# Chromosomal Breakpoint Detection in Human Cancer

Kees Jong[1], Elena Marchiori[1], Aad van der Vaart[1], Bauke Ylstra[2], Marjan Weiss[2], and Gerrit Meijer[2]

[1] Department of Mathematics and Computer Science
Free University Amsterdam, The Netherlands
cjong,elena,aad@cs.vu.nl
[2] VU University Medical Center
Free University Amsterdam, The Netherlands
b.ylstra,ga.meijer,mm.weiss@vumc.nl

**Abstract.** Chromosomal aberrations are differences in DNA sequence copy number of chromosome regions [3]. These differences may be crucial genetic events in the development and progression of human cancers. Array Comparative Genomic Hybridization is a laboratory method used in cancer research for the measurement of chromosomal aberrations in tumor genomes. A recurrent aberration at a particular genome location may indicate the presence of a tumor suppressor gene or an oncogene. The goal of the analysis of this type of data includes detection of locations of copy number changes, called breakpoints, and estimate of the values of the copy number value before and after a change. Knowing the exact locations of a breakpoint is important to identify possibly damaged genes. This paper introduces genetic local search algorithms to perform this task.

## 1 Introduction

Array Comparative Genomic Hybridization (array-CGH) is an approach for genome-wide scanning of differences in chromosomal copy numbers. Normal human cells contain two copies of each of the 22 non-sex chromosomes. In tumor cells one or both copies of parts of chromosomes may be deleted or duplicated. Chromosomal copy numbers are defined to be 2 for normal cells, 1 or 0 for single and double deletions and 3 and higher for single copy gains and higher level amplifications. Ideally the purpose of array-CGH is to construct a graph of the copy numbers for a selection of clones (normal mapped chromosomal sequences, i.e. small pieces of DNA) as a function of position of the clone on the genome. DNA copy-number aberrations are used in cancer research, for instance, by searching for novel genes implicated in cancer by analyzing those genes located in regions

[3] Aberrations can occur without change of copy number, but these aberrations are not the subject of this paper.

with abnormal copy numbers. It is therefore of fundamental relevance to identify as precise as possible chromosomal regions with abnormal copy numbers.

Because copy numbers cannot be measured directly, tumor cells are compared to normal cells. A large number (up to 2500) of clones are printed on a glass slide (micro array), which is next treated with a mixture of DNA originating from tumor and normal cells, both cut into fragments. Before applying the DNA mixture to the micro array the two types of DNA are labelled red (Cy5) and green (Cy3), respectively. The labelled fragments hybridize ("stick") to a spot on the array with a matching DNA sequence. The measured red/green ratio for each of the spots on the array is roughly proportional to the quotient of copy numbers for tumor and normal tissue. This experiment is repeated for a number of tumors. A more elaborate introduction to array-CGH can be found in [5].

Unfortunately, a sample of cells taken from a tumor will generally consist of multiple cell types, which may differ in their chromosomal copy numbers. In particular, the sample usually consists of tumor cells and admixed normal cells. In some cases the sample may also contain tumor cells that are intermediate in the development of the tumor and have fewer copy number changes. It is assumed that the actual tumor cells occur by far the most in the sample. In our case the experimental samples were selected by the pathologist to have more than 70% tumor cells. The values found in the experiment then represent the copy numbers of the actual tumor cells plus some "noise" generated by the normal cells and some experimental noise.

When the observed relative copy numbers (or their logarithms) for the clones are ordered by location on the genome, the values form "clouds" with different means, supposedly reflecting different levels of copy numbers. We introduce a "smoothing" algorithm that tries to adjust the observed array-CGH values such that they represent the copy number of the most common tumor cells. That is, the algorithm tries to set the values to the means of the "clouds". Since the copy number of a clone is always quite small (normally 2, varying from 0 to about 10), we would like to set means of "clouds" that are close to the same value, because they represent the same copy number. Next we also want the number of value changes ("breakpoints") to be small.

The problem can be formalized as model fitting to search for most-likely-fit model given the data. A model describes a number of breakpoints, a position for each, and parameters of the distribution of copy number for each. Then one has to estimate the real parameters of the model from the observed array-CGH values.

We assume that the data are generated by a Gaussian process and use the maximum likelihood criterion for measuring the goodness of a partition, adjusted with a penalization term for taking into account model dimension. We introduce a local search procedure that searches for a most probable partition of the data using $N$ breakpoints, for a given $N$. The procedure is incorporated into a genetic algorithm that evolves a population of partitions with possibly different number of breakpoints that may vary during execution. We design two

algorithms based on this approach. The first one is a genetic local search algorithm that iteratively selects two 'good' chromosomes, generates two offspring using uniform crossover, applies mutation and the local search procedure to the offsprings and inserts them in the (worst chromosomes of the) population. The second algorithm generates only one offspring, applies local search and tries to further optimize the offspring with an ad-hoc procedure.

We analyze the performance of these algorithms on array-CGH measurements for 9 gastric cancer tumors. For each chromosome of a tumor, we compare the smoothing of the two GAs, the multi-start `LS` and the SA algorithm. The best algorithm we compare with the expert.

## 2 Breakpoint Detection

In our CGH experiments copy numbers are measured for approximately 2200 clones spread along the genome. We apply our algorithm to each of the 23 chromosomes separately. Denote by $x_1, \ldots, x_n$ the measured CGH values for a given chromosome. The main goal is to cluster these values in a small number of clusters $(x_1, \ldots, x_{y_1}), (x_{y_1+1}, \ldots, x_{y_2}), \ldots, (x_{y_N+1}, \ldots, x_n)$ such that the copy numbers of the clones in each cluster are identical. We refer to the indices $y_0 = 0 < y_1 < \cdots < y_N < n = y_{N+1}$ as breakpoints.

Our algorithm is motivated by the working hypothesis that the measured value $x_j$ is equal to the relative copy number of clone $j$ plus random noise that is independent across clones. Thus our model stipulates that for $y_{i-1} < j \leq y_i$ the observed CGH value $x_j$ can be considered as drawn from a normal distribution with mean $\mu_i$ and variance $\sigma_i^2$ particular to the $i$th cluster. This leads to the likelihood function

$$\prod_{i=1}^{y_1} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2} \ldots \prod_{i=y_N+1}^{n} \frac{1}{\sigma_{N+1}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_{N+1}}{\sigma_{N+1}}\right)^2}$$

The maximum likelihood estimators are the parameter values for which this expression is maximal. Given breakpoints $y_0 = 0 < y_1 < \cdots < y_N < n = y_{N+1}$ the maximization relative to the $\mu_i$ and $\sigma_i^2$ is equivalent to performing maximum likelihood estimation on each of the samples $x_{y_{i-1}+1}, \ldots, x_{y_i}$ separately, which leads to the usual estimates

$$\hat{\mu}_i = \frac{1}{y_i - y_{i-1}} \sum_{j=y_{i-1}+1}^{y_i} x_j, \qquad \hat{\sigma}_i^2 = \frac{1}{y_i - y_{i-1}} \sum_{j=y_{i-1}+1}^{y_i} (x_j - \hat{\mu}_i)^2.$$

Reinserting these values into the likelihood we are, after some simplification, left with

$$\frac{1}{\hat{\sigma}_1^{y_1}\sqrt{2\pi}^{y_1}} e^{-1/2} \cdots \frac{1}{\hat{\sigma}_{N+1}^{n-y_N}\sqrt{2\pi}^{n-y_N}} e^{-1/2}.$$

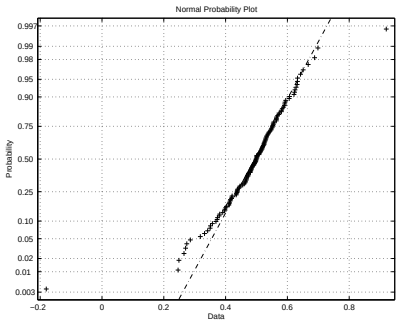
The next step is to find suitable breakpoints by maximizing this relative to $y_1, \ldots, y_N$. Equivalently, we minimize minus the logarithm, which up to an additive constant is equal to

$$\sum_{i=1}^{N+1} (y_{i+1} - y_i) \log \hat{\sigma}_i$$

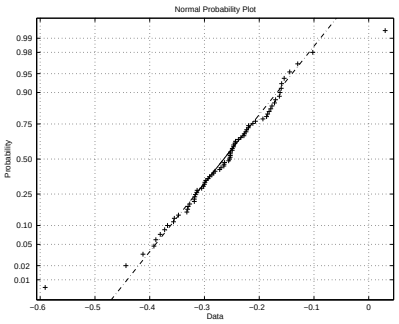
Note here that the $\hat{\sigma}_i$ in this expression also depends on the choice of $y_1, \ldots, y_N$. However, it is obvious that the highest value of the likelihood is obtained by choosing the highest possible number of breakpoints, as this gives more flexibility in choosing the parameters $\mu_i$ and $\sigma_i$. The last minimization step is therefore not well defined. We remedy this by adding a penalty to the criterion, in order to discourage a large number of breakpoints. A simple penalty of the form $\lambda N$, for $\lambda$ a suitable constant, performed well in our experiments. This leads to the following function to be minimized.

$$f(y_1, \ldots, y_N) = \sum_{i=1}^{N+1} (y_{i+1} - y_i) \log \hat{\sigma}_i + \lambda N \tag{1}$$

If we consider there to be $3N$ parameters ($2N$ continuous parameters and $N$ breakpoints), then the choices $\lambda = (3/2) \log n$ and $\lambda = 3$ correspond to Bayesian information criterion [7] and Akaike information criterion [1], respectively. In our experiments the choice $\lambda = 10$ was appropriate.




Chrm 8, tumor tissue, a gain




Chrm 14,tumor tissue, a loss

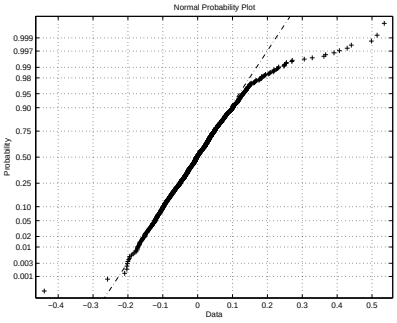
The assumptions of normality and independence may be slightly violated, as illustrated by the normal probability plots [6] (data are normally distributed when they lay near the dotted line). This holds also for normal tissues, as shown by the corresponding plot.

Nevertheless, in our experiments the resulting criterion gives adequate results.

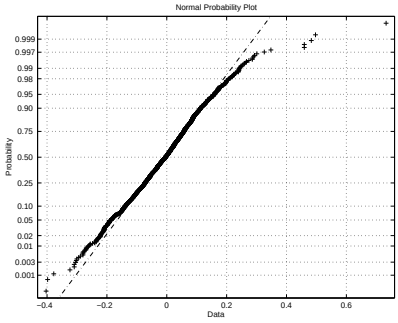
The most obvious violations of the normality assumption are caused by amplifications. Results can be further improved if the clones in amplification areas

are removed from the data. However, it is not easy to define an amplification unambiguously. An amplification area starts with a "big" increase of CGH value, last for only a "few" clones, after which the CGH values decrease "steeply". The number of clones for which an amplification lasts at most is a parameter. The increase and decrease of the value that are at least necessary to form an amplification depend on all value changes between consecutive clones. Say the average value change is $\overline{d}$ and the standard deviation of the value changes is $s_d$. Then the criterion is $\frac{d_i - \overline{d}}{s_d} \geq T$, where $T$ is a parameter.

In the sequel, we do not apply pre-processing for dealing with amplifications.




Chrms 1 to 22, tumor tissue, all normal




Chrms 1 to 22, normal tissue

## 3 The Genetic / Local Search Algorithms

The local search algorithm takes as input the CGH data $l = x_1 \dots x_n$ of one chromosome, a number $N$ of randomly generated indices $y_1, \dots, y_N \in [1, n]$ indicating potential (locations of) breakpoints, and updates repeatedly the breakpoints (locations) in order to minimize the function $f$ given in (1), where the first term is the negative log-likelihood of the data and the second one is a penalization with parameter $\lambda$ which punishes partitions containing many breakpoints. The algorithm uses $f$ as scoring function. At every iteration an update rule is applied to each breakpoint, selected randomly. The update rule chooses randomly a direction (left or right) and moves the breakpoint location of one position in that direction only if the move improves the scoring value (that is if $f$ decreases), otherwise it moves the breakpoint of one place in the opposite direction if this yields an improvement. The iterative process terminates when the application of the update rule to each breakpoint does not improve the scoring. We call this algorithm `LS`. We use `LS` in a multi-start local search algorithm, and as local optimizer in the two heuristic algorithms described in the sequel.

Genetic local search algorithms, also called memetic algorithms [4], use local search for optimizing the population after the application of the genetic operators. So at each iteration of the evolutionary process the population consists of a

set of local optima. We introduce the two memetic algorithms illustrated below for identifying breakpoints in array-CGH data of a chromosome, called `GLS` and `GLSo`, respectively.

In order to avoid confusion, in the sequel we say 'individual' instead of the standard genetic algorithms term 'chromosome' for indicating an element of the population.

```
GLS
{
 generate initial population
 while (termination criterion not satisfied)
 {
   select two parents from population using roulette wheel
   generate offsprings using uniform crossover
   apply mutation to each offspring
   apply LS to each offspring
   replace two worst individuals of population with offsprings
  }
}
```

```
GLSo
{
 generate initial population
 while (termination criterion not satisfied)
 {
   select two parents from population using roulette wheel
   generate offspring using OR crossover
   apply LS to offspring
   apply JOIN to offspring
   replace worst individual of population with offspring
  }
}
```

Our genetic algorithms use a representation where an individual is a bit string denoting chromosome locations with a 1 in each location containing a breakpoint and a 0 elsewhere. The fitness function to be minimized is the score function (1). The initial population is constructed as follows. For each $N$ in a fixed range, a number $k$ of elements is generated, where an element is a bit string with $N$ 1's randomly placed. The local search `LS` is applied to each individual.

`GLS` uses (blind) uniform crossover, while mutation randomly decides whether to add or remove a breakpoint and then applies the chosen operation (that is flipping the value of the selected individual location). The 'remove' operation consists of removing the breakpoint that yields the best fitness function score. Note that this operation is applied even if it does not decrease the fitness of the individual. The 'add' operation selects the segment (a region between two consecutive ones) with relative chromosonal array-CGH region (set of clones

values) having the highest standard deviation, and places a breakpoint in the middle of that region.

The termination criterion is satisfied when either a maximum number of iterations is reached or when the fitness of the best individual does not decrease and there is no pair of corresponding clones in the population having a difference in smoothed value of more than 0.01. The smoothed value of a clone is the mean value of the (chromosomal array-CGH region corresponding to the) segment containing that clone.

`GLSo` generates one offspring per iteration by selecting two individuals and constructing one offspring by taking the union of their breakpoints (by performing a bitwise OR of the two individuals). Then the offspring is optimized using `LS` and further optimized by removing breakpoints using the `JOIN` procedure. The `JOIN` procedure repeatedly selects the breakpoint whose removal yields the biggest improvement (decrease) of the fitness function, and continues until the fitness does not decrease anymore.

## 4 Experimental Results

Genomic DNA was isolated from snap-frozen tumor samples taken from gastrectomy specimens. The samples were obtained from the archives of the department of Pathology of the VU University Medical Center. Array-CGH experiments were performed according to [8] and ratio measurements according to [3]. The scanning array comprised DNA from 2275 BAC and P1 clones spotted in triplicate, evenly spread across the whole genome at an average resolution of 1.4 Mb. Chromosome X-clones were discarded from further analysis since all tumor samples were hybridized to male reference DNA, leaving 2214 clones per array to be evaluated. Each clone contains at least one STS for linkage to the sequence of the human genome. These data is analyzed in [9].

The 9 tumors used to test our method are all gastric tumors. A manual smoothing for these tumors, carried out by the expert B. Ylstra, is used to assess the performance of the algorithms and the maximum likelihood function as approximation for the expert. We run our algorithms on each chromosome of these 9 tumors, for a total of 207 chromosomes containing an average of about 100 clones.

The following GA parameter setting is chosen. The initialization generates 40 individuals containing $N$ breakpoints, with $N$ that varies from 1 to 10. An individual with 0 breakpoints is also added, thus giving a total of 401 individuals. The maximum number of iterations allowed is 100000. Crossover and mutation rates are equal to 1.

The multi-start `LS` performs 100000 plus 1 runs, where the number $N$ of breakpoints varies from 1 to 20, with an equal number of runs assigned to each value of $N$. Also a run with 0 breakpoints is done. The final result is the solution with best score over the runs.

We compare the performance of GLSo, GLS, multi-start LS, and a multi-start variant of LS based on simulated annealing (SA). The annealing schedule of SA is as follows. The starting temperature is 100000. After 10000 changes of breakpoint location it cools down to 0.00001. After each change the actual temperature is divided by $10^{10^{-3}}$. After 10000 changes we make the algorithm behave exactly like LS. The other settings are similar to the multi-start LS, except that it only performs 2001 runs to make the comparison more fair in terms of computation time.

We compare the performance of the four algorithms in minimizing the function (1) by the median and mean values obtained for the $9 \times 23 = 207$ chromosomes in our gastric tumors. In the tables below mLS and mSA denote multi-start LS and SA, respectively. As shown in the following table method GLS performs best according to this criterion followed by the second genetic algorithm GLSo.

| Algorithm | Median | Mean |
|---|---|---|
| mLS | -192.99 | -218.77 |
| mSA | -193.29 | -220.07 |
| GLSo | -194.47 | -220.83 |
| GLS | -196.00 | -223.08 |

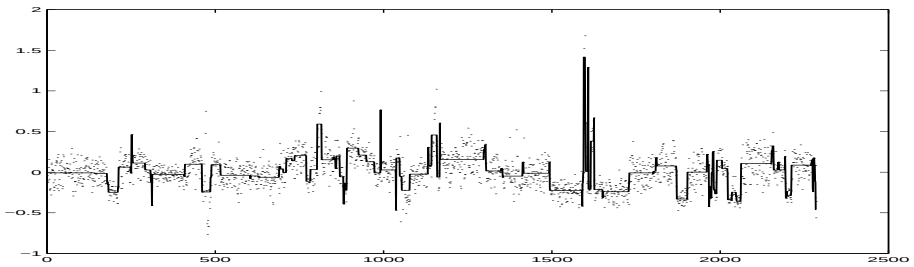
At closer inspection the nature of the differences in performance of the four algorithms vary considerably over the 207 chromosomes. For 22 chromosomes all four algoritms yield an identical smoothing, and for as many as 76 chromosomes the smoothings produced by at least three of the four algorithms are identical. For this reason we also investigated the differences in fitness for the 207 chromosomes for each pair of algorithms. Medians, means, 20% trimmed means, and the number of chromosomes with identical smoothing are given in the following table, together with the p-values of the sign test. The latter test (cf. [6]) indicates if the observed differences (in obtained fitness values) between the pairs of algorithms are statistically significant, under the assumption that the 207 chromosomes can be considered a random sample of chromosomes.

| Algorithm | Median | Mean | Trimmed Mean | # Zeros | P-value |
|---|---|---|---|---|---|
| mLS-mSA | 0.00 | 1.30 | 0.52 | 67 | 0 |
| mLS-GLSo | 0.00 | 2.06 | 0.22 | 50 | 0.63 |
| mLS-GLS | 0.70 | 4.31 | 1.84 | 81 | 0 |
| mSA-GLSo | 0.00 | 0.76 | -0.08 | 31 | 0.47 |
| mSA-GLS | 0.79 | 3.01 | 1.66 | 62 | 0 |
| GLSo-GLS | 0.24 | 2.25 | 1.50 | 48 | 0 |

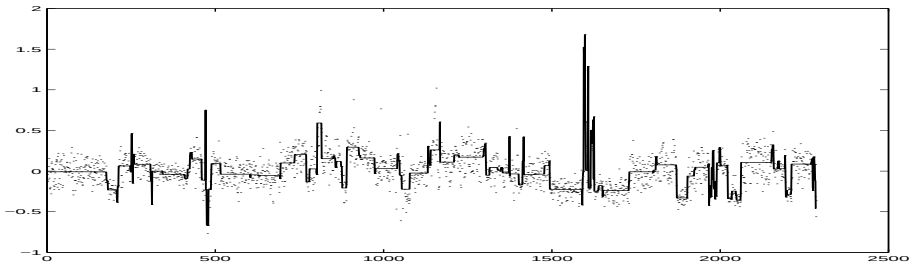
From these numbers we conclude again that GLS is best, followed by GLSo, mSA and mLS. Furthermore, the superior performance of GLS is statistically significant, whereas the observed differences between the other methods may not be replicable on data of additional tumors.

To illustrate the results of the four algorithms we show below pictures of the smoothing/breakpoints found by each of algorithms on one of the tumors in

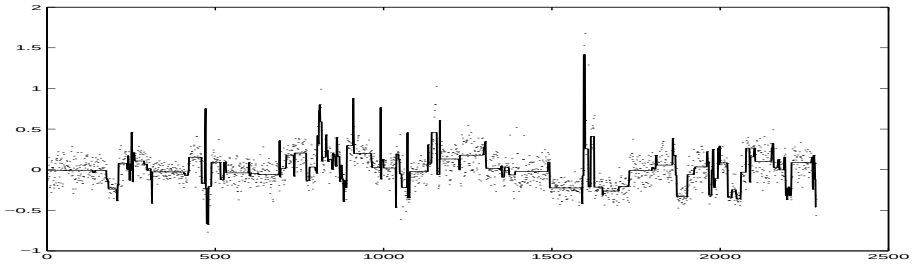
which various types of chromosomal aberrations (gain, loss and amplification) occur.



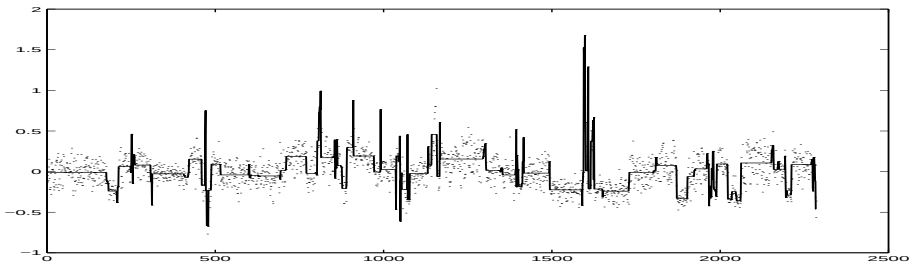

Tumor 2008c, dots are raw data, line is result of multi-start LS.




Tumor 2008c, dots are raw data, line is result of multi-start SA.



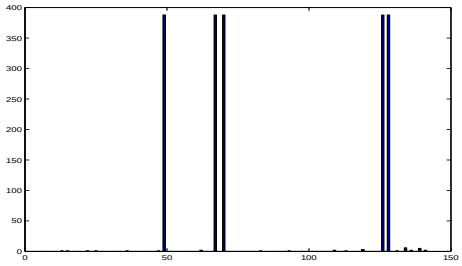

Tumor 2008c, dots are raw data, line is result of GLSo.



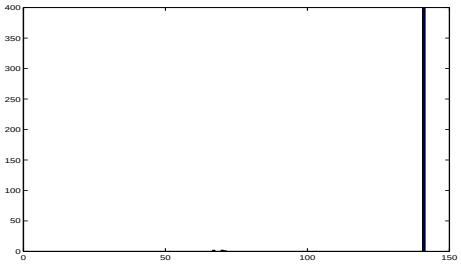

Tumor 2008c, dots are raw data, line is result of GLS.

In order to assess the convergence behaviour of the genetic algorithms we give below plots of a typical histogram of the distribution of breakpoints within the

solutions of the pool after the stopping criterion is satisfied. The plots indicate that the evolutionary process ends with individuals with breakpoints in nearby locations. Observe that the stopping criterion is such that shifting a breakpoint within an area that has no clear breakpoint stops the iterations. In such a case there is "no clear breakpoint", meaning that the means of the two corresponding segments in all individuals are close. This may cause the algorithm to stop after a few iterations even if the individuals have breakpoints in different locations.

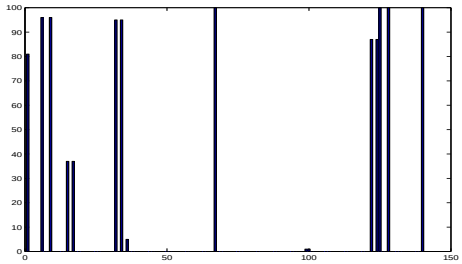


GLSo, tumor 2730, chrm 1

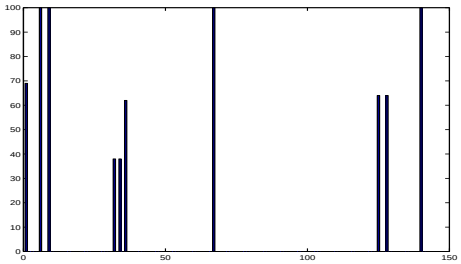


GLS, tumor 2730, chrm 1

Next, we compare the robustness of the genetic algorithms, that is the sensitivity of the outcome to the initialization and other random operators used. Below we plot a typical histogram of the location of breakpoints of the best individual of the final population over 100 runs of the genetic algorithms on chromosome 1.



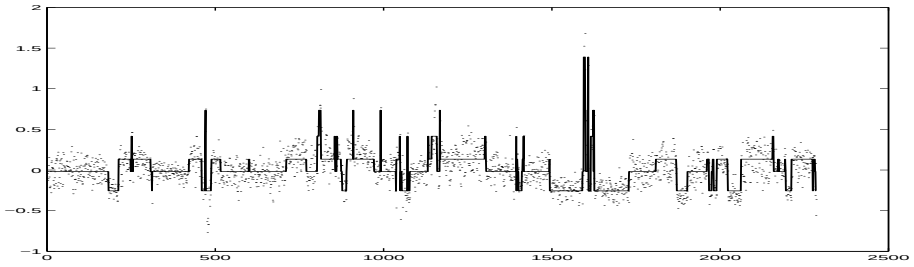
GLSo, Tumor 2730, chrm 1



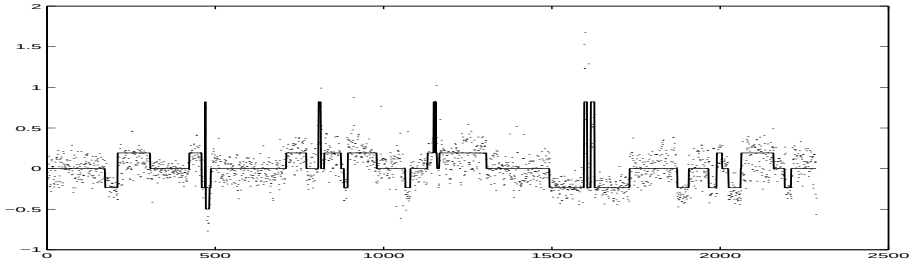
GLS, Tumor 2730, chrm 1

Finally, we compare the smoothings and breakpoints obtained by `GLS` with those manually produced by the expert. The manual smoothings have been built under the assumption that there is a small number of different smoothing levels, reflecting the observation that few copy number values are present in chromosomes. In order to incorporate this constraint in our method, we perform a post processing step that joins close smoothing levels. To this aim the k-means algorithm is applied to the set of CGH values generated by running `GLS` over all the chromosomes of a tumor, and then the resulting smoothing levels that are closer than a fixed threshold are joined. Over all tumors the average difference between the values of the clones is 0.0513, indicating that `GLS` followed by post processing (denoted below by `GLS`-pp) is a satisfactory approximation of the manual

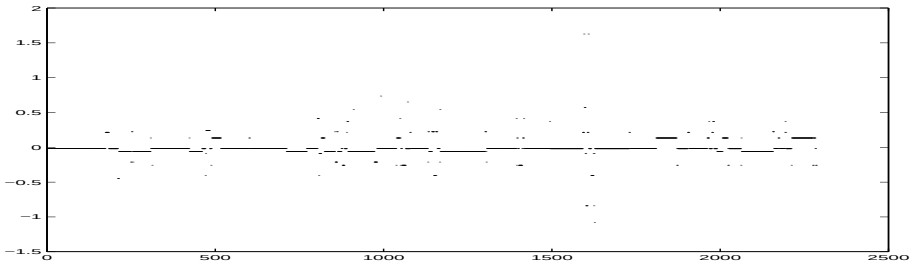
smoothing. GLS seems more sensitive to outliers. This can be explained by the fact that the expert sometimes knows an outlier is meaningless and so ignores it.




Tumor 2730, dots are raw data, line is result of GLS-pp.




Tumor 2730, dots are raw data, line is result of the manual smoothing.




Tumor 2730, dots are the difference between GLS-pp and the manual smoothing.

## 5 Discussion

The results of the experiments indicate that GLS performs better than the other algorithms in minimising function (1).

Both GAs converge within the maximum number of iterations in case the data contains clear breakpoints. The stopping criterion prevents the algorithm from searching for optimal locations of breakpoints that are not clear.

GLS-pp finds smoothings that are very similar to the manual smoothings. It should be noted that an expert produces smoothings based on more information

than just the CGH values. An expert also keeps in mind information like misplacement of clones on the genome and recurring aberrations of clones due to known experimental artefacts. From the normality plots shown it seems that the final expert smoothings are reasonably well normally distributed. Lacking data combining CGH values with known copy numbers in cell types and frequencies of cell types in samples, we were not able to test the suitability of our model to remove noise from the experiment and some cells of types that occur in small numbers. This remains an open problem for future research.

We conclude with some words on related work. To the best of our knowledge, nothing has yet being published on automatic breakpoint detection and estimation of copy number values. We are aware of work in progress carried out at UCSF Cancer Center by Jane Fridlyand, who is trying to use Hidden Markov Models to tackle this problem, and at the Memorial Sloan-Kettering Cancer Center by Adam Olshen who is using change-points and Markovian methods [2].

## References


1. H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. Petrox and F. Caski, editors, *Second International Symposium on Information Theory*, page 267, 1973.
2. J. Fridlyand. Personal communication, 2002.
3. A. N. Jain, T. A. Tokuyasu, A. M. Snijders, R. Segraves, D. G. Albertson, and D. Pinkel. Fully automatic quantification of microarray image data. *Genome research*, 12:325–332, 2002.
4. P. Moscato. On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. Technical report, Caltech Concurrent Computation Program, Californian Institute of Technology, U.S.A., TR No790 1989.
5. D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, S.H. Dairkee, B.M. Ljung, J.W. Gray, and D.G. Albertson. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20, 1998.
6. J.A. Rice. *Mathematical Statistics and Data Analysis Second Edition.* Duxbury Press, 1995.
7. G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
8. A.M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A.K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J.P. Yue, J.W. Gray, A.N. Jain, D. Pinkel, and D.G. Albertson. Assembly of microarrays for genome-wide measurement of dna copy number by cgh. *Nature Genetics*, 29:263–264, 2001.
9. M.M. Weiss, E.J. Kuipers, C. Postma, A. M. Snijders, I. Siccama, D. Pinkel, J. Westerga, S.G.M. Meuwissen, D. G. Albertson, and G.A. Meijer. Genomic profiling of gastric cancer predicts lymph node status survival. *Oncogene*, 2003. In press.