or smoothed means) are the most direct and usable answer to this problem (Broët and Richardson, 2006). Probabilities can be used in contexts from basic research to clinical applications (Lockwood *et al*., 2006; Pinkel and Albertson, 2005) so that a clinician might require high certainty of alteration of a specific gene before invasive procedures, whereas a basic researcher can consider for further study genes that show only a moderate probability of alteration. In addition, many aCGH platforms have probes located at variable distances, which should be incorporated in the analysis (Broët and Richardson, 2006; Lockwood *et al*., 2006). A variety of methods have been developed for the analysis of aCGH data (see reviews in Lai *et al*., 2005; Rueda and Diaz-Uriarte, 2007a,b; Willenbrock and Fridlyand, 2005), but most of them do not return probabilities of alteration nor make use of the distance between probes. The few approaches that return probabilities of alteration either do not use distance between probes, or fix the number of possible states of alteration to three or four, a biologically unrealistic assumption. In addition to locating probes that show copy number changes, the identification of common or recurrent regions of alteration is one frequent study objective: the regions more likely to harbor disease-critical genes are those that are recurrent or common among samples (Diskin *et al*., 2006; Pinkel and Albertson, 2005). The identification of these regions should use the information about the probability of alteration (to avoid giving the same weight to probes with strong and weak evidence of alteration), and should allow the discovery of regions over subsets of samples as it is known that many complex diseases, such as cancer or autism, are composed of subtypes of syndromes (Sebat, 2007). Most available methods for locating common regions (Klijn *et al*., 2008; Rouveirol *et al*., 2006; Shah *et al*., 2007; Taylor *et al*., 2008), do not allow for among-subject heterogeneity nor use probabilities. Finally, many of the existing methods are not always readily and freely available like those on CRAN, or as easy to use without forcing (often arbitrary) choices on the user. We have developed a freely available R package, RJaCGH, for the analysis of aCGH data that incorporates distance between probes, returns probabilities of alteration and allows the identification of recurrent regions of CNA.

2 RESULTS

To estimate probabilities of copy number changes, we use a non-homogeneous Hidden Markov model (HMM) with an unknown number of hidden states fitted via Reversible Jump Markov Chain Monte Carlo (Cappé *et al*., 2005). By using a non-homogeneous HMM, we can account for the variable distance between probes/genes and Reversible Jump allows us to use HMMs without fixing the number of hidden states. By exploring the full posterior probabilities and retaining the probabilities of models of different sizes, we can employ Bayesian model averaging (Hoeting *et al*., 1999), thus incorporating model uncertainty and not conditioning our inferences to the selection of a particular model. The statistical model is described in Rueda and Diaz-Uriarte (2007a), where it is shown that the method performs as well as, or better than, the competing methods ACE (Lingjaerde *et al*., 2005), BioHMM (Marioni *et al*. 2006), HMM (Fridlyand *et al*., 2004), CGHseg (Picard *et al*., 2005), DNAcopy (Venkatraman and Olshen, 2007) and GLAD (Hupé *et al*., 2004) in terms of calling gains and losses, and

*To whom correspondence should be addressed.

†Present address: Breast Cancer Functional Genomics, Cancer Research UK, Cambridge, UK

the performance advantage increases as the variability in inter-probe distance increases.

For the identification of recurrent regions of CNA, we have developed two algorithms, **pREC-A** and **pREC-S** (fully described in the documentation of the program and as technical report from http://biostats.bepress.com/cobra/ps/art43/). **pREC-A** (probabilistic recurrent copy number regions, common threshold over all arrays) does not allow for among-subject heterogeneity and is, thus, similar in objectives to previous approaches except for the fact that we explicitly use probabilities. **pREC-S** (probabilistic recurrent copy number regions, subsets of arrays), identifies common regions over subsets of arrays; alternatively, we can think of this algorithm as identifying subsets of arrays that share regions of alteration. This is a novel algorithm, explicitly targeted to incorporate heterogeneity and use probabilities. Both methods use probabilities of alteration as returned by the non-homogeneous HMM. No hard thresholds are imposed, and thus the user decides what constitutes sufficient evidence (in terms of probability of alteration) to call a probe gained (or lost). The probabilities that we use are not the marginal probabilities of alteration but the joint probabilities of alteration of a region of probes. Our approach incorporates both within- and among-array variability: we use the information on the certainty of each call of gain/loss (i.e. the probability) in all computations of recurrent regions. Moreover, using probabilities of alteration (instead of magnitude of change), in addition to differentiating between evidence of alteration and estimated fold change, prevents inter-array differences in range of $\log_2$ ratios and tissue mixture to get confounded with evidence of alteration. Finally, both algorithms use at most two parameters and their biological meaning is immediate: probability of alteration, and number of samples that share an alteration. We can use the output of **pREC-S** as the basis for clustering and to display patterns of groupings of arrays; an example is shown in the documentation of the program.

The RJaCGH method has been implemented as an R package (R Development Core Team, 2006). All of the MCMC code for the HMM as well as the two algorithms for common regions have been implemented in C (dynamically loaded from R) for speed. The program is available from the standard R repositories (e.g. http://cran.r-project.org/web/packages/) under the GPL (v. 3) license and has been submitted to BioConductor. The package depends on no additional software (besides R itself). The flexibility and comprehensiveness of RJaCGH does have a computational cost: estimation of probabilities by RJaCGH is considerably slower than segmentation by alternative approaches. If probabilities of alteration are desired (but finding recurrent regions or incorporating distance between probes is not needed), the bcp method of (Erdman and Emerson, 2007, 2008) is a much faster alternative. **pREC-A** and **pREC-S**, once the probabilities have been obtained, are very fast (on the order of seconds to a few minutes for datasets that include 50–70 samples).

ACKNOWLEDGEMENTS

J. Fadista, A. Ivens and D. Grove for comments and bug report of the package. Three reviewers for comments on the ms.

Funding: Fundacion de Investigacion Medica Mutua Madrileña, RTIC COMBIOMED (RD07/0067/0014) Spanish Health Ministry, Supercomputacion y Ciencia (CSD2007-00050) Spanish Ministry of Education and Science, Spanish National Bioinformatics Institute (www.inab.org) a platform of Genoma España.

Conflict of Interest: none declared.

REFERENCES

Broët,P. and Richardson,S. (2006) Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics*, **22**, 911–918.

Cappé,O. *et al.* (2005) *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer, New York.

Diskin,S. *et al.* (2006) Stac: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res.*, **16**, 1149–1158.

Erdman,C. and Emerson,J. W. (2007) bcp: an R package for performing a Bayesian analysis of change point problems. *J. Stat. Softw.*, **23**, 1–13.

Erdman,C. and Emerson,J. W. (2008) A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, **24**, 2143–2148.

Fridlyand,J. *et al.* (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivar. Anal.*, **90**, 132–153.

Hoeting,J.A. *et al.* (1999) Bayesian model averaging: a tutorial (with discussion). *Stat. Sci.*, **14**, 382–417.

Hupé,P. *et al.* (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Klijn,C. *et al.*(2008) Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res.*, **36**.

Lai,W.R.R. *et al.* (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Lingjaerde,O.C. *et al.* (2005) CGH-explorer: a program for analysis of array-CGH data. *Bioinformatics*, **21**, 821–822.

Lockwood,W.W. *et al.* (2006) Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur. J. Hum. Genet.*, **14**, 139–148.

Marioni,J.C. *et al.* (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

McCarroll,S.A. and Altshuler,D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**(Suppl. 7), S37–S42.

Picard,F. *et al.* (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Pinkel,D. and Albertson,D. G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37 (Suppl.)**, S11–S17.

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria, ISBN 3-900051-07-0.

Rouveirol,C. *et al.* (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 2066–2073.

Rueda,O.M. and Diaz-Uriarte,R. (2007a) Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Comput. Biol.*, **3**, e122.

Rueda,O.M. and Diaz-Uriarte,R. (2007b) A response to Yu et al. 'a forward-backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (SNP) array', BMC bioinformatics 2007, 8: 145. *BMC Bioinformatics*, **8**, 394.

Sebat,J. (2007) Major changes in our dna lead to major changes in our thinking. *Nat. Genet.*, **39**, S3–S5.

Shah,S. *et al.* (2007) Modeling recurrrent CNA copy number alterations in array CGH data. *Bioinformatics*, **23**, i450–i458.

Taylor,B.S.S. *et al.* (2008) Functional copy-number alterations in cancer. *PLoS ONE*, **3**.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.