Disclosed: An efficient depth-first, top-down algorithm for mining disjunctive closed itemsets in high-dimensional data
Supplementary Material - Summary of data sets

Renato Vimieiro and Pablo Moscato

Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine

The University of Newcastle

Hunter Medical Research Institute

Lot 1, Kookaburra Circuit, New Lambton Heights, NSW, 2305, Australia.

Email: {renato.vimieiro, pablo.moscato}@newcastle.edu.au

Characteristics of the data sets of the experiments to assess the performance of Disclosed

Table A-1: Description of the sources of the data sets used in the experiments for assessing the performance of Disclosed. Many original files were converted to WEKA's ARFF format by The Bioinformatics Group of Seville–Spain (`http://www.upo.es/eps/bigs/`). Content of original files and ARFF versions were matched before conducting experiments.

Name	Source	URL	Format
ALL–AML	Brunet et al. (2004)	`http://www.pnas.org/content/101/12/4164/suppl/DC1` `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`	TSV
Colon	Alon et al. (1999)	`http://genomics-pubs.princeton.edu/oncology/affydata/index.html` `http://www.upo.es/eps/bigs/dataSet/colon.arff`	TSV ARFF
Embryo (data set 'C')	Pomeroy et al. (2002)	`http://www.broadinstitute.org/mpr/publications/projects/CNS/` `http://www.upo.es/eps/bigs/dataSet/dataset_C.arff`	RES ARFF
GDS963	Strunnikova et al. (2005)	`http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS963`	SOFT
GDS2200	Nindl et al. (2006)	`http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2200`	SOFT
GDS2250	Richardson et al. (2006)	`http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2250`	SOFT
GDS2519	Scherzer et al. (2007)	`http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2519`	SOFT
GDS2545	Chandran et al. (2007)	`http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2545`	SOFT
GDS2821	Lesnick et al. (2007)	`http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2821`	SOFT
GDS2941	Lockstone et al. (2007)	`http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2941`	SOFT
Leukemia	Golub et al. (1999)	`http://www.broadinstitute.org/mpr/publications/projects/Leukemia/data_set_ALL_AML_train.tsv` `http://www.upo.es/eps/bigs/dataSet/leukemia_train_38x7129.arff`	TSV ARFF
Lymphoma	Alizadeh et al. (2000)	`http://llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt` `http://www.upo.es/eps/bigs/dataSet/Lymphoma96x4026+9classes.arff`	CDT ARFF
Promoters	Frank and Asuncion (2010)	`http://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Promoter+Gene+Sequences)`	CSV

Table A-2: Characteristics of the ALL–AML data sets considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $RelativeDensity = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $RelativeGlobalDensity = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
38–34	4812	38	0	0	0.0	0.0	0.0
33	4812	38	1	33	1.0	33.0	0.00018
32	4812	38	3	38	0.851	32.3	0.00053
31	4812	38	4	38	0.842	32.0	0.0007
30	4812	38	7	38	0.82	31.1	0.00119
29	4812	38	23	38	0.78	29.7	0.00373
28	4812	38	52	38	0.756	28.7	0.00817
27	4812	38	126	38	0.729	27.7	0.0191
26	4812	38	1010	38	0.69	26.2	0.145
25	4812	38	1268	38	0.683	26.0	0.18
24	4812	38	1451	38	0.677	25.7	0.204
23	4812	38	1578	38	0.671	25.5	0.22
22	4812	38	1703	38	0.664	25.2	0.235
21	4812	38	1809	38	0.658	25.0	0.247
20	4812	38	1914	38	0.651	24.7	0.259
19	4812	38	2031	38	0.642	24.4	0.271
18	4812	38	2126	38	0.634	24.1	0.28
17	4812	38	2223	38	0.626	23.8	0.289
16	4812	38	2298	38	0.619	23.5	0.296
15	4812	38	2384	38	0.611	23.2	0.303
14	4812	38	2450	38	0.605	23.0	0.308
13	4812	38	2524	38	0.597	22.7	0.313
12	4812	38	2627	38	0.586	22.3	0.32
11	4812	38	2805	38	0.567	21.6	0.331
10	4812	38	4373	38	0.458	17.4	0.416
9	4812	38	4612	38	0.447	17.0	0.428
8	4812	38	4694	38	0.443	16.8	0.432
7	4812	38	4750	38	0.44	16.7	0.434
6–4	4812	38	4777	38	0.438	16.6	0.435
3–1	4812	38	4809	38	0.436	16.6	0.436

Table A-3: Characteristics of the Colon data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $RelativeDensity = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $RelativeGlobalDensity = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
62–46	2000	62	0	0	0.0	0.0	0.0
45	2000	62	1	45	1.0	45.0	0.000363
44	2000	62	2	53	0.84	44.5	0.000718
43	2000	62	6	61	0.713	43.5	0.0021
42	2000	62	15	62	0.687	42.6	0.00515
41	2000	62	34	62	0.673	41.7	0.0114
40	2000	62	434	62	0.647	40.1	0.14
39	2000	62	507	62	0.645	40.0	0.163
38	2000	62	535	62	0.643	39.9	0.172
37	2000	62	566	62	0.64	39.7	0.181
36	2000	62	589	62	0.638	39.6	0.188
35	2000	62	615	62	0.635	39.4	0.195
34	2000	62	639	62	0.632	39.2	0.202
33	2000	62	666	62	0.628	38.9	0.209
32	2000	62	693	62	0.623	38.7	0.216
31	2000	62	716	62	0.619	38.4	0.222
30	2000	62	736	62	0.616	38.2	0.227
29	2000	62	762	62	0.611	37.9	0.233
28	2000	62	792	62	0.605	37.5	0.239
27	2000	62	812	62	0.601	37.2	0.244
26	2000	62	837	62	0.595	36.9	0.249
25	2000	62	856	62	0.591	36.6	0.253
24	2000	62	879	62	0.586	36.3	0.257
23	2000	62	896	62	0.581	36.0	0.26
22	2000	62	1505	62	0.49	30.4	0.369
21	2000	62	1623	62	0.479	29.7	0.389
20	2000	62	1787	62	0.464	28.8	0.415
19	2000	62	1869	62	0.457	28.4	0.428
18	2000	62	1940	62	0.451	28.0	0.438
17	2000	62	1966	62	0.449	27.8	0.441
16	2000	62	1981	62	0.448	27.8	0.443
15–1	2000	62	1992	62	0.446	27.7	0.445

Table A-4: Characteristics of the Embryo data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $RelativeDensity = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $RelativeGlobalDensity = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
59–46	7129	59	0	0	0.0	0.0	0.0
45	7129	59	4	53	0.849	45.0	0.000428
44	7129	59	15	58	0.763	44.3	0.00158
43	7129	59	52	59	0.735	43.4	0.00536
42	7129	59	108	59	0.723	42.7	0.011
41	7129	59	258	59	0.707	41.7	0.0256
40	7129	59	469	59	0.694	40.9	0.0456
39	7129	59	3118	59	0.666	39.3	0.291
38	7129	59	3212	59	0.665	39.3	0.3
37	7129	59	3303	59	0.664	39.2	0.308
36	7129	59	3394	59	0.663	39.1	0.316
35	7129	59	3491	59	0.661	39.0	0.324
34	7129	59	3613	59	0.658	38.8	0.333
33	7129	59	3723	59	0.655	38.7	0.342
32	7129	59	3860	59	0.651	38.4	0.353
31	7129	59	3993	59	0.647	38.2	0.362
30	7129	59	4142	59	0.642	37.9	0.373
29	7129	59	4313	59	0.636	37.5	0.385
28	7129	59	4494	59	0.629	37.1	0.397
27	7129	59	4656	59	0.623	36.8	0.407
26	7129	59	4804	59	0.618	36.5	0.416
25	7129	59	5032	59	0.609	35.9	0.43
24	7129	59	5249	59	0.601	35.4	0.442
23	7129	59	5505	59	0.591	34.9	0.456
22	7129	59	5801	59	0.58	34.2	0.472
21	7129	59	6035	59	0.571	33.7	0.483
20	7129	59	7061	59	0.537	31.7	0.532
19	7129	59	7105	59	0.536	31.6	0.534
18–1	7129	59	7120	59	0.536	31.6	0.535

Table A-5: Characteristics of the GDS963 data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $RelativeDensity = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $RelativeGlobalDensity = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
36	12557	36	1	36	1.0	36.0	7.96e-05
35	12557	36	2	36	0.986	35.5	0.000157
34	12557	36	3	36	0.972	35.0	0.000232
33	12557	36	5	36	0.95	34.2	0.000378
32	12557	36	19	36	0.905	32.6	0.00137
31	12557	36	31	36	0.888	32.0	0.00219
30	12557	36	47	36	0.869	31.3	0.00325
29	12557	36	84	36	0.841	30.3	0.00563
28	12557	36	157	36	0.812	29.2	0.0101
27	12557	36	249	36	0.789	28.4	0.0156
26	12557	36	355	36	0.769	27.7	0.0217
25	12557	36	530	36	0.744	26.8	0.0314
24	12557	36	748	36	0.722	26.0	0.043
23	12557	36	1079	36	0.696	25.1	0.0598
22	12557	36	1652	36	0.667	24.0	0.0877
21	12557	36	2537	36	0.638	23.0	0.129
20	12557	36	3988	36	0.608	21.9	0.193
19	12557	36	6655	36	0.576	20.7	0.305
18–1	12557	36	12557	36	0.54	19.4	0.54

Table A-6: Characteristics of the GDS2200 data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $RelativeDensity = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $RelativeGlobalDensity = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
15–12	22215	15	0	0	0.0	0.0	0.0
11	22215	15	44	11	1.0	11.0	0.00145
10	22215	15	238	15	0.679	10.2	0.00727
9	22215	15	621	15	0.63	9.45	0.0176
8	22215	15	1278	15	0.58	8.71	0.0334
7	22215	15	2405	15	0.527	7.91	0.0571
6	22215	15	12391	15	0.425	6.37	0.237
5	22215	15	15431	15	0.407	6.1	0.282
4	22215	15	19189	15	0.379	5.69	0.328
3	22215	15	21523	15	0.36	5.4	0.349
2–1	22215	15	22109	15	0.354	5.31	0.352

Table A-7: Characteristics of the GDS2250 data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $RelativeDensity = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $RelativeGlobalDensity = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
47–34	54613	47	0	0	0.0	0.0	0.0
33	54613	47	3	38	0.868	33.0	3.86e-05
32	54613	47	10	38	0.85	32.3	0.000126
31	54613	47	21	38	0.832	31.6	0.000259
30	54613	47	32	38	0.817	31.1	0.000387
29	54613	47	47	38	0.8	30.4	0.000557
28	54613	47	85	38	0.772	29.3	0.000971
27	54613	47	133	38	0.75	28.5	0.00148
26	54613	47	193	38	0.729	27.7	0.00208
25	54613	47	289	38	0.706	26.8	0.00302
24	54613	47	430	38	0.681	25.9	0.00434
23	54613	47	662	45	0.553	24.9	0.00642
22	54613	47	1120	47	0.504	23.7	0.0103
21	54613	47	2293	47	0.475	22.3	0.0199
20	54613	47	2678	47	0.468	22.0	0.0229
19	54613	47	4350	47	0.443	20.8	0.0353
18	54613	47	29890	47	0.392	18.4	0.214
17	54613	47	30156	47	0.392	18.4	0.216
16	54613	47	30329	47	0.391	18.4	0.217
15	54613	47	30500	47	0.391	18.4	0.218
14	54613	47	30700	47	0.39	18.3	0.219
13	54613	47	30988	47	0.389	18.3	0.221
12	54613	47	31410	47	0.387	18.2	0.223
11	54613	47	32091	47	0.384	18.1	0.226
10	54613	47	33242	47	0.378	17.8	0.23
9	54613	47	35546	47	0.366	17.2	0.238
8	54613	47	41834	47	0.337	15.8	0.258
7	54613	47	48613	47	0.31	14.6	0.276
6	54613	47	51189	47	0.301	14.2	0.282
5	54613	47	52280	47	0.297	14.0	0.284
4	54613	47	52735	47	0.295	13.9	0.285
3	54613	47	53346	47	0.293	13.8	0.286
2–1	54613	47	54599	47	0.287	13.5	0.287

Table A-8: Characteristics of the GDS2519 data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $\mathit{RelativeDensity} = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $\mathit{RelativeGlobalDensity} = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
105–52	22215	105	0	0	0.0	0.0	0.0
51	22215	105	14	59	0.864	51.0	0.000306
50	22215	105	2834	79	0.633	50.0	0.0608
49	22215	105	4046	81	0.614	49.7	0.0862
48	22215	105	4573	81	0.611	49.5	0.0971
47	22215	105	4830	81	0.61	49.4	0.102
46	22215	105	4962	81	0.608	49.3	0.105
45	22215	105	5045	81	0.608	49.2	0.106
44	22215	105	5094	81	0.607	49.2	0.107
43–35	22215	105	5114	81	0.607	49.1	0.108
34	22215	105	5188	92	0.532	48.9	0.109
33	22215	105	11772	99	0.404	40.0	0.202
32	22215	105	14210	99	0.39	38.6	0.235
31	22215	105	15061	99	0.386	38.2	0.247
30	22215	105	15396	99	0.384	38.0	0.251
29	22215	105	15504	99	0.384	38.0	0.252
28–23	22215	105	15542	99	0.383	38.0	0.253
22	22215	105	20775	105	0.323	33.9	0.302
21	22215	105	21958	105	0.317	33.2	0.313
20–1	22215	105	22170	105	0.315	33.1	0.315

Table A-9: Characteristics of the GDS2545 data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $\mathit{RelativeDensity} = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $\mathit{RelativeGlobalDensity} = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
171–69	12558	171	0	0	0.0	0.0	0.0
68	12558	171	1	68	1.0	68.0	3.17e-05
67	12558	171	2	70	0.964	67.5	6.29e-05
66	12558	171	19	85	0.778	66.2	0.000585
65	12558	171	2216	92	0.707	65.0	0.0671
64	12558	171	2983	139	0.466	64.8	0.0899
63	12558	171	4706	139	0.461	64.1	0.14
62	12558	171	5181	139	0.46	63.9	0.154
61	12558	171	5340	139	0.459	63.8	0.159
60	12558	171	5391	139	0.459	63.8	0.16
59–27	12558	171	5411	139	0.459	63.8	0.161
26	12558	171	5461	164	0.387	63.5	0.162
25	12558	171	9536	164	0.287	47.1	0.209
24	12558	171	10225	164	0.277	45.5	0.217
23	12558	171	10431	164	0.275	45.1	0.219
22–19	12558	171	10508	164	0.274	44.9	0.22
18	12558	171	12274	171	0.24	41.0	0.235
17–1	12558	171	12495	171	0.237	40.6	0.236

Table A-10: Characteristics of the GDS2821 data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $\mathit{RelativeDensity} = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $\mathit{RelativeGlobalDensity} = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
25	54277	25	0	0	0.0	0.0	0.0
24	54277	25	11	25	0.96	24.0	0.000195
23	54277	25	41	25	0.931	23.3	0.000703
22	54277	25	131	25	0.896	22.4	0.00216
21	54277	25	402	25	0.858	21.5	0.00636
20	54277	25	1157	25	0.82	20.5	0.0175
19	54277	25	2938	25	0.784	19.6	0.0424
18	54277	25	6906	25	0.747	18.7	0.0951
17	54277	25	15797	25	0.709	17.7	0.206
16	54277	25	35434	25	0.671	16.8	0.438
15	54277	25	36000	25	0.67	16.7	0.444
14	54277	25	36806	25	0.667	16.7	0.453
13	54277	25	38127	25	0.662	16.6	0.465
12	54277	25	40103	25	0.653	16.3	0.483
11	54277	25	43022	25	0.639	16.0	0.506
10	54277	25	47423	25	0.617	15.4	0.539
9–1	54277	25	54277	25	0.584	14.6	0.584

Table A-11: Characteristics of the GDS2941 data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $\mathit{RelativeDensity} = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $\mathit{RelativeGlobalDensity} = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
15	22215	15	649	15	1.0	15.0	0.0292
14	22215	15	1884	15	0.956	14.3	0.0811
13	22215	15	3620	15	0.913	13.7	0.149
12	22215	15	5909	15	0.869	13.0	0.231
11	22215	15	8671	15	0.826	12.4	0.322
10	22215	15	11630	15	0.786	11.8	0.411
9	22215	15	14935	15	0.744	11.2	0.5
8	22215	15	19722	15	0.693	10.4	0.615
7–1	22215	15	22215	15	0.668	10.0	0.668

Table A-12: Characteristics of the Leukemia data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $\mathit{RelativeDensity} = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $\mathit{RelativeGlobalDensity} = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
37	7129	36	0	0	0.0	0.0	0.0
36	7129	36	2	36	1.0	36.0	0.000281
35	7129	36	3	36	0.991	35.7	0.000417
34	7129	36	10	36	0.958	34.5	0.00134
33	7129	36	18	36	0.94	33.8	0.00237
32	7129	36	36	36	0.914	32.9	0.00462
31	7129	36	68	36	0.889	32.0	0.00848
30	7129	36	134	36	0.862	31.0	0.0162
29	7129	36	262	36	0.834	30.0	0.0307
28	7129	36	492	36	0.808	29.1	0.0558
27	7129	36	970	36	0.779	28.1	0.106
26	7129	36	3582	36	0.738	26.6	0.371
25	7129	36	3723	36	0.736	26.5	0.384
24	7129	36	3831	36	0.734	26.4	0.394
23	7129	36	3937	36	0.732	26.3	0.404
22	7129	36	4055	36	0.728	26.2	0.414
21	7129	36	4179	36	0.724	26.1	0.424
20	7129	36	4310	36	0.719	25.9	0.434
19	7129	36	4416	36	0.714	25.7	0.442
18	7129	36	4532	36	0.709	25.5	0.45
17	7129	36	4682	36	0.701	25.2	0.46
16	7129	36	4826	36	0.693	25.0	0.469
15	7129	36	5005	36	0.683	24.6	0.48
14	7129	36	5188	36	0.673	24.2	0.49
13	7129	36	5414	36	0.66	23.8	0.501
12	7129	36	5654	36	0.646	23.3	0.512
11	7129	36	5878	36	0.633	22.8	0.522
10	7129	36	7110	36	0.572	20.6	0.57
9–1	7129	36	7124	36	0.571	20.6	0.571

Table A-13: Characteristics of the Lymphoma data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $RelativeDensity = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $RelativeGlobalDensity = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
46–23	4026	46	0	0	0.0	0.0	0.0
22	4026	46	2	25	0.88	22.0	0.000238
21	4026	46	8	34	0.625	21.2	0.000918
20	4026	46	15	37	0.559	20.7	0.00167
19	4026	46	126	45	0.427	19.2	0.0131
18	4026	46	904	46	0.395	18.2	0.0887
17	4026	46	1088	46	0.391	18.0	0.106
16	4026	46	1222	46	0.386	17.8	0.117
15	4026	46	1324	46	0.381	17.5	0.125
14	4026	46	1417	46	0.376	17.3	0.132
13	4026	46	1495	46	0.371	17.1	0.138
12	4026	46	1579	46	0.366	16.8	0.143
11	4026	46	1671	46	0.359	16.5	0.149
10	4026	46	1759	46	0.351	16.2	0.154
9	4026	46	1858	46	0.343	15.8	0.158
8	4026	46	1972	46	0.333	15.3	0.163
7	4026	46	2206	46	0.314	14.5	0.172
6	4026	46	2358	46	0.302	13.9	0.177
5	4026	46	2500	46	0.291	13.4	0.181
4	4026	46	2657	46	0.279	12.8	0.184
3	4026	46	2974	46	0.256	11.8	0.189
2	4026	46	3496	46	0.225	10.3	0.195
1	4026	46	3867	46	0.205	9.44	0.197

Table A-14: Characteristics of the Promoters data set considering the variations of minimum support thresholds. *Number of Features (Samples)* refers to the total number of features (samples) in the original data set. *Relative Number of Features* is the number of features that surpass the minimum support threshold, the *Relative Number of Samples* is the number of samples associated with the features with support greater than the minimum. *Relative Density* is the density in the data set formed with features surpassing minimum support and their associated samples — $RelativeDensity = R'/S' \times F'$, where (S', F', R') is the data set induced by the minimum support. *Relative Global Density* is the density of the induced data set with respect to the original data set — $RelativeGlobalDensity = R'/S \times F$.

Minimum Support	Number of Features	Number of Samples	Relative Number of Features	Relative Number of Samples	Relative Density	Relative Avg. Feature Length	Relative Global Density
106–55	228	106	0	0	0.0	0.0	0.0
54	228	106	2	70	0.771	54.0	0.00447
53–46	228	106	3	81	0.663	53.7	0.00666
45	228	106	4	90	0.572	51.5	0.00852
44	228	106	5	94	0.532	50.0	0.0103
43	228	106	6	96	0.509	48.8	0.0121
42	228	106	9	102	0.456	46.6	0.0173
41	228	106	11	105	0.434	45.5	0.0207
40	228	106	12	105	0.429	45.1	0.0224
39	228	106	13	105	0.425	44.6	0.024
38	228	106	17	105	0.41	43.1	0.0303
37	228	106	18	105	0.407	42.7	0.0318
36	228	106	22	105	0.395	41.5	0.0378
35	228	106	27	105	0.384	40.3	0.045
34	228	106	36	106	0.365	38.7	0.0577
33	228	106	40	106	0.36	38.1	0.0631
32	228	106	49	106	0.349	37.0	0.0751
31	228	106	58	106	0.34	36.1	0.0866
30	228	106	65	106	0.334	35.4	0.0953
29	228	106	78	106	0.324	34.4	0.111
28	228	106	89	106	0.317	33.6	0.124
27	228	106	108	106	0.306	32.4	0.145
26	228	106	120	106	0.3	31.8	0.158
25	228	106	129	106	0.295	31.3	0.167
24	228	106	149	106	0.286	30.3	0.187
23	228	106	155	106	0.283	30.0	0.193
22	228	106	174	106	0.275	29.2	0.21
21	228	106	185	106	0.271	28.7	0.22
20	228	106	194	106	0.267	28.3	0.227
19	228	106	201	106	0.264	28.0	0.232
18	228	106	206	106	0.261	27.7	0.236
17	228	106	215	106	0.257	27.3	0.243
16	228	106	216	106	0.257	27.2	0.243
15	228	106	220	106	0.255	27.0	0.246
14	228	106	224	106	0.252	26.8	0.248
13	228	106	226	106	0.251	26.6	0.249
12–1	228	106	227	106	0.251	26.6	0.25

References

Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt (2000, Feb). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature 403*(6769), 503–511.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences 96*(12), 6745–6750.

Brunet, J.-P., P. Tamayo, T. R. Golub, and J. P. Mesirov (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences 101*(12), 4164–4169.

Chandran, U., C. Ma, R. Dhir, M. Bisceglia, M. Lyons-Weiler, W. Liang, G. Michalopoulos, M. Becich, and F. Monzon (2007). Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer 7*(1), 64.

Frank, A. and A. Asuncion (2010). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999, Oct). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*(5439), 531–537.

Lesnick, T. G., S. Papapetropoulos, D. C. Mash, J. Ffrench-Mullen, L. Shehadeh, M. de Andrade, J. R. Henley, W. A. Rocca, J. E. Ahlskog, and D. M. Maraganore (2007, 06). A genomic pathway approach to a complex disease: Axon guidance and parkinson disease. *PLoS Genetics 3*(6), e98.

Lockstone, H. E., L. W. Harris, J. E. Swatton, M. T. Wayland, A. J. Holland, and S. Bahn (2007). Gene expression profiling in the adult down syndrome brain. *Genomics 90*(6), 647–660.

Nindl, I., C. Dang, T. Forschner, R. J. Kuban, T. Meyer, W. Sterry, and E. Stockfleth (2006). Identification of differentially expressed genes in cutaneous squamous cell carcinoma by microarray expression profiling. *Molecular Cancer 5*, 30.

Pomeroy, S. L., P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub (2002, Jan). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature 415*(6870), 436–442.

Richardson, A. L., Z. C. Wang, A. De Nicolo, X. Lu, M. Brown, A. Miron, X. Liao, J. D. Iglehart, D. M. Livingston, and S. Ganesan (2006, Feb). X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell 9*(2), 121–32.

Scherzer, C. R., A. C. Eklund, L. J. Morse, Z. Liao, J. J. Locascio, D. Fefer, M. A. Schwarzschild, M. G. Schlossmacher, M. A. Hauser, J. M. Vance, L. R. Sudarsky, D. G. Standaert, J. H. Growdon, R. V. Jensen, and S. R. Gullans (2007). Molecular markers of early Parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences 104*(3), 955–960.

Strunnikova, N., S. Hilmer, J. Flippin, M. Robinson, E. Hoffman, and K. G. Csaky (2005). Differences in gene expression profiles in dermal fibroblasts from control and patients with age-related macular degeneration elicited by oxidative injury. *Free Radical Biology and Medicine 39*(6), 781–796.