BioTechnologia vol. 92(3) • pp. 283-295 • 2011
Journal of Biotechnology, Computational Biology and Bionanotechnology RESEARCH PAPER

A stable density approach to probe selection for a custom aCGH design

TOMASZ GAMBIN [1]*, PAWEŁ STANKIEWICZ [2,3], ANNA GAMBIN [4,5]

[1] Institute of Computer Science, Warsaw University of Technology, Warszawa, Poland

[2] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, U.S.A.

[3] Department of Medical Genetics, Institute of Mother and Child, Warszawa, Poland

[4] Institute of Informatics, University of Warsaw, Warszawa, Poland

[5] Mossakowski Medical Research Centre, Polish Academy of Sciences, Warszawa, Poland

* Corresponding author: tgambin@ii.pw.edu.pl

Abstract

Usage of the custom aCGH design provides the ability to detect copy number alternations (CNAs) with a very high resolution in almost every genomic region. Since the detection rate of CNAs greatly depends on the quality of the design, it is crucial to ensure the optimal probe coverage for the regions of interests. In this paper, we focus on the problem of finding the best possible probe coverage for a given region and a predefined set of probes. The lack of available probes at some places is a considerable problem which may lead to the decrease in the detection rate of CNAs. We propose a new approach towards the generation of coverage with stable probe density. The developed algorithms attempt to compensate the lack of probes in poorly covered regions by selecting additional probes from their nearest neighborhood. Here we introduce evaluation measures that reflect density stabilization along the covered region. Finally, we use these measures to compare our algorithms with other standard approaches for coverage preparation.

Key words: custom aCGH design, probe selection, Copy Number Alternations

Introduction

DNA copy number alternations (CNAs) which cause the gain or loss of chromosomal material are associated with many types of genomic disorders like mental retardation, congenital malformations, and autism (Lupski, 2009; Shaw et al., 2004). Moreover, genetic aberrations are the characteristic of many type of cancers and are thought to drive several cancer pathogenesis processes (O'Hagan et al., 2003; Snijders et al., 2005; Wang et al., 2006; Lai et al., 2007).

Array comparative genomic hybridization (aCGH) has become the commonly used technique for the identification of CNAs in human genomes (Pollack et al., 1999; Perry et al., 2008). In typical experiments, two DNA samples (e.g. a diseased patient vs. a healthy donor; or normal tissue vs. malignant tissue) are differentially labeled using different fluorophores, and then hybridized to an array. Signal fluorescent intensities of each spot from both samples are considered to be proportional to the copy-number ratio between respective genomic sequence.

Microarrays that contain the required probe dense representation of the genome or larger blocks are typically referred to as tilling arrays (Schliep and Krause, 2008). Most of the works that concern the probe selection and coverage optimization are dedicated to a wide range applications of tilling arrays. aCGH is only one of the applications wherein tilling arrays are used. Besides the aCGH, there are also: ChIP-on-chip for identification of protein biding sides, transcriptome mapping for finding gene expression under various conditions, MeDIP-chip for methylation mapping and DNase-chip. In these applications, different aspects of probes selection are important. In this paper, we focus on designing the CGH array for cytogenetic studies.

In general, when designing the CGH array, the interest and focus lies in the selection of "highest quality" probes with the "best possible" coverage (Lipson et al.,

2007). However, in the real-life designs, there are a number of limitations and prerequisites that have to be taken into consideration.

Among them, there is a limitation as to the number of probes that can be placed on the microarray. Typically, an array consists of a few thousands to several millions of oligonucleotides. If a microarray with more probes is available, more regions can be covered and/or a greater density can be used.

Probes designed for two different genomic locations may differ in their sensitivity, specificity, or other properties. Typically, all these properties are gathered into a single score which reflects the probe quality.

The distribution of probe scores in the genome may have a great impact on the process of probe selection, for example it is pointless to select probes of unacceptable low quality, because they do not provide any meaningful information. In fact, there are a number of regions in the genome, such as segmental duplications, copy number variations, for which it is impossible to prepare well-performing oligonucleotides.

Related research. The process of designing optimal oligonucleotides for a given genome has been reported in many previous papers (Mei, 2003; Li et al.; Gräf et al., 2007). Therefore, there are established databases with predefined sets of probes. The database used in our research consists of 26 millions oligos provided by Agilent Technologies manufacturer, and is accessible through the eArray (http://earray.chem.agilent.com/earray), web-based program. The quality of the probes in the database was confirmed in the experimental studies.

In this paper, we focus on the coverage preparation, i.e. selection of probes for a particular array design. A short summary of the previous research in this area is presented below.

The naive tilling method was used by by Seligner for preparing an array for the transcriptome analysis of *E. coli* (Selinger et al., 2000). The array was designed by selecting oligonucleotides at every sixth base pair for intergenic regions, and every 60th base pair in the coding regions. Similar approaches were used to design a microarray for one of the human chromosomes (Kapranov et al., 2002) and for the whole human genome (Kapranov et al., 2007). The imperfection of these methods is that they do not consider probes performance and may lead to selecting oligos with non-unique sequences, and/or poor hybridization properties.

Fixed window approach was implemented in the Array Design (Gräf et al., 2007). It selects oligonucleotides of better quality in each window of the tilling path. When a wider window is employed, more care can be taken to select unique probes of better performance than it is in the case of naive tilling. On the other hand, the resolution of coverage is limited to the selected size of window.

The method to maximize the resolution proposed in (Lipson et al., 2007), works on a subset of high quality probes. The algorithm selects oligonucleotides that ensure the most homogenous distribution of the tilling path. The main constraints arise from the positioning of the oligos and not from their quality. This paper is one of the few works which are addressed specifically to aCGH custom design.

The approach that considers the problem of the optimization of tilling path has been presented in (Schliep and Krause, 2008). The main goal of this algorithm is to find a trade-off between selecting the most homogeneous and the best quality probes. The authors formulate and resolve the minimal-cost tilling path problem for the selection of oligonucleotides from a set of candidates.

An example of the mixed method is presented in (Hovik and Chen, 2010). The authors propose a two-stage approach to the problem of probe selection. In the first "sequential" stage, probes are selected from the sequence windows tiled alongside the genome. Subsequently, the algorithm tries to fill the largest gaps between adjacent probes with additional oligonucleotides. This procedure is continued until a predefined number of probes are reached.

Motivation. When defining the problem of optimal probe selection, the purpose for which the array is designed, must be considered. In a previous study, authors usually have made an assumption that the primary usage of aCGH is to pinpoint the genomic breakpoints with the best possible accuracy (Lipson et al., 2007). This assumption obviously leads to a design that minimizes the uncertainty at which the breakpoints are mapped. In fact, the accurate breakpoint mapping is crucial in the following aCGH applications.

While analyzing the data from the cancer samples with a high rate of rearrangements, the comparison of breakpoint locations between samples may provide the knowledge about the type of tumor or cancer stage. Moreover, a detailed analysis of the short genomic inter-

vals is also required when one wishes to determine the accurate breakpoint mapping, e.g. to design PCR primers).

On the other hand, there are a number of scenarios where the accurate breakpoint pinpointing becomes a secondary issue. For instance, in the whole genome cytogenetics studies, the detection of all CNAs that are present in the test sample is more important than precise location of rearrangement breakpoints. Once the CNA is detected using aCGH, several other methods can be used to confirm its existence (e.g. fluorescence in situ hybridization) and location of the exact positions of the rearrangement breakpoints, if needed (e.g. by sequencing the entire region).

Our results. In this paper, we focus on the aforementioned cytogenetic applications and propose an approach to generate the coverage with a stable probe density, i.e. we ensure the best possible detection rate of each subinterval of the region that is being covered.

Implemented algorithms were applied to generate the coverage of the sample region from the chromosome 19 of the human genome. The results were evaluated with measures of the stability of the coverage density. For this purpose, we introduced two measures: k-δ density and k-probes distance. Finally, we compared the coverage generated by our algorithms with the coverage generated by the algorithm proposed in (Lipson et al., 2007), which implements the approach to maximize the resolution.

Methods

In this paper, we discuss the problem of probe selection from the subset of high quality probes. Although we describe the coverage of a single continuous DNA section, our algorithms can be easily adapted for the preparation of the design for the set of non-continuous regions.

The goal was to obtain a coverage from the given region, with most stable density, so that all genomic intervals of the equal length, located inside the region would be represented by the same number of probes. This was to ensure the best possible detection of each genomic subregion.

However, because of the lack of oligos in some regions, we could not avoid the fact that in some intervals, high quality probes were underrepresented. The solutions proposed in the previous works usually attempted to overcome this issue by either ignoring these gaps (Lipson et al., 2007) or by using lower quality probes (Schliep and Krause, 2008; Thomassen et al., 2009; Hovik and Chen, 2010).

In our case, we assumed a fixed set of the available probes, i.e. it is impossible to obtain additional probes from gap regions. We propose an approach, wherein the lack of probes in some intervals is compensated with oligos located in the nearest neighborhood of these gaps. Although we did not affect the coverage of gaps, we have provided the opportunity to detect CNAs in the possible smallest wrapping interval of these regions.

Figure 1 shows the advantages of the stable density approach over the solutions which do not consider poorly covered regions but maximize the coverage resolution outside the gaps. The coverage shown in subfigure a) was designed to optimize the probe resolution. Besides the gap, distances among the oligos are uniformly distributed in order to ensure the most accurate breakpoint mapping. Subfigure c) illustrates our stable density approach. The lack of probe around the central location is compensated by an additional oligo from the left neighborhood of the gap. Subfigures c) and d) present results of aCGH experiment performed on subfigures a) and c), respectively. Two deviated probes presented in subfigure b) are usually insufficient to determine the duplication, while three probes with log2-ratio above the threshold (subfigure d) allow calling the aberration, which overlaps the gap.

Poor availability of probes in some locations is usually caused by the presence of repetitive sequences in these regions. Consequently, the likelihood of rearrangements in such regions is greater than in other genomic locations (Stankiewicz and Lupski, 2010). It is becoming extremely important to pay a special attention on poorly covered regions (Hovik and Chen, 2010). Moreover, most of the eukaryotic non-protein coding sequences are low complexity sequences, such as repetitive elements (Ahnert et al., 2008). Many of these sequences are functionally active and have been found to play important regulatory roles (Mercer et al., 2010; Costa, 2010).

Mathematical formulation. In the following definitions, let us fix the genomic interval G of the length $|G|$ that should be covered with n probes. Then the desired space between two consecutive probes is:

$$\delta = |G| / n \tag{1}$$

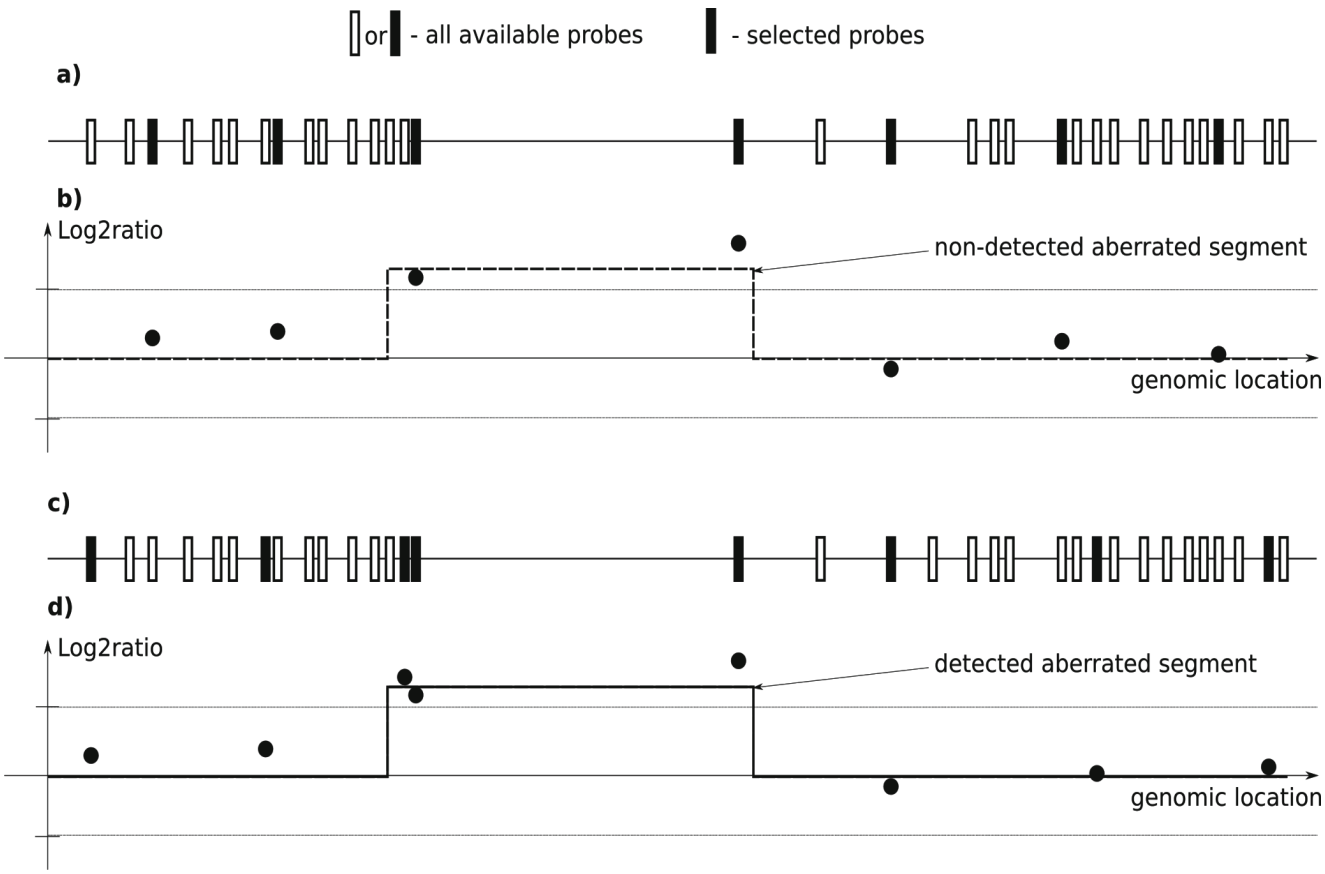

Fig. 1. Comparison of two approaches to coverage preparation

Let us define the property of *k-probes desired density* as follows.

Definition 1. The genomic interval (g_i, x_j) has a property of k-probes desired density, denoted by μ_k, if following conditions are satisfied:

1. The interval (g_i, x_j) is covered by exactly k probes
2. $|(g_i, x_j)| \leq = k * \delta$.

Now, we define the concept of *minimal right neighborhood with k-probes desired density*, denoted by $R^{\mu_k}(g_i)$.

Definition 2. Minimal right neighborhood $R^{\mu_k}(g_i)$ of genomic location g_i with k-probes desired density is an interval (g_i, x_j), such as:

1. The interval (g_i, x_j) has a property μ_k.
2. There exists no interval (g_i, x_l) with property μ_m, where $m \leq k$.

Similarly, we define the *minimal left neighborhood with k-probes desired density*, denoted by $L^{\mu_k}(g_i)$.

Definition 3. Minimal left neighborhood $L^{\mu_k}(g_i)$ of genomic location g_i with k-probes desired density is an interval (x_j, g_i), such as:

1. The interval (x_j, g_i) has a property μ_k.
2. There exists no interval (x_l, g_i) with property μ_m, where $m \leq k$.

Finally, we define the *minimal neighborhood of genomic location gi with k-probes desired density*, denoted by $N^{\mu_k}(g_i)$, as a shorter interval of either: $L^{\mu_k}(g_i)$ or $R^{\mu_k}(g_i)$.

The concept of minimal neighborhoods is illustrated in Figures 2 and 3. Note that if there is no limitation of probes availability, the following holds: the length of minimal neighborhood $|N^{\mu_k}(g_i)| \leq \delta$, for each genomic location.

Algorithms. Selecting all probes from the interval $N^{\mu_k}(g_i)$ ensures that k-probes desired density is reached in the possibly nearest neighborhood of the location g_i. Based on this observation, we developed our algorithms, which try to keep the similar average density along the entire region of coverage.

Let us denote by theoretical backbone Φ, a series of uniformly distributed locations across the region that is being covered, where the distance between every two consecutive locations equals δ.

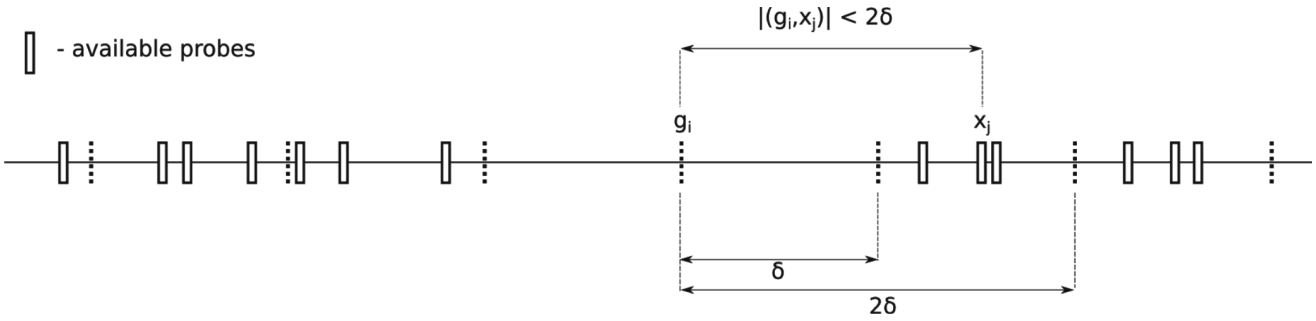

Fig. 2. Interval (g_i, x_j) is the *minimal right neighborhood* $R^{\mu_k}(g_i)$ *of genomic location gi with k-probes desired density*, i.e. there are two probes inside the interval (g_i, x_j) and $|(g_i, x_j)| < 2 * \delta$

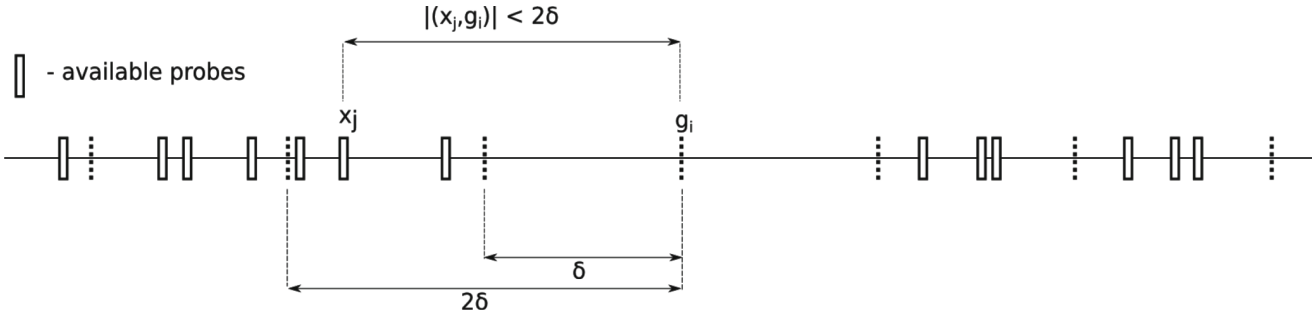

Fig. 3. Interval (x_j, g_i) is the *minimal left neighborhood* $L^{\mu_k}(g_i)$ *of genomic location gi with k-probes desired density*, i.e. there are two probes inside the interval (x_j, g_i) and $|(x_j, g_i)| < 2 * \delta$

Problem 1. Given a genomic region $G = (g_{beg}, g_{end})$, the theoretical backbone of this region Φ and all available probes inside the region, find a coverage which satisfies the following: for each location gi in Φ all probes belonging to the interval $N^{\mu_k}(g_i)$ are selected for the coverage.

Algorithm 1: Find stable density *n*-cover

```
1: FindStableDensityCover(S,n)
2: δ ← (g_end − g_beg) / n
3: Φ ← {g_beg, g_beg + δ, g_beg + 2 * δ, . . . , g_beg
   + (k − 1) * δ, g_end}
4: for all g_i in Φ do
5:     Set p to be the closest element of S to g_i
6:     C ← C + p
7:     S ← S − p
8: end for
9: return C
```

The idea of the Algorithm 1 is straightforward, and thus we call it the naive algorithm. For each genomic location in Φ, starting from the first probe, the algorithm takes the closest available oligo. This probe is added to the result set and removed from the list of available probes. Then the algorithm proceeds to the next location in Φ and repeats the previous step, until it reaches the last position of theoretical backbone.

This algorithm works fine in most of the cases; however, it does not always return the optimal results, in reference to the Problem 1. Scenario in which the naive algorithm produce non-optimal coverage is presented in Figure 5, i.e. the interval $N^{\mu_k}(g_i)$ determined for all available probes differ from the interval $N^{\mu_k}(g_i)$ determined for all selected probes.

It is worth noting that the outcome of the naive algorithm depends on the direction in which the subsequent locations from Φ are processed. Let us consider the set of available probes presented in Figure 5, when we force the algorithm to process locations from the right to the left, then the coverage returned by this algorithm is optimal.

This observation leads us to the Algorithm 2, which we call the optimal algorithm that does not assume any order of processing Φ locations.

In the first substep, the algorithm finds the nearest probe for each unassigned location in Φ. After that some probes could become assigned to more than one loca-

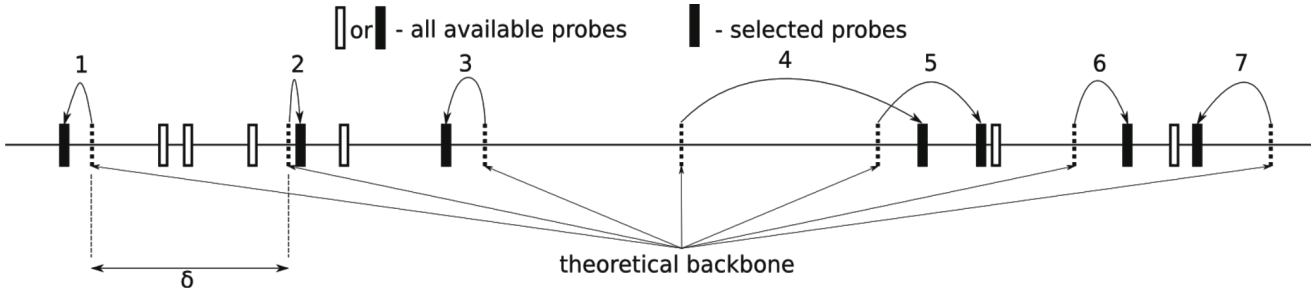

Fig. 4. Illustration of the naive algorithm (Algorithm 1). Arrows indicate assignments of probes to the positions from Φ. Numbers above arrows correspond to the subsequent steps of the algorithm. In each iteration, the nearest available probe to the ith location of Φ is selected

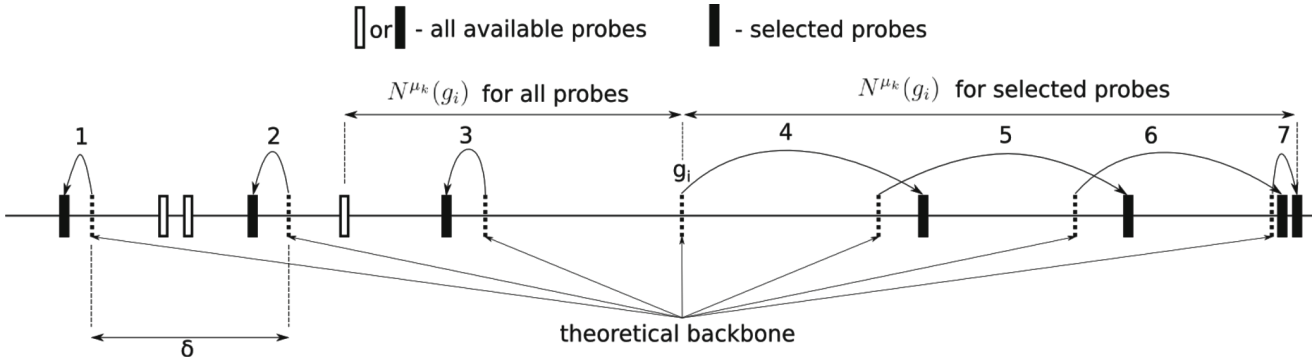

Fig. 5. Example of the coverage generated by the naive algorithm (Algorithm 1), where the minimal neighborhood of the location g_i with k-probes desired density was not optimized

tion. These conflicts are resolved in the second substep. For each probe assigned to more than one location, the algorithm leaves only that assignment for which there is the shortest distance between this probe and location.

Locations that remain unassigned are processed in the next iteration. The algorithm repeats these substeps until all locations in Φ are assigned to probes.

The idea of optimal algorithm is illustrated in the Figure 6. One conflict occurred in the first iteration, i.e. a single probe was assigned to two distinct locations (g_i and g_{i-1}). Because the probe was situated further to g_i than to g_{i-1}, the location g_i remains unassigned. Finally, in the second iteration, the other probe was successfully assigned to g_i. In contrast to the results from naive algorithm, this procedure generates the coverage for which the interval $N^{\mu_k}(g_i)$ determined for all available probes and the interval $N^{\mu_k}(g_i)$ determined for all selected probes are identical.

Now we will prove that Algorithm 2 provides optimal result. Below, we use "('and')" to indicate that endpoints of interval are excluded. Otherwise we use "['and']".

Algorithm 2: Find optimal stable density n-cover

```
1: FindOptimalStableDensityCover(S,n)
2: δ ← (g_end − g_beg) / n
3: Φ ← {g_beg, g_beg + δ, g_beg + 2 ∗ δ, . . . , g_beg
   + (n − 1) ∗ δ, g_end}
4: G ← Φ
5: while G is not empty do
6:     for all g_i in G do
7:         Set p to be the closest element of S to g_i and
           associate p with g_i.
8:     end for
9:     for all p in S, associated with at least one location
       from G do
10:        C ← C + p
11:        S ← S − p
12:        From all locations associated to p, select the
           one which is the closest to p and remove it
           from G.
13:    end for
14: end while
15: return C
```

Claim 1. Consider a point $g_i \in \Phi$ and a probe p, located at x_j as a candidate probe to associate with g_i. We assume that $x_j > g_i$. Then the following holds:

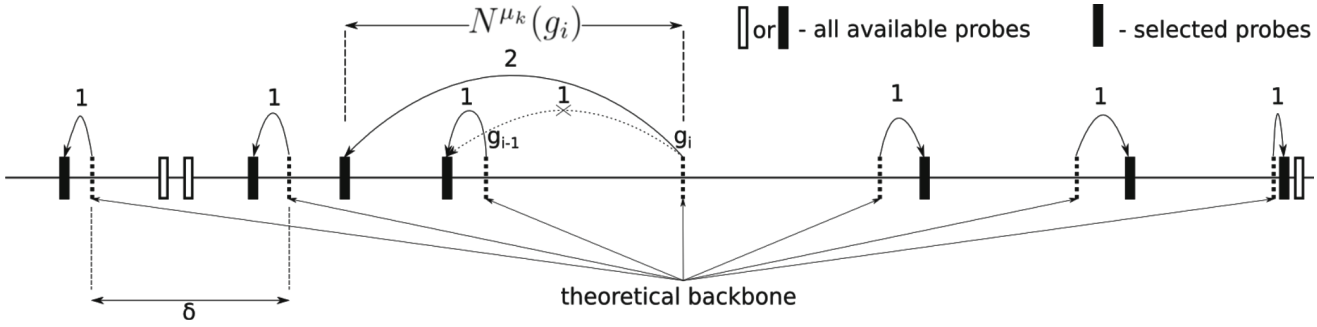

Fig. 6. Illustration of the optimal algorithm (Algorithm 2). Arrows indicate assignments of probes to the positions from Φ. Numbers above arrows correspond to the subsequent iterations of algorithm

(i) The probe p can be assigned to the location g_i only if every position from Φ located within the interval (g_i, x_j) has already an associated probe. Moreover, all of these probes are located within the interval (g_i, x_j).

(ii) The interval $[g_i, x_j]$ has a property of μ_k, (i.e. a k-probes desired density).

(iii) All probes belonging to the interval $N^{\mu_k}(g_i)$ are selected for the coverage C, which is returned by the Algorithm 2.

Proof of Claim 1(i). There are two possibilities in which the Claim 1(i) is not true:

1. At least one of the positions from Φ located within the interval (g_i, x_j) has no associated probes.
2. At least one of the probes assigned to the positions from the interval (g_i, x_j) is located outside this interval.

Let us consider the first possibility. In such a case, the probe p cannot be assigned to g_i, because there exists an unassigned location from Φ which is closer to p, than g_i, which contradicts with the line 12 of the Algorithm 2.

Now, let us consider the second possibility. Let us denote the set of probes by P, associated with these locations from Φ which are located inside the interval (g_i, x_j). Let us assume that the location of one of these probes is $< g_i$. This is impossible, because such a probe would be assigned to the location gi in the previous step of Algorithm 2, since the gi is closer to this probe. The probe from P cannot be located outside the right boundary of the interval (g_i, x_j) either, since there exists the unassigned probe p which is closer to any location from (g_i, x_j).

*Proof of (ii).Let us a*ssume, that $|[g_i, x_j]| < k * \delta$. Because of the Claim 1(i), the number of probes, that are already assigned to positions from Φ and are located within the interval $[g_i, x_j)$ is $\geq k - 1$. This implies that if the probe p is assigned to g_i, there are at least k probes selected for the coverage, located within the interval $[g_i, x_j]$, and thus this interval has a property of μ_k.

Proof of (iii). Let us denote the length of interval $[g_i, x_j]$ by ε. Since the probe p is the nearest unassigned probe to the location g_i, then all available probes located within the interval $(g_i - \varepsilon, g_i + \varepsilon)$ must also be selected for the coverage.

Since the interval $[g_i, x_j]$ has a property of μ_k (see the Claim 1(ii)) and all probes from the neighborhood of g_i smaller than 2ε are selected for the coverage, then we can conclude that probes from the interval $N^{\mu_k}(g_i)$ are also selected.

Analogously, it can be shown that Claims 1(i), 1(ii) and 1(iii) remain true when $x_j < g_i$.

Complexity issues. Let us denote the number of locations from Φ by n and the number of avaiable probe by m. To ensure the efficiency of the naive algorithm, a list should be implemented that combines locations from Φ and probes. Then the list is sorted according to genomic positions of elements. Based on this data structure, the whole algorithm can be processed in n steps of the constant time. Thus, the running time of naive algorithm depends on the time of combined list sorting and equals to $O((n + m) \log (n + m))$.

The optimal algorithm can be implemented in a similar way. However, in the worst case, additional n^2 steps may be needed, because of the possible conflicts that may occur in each iteration. The running time of the optimal algorithm is $O((n + m) \log (n + m) + n^2)$.

Coverage evaluation measures. Evaluation measures that are used for validation of the design can be divided into measures which describe probe performance

and measures which reflect the correctness of the probes distribution. The following measures are related to the evaluation of probe performance: average probe score, distribution of probes melting temperature, percentage of non-unique probes or percentage of GC content in the coverage (e.g. see Lemoine et al., 2009). Typical measures which are used for the validation of the coverage distribution are as follows: average probe spacing (average distance between two consecutive probes) (Lipson et al., 2007; Lemoine et al., 2009), percentage of copy number variation in coverage or breakdown analysis of the target elements of coverage.

Because our solution is considered to work with a predefined set of high quality probes, we focused on the validation of coverage distribution. Since typical measures proposed in previous works do not reflect the stabilization of density along the covered region, in this paper, we propose the following approaches to the coverage evaluation:

- *k-δ density of genomic location* – computed for each position g_i from the covered region as a number of probes which are located inside the interval $(g_i - k * \delta / 2, g_i + k * \delta / 2)$ divided by k.
- k-probes distance – computed for each probe p from the coverage as a distance between p and k-th right nearest neighbor of p. We also used the *k-probes normalized distance*, which we define as a *k-probes distance* divided by k.

Graphical interpretation of these measures is presented in the Figure 7. It is worth noting that since we want to keep a stable density along the covered region, the desired value of the *k-δ density* is 1, and the desired *k-probes normalized distance* equals to δ.

In order to obtain global coefficients for the entire coverage, we used the root mean square deviation (RMSD), which refer to the deviations of either *k-δ density* or *k-probes normalized distance* from their desired values. The RMSD, for observations X_i and desired value μ *were* defined as:

$$RMSD(X_i, \mu) = \sqrt{MSD(X_i, \mu)} = \sqrt{E((X_i - \mu)^2)} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}}$$

In this paper we consider *RMSD (k-δ densities,* 1*)* and *RMSD (k-probes normalized distances, δ)*.

Comparison with existing method. We compared our algorithms with the common approach of coverage preparation, which maximizes the resolution outside the poorly covered regions. For comparison, we selected the algorithm proposed in (Lipson et al., 2007), because unlike other approaches it was developed with focus on issues inherent in the aCGH technology.

Authors clearly defined the problem of coverage, by introducing the concept of the *whenever possible* (WP) ε-cover in following way:

Definition 4. Given a genomic region G, a set of candidate probes P and a parameter ε, a subset $(C = (c_1, \ldots, c_k)) \subseteq P$ is a *WP* ε-cover of G with respect to P, if for any genomic location $x \in G$ with respect to P, the following holds. Let c_i and c_{i+1} be the two selected probes closest to x from the left and from the right, respectively (if $x < c_1$ then c_0 is set to be the left-end of G, and for $x > c_k$, c_{k+1} is the right and of G). The one of the following holds:

1. $c_{i+1} - c_i \leq \varepsilon$ (i.e. the flanking selected probes are within ε distance of each other), or
2. there is no candidate probe between c_i and c_{i+1}. For such a cover C, we say that the resolution of C is ε.

To find the *WP* ε-cover, authors of (Lipson et al., 2007) developed Algorithm 3, which is called the FindMinCover algorithm. The results of the comparison of our naive and optimal algorithms with the FindMinCover is presented in the results section.

Algorithm 3: Find a minimal size WP ε-cover

```
1: FindMinCover(ε)
2: c_0 ← g_beg
3: i ← 0
4: while (g_beg − c_i) > ε do
5:     Set C_i to be the rightmost candidate probe following c_i such that (c_{i+i} − c_i) ≤ ε
6:     If no such probe exists: c_{i+1} is the next candidate probe to the right of c_i
7:     i ← i + 1
8: end while
9: return C = {c_1, ..., c_i}
```

Results

Experimental setup. We prepared 3 coverages for the 22 Mbp fragment of chromosome 19 of the human genome, using our naive and optimal algorithms and FindMinCover.

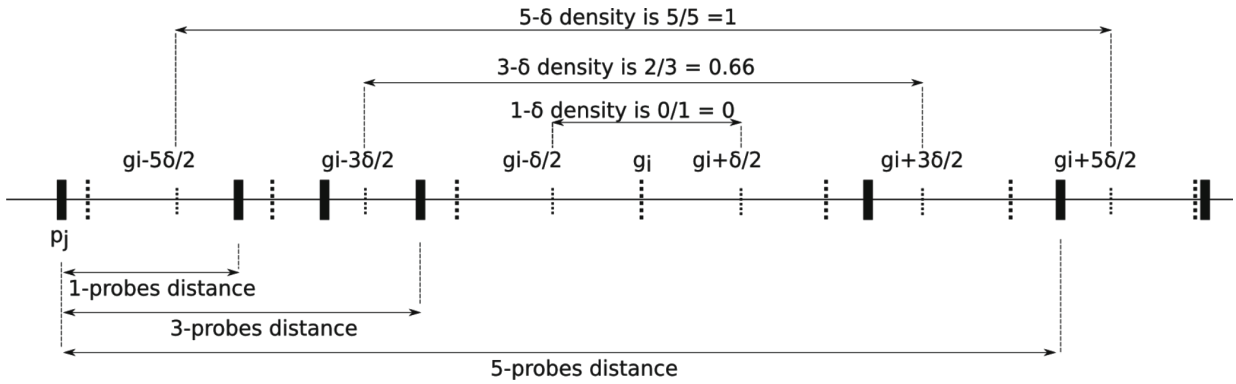

Fig. 7. *k-δ density of genomic location* g_i (shown on the top of figure) and *k-probes distance* for the probe p_j (shown at the bottom of the figure). Measures are presented for k = 1,3 and 5. In this case the desired value of density is reached for k = 5

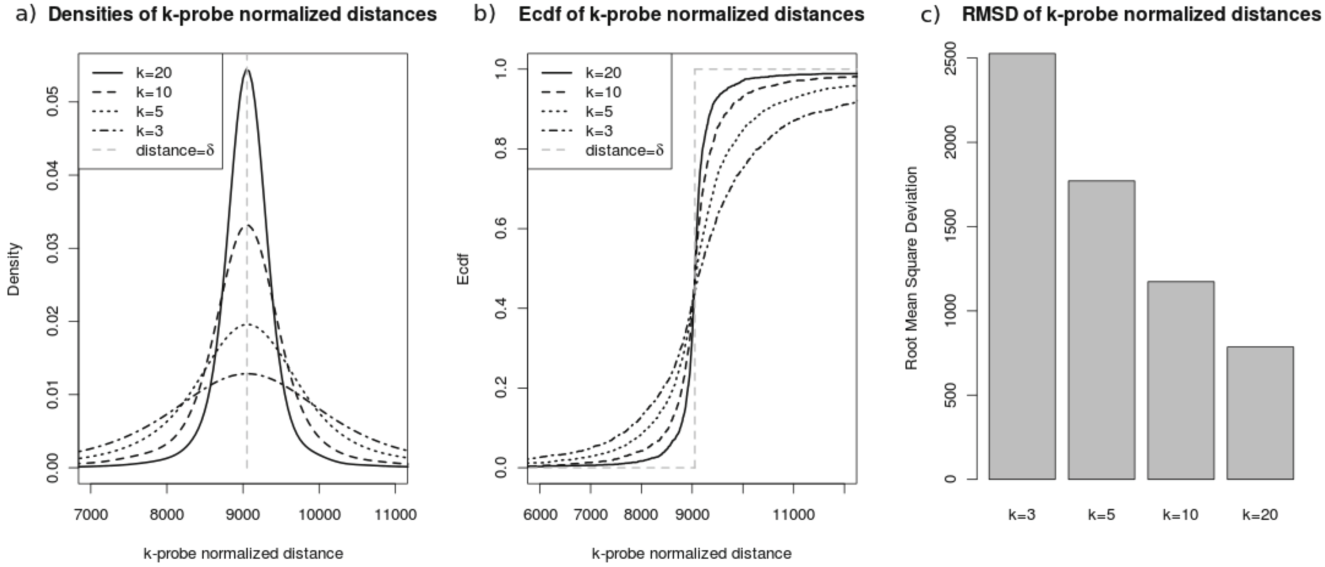

Fig. 8. Subfigure a) presents the densities of *k-δ* densities, for k = 3, 5, 10 and 20. The vertical gray line corresponds to the desired distance δ. Similarly, the ecdf's are plotted in the subfigure b) and the RMSDs are presented in the subfigure c)

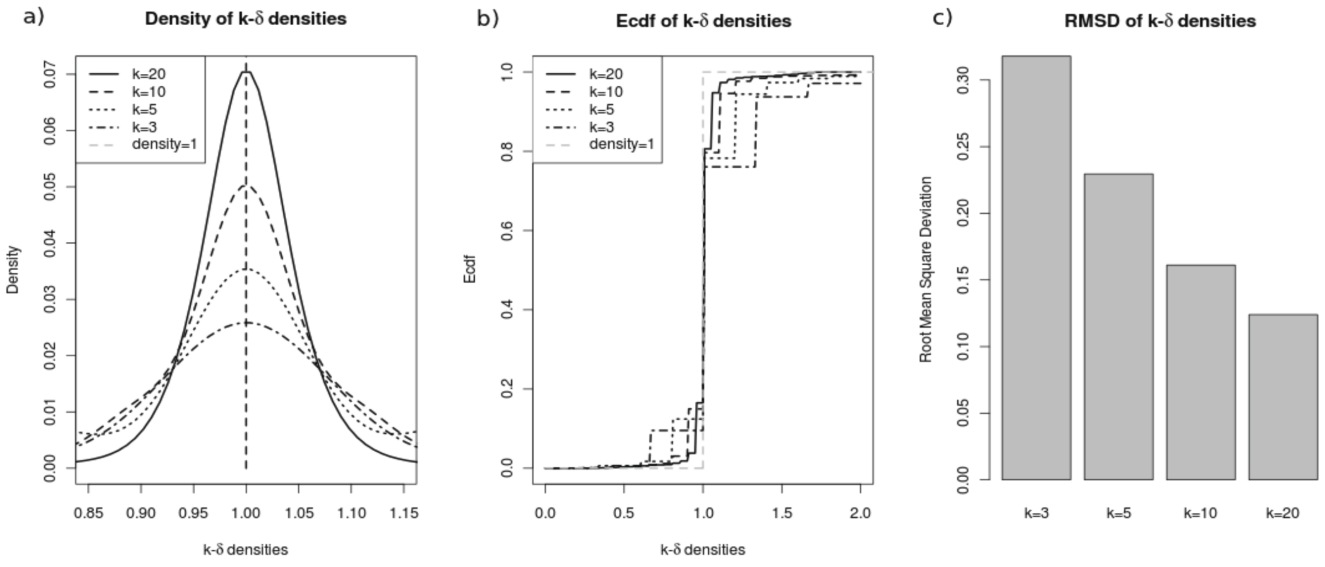

Fig. 9. Subfigure a) presents the densities of *k-δ* densities, for k = 3, 5, 10 and 20. The vertical gray line corresponds to the desired distance δ. Similarly, the ecdfs are plotted in the subfigure b) and the RMSDs are presented in the subfigure c)

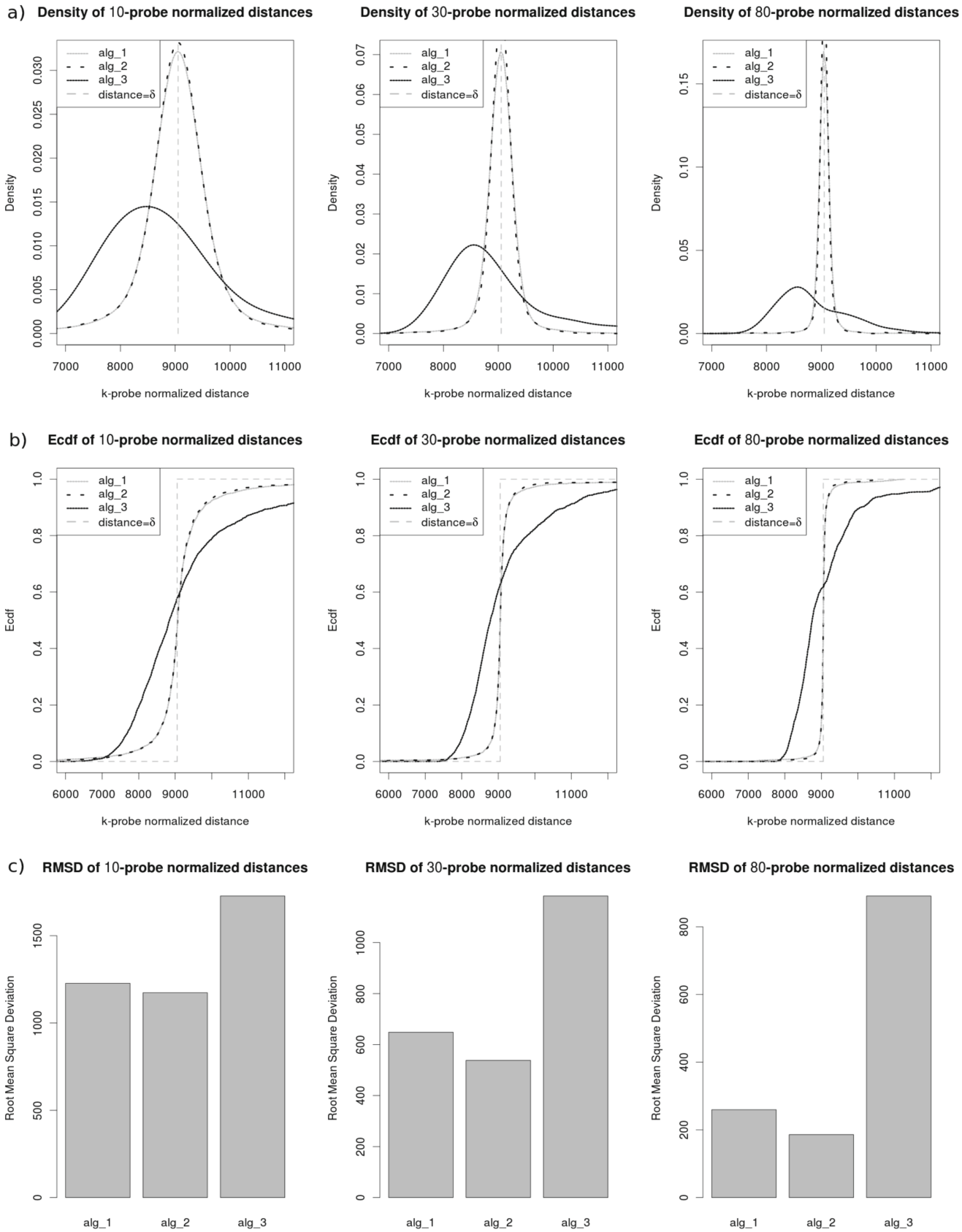

Fig. 10. Comparison of our naive (alg 1) and optimal (alg 2) algorithms to FindMinCover (alg 3). The Figures show the following statistics for the k-probes normalized distances: densities – row a) ecdfs – row b) and RMSDs – row c). Successive columns corresponds to $k = 10$, 30 and 80

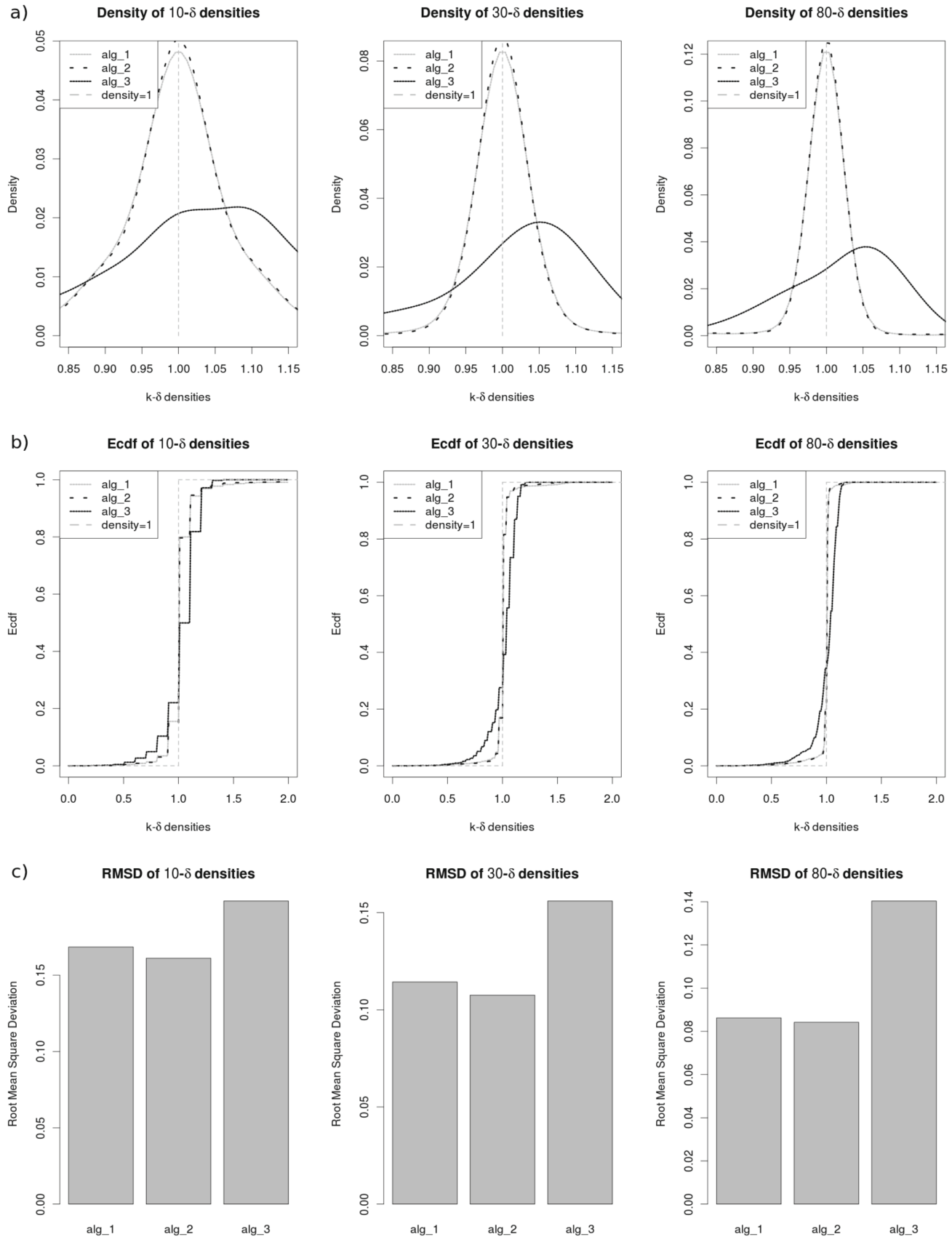

Fig. 11. Comparison of our naive (alg 1) and optimal (alg 2) algorithms to FindMinCover (alg 3). The Figures show the following statistics for the k-δ densities: densities – row a), ecdfs – row b) and RMSDs – row c). Successive columns corresponds to k = 10, 30 and 80

The predefined set of about 177 K of probes was obtained from Agilent probes database accessible through eArray. From this set, 20 K well-performing probes (with scores above 0.8) were selected.

The first coverage was computed using FindMinCover algorithm, with the parameter $\varepsilon = 10{,}000$. The obtained coverage consisted of 2453 oligos. This number was used as an input parameter n for our naive and optimal algorithms. With respect to these constraints, the desired distance between two consecutive probes was fixed to 9052 bp.

Outcome. In Figures 8 and 9, the statistics of *k-probes normalized distance* and *k-δ density* were presented for four different values of k. All of these statistics were computed on the basis of the results obtained by the optimal algorithm (Algorithm 2).

According to our observation, while increasing the k, distributions converge to their desired shape, and the RMSD tends to zero. This is an important remark, because it indicates that the stability of probe densities in the coverage increases monotonically with an increase in k.

The other conclusion is that there exists a correspondence between *k-probes normalized distance* and *kδ density* measures, i.e. for a given k the distributions shown in Figure 8 and 9 are similar with respect to their shapes.

However, in a specific situation, the usage of one measure can be more convenient than the other. For instance, the *k-δ density* measure shows the local density for each single position of the covered interval. The *k-probes normalized distance* measure presents the similar information; however, it is limited to the locations of selected oligos. Because of this, the first measure is better, for example when one analyzes a short fragment of coverage.

On the other hand, the *k-probes normalized distance* is more informative, when one wishes to investigate the entire coverage using a low value of k. Because of the more discrete nature of *k-δ density* in comparison to *k-probes normalized distance*, the empirical cumulative distribution function (ecdf) computed for the second measure is easier to interpret (compare ecdfs presented in figures 8 and 9; for lower values of k, the *k-δ density* can take only certain values, which correspond to the step-function character of ecdf plots).

Finally, in figures 10 and 11, we present the plots of the comparison of our naive and optimal algorithms and FindMinCovered algorithm, proposed by Lipson and coworkers (Lipson et al., 2007).

The comparison has revealed that our naive and optimal algorithms outperform the FindMinCover, with respect to stabilization of probe densities. The superiority of our algorithms increases with the increasing k. As expected, the differences in performance are more visible when comparing *k-probes distances* rather than comparing *k-δ densities*.

There are almost no differences in the distribution of evaluation measures based on the results from the naive and optimal algorithm; however, the optimal algorithm slightly outperforms the naive version, when the RMSDs are compared.

Conclusions

Because of the noise and technical issues, the detection of CNAs is usually limited to regions that are covered by sufficient number of probes. The number of deviated probes necessary to confirm the presence of aberrated segment may depend on: the selected technology, the quality of experiments or type of sample. In typical well-performed aCGH experiments, 3 to 5 probes are enough to determine the CNA in a given region. However, several thousands of oligos may be required to confirm the aberration, when one works with extremely noisy data, like in case of Single Cell Chip (see Fiegler et al., 2007), where DNA sample are obtained from a single cell.

This problem has been noticed by authors of the software for analyzing aCGH results. In particular, CBS segmentation algorithm (Olshen et al., 2004), allows to define the minimal number of consecutive probes of which a segment is composed. In consequence, aberrations in regions that are not covered with this number of probes, become undetectable.

Therefore, the application of our approach to the coverage preparation may significantly increase the performance of CNAs detection, especially near regions of poor probes availability.

Acknowledgments

This research has been supported in part by the Polish Ministry of Science and Educations grants N301 065236, N516 531839 and R13-0005-04/2008. It was also supported by the Foundation for Polish Science and the European Union in the framework of European Social Fund through the Warsaw University of Technology Development Programme.

References

Ahnert S.E., Fink T.M.A., Zinovyev A. (2008) *How much non-coding DNA do eukaryotes require?* J. Theor. Biol. 252: 587-592.

Costa F.F. (2010) *Non-coding RNAs: meet thy masters*. BioEssays 32: 599-608.

Fiegler H., Geigl J.B., Langer S., Rigler D., Porter K., Unger K., Carter N.P., Speicher M.R. (2007) *High resolution array-CGH analysis of single cells*. Nucl. Acids Res. 35: e15-e15.

Gräf S., Nielsen F.G.G., Kurtz S., Huynen M.A., Birney E., Stunnenberg H., Flicek P. (2007) *Optimized design and assessment of whole genome tiling arrays*. Bioinformatics 23: i195-i204.

Hovik H., Chen T. (2010) *Dynamic probe selection for studying microbial transcriptome with high- density genomic tiling microarrays*. BMC Bioinformatics 11: 82.

Kapranov P., Cawley S.E., Drenkow J., Bekiranov S., Strausberg R.L., Fodor S.P.A., Gingeras T.R. (2002) *Large-Scale transcriptional activity in chromosomes 21 and 22*. Science 296: 916-919.

Kapranov P., Cheng J., Dike S., Nix D.A., Duttagupta R., Willingham A.T., Stadler P.F., Hertel J., Hackermüller J., Hofacker I.L. et al. (2007) *RNA maps reveal new RNA classes and a possible function for pervasive transcription*. Science 316: 1484-1488.

Lai C., Horlings H.M., van de Vijver M.J., van Beers E.H., Nederlof P.M., Wessels L.F., Reinders M.J. (2007) *SIRAC: supervised identification of regions of aberration in aCGH datasets*. BMC Bioinformatics 8: 422.

Lemoine S., Combes F., Crom S.L. (2009) *An evaluation of custom microarray applications: the oligonucleotide design challenge*. Nucl. Acids Res. 37: 1726-1739.

Li X., He Z., Zhou J. (2005) *Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation*. Nucl. Acids Res. 33: 6114-6123.

Lipson D., Yakhini Z., Aumann Y. (2007) *Optimization of probe coverage for high-resolution oligonu- cleotide acgh*. Bioinformatics 23: e77-83.

Lupski J.R. (2009) *Genomic disorders ten years on*. Genome Med. 1: 42.

Mei R. (2003) *Probe selection for high-density oligonucleotide arrays*. Proc. Nat. Acad. Sci. 100: 11237-11242.

Mercer T.R., Qureshi I.A., Gokhan S., Dinger M.E., Li G., Mattick J.S., Mehler M.F. (2010) *Long non- coding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation*. BMC Neurosci. 11: 14.

O'Hagan R.C., Brennan C.W., Strahs A., Zhang X., Kannan K., Donovan M., Cauwels C., Sharpless N.E., Wong W.H., Chin L. (2003) *Array comparative genome hybridization for tumor classification and gene discovery in mouse models of malignant melanoma*. Cancer Res. 63: 5352-5356.

Olshen A.B., Venkatraman E.S., Lucito R., Wigler M. (2004) *Circular binary segmentation for the analysis of array-based dna copy number data*. Biostatistics (Oxford, England) 5: 557-572.

Perry G.H., Ben-Dor A., Tsalenko A., Sampas N., Rodriguez-Revenga L., Tran C.W., Scheffer A., Steinfeld I., Tsang P., Yamada N.A. et al. (2008) *The fine-scale and complex architecture of human copy-number variation*. Amer. J. Human Gen. 82: 685-695.

Pollack J.R., Perou C.M., Alizadeh A.A., Eisen M.B., Pergamenschikov A., Williams C.F., Jeffrey S.S., Botstein D., Brown P.O. (1999) *Genome-wide analysis of dna copy-number changes using cdna microarrays*. Nature Gen. 23: 41-46.

Schliep A., Krause R. (2008) *Efficient algorithms for the computational design of optimal tiling arrays*. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 5: 557-567.

Selinger D.W., Cheung K.J., Mei R., Johansson E.M., Richmond C.S., Blattner F.R., Lockhart D.J., Church G.M. (2000) *RNA expression analysis using a 30 base pair resolution escherichia coli genome array*. Nat. Biotech. 18: 1262-1268.

Shaw C.J., Shaw C.A., Yu W., Stankiewicz P., White L.D., Beaudet A.L., Lupski J.R. (2004) *Comparative genomic hybridisation using a proximal 17p bac/pac array detects rearrangements responsible for four genomic disorders*. J. Med. Genet. 41: 113-119.

Snijders A.M., Schmidt B.L., Fridlyand J., Dekker N., Pinkel D., Jordan R.C.K., Albertson D.G. (2005) *Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma*. Oncogene 24: 4232-4242.

Stankiewicz P., Lupski J.R. (2010) *Structural variation in the human genome and its role in disease*. Ann. Rev. Med. 61: 437-455.

Thomassen G.O.S., Rowe A.D., Lagesen K., Lindvall J.M., Rognes T. (2009) *Custom design and analysis of High-Density oligonucleotide bacterial tiling microarrays*. PLoS ONE 4: e5943.

Wang Y., Makedon F., Pearlman J. (2006) *Tumor classification based on dna copy number aberrations determined using snp arrays*. Oncol. Rep. 15 Spec no.: 1057-1059.