

Genome analysis

CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms

Weil Lai¹, Vidhu Choudhary¹ and Peter J. Park^{1,2,*}¹Harvard-Partners Center for Genetics and Genomics and ²Informatics Program, Children's Hospital, Boston, MA 02139, USA

Received on November 23, 2007; revised on January 21, 2008; accepted on February 15, 2008

Advance Access publication February 22, 2008

Associate Editor: Keith Crandall

ABSTRACT

Summary: Accurate estimation of DNA copy numbers from array comparative genomic hybridization (CGH) data is important for characterizing the cancer genome. An important part of this process is the segmentation of the log-ratios between the sample and control DNA along the chromosome into regions of different copy numbers. However, multiple algorithms are available in the literature for this procedure and the results can vary substantially among these. Thus, a visualization tool that can display the segmented profiles from a number of methods can be helpful to the biologist or the clinician to ascertain that a feature of interest did not arise as an artifact of the algorithm. Such a tool also allows the methodologist to easily contrast his method against others.

We developed a web-based tool that applies a number of popular algorithms to a single array CGH profile entered by the user. It generates a heatmap panel of the segmented profiles for each method as well as a consensus profile. The clickable heatmap can be moved along the chromosome and zoomed in or out. It also displays the time that each algorithm took and provides numerical values of the segmented profiles for download. The web interface calls algorithms written in the statistical language R. We encourage developers of new algorithms to submit their routines to be incorporated into the website.

Availability: <http://compbio.med.harvard.edu/CGHweb>**Contact:** peter_park@harvard.edu**1 INTRODUCTION**

Array comparative genomic hybridization (CGH) is a technique for genome-wide measurement of the DNA copy number on a microarray (Pinkel *et al.*, 1998). With the availability of high-resolution tiling arrays, variations in the copy number can be captured with an unprecedented accuracy. This technology is most often used to characterize chromosomal instability in the cancer genome, but recent work on 270 individuals from four populations (HapMap collection) has found that natural copy number polymorphisms also exist to a much greater extent than expected (Redon *et al.*, 2006).

A successful array CGH experiment requires several components. First, it is important to obtain a homogeneous sample of interest with an appropriate control. Given the large number of

copy number polymorphisms, getting a normal sample from the same patient is ideal. For tumors, it is often difficult to ascertain whether a 'normal' sample often obtained from a nearby location is truly normal; in this case, DNA from the blood may have to be used. Second, the hybridization experiment must be carried out properly, on arrays of sufficient resolution. For instance, BAC-based arrays may not be sufficient, if the goal is to detect small alterations. The last and frequently the most difficult component is the statistical analysis and interpretation of the resulting data.

2 RESULTS

The main issue in analysis is to segment the sequence of log-ratios along the chromosome into regions of amplification, deletion or no change. There has been extensive work in this field, with many methods derived from existing techniques in other fields. For instance, the segmentation problem can be reformulated as a change-point problem in statistics (Olshen *et al.*, 2004) or an optimization problem in engineering to be solved by dynamic programming (Autio *et al.*, 2003). Given a wide range of choices, comparative analysis of these methods has been useful for the practitioner who must decide among all the choices (Lai *et al.*, 2005, Willenbrock and Fridlyand, 2005). However, a choice based on such a study does not guarantee that the algorithm being applied is the most appropriate one for a specific dataset—it is possible that the feature that the user sees in his data may be an artifact of that particular algorithm.

The web-based tool we developed alleviates this problem. It takes an input profile from the user and applies up to 10 different algorithms (Fig. 1). The resulting profiles are returned in a heatmap for easy comparison (Fig. 2). The user can then see whether a particular aberration that he is interested in pursuing further has been found by other algorithms as well. Other features of the software include the following: a heatmap display of gain/loss determined by user-defined cut-off; a consensus profile (average of all segmented profiles); a tabular summary of the aberrations found; example datasets from BAC, Agilent and Nimblegen arrays; a bargraph displaying the time taken by each algorithm; buttons to zoom in/out and move along the chromosome; clickable map that takes the user to the UCSC genome browser for a specific region and a zipped file containing predicted values at all probes for download. A few web-based interfaces are available for array CGH data,

*To whom correspondence should be addressed.

The screenshot shows the CGHweb interface. At the top, there's a logo and a tagline: 'Enabling users to analyze their array-CGH data with multiple algorithms simultaneously.' Below this, the 'Select algorithm(s) to use' section has three radio buttons: 'Select our 3 preferred algorithms', 'Select all algorithms', and 'Clear selected algorithms'. There are several checkboxes for different algorithms: BAC, Agilent, Nimblegen, Circular Binary Segmentation, Forward-Backward Fragment-Annealing Segmentation, Fused Lasso (qf/Lasso), Gaussian Model with Adaptive Penalty (Pisard et al.), GLAD, Locally weighted scatterplot smoother, Wavelet smoothing, Quantile Smoothing, and Running Average. Each algorithm has associated parameters that can be tuned, such as Alpha, Delta, FDR, Rho, Rho2, Rho3, Rho4, Rho5, Rho6, Rho7, Rho8, Rho9, Rho10, Rho11, Rho12, Rho13, Rho14, Rho15, Rho16, Rho17, Rho18, Rho19, Rho20, Rho21, Rho22, Rho23, Rho24, Rho25, Rho26, Rho27, Rho28, Rho29, Rho30, Rho31, Rho32, Rho33, Rho34, Rho35, Rho36, Rho37, Rho38, Rho39, Rho40, Rho41, Rho42, Rho43, Rho44, Rho45, Rho46, Rho47, Rho48, Rho49, Rho50, Rho51, Rho52, Rho53, Rho54, Rho55, Rho56, Rho57, Rho58, Rho59, Rho60, Rho61, Rho62, Rho63, Rho64, Rho65, Rho66, Rho67, Rho68, Rho69, Rho70, Rho71, Rho72, Rho73, Rho74, Rho75, Rho76, Rho77, Rho78, Rho79, Rho80, Rho81, Rho82, Rho83, Rho84, Rho85, Rho86, Rho87, Rho88, Rho89, Rho90, Rho91, Rho92, Rho93, Rho94, Rho95, Rho96, Rho97, Rho98, Rho99, Rho100. The 'Specify plotting parameters' section has a 'Minimum Log Ratio' field set to 0.2 and a 'Threshold' field set to 0.2. There's also an 'Email Address (required)' field and a 'Data File' section with 'Upload your data file' and 'Upload example data file' buttons. At the bottom, there are 'Restore default settings' and 'Submit Query' buttons.

Fig. 1. This front page lists the algorithms along with parameters that can be tuned. Examples from BAC, Agilent and Nimblegen data can be uploaded easily.

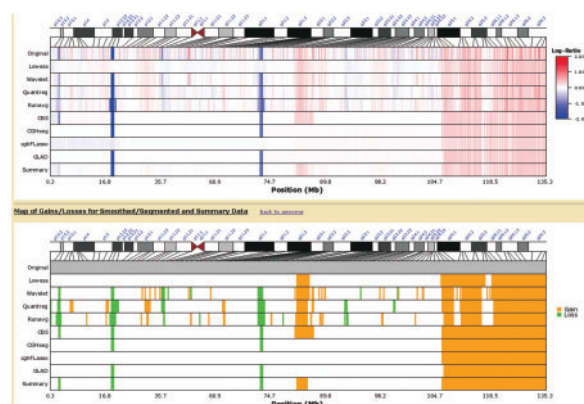


Fig. 2. Results page (zoomed in to a small region). Output from all of the algorithms are shown in a heatmap (top panel); gain and loss are called using a user-specified threshold (bottom panel). Discrepancies among different algorithms can be easily detected.

including CAPweb (Liva *et al.*, 2006), ISACGH (Conde *et al.*, 2007) and Asterias (Diaz-Urriarte *et al.*, 2007), but CGHweb provides the most comprehensive list of algorithms and a convenient interface for navigating through the results.

To make this tool a useful resource for the community, our goal is to incorporate as many algorithms as we can. Because it is not possible for any one group to implement all the algorithms, we have defined a function call with a set of arguments (details available on the website) and we encourage developers of new algorithms to create and functions according to this specification. We have chosen the R language because it is most widely used for microarray analysis and wrappers can be written easily for routines in C. Source code is available for those interested in local installation.

3 DISCUSSION AND CONCLUSION

Some analytical issues have not been fully resolved in the literature. For instance, it is well-known from gene expression

studies that log-ratios derived from low intensity signals are unreliable and that a local variance correction can ameliorate this problem (Colantuoni *et al.*, 2002). Few CGH algorithms account for this in the segmentation process. The effect of spatial smoothing applied in combination with segmentation also has not been carefully explored. CGHweb, however, leaves it to the user and the algorithm to make any desired transformation of the data. Deriving a consensus profile from multiple samples is also an important issue (e.g. Engler *et al.*, 2006; Diskin *et al.*, 2006), but that area is less developed and is not addressed here beyond simple pointwise averaging. Recently, an algorithm based on pointwise averaging was shown to have good performance (Beroukhim *et al.*, 2007). This suggests that pointwise averaging may provide a reasonable solution for balancing the importance of amplitude and frequency of alterations.

The CGHweb interface collects results from multiple algorithms and allows developers to submit their new algorithms. This site makes it possible for the user who is not familiar with programming to ascertain a segmentation profile via multiple methods. It also facilitates comparison of a novel method to the existing ones, thus setting a higher standard to which previously untested methods should be measured.

ACKNOWLEDGEMENTS

This work was funded by NIH through the Cancer Genome Characterization Center grant (1U24 CA126554).

Conflict of Interest: none declared.

REFERENCES

- Autio, R. *et al.* (2003) CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics*, **19**, 1714–1715.
- Beroukhim, R. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA*, **104**, 20007–20012.
- Colantuoni, C. *et al.* (2002) SNOMAD (Standardization and Normalization of MicroArray Data): web-accessible gene expression data analysis. *Bioinformatics*, **18**, 1540–1541.
- Conde, L. *et al.* (2007) ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling. *Nucleic Acids Res.*, **35**(Web Server issue), 81–85.
- Diaz-Urriarte, R. *et al.* (2007) Asterias: integrated analysis of expression and a CGH data using an open-source, web-based, parallelized software suite. *Nucleic Acids Res.*, **35**(Web Server issue), 75–80.
- Diskin, S.J. *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.
- Engler, D.A. *et al.* (2006) A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, **7**, 399–421.
- Lai, W.R. *et al.* (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.
- Liva, S. *et al.* (2006) CAPweb: a bioinformatics CGH array Analysis Platform. *Nucleic Acids Res.*, **34**(Web Server issue), 477–481.
- Olshen, A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
- Pinkel, D. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.
- Redon, R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
- Willenbrock, H. and Fridlyand, J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.