Pattern Recognition In Clinical Data

Saket Choudhary
*Dual Degree Project*

Guide: Prof. Santosh Noronha

October 30, 2013
INTRODUCTION

INTRODUCTION
Objective

SIGNIFICANT MUTATIONS
Next Generation Sequencing
Motivation
Computational Methods for Driver Detection

VIRAL GENOME DETECTION
Workflow

REPRODUCIBILITY
Reproducibility

CONCLUSIONS
Wrapping up
**Objective**
Objective

Next Generation Sequencing & Cancer Research
**Objective**

Tools for Biologists Reproducible analysis

Next Generation Sequencing & Cancer Research
Objective

Tools for Biologists Reproducible analysis

Better/New Algorithms

Next Generation Sequencing & Cancer Research
OBJECTIVE

- Tools for Biologists Reproducible analysis
- Better/New Algorithms
- Next Generation Sequencing & Cancer Research
- Driver & Passenger Mutation Detection
OBJECTIVE

Tools for Biologists Reproducible analysis

Next Generation Sequencing & Cancer Research

Better/New Algorithms

Driver & Passenger Mutation Detection

Galaxy Tools
**OBJECTIVE**

- Driver & Passenger Mutation Detection
- Galaxy Tools
- Viral Genome Integration
- Tools for Biologists Reproducible analysis
- Better/New Algorithms
- Next Generation Sequencing & Cancer Research
- Benchmarking Alignment tools
- Ensembl Method
- Lit. Survey
**Objective**

- Galaxy Workflow
- Viral Genome Integration
- Tools for Biologists Reproducible analysis
- Driver & Passenger Mutation Detection
- Galaxy Tools
- Better/New Algorithms
- Next Generation Sequencing & Cancer Research
Objective

Tools for Biologists Reproducible analysis

Next Generation Sequencing & Cancer Research

Better/New Algorithms

Benchmarking Alignment tools

Viral Genome Integration

Galaxy Workflow

Galaxy Tools

Driver & Passenger Mutation Detection

Ensembl Method

Lit. Survey

Next Generation Sequencing & Cancer Research

Better/New Algorithms

Benchmarking Alignment tools

Viral Genome Integration

Galaxy Workflow

Galaxy Tools

Driver & Passenger Mutation Detection

Ensembl Method

Lit. Survey
**Objective**

- **Next Generation Sequencing & Cancer Research**
- **Driver & Passenger Mutation Detection**
- **Galaxy Tools**
- **Viral Genome Integration**
- **Galaxy Workflow**
- **Tools for Biologists Reproducible analysis**
- **BWA v/s BWA-PSSM**
- **Benchmarking Alignment tools**
- **Better/New Algorithms**
- **Ensembl Method**
- **Lit. Survey**
Objective

Tools for Biologists Reproducible analysis

Next Generation Sequencing & Cancer Research

Better/New Algorithms

Driver & Passenger Mutation Detection

BWA v/s BWA-PSSM

Benchmarking Alignment tools

Galaxy Tools

Viral Genome Integration

Galaxy Workflow

Ensembl Method

Lit. Survey
Objective

Tools for Biologists Reproducible analysis

Next Generation Sequencing & Cancer Research

Better/New Algorithms

Driver & Passenger Mutation Detection

BWA v/s BWA-PSSM

Benchmarking Alignment tools

Ensemble Method

Galaxy Tools

Viral Genome Integration

Galaxy Workflow

Driver & Passenger Mutation Detection

Galaxy Tools

Ensemble Method
**Objective**

- Tools for Biologists Reproducible analysis
- Galaxy Workflow
- Viral Genome Integration

- Next Generation Sequencing & Cancer Research
- Galaxy Tools
- Driver & Passenger Mutation Detection

- Better/New Algorithms
- BWA v/s BWA-PSSM
- Benchmarking Alignment tools

- Ensembl Method
- Lit. Survey
- Driver & Passenger Mutation Detection
Next Generation Sequencing
Next Generation Sequencing
Next Generation Sequencing
Next Generation Sequencing
NGS: Why?

- **Molecular Approach**: Study of variations at the ‘base’ level
- **Low Cost**: 1000$ genome
- **Faster**: Quicker than traditional sequencing techniques
NGS: Why?

- **Molecular Approach**: Study of variations at the ‘base’ level
- **Low Cost**: $1000 genome
- **Faster**: Quicker than traditional sequencing techniques
**NGS: Why?**

- **Molecular Approach:** Study of variations at the 'base' level
- **Low Cost:** 1000$ genome
- **Faster:** Quicker than traditional sequencing techniques
NGS: WHERE?

- Study variations, genotype-phenotype association
- Look for ‘markers of diseases’
- Prognosis
NGS: Where?

- Study variations, genotype-phenotype association
- Look for ‘markers of diseases’
- Prognosis
NGS: Where?

- Study variations, genotype-phenotype association
- Look for ‘markers of diseases’
- Prognosis
NGS: Mutations

- $3 \times 10^9$ base pairs
- We are all 99.9% similar, at DNA level
- More than 2 million SNPs
- No particular pattern of SNPs
- If a certain mutation causes a change in an amino acid, it is referred to as non synonymous (nsSNV)
Drivers and Passengers I

Cancer is known to arise due to mutations
Not all mutations are equally important!

Somatic Mutations
Set of mutations acquired after zygote formation, over and above the germline mutations

Driver Mutations
Mutations that confer growth advantages to the cell, being selected positively in the tumor tissue
Drivers and Passengers

Drivers are **NOT** simply *loss of function* mutations, but more than that:

- **Loss of function**: Inactivate tumor suppressor proteins
- **Gain of function**: Activates normal genes transforming them to oncogenes
- **Drug Resistance Mutations**: Mutations that have evolved to overcome the inhibitory effect of drugs
Drivers and Passengers

Drivers are **NOT** simply *loss of function* mutations, but more than that:

- **Loss of function**: Inactivate tumor suppressor proteins
- **Gain of function**: Activates normal genes transforming them to oncogenes
- **Drug Resistance Mutations**: Mutations that have evolved to overcome the inhibitory effect of drugs
Drivers and Passengers

Drivers are NOT simply loss of function mutations, but more than that:

▶ **Loss of function**: Inactivate tumor suppressor proteins
▶ **Gain of function**: Activates normal genes transforming them to oncogenes
▶ **Drug Resistance Mutations**: Mutations that have evolved to overcome the inhibitory effect of drugs
Driver Mutations: Why?

Identify **driver mutations** → better therapeutic targets
But how does one zero down upon the exact set? → experiments are too costly, probably infeasible for 2 million+ SNPs → Leverage computational analysis

- **Low cost of NGS comes with a heavier roadblock of data analysis**
- Searching among 2 million+ SNPs is a non-trivial, and a computationally intensive problem
- Softwares have a low consensus ratio amongst them selves ←→ Defining a driver, computationally is non-trivial
- However there is no tool that allows one to visualise the results on an input across the cohort of tools
**Driver Mutations: Why?**

Identify **driver mutations** → better therapeutic targets
But how does one zero down upon the exact set? → experiments are too costly, probably infeasible for 2 million+ SNPs → Leverage computational analysis

- Low cost of NGS comes with a heavier roadblock of data analysis
- Searching among 2 million+ SNPs is a non-trivial, and a computationally intensive problem
- Softwares have a low consensus ratio amongst them selves ↔ Defining a driver, computationally is non-trivial
- However there is no tool that allows one to visualise the results on an input across the cohort of tools
Driver Mutations: Why?

Identify **driver mutations** → better therapeutic targets
But how does one zero down upon the exact set? → experiments are too costly, probably infeasible for 2 million+ SNPs → Leverage computational analysis

▶ Low cost of NGS comes with a heavier roadblock of data analysis
▶ Searching among 2 million+ SNPs is a non-trivial, and a computationally intensive problem
▶ Softwares have a low consensus ratio amongst them selves ←→ Defining a driver, computationally is non-trivial
▶ However there is no tool that allows one to visualise the results on an input across the cohort of tools
Driver Mutations: Why?

Identify **driver mutations** $\rightarrow$ better therapeutic targets
But how does one zero down upon the exact set? $\rightarrow$
experiments are too costly, probably infeasible for 2 million+
SNPs $\rightarrow$ Leverage computational analysis

- Low cost of NGS comes with a heavier roadblock of data
  analysis
- Searching among 2 million+ SNPs is a non-trivial, and a
  computationally intensive problem
- Softwares have a low consensus ratio amongst them selves
  $\leftarrow$ Defining a driver, computationally is non-trivial
- However there is no tool that allows one to visualise the
  results on an input across the cohort of tools
Machine Learning I

Two datasets:

- **Training**: *Labeled* dataset, containing a table of features with mutations labelled as "drivers/passengers"
- **Test**: 'Learning' from training dataset, test the prediction model

Table: Training Dataset

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>Ref</th>
<th>Alt</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27822</td>
<td>A</td>
<td>G</td>
<td>Driver</td>
</tr>
<tr>
<td>1</td>
<td>27832</td>
<td>T</td>
<td>G</td>
<td>Driver</td>
</tr>
<tr>
<td>2</td>
<td>47842</td>
<td>G</td>
<td>C</td>
<td>Passenger</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Machine Learning II

Table: Test Dataset

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>Ref</th>
<th>Alt</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27824</td>
<td>A</td>
<td>G</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>47832</td>
<td>T</td>
<td>G</td>
<td>?</td>
</tr>
</tbody>
</table>
Machine Learning: Feature Selection I

Machine Learning relies on a set of features for training. Redundant features should be avoided.

CHASM [1] makes use of $p(X_i)$ represents the probability of occurrence of an event $X_i$.

Considering a series of events $X_1, X_2, X_3..., X_n$ analogous ‘series of packets’ in communication theory, the information received at each step can be quantified on a log scale by:

$$\frac{1}{\log_2(X_i)} = -\log_2(p(X_i))$$  \hspace{1cm} (1)$$

The expected value of information from a series of events is called Shannon entropy: $H(X)$:

$$H(X) = - \sum_i p(X_i) \log_2 p(X_i)$$  \hspace{1cm} (2)$$
**Machine Learning: Feature Selection II**

Mutual Information between two random variables $X, Y$ is defined as the amount of information gained about random variable $X$ due to additional information gained from the second, $Y$:

$$I(X, Y) = H(X) - H(X | Y)$$  \hspace{1cm} (3)

Here:
- $X$: Class Label [Driver/Passenger]
- $Y$: Predictive Feature

and hence $I(X, Y)$ represents how much information was gained about the class label $Y$ from knowledge of a feature $X$

Simplifying:

$$I(X, Y) = \sum p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$  \hspace{1cm} (4)
FUNCTIONAL IMPACT I

▶ If a certain mutation confers an advantage to the cell in terms of replication rate, it is probably going to be selected while all those mutations that reduce its fitness have a higher chance of being eliminated from the population.

▶ Certain residues in a MSA of homologous sequences are more conserved than others. A highly conserved if mutated is possibly going to cost a lot since what had ‘evolved’ is disturbed!

▶ Scores can be assigned based on this “conservation” parameter.
FUNCTIONAL IMPACT II

Figure: SIFT [?] algorithm

1. User inputs query sequence

>FASTA header
I R R L R P M D

2. SIFT searches protein databases for related sequences

3. SIFT builds a sequence alignment

4. SIFT calculates conservation value and scaled probability for each position

Hydrophobic conserved
Highly conserved
Unconserved

5. SIFT makes predictions

Probability < cutoff

No

Tolerated

Yes
Some of the common tools/algorithms used for driver mutation prediction:

- SIFT
- Polyphen
- Mutation Assesor
- TransFIC
- Condel
FRAMEWORK FOR COMPARING VARIOUS TOOLS I

- Different tools use different formats, give different outputs for similar input
- Running analysis on multiple tools → keep shifting data formats
- Concordance?

### Polyphen2 Input

<table>
<thead>
<tr>
<th>Location</th>
<th>SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1:888659</td>
<td>T/C</td>
</tr>
<tr>
<td>chr1:1120431</td>
<td>G/A</td>
</tr>
<tr>
<td>chr1:1387764</td>
<td>G/A</td>
</tr>
<tr>
<td>chr1:1421991</td>
<td>G/A</td>
</tr>
<tr>
<td>chr1:1599812</td>
<td>C/T</td>
</tr>
<tr>
<td>chr1:1888193</td>
<td>C/A</td>
</tr>
<tr>
<td>chr1:1900186</td>
<td>T/C</td>
</tr>
</tbody>
</table>
**FRAMEWORK FOR COMPARING VARIOUS TOOLS II**

<table>
<thead>
<tr>
<th>SIFT Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,888659,T,C</td>
</tr>
<tr>
<td>1,1120431,G,A</td>
</tr>
<tr>
<td>1,1387764,G,A</td>
</tr>
<tr>
<td>1,1421991,G,A</td>
</tr>
<tr>
<td>1,1599812,C,T</td>
</tr>
<tr>
<td>1,1888193,C,A</td>
</tr>
<tr>
<td>1,1900186,T,C</td>
</tr>
</tbody>
</table>
**Driver Mutations: Tools don’t agree**

X Axis: Condel Score  
Y Axis: MA Score
Solution?:
Galaxy[?], an open source web-based platform for bioinformatics, makes it possible to represent the entire data analysis pipeline in an intuitive graphical interface.

Figure: Galaxy Workflow polyphen2 algorithm
Run all tools in one go:

Figure: Run all tools
Compare all tools:

Figure: Compare all tools
Cervical cancers have been proven to be associated with Human Papillomavirus (HPV). Cervical cancer datasets from Indian women were put through an analysis to detect:

1. Any possible HPV integration
2. Sites of HPV integration

Who Cares?

- Replacing whole genome sequencing, by targeted sequencing at the sites where these viruses have been detected in a cohort of samples, thus speeding up the whole process.
**Introducción**

**Significativos Mutaciones**

**Viral Genoma Detección**

**Reproductibilidad**

**Conclusión**

1. **RawData**
   - Aligned data to human genome reference
2. **Extract**
   - Unmapped regions
3. **Align**
   - Unmapped regions to Virus genome
4. **BLAST**
RawData

Align data to human genome reference

Extract unmapped regions

Align unmapped regions to Virus genome

BLAST
RawData → Align data to human genome reference → Extract unmapped regions → Align unmapped regions to Virus genome → BLAST
RawData

Align data to human genome reference

Extract unmapped regions

Align unmapped regions to Virus genome

BLAST
**RawData**

1. Align data to human genome reference
2. Extract unmapped regions
3. Align unmapped regions to Virus genome
4. BLAST
RawData → Align data to human genome reference → Extract unmapped regions → Align unmapped regions to Virus genome → BLAST
Galaxy Workflow
Figure: Aligned HPV genomes
Reproducibility

- In pursuit of novel 'discovery', standardizing the data analysis pipeline is often ignored, leading to dubious conclusions
- Analysis should be reproducible and above all, correct
- Parameter’s values can change the results by a big factor, they need to be documented/logged
- Garbage in, Garbage out
CONCLUSIONS

With the Galaxy tool box for identification of significant mutations and the study of the science behind the methods, the next steps would be to:

▶ **Open source the toolbox to the community:** A tool makes little sense if it is not in a usable form, community feedback will be used to add more tools and improve the existing ones

▶ **A new method for driver mutation prediction:** all the methods have low level of concordance. A new method that takes into account the available data at all levels: mutations, transcriptome and micro array data is possible. With the Galaxy toolbox in place, it would be possible to integrate information at various levels
FUTURE WORK

- Develop an algorithm that integrates machine learning approach with functional approach by zeroing down upon only those attributes that are known to have an impact
- The algorithm would also account for information at other levels: RNA expressions, Clinical data.
- Integrating information at all levels would provide a deeper insight
- The developed Galaxy toolbox will be used as the basic framework for integrating information
REFERENCES I