Rule mining

Oleg Shpynov

JetBrains Biolabs

March 20, 2015

Agenda

- Association rule
- Support-Confidence framework
- Lift, Conviction
- Comparison[1]
- Appilication to Epigenetics

[1] Comparing Rule Measures for Predictive Association Rules http://www.di.uminho.pt/~pja/ps/conviction.pdf

Notions

- T - set of items.
- Transaction d - subset of T
- DataBase D - list of transcation $d_i \subseteq T$
- Association Rule $X \to Y$.
  $X \subseteq T$, $Y \subseteq T$, $X \cap Y = \emptyset$
- $supp(X) = \frac{\#\{d_i | X \subseteq d_i\}}{\#\{D\}} = P(X)$
- $conf(X \to Y) = \frac{supp(X \to Y)}{supp(X)} = \frac{P(X \wedge Y)}{P(X)} = P(Y|X)$
- $lift(X \to Y) = lift(Y \to X) = \frac{conf(X \to Y)}{supp(Y)} = \frac{P(X \wedge Y)}{P(X)P(Y)}$
- $conviction(X \to Y) = \frac{1 - supp(Y)}{1 - conf(X \to Y)} = \frac{P(X)P(\neg Y)}{P(X \wedge \neg Y)}$

Measures[2]

Table 2. Association rule quality measures

	Definition	Co-domain
BF	$\frac{n_{ab}n_{\bar{b}}}{n_b n_{a\bar{b}}}$	$[0, +\infty[$
CenConf	$\frac{nn_{ab}-n_a n_b}{nn_a}$	$[-\frac{n_b}{n}, \frac{n_{\bar{b}}}{n}]$
Conf	$\frac{n_{ab}}{n_a}$	$[0, 1]$
Conv	$\frac{n_a n_{\bar{b}}}{nn_{a\bar{b}}}$	$[\frac{n_{\bar{b}}}{n}, +\infty[$
ECR	$\frac{n_{ab}-n_{a\bar{b}}}{n_{ab}} = 1 - \frac{1}{\frac{n_a}{n_{a\bar{b}}}-1}$	$]-\infty, 1]$
EII	$\left\{[(1-h_1(\frac{n_{a\bar{b}}}{n})^2)(1-h_2(\frac{n_{a\bar{b}}}{n})^2)]^{1/4}\text{IntImp}\right\}^{1/2}$	$[0, 1]$
IG	$\log(\frac{nn_{ab}}{n_a n_b})$	$]-\infty, \log\frac{n}{n_b}]$
-ImpInd	$\frac{n_a n_b - nn_{ab}}{\sqrt{nn_a n_{\bar{b}}}}$	$[-\frac{\sqrt{n_a}n_b}{\sqrt{nn_{\bar{b}}}}, \sqrt{\frac{n_a n_{\bar{b}}}{n}}]$
IntImp	$P\left[N(0,1) \geq \text{ImpInd}\right]$	$[0, 1]$
Kappa	$2\frac{nn_{ab}-n_a n_b}{nn_a+nn_b-2n_a n_b}$	$[-2\frac{n_a n_b}{n_a n_{\bar{b}}+n_{\bar{a}} n_b}, 2\frac{n_a n_{\bar{b}}}{n_a n_{\bar{b}}+n_{\bar{a}} n_b}]$
Lap	$\frac{n_{ab}+1}{n_a+2}$	$[\frac{1}{n_a+2}, \frac{n_a+1}{n_a+2}]$
LC	$\frac{n_{ab}-n_{a\bar{b}}}{n_b}$	$[-\frac{n_a}{n_b}, \frac{n_a}{n_b}]$
Lift	$\frac{nn_{ab}}{n_a n_b}$	$[0, \frac{n}{n_b}]$
Loe	$\frac{nn_{ab}-n_a n_b}{n_a n_{\bar{b}}}$	$[-\frac{n_b}{n_{\bar{b}}}, 1]$
PDI	$P\left[\mathcal{N}(0,1) > \text{ImpInd}^{CR/\mathcal{B}}\right]$	$]0, 1[$
PS	$n_{ab} - \frac{n_a n_b}{n}$	$[-\frac{n_a n_b}{n}, \frac{n_a n_{\bar{b}}}{n}]$
r	$\frac{nn_{ab}-n_a n_b}{\sqrt{nn_a n_b n_{\bar{a}}.n_{\bar{b}}}}$	$[-\sqrt{\frac{n_a n_b}{nn_{\bar{a}} n_{\bar{b}}}}, \sqrt{\frac{n_a n_{\bar{b}}}{nn_{\bar{a}} n_b}}]$
Seb	$\frac{n_{ab}}{n_{a\bar{b}}}$	$[0, +\infty[$
Sup	$\frac{n_{ab}}{n}$	$[0, \frac{n_a}{n}]$
Zhang	$\frac{nn_{ab}-n_a n_b}{\max\{n_{ab}n_{\bar{b}}, n_b n_{a\bar{b}}\}}$	$[-1, 1]$

[2] Guillet Hamilton - Quality Measures in Data Mining (2007)

Epigenetics

DataBases

- Binned
- Gene Locus (TSS in BivalentHCPExperiment)
- any locations (in this terms any kind of LocationsRule can be reproduced)

Predicates

- CpG predicate
- Enrichemnt predicate max vote within (DZip)? PoissonHMM model
- ChromHMM predicate location is marked if $> 50\%$ intersection with state

Hypotheses

- Cpg content $\rightarrow$ Histone enrichment
- Histone enrichment $\rightarrow$ Histone enrichment
- ChromHMM State $\rightarrow$ Histone enrichment
- Histone enrichment $\rightarrow$ ChromHMM State

Database Rule

Hypothesis $Condition \rightarrow Target$ on DataBase D.

- What if *Target* is property of D itself?

Rule

Hypothesis $Condition \rightarrow Target$ on DataBase D.

- Measure $conviction(Condition \rightarrow Target)$ on D.
- What if we get good value by chance?

Mean, SD, Z-score

Consider we have rule $X \to Y$ and metrics function F. We compute $F(X \to Y)$. If the result is good by chance?
Bootstrap:

- C times sample dataset D' from D
- Compute $F(X \to Y)$ on D'
- Compute $zScore(F, X \to Y, D) = \frac{F(X \to Y)_D - mean(F)_{D'}}{sd(F)_{D'}}$
- *zScore* shows how stable is association rule in terms of metrics F
- In case of HUGE D let's deal with $mean(F)_D$ and $sd(F)_{D'}$ instead of *zScore*

NOTE: in case of database rule we use resampling within WHOLE genome rather than D'.

Resampling

- Problem
  Consider rule $F(X \rightarrow Y)$ on D.
  In case when $support(Y) \approx 1$, $conviction(F(X \rightarrow Y))$ is proportional to $support(X)$, which in turn is likely to be proportional to $\#(D)$.
  Once we are doing resampling *conviction* from D‘ we get both *mean* and *sd* proportional to D‘.
- Solution
  Use resampling of the same size (boostrap).
- Roadmap
  Use bootstrap to estimate $mean(F)_D$ and $sd(F)_D$,
  $mean(support(Target))_{D_{genomic}}$ and $sd(Target)_{D_{genomic}}$.

Other metricss

According to publication[3] several interesting rule approach are introduces. Consider the following two realistic scenarios for the analysis:

- Sc1: The expert E_r tolerates the appearance of a certain number of counter-examples$X \wedge \neg Y$ to a decision rule. In this case, the rejection of a rule is postponed until enough counter-examples are found.
- Sc2: The expert E_r refuses the appearance of too many counter-examples to a decision rule. The rejection of the rule must be done rapidly with respect to the number of counter-examples.

According to these scenarios:

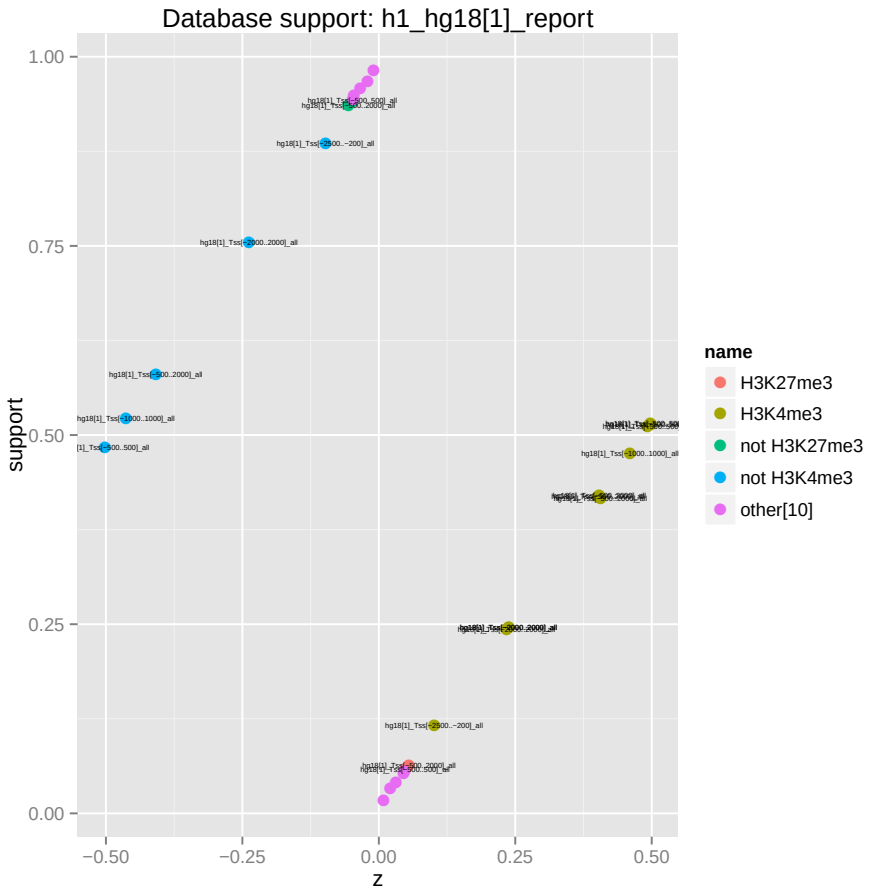
- Baeysian Factor and Conviction are leading Sc2.
- Loe is well placed in both scenarios. It stands for a good compromise.

[3] Guillet Hamilton - Quality Measures in Data Mining (2007)

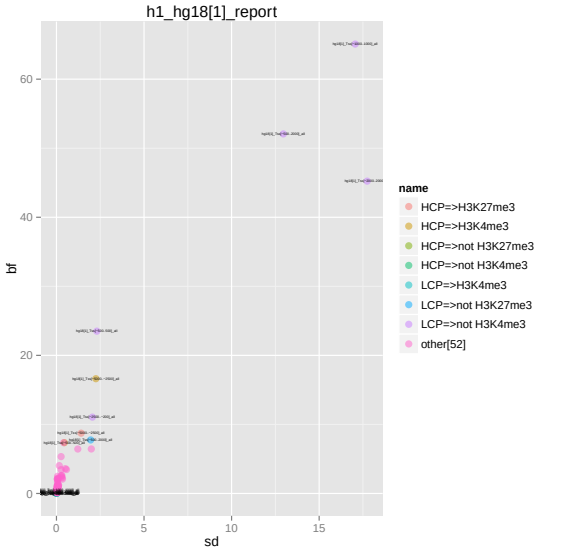
Formulas

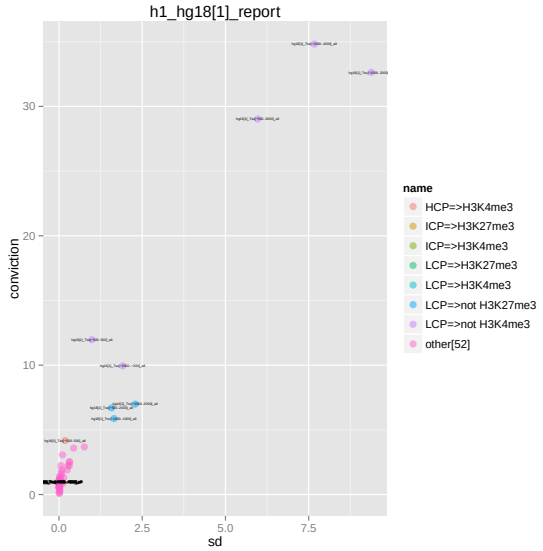
- $conviction(X \to Y) = \frac{1-supp(Y)}{1-conf(X \to Y)} = \frac{P(X)P(\neg Y)}{P(X \wedge \neg Y)}$
- $BF(X \to Y) = \frac{P(X \wedge Y)P(\neg Y)}{P(X)P(X \wedge \neg Y)}$
- $LOE(X \to Y) = \frac{nsup(X \wedge Y) - sup(X)sup(Y)}{sup(X)sup(\neg Y)}$

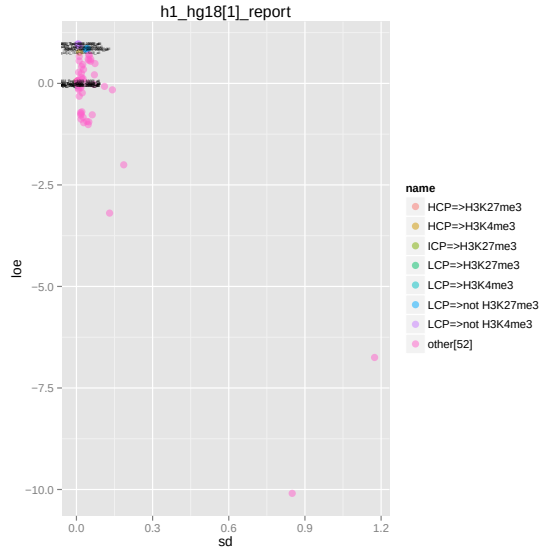
BivalentHCPExperiment TOP 20 Rules



database	name	support	z
hg18[1]_Tss[−500..500]_all	not H3K4me3	0.48340721	−0.50182330
hg18[1]_Tss[−500..500]_all	H3K4me3	0.51456355	0.49836148
hg18[1]_Tss[−500..500]_all	H3K4me3	0.51493894	0.49728261
hg18[1]_Tss[−500..500]_all	H3K4me3	0.51153023	0.49222333
hg18[1]_Tss[−1000..1000]_all	not H3K4me3	0.52238052	−0.46315876
hg18[1]_Tss[−1000..1000]_all	H3K4me3	0.47600030	0.46032701
hg18[1]_Tss[−500..2000]_all	not H3K4me3	0.58003166	−0.40840962
hg18[1]_Tss[−500..2000]_all	H3K4me3	0.41697874	0.40613100
hg18[1]_Tss[−500..2000]_all	H3K4me3	0.41986884	0.40409610
hg18[1]_Tss[−500..2000]_all	H3K4me3	0.41904870	0.40188767
hg18[1]_Tss[−2000..2000]_all	H3K4me3	0.24568240	0.23812253
hg18[1]_Tss[−2000..2000]_all	not H3K4me3	0.75431609	−0.23789031
hg18[1]_Tss[−2000..2000]_all	H3K4me3	0.24633539	0.23757247
hg18[1]_Tss[−2000..2000]_all	H3K4me3	0.24380335	0.23412316
hg18[1]_Tss[−2500..−200]_all	H3K4me3	0.11609561	0.10138927
hg18[1]_Tss[−2500..−200]_all	not H3K4me3	0.88533404	−0.09735037
hg18[1]_Tss[−500..2000]_all	H3K27me3	0.06391226	0.05534517
hg18[1]_Tss[−500..2000]_all	not H3K27me3	0.93537766	−0.05521948
hg18[1]_Tss[−500..500]_all	not H3K27me3	0.94170963	−0.04829908
hg18[1]_Tss[−500..500]_all	H3K27me3	0.05815167	0.04829085







database	name	bf	sd
hg18[1]_Tss[−1000..1000]_all	LCP=>not H3K4me3	65.08536833	17.058411196
hg18[1]_Tss[−500..2000]_all	LCP=>not H3K4me3	52.06879655	12.955747583
hg18[1]_Tss[−2000..2000]_all	LCP=>not H3K4me3	45.23124861	17.742208560
hg18[1]_Tss[−500..500]_all	LCP=>not H3K4me3	23.50585637	2.306644524
hg18[1]_Tss[−5000..−2500]_all	HCP=>H3K4me3	16.62097961	2.249341913
hg18[1]_Tss[−2500..−200]_all	LCP=>not H3K4me3	11.05450693	2.052467677
hg18[1]_Tss[−5000..−2500]_all	HCP=>H3K27me3	8.72969029	1.409297585
hg18[1]_Tss[−500..2000]_all	LCP=>not H3K27me3	7.74824891	1.953177514
hg18[1]_Tss[−500..500]_all	HCP=>H3K4me3	7.36253785	0.451550290
hg18[1]_Tss[−1000..1000]_all	LCP=>H3K4me3	0.01712283	0.003020593
hg18[1]_Tss[−500..2000]_all	LCP=>H3K4me3	0.02037901	0.003903056
hg18[1]_Tss[−500..500]_all	LCP=>H3K4me3	0.04185418	0.004605073
hg18[1]_Tss[−5000..−2500]_all	HCP=>not H3K4me3	0.05938941	0.005778953
hg18[1]_Tss[−2500..−200]_all	HCP=>not H3K4me3	0.13777981	0.006561686
hg18[1]_Tss[−500..2000]_all	HCP=>not H3K4me3	0.24541790	0.006753933
hg18[1]_Tss[−2000..2000]_all	LCP=>H3K4me3	0.02602945	0.007319790
hg18[1]_Tss[−2000..2000]_all	HCP=>not H3K4me3	0.39432824	0.008290350
hg18[1]_Tss[−500..500]_all	HCP=>not H3K4me3	0.13707598	0.008821247
hg18[1]_Tss[−1000..1000]_all	HCP=>not H3K4me3	0.19133806	0.009348596
hg18[1]_Tss[−1000..1000]_all	HCP=>not H3K27me3	0.48883961	0.012526217

database	name	conviction	sd
hg18[1]_Tss[−1000..1000]_all	LCP=>not H3K4me3	34.8208760	7.671513719
hg18[1]_Tss[−2000..2000]_all	LCP=>not H3K4me3	32.6352665	9.384016041
hg18[1]_Tss[−500..2000]_all	LCP=>not H3K4me3	29.0243058	5.968413482
hg18[1]_Tss[−500..500]_all	LCP=>not H3K4me3	11.9704296	0.992568829
hg18[1]_Tss[−2500..−200]_all	LCP=>not H3K4me3	9.9401796	1.913823888
hg18[1]_Tss[−2000..2000]_all	LCP=>not H3K27me3	6.9896202	2.292747942
hg18[1]_Tss[−500..2000]_all	LCP=>not H3K27me3	6.6964655	1.579993848
hg18[1]_Tss[−1000..1000]_all	LCP=>not H3K27me3	5.8735796	1.655363835
hg18[1]_Tss[−500..500]_all	HCP=>H3K4me3	4.1727057	0.185056179
hg18[1]_Tss[−5000..−2500]_all	LCP=>H3K27me3	0.9917297	0.001214544
hg18[1]_Tss[−5000..−2500]_all	LCP=>H3K4me3	0.9796646	0.001855044
hg18[1]_Tss[−5000..−2500]_all	ICP=>H3K27me3	1.0044879	0.002262248
hg18[1]_Tss[−2500..−200]_all	LCP=>H3K27me3	0.9750451	0.002935859
hg18[1]_Tss[−500..500]_all	LCP=>H3K27me3	0.9595320	0.002941297
hg18[1]_Tss[−2000..2000]_all	ICP=>H3K27me3	0.9641219	0.003000394
hg18[1]_Tss[−1000..1000]_all	LCP=>H3K27me3	0.9542729	0.003027604
hg18[1]_Tss[−2000..2000]_all	LCP=>H3K27me3	0.9553151	0.003122753
hg18[1]_Tss[−5000..−2500]_all	ICP=>H3K4me3	1.0012218	0.003157894
hg18[1]_Tss[−500..2000]_all	LCP=>H3K27me3	0.9461263	0.003324978
hg18[1]_Tss[−2500..−200]_all	LCP=>H3K4me3	0.8968385	0.003588694

database	name	loe	sd
hg18[1]_Tss[−1000..1000]_all	LCP=>not H3K4me3	9.689832e−01	0.004727505
hg18[1]_Tss[−500..2000]_all	LCP=>not H3K4me3	9.679461e−01	0.006715466
hg18[1]_Tss[−2000..2000]_all	LCP=>not H3K4me3	9.663253e−01	0.008979424
hg18[1]_Tss[−500..500]_all	LCP=>not H3K4me3	9.135322e−01	0.007764123
hg18[1]_Tss[−2500..−200]_all	LCP=>not H3K4me3	8.981624e−01	0.013555177
hg18[1]_Tss[−500..2000]_all	LCP=>not H3K27me3	8.563724e−01	0.039216454
hg18[1]_Tss[−2000..2000]_all	LCP=>not H3K27me3	8.543292e−01	0.044156056
hg18[1]_Tss[−1000..1000]_all	LCP=>not H3K27me3	8.266021e−01	0.038259117
hg18[1]_Tss[−500..500]_all	HCP=>H3K4me3	7.654179e−01	0.012240568
hg18[1]_Tss[−5000..−2500]_all	LCP=>H3K27me3	−8.667886e−03	0.001569423
hg18[1]_Tss[−5000..−2500]_all	LCP=>H3K4me3	−2.106586e−02	0.002283535
hg18[1]_Tss[−2000..2000]_all	HCP=>H3K27me3	5.138642e−02	0.002735378
hg18[1]_Tss[−2500..−200]_all	LCP=>H3K27me3	−2.554546e−02	0.002759106
hg18[1]_Tss[−5000..−2500]_all	ICP=>H3K27me3	2.817012e−03	0.002948829
hg18[1]_Tss[−500..2000]_all	LCP=>H3K27me3	−5.754226e−02	0.003076083
hg18[1]_Tss[−2500..−200]_all	ICP=>H3K27me3	6.471627e−07	0.003127994
hg18[1]_Tss[−1000..1000]_all	LCP=>H3K27me3	−4.853927e−02	0.003297057
hg18[1]_Tss[−500..500]_all	LCP=>H3K27me3	−4.305592e−02	0.003304023
hg18[1]_Tss[−1000..1000]_all	HCP=>H3K27me3	5.550393e−02	0.003389775
hg18[1]_Tss[−2000..2000]_all	LCP=>H3K27me3	−4.777833e−02	0.003528619

Metrics

Everything is quite similar.

Stability *conviction* vs *BF* vs *LOE*

TSS[-1000, 1000]

atabase	name	bf		name	convitabase	name
s[−1000..1000]_all	CTCF=>Input_Sonicated	6627.720	000]_all	H3K27me3 and H3K4me3 or H3K4me1 and H3K9Ac=>H3K4me2	388.75[−1000..1000]_all	H3K27me3 and H3K36me3=>H4K20me1
s[−1000..1000]_all	Input_Sonicated=>CTCF	5832.220	2000]_all	H3K27me3 and H3K4me3 or H4K20me1 and H3K9Ac=>H3K4me2	319.23[−1000..1000]_all	CTCF or H3K27me3 and H3K36me3=>H4K20me1
s[−1000..1000]_all	CTCF and H3K27me3=>Input_Sonicated	5061.174	1000]_all	not H3K27me3 and Input_Sonicated or H3K4me1 and H3K9Ac=>H3K4me2	302.81[−1000..1000]_all	CTCF or H3K27me3 and H3K27ac=>H4K20me1
s[−1000..1000]_all	H3K27me3 and Input_Sonicated=>CTCF	4992.886	1000]_all	H3K27me3 and H3K27ac or H3K4me1 and H3K9Ac=>H3K4me2	302.24[−1000..1000]_all	H3K36me3 and H3K9Ac=>H4K20me1
s[−1000..1000]_all	CTCF or H3K27me3 and H3K27ac=>Input_Sonicated	2544.685	2000]_all	H3K27me3 and Input_Sonicated or H3K4me1 and H3K9Ac=>H3K4me2	300.69[−1000..1000]_all	H3K27me3 and H3K27ac or H3K36me3 and H3K9Ac=>H4K20me1
s[−1000..1000]_all	H3K27me3 and H3K27ac=>Input_Sonicated	2488.021	2000]_all	H3K4me1 and H3K9Ac=>H3K4me2	299.78[−1000..1000]_all	H3K27me3 and H3K27ac or Input_Sonicated=>H4K20me1
s[−1000..1000]_all	H3K27me3 and H3K27ac=>CTCF	2137.571	1000]_all	not CTCF and H3K9Ac or Input_Sonicated=>H3K4me2	288.56[−1000..1000]_all	H3K27me3 and H3K36me3 or Input_Sonicated=>H4K20me1
s[−1000..1000]_all	H3K27me3 and H3K27ac or Input_Sonicated=>CTCF	2079.140	1000]_all	not CTCF and H3K9Ac or H3K27me3 and H3K4me3=>H3K4me2	280.01[−1000..1000]_all	H3K27me3 and Input_Sonicated or H3K36me3 and H3K9Ac=>H4K20me1
s[−1000..1000]_all	CTCF or not H3K4me2 and H3K9Ac=>Input_Sonicated	2021.590	1000]_all	H4K20me1 and H3K4me3 or H3K4me1 and H3K9Ac=>H3K4me2	276.45[−1000..1000]_all	not H3K27me3 and Input_Sonicated or H3K36me3 and H3K9Ac=>H4K20m
s[−1000..1000]_all	not H3K36me3=>CTCF	0.000	000]_all	H3K4me2 or not H3K9Ac=>CTCF	1.00[−1000..1000]_all	H3K27me3 and H3K36me3=>H3K4me1
s[−1000..1000]_all	H3K27me3 and not H3K36me3=>CTCF	0.000	000]_all	H3K4me2 or not H3K9Ac=>Input_Sonicated	1.00[−1000..1000]_all	CTCF or H3K27me3 and H3K36me3=>H3K4me1
s[−1000..1000]_all	not H3K27me3 and not H3K36me3=>CTCF	0.000	000]_all	H3K4me2 or not H3K27ac and not H3K9Ac=>Input_Sonicated	1.00[−1000..1000]_all	H3K27me3 and H3K27ac=>H3K4me1
s[−1000..1000]_all	H3K27me3 and not H4K20me1=>CTCF	0.000	000]_all	H3K4me2 and not Input_Sonicated or not H3K4me3=>CTCF	1.00[−1000..1000]_all	CTCF or H3K27me3 and H3K27ac=>H3K4me1
s[−1000..1000]_all	not H3K27me3 and not H4K20me1=>CTCF	0.000	000]_all	H3K4me2 or not H3K4me3=>CTCF	1.00[−1000..1000]_all	H3K27me3 and H3K9Ac=>H3K4me1
s[−1000..1000]_all	not H3K36me3 and not H4K20me1=>CTCF	0.000	000]_all	H3K4me2 or not H3K4me3 and not H3K9Ac=>Input_Sonicated	1.00[−1000..1000]_all	CTCF or H3K27me3 and H3K9Ac=>H3K4me1
s[−1000..1000]_all	H3K36me3 and not H4K20me1=>CTCF	0.000	000]_all	not H3K27me3 and not Input_Sonicated or H3K4me1=>CTCF	1.00[−1000..1000]_all	H3K27me3 and H3K9Ac or Input_Sonicated=>H3K4me1
s[−1000..1000]_all	not H4K20me1=>CTCF	0.000	000]_all	H3K4me2 or not H3K27ac=>CTCF	1.00[−1000..1000]_all	H3K27me3 and H3K27ac or Input_Sonicated=>H3K4me1
s[−1000..1000]_all	not H3K36me3 and H4K20me1=>CTCF	0.000	000]_all	H3K4me2 or not H3K4me3=>Input_Sonicated	1.00[−1000..1000]_all	H3K27me3 and H3K36me3 or Input_Sonicated=>H3K4me1
s[−1000..1000]_all	not H3K36me3 or not H4K20me1=>CTCF	0.000	000]_all	H3K4me2 or not H3K4me3 and not H3K9Ac=>CTCF	1.00[−1000..1000]_all	CTCF=>H3K4me2
s[−1000..1000]_all	H3K27me3 and not H4K20me1 or not H3K36me3=>CTCF	0.000	000]_all	H3K4me2 or not H3K27ac=>Input_Sonicated	1.00[−1000..1000]_all	H3K27me3 and H3K4me3=>H3K4me2

Genome Binned[200]

se	name	bf	ase	name	conviction	database	name	loe	s
n_200	H3K4me1 and H3K4me3=>H3K4me2	2543.285	bin_200	H3K4me1 and H3K4me3=>H3K4me2	68.30338	g18[1]_bin_200	H3K27me3 and H3K9Ac=>H4K20me1	1	0
n_200	H3K4me3 and not H3K9Ac=>H3K4me2	2508.343	bin_200	H3K4me3 and not H3K9Ac=>H3K4me2	65.43184	g18[1]_bin_200	CTCF and H3K27me3=>H3K4me1	1	0
n_200	H3K27ac and H3K9Ac=>H3K4me3	2364.503	bin_200	H3K27me3 and H3K27ac or H3K4me1 and H3K4me3=>H3K4me2	63.04351	g18[1]_bin_200	CTCF and H3K36me3=>H3K4me1	1	0
n_200	H3K27me3 and H3K27ac or H3K4me1 and H3K4me3=>H3K4me2	2337.043	bin_200	H3K27me3 and H3K27ac or H3K4me3 and not H3K9Ac=>H3K4me2	62.79151	g18[1]_bin_200	CTCF and H3K36me3 or H3K27me3 and H3K27ac=>H3K4me1	1	0
n_200	H3K27me3 and H3K27ac or H3K4me3 and not H3K9Ac=>H3K4me2	2333.948	bin_200	H3K27me3 and H3K27ac or H4K20me1 and H3K4me3=>H3K4me2	62.56479	g18[1]_bin_200	CTCF and H3K27me3 or not H3K4me3 and H3K9Ac=>H3K4me1	1	0
n_200	H3K27me3 and H3K27ac or H4K20me1 and H3K4me3=>H3K4me2	2293.515	bin_200	CTCF and H3K27ac or H3K4me1 and H3K4me3=>H3K4me2	56.62230	g18[1]_bin_200	CTCF and H3K36me3 or H3K27me3 and H3K9Ac=>H3K4me1	1	0
n_200	H3K27me3 and H3K27ac or H3K4me2 and H3K9Ac=>H3K4me3	2111.752	bin_200	H3K27me3 and H3K9Ac or H3K4me1 and H3K4me3=>H3K4me2	54.16963	g18[1]_bin_200	CTCF and H3K36me3 or not H3K4me3 and H3K9Ac=>H3K4me1	1	0
n_200	CTCF and H3K27ac or H3K4me1 and H3K4me3=>H3K4me2	2067.893	bin_200	H4K20me1 and H3K4me3=>H3K4me2	53.52598	g18[1]_bin_200	H3K27me3 and H3K27ac or not H3K4me3 and H3K9Ac=>H3K4me1	1	0
n_200	H3K27me3 and H3K27ac or not H3K4me1 and H3K9Ac=>H3K4me3	1986.951	bin_200	not CTCF and H3K4me3 or not H4K20me1 and H3K9Ac=>H3K4me2	51.67541	g18[1]_bin_200	not H3K4me1 and H3K9Ac=>H3K4me3	1	0
n_200	H3K27me3 and not H3K36me3=>CTCF	0.000	bin_200	not CTCF=>Input_Sonicated	1.00000	g18[1]_bin_200	H3K27me3 and H3K27ac or not H3K4me1 and H3K9Ac=>H3K4me3	1	0
n_200	H3K27me3 and not H4K20me1=>CTCF	0.000	bin_200	not CTCF and not H3K27me3=>Input_Sonicated	1.00000	g18[1]_bin_200	not CTCF=>Input_Sonicated	0	0
n_200	H3K36me3 and not H3K4me1=>CTCF	0.000	bin_200	not H3K27me3=>Input_Sonicated	1.00000	g18[1]_bin_200	CTCF=>Input_Sonicated	0	0
n_200	H3K27me3 and not H3K4me1=>CTCF	0.000	bin_200	not CTCF or not H3K27me3=>Input_Sonicated	1.00000	g18[1]_bin_200	not CTCF and not H3K27me3=>Input_Sonicated	0	0
n_200	H3K27me3 and not H4K20me1 or H3K36me3 and not H3K4me1=>CTCF	0.000	bin_200	CTCF or not H3K27me3=>Input_Sonicated	1.00000	g18[1]_bin_200	H3K27me3=>Input_Sonicated	0	0
n_200	CTCF and not H3K36me3=>H3K27me3	0.000	bin_200	not CTCF or H3K27me3=>Input_Sonicated	1.00000	g18[1]_bin_200	CTCF and not H3K27me3=>Input_Sonicated	0	0
n_200	CTCF and not H4K20me1=>H3K27me3	0.000	bin_200	not CTCF and H3K36me3=>Input_Sonicated	1.00000	g18[1]_bin_200	CTCF and H3K27me3=>Input_Sonicated	0	0
n_200	CTCF and not H3K4me1=>H3K27me3	0.000	bin_200	CTCF and not H3K36me3=>Input_Sonicated	1.00000	g18[1]_bin_200	not H3K27me3=>Input_Sonicated	0	0
n_200	not H3K36me3 and H3K27ac=>H3K27me3	0.000	bin_200	not CTCF and not H3K36me3=>Input_Sonicated	1.00000	g18[1]_bin_200	not CTCF and H3K27me3=>Input_Sonicated	0	0
n_200	not H3K4me3 and H3K27ac=>H3K27me3	0.000	bin_200	not H3K27me3 and not H3K36me3=>Input_Sonicated	1.00000	g18[1]_bin_200	not CTCF or not H3K27me3=>Input_Sonicated	0	0
n_200	not H4K20me1 and H3K27ac=>H3K27me3	0.000	bin_200	not H3K36me3=>Input_Sonicated	1.00000	g18[1]_bin_200	CTCF or not H3K27me3=>Input_Sonicated	0	0

Measures

conviction and *LOE* are stable on different datasets (same targets for top rules). *BF* differs.

Description

Ideal case: we have no counter examples $P(A|\neg B)$, target is fully covered by condition high $P(A|B)$, rule has high support $P(A)$. Proportion $\frac{P(A|B)}{P(A|\neg B)}$ is exactly BF.
In case we have no counter examples and high rule applicability we get *conviction*.
Lets introduce new metrics **Description**!
$Description(A \to B) = \frac{P(A)P(A|B)}{P(A|\neg B)}$.
Problem: in case if we have 1 and 5 counterexamples among 1000+ supporting items, any metrics shouldn't change in 5 fold way.
Introduce corrected number of counter examples as $corrected\_counter\_examples(A \to B) = counter\_examples(A \to B) + 5\% support(A \to B)$.

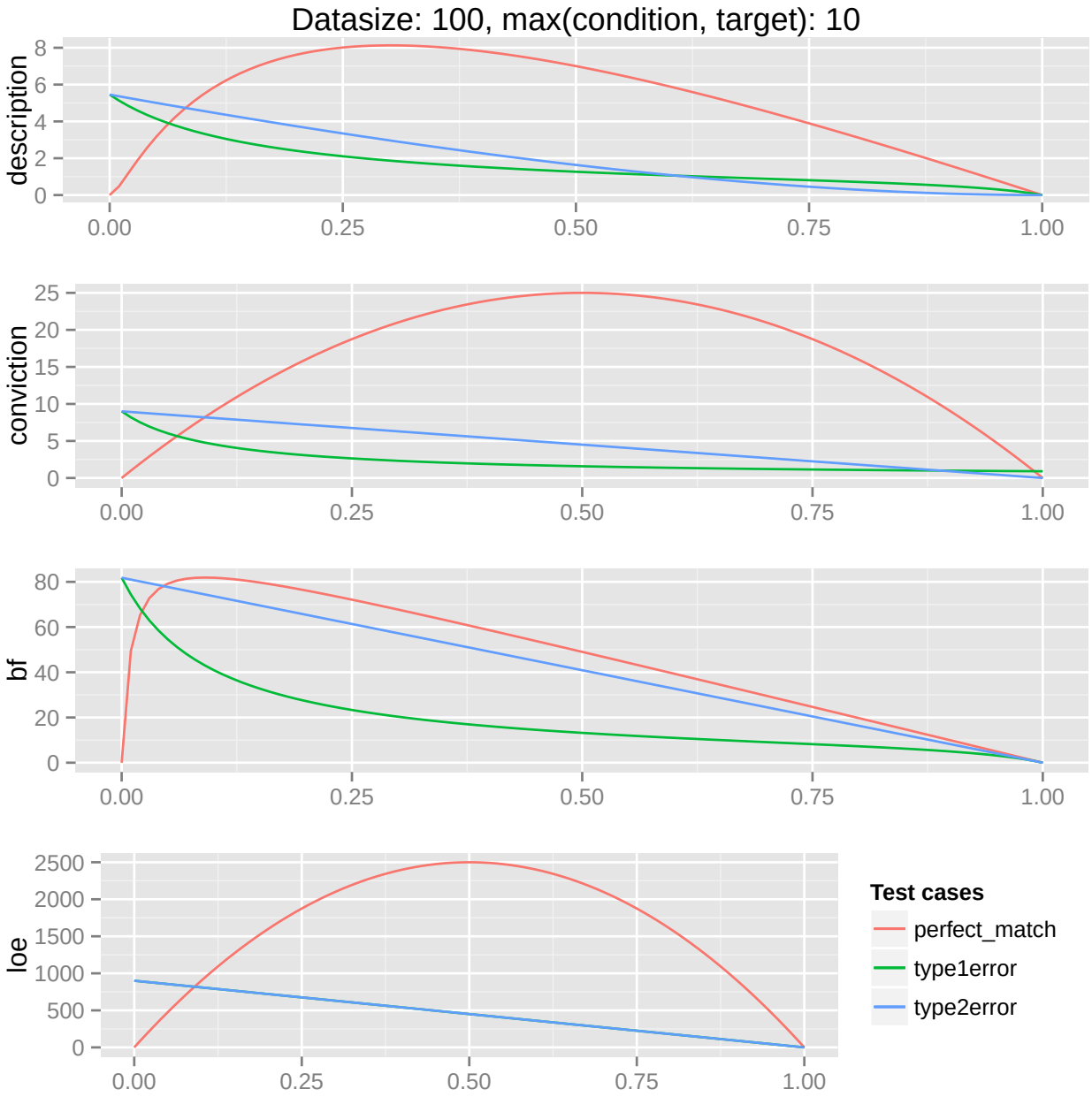
Metrics test

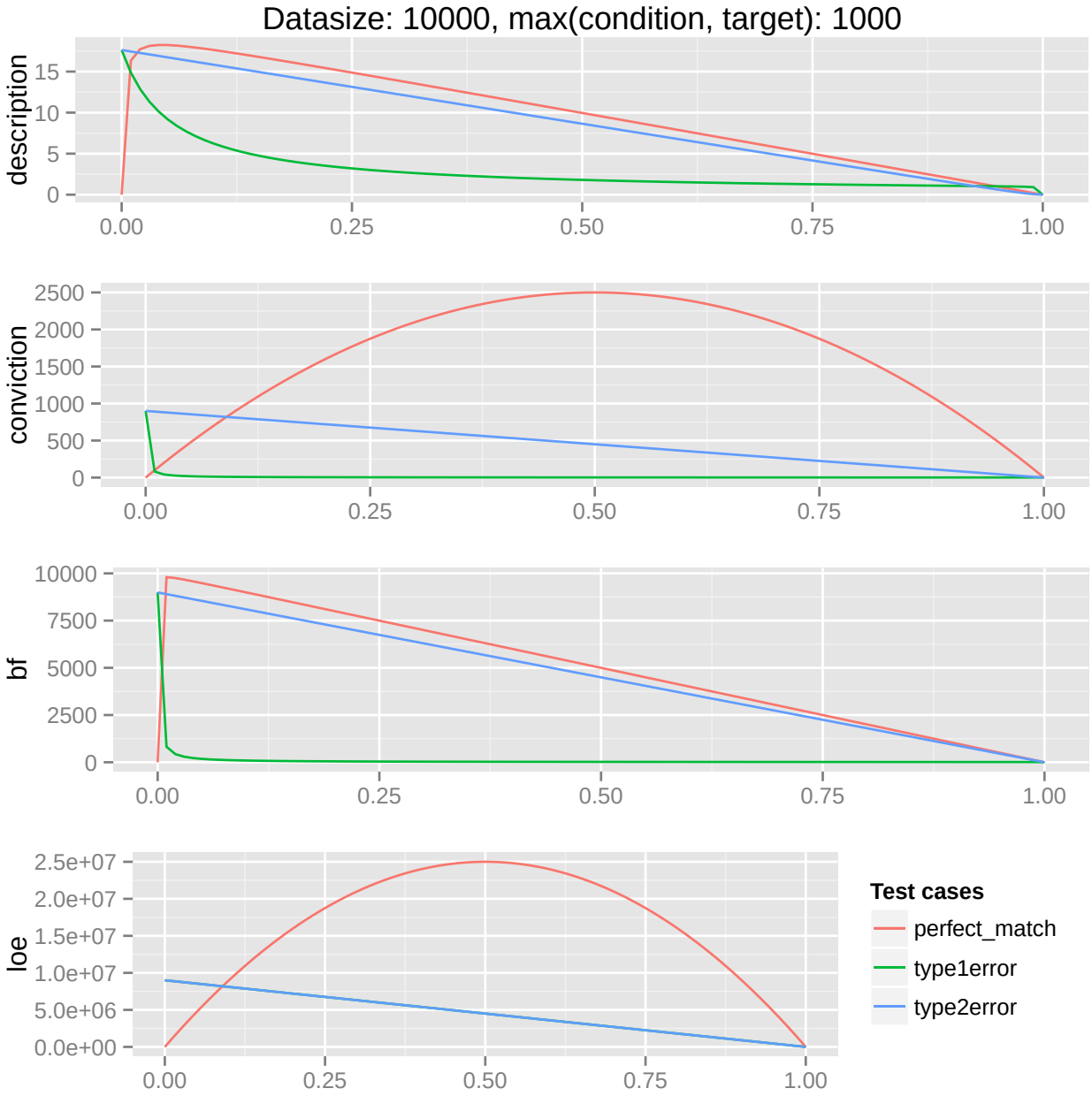
Lets consider the following cases

- Perfect match, i.e. $condition = target = both$
- Type 1 error, i.e. $condition = TRUE$, $target = FALSE$
- Type 2 error, i.e. $condition = FALSE$, $target = TRUE$

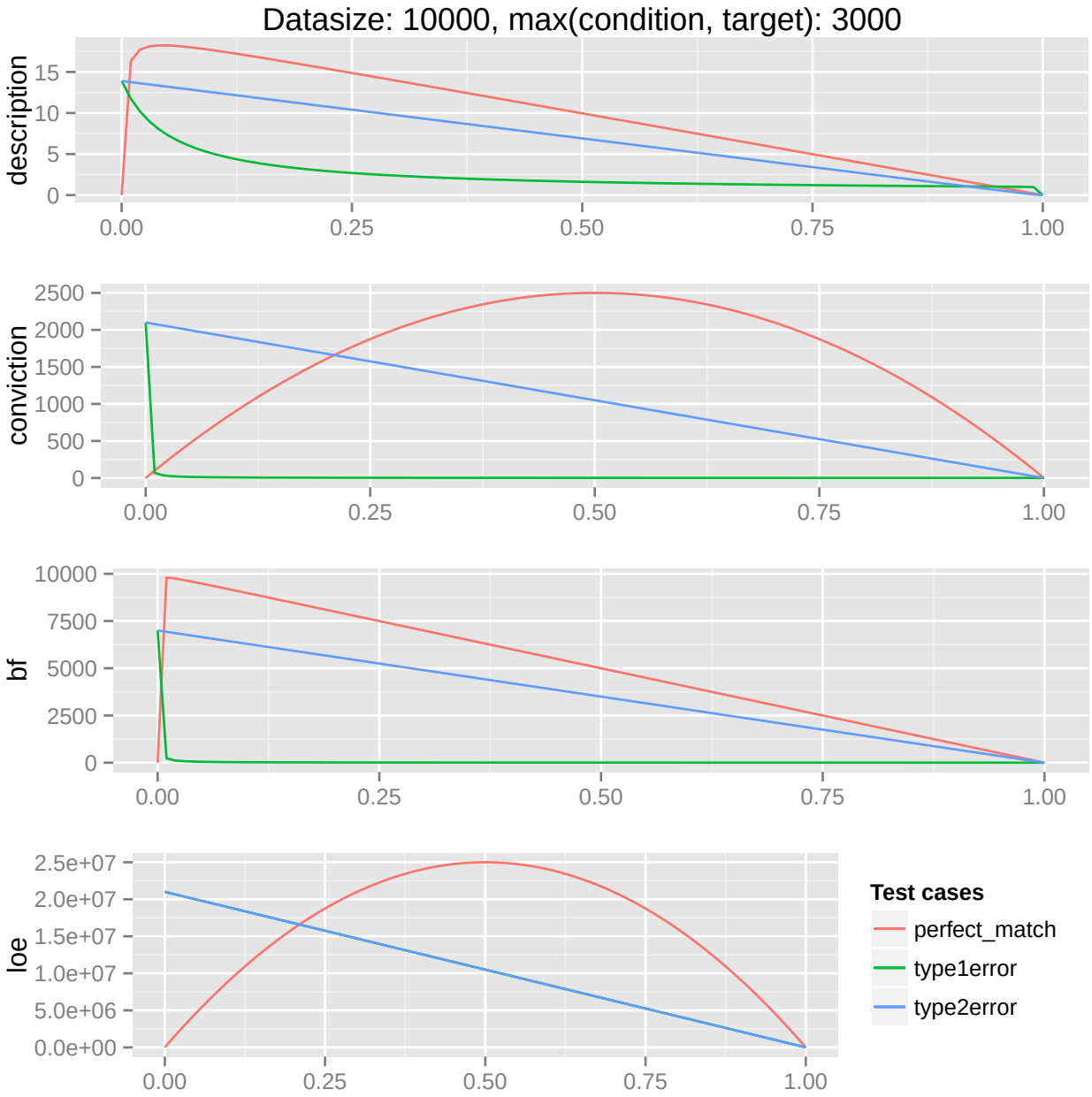
Lets analyze dependence of *Conviction*, *BF*, *LOE* and *Description* on different error types.

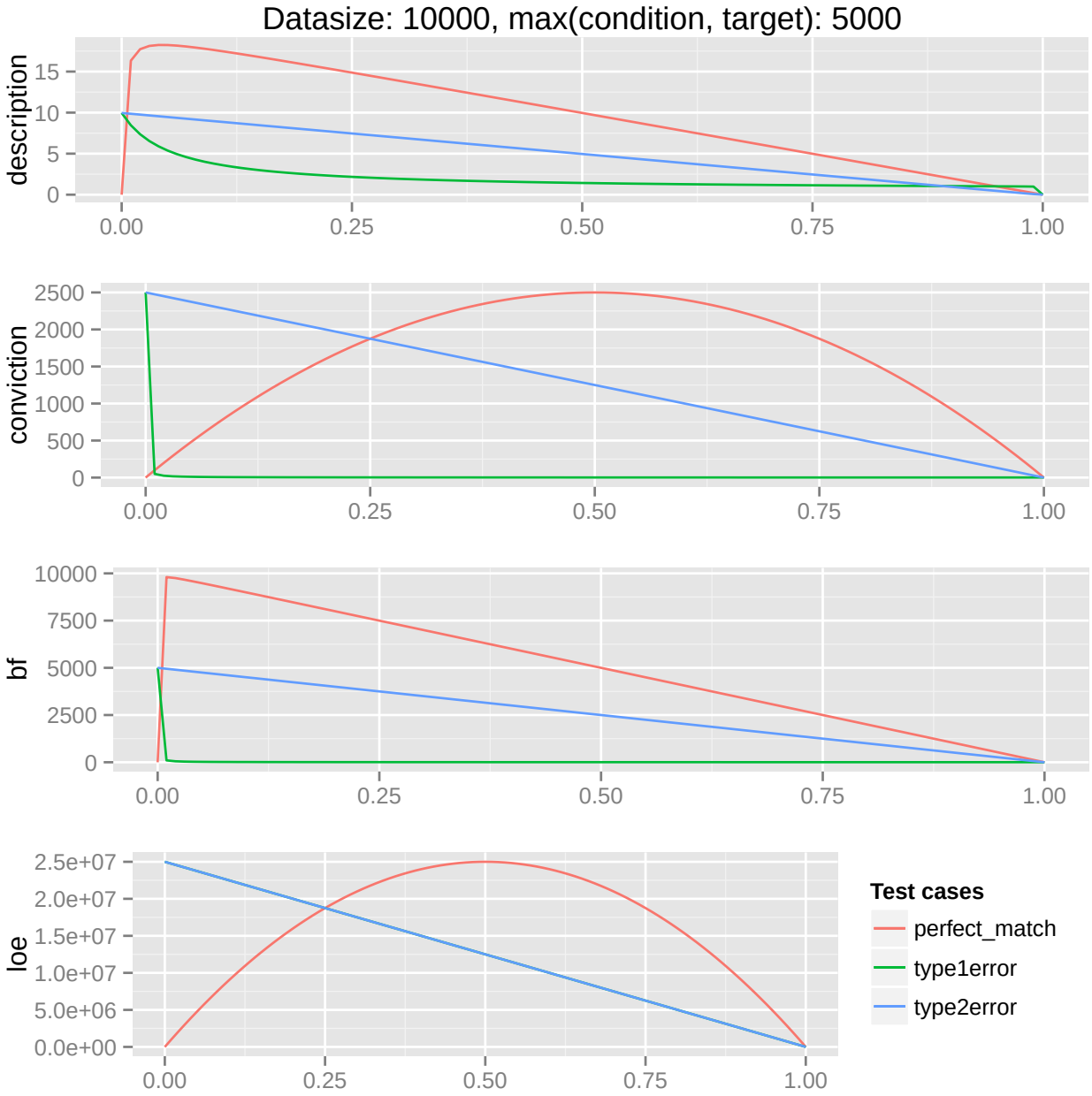
Small database

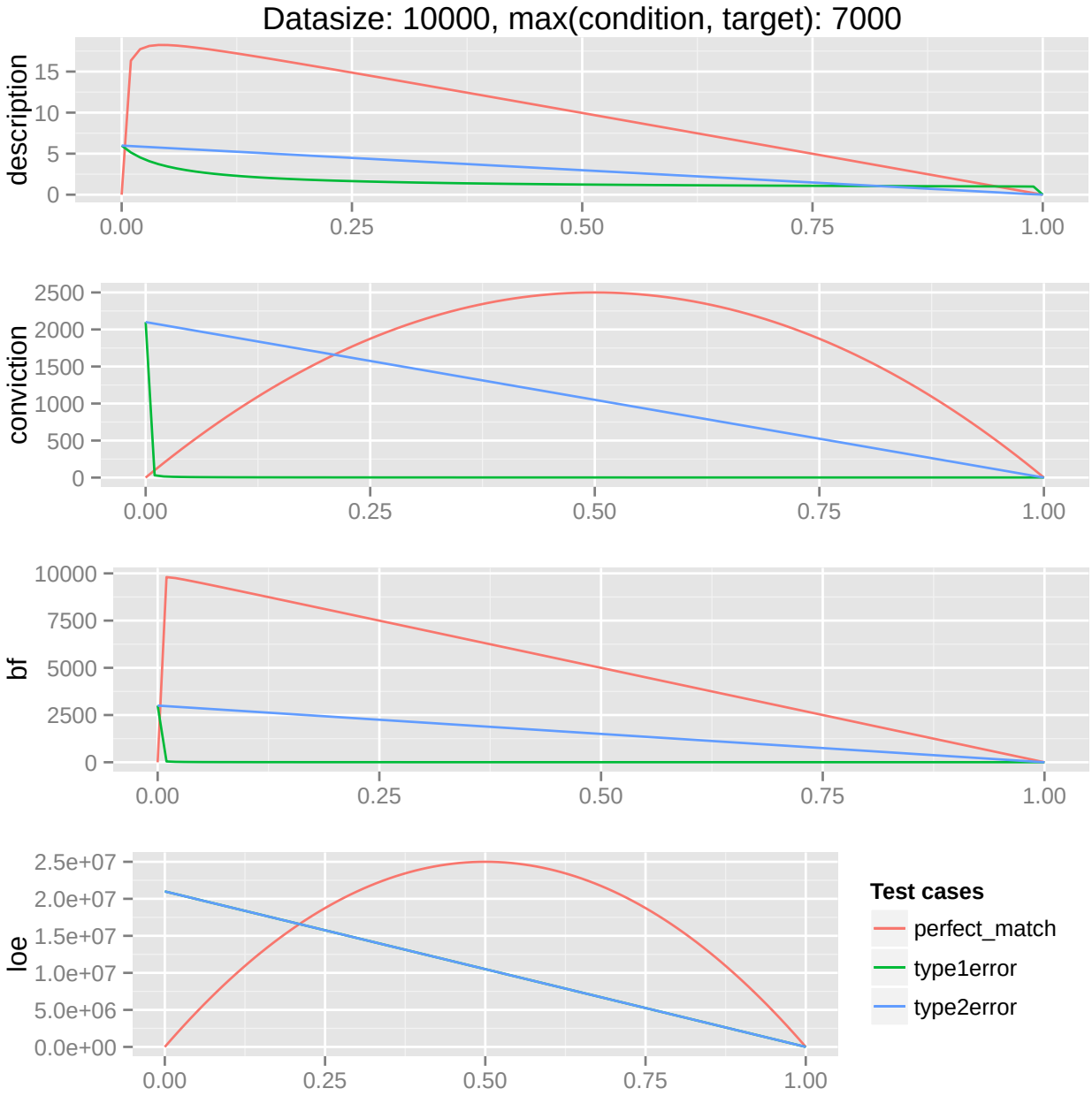



Datasize: 10000, max(condition, target): 1000
description
conviction
bf
loe
0.00
0.25
0.50
0.75
1.00
Test cases
perfect_match
type1error
type2error

Experiments database




Datasize: 10000, max(condition, target): 5000
description
conviction
bf
loe
0.00
0.25
0.50
0.75
1.00
Test cases
perfect_match
type1error
type2error


Datasize: 10000, max(condition, target): 7000
description
conviction
bf
loe
0.00
0.25
0.50
0.75
1.00
Test cases
perfect_match
type1error
type2error

Metrics analysis

- *Description*[4] doesn't depend on size.
- All of them scores 100% perfect match rules as uninteresting.
- All of them have linear dependency on type2 errors.
- *Description* tolerates small number of type1 error.
- $Max(description)$ is reached at $\approx$ 10-30%.

[4] *Description* is the only one metrics to pass all the greedy optimization tests

Branch and bounds

- Ignore targets predicate in condition
- Ignore atomic predicates with empty or full support
- AND: consider we have formula p and atomic predicate a
  Ignore if
  1. $sup(p \wedge a) = sup(p)$
  2. $sup(p \wedge a) = sup(a)$
  3. $sup(p \wedge a) = \emptyset$
- OR: consider we have formula p and atomic predicate a
  Ignore if
  1. $sup(p \vee a) = sup(p)$
  2. $sup(p \vee a) = sup(a)$
  3. $sup(p \vee a) = D$
- TODO: at the moment we ignore atomic predicate injection inside OR formulas
- TODO: take into account *measure_function*

Greedy iteractive rule search

Problem: find best rule in terms of function f.

- target B and atomic predicates *Atomics*
- set current best condition NULL
- while we can improve
- take current best condition A and for each atomic a not in *Atomics*
- try to "inject" a and $\neg a$ into A
- choose current best condition from rule with max f

function Inject(A, a)

- if A is NULL then a and $\neg a$
- else for each subformula: remove, replace, apply or, apply and with a or $\neg a$

Convergence

Problem:
We can get micro improvements while modifying rule (add or remove), which leads to enourmous predicates size.
Solution:
Use fine for predicates complexity. This prevents us from small fluctuations.
Fine function $fine(rule) = 5\%size(condition)^2 + 5\%size(target)^2$ works fine and passes all the optimization tests.

Greedy search example

Database $= [0; 1000)$, Predicates $= \{[i * 100; (i + 1) * 100), i \in 0...9\}$
Condition

- $[200; 300) \vee [300; 400) \vee [400; 500) \rightarrow [200; 500)$
- $[200; 300) \vee [300; 400) \vee [400; 500) \rightarrow [190; 510)$
- $[200; 300) \vee [300; 400) \vee [400; 500) \rightarrow [210; 490)$
- $[300; 400) \rightarrow [215; 485)$ too many counter-examples.

Target

- $[200; 500) \rightarrow [200; 300) \vee [300; 400) \vee [400; 500)$
- $[190; 510) \rightarrow [200; 300) \vee [300; 400) \vee [400; 500)$
- $[210; 490) \rightarrow [200; 300) \vee [300; 400) \vee [400; 500)$
- $[185; 515) \rightarrow [100; 200) \vee [200; 300) \vee [300; 400) \vee [400; 500) \vee [500; 600)$ too many counter-examples.

Greedy iteractive condition search

Problem: find best rule in terms of function f.

- target B and atomic predicates *Atomics*
- set current best condition $C = \text{NULL}$
- set current target $T = B$
- find best condition $C1$ for current target
- set current target $T1 = T \wedge \neg C1$
- find best condition $C2$ for $T1$
- check if rule $C1 \vee C2 \rightarrow B$ has higher score f
- etc

Greedy search condition example

Database $= [0; 1000)$, Predicates $= \{[i * 100; (i + 1) * 100), i \in 0...9\}$

Greedy search

- $[200; 300) \vee [300; 400) \rightarrow [200; 450)$

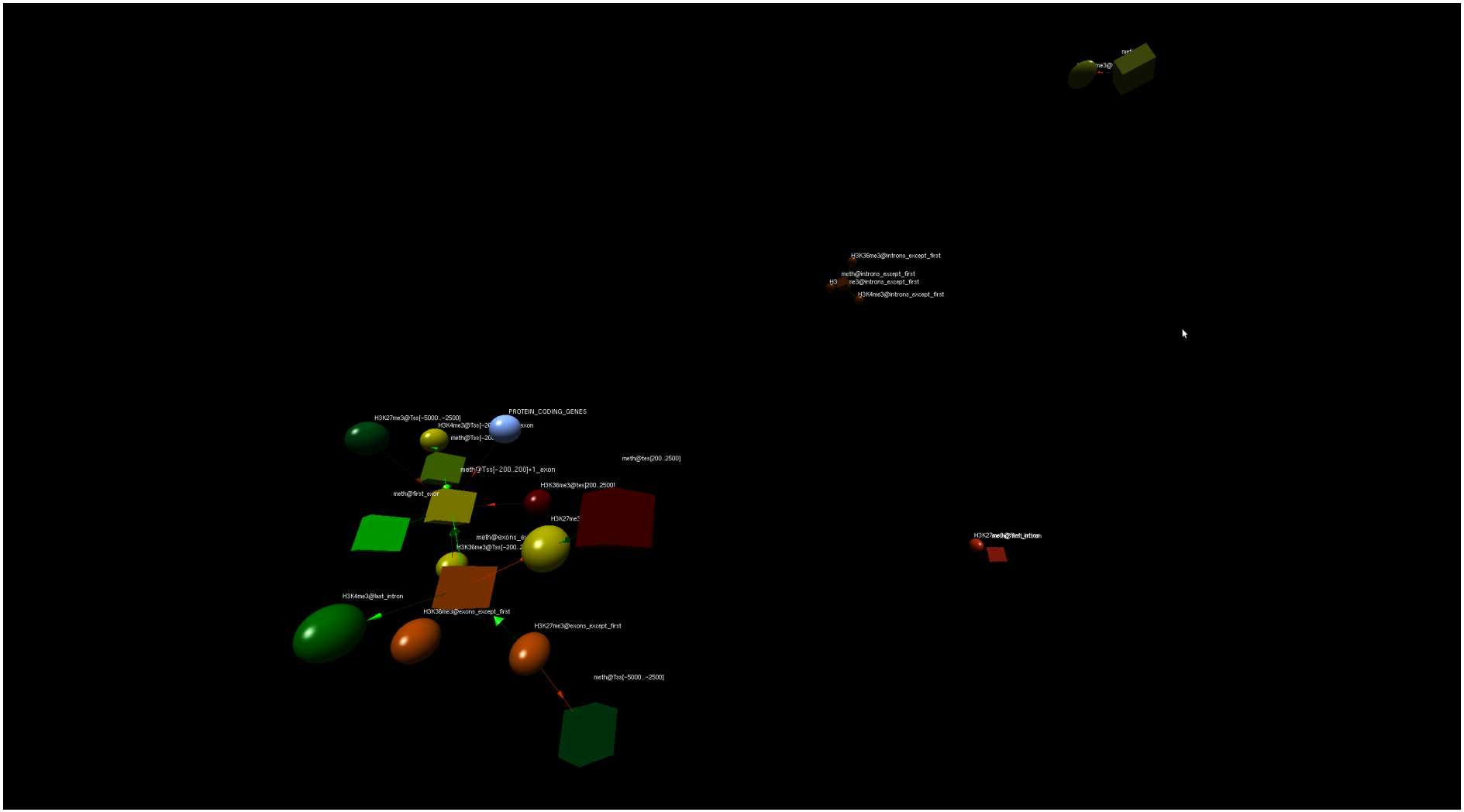
Condition optimization

- $[200; 300) \vee [300; 400) \rightarrow [200; 450)$
- Set $[200; 450) \wedge not[200; 300) \wedge not[300; 400)$ as new target.
- Optimization:
  $[400; 500) \wedge not[450; 1000) \rightarrow [200; 450) \wedge not[200; 300) \wedge not[300; 400)$
- New rule with better score
  $[200; 300) \vee [300; 400) \vee [400; 500) \wedge not[450; 1000) \rightarrow [200; 450)$

Massive rules visualization

DataSet GSE47819.
To examine aging-related alterations, we profiled transcriptome, DNA methylome and the principal regulatory chromatin marks in HSCs, specifically H3K4me3, H3K27me3 and H3K36me3 marks in young (4mo) and old (24mo) murine hematopoietic stem cells(HSCs).

GSE47189 meth influence



Rule tracing/database optimization

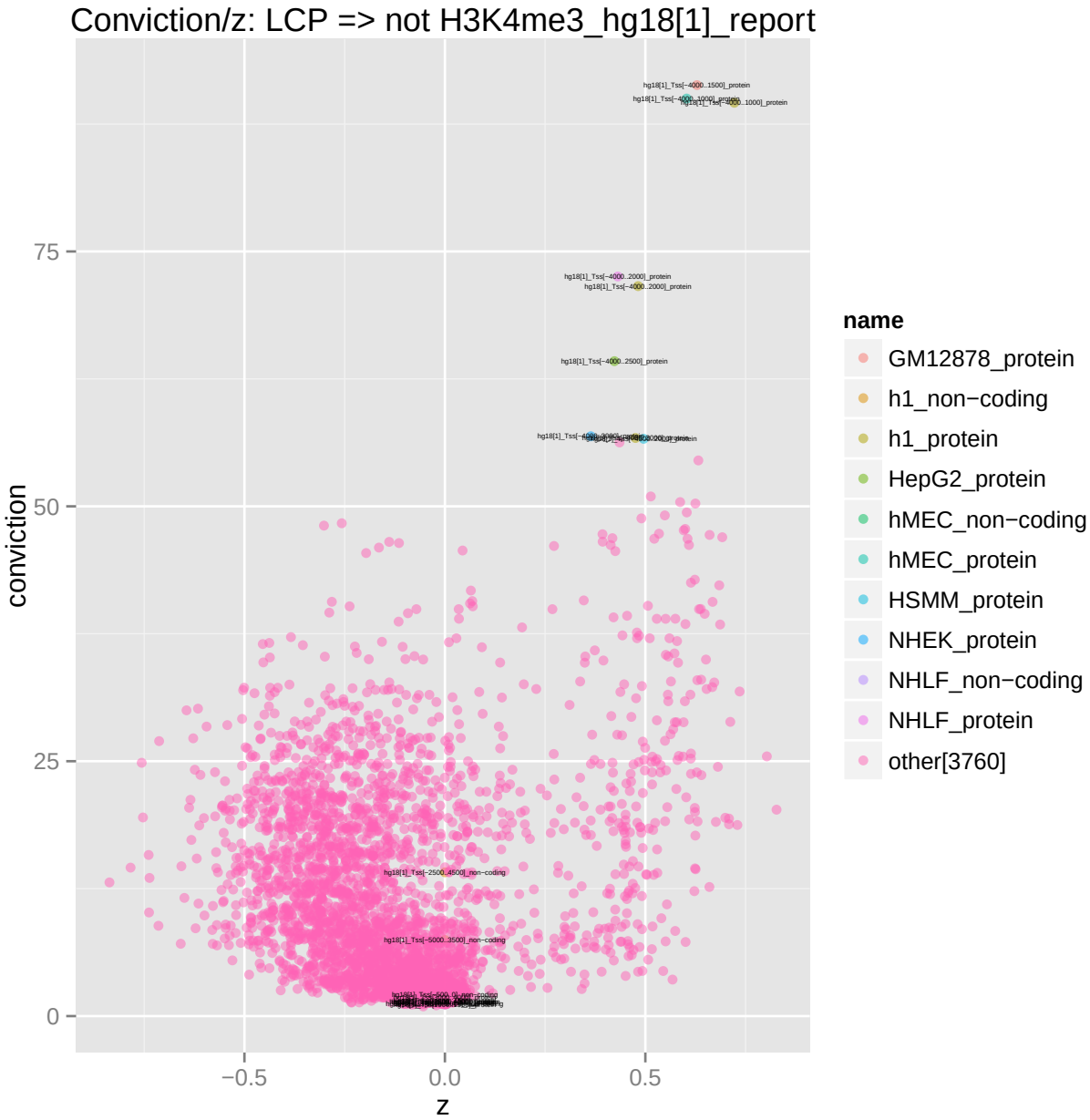
Tracing

- Fix rule of interest
- Optimize metricss on different databases

RuleTracingExperiment

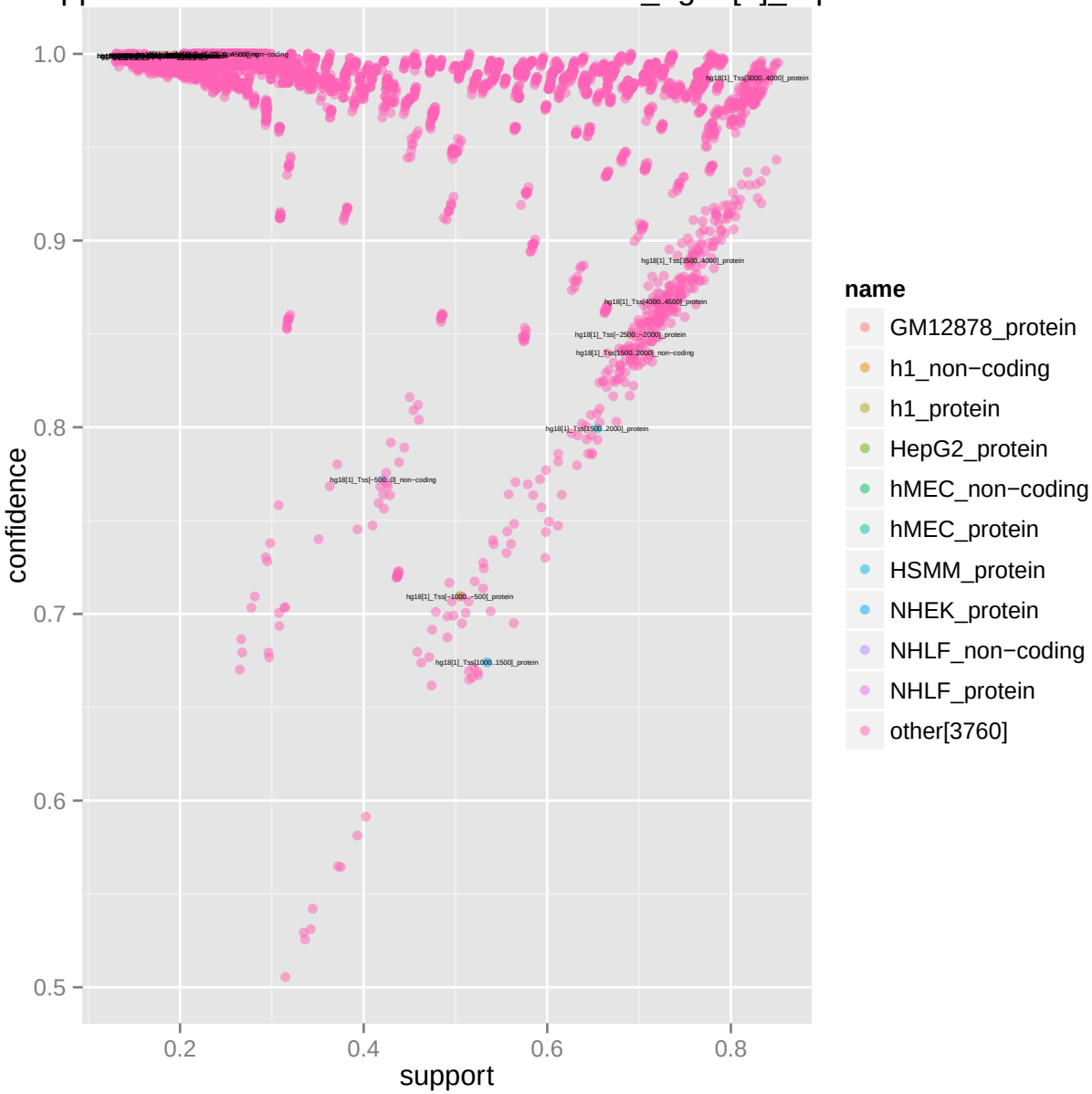
- LCP $\rightarrow$ not H3K4me3
- GSE26320 cells
- protein coding, non protein c oding genes
- TSS[$-l$, r], l, r in range [-5000, 5000], step 500
- Plot conviction(color), z(size) for l, r for each cell line and genes class

RuleTracingExperiment LCP → not H3K4me3

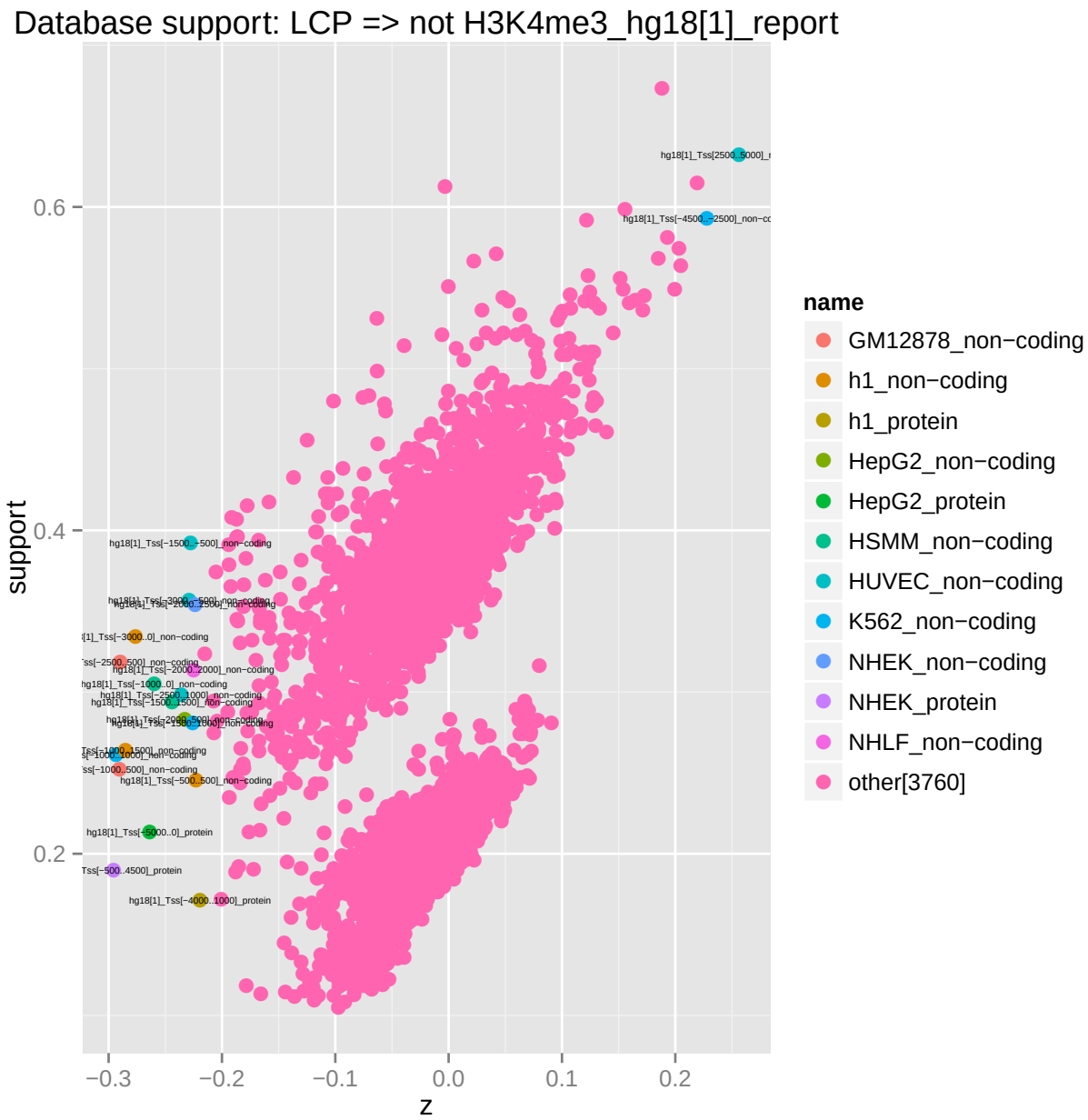


database	name	conviction	z
hg18[1]_Tss[−4000..1500]_protein	GM12878_protein	91.315129	6.287098e−01
hg18[1]_Tss[−4000..1000]_protein	hMEC_protein	89.997232	6.029991e−01
hg18[1]_Tss[−4000..1000]_protein	h1_protein	89.587085	7.219069e−01
hg18[1]_Tss[−4000..2000]_protein	NHLF_protein	72.546125	4.312693e−01
hg18[1]_Tss[−4000..2000]_protein	h1_protein	71.601476	4.819167e−01
hg18[1]_Tss[−4000..2500]_protein	HepG2_protein	64.238007	4.228085e−01
hg18[1]_Tss[−4000..3000]_protein	NHEK_protein	56.873801	3.639819e−01
hg18[1]_Tss[−4000..3000]_protein	h1_protein	56.705535	4.753776e−01
hg18[1]_Tss[−4500..2000]_protein	HSMM_protein	56.623616	4.957172e−01
hg18[1]_Tss[4000..4500]_protein	HepG2_protein	1.490480	−6.106247e−06
hg18[1]_Tss[−2500..−2000]_protein	GM12878_protein	1.450231	−8.043400e−06
hg18[1]_Tss[3000..4000]_protein	h1_protein	1.893227	−4.881634e−05
hg18[1]_Tss[3500..4000]_protein	HepG2_protein	1.449352	−6.609164e−05
hg18[1]_Tss[−1000..−500]_protein	h1_protein	1.394929	7.582398e−05
hg18[1]_Tss[1000..1500]_protein	NHEK_protein	1.159202	1.014114e−04
hg18[1]_Tss[−2500..4500]_non−coding	h1_non−coding	14.104046	1.045764e−04
hg18[1]_Tss[1500..2000]_protein	HSMM_protein	1.311655	−1.282420e−04
hg18[1]_Tss[−500..0]_non−coding	NHLF_non−coding	2.117961	1.308101e−04
hg18[1]_Tss[1500..2000]_non−coding	hMEC_non−coding	1.250506	−1.334606e−04
hg18[1]_Tss[−5000..3500]_non−coding	h1_non−coding	7.468208	−1.360709e−04

Support/confidence: LCP => not H3K4me3_hg18[1]_report

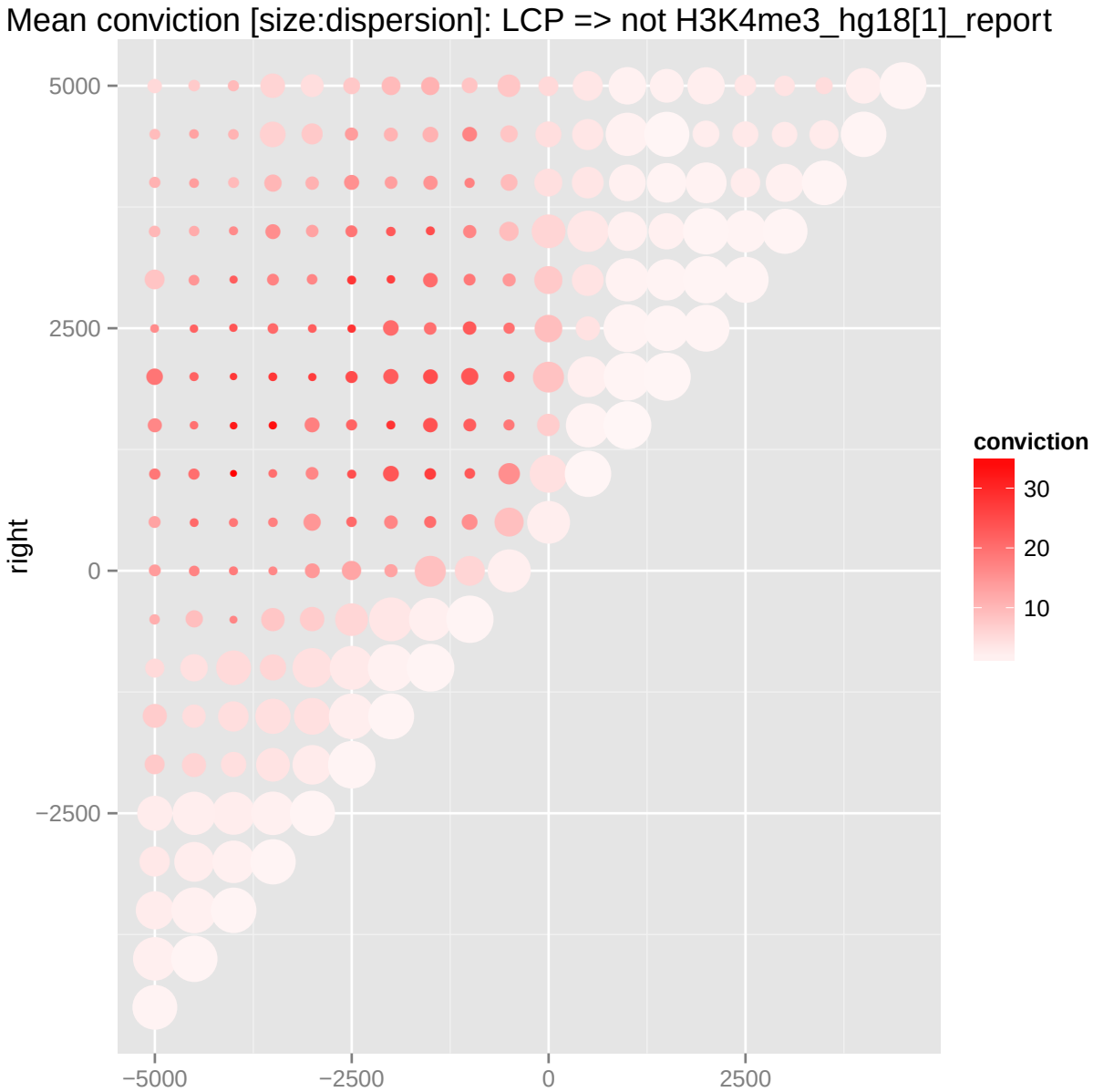


We see that NO target can be considered as Database rule.



TOP parameters

Mean conviction: color, sd: size. Best value: big red circle.



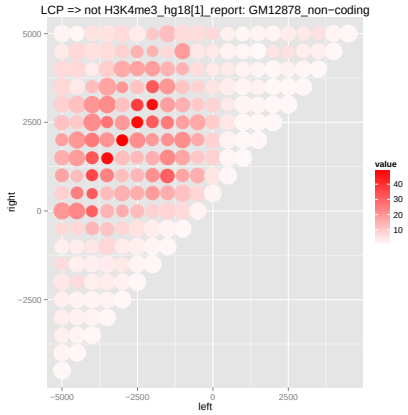
TOP parameters

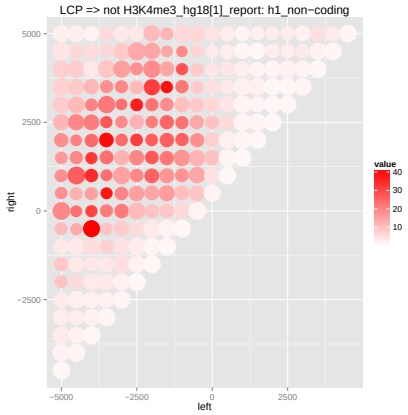
conviction	sd	left	right
35.552055604027	21.8114522059561	−4000	1000
33.468544656562	12.5514195167947	−3500	1500
32.0481032880794	15.7010930684938	−4000	1500
28.3141783257703	11.6815402771032	−2500	2500
28.2371957131326	10.0880371448255	−2000	1500
27.9535980808129	9.94518598066651	−2500	3000
27.8637237251192	17.0190892396959	−4000	2000
27.8593157458446	10.6632319710853	−3500	2000
27.174727972523	12.5315517631142	−3000	2000
26.8545813299393	5.75217268580432	−1500	1000
26.7836472429554	11.3579210856414	−2000	3000
25.0160251974383	5.1951127978949	−2500	2000
24.7866185692861	9.9066677194095	−1500	3500
24.7362520703767	3.53441822422071	−1500	2000
24.5160214404823	9.67291868981676	−2500	1000
24.0369931421477	3.63566946628666	−1500	1500
23.6445545416502	2.66293043713036	−1000	2000
23.5675552465854	12.1998091657517	−4000	2500
23.2911651807902	3.10321385254426	−2000	1000
23.2425317918083	6.88882148901572	−1000	1000

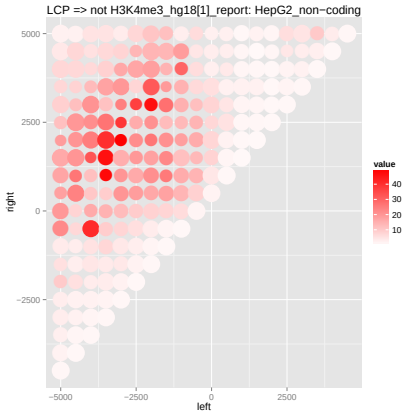
RuleTracingExperiment LCP → not H3K4me3

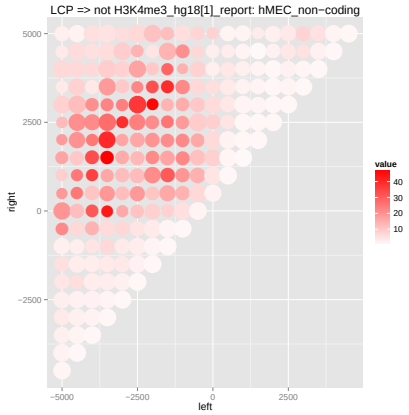
Conviction:color, Z:size - the bigger the smallest $abs(z)$

Non-coding

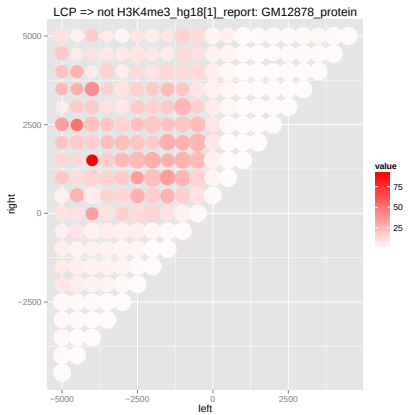


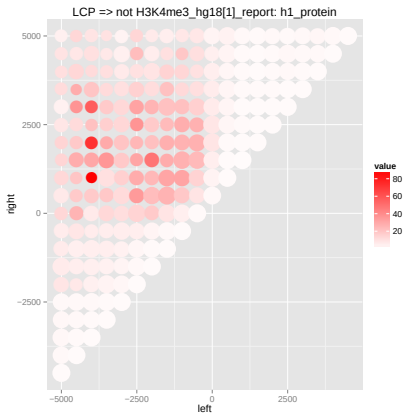


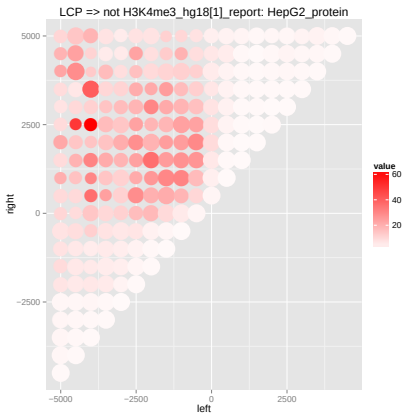


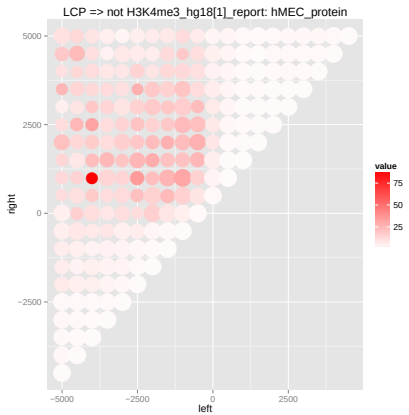


Coding









etc...

Hypothesis testing roadmap

- Select list of atomic Predicates to test.
- Select appropriate list of databases for predicates.
- Process rule mining (all vs all now, more complex in future).
- Select interesting rules(by top conviction/z, support/confidence).
  Select *TOP_RULES* with **max(conviction)** and **min(abs(z))**,
  higlight them on the plot, merge others.
- Select database rules by **max(support)** and **max(abs(z))** vs genomic.
- If target has **support** $\approx$ **1** and **high abs(z)**, ignore it.
- Optimize datatabase parameters for interesting rules.
- Process rule mining of optimized database, etc.

References

- Guillet Hamilton - Quality Measures in Data Mining (2007)
- Comparing Rule Measures for Predictive Association Rules `http://www.di.uminho.pt/~pja/ps/conviction.pdf`
- Measures overview `http://michael.hahsler.net/research/association_rules/measures.html`
- Web Data mining - Bing Liu `http://link.springer.com/book/10.1007/978-3-540-37882-2`