

# Frequent Itemsets for Genomic Profiling

Jeannette M. de Graaf<sup>1</sup>, Renée X. de Menezes<sup>2,3</sup>,  
Judith M. Boer<sup>2</sup>, and Walter A. Kusters<sup>1</sup>

<sup>1</sup> Leiden Institute of Advanced Computer Science,  
Universiteit Leiden, Leiden, The Netherlands  
{`graaf`, `kusters`}@liacs.nl

<sup>2</sup> Center for Human and Clinical Genetics,  
Leiden University Medical Center, Leiden, The Netherlands  
{`r.x.menezes`, `j.m.boer`}@lumc.nl

<sup>3</sup> Laboratory of Pediatrics, Erasmus Medical Center,  
Rotterdam, The Netherlands

**Abstract.** Frequent itemset mining is a promising approach to the study of genomic profiling data. Here a dataset consists of real numbers describing the relative level in which a clone occurs in human DNA for given patient samples. One can then mine, for example, for sets of samples that share some common behavior on the clones, i.e., gains or losses. Frequent itemsets show promising biological expressiveness, can be computed efficiently, and are very flexible. Their visualization provides the biologist with useful information for the discovery of patterns. Also it turns out that the use of (larger) frequent itemsets tends to filter out noise.

## 1 Introduction

*Frequent itemsets* are often used in Data Mining research [11]; they can supplement the more traditional statistical approach [2]. The concept is simple, many efficient algorithms are devised to detect different types of frequent itemsets, and there is a rich literature describing associated topics. For instance, many researchers dealt with the problem of finding interesting sets, and the fuzzy logic approach also gave a new impetus. The most well-known application is in the area of market-basket analysis. In this case a frequent itemset is a set of products that is often purchased together. From such a set one can easily deduce *association rules* of the form “if one buys  $X$ , one (often) buys  $Y$  too”.

In this paper we apply the frequent itemset approach to explore copy number changes in the genome. Chromosomal instability in tumors leads to DNA copy number alterations with associated gain or loss of genes important in tumor development [5]. Array-based comparative genomic hybridization (array CGH) allows for high-throughput genome-wide screening of these DNA copy number changes [1,7,9]. Typically, these experiments involve co-hybridization of a few hundred fluorescently labeled patient DNA samples with normal reference DNA onto microarrays containing several thousands of large-insert genomic clones

(relatively short pieces of DNA) such as bacterial artificial chromosomes (BACs). The resulting dataset is a database of clones, each consisting of a few hundred real numbers. Any such number describes the normalized log<sub>2</sub>-ratio of the number of clone copies found in a given patient sample compared with the reference DNA. When there is no copy number change in the patient DNA, the log<sub>2</sub>-ratio is expected to be 0 (no change). When the log<sub>2</sub>-ratio lies above a certain threshold or below another fixed threshold, the patient has a *gain* or a *loss*, respectively, for this clone. In principle, the boundaries between no change and change are very strict, allowing discretization of the data. However, factors such as tissue heterogeneity (i.e., a loss or a gain is present in only a subset of the cells) and the use of amplification procedures introduce more variation in measurements, making such boundaries less strict. And finally, we have the usual problems like measurement noise.

The database records can be viewed in (at least) two different ways. First, one can look at the clones as transactions, and view the samples as items; this is called here the *frequent sample sets model*. Note that this is the way in which the data is usually presented. Second, one can also see the samples as transactions, and the clones as items; this we call the *frequent clone sets model*. In this paper we treat both approaches, with emphasis on the first one. If we adhere to the first choice, we are interested in groups of samples, where the group elements share some common behavior; for the second choice, we try to find associations between clones. We shall provide many examples of the use of frequent itemsets in this biological setting.

For related work we refer to [8], where — among other things — minimal and related gain and loss zones are detected using frequent pattern mining. In [10] a method is discussed that deals with finding interesting association rules. The first step is to generate frequent itemsets. From these one can deduce a huge amount of association rules. The authors deal with a method to filter out, after all rules are discovered, the most interesting rules for biologists.

In the current paper we generate the frequent itemsets and extract useful information from those. We also show that frequent itemsets can be used to reduce the effect of noise. We mainly focus on visualizations, which are easily made and from which biologists can deduce information about certain relations between clones or patient samples. Our method is meant to be used as an exploratory tool, aiming at pattern discovery, that can be used in combination with other methods.

We shall not treat the database in any detail, but rather refer to the paper where it originates from [6]. In a few places we shall provide the necessary biological background, and we mention the biological consequences of the proposed methods. Anyway, there is a lot of data preparation involved, apart from some trivial data cleaning. In particular we mention the problem of distinguishing change from no change, as mentioned above, which is both of technical as well as biological nature.

The paper is organized in the following way. We first describe the method and illustrate it by using artificial data (Section 2 and Section 3), for both models

mentioned earlier. In Section 4 we apply the techniques to the real life data from [6] and focus on the biological consequences of the proposed methods. We end with some conclusions and issues for further research.

## 2 Frequent Itemsets

Suppose we have a dataset  $\mathcal{D}$  consisting of subsets (usually called *itemsets*) of a given finite set  $\mathcal{I}$ . The subsets have unique identifiers, so multiple occurrences of the same subset may appear. It is also possible to consider the dataset as a (time ordered) series, but this viewpoint is not taken here.

For any subset  $S$  of  $\mathcal{I}$  we define its *support* as the number of elements in  $\mathcal{D}$  that contain  $S$ . An itemset is called *frequent* if its support is larger than or equal to a pre-given support threshold *minsup*. If an itemset has  $k$  elements, it is called a  $k$ -itemset.

The first main problem in frequent itemset mining is to find all frequent itemsets for a given  $\mathcal{D}$  and *minsup*. There exist many efficient implementations to tackle this problem. The fastest ones rely on so-called FP-trees and use the Apriori property [11]. For the experiments we used the implementation from [3].

In this paper we focus on data from array CGH studies. Here the original database consists of real numbers, but it is discretized to a database describing if a sample has a gain on a clone or not, or to a database showing if a sample has a loss on a clone or not. This is done because in CGH analysis one is often more interested in whether or not a patient has a gain (loss) at some clone, and not in the exact value. So the database consists of itemsets that are either sets of samples that have higher (lower) value than normal for a given clone, or sets of clones that have higher (lower) value than normal for a given sample. depending on whether we are more interested in the gains or the losses. In the first case, the clone is the identifier of the itemset, in the second case the sample is the identifier. One can think of the database as a two-dimensional array where rows correspond to clones and columns to samples (or the other way round in the second case). The transformed database contains only zeros and ones. If a clone occurs more (less) than normal for a given patient (its value being higher (lower) than some threshold), it is assigned a one on the corresponding array position, otherwise a zero.

## 3 Simulated Data

A dataset with similar structure to the one from array CGH studies was simulated as follows. A total of 150 samples, with 3200 observations per sample, are divided into three main groups of 50 samples each. Samples in each group are characterized by having in common a specific copy number effect in one of the chromosomes, as well as other effects in other chromosomes, as summarized in Table 1. The effect is assumed to hold for a given number of consecutive clones (shown between brackets) at the beginning of the affected chromosomes.

These three groups can be thought of as referring to patients with the same disease, but different genotypes. This is observed for example in many cancers, where various genotypic mechanisms can lead to the same result, as in the same kind of cancer. It is important to identify these different mechanisms since they often are associated with varying susceptibility levels to treatments and, as a consequence, varying chances of recovery. Sometimes these mechanisms share part of their structure, but differ in other parts.

Unaffected clone intensities are assumed to be independent of each other and to follow a normal distribution with mean 0 and standard deviation 1. Affected clone intensities are also assumed to be independent of each other and have a normal distribution with standard deviation 1, but their mean is taken as either 3 (if effect is a gain) or  $-3$  (if effect is a loss).

**Table 1.** Summary of simulated effects (G = gain, L = loss)

Samples affected	Chromosomes								Gains/losses (total)
	1	3	7	10	11	13	18	20	
136–150		G(60)		G(40)			G(30)		130/0
121–135		G(60)				L(50)			60/50
101–120		G(60)							60/0
91–100									0/0
76–90	L(80)				L(60)				0/140
61–75	L(80)		G(50)						50/80
51–60	L(80)								0/80
36–50		G(60)					G(30)	G(20)	110/0
21–35		G(60)			L(50)				60/50
1–20		G(60)							60/0

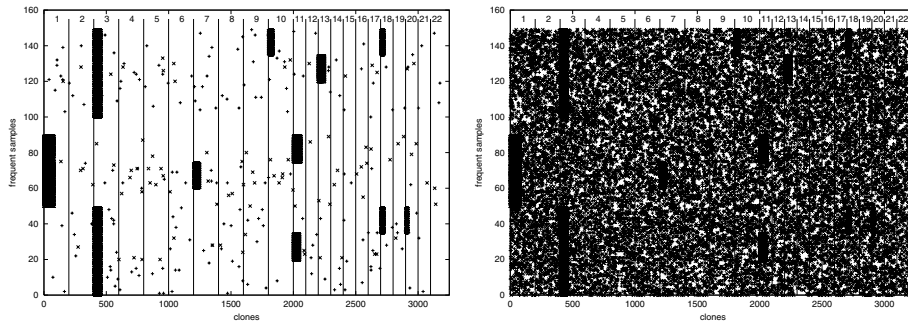
In order to evaluate the effect of having more or less noise in the data, we have also simulated a dataset with the same structure and effects, where the standard deviation of the measurements was 0.6 instead of 1. This dataset is referred to as the *ideal dataset*: it corresponds to an “ideal” scenario, where there is very good separation between measurements with copy number and without. Of course, in such a case no special method has to be used to identify effects. In practice, however, it is more common to observe datasets with less perfect separation, as the first one. This dataset is called the *noisy dataset*.

### 3.1 The Frequent Sample Sets Model

We now regard the database as an ordered series of 3200 clones. Each record (i.e., clone, transaction) consists of 150 real numbers, corresponding to the samples (patients). As mentioned before we first transform the database into a database of zeros and ones after defining suitable thresholds for gains and losses.

In order to obtain insight in the data, and also to give a first (simple) application of frequent itemsets, in Figure 1 we show all frequent samples, i.e.,

1-itemsets. In the left hand side picture we have the ideal dataset, in the right hand side picture the noisy dataset. In a sense, these pictures give simple snapshots of the entire dataset: the picture on the left clearly reflects Table 1, while the dense regions of +’s and ×’s in the picture on the right also do so, but less convincing. The vertical lines denote the chromosome boundaries, with the chromosome numbers on top. We show the 1-itemsets for gains and the 1-itemsets for losses in one picture; gains (value  $> 2.0$ ) have +’s, losses (value  $< -2.0$ ) have ×’s. If a 1-itemset  $\{i\}$  has at least  $minsup = 30$  gains, those gains are plotted horizontally at  $y$ -level  $i$ , and similarly for the losses. For example, sample 80 has a series of ×’s for chromosomes 1 and 11, and single ×’s for clone 1306 (in chromosome 7) and clone 1623 (in chromosome 9). For the ideal dataset there are 115 frequent 1-itemsets for gains (meaning there are 115 patient samples having a gain on at least 30 clones), and 70 for losses — as expected. The noisy dataset has 150 frequent 1-itemsets for both, or equivalently: every 1-itemset is frequent here, so every patient has at least 30 gains and 30 losses!



**Fig. 1.** Frequent samples (1-itemsets) for ideal (left) and noisy (right) dataset; gains (+) and losses (×)

The left hand side picture clearly shows the expected effects, the right hand side picture is much more diffuse. The supports on the left (the numbers of +/×’s in a single row) are smaller and the distribution is more crisp.

Note that these two pictures are the only ones that contain two types of itemsets in one image. In order to be frequent, an itemset should have at least some minimum number of gains or losses (but not together). In the sequel we also mention “combined gains and losses”, which means that we add the numbers of gains and losses.

We now try to find larger sets of samples that share some common behavior, i.e., we look at  $k$ -itemsets with  $k > 1$ . We first examine gains; we let  $minsup = 60$ . In the plots from Figure 2 we depict the frequent 2-itemsets. Every horizontal series of +’s indicates the clones that have gains for *both* samples in the set. The frequent itemsets are depicted in the order in which they are generated by the algorithm from [3]; roughly speaking, larger supports occur for the higher

numbered sample sets. Neighbouring sample sets usually have a non-empty intersection in this order (which is not the case if they are ordered by support). Again, the left hand side picture is for the ideal dataset. In this case we have 443 frequent 2-itemsets; the 2-itemset  $\{138, 140\}$  has the highest support: 122. This means that there are 122 clones on which sample 138 and sample 140 both have gains. Note that the gains series on chromosome 7 is not visible, since its length (50 clones) is smaller than *minsup* and samples 61–75 have no other gains. Therefore no combination of two samples from 61–75 (the only samples that have gains on chromosome 7) can reach the threshold 60. Furthermore, the samples 61–75 will not occur in any of the 443 frequent 2-itemsets.

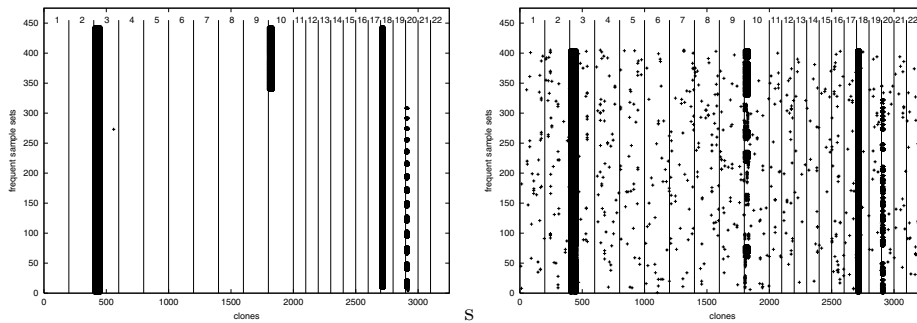


Fig. 2. Frequent 2-itemsets for ideal (left) and noisy (right) dataset; gains

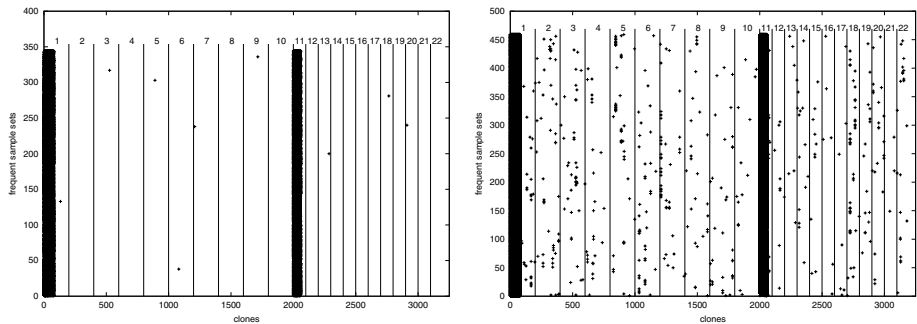


Fig. 3. Frequent sample sets (3-itemsets) for noisy dataset; losses; left: loss if value  $< -2.0$ , right: loss if value  $< -1.5$

For the noisy dataset (Figure 2, right) there are 405 frequent 2-itemsets; the 2-itemset  $\{139, 141\}$  has the highest support: 105. Like in the ideal case we do not see the gains at chromosome 7 here either.

This example also reveals that a larger value of the size of the itemsets allows for better pictures, in particular for the noisy case. Patterns are much more visible now. In a next step one might decide to study chromosomes 3, 10,

18 and 20 in more detail, e.g., using the same method again on these specific chromosomes.

As a final picture we show the frequent 3-itemsets (at least having 80 common losses) for the noisy dataset, see Figure 3. In the left plot we have a loss if the dataset value is smaller than  $-2.0$  (344 itemsets), in the right plot if the value is smaller than  $-1.5$  (459 itemsets). This shows the dependence on the threshold defining gains/losses. Note that losses on chromosome 13 are not visible, because they only occur on 50 clones, and for samples which do not have losses elsewhere. It appears that the itemsets show a lot of overlap — a phenomenon that emerges even more for larger values of the itemset size.

### 3.2 The Frequent Clone Sets Model

As said in the introduction, we can also look at the database as being a series of samples. In that case we are interested in sets of clones that behave in a similar way, e.g., are all gains on at least some *minsup* common samples.

The picture below (Figure 4, left), which is just a “random” example, shows the 1136 frequent 7-itemsets (so each itemset consists of 7 clones) where each element has a value larger than 2.0 on at least *minsup* = 50 common (among the 7 elements) samples, for the noisy dataset. On the right the 7 elements are plotted for these sets. As observed above, there is a lot of overlap present here. Furthermore note that only the clones at the beginning of chromosome 3 are gains for at least 50 patient samples, which is consistent with Table 1.

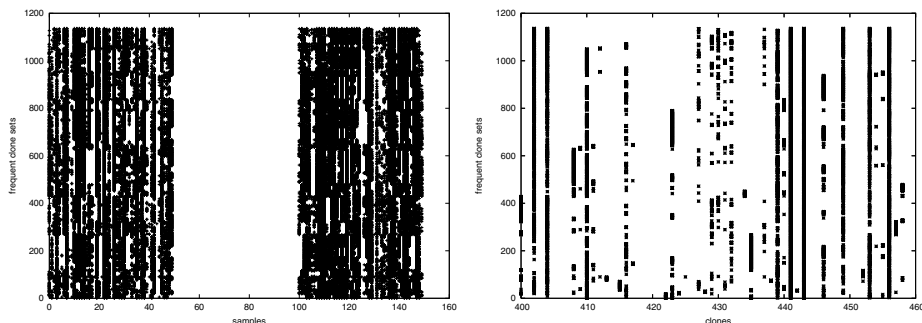


Fig. 4. Frequent clone sets (7-itemsets) for noisy dataset; gains; right: the set elements

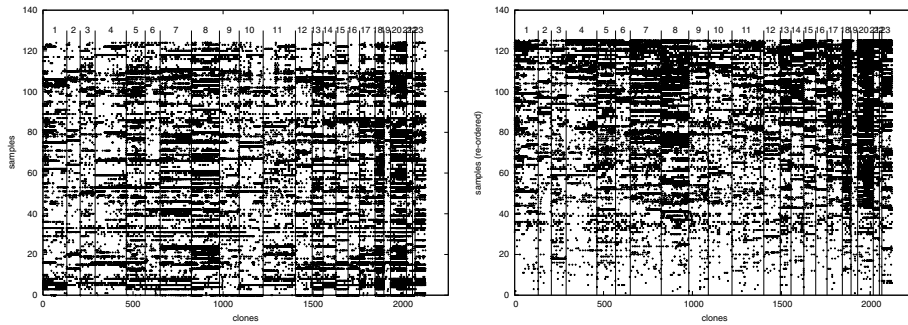
## 4 Application to Colon Cancer Data

Nakao et al. [6] analyzed copy number changes in the genomes of 125 colorectal tumors using array CGH on microarrays containing 2463 BAC clones that covered the human genome at 1.5 Mb resolution. Their publicly available dataset contains normalized log<sub>2</sub>-ratios for 2124 clones (after filtering), located on chromosomes 1–22 and the X-chromosome (here referred to as 23). In this dataset

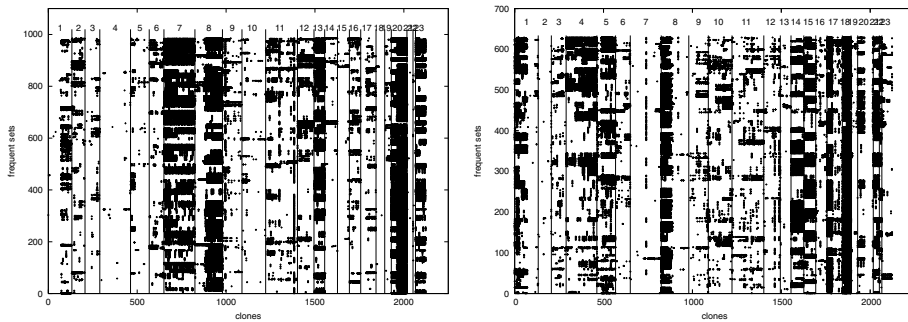
any value larger than 0.225 is considered as a gain, any value smaller than  $-0.225$  is considered as a loss. This threshold corresponds to values between 2 and 3 standard deviations from the mean. The total number of gains and losses varies between 2 and 1020 per sample.

The authors concluded that the majority of clones were infrequently gained or lost, with 95% of the changes occurring less than 35% of the time. However, high-frequency gains were detected on chromosomes 7p (35%), 7q (35%), 8q (42%), 11q (35%) and 20q (65%), and high-frequency losses were detected on 5q (35%), 8p (37%), 17p (46%), 18p (49%), 18q (60%), and 21q (35%). The distribution of alterations over the individual patients was not explored.

In Figure 5 we depict all 125 1-itemsets (combined gains and losses). In the left panel the samples are shown in the order in which they occur in the original dataset; in the right panel they are ordered with respect to their support. The 1-itemset  $\{53\}$  has the highest support: 1020.



**Fig. 5.** All 1-itemsets, combined gains and losses; left: original order, right: ordered with respect to support



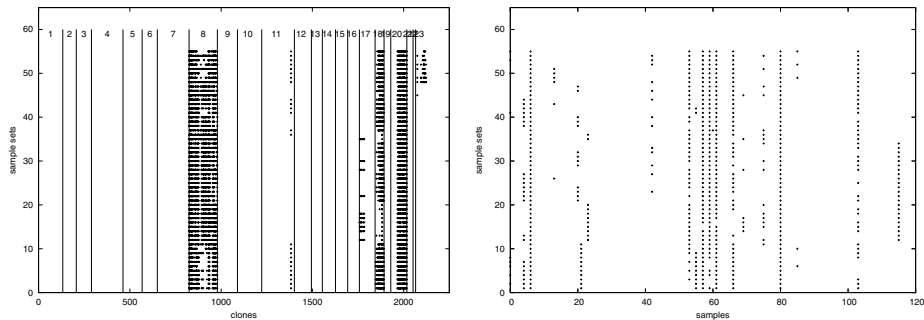
**Fig. 6.** Frequent sample sets; 2-itemsets; left: gains, right: losses

In Figure 6 we show the 985 frequent 2-itemsets for gains (left) and the 629 frequent 2-itemsets for losses (right), both for  $minsup = 100$ . Again a larger value of the itemset size gives rise to a clearer picture, showing e.g. common



regions of gains on chromosomes 7, 8, 13, 20 and 23, which is consistent with the conclusions in [6]. However, the results for the synthetic noisy dataset are more outspoken, due to the random nature of this set.

This becomes even more apparent if we consider the 55 10-itemsets ( $minsup = 100$ ), see Figure 7. To the left we see the usual plot, showing a very small region in chromosome 11 having a gain, also detected in [6]. This region was not so clear from Figure 6, showing the importance of studying larger itemsets and thus filtering out more noise. To the right we plot for each set its 10 elements. This picture shows that the sets have quite a lot in common. It could have been worse: in the current situation there are no 11-itemsets; the 10-itemsets are all *maximal* (i.e., all their supersets are infrequent) and hence *closed* (i.e., all their supersets have lower support). The number of frequent itemsets depends on their size and on the support threshold  $minsup$ , as shown in Table 2. It is a challenging task to find combinations that give rise to interesting visualizations.



**Fig. 7.** Frequent sample sets (10-itemsets); combined gains and losses; right: the set elements

From the biological viewpoint, the visualizations yield relevant information about the dataset, which is commonly hard to obtain. First, from Figure 5 no clear patterns emerge, a common feature of tumor samples. As they grow, tumors accumulate genomic changes, each tumor accumulating different changes. Most of these changes occurring during tumor growth are believed to be results of random processes, adding noise to the decisive changes that turned the tissue into a tumor in the first place. Then in Figure 6 (left) it shows that, by focusing on 2-itemsets with gains, the noise is filtered out and some patterns become evident, such as gains in chromosomes 7, 8, 13 and 20. By then progressively increasing the value of the itemset size, noise is step-by-step being filtered out and only the most consistent patterns remain. Indeed, only 20 of the 125 samples (16%) in the dataset contribute to the 10-itemsets represented in Figure 7, but these have a consistent pattern of copy number changes in chromosomes 8, 18 and 20. Also chromosomes 11 and 17 show some activity. Changes in chromosome 23 (the X chromosome) are often ignored, as they mostly indicate that the sample and the control are of opposite genders, which is not of main interest.

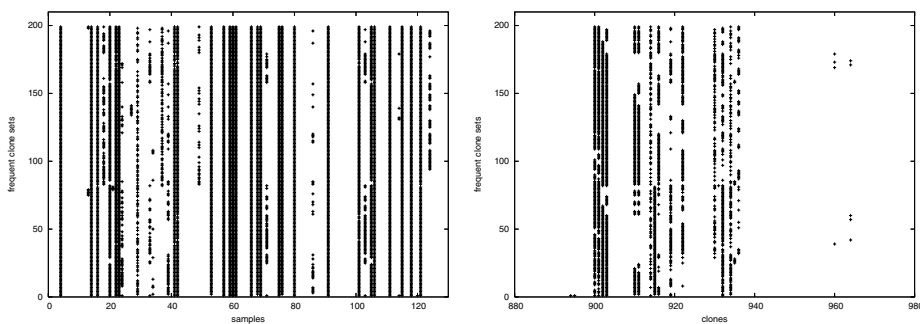
**Table 2.** Number of frequent itemsets for different size and *minsup*: gains/losses/combined gains and losses

Size	<i>minsup</i>		
	80	90	100
1	90/84/97	87/79/96	84/76/95
2	1519/1002/3196	1236/800/2942	985/629/2743
3	6281/2222/37417	3675/1282/29634	2036/726/23228
4	10135/1618/179576	4001/647/112866	1601/285/71539
5	8090/621/425627	2147/213/210119	546/79/103318
6	3692/185/581939	556/52/220138	39/12/83637
7	972/30/507966	55/4/148282	0/0/43049
8	155/1/300636	2/0/65081	0/0/12865
9	9/0/117955	0/0/16428	0/0/1739
10	0/0/27494	0/0/1864	0/0/55
11	0/0/3048	0/0/43	0/0/0
12	0/0/79	0/0/0	0/0/0

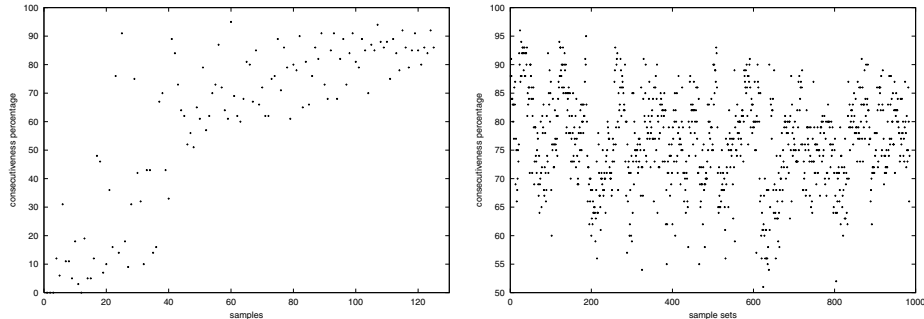
We now look at the frequent clone sets model. Experiments showed that chromosome 20 was really dominant. Taking into account only clones 1–1800, finer patterns on other chromosomes can be discovered. As an example we show the 199 9-itemsets for gains (Figure 8), with *minsup* = 30. The right picture has the set elements (cf. Figure 4), all on chromosome 8. The four neighbouring clones near 900 are indeed of biological interest.

It is possible to use the frequent itemset approach for the discovery of particular phenomena. For example, there is exactly one 4-itemset, the set of samples {53, 59, 66, 80}, having 300 or more common gains and losses.

If one keeps track of the distance between consecutive common gains (and/or losses) one can order the frequent itemsets found. For example, for the 4-itemset mentioned before, 69% of the 313 common gains and losses are really consecutive; if one allows for at most one intermediate normal clone (a so-called *gap*), this percentage rises to 87%. In Figure 9 above we plot these last percentages for all



**Fig. 8.** Frequent clone sets (9-itemsets); gains; right: the set elements



**Fig. 9.** Consecutiveness percentages; left: 1-itemsets, gains and losses; right: 2-itemsets, gains (sets on  $x$ -axis)

125 1-itemsets (gains and losses; ordered on the  $x$ -axis with respect to increasing support; left) and the 985 frequent 2-itemsets (gains;  $minsup = 100$ ; right). Efficiently incorporating consecutiveness into frequent itemset mining seems non-trivial and is left for future work.

## 5 Conclusion and Further Research

We presented a method to discover patterns in array CGH datasets. We make use of frequent itemset mining in order to obtain combinations of samples or clones that share some common behavior. The method is flexible, fast (the generation of a picture usually takes a few seconds), capable of dealing with noise, and allows for different types of post-processing. In contrast with many other techniques the method is largely unsupervised, and allows for individual patient tracking.

Once given the frequent itemsets, one can use many different Data Mining techniques. It is for instance possible to use Self Organizing Maps (SOMs) and the like in order to obtain visualizations. In Figure 10 the 55 10-itemsets from Figure 7 are embedded in the unit square, using a push-and-pull network [4]. The Euclidean distance between embedded data points in the plain resembles the “gains and losses distance”, obtained by squaring the difference in numbers of gains and losses on the different chromosomes.

We are very interested to extend the frequent itemset analysis to amplified array CGH data, which is more noisy due to reproducible ratio distortions resulting from differential processing of repetitive and polymorphic regions by the amplification enzyme [1]. In this dataset, the boundaries depend on the clone at hand, and new techniques are needed to deal with this varying boundary value issue. Perhaps fuzzy logic might be useful. We would also like to add clinical data such as stage of the tumor or age of the patient, expressed in association rules with attached interestingness measures. Finally, we will explore application of frequent itemsets to other types of genomic data, such as single nucleotide polymorphism genotyping data.

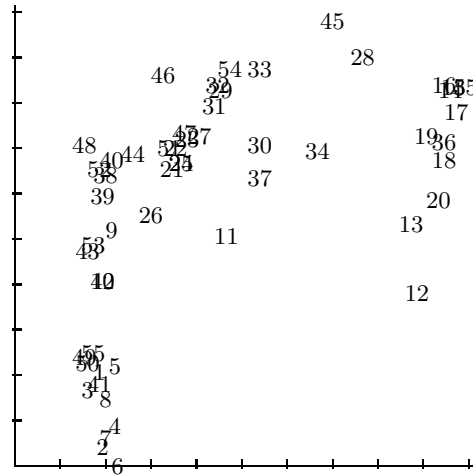


Fig. 10. Distance preserving embedding of the 55 10-itemsets from Figure 7

## References

1. J. Cardoso, L. Molenaar, R.X. de Menezes, C. Rosenberg, H. Morreau, G. Möslein, R. Fodde and J.M. Boer, Genomic Profiling by DNA Amplification of Laser Capture Microdissected Tissues and Array CGH, *Nucleic Acids Research* 32 (2004) e146.1–146.13.
2. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.
3. W.A. Kusters and W. Pijls, Apriori: A Depth First Implementation, FIMI'03, Workshop on Frequent Itemset Mining Implementations 2003; CEUR Workshop Proceedings (online; B. Goethals and M.J. Zaki (eds.)).
4. W.A. Kusters and M.C. van Wezel, Competitive Neural Networks for Customer Choice Models, pp. 41–60 in: *E-Commerce and Intelligent Models* (J. Segovia, P.S. Szczepaniak and M. Niedzwiedzinski (eds.)), Physica Verlag, Springer, 2002.
5. C. Lengauer, K. Kinzler and B. Vogelstein, Genetic Instabilities in Human Cancers. *Nature* 396 (1998) 643–649.
6. K. Nakao, K.R. Mehta, J. Fridlyand, D.H. Moore, A.N. Jain, A. Lafuente, J.W. Wiencke, J.P. Terdiman and F.M. Waldman, High-resolution Analysis of DNA Copy Number Alterations in Colorectal Cancer by Array-based Comparative Genomic Hybridization, *Carcinogenesis* 25 (2004) 1345–1357.
7. D. Pinkel, R. Segev, 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) 207–211.
8. C. Rouveirol and F. Radvanyi, Local Pattern Discovery in Array-CGH Data, Proceedings Dagstuhl Workshop on Detecting Local Patterns, (J.F. Boulicaut, K. Morik and A. Siebes (eds.)), to appear in *Lecture Notes in Artificial Intelligence*, Springer, 2005.

9. S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, L. Nickolenko, A. Benner, H. Dohner, T. Cremer and P. Lichter, Matrix-based Comparative Genomic Hybridization: Biochips to Screen for Genomic Imbalances, *Genes Chromosomes Cancer* 20 (1997) 399-407.
10. A. Tuzhilin and G. Adomavicius, Handling Very Large Numbers of Association Rules in the Analysis of Microarray Data, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 396-404, ACM Press, 2002.
11. C. Zhang and S. Zhang, *Association Rule Mining, Models and Algorithms*, Lecture Notes in Artificial Intelligence 2307, Springer, 2002.