Saudi Journal of Biological Sciences 27 (2020) 3647–3654

Contents lists available at ScienceDirect

Saudi Journal of Biological Sciences

journal homepage: www.sciencedirect.com

Original article

Fuzzy methods for the detection of copy number variations in comparative genomic hybridization arrays

Ahmad AlShibli [a,*], Hassan Mathkour [b]

[a] Department of computer science, College of computer and information sciences, King Saud University, Riyadh, Saudi Arabia
[b] Department of computer science, College of computer and information sciences, King Saud University, Riyadh, Saudi Arabia

ARTICLE INFO

Article history:
Received 12 November 2019
Revised 2 August 2020
Accepted 3 August 2020
Available online 13 August 2020

Keywords:
Copy Number Variations
Fuzzy Inference System
Fuzzy c-means clustering
Membership function

ABSTRACT

Genomic copy number variations (CNVs) are considered as a significant source of genetic diversity and widely involved in gene expression and regulatory mechanism, genetic disorders and disease risk, susceptibility to certain diseases and conditions, and resistance to medical drugs. Many studies have targeted the identification, profiling, analysis, and associations of genetic CNVs. We propose herein two new fuzzy methods, taht is, one based on the fuzzy inference from the pre-processed input, and another based on fuzzy C-means clustering. Our solutions present a higher true positive rate and a lower false negative with no false positive, efficient performance and consumption of least resources.

1. Introduction

The human genome sequence is subject to variations of different sizes, ranging from a single nucleotide base to an entire chromosome (Montgomery et al., 2013). Structural variations are defined as those with lengths that exceed 1000 bases (Guan and Sung, 2016; Perry, 2008). Copy number variations/alterations (CNVs/CNAs), are considered among the most important structural variations (Montgomery et al., 2013; Macé et al., 2018) because they are located in 12% of human genomes (Redon et al., 2006), influence the protein expression process (Flores et al., 2013; Shao et al., 2019), known to affect susceptibility to some disorders (Shlien and Malkin, 2009; Wang et al., 2019; Yan et al., 2018; Kendall et al., 2019), correlated with several diseases (Usher and McCarroll, 2015) such as cancer, diabetes and autism, and suggested to be associated with drug resistance in some health conditions (Wang et al., 2019).

The remainder of this section provides a review of the previous work to discover copy number variations in comparative genomic hybridization array (arrayCGH) data. Section 2 describes the dataset, the process of preparing the experiment data, and the solutions we propose to solve the detection problem. Section 3 specifies how the experiments are conducted and presents the results. Section 4 discusses our findings and the comparison used to evaluate our solutions. Section 5 summarizes our conclusions. Finally, Section 6 sheds light on the future work.

1.1. Problem statement

The computational discovery of CNVs has been an important research subject for over a decade. With the advances in biotechnology and the emergence of high-throughput platforms, an increasing number of studies are targeting this problem with new methodologies and improvements of existing ones. The following three main platforms may lend their output for the computational inference of genetic copy numbers: single nucleotide polymorphism (SNP) arrays, microarray comparative genomic hybridization (array CGH), and Next Generation Sequencing (NGS).

In array CGH, two DNA pools (i.e., a test and a control) are labeled differently with fluorescent dyes and hybridized to an array. After the array processing, its image is captured and digitally analyzed to quantify the color intensity at each array site. This combined intensity is seen as the ratio between the test and control fluorescence intensities, which consequently represents the copy number ratio between the two clones for each target (Yao et al., 2018; Kallioniemi et al., 1992).

An array CGH platform makes it possible to map CNVs to a DNA sequence and allows for increasingly high resolutions with the past advancements of the microarray technology. This platform is technically not the most advanced one. Nevertheless, with the tremendous repository of array data that were accumulated over the years and the rapidly increasing resolution with the falling cost of

* Corresponding author at: P.O. Box 230734, Riyadh, Saudi Arabia.
E-mail addresses: alshibli@ksu.edu.sa (A. AlShibli), mathkour@ksu.edu.sa (H. Mathkour).

https://doi.org/10.1016/j.sjbs.2020.08.007
1319-562X/© 2020 The Author(s). Published by Elsevier B.V. on behalf of King Saud University.

running nowadays, microarrays keep the array CGH in the center of CNV detection studies and make it a central tool for analyzing structural variations and verifying the findings of newer technologies (Shah et al., 2006).

1.2. Related work

The subsequent sections will present an extensive survey of the computational methods of detecting CNVs using the array CGH platform in the literature. We chose to better compare and contrast their underlying mechanisms:

a. Smoothing-based methods
b. Clustering-based methods
c. Hidden Markov Model-based methods
d. Recursive segmentation methods

1.2.1. Smoothing-based methods

These methods try to restore the original signal by assuming a Gaussian noise interference. Usually, the first step is filtering out the outlying observations interrupting its consistent neighbors. The next step is to fit continuous curves to the long bursts of these neighbors. The points within these curves demonstrate a consistent likelihood measure among their neighbors. A sharp change of the likelihood marks a boundary, that is the end of the curve and the start of the next one.

Jong et al. (2003) proposed two genetic local search algorithms (Moscato, 1989) to find the most N probable breakpoints yielding clusters of clones of the same copy number value. These algorithms use a fitness function that combines the maximum-likelihood and a penalty adjustment to discourage high model dimensionality.

Hupé et al. (2004) used an iterative adaptive weight smoothing technique (Polzehl and Spokoiny, 2000) that tries to find piecewise approximated constant functions for regions with maximum local likelihood while preserving the contrasts on both sides of the breakpoint.

Eilers and de Menezes (2005) smoothed the array data using a penalized quantile regression algorithm (Eilers, 2003). This method helped in the visual detection of changes in the smoothed median trend. However, the choice of the penalization parameter was manually made, and the model significantly relied on the robustness of the patterns within the data.

Stamoulis and Betensky (2011) applied signal decomposition (SD) in the first step to increase the signal-to-noise ratio (SNR). A second waveform-matching filter (MF) step was then taken to maximize the SNR in the matching regions. Considering the dependency of the algorithm's detection sensitivity on the frequency of the underlying CNVs, a modification was introduced (Stamoulis and Betensky, 2016) to optimize the signal decomposition matched filtering (SDMF) method such that it compares the SNR changes to a threshold selected by studying the behavior of the algorithm through a series of simulations.

1.2.2. Clustering-based methods

Autio et al. (2003) proposed a three-stage visual segmentation tool, called CGH-Plotter, that uses a moving median/mean filter to denoise the signal, three-means clustering to discover the number of changes per chromosome, and dynamic programming that relies on the minimum mean square sum to recognize the change points.

Wang et al. (2005) proposed the Cluster Along Chromosome (CLAC) hierarchical algorithm that starts with each probe in a separate cluster, then repeatedly merges the adjacent clusters of a small relative distance. Three measurements were used to select the clusters of interest, namely, the maximum relative distance between two nodes in the cluster, size of all sub trees with respect to each node, and mean values of the log ratios of all subtrees' leave nodes.

1.2.3. Hidden Markov model (HMM) methods

The basic idea in using HMM methods, is to have the underlying copy number at each chromosome loci represented by a hidden state. The goal then becomes inferring these hidden states starting from the initial state and moving iteratively to the following states using a set of transition probability matrices and suitable emission probability functions.

Fridlyand et al (2004) fitted a k-state HMM to break clones into segments of the same real copy numbers, where the number of states (k) was determined by trying all values between 1 and kmax, which is the number of clones on the chromosome, penalizing each trial with a negative log-likelihood, and then choosing the model with a minimum penalty. The neighboring segments with median distance less than a given threshold were then merged, reducing the k. In the second stage, the clone's copy number was estimated by removing the outliers and considering the following three factors: 1) whole chromosomal changes; 2) focal amplifications and aberrations; and 3) double stranded breaks.

This work was extended by Marioni et al (2006) to incorporate the distance between the neighboring segments and the close quality using a homogenous HMM method, called BioHMM. The transition probability in this method depends on the distance between the midpoints of two adjacent segments; hence, the link between far-apart consecutive segments will have less significance.

Shah et al (2006) proposed another modification to fix the over-segmentation problem that results from sensitivity to outliers. This issue was tackled at two levels. The previously assumed Gaussian model for observation was replaced by two Gaussian components (i.e., one for the recovered state and one for the outliers), which enables the incorporation of prior knowledge on the frequencies of the copy number polymorphisms (CNPs) in the outlier component at the next level.

Mahmud and Schliep (2011) tried to improve the computational demand of these modifications by integrating a KD-tree algorithm in CMCM sampling to approximate the data into compressed blocks. Wiedenhoeft et al. (2016) noted that this enhancement imposes rigidness of the compression block sizes that was not CNVs' nature and an inherent tendency for overfitting or weak clustering. Hence, they combined a Haar wavelet smoother with HMM segmenting, which helped shift the heavy computational effort from obvious CNVs to problematic ones.

1.2.4. Recursive segmentation methods

Olshen et al (2004) used circular binary segmentation (CBS) to analyze the copy number data through a recursive division of chromosomes into segments, in which each has the same distribution function for its intensities. The algorithm requires quadratic complexity computations for verification; thus, Venkatraman and Olshen (2007) proposed hybrid modification to achieve a linear time computation. This method was further improved by Picard et al. (2005).

The current methods in the above categories tried to restore the original signal by combining the inferred characteristics of the series density distribution and some assumptions about the distributions then predicting breakpoints wherever the distribution characteristics change. The methods we propose in the following section do not require knowledge nor make assumptions about the underlying signal distribution.

Similar problems in other domains used some of the previous solutions, however, they successfully employed other methods. in the rest of this paper we propose, test and compare two new fuzzy methods.

2. Material and methods

We start by specifying the dataset and the basis of our selection. We then try to extrapolate the literature to show how researchers in other domains tackled similar problems, right up to the formation of our solution to the problem.

2.1. Dataset and data preparation

In this study, we used the Coriell array CGH data, consisting of 15 cell line samples. This data set is referred to as a gold standard in studies of array CGH data analysis (Fridlyand et al., 2004; Hupé et al., 2004; Carter, 2007; Sheha et al., 2016; Yin and Li, 2010). For each sample, and on each of the 23 chromosomes, the genomic positions of the array probes and their corresponding normalized log2 ratios were stored in parallel vectors.

What distinguishes the Coriell dataset is the fact that the locations of the CNVs in the cell lines have been cryogenically verified (Snijders et al., 2001).

Each sample is a pure cell line with two sets of chromosomes (diploid). The cell lines were produced from fibroblast cells (12 samples), chorionic villi cells (2 samples) and one from lymphoblast. The strains were individually hybridized with bacterial artificial chromosome (BAC) and spotted on 2276 probe microarrays in triplicate (Hupé et al., 2004).

2.2. Fuzzy inference method of copy number variations

The fuzzy inference system (FIS) uses fuzzy arithmetic and a set of fuzzy conditional rules dealing with fuzzy input and fuzzy output, rendering all mappings to a degree-based dependency. The three main stages for processing data in the FIS are as follows:

1) Fuzzifying the input: a choice of membership function is used to convert each variable from its crisp into a degree of membership in the fuzzy set, using a choice of membership function.
2) Applying fuzzy conditional rules: this stage comprises of three steps by itself, namely the application of fuzzy logical operators to extract a binary value summarizing the input, assignment of a rule-based weight to the result, and aggregation of the output of all rules for each output variable.
3) Defuzzifying the output variable: the output variable is defuzzified by converting the aggregated fuzzy value into a crisp number.

When designing a fuzzy system, the choice of the membership function plays a significant role in system effectiveness. Zhao and Bose (2002) evaluated the sensitivity and effectiveness of 12 common types of membership functions, and concluded that triangular and trapezoidal membership functions yield a dominant performance.

Fuzzy inference systems are widely used in segmentation problems (Canny, 1986; Chaira, 2010; Shah et al., 2013; Haq et al., 2015), but to our knowledge, the problem of CNV identification has not so far benefited from the potentials of this method.

We describe herein a fuzzy identification approach for the discovery of CNVs. Unlike the reviewed statistical methods, our algorithm does not require a prior knowledge on the distributions of the signal or its segments. Fig. 1 presents the block diagram of the proposed system, and highlights its main stages and functional components, which will be explained afterwards.

2.3. Fuzzy c-means clustering of copy number variations

The traditional Hard C-Means (HCM) clustering algorithm, which is also known as K-means clustering, partitions observations into K mutually exclusive clusters based on the distances between data points. Each cluster is characterized by its centroid; hence, the distances between the cluster points and its centroid are minimal, while those distances between the cluster points and the other clusters' centroids are maximal.

On the contrary, fuzzy c-means (FCM) is a clustering method that groups observations into N clusters and allows each observation to belong to each cluster with different membership degrees (Dunn, 1973; Bezdek, 1981). HCM can then be defined as FCM with a Boolean membership (i.e., each observation belongs to one cluster with a membership degree of 1, and to other clusters with member degrees equal to 0).

Our second solution to the discovery of the CNV problem uses fuzzy C-means clustering to partition the input into segments of similar neighboring points. Pre-processing the input requuires the application of smoothing, interpolating and curve fitting alternative steps as in the FIS-based solution. However, the differentiation herein aims to catch the main changing parts of the curve and the number of clusters (Fig. 2).

Fig. 3 shows the main steps and processing components of the FCM-based solution for CNVs discovery.

While FIS provide ease of application and transparency and flexibility in number of inputs and classes, it requires balancing the number of rules, which may grow precipitately. On the other hand, FCM algorithm reduces intra-cluster variances, but as the number of clusters increases, its efficiency depends on the parameters' initialization.

3. Results

The smoothing step greatly affects the result; therefore we experimented several smoothing algorithms to choose that which better serves the processing. What we were looking for is a smoothing algorithm that eliminates outliers and ignores small frequent fluctuations, but preserves the general pattern of the signal with a sufficient precision level. Fig. 4 shows a visual comparison of five smoothing algorithms: quadratic regression (loess), move mean, move median, Savitzky-Golay filter (golay), and linear regression (lowess).

The input observations were not uniformly spaced; thus, we could not evaluate the results at the intermediate positions (i.e., between the actual observations). To overcome this problem, we assumed that a function f(x) approximates the input data and can be relied on to represent all points even if they do not appear in the input. Such a function can be obtained by interpolation and curve fitting.

The intermediate date was estimated through interpolation, assuring that the interpolating function will typically pass through the given input data by fitting a different 3rd-degree polynomial between each pair of data points for curves (cubic spline).

Unlike interpolation, curve fitting does not guarantee that the fitted function will pass through the input data. Instead, a function is considered a good fit to the data in some sense by means of approximation.

The Sugeno fuzzy model was mainly selected for the following reasons:

1) linearity of the time complexity for defuzzifying the output, and
2) our application's requirement of constant values for the output, which is one of the attributes of the Sugeno model.

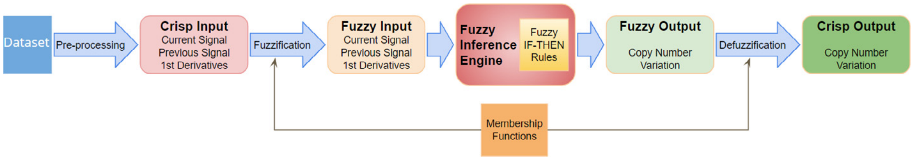
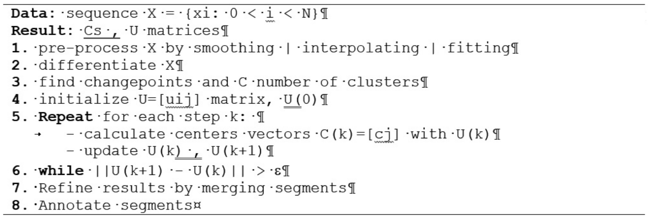


Fig. 1. Block diagram of the FIS-based solution.



```
Data:·sequence·X·=·{xi:·0·<·i·<·N}¶
Result:·Cs·,·U·matrices¶
1.·pre-process·X·by·smoothing·|·interpolating·|·fitting¶
2.·differentiate·X¶
3.·find·changepoints·and·C·number·of·clusters¶
4.·initialize·U=[uij]·matrix,·U(0)¶
5.·Repeat·for·each·step·k:·¶
  →   –·calculate·centers·vectors·C(k)=[cj]·with·U(k)¶
      –·update·U(k)·,·U(k+1)¶
6.·while·||U(k+1)·–·U(k)||·>·ε¶
7.·Refine·results·by·merging·segments¶
8.·Annotate·segments¤
```

Fig. 2. FCM-based algorithm for calling CNVs from array CGH.

Fig. 5 shows the structure of the surgeon fuzzy inference system with three input (i.e., $\log_2$ ratio of the Current locus, $\log_2$ ratio of the Previous locus and 1^{st} Derivative and) and one output (copy number Variation).

For the first and second inputs (Log_2 Ratio of the current locus, and the previous locus), Gaussian combination membership functions were chosen for the "Low" and "High" sets, while Gaussian distribution function was the membership function of the "Base" set. These selections were based on experimentation. For the third input (i.e., First Derivative), trapezoidal functions were chosen for both "Steady" and "Opima" sets.

A set of logical rules was compiled and tested. These rules mapped fuzzified input to the output (Table 1).

The experiments were conducted with three different preprocessing configurations. First, we smoothed the data using locally weighted scatter plot smoothing. In the second run, we used cubic spline interpolation to estimate the values between the input data. Finally, we used polynomial curve fitting (Table 2). Fig. 5 and Fig. 6 shows the results of applying the FIS-based solution to chromosome 17 of sample GM13031 and chromosome 9 of sample GM01750 respectively.

The setup in this experiment was similar to that of the FIS-based experiment. The three different preparatory processes of the input are as follows: locally weighted scatter plot smoothing, cubic spline interpolation, and polynomial curve fitting.

The absolute values of the normalized first derivative were used to find the major changepoints. The small changes in the derivatives and minor local extrema were suppressed. In contrast, the higher derivative values were boosted and magnified. The precise locations of the change point were not of interest. The processing step mainly aims to estimate the number of differentiated pieces of data fed as a number of clusters to the next step.

The fuzzy C-means clustering was performed with the following parameters:

N : number of clusters estimated in the previous step;

E : exponent of the fuzzy membership matrix U used to control the amount of fuzzy overlap between clusters; we experimented with E = 2 and E = 3, and the reported results corresponded to the latter;

I : maximum number of iterations set to 100; and

ε : minimum improvement in the objective function between two consecutive iterations set to 0.00001

Fig. 7 and Fig. 8 show the results of applying the FIS-based solution to chromosome 17 of sample GM13031 and chromosome 9 of sample GM01750 respectively.

4. Discussion

When we evaluated the results of our method, we adopted the following three criteria borrowed from edge detection in the images (Canny, 1986):

1) Low rate for false positives and false negatives.
2) Minimal distance between the detected breakpoint and the center of the edge.
3) Each breakpoint must be called once and once only.

First, we compared our findings with the published verifications of the studied dataset. The online supplemental material of (Snijders et al., 2001) summarizes the CNVs in the curated dataset. We reported the type (loss or gain) and the start and end positions of the segment for each detected CNV to perform an assessment based on the second criterion. We also calculated the distance between our detected boundaries and the verified ones. Futhermore, we calculated the ratio of the error to the segment size to put the distance error in perspective when assessing compliance with the criteria.

Note that the distance was calculated according to the positions stored in the dataset. That is, we found the distance between the breakpoint resulted in our approach and the closest point in the dataset to the published breakpoint.

As shown, no false positives existed. Our method found all CNVs except for the deletion of chromosome 15 of sample (GM07081). Other methods also encountered this alteration (Hupé et al., 2004). We believe that the microarray processing of this sample did not yield an adequate representation of the genetic content. Access to the detail results of the actual sample triplicates is needed to confirm this speculation. This option is not available in the time being.

Table 3 presents the distance error in terms of the number of missing probes between the breakpoint we found and the published one. Probes are not evenly distributed on the genomic strand. Moreover, the dataset suffers from missing data. In this case, the number of probes that the algorithm skipped would more accurately characterize the error. Fig. 9 explains this point.

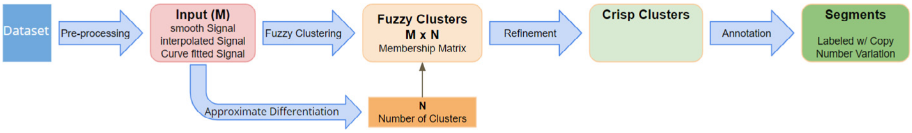


Fig. 3. Block diagram of the FCM-based solution.

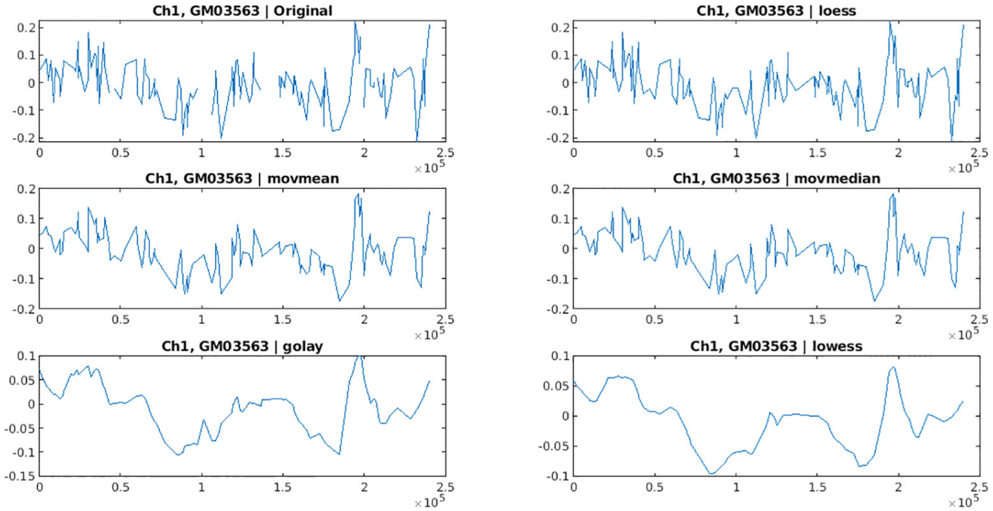


Fig. 4. Comparison of five smoothing methods.

Table 1
Fuzzy rules for the inference system.

		Current		Previous		Derivative		Variation
01	IF	LOW	AND	LOW			THEN	LOSS
02	IF	BASE	AND	BASE			THEN	DIPLOID
03	IF	HIGH	AND	HIGH			THEN	GAIN
04	IF	LOW	AND	BASE	AND	OPTIMA	THEN	LOSS
05	IF	HIGH	AND	BASE	AND	OPTIMA	THEN	GAIN
06	IF	LOW	AND	BASE	AND	STEADY	THEN	DIPLOID
07	IF	HIGH	AND	BASE	AND	STEADY	THEN	DIPLOID
08	IF	BASE	AND	LOW	AND	OPTIMA	THEN	DIPLOID
09	IF	HIGH	AND	LOW	AND	OPTIMA	THEN	HIGH
10	IF	BASE	AND	LOW	AND	STEADY	THEN	LOW
11	IF	HIGH	AND	LOW	AND	STEADY	THEN	LOW
12	IF	LOW	AND	HIGH	AND	OPTIMA	THEN	LOW
13	IF	BASE	AND	HIGH	AND	OPTIMA	THEN	DIPLOID
14	IF	LOW	AND	HIGH	AND	STEADY	THEN	HIGH
15	IF	BASE	AND	HIGH	AND	STEADY	THEN	HIGH

Table 2
Distance error means for boundaries of the variation segments.

	FIS			FCM		
	smooth	interpolate	fitting	smooth	interpolate	fitting
DF	21					
Error mean	7.87%	16.83%	16.17%	9.99%	11.14%	20.92%
Error SD	0.420	−0.168	−0.162	−0.100	−0.111	−0.209
t-value	1.791	1.842	1.822	2.141	1.822	2.256
p-value	0.044	0.040	0.041	0.022	0.044	0.0174

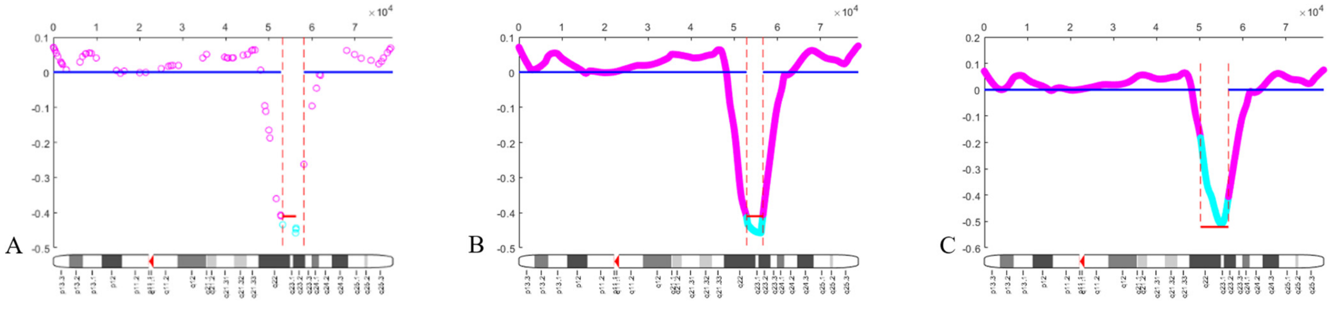


Fig. 5. Application of the FIS-based solution to the A) smoothed, B) interpolated and, C) curve fitted, data of chromosome 17 of sample 13031.

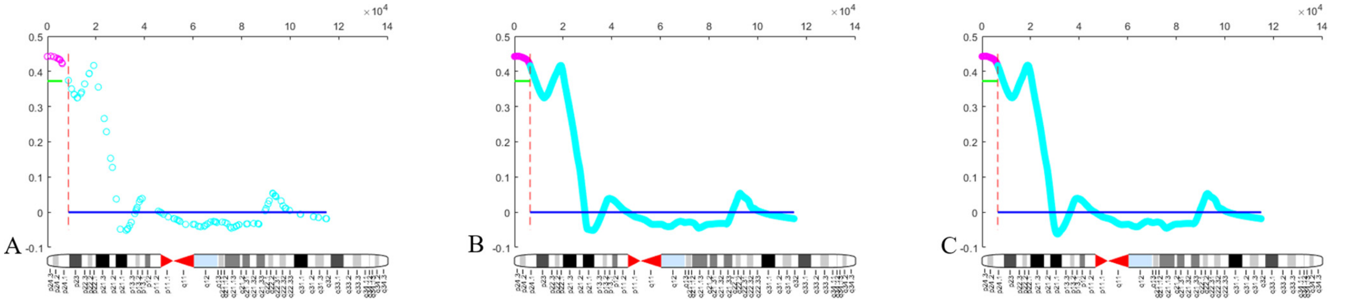


Fig. 6. Application of the FCM-based solution to the A) smoothed, B) interpolated and, C) curve fitted, data of chromosome 9 of sample GM01750.

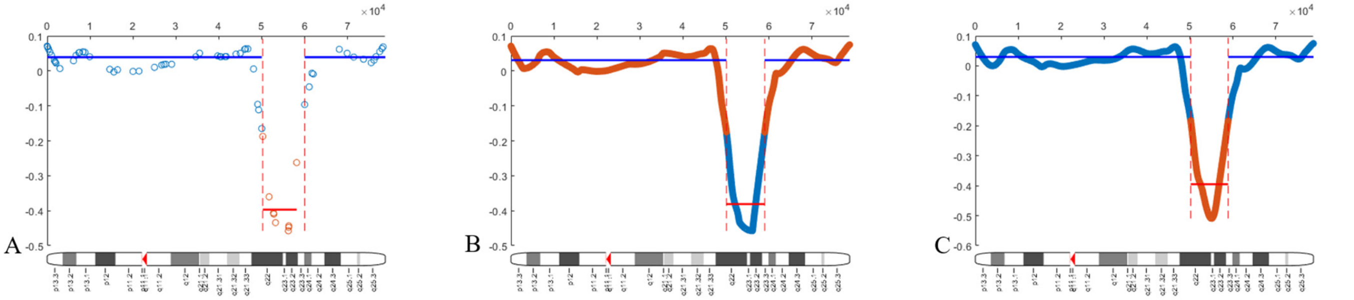


Fig. 7. Application of the FCM-based solution to the A) smoothed, B) interpolated and, C) curve fitted, data of chromosome 17 of sample GM13031.

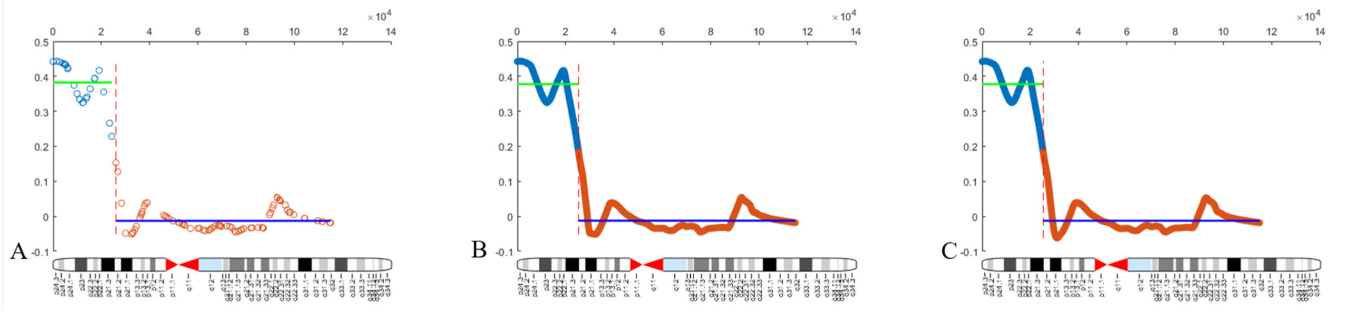


Fig. 8. Application of the FCM-based solution to the A) smoothed, B) interpolated and, C) curve fitted, data of chromosome 9 of sample GM01750.

Table 3
Distance error expressed in terms of missing probes.

Sample	Ch	Distance (kb)	Distance (probes)
01GM03563	9	2528	1
03GM05296	10	2780	2
03GM05296	10	1500	2
03GM05296	11	1490	2
05GM01750	9	9144	9
05GM01750	14	28,030	19
06GM03134	8	3940	3
07GM13330	1	11,500	11
07GM13330	4	4900	12
09GM01535	5	11,200	12
10GM07081	7	2100	3
14GM13031	17	1050	1
15GM01524	6	4190	2
15GM01524	6	8700	5

We compared our results with three other methods that used the same benchmark. Table 4 shows that our fuzzy approaches outperformed the considered methods in all comparison factors: true positives, false positives and false negatives.

We also calculated the precision, recall, and F-measure to better demonstrate how our solutions perform compared to the selected methods (Table 5).

5. Conclusions

The literature review showed that array CGH is still a potential platform for the discovery and analysis of copy number variations. Furthermore, the current detection methods do not offer a precise efficient solution to the problem, and there is plenty of room for new solutions. In general, fuzzy inference systems are important tools in segmentation problems.

We designed a fuzzy-based algorithm to detect the copy number variations in the array CGH data. The results that our solutions can detect all known CNVs, except for one alteration known to be problematic. Our methods did not report any false detection. Furthermore, they demonstrated the highest recall of 96% and the highest F-measure of 98% when compared to the prevailing method in each category of the aCGH-based CNV detection methods. The performance of our solution was outstanding because most of the calculations were done in linear time.

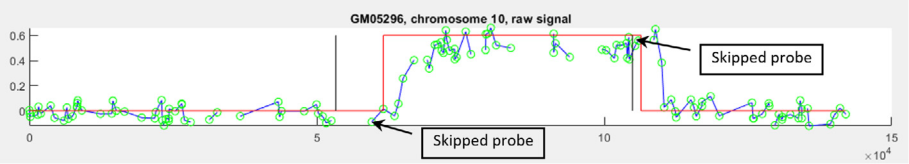


Fig. 9. Counting skipped probes as distance errors.

Table 4
Comparison of our results with those of other methods in terms of true positives, false positives and false negatives.

	Fuzzy	Hupé et al	CBS	CRF_CNV
Discovered CNVs	22	36	28	21
True positives	22	21	12	20
False positives	0	15	16	0
False negatives	1	2	9	2

Table 5
Comparison of our results with those of other methods in terms of precision, recall and F-measure.

	Fuzzy	Hupé et al	CBS	CRF_CNV
Precision	1	0.58	0.43	0.95
Recall	0.96	0.91	0.52	0.87
F-measure	0.98	0.71	0.47	0.91

6. Future work

This work showed methods that were proved successful in similar problems in other domains, which can be exploited and have the potential to excel. In the future, we intend to investigate more solutions that were studied for problems like segmentation and detecting breakpoints in domains, such as data and time series and images. The proposed fuzzy-based method can also be applied to other platforms. One important area to investigate it to support the detection of CNVs from array CGH by including related data from other studies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Deanship of Scientific Research for funding and supporting this research through the initiative of DSR Graduate Students Research Support (GSR), and RSSU at King Saud University for their technical support.

References

Autio, R., Hautaniemi, S., Kauraniemi, P., Yli-Harja, O.P., Astola, J., Wolf, M., Kallioniemi, A., 2003. CGH-Plotter: MATLAB toolbox for CGH-data analysis. Bioinformatics 19, 1714–1715. https://doi.org/10.1093/bioinformatics/btg230.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms.

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8 (6), 679–698.

Carter, N.P., 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet. 39 (7), 16–21.

Chaira, T., 2010. Intuitionistic Fuzzy Segmentation of Medical Images. IEEE Trans. Biomed. Eng. 57 (6), 1430–1436.

Dunn, J.C., 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. J. Cybern. 3 (3), 32–57. https://doi.org/10.1080/01969727308546046.

Eilers, P.H.C., 2003. A perfect smoother. Anal. Chem. 75, 3631–3636. https://doi.org/10.1021/ac034173t.

Eilers, P.H.C., de Menezes, R.X., 2005. Quantile smoothing of array CGH data. Bioinformatics 21, 1146–1153. https://doi.org/10.1093/bioinformatics/bti148.

Flores, M., Hsiao, T.H., Chiu, Y.C., Chuang, E.Y., Huang, Y., Chen, Y., 2013. Gene regulation, modulation, and their applications in gene expression data analysis. Adv. Bioinformatics 2013, 360678. https://doi.org/10.1155/2013/360678.

Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G., Jain, A.N., 2004. Hidden Markov models approach to the analysis of array CGH data. J. Multivar. Anal. 90, 132–153. https://doi.org/10.1016/j.jmva.2004.02.008.

Guan, P., Sung, W.-K., 2016. Structural variation detection using next-generation sequencing data: a comparative technical review. Methods 102, 36–49. https://doi.org/10.1016/j.ymeth.2016.01.020.

Haq, I., Anwar, S., Shah, K., Khan, M.T., Shah, S.A., 2015. Fuzzy Logic Based Edge Detection in Smooth and Noisy Clinical Images. PLOS ONE 10 (9).

Hupé, P., Stransky, N., Thiery, J.-P., Radvanyi, F., Barillot, E., 2004. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20, 3413–3422. https://doi.org/10.1093/bioinformatics/bth418.

Yin, X.-L., Li, J., 2010. Detecting copy number variations from array CGH data based on a conditional random field model. J. Bioinform. Comput. Biol. 8 (2), 295–314.

Jong, K., Marchiori, E., van der Vaart, A., Ylstra, B., Weiss, M., Meijer, G., 2003. Chromosomal Breakpoint Detection in Human Cancer, in: Cagnoni, S., Johnson, C.G., Cardalda, J.J.R., Marchiori, E., Corne, D.W., Meyer, J.-A., Gottlieb, J., Middendorf, M., Guillot, A., Raidl, G.R., Hart, E. (Eds.), Applications of Evolutionary Computing, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 54–65.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., Pinkel, D., 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258, 818–821. https://doi.org/10.1126/science.1359641.

Kendall, K.M., Rees, E., Bracher-Smith, M., Legge, S., Riglin, L., Zammit, S., O'Donovan, M.C., Owen, M.J., Jones, I., Kirov, G., Walters, J.T.R., 2019. Association of rare copy number variants with risk of depression. JAMA Psychiatry. https://doi.org/10.1001/jamapsychiatry.2019.0566.

Macé, A., Kutalik, Z., Valsesia, A., 2018. Copy number variation. Methods Mol. Biol. 1793, 231–258. https://doi.org/10.1007/978-1-4939-7868-7_14.

Mahmud, M.P., Schliep, A., 2011. Fast MCMC sampling for hidden Markov models to determine copy number variations. BMC Bioinf. 12, 428. https://doi.org/10.1186/1471-2105-12-428.

Marioni, J.C., Thorne, N.P., Tavaré, S., 2006. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics 22, 1144–1146. https://doi.org/10.1093/bioinformatics/btl089.

Montgomery, S.B. et al., 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res. 23, 749–761. https://doi.org/10.1101/gr.148718.112.

Moscato, P., 1989. On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts - Towards Memetic Algorithms.

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 5, 557–572. https://doi.org/10.1093/biostatistics/kxh008.

Perry, G.H., 2008. The evolutionary significance of copy number variation in the human genome. Cytogenet Genome Res 123, 283–287. https://doi.org/10.1159/000184719.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., Daudin, J.-J., 2005. A statistical approach for array CGH data analysis. BMC Bioinf. 6, 27. https://doi.org/10.1186/1471-2105-6-27.

Polzehl, J., Spokoiny, V.G., 2000. Adaptive weights smoothing with applications to image restoration. J. R. Stat. Soc. 62, 335–354. https://doi.org/10.1111/1467-9868.00235.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., González, J.R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, Junjun, Zerjal, T., Zhang, Jane, Armengol, L., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W., Hurles, M.E., 2006. Global variation in copy number in the human genome. Nature 444, 444–454. https://doi.org/10.1038/nature05329.

Shah, J., Patel, N., Tandel, H., Soni, N., Prajapati, G., 2013. A Hybrid Approach For Edge Detection Using Fuzzy Logic And Canny Method. Int. J. Eng. Res. Technol. 2, 4.

Shah, S.P., Xuan, X., DeLeeuw, R.J., Khojasteh, M., Lam, W.L., Ng, R., Murphy, K.P., 2006. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 22, e431–e439. https://doi.org/10.1093/bioinformatics/btl238.

Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D., Fan, X., 2019. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med. Genet. 20, 1–14. https://doi.org/10.1186/s12881-019-0909-5.

Sheha, M.A., Mabrouk, M.S., Elhefnawi, M., 2016. Detecting and analyzing copy number alternations in array-based cgh data. Biomed. Eng. Appl. Basis Commun. 28 (6).

Shlien, A., Malkin, D., 2009. Copy number variations and cancer. Genome Med. 1, 62. https://doi.org/10.1186/gm62.

Stamoulis, C., Betensky, R.A., 2016. Optimization of signal decomposition matched filtering (SDMF) for improved detection of copy-number variations. IEEE/ACM Trans. Comput. Biol. Bioinform. 13, 584–591. https://doi.org/10.1109/TCBB.2015.2448077.

Snijders, A.M., Nowak, N., Segraves, R., et al., 2001. Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 29, 263–264. https://doi.org/10.1038/ng754.

Stamoulis, C., Betensky, R.A., 2011. A novel signal processing approach for the detection of copy number variations in the human genome. Bioinformatics 27, 2338–2345. https://doi.org/10.1093/bioinformatics/btr402.

Usher, C.L., McCarroll, S.A., 2015. Complex and multi-allelic copy number variation in human disease. Brief. Funct. Genomics 14, 329–338. https://doi.org/10.1093/bfgp/elv028.

Venkatraman, E.S., Olshen, A.B., 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23, 657–663. https://doi.org/10.1093/bioinformatics/btl646.

Wang, K., Yu, X., Jiang, Hongwei, Huang, J., Wang, H., Jiang, Hongyu, Wei, S., Liu, L., 2019a. Genome-wide expression profiling-based copy number variations and colorectal cancer risk in Chinese. Mol. Carcinog. 58, 1324–1333. https://doi.org/10.1002/mc.23015.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., Tibshirani, R., 2005. A method for calling gains and losses in array CGH data. Biostatistics 6, 45–58. https://doi.org/10.1093/biostatistics/kxh017.

Wang, Y., Li, Y., Chen, Y., Zhou, R., Sang, Z., Meng, L., Tan, J., Qiao, F., Bao, Q., Luo, D., Peng, C., Wang, Yaoshen, Luo, C., Hu, P., Xu, Z., 2019b. Systematic analysis of copy-number variations associated with early pregnancy loss. Ultrasound Obstet. Gynecol. https://doi.org/10.1002/uog.20412.

Wiedenhoeft, J., Brugel, E., Schliep, A., 2016. Fast bayesian inference of copy number variants using hidden markov models with wavelet compression. PLoS Comput. Biol. 12. https://doi.org/10.1371/journal.pcbi.1004871.

Yan, Y.-X., Li, J.-J.-H., Xiao, H.-B., Wang, S., He, Y., Wu, L.-J., 2018. Association analysis of copy number variations in type 2 diabetes-related susceptible genes in a Chinese population. Acta Diabetol. 55, 909–916. https://doi.org/10.1007/s00592-018-1168-1.

Yao, R., Yu, T., Qing, Y., Wang, J., Shen, Y., 2018. Evaluation of copy number variant detection from panel-based next-generation sequencing data. Mol. Genet. Genomic Med. 7. https://doi.org/10.1002/mgg3.513.