USING HADOOP TO IDENTIFY FALSE POSITIVES IN PYROPRINTING

Colin Adams & Skyler Gordon

IAB Presentation

Spring 2016
AGENDA

• Background
  • *E. coli* and Pyroprinting

• Implementation
  • Hadoop

• Results
BACKGROUND
**E. COLI DNA**

- **Structure**
  - Circular chromosome
  - Seven copies of rRNA operons
  - Operon structure

- **Interested in copies of ITS1 and ITS2 alleles**

- **Allele Ratio**
  - 7 - all seven alleles are same sequence
  - 6:1 – six of one allele, one different
  - 5:2 – five of one allele, two of another
  - ...
  - 1:1:1:1:1:1:1 – every allele is different

• Rapid, inexpensive strain typing method
• Creates fingerprint from ITS1 or ITS2 alleles
• Two *E. coli* with matching pyroprints considered same strain

**Question:** How often do the pyroprints between *E. coli* match but the allele ratio or sequences differ? In other words, how frequent is a false positive?

SOLUTION STRUCTURE

- Determine alleles and allele ratio for each *E. coli* isolate
  - If two isolates have matching pyroprints but different alleles/ratio → false positive

- Process
  - Choose 100 *E. coli* isolates
  - Amplify & Modify ITS Regions (PCR)
  - Have sequenced (Illumina/Mr. DNA)
    - 5-20 million sequences
  - Analyze sequences (Hadoop)
    - Count alleles for each isolate and region
    - Determine allele ratio

```
TAT ATCG T GGA...C ACGTACGT...
```

| barcode | its1/2 primer | allele |
IMPLEMENTATION
HADOOP BACKGROUND

• Distributed storage and processing framework
• Simple data processing pipeline
  • Map(obj) -> <key, value>…
    • Transforms each input object into any number of <key, value> pairs
  • Shuffle
    • Map() output is shuffled so all objects with same key are grouped for Reduce()
  • Reduce(key, values…) -> <key, value>…
    • Performs reduction on values with matching key and outputs any number of results
**FIRST HADOOP JOB**

**Map**

- Text File (10m)
  - TATATCG...
  - Attempts to parse into Sequence object

**Reduce**

- **Valid**
  - Sums occurrences
  - Outputs individual

- **Invalid**
  - Outputs individual

Possible Validity Codes:
- **Valid**
- Invalid Barcode
- Invalid Region*
- Too Short

*Configurable Approx. Match

**Valid:**

- **GTCA:** ITS1: CTCC...GT
- **VALUE:**
  - TATGTCATGGAA...ACCTCC...GT

**Invalid:**

- **KEY:** GTCA: ITS1
- **VALUE:** CTCC...GT: 1
SECOND HADOOP JOB

Text File (Prior)

KEY:
  barcode:region
VALUE:
  allele:count
...

Identity Map

Mutated
Mutated Allele count < 1% of Total
Result
Analyze ratio of other allele counts

Analysis:
• Look at allele ratios w/ same number
  • E.g. 2 alleles = 6:1, 5:2, 4:3
• Run Chi-Square w/ expected counts
• If there is only one that isn’t rejected, it is valid

Mutated
“barcode”:
“region”:
“allele”:
“count”:

Result
“barcode”:
“region”:
“statisticalValidity”:
“pValue”:
“ratios”:
“alleleCountRatios”: []

Identity Map

GTCA:ITS1   CTCC...GT:40
GTCA:ITS1   CTCC...AT:30

{  
  “barcode”: GTCA,
  “region”: ITS1,
  “statisticalValidity”: true,
  “pValue”: 1.0,
  “ratios”: [4:3],
}

“alleleCountRatios”: [
  {“allele”: “CTCC...GT”, “count”: 40, “ratio”: 4},
  {“allele”: “CTCC...AT”, “count”: 30, “ratio”: 3}
]
RESULTS
{"catalogNumber":9,"region":"ITS1","strainName":"Av-009","statisticallyValid":false,"ratios":['THREE_ONE_ONE_ONE_ONE'],"pValue":0.0,"barcode":"GTCA","alleleCountRatios":[]}

{"catalogNumber":9,"region":"ITS1","strainName":"Cw-053","statisticallyValid":false,"ratios":['THREE_ONE_ONE_ONE_ONE'],"pValue":0.0,"barcode":"CGAG","alleleCountRatios":[]}

{"catalogNumber":9,"region":"ITS1","strainName":"Pg-085","statisticallyValid":false,"ratios":['THREE_ONE_ONE_ONE_ONE'],"pValue":0.0,"barcode":"GCCA","alleleCountRatios":[]}
**PERFORMANCE**

Sequential & Parallel
- Single-Machine
  - Intel i7
  - 8 GB RAM
- Java 8 Streams API

Hadoop
- Google Cloud Compute
  - 1 Master
  - 3 Workers
  - 8 GB RAM each
LESSONS LEARNED

COLIN

• Technical
  • Hidden Complexity
  • Sequential → Hadoop
  • Cloud services
• Interdisciplinary
  • Be involved early and often
  • Many problems need CS
  • Efficiency

SKYLER

• Future Implications
THANK YOU