Glove in PubMed text

张亚亮

2018/4/27

预处理代码

- 处理从Pubmed得到的数据
- 提取摘要
- 数据预处理

pubmed-medline (not abstract)

```
PMC - PMC4290828
PMID- 25633159
IS  - 1465-6906 (Print)
IS  - 1465-6914 (Electronic)
VI  - 15
IP  - 12
DP  - 2014
TI  - An investigation of biomarkers derived from legacy microarray data for
their
      utility in the RNA-seq era.
LID - 3273
AB  - Background: Gene expression microarray has been the primary biomarker
platform
      ubiquitously applied in biomedical research, resulting in enormous data,
      predictive models, and biomarkers accrued. Recently, RNA-seq has
      looked likely to
      replace microarrays, but there will be a period where both
      technologies co-exist.
      This raises two important questions: Can microarray-based models and
      biomarkers
      be directly applied to RNA-seq data? Can future RNA-seq-based
      predictive models
      and biomarkers be applied to microarray data to leverage past
      investment?
      Results: We systematically evaluated the transferability of predictive
      models and
      signature genes between microarray and RNA-seq using two large
      clinical data
      sets. The complexity of cross-platform sequence correspondence was
      considered in
      the analysis and examined using three human and two rat data sets, and
      three
      levels of mapping complexity were revealed. Three algorithms
      representing
      different modeling complexity were applied to the three levels of
      mappings for
      each of the eight binary endpoints and Cox regression was used to
      model survival
      times with expression data. In total, 240,096 predictive models were
      examined.
```

abstract

1. Nat Commun. 2018 Apr 11;9(1):1402. doi: 10.1038/s41467-018-03635-9.

PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition.

Durham TJ(1), Libbrecht MW(1), Howbert JJ(1), Bilmes J(2), Noble WS(3)(4).

Author information:
(1)Department of Genome Sciences, University of Washington, Foege Building S-250,
Box 355065, 3720 15th Ave NE, Seattle, WA, 98195, USA.
(2)Department of Electrical Engineering, University of Washington, Paul Allen Center AE100R, Box 352500, 185 Stevens Way, Seattle, WA, 98195, USA.
(3)Department of Genome Sciences, University of Washington, Foege Building S-250,
Box 355065, 3720 15th Ave NE, Seattle, WA, 98195, USA. william-noble@uw.edu.
(4)Department of Computer Science and Engineering, University of Washington, Foege Building S-250, Box 355065, 3720 15th Ave NE, Seattle, WA, 98195, USA.
william-noble@uw.edu.

The Encyclopedia of DNA Elements (ENCODE) and the Roadmap Epigenomics Project seek to characterize the epigenome in diverse cell types using assays that identify, for example, genomic regions with modified histones or accessible chromatin. These efforts have produced thousands of datasets but cannot possibly measure each epigenomic factor in all cell types. To address this, we present a method, PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition (PREDICTD), to computationally impute missing experiments. PREDICTD leverages an elegant model called "tensor decomposition" to impute many experiments simultaneously. Compared with the current state-of-the-art method, ChromImpute, PREDICTD produces lower overall mean squared error, and combining the two methods yields further improvement. We show that PREDICTD data captures enhancer activity at noncoding human accelerated regions. PREDICTD provides reference imputed data and open-source software for investigating new cell types, and demonstrates the utility of tensor decomposition and cloud computing, both promising technologies for bioinformatics.

result – zs-Glove - RNAseq

- key:RNAseq 5000ABs method:zs-Glove

- key:RNAseq 8000ABs method:zs-Glove

- 耗时太长

result –Standford-Glove -RNAseq

- key:RNAseq 8000ABs method:Glove1.2
- key:RNAseq 60000ABs method:Glove1.2

- key:lung cancer gene mutation 100000ABs method:Glove1.2

- key:chipseq 800ABs method:Glove1.2

- key:cancer gene phenotype 290000ABs method:Glove1.2

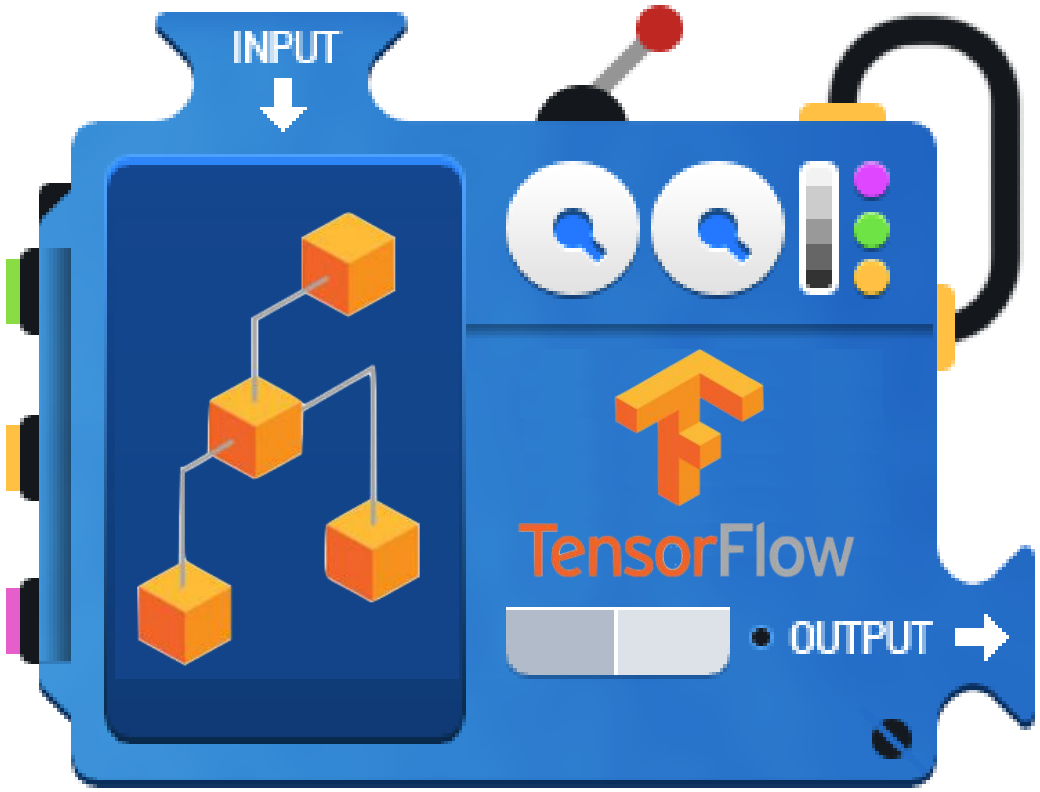
相似性计算-章胜同学的代码（修改）

- 向量的余弦相似性

TensorFlow Mechanics

2 feed data and run graph (operation)
sess.run (op, feed_dict={x: x_data})

1 Build graph using
TensorFlow operations



3 update variables
in the graph
(and return values)

thanks for watching