Using Data Mining and OLAP to Discover Patterns in a Database of Patients with Y-Chromosome Deletions

Saso DZEROSKI[(1)], Dimitar HRISTOVSKI[(2)], Borut PETERLIN[(3)]
(1) Jozef Stefan Institute, Jamova 39, Ljubljana, Slovenia
(2)Institute of Biomedical Informatics, Faculty of medicine, University of Ljubljana, Slovenia
(3) Division of medical genetics, UMC, Slajmerjeva 3, Ljubljana, Slovenia

ABSTRACT

The paper presents a database of published Y chromosome deletions and the results of analyzing the database with data mining techniques. The database describes 382 patients for which 177 different markers were tested: 364 of the 382 patients had deletions. Two data mining techniques, clustering and decision tree induction were used. Clustering was used to group patients according to the overall presence/absence of deletions at the tested markers. Decision trees and On-Line-Analytical-Processing (OLAP) were used to inspect the resulting clustering and look for correlations between deletion patterns, populations and the clinical picture of infertility. The results of the analysis indicate that there are correlations between deletion patterns and patient populations, as well as clinical phenotype severity.

INTRODUCTION

Deletions of the Y chromosome are an important cause of male infertility [1]. Testing for deletions is important for etiological diagnosis of male infertility, as well as for the prevention of iatrogenic transmission of mutations through assisted reproduction techniques to the offspring. Over 150 Y chromosome specific DNA markers have been used in different studies on patients from different regions to detect deletions. However, no mechanism of deletions, no correlation between deletion patterns and population origins and no correlation between deletion patterns and severity of clinical phenotype has been found yet. We have created and analyzed an extensive database of published Y chromosome deletions with the aim to investigate the possible correlation between deletion patterns, population origins and the severity of the clinical phenotype.

To analyse this database, we use data mining methods [2], which allow us to discover hidden patterns from facts recorded in a database. In particular, we use two types of data mining methods: clustering and decision tree induction. Clustering methods group similar objects together. Decision trees build concise classifiers that discriminate among given groups of objects.

Hierarchical clustering was used to group patients according to the overall presence/absence of deletions at the tested markers. To summarize the differences among the clusters, we build decision trees that discriminate among them. To obtain a more detailed picture of the characteristics of patients in each cluster, we use On-Line-Analytical-Processing (OLAP) [3] to inspect the resulting clustering.

OLAP allows multi-dimensional views of data, where aggregations are permitted along hierarchical dimensions. One dimension used in our case was the geographic one, where countries (and the corresponding patient populations) belong to continents and this forms a three-level hierarchy (world, continents, countries). Another hierarchical dimension was the clustering generated during the data mining phase. We can thus view properties we are interested in (such as the number of patients or percent of deletions) aggregated across the two dimensions: e.g., we can view the number of patients from each population that belong to a given cluster.

The clusters created in our analysis indicate correlations between deletion patterns and population origins, as well as the severity of the clinical phenotype. These correlations need to be confirmed by further clinical studies on new patients with deletions of the Y chromosome.

70	GY6	DYS135	112	273	154	208	OX7
72	DBY	DYS131/A	121	274	DYS7C	DAZ	160
AMEL-Y	494cen	DYF27/A	122	140	147	205-DAZ	
PRY	494-130K	210-STSP	123	141	240	254-DAZ	
75	UTY	100	SMCY	DYS7E	245	255-DAZ	
TSPY	Y6HP35pr	101	DYS107	142	203	624-DAZ	
78	88	102	124	143	242	276-DAZ	
TTY1	165	103	125	RBM1	148	277-DAZ	
TTY2	TB4Y	104	eIF-1AY	RBM2	220a	279	
79	89	105	126	GY48	262a	283-DAZ	
183	90	CDY	127	144	221a	SPGY	
81	182	XKRY	128	145	233a	243	
DYS140	151	106	129	DYS20	238	DYS12	
82	91	107	130	DYS7	146	248	
83	DYS139	135	55	DYS21	232a	236	
203tel	94	108	131	153	239	267	
86	BPY1	109	132	152	257	269	
85	95	110	133	150	249	BPY2	
84	DYS109	113	134	220	156	202	
DFFRY	161	114	164	Fr15II	224	247	
DF5'	97	116	136	232	231	157	
DF4,1	98	117	207	262	204	158	
DFJ/D	99-STSP	Y6PHc54pr	138	221	201	159	
DF3,1	DYS134	118	139	233	206	166	
87	DYS132	119	272	155	149	167	

Table 1. The names of the DNA markers used for testing deletions of the Y chromosome.

THE DATABASE

The data we used come from papers indexed by the MEDLINE bibliographic database (a list of which is available on request from the authors). From 34 published papers, we extracted data on 382 patients. Note that 364 of these had deletions. About 5% of infertile men carry deletions in the Y chromosome. As most studies have analyzed under 200 patients, usually not more than 10 patients with deletions were described in each article.

In the studies from which the patient data were collected, a total of 177 DNA markers were used. A typical study would test about 40 markers. The list of all 177 DNA markers is given in Table 1 (see also [4]). The data was stored in a MS Acess table with 177 columns and 382 rows. The value – for a given DNA marker indicated that the corresponding deletion was present, + that the deletion was not present, and blank that the marker was not tested for the patient in question. The average number of markers tested per patient was 43.8.

GROUPING PATIENTS WITH CLUSTERING

Cluster analysis [5] divides data points into groups of points that are "close" to each other. We used cluster analysis to group patients into clusters according to their deletion patterns. The clustering method we used was hierarchical agglomerative clustering. It starts with every data point being a cluster and repeatedly aggregates the most similar (least dissimilar) groups together until there is just one big group. The number of groups can be chosen subsequently.

A dissimilarity measure between patients was defined for the purpose of clustering. Assuming the outcomes of testing the 177 markers on patients P_1 and P_2 are ($p_{1,1}$, …, $p_{1,177}$) and ($p_{2,1}$, …, $p_{2,177}$) their dissimilarity $d_{1,2}$ is defined as the ratio between $D_{1,2}$, the number of markers where P_1 and P_2 have both been tested and have yielded the same outcome and N, the number of markers where both P_1 and P_2 have been tested. $d_{1,2}$ can take values between 0 and 1. If N=0, i.e. P_1 and P_2 have not been tested together on any marker, we set the dissimilarity to 1.

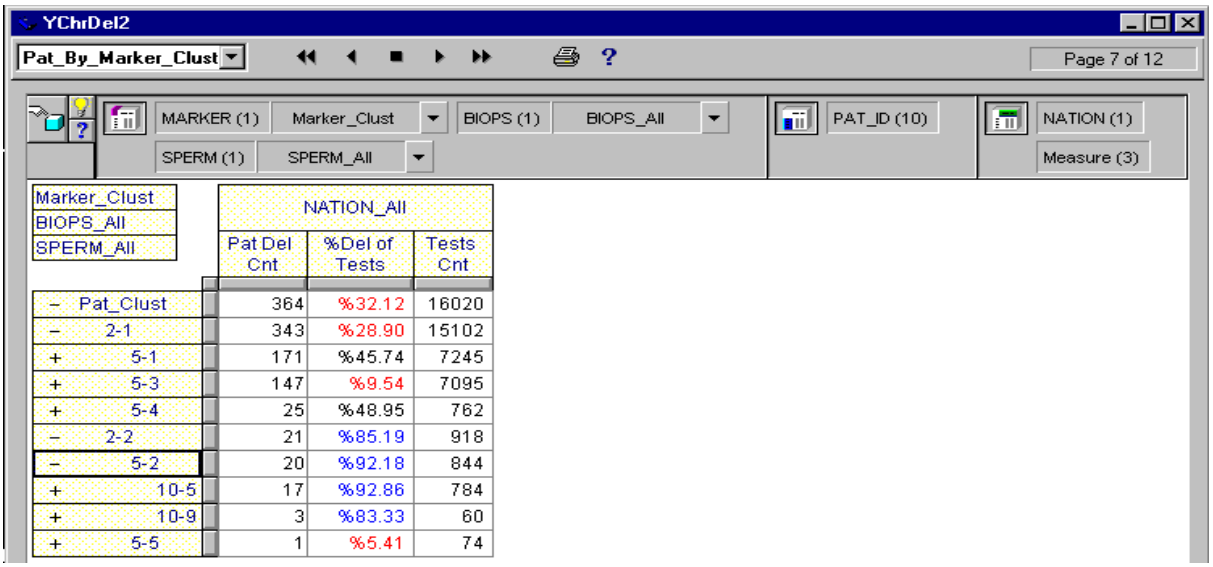


Figure 1. An OLAP view of the number of patients, the number of marker tests and the percentage of tests indicating deletions across patient clusters. The hierarchy of patient clusters is shown on the left.

Figure 2. An OLAP view of the number of patients across patient clusters and severity of clinical phenotype. The hierarchy of patient clusters is shown on the left. The first column gives information on patients of all types of severity, next four columns give information on patients of the respective severity of clinical phenotype.

After calculating the dissimilarity between each pair of patients, we performed hierarchical agglomerative clustering as described above. After clustering was complete, we decided on several numbers of clusters: 2, 5, 10, 20 and 30 clusters. These form a hierarchy: clusters 5-1, 5-3 and 5-4 are subclusters of cluster 2-1, while clusters 5-2 and 5-5 are subclusters of cluster 2-2. This hierarchy of clusters is used as a hierarchical dimension for summarizing cluster properties with OLAP.

Figure 1 displays an OLAP view, where the measures of interest (number of patients, number of marker tests, and percent of positive marker outcomes/deletions) are shown for (a part of) the hiererachy of patient clusters. Of the 364 patients, 343 belong to cluster 2-1 and 21 to 2-2. (Although we clustered all 382 patients, only the 364 patients with at least one deletion are shown in the Figures). The top level of the patient cluster hierarchy (all patients) is shown, as well as the next two levels (2 and 5 clusters). Cluster 5-2 is further broken down into clusters 10-5 and 10-9 at the next hierarchical level.

Overall, of the 16020 marker tests, some 32% have indicated deletions. In the cluster 2-1, this percentage is slightly lower (29%), while in cluster 2-2, it is much higher (85%). This percentage is still higher (92%) in the subcluster 5-2, which comprises 20 of the 21 patients in cluster 2-2.

Referring to Figure 2, where patients are broken down according to the severity of the clinical phenotype, we can see that all patients in cluster 5-2, for which the severity of the clinical phenotype is known, have azoospermia (AZO). Referring to Figure 3, where patients are broken down according to populations/geography, we can see that cluster 5-2 has 11 European, 6 American, 2 Asian, and 1 other patient. The percent of European patients here is 55%, as compared to the overall 45%.

Another interesting cluster is 5-3. This is a large cluster (147 of the 364 patients) with surprisingly few deletions (less than 10%). It has an unusually high relative number of patients with OAT (oligo-asteno-terato-azoospermia): while there are overall three times more patients with AZO as compared to OAT, in this cluster the numbers are comparable (41 and 47). This cluster has a higher percentage of Americans and lower percentage of Asians as compared to the entire dataset.

DESCRIBING THE PATIENT CLUSTERS WITH DECISION TREE INDUCTION

Classification trees, often called decision trees [5] predict the value of a discrete dependent variable with a finite set of values (called class) from the values of a set of independent variables (called attributes), which may be either continuous or discrete. Data represented in the form of a table, can be used to learn or automatically construct a decision tree. In the table, each row (example) has the form: $(x_1, x_2, ..., x_N, y)$, where x_i are values of the attributes (e.g., the outcomes of applying different DNA markers) and y is the value of the class (e.g., the severity of the clinical phenotype: azoospermia /*azo*/ or oligo-asteno-terato-azoospermia /*oat*/).

The induced (learned) decision tree has a test in each inner node that tests the value of a certain attribute, and in each leaf a value for the class. Given a new example for which the value of the class should be predicted, the tree is interpreted from the root. In each inner node the prescribed test is performed and according to the result of the test the corresponding left or right subtree is selected. When the selected node is a leaf then the value of the class for the the new example is predicted according to the class in the leaf.

In our case, we took the cluster to which a patient was assigned as the class. We considered two different problems: one for the second level of the clustering hierarcy (two clusters) and one for the third level (5 clusters). The attributes were the same in both cases: these were the 177 markers.

Below we give the tree for discriminating between the two clusters 2-1 and 2-2, rewritten in the form of if-then-else rules:

```
IF D84 = no THEN Cluster = 2-1
ELSE IF D84 = yes THEN
   IF D160 = no THEN Cluster = 2-1
   ELSE IF D160 = yes THEN
      IF D153 = yes THEN Cluster = 2-2
      ELSE IF D153 = no THEN Cl.=2-1
```

The tree states that patients of cluster 2-2 are characterized by the presence of deletions at each of the following three markers: DYS84, DYS160 and DYS153. These three markers are placed respectively at the proximal part (DYS84), the middle region (DYS153) and the distal part (DYS160) of the Y q-arm. The presence of deletions at all three markers simultaneously indicates the presence of large deletions, leading to severe phenotype of male infertility.

CONCLUSIONS

Using data from papers indexed by the MEDLINE bibliographic database, we have created a database containing information on patients with Y chromosome deletions, including results of testing markers on patients and the severity of the clinical phenotype. We have analysed this data with data mining techniques, in particular clustering and decision-tree induction, and inspected the data mining results using On-Line Analytical Processing (OLAP).

The obtained clusters indicate correlations between deletion patterns and population origins. This is most evident in the case of patients with relatively few deletions, where the Asian population was underrepresented as compared to Americans and Europeans. This could mean that the Y chromosome population background influences the mechanism(s) of deletions. This has to be confirmed by further studies, based on analysis of Y chromosome genetic polymorphisms in the respective populations.

The clusters also indicate strong correlation between deletion patterns and the severity of the clinical phenotype. Most notably, patients where a high proportion of the tested markers has shown deletions tend to have a more severe clinical phenotype. The clinical relevance of these findings need to be tested on new patients with deletions.

References

[1] Peterlin B et al. Sterility associated with Y chromosome abnormalities. In: C. Barratt, C.De Jonghe, D. Mortimer, j. Parinaud. Genetics of human male fertility. Sevres: EDK. 1997:66-75

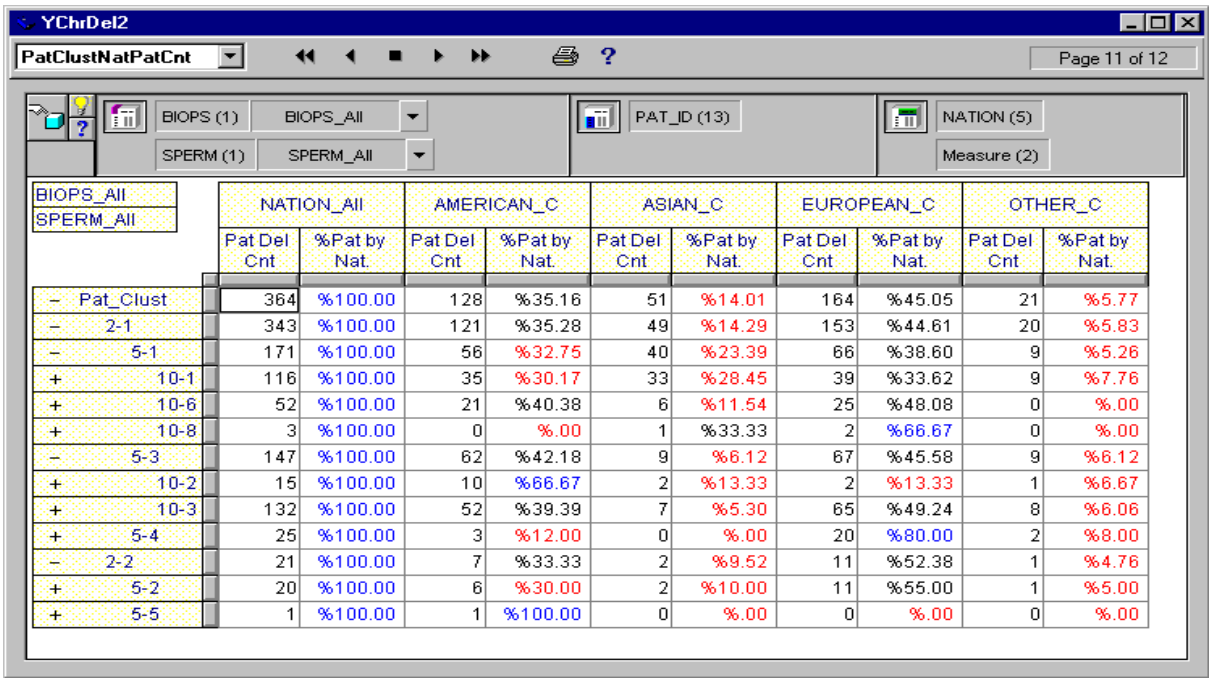
[2] Adriaans P, Zantinge D. (1996) Data Mining. Addison-Wesley, Reading.

[3] Codd EF, Codd SB, and Sally CT (1993) Beyond decision support. Computerworld 27(30).

[4] Dzeroski S, Hristovski D, Kunej T, Peterlin B (2000) A data mining approach to the development of a diagnostic test for male infertility. Submitted (MIE2000, Hanover).

[5] Everitt B. (1980). Cluster Analysis (second edition). Halsted, New York.

[6] Quinlan JR. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco.



BIOPS_All / SPERM_All	NATION_All		AMERICAN_C		ASIAN_C		EUROPEAN_C		OTHER_C	
	Pat Del Cnt	%Pat by Nat.	Pat Del Cnt	%Pat by Nat.	Pat Del Cnt	%Pat by Nat.	Pat Del Cnt	%Pat by Nat.	Pat Del Cnt	%Pat by Nat.
– Pat_Clust	364	%100.00	128	%35.16	51	%14.01	164	%45.05	21	%5.77
– 2-1	343	%100.00	121	%35.28	49	%14.29	153	%44.61	20	%5.83
– 5-1	171	%100.00	56	%32.75	40	%23.39	66	%38.60	9	%5.26
+ 10-1	116	%100.00	35	%30.17	33	%28.45	39	%33.62	9	%7.76
+ 10-6	52	%100.00	21	%40.38	6	%11.54	25	%48.08	0	%.00
+ 10-8	3	%100.00	0	%.00	1	%33.33	2	%66.67	0	%.00
– 5-3	147	%100.00	62	%42.18	9	%6.12	67	%45.58	9	%6.12
+ 10-2	15	%100.00	10	%66.67	2	%13.33	2	%13.33	1	%6.67
+ 10-3	132	%100.00	52	%39.39	7	%5.30	65	%49.24	8	%6.06
+ 5-4	25	%100.00	3	%12.00	0	%.00	20	%80.00	2	%8.00
– 2-2	21	%100.00	7	%33.33	2	%9.52	11	%52.38	1	%4.76
+ 5-2	20	%100.00	6	%30.00	2	%10.00	11	%55.00	1	%5.00
+ 5-5	1	%100.00	1	%100.00	0	%.00	0	%.00	0	%.00

Figure 3. An OLAP view of the number of patients across patient clusters and population origins. The hierarchy of patient clusters is shown on the left. The first column gives information on patients of all nations, next four columns give information on patients of the respective population groups.