STANDING ON THE SHOULDERS OF GIANTS

HOW MASSIVE KNOWLEDGE-BASES ARE TRANSFORMING DATA ANALYTICS IN BIOLOGY

SAURABH SINHA

PROFESSOR OF COMPUTER SCIENCE
AND THE CARL R. WOESE INSTITUTE FOR GENOMIC BIOLOGY
CO-DIRECTOR & RESEARCH PI, NIH BD2K CENTER OF EXCELLENCE, UIUC & MAYO CLINIC
IT’S, UH, “GENOMICAL”
PERSPECTIVE

Big Data: Are We Genomical?

Zachary D. Stephens¹, Skyler L. Edgington², Miles J. Efron³, Ravishankarra⁴, E. Robinson⁵

Roy H. Campbell⁶, Chengxiang Zhai⁷, Saurabh Sinha⁸, Gene

Growth of DNA Sequencing

Worldwide Annual Sequencing Capacity

IN THE BEGINNING THERE WAS THE GENOME
YOUR GENOME = YOUR CELL’S SOFTWARE

main()
ACGGACG
GATTGGA
CGAAGAACool
GAMES)
else
se
“MACHINE” CODE IS NOT VERY USEFUL
SO PEOPLE STARTED PROFILING THE CODE
ALL ROADS LEAD TO A SPREADSHEET
<table>
<thead>
<tr>
<th>Genes</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tissue 1</td>
</tr>
<tr>
<td>Gene 1</td>
<td>20</td>
</tr>
<tr>
<td>Gene 2</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>100</td>
</tr>
<tr>
<td>Gene 20000</td>
<td>20</td>
</tr>
<tr>
<td>Genes</td>
<td>Conditions</td>
</tr>
<tr>
<td>----------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>Patient 1</td>
</tr>
<tr>
<td>Gene 1</td>
<td>20</td>
</tr>
<tr>
<td>Gene 2</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>100</td>
</tr>
<tr>
<td>Gene 20000</td>
<td>20</td>
</tr>
</tbody>
</table>
### SPREADSHEET ANALYTICS (A.K.A. BIOINFORMATICS)

#### Table

<table>
<thead>
<tr>
<th></th>
<th>Patient 1</th>
<th>Patient 2</th>
<th>...</th>
<th>Patient 400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>20</td>
<td>5</td>
<td>23</td>
<td>37</td>
</tr>
<tr>
<td>Gene 2</td>
<td>10</td>
<td>17</td>
<td>201</td>
<td>29</td>
</tr>
<tr>
<td>...</td>
<td>100</td>
<td>102</td>
<td>99</td>
<td>84</td>
</tr>
<tr>
<td>Gene 20000</td>
<td>20</td>
<td>45</td>
<td>74</td>
<td>62</td>
</tr>
</tbody>
</table>

#### Diagrams
- **Regression**
- **Classification**
- **Clustering**
“A GOOD DECISION IS BASED ON KNOWLEDGE AND NOT ON NUMBERS”
- PLATO
THE GWAS STORY

I1: AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I2: AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I3: AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I4: AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I5: AACGAGCTAGCGATCGATCGACACGACTACGAGGT +
I6: AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I7: AACGAGCTAGCGATCGATCGACTACGACTACGAGGT
I8: AACGAGCTAGCGATCGATCGACACGACTACGAGGT +

Disease?

Doesn’t work as well as we’d like it to
SEEK MODULES, NOT INDIVIDUAL GENES

Many ways to ‘break’ the code
BIOLOGICAL NETWORKS GALORE
BIOLOGICAL NETWORKS GALORE

‘Knowledge Network’:
3M nodes
80M edges
82 edge types
GIANT NETWORK GUIDES BIOLOGICAL ANALYSIS

Knowledge network + user spreadsheet
EXAMPLE: FINDING GENES INFLUENCING DRUG RESPONSE

- Individuals
- Drug Response
- Correlation
- Drug Related Genes
- Random Walk w/ Restart
- Ranking of genes by relevance to drug response
EXAMPLE: FINDING GENES INFLUENCING DRUG RESPONSE

Validated 17 genes for several cancer drugs
CYBERINFRASTRUCTURE
HOW BIOLOGISTS DO BIOINFORMATICS TODAY

• HIRE BIOINFORMATICIAN OR SEEK BIOINFORMATICS COLLABORATOR.

• DELEGATE THE FOLLOWING:
  • DOWNLOAD AND INSTALL CODE.
  • BUY COMPUTE CLUSTERS
  • RUN CODE ON CLUSTER

IN SHORT, PAINFUL.
A RANDOM WALK in the CLOUDS
CLOUD-BASED KNOWLEDGE ENGINE FOR GENOMICS

User Spreadsheet

Knowledge Network

Analytics

Classification

Clustering

Regression

Decision-tree
Meta-paths
PCA
Graph-mining
Elastic-net
Network-smoothing
Support-Vector-Machine
Lasso
Hierarchical-clustering
NMF
Lars
Dimensionality-reduction
Feature-selection
Random-walk
Linear-regression
Compressed-sensing
COMPLEX WORKFLOWS ON THE CLOUD

Network-based stratification of tumor mutations
Hofree et al. *Nature Methods* 2013

**Spreadsheets**

**Knowledge Network**

**Network Smoothing**

**Random Walk with Restart**

**Aggregate Subtypes**

**Hierarchical Clustering**

**Network NMF**

**Clustering Algorithm**

**Docker Containers**

**Easy, parallel exploration of workflow variants**
SOFTWARE IS ONLY AS GOOD AS ITS FRONT END
CARL R. WOESE INSTITUTE FOR GENOMIC BIOLOGY

NIH

CS @ ILLINOIS