Rule mining

Oleg Shpynov

JetBrains Biolabs

November 19, 2014

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 $X \to Y$.
  $X \subseteq T$, $Y \subseteq T$, $X \cap Y = \emptyset$
- $supp(X) = \frac{\#\{d_i | X \subseteq d_i\}}{\#\{D\}} = P(X)$
- $conf(X \to Y) = \frac{supp(X \to Y)}{supp(X)} = \frac{P(X \wedge Y)}{P(X)} = P(Y|X)$
- $lift(X \to Y) = lift(Y \to X) = \frac{conf(X \to Y)}{supp(Y)} = \frac{P(X \wedge Y)}{P(X)P(Y)}$
- $conviction(X \to Y) = \frac{1 - supp(Y)}{1 - conf(X \to Y)} = \frac{P(X)P(\neg Y)}{P(X \wedge \neg Y)}$

Measurements

Measure	Definition	Range
confidence	$conf(A \to C) = \frac{sup(A \cup C)}{sup(A)}$	$[0, 1]$
Laplace	$lapl(A \to C) = \frac{sup(A \cup C)+1}{sup(A)+2}$	$[0, 1[$
lift	$lift(A \to C) = \frac{conf(A \to C)}{sup(C)}$	$[0, +\infty[$
conviction	$conv(A \to C) = \frac{1-sup(C)}{1-conf(A \to C)}$	$[0.5, +\infty[$
leverage	$leve(A \to C) = sup(A \cup C) - sup(A) \times sup(C)$	$[-0.25, 0.25]$
χ^2	$\chi^2(A \to C) = N \times \sum_{X \in \{A, \neg A\}, Y \in \{C, \neg C\}} \frac{(sup(X \cup Y) - sup(X).sup(Y))^2}{sup(X) \times sup(Y)}$	$[0, +\infty[$
Jaccard	$jacc(A \to C) = \frac{sup(A \cup C)}{sup(A)+sup(C)-sup(A \cup C)}$	$[0, 1]$
cosine	$cos(A \to C) = \frac{sup(A \cup C)}{\sqrt{sup(A) \times sup(C)}}$	$[0, 1]$
ϕ-coeff	$\phi(A \to C) = \frac{leve(A \to C)}{\sqrt{(sup(A) \times sup(C)) \times (1-sup(A)) \times (1-sup(C))}}$	$[-1, 1]$
mutual inf.	$MI(A \to C) = \frac{\sum_i \sum_j sup(A_i \cup C_j) \times log(\frac{sup(A_i \cup C_j)}{sup(A_i) \times sup(C_j)})}{min(\sum_i -sup(A_i) \times log(sup(A_i)), \sum_j -sup(C_j) \times log(sup(C_j)))}$	$[0, 1]$

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 => Histone enrichment
- Histone enrichment => Histone enrichment
- ChromHMM State => Histone enrichment
- Histone enrichment => ChromHMM State

Hypothesis $Condition => Target$ on DataBase D.
DATABASE RULE

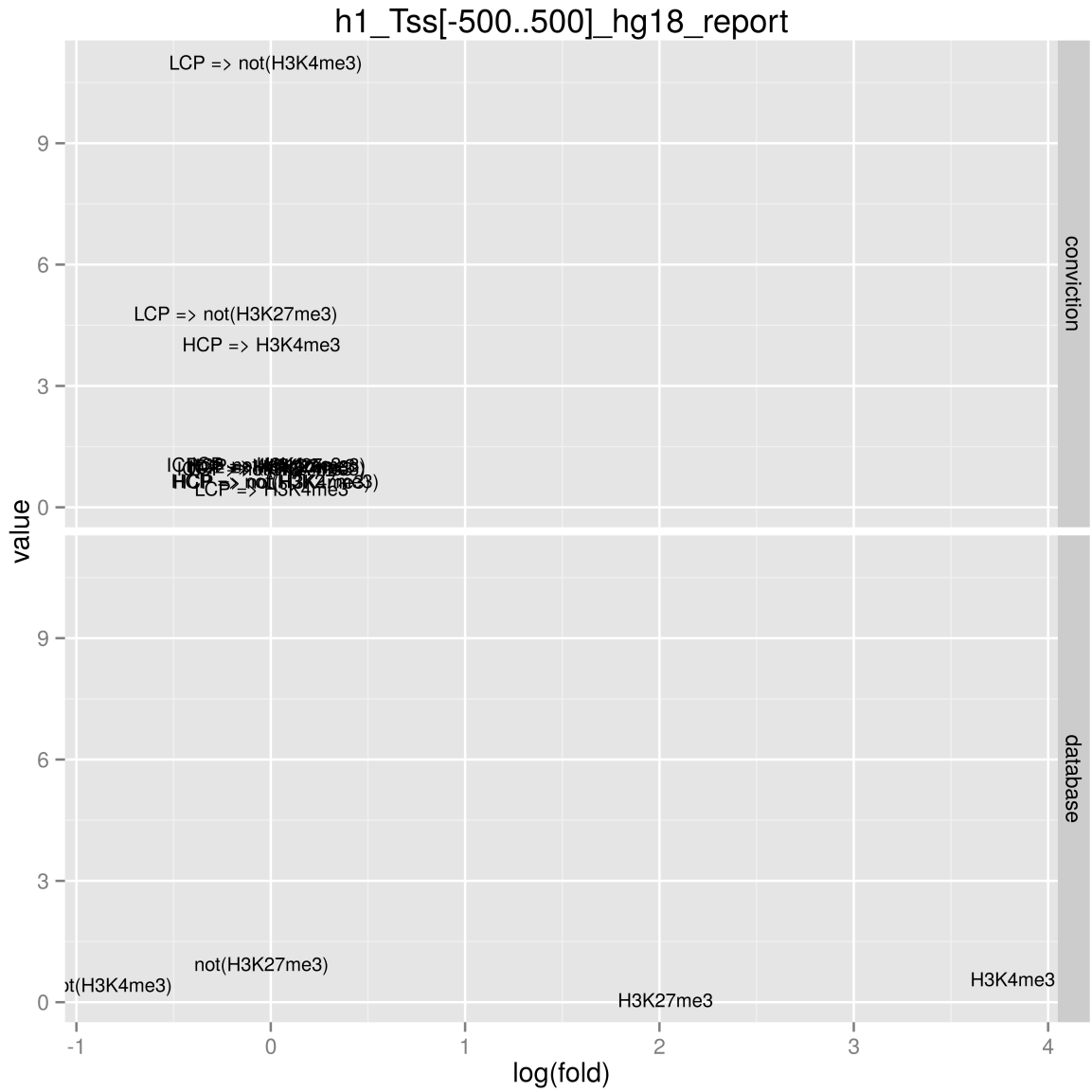
- What if $Target$ is property of D itself?
- Compute fold change $confidence(TRUE => Target)$ on D vs $D.sampleGenomic()$.
- If $foldChange > threshold$ then $Target$ is a D property!

Hypothesis $Condition => Target$ on DataBase D.
RULE

- Measure $conviction(Condition => Target)$ on D.
- What if we get good value by chance?
- We can sample and get confidence intervals for *conviction* - check how stable it is.
- Check foldChange on D vs $D.sample$
- If $foldChange < threshold$ then we found a RULE!

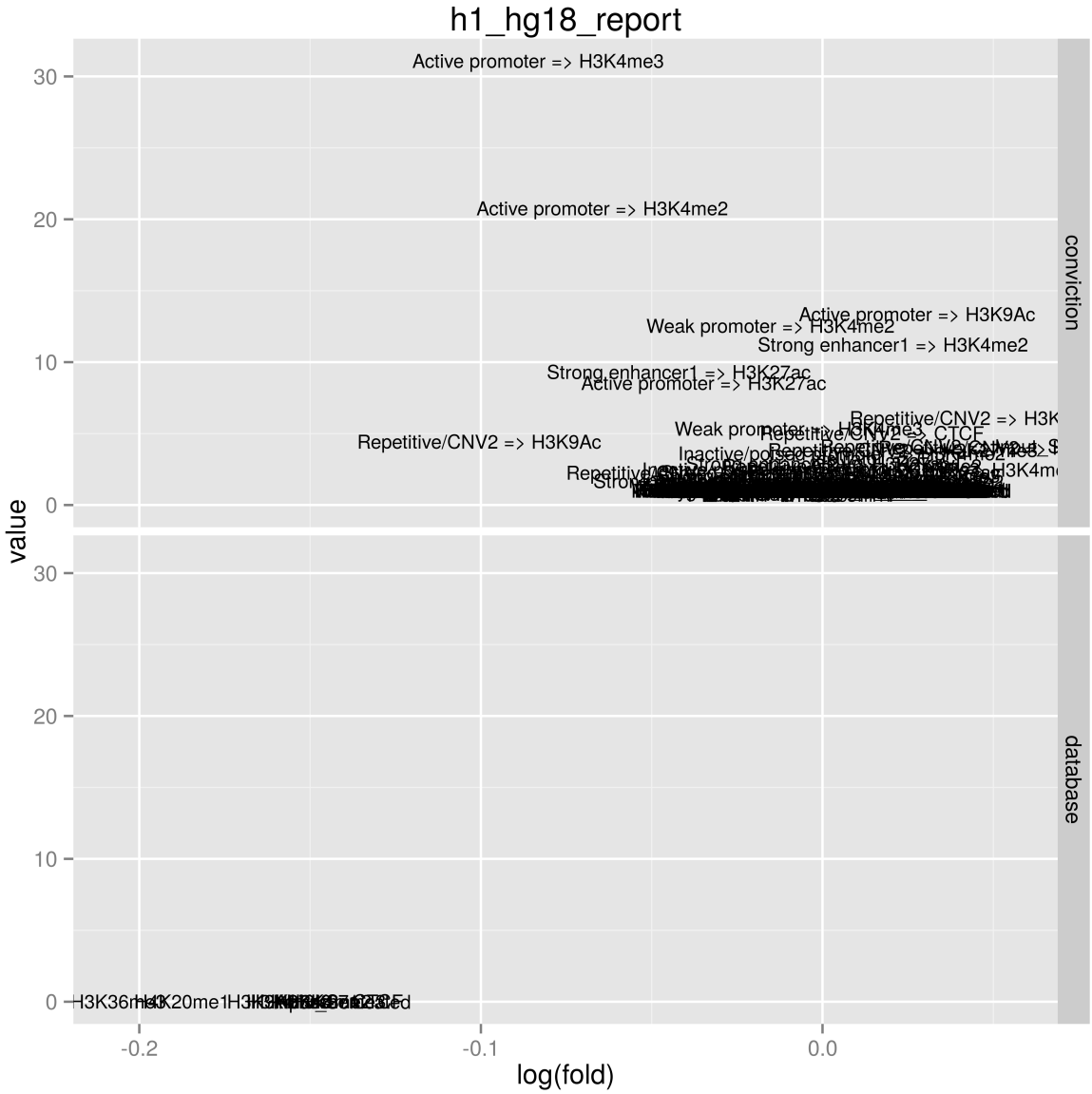
BivalentHCPExperiment

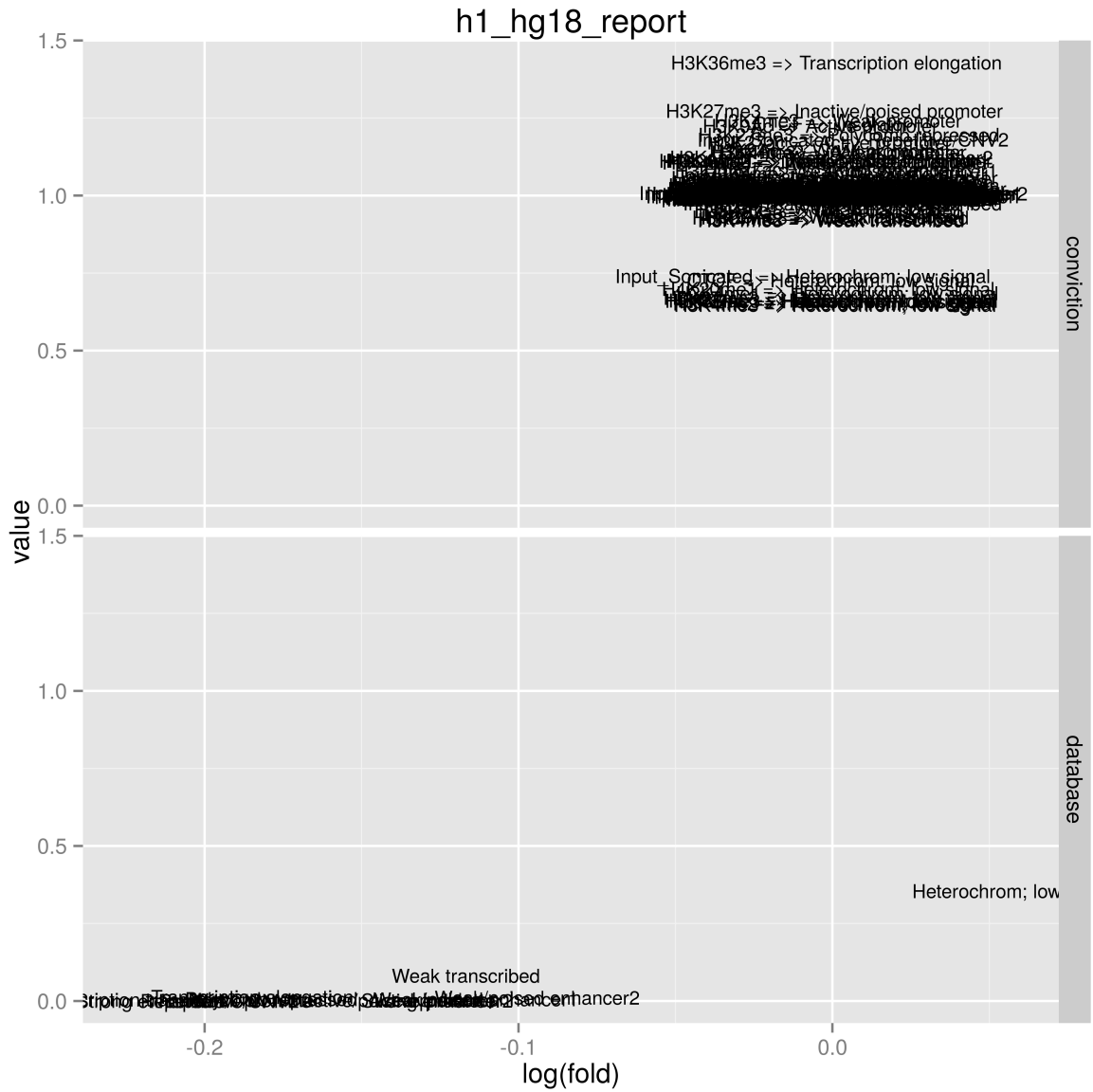
- FOUND DATASET RULE: Tss[-500..500]_protein: H3K4me3 is 43.49834364520371 more often vs genomic [0.569591251275534 8]
- FOUND RULE: $HCP => H3K4me3$
  DATA: Tss[-500..500]_protein [51939]
  CONDITION: HCP 0.4528581605344731 4 [23521]
  TARGET: $H3K4me3$ 0.569591251275534 8 [29584]
  CONFIDENCE: $H3K4me3|HCP$: 0.893116789252157 7 [21007]
  LIFT: 1.567995974748776 8
  CONVICTION: 4.026906992342143
- FOUND RULE: $LCP => not(H3K4me3)$
  DATA: Tss[-5000..-2500]_protein [51939]
  CONDITION: LCP 0.613931727603534 9 [31887]
  TARGET: $not(H3K4me3)$ 0.970292073393788 9 [50396]
  CONFIDENCE: $not(H3K4me3)|LCP$: 0.996173989400069 [31765]
  LIFT: 1.026674355810980 8
  CONVICTION: 7.764726686002016 5
- ...


h1_Tss[-500..500]_hg18_report
LCP => not(H3K4me3)
LCP => not(H3K27me3)
HCP => H3K4me3
LCP => H3K4me3
conviction
database
value
log(fold)
not(H3K27me3)
H3K27me3
H3K4me3
0
3
6
9
-1
0
1
2
3
4

ChromHMMChipSeqExperiment

- $ChipSeq => ChipSeq$
  $H3K4me3 => H3K4me2$
- $ChromHMM => ChipSeq$
  H3K4me2 as result for many rules with high conviction
- $ChipSeq => ChromHMM$
  No rules found.


h1_hg18_report
Active promoter => H3K4me3
Active promoter => H3K4me2
Active promoter => H3K9Ac
Weak promoter => H3K4me2
Strong enhancer1 => H3K4me2
Strong enhancer1 => H3K27ac
Active promoter => H3K27ac
Repetitive/CNV2 => H3K
Repetitive/CNV2 => H3K9Ac
conviction
database
value
30
20
10
0
-0.2
-0.1
0.0
log(fold)


h1_hg18_report
H3K36me3 => Transcription elongation
H3K27me3 => Inactive/poised promoter
Input_Saturated => Heterochrom; low signal
conviction
database
value
Heterochrom; low
Weak transcribed
1.5
1.0
0.5
0.0
-0.2
-0.1
0.0
log(fold)

References

- 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`