



*Comparative Data Mining for
Microarrays*

A Case Study Based on Multiple Myeloma

Michael Waddell and David Page

University of Wisconsin, USA

Fenghuang Zhan, Bart Barlogie and John Shaughnessy, Jr.

University of Arkansas for Medical Sciences, USA

Johanna Hardin

Fred Hutchinson Cancer Research Center, USA

James Cussens

University of York, UK





Motivation

- Supervised machine learning and data mining tools have the potential to:
 - Uncover new therapeutic targets.
 - Predict how patients will respond to specific treatments.
 - Uncover regulatory relationships among genes in normal and disease situations.
- Research is needed to guide methodological decisions when using these algorithms on microarray data.



Research Questions

- Which supervised data mining algorithms are **most appropriate** for microarray data?
 - Trees
 - Boosted Trees
 - Support Vector Machines (SVMs)
 - Unweighted Ensembles of Voters (EOVs)
 - Bayesian Networks (BNs) and Naïve BNs
- Should we **omit** certain genes based on low or negative expression levels (e.g., in AffymetrixTM data)?



Data Set

- Gene expression **microarray data** on 105 highly purified plasma cell samples
 - 74 newly diagnosed **Multiple Myeloma** patients
 - 31 **normal**, healthy donors
- **Multiple Myeloma** is an incurable malignancy of immunoglobulin secreting plasma cells that grow and expand in the bone marrow.
- *The dataset is publically available at <http://lambertlab.uams.edu/publicdata.htm>*

Input file form used

	Accession Number				class			
	A28202		AB00014		AB00015		...	
Person 1	P	1142.0	A	321.0	P	2567.2	...	myeloma
Person 2	A	-586.3	P	586.1	P	759.0	...	normal
Person 3	A	105.2	A	559.3	P	3210.7	...	myeloma
Person 4	P	-42.8	A	692.1	P	812.0	...	normal
...

Figure 1: Illustration of the input file form for data mining runs. For each Accession Number, there is an Absolute Call of either Absent (A) or Present (P) and an Average Distance value. AD compares hybridization with 25-mers that are known to appear in a gene against hybridization with the same 25-mers except that the middle (13th) base has been changed to its complement.



Lessons Learned

- A directly comprehensible model exposes “trivially-accurate genes.”
- EOV and BN learning are quite accurate and provide **more insight** than other methods.
- **Consistency** is a more important predictor than magnitude.
- **Throwing out** low or negative values from AffymetrixTM data is a **mistake**; as is using only Absolute Call (AC) values.

"Trivially-Accurate" Genes

- Genes that provide **no new insight** but are highly-accurate due to the nature of the disease or of sample collection.
- Plasma cells (normally **polyclonal**) in a patient with **Multiple Myeloma** are **monoclonal**.
- This lack of variability is a **very good indicator** of the disease, but provides **no new insight**.
- If the goal is to obtain insight, then trivially-accurate genes should be **removed** through consultation with a **domain expert**.



Insight of EOV and BN

- **Voting** and **Bayes nets** provide greater insight than other approaches.
- In both cases, we see **directly** which genes are important.
- Boosted Trees and SVMs **cannot** provide this type of direct insight.
- Single decision trees are also insightful, but provide **much less information**.



Insight of EOV and BN

- Bayes Nets (non-naïve) expose **coorelations** among genes, but **discard** some important genes.
- Naïve Bayes nets and voting **don't discard** any important genes, but **ignore** coorelations among genes.
- The **type** of insight needed will determine whether voting or Bayes nets is best for a particular application.



Consistency

- The traditional "fold-change" measure for comparing gene expression looks for large differences between some samples.
- Information gain looks for a high level of consistency in differential expression instead.
- Thus, considering consistency alone is far better than considering magnitude alone.
- Further studies will determine whether or not SVMs have an advantage over other methods by being able to consider both.



Throwing Out Data

- A **negative AD** means that the "mismatch" 25-mers tend to hybridize more than the "perfect-match" ones.
- Also, some believe that AC data is **less noisy** and less prone to **over-fitting** than AD data.
- Therefore, some researchers **ignore negative values** of AD or use **only** the AC data.
- In this study, 4 of our top 8 genes have **negative** split-points, and **all** of our top 70 genes are AD values.

Summary of Accuracies

Method	AC Only	AC+AD
Trees	90.5	98.1
Boosted Trees	97.1	99.0
SVM	100.0	100.0
EOV	93.3	100.0
Naïve BN	98.1	100.0
BN	95.2	100.0

Figure 2: Summary of accuracies, by leave-one-out cross-validation, of the 6 data mining techniques. The column labeled “AC Only” gives performance when using only the AffymetrixTM Absolute (Absent-Present) Call. “AC+AD” uses both Absolute Call and Average Difference. Note: The BN for “AC+AD” was a Naïve Bayes net containing 30 genes. This is similar, but not identical to the Naïve BN for “AC+AD” which contained 70 genes.

The 8 Best “AD” Genes

Score	Gene	Accession Number	Split	MH	ML	NH	NL
0.80	<i>APOA2</i>	X04898	-777	74	0	1	30
0.74	<i>HERV K22 pol</i>	K03498	637	3	71	31	0
0.70	<i>TERT</i>	AF015950	-1610	70	4	0	31
0.70	<i>UMOD</i>	M15881	1119.1	0	74	28	3
0.70	<i>CDH4</i>	L34059	-278	74	0	3	28
0.66	<i>ACTR1A</i>	Z14978	3400.6	3	71	30	1
0.66	<i>MASP1</i>	D17525	-536.6	71	3	1	30
0.65	<i>PTPN21</i>	X79510	1256.1	6	68	31	0

Figure 3: The 8 genes with the top information gain scores according to Absolute Call(AC) or Average Difference(AD) are all AD features. “Score” is the information gain score. “Split” is the AD value at which the split is made. MH is the number of samples of class “Myeloma” (M) and an AD above the split value. ML is the number of samples of class M and an AD below the split value. NH and NL are analogous for samples of class “Normal.”

The 8 Best “AC” Genes

Score	Gene	Accession Number	MH	ML	NH	NL
0.45	<i>H1F2</i>	X57129	57	17	0	31
0.44	<i>NCBP2</i>	D59253	57	17	0	31
0.43	<i>SM15</i>	U73167	56	18	0	31
0.43	<i>GCN5L2</i>	U57316	56	18	0	31
0.41	<i>MHC2 beta W52</i>	HT3779	12	62	29	2
0.41	<i>RNASE6</i>	U64998	15	59	30	1
0.41	<i>TNFRSF7</i>	M63928	15	59	30	1
0.41	<i>SDF1</i>	L36033	15	59	30	1

Figure 4: The 8 genes with the top information gain scores using only the Absolute Calls (AC). “Score” is the information gain score. MH is the number of samples of class “Myeloma” (M) and an AD above the split value. ML is the number of samples of class M and an AD below the split value. NH and NL are analogous for samples of class “Normal.”

Decision Trees

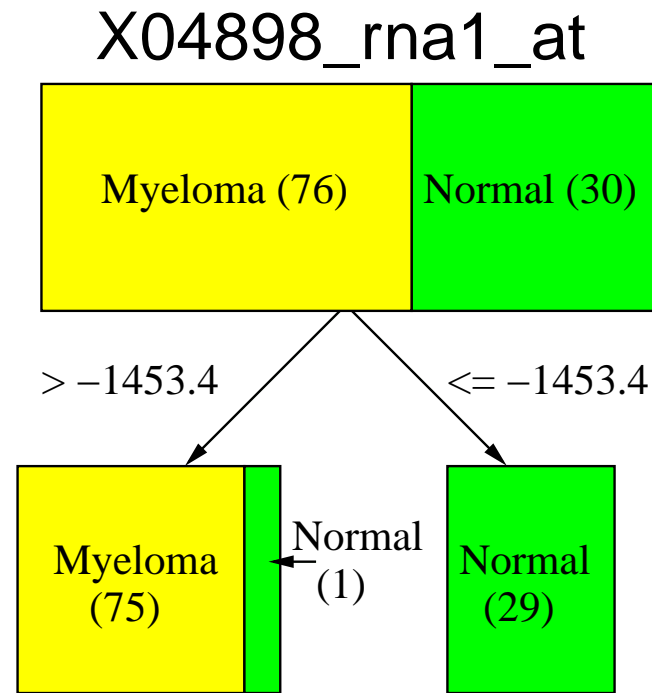


Figure 5: Example decision tree for “AD+AC” data. Single decision trees are directly interpretable, but provide very limited information because of the emphasis that they place on single “good-predictors.” Because this decision tree has only one “level,” it is also referred to as a “decision stump.” See Figure 6 for examples of multi-level decision trees.

Boosted Decision Trees

Decision tree 0:

AC_U78525_at = M: myeloma (0)
AC_U78525_at = A: normal (21/1)
AC_U78525_at = P:
...AC_M62505_at = M: myeloma (0)
AC_M62505_at = P: normal (4)
AC_M62505_at = A:
...AC_AF002700_at = P: myeloma (0)
AC_AF002700_at = M: normal (2)
AC_AF002700_at = A:
...AC_U97188_at = M: myeloma (0)
AC_U97188_at = P: normal (2)
AC_U97188_at = A:
...AC_HG415-HT415_at = M: myeloma (0)
AC_HG415-HT415_at = A: myeloma (72)
AC_HG415-HT415_at = P: normal (3/1)

Decision tree 4:

AC_M63928_at = M: myeloma (0)
AC_M63928_at = A: myeloma (61.6/0.7)
AC_M63928_at = P:
...AC_D59253_at = M: normal (0)
AC_D59253_at = A: normal (32.8/1.3)
AC_D59253_at = P: myeloma (9.6)

Decision tree 8:

AC_D50925_at = M: myeloma (0)
AC_D50925_at = A:
...AC_U78525_at = M: myeloma (0)
: AC_U78525_at = A: normal (14.3/1.4)
: AC_U78525_at = P: myeloma (61.6/1.7)
AC_D50925_at = P:
...AC_X89750_at = A: myeloma (3.1)
AC_X89750_at = P: normal (23.4/0.6)
AC_X89750_at = M: normal (1.6)

Decision tree 1:

AC_U09579_at = M: myeloma (0)
AC_U09579_at = A: normal (23.7/3.5)
AC_U09579_at = P:
...AC_M69197_xpt2_s_at = M: myeloma (0)
AC_M69197_xpt2_s_at = P: normal (2.6)
AC_M69197_xpt2_s_at = A:
...AC_M60331_at = M: myeloma (0)
AC_M60331_at = A: myeloma (73.3/0.9)
AC_M60331_at = P: normal (4.4/1.8)

Decision tree 5:

AC_U73167_cds5_at = M: normal (0)
AC_U73167_cds5_at = P: myeloma (57.8)
AC_U73167_cds5_at = A:
...AC_U64998_at = M: normal (0)
AC_U64998_at = A: myeloma (20.8/0.8)
AC_U64998_at = P: normal (25.4/0.8)

Decision tree 9:

AC_L36033_at = A: myeloma (50.1)
AC_L36033_at = M: normal (1.7)
AC_L36033_at = P:
...AC_U64998_at = M: normal (0)
AC_U64998_at = A: myeloma (9.5)
AC_U64998_at = P:
...AC_U73167_cds5_at = M: normal (0)
AC_U73167_cds5_at = A: normal (38.9)
AC_U73167_cds5_at = P: myeloma (3.7)

Decision tree 2:

AC_M55267_at = M: myeloma (0)
AC_M55267_at = A:
...AC_HG2441-HT2537_s_at = M: normal (0)
: AC_HG2441-HT2537_s_at = A: myeloma (7.5/1.2)
: AC_HG2441-HT2537_s_at = P: normal (25.4)
AC_M55267_at = P:
...AC_M54992_at = A: myeloma (64.3/0.6)
AC_M54992_at = P: normal (5.6/1.2)
AC_M54992_at = M: myeloma (1.2/0.6)

Decision tree 6:

AC_L05779_at = M: myeloma (0)
AC_L05779_at = P: myeloma (54.8/1.8)
AC_L05779_at = A:
...AC_L06845_at = M: normal (0)
AC_L06845_at = A: normal (38.3/3.6)
AC_L06845_at = P: myeloma (10.9)

Decision tree 3:

AC_X57129_at = M: myeloma (0)
AC_X57129_at = P: myeloma (62.6)
AC_X57129_at = A:
...AC_U73167_cds5_at = M: normal (0)
AC_U73167_cds5_at = P: myeloma (8.6)
AC_U73167_cds5_at = A:
...AC_L42379_at = M: normal (0)
AC_L42379_at = A: normal (29.3/0.7)
AC_L42379_at = P: myeloma (3.6/0.7)

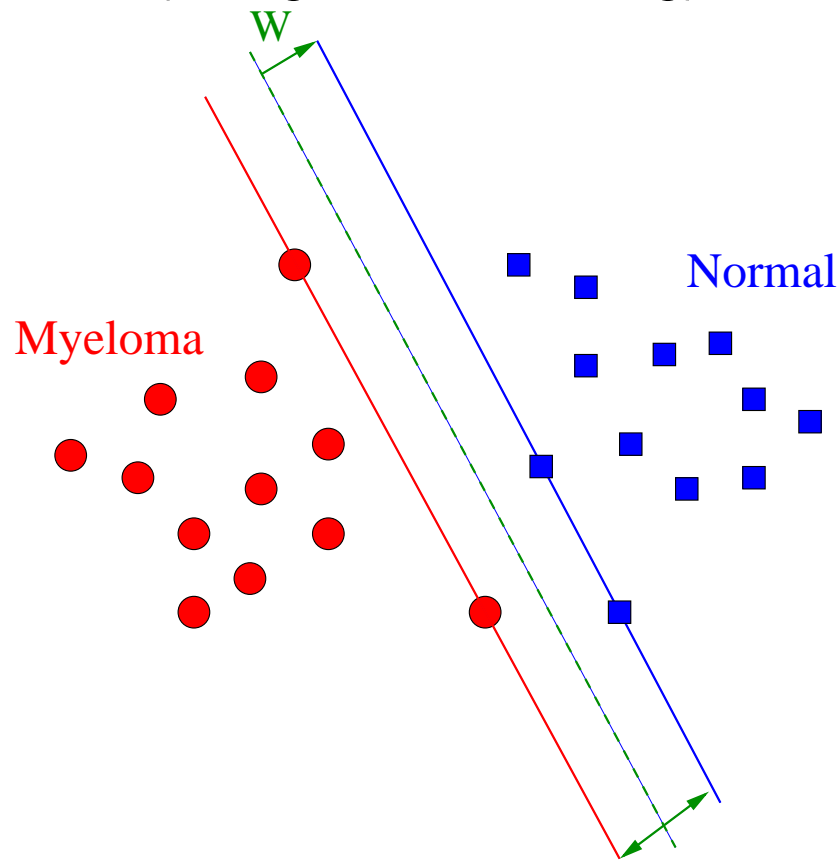
Decision tree 7:

AC_D87447_at = P: myeloma (61.5/3.4)
AC_D87447_at = M: normal (0.6)
AC_D87447_at = A:
...AC_M17733_at = M: normal (0)
AC_M17733_at = A: myeloma (8.7)
AC_M17733_at = P: normal (33.2/1.7)

Figure 6: An example boosted decision tree for “AC Only.” A boosted decision tree is a collection of individual trees that “vote” on the final decision.

Support Vector Machines

Maximizing the Margin between Bounding Planes
(Mangasarian & Fung)



Support Vector Machines

$$\begin{aligned}\alpha_4 &= 3.5627244229813595396438711403089 \times 10^{-12} & \alpha_{43} &= 5.8862508877702210773416306782626 \times 10^{-12} \\ \alpha_5 &= 2.5540568853697465261039034298171 \times 10^{-12} & \alpha_{46} &= 3.1304221192815217402053350048013 \times 10^{-11} \\ \alpha_9 &= 2.9743480105240216300258073135745 \times 10^{-13} & \alpha_{47} &= 3.2245338697842914165727392562437 \times 10^{-13} \\ \alpha_{12} &= 1.622625401180709946285813013707 \times 10^{-12} & \alpha_{50} &= 2.9936090017496582806141789794832 \times 10^{-13} \\ \alpha_{14} &= 1.3482092528423377643109827323138 \times 10^{-11} & \alpha_{61} &= 2.5847971650904005330972452017409 \times 10^{-11} \\ \alpha_{21} &= 1.6316999160198477346554084137298 \times 10^{-12} & \alpha_{64} &= 1.1167616737173964582930202218717 \times 10^{-11} \\ \alpha_{23} &= 3.5300503605894770301449329344089 \times 10^{-12} & \alpha_{67} &= 2.7643880608667294064813714409901 \times 10^{-11} \\ \alpha_{25} &= 1.04575860063064744533551868498 \times 10^{-11} & \alpha_{71} &= 1.6482209938445251625649549522619 \times 10^{-13} \\ \alpha_{28} &= 8.2915004716070301657692546639064 \times 10^{-12} & \alpha_{75} &= -1.2294121053634246621270117874689 \times 10^{-11} \\ \alpha_{30} &= 7.090043266776840637820698577112 \times 10^{-12} & \alpha_{78} &= -4.9558230050706333672931139006264 \times 10^{-11} \\ \alpha_{31} &= 5.1970502257530206419876911029797 \times 10^{-11} & \alpha_{82} &= -1.8742092610524455666809308268549 \times 10^{-12} \\ \alpha_{36} &= 1.4778770472298374719540378223604 \times 10^{-11} & \alpha_{86} &= -1.9287866684058292506270398519414 \times 10^{-11} \\ \alpha_{37} &= 2.4602889365144345791421943611519 \times 10^{-11} & \alpha_{95} &= -1.8838354888680847573119814892997 \times 10^{-11} \\ \alpha_{40} &= 3.6095049887373241944752884233444 \times 10^{-11} & \alpha_{96} &= -4.0853582517483234412572852483917 \times 10^{-11} \\ \alpha_{41} &= 2.5882221470086245850770783624996 \times 10^{-11} & \alpha_{97} &= -2.9231839203610967212820317905043 \times 10^{-11} \\ \alpha_{42} &= -1.0801101194561801851440636655744 \times 10^{-10} & \alpha_{103} &= -2.8536609371763672774136000332963 \times 10^{-11}\end{aligned}$$

Figure 7: The values of α_i for the Support Vector Machine learned on the “AC+AD” data. (Note: α_i for any i not listed is equal to zero.)



Ensembles of Voters

- An ensemble of voters is a collection of “**decision stumps**” (see Figure 5.)
- For the “AC+AD” data, the top 1% (70) genes were selected by **information gain**. (All 70 were AD values.)
- Each of these 70 genes was given **one vote** on whether or not a patient has Multiple Myeloma based upon its splitpoint.

Top 70 Genes

All of the top 70 (by information gain) clones were AD values. The adjusted p-value gives the probability that a gene would look as one-sided as we see.

InfoGain	Gene	Accession						Adjusted p-value
		Number	Split	MH	ML	NH	NL	
0.802422	<i>APOA2</i>	X04898	-777	74	0	1	30	1.363×10^{-21}
0.735975	<i>HERV K22 pol</i>	K03498	637	3	71	31	0	1.174×10^{-19}
0.704489	<i>TERT</i>	AF015950	-1610	70	4	0	31	1.056×10^{-18}
0.701219	<i>UMOD</i>	M15881	1119.1	0	74	28	3	1.364×10^{-18}
0.701219	<i>CDH4</i>	L34059	-278	74	0	3	28	1.364×10^{-18}
0.664859	<i>ACTR1A</i>	Z14978	3400.6	3	71	30	1	7.853×10^{-18}
0.664859	<i>MASP1</i>	D17525	-536.6	71	3	1	30	7.853×10^{-18}
0.650059	<i>PTPN21</i>	X79510	1256.1	6	68	31	0	4.950×10^{-17}
0.650059	<i>TCEB3</i>	L47345	1451.4	6	68	31	0	4.950×10^{-17}
0.63397	<i>SDF1</i>	L36033	2178.6	4	70	30	1	6.783×10^{-17}
0.625966	<i>TNFRSF7</i>	M63928	8129	7	67	31	0	2.758×10^{-16}
0.625966	<i>UROD</i>	M14016	2718.9	7	67	31	0	2.758×10^{-16}
0.625966	<i>KIAA0135</i>	D50925	788.7	7	67	31	0	2.758×10^{-16}
0.625966	<i>KIAA0133</i>	D50923	-1517	67	7	0	31	2.758×10^{-16}
0.612641	<i>PML</i>	M79463	88.7	71	3	2	29	2.657×10^{-16}
0.612641	<i>IFNA4</i>	M27318	415	3	71	29	2	2.657×10^{-16}
0.612641	<i>PPP2R5D</i>	L76702	-1601.8	71	3	2	29	2.657×10^{-16}

Top 70 Genes

InfoGain	Gene	Accession Number	Split	MH	ML	NH	NL	Adjusted p-value
0.606152	<i>H2BFQ</i>	X57985	506.3	69	5	1	30	4.821×10^{-16}
0.606152	<i>ABL1</i>	X16416	709	69	5	1	30	4.821×10^{-16}
0.606152	<i>DCTD</i>	L39874	3545	69	5	1	30	4.821×10^{-16}
0.606152	<i>H2AFO</i>	L19779	7368	69	5	1	30	4.821×10^{-16}
0.603481	<i>CNGB1</i>	U58837	173.8	8	66	31	0	1.379×10^{-15}
0.603481	<i>ADRA1B</i>	HT4369	994	8	66	31	0	1.379×10^{-15}
0.603481	<i>RAD23A</i>	D21235	5044	66	8	0	31	1.379×10^{-15}
0.582583	<i>NNT</i>	U40490	-2.1	74	0	6	25	5.819×10^{-15}
0.582366	<i>DUSP7</i>	X93921	1088	9	65	31	0	6.282×10^{-15}
0.58236	<i>MAPK12</i>	X79483	1814.7	4	70	29	2	2.201×10^{-15}
0.580711	<i>H326</i>	U06631	1501	68	6	1	30	2.934×10^{-15}
0.580711	<i>APOE</i>	M12529	1044	6	68	30	1	2.934×10^{-15}
0.580711	<i>SLC34A1</i>	L13258	732.5	6	68	30	1	2.934×10^{-15}
0.562437	<i>H2BFH</i>	Z80780	455.7	64	10	0	31	2.639×10^{-14}
0.562437	<i>H1F2</i>	X57129	455	64	10	0	31	2.639×10^{-14}
0.562437	<i>CTSH</i>	X16832	3458.7	10	64	31	0	2.639×10^{-14}
0.562437	<i>RXRB</i>	U41068	2555	10	64	31	0	2.639×10^{-14}
0.562437	<i>CD81</i>	M33680	9422	10	64	31	0	2.639×10^{-14}
0.562437	<i>DSC3</i>	D17427	1172.6	10	64	31	0	2.639×10^{-14}

Top 70 Genes

InfoGain	Gene	Accession Number	Split	MH	ML	NH	NL	Adjusted p-value
0.561439	<i>ABCB10</i>	U18237	-284	73	1	5	26	1.137×10^{-14}
0.557191	<i>SCAP</i>	D83782	-662	67	7	1	30	1.571×10^{-14}
0.555139	<i>ITGB2</i>	X64072	230	5	69	29	2	1.501×10^{-14}
0.543549	<i>HRAS</i>	V00574	117.7	63	11	0	31	1.031×10^{-13}
0.543549	<i>ITS1</i>	U13369	353	11	63	31	0	1.031×10^{-13}
0.543549	<i>KIAA00167</i>	D28589	4334	63	11	0	31	1.031×10^{-13}
0.537806	<i>S100A5</i>	Z18954	2142	4	70	28	3	4.810×10^{-14}
0.537806	<i>MAS1</i>	M13150	-433	70	4	3	28	4.810×10^{-14}
0.537806	<i>PPY</i>	M11726	1366.6	4	70	28	3	4.810×10^{-14}
0.535273	<i>OR1D2</i>	X65857	617.6	8	66	30	1	7.557×10^{-14}
0.535273	<i>NARS</i>	U79254	4613	66	8	1	30	7.557×10^{-14}
0.535273	<i>ATR</i>	U49844	1128.8	66	8	1	30	7.557×10^{-14}
0.535273	<i>FCER1G</i>	M33195	805.2	8	66	30	1	7.557×10^{-14}
0.535273	<i>TCN2</i>	L02648	1481	8	66	30	1	7.557×10^{-14}
0.530287	<i>ALDOC</i>	X05196	3226	6	68	29	2	8.776×10^{-14}
0.530287	<i>GNRH1</i>	X01059	297.8	6	68	29	2	8.776×10^{-14}
0.530287	<i>GNL1</i>	HT3404	2757	6	68	29	2	8.776×10^{-14}
0.525585	<i>MHC2 beta W52</i>	HT3779	757	12	62	31	0	3.782×10^{-13}
0.525585	<i>HML2</i>	D50532	1255	12	62	31	0	3.782×10^{-13}

Top 70 Genes

InfoGain	Gene	Accession Number	Split	MH	ML	NH	NL	Adjusted p-value
0.515204	<i>RFX2</i>	HT3504	-4294.2	74	0	8	23	7.073×10^{-13}
0.514718	<i>H4FL</i>	Z80788	-346	65	9	1	30	3.313×10^{-13}
0.514718	<i>ABCC1</i>	L05628	-938.8	65	9	1	30	3.313×10^{-13}
0.514718	<i>KIAA0084</i>	D42043	1375.7	9	65	30	1	3.313×10^{-13}
0.511192	<i>PSEN1</i>	U40380	3746	5	69	28	3	3.147×10^{-13}
0.508447	<i>DPYSL2</i>	U97105	305.2	13	61	31	0	1.309×10^{-12}
0.508447	<i>CD79A</i>	U05259	9482.1	13	61	31	0	1.309×10^{-12}
0.508447	<i>ITGB2</i>	M15395	146	13	61	31	0	1.309×10^{-12}
0.507348	<i>ATIC</i>	D82348	3516	67	7	2	29	4.517×10^{-13}
0.497793	<i>anonymous gene</i>	L18972	1737.5	4	70	27	4	7.958×10^{-13}
0.495344	<i>NAP1L4</i>	U77456	1745	64	10	1	30	1.339×10^{-12}
0.495344	<i>ETV4</i>	U18018	1331	10	64	30	1	1.339×10^{-12}
0.492055	<i>MYO1F</i>	X98411	1744.8	14	60	31	0	4.301×10^{-12}
0.492055	<i>PLXNB1</i>	X87904	-600.5	60	14	0	31	4.301×10^{-12}
0.492055	<i>TXN</i>	X77584	11024	60	14	0	31	4.301×10^{-12}

Naïve Bayesian Networks

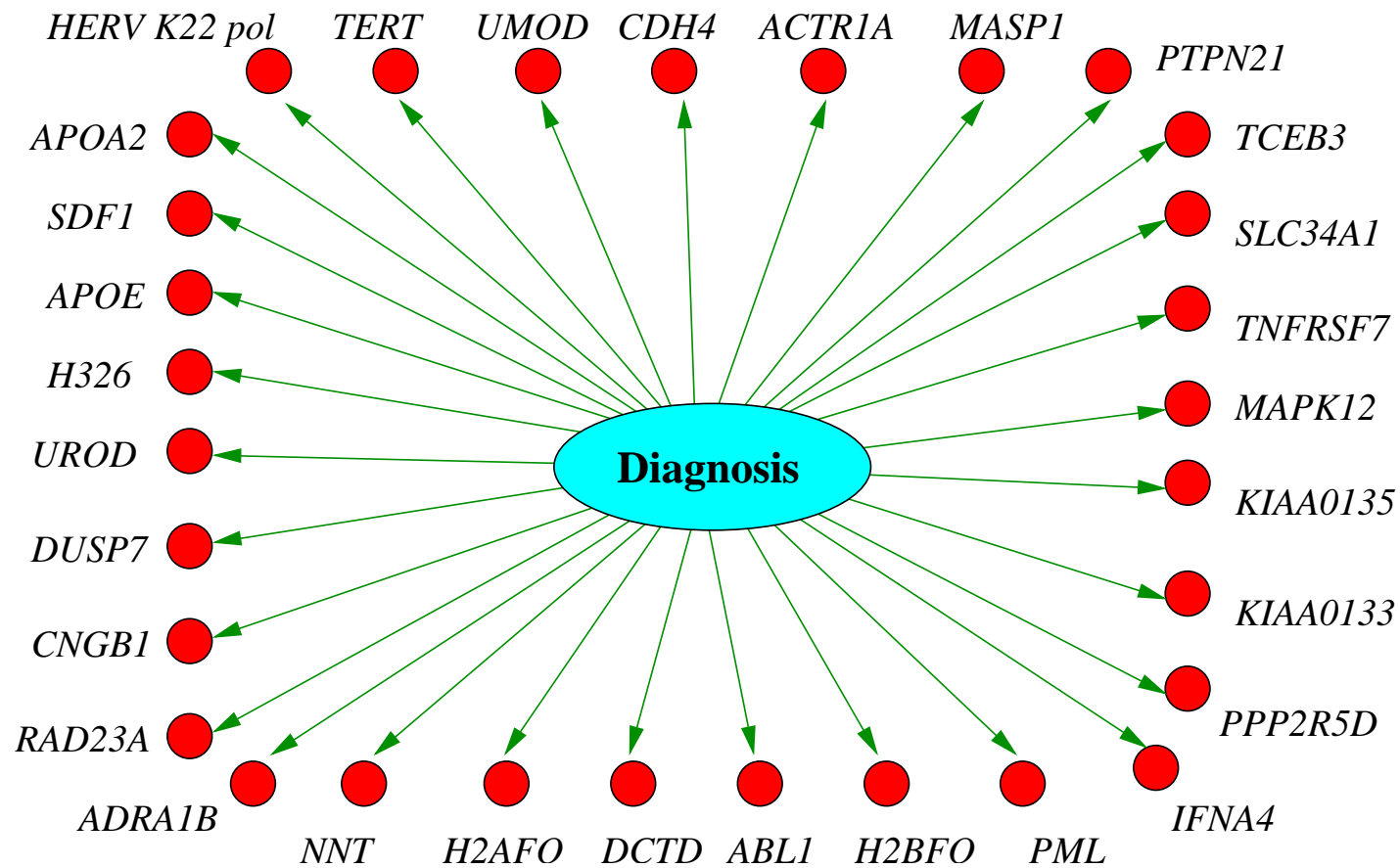


Figure 8: The Bayes net learned from AC+AD had the structure of a Naïve Bayes net.

Bayesian Networks

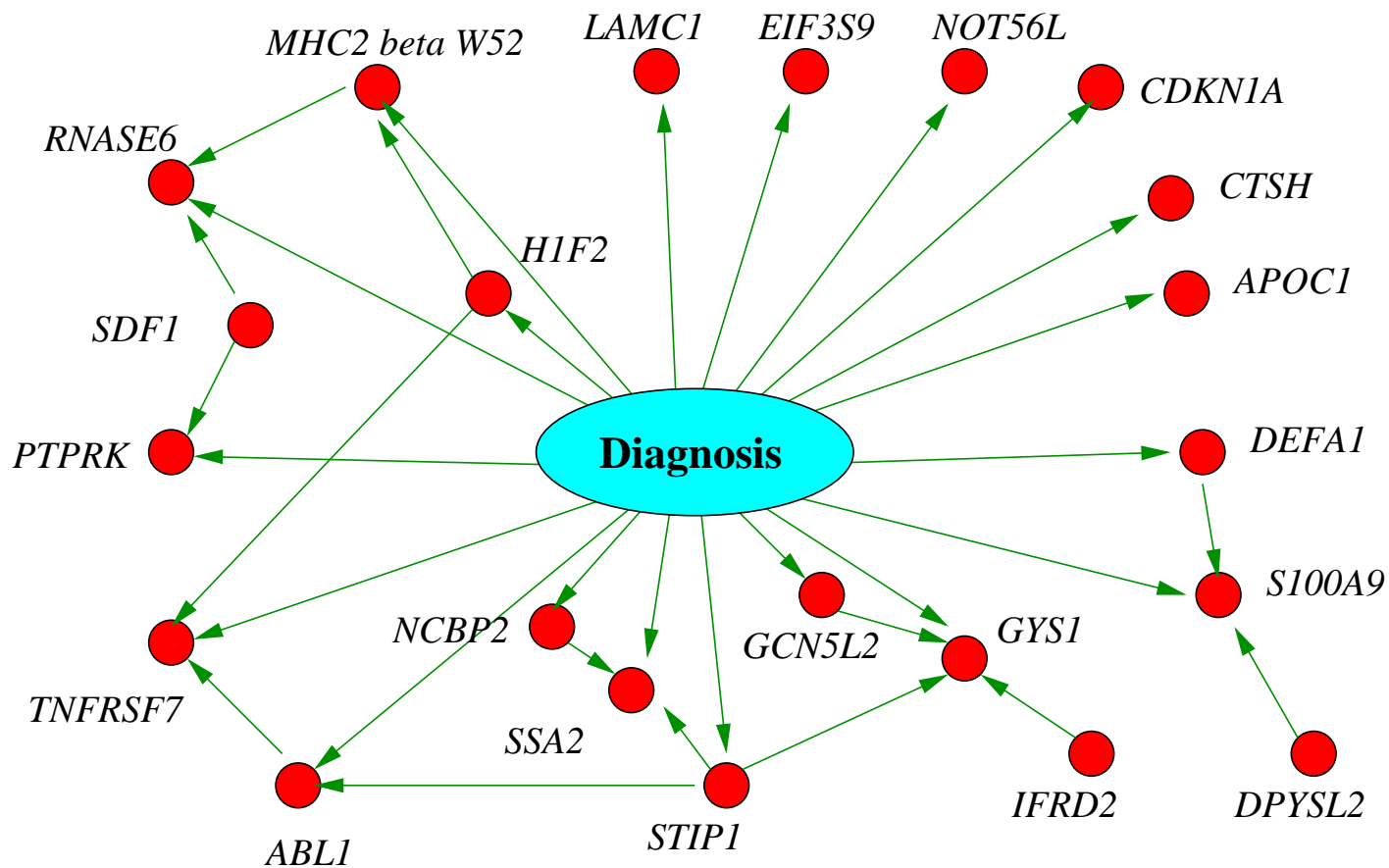


Figure 9: The structure of the Bayes net learned from AC only.

Resources

● Trees and Boosted Trees

- C5.0 (<http://www.rulequest.com>)

● Support Vector Machines

- SVM^{light} (<http://svmlight.joachims.org>)

- Used linear SVMs which worked better than Gaussian kernels

● Unweighted Ensembles of Voters and Naïve Bayesian Networks

- ensemble of voting decision stumps using 1% of the genes as stumps
- EOV (<http://www.biostat.wisc.edu/~mwaddell/eov.html>)

● Bayesian Networks

- BayesNet PowerPredictor (<http://www.cs.ualberta.ca/~jcheng>)
- The data was discretized and limited to the top 30 features (by information gain) due to the limitations of the system.