Visualizing Co-occurrence of Events in Populations of Viral Genome Sequences

Alper Sarikaya¹, Michael Correll², Jorge M. Dinis¹, David H. O’Connor¹,³, and Michael Gleicher¹

¹ University of Wisconsin-Madison
² University of Washington
³ Wisconsin National Primate Center

http://graphics.cs.wisc.edu/Vis/CoocurViewer/

@yleralp
http://cs.wisc.edu/~sarikaya/
Outline

Biological Background
bound our design space and exploration

Displaying occurrence relationships (in biology)
similar visual metaphors and related workflow techniques

MatrixViewer
exploring design decisions in the first iteration, learning from analyst confusion

CooccurViewer
analyst-guided exploration of ‘interesting’ co-occurrences

Case Study, Future Work
application to virology workflow, application to other data domains
RNA viruses are very error prone in replication lacks the error-checking of double-stranded DNA

Viruses accumulate variation to help its survival known as “viral fitness,” this can identify how a virus can adapt to new challenges from immune response

Influenza, H1N1, Zika are hard to eliminate these RNA viruses all have a wide swath of variation, and therapies for these viruses attack essential viral function
Timeline of RNA virus infection

Original virus

Infection

Population of viral genomes
(lots of variants!)
The Analysis Goal

Discover where functional shifts are occurring driven by virologist intuition coupled with previous work

Identify ‘co-occurrences’ of mutations in genome epistasis can identify functional shifts that are preferred in the given environment

Identify groups of like-behaving subpopulations understand how mutated viruses conserve vital functionality in order to target therapies
The Analysis Goal

Discover where functional shifts are occurring driven by virologist intuition coupled with previous work

**Identify ‘co-occurrences’ of mutations in genome**
epistasis can identify functional shifts that are preferred in the given environment

**Identify groups of like-behaving subpopulations**
understand how mutated viruses conserve vital functionality in order to target therapies
The Analysis Goal

Discover where functional shifts are occurring driven by virologist intuition coupled with previous work

Identify ‘co-occurrences’ of mutations in genome epistasis can identify functional shifts that are preferred in the given environment

Identify groups of like-behaving subpopulations understand how mutated viruses conserve vital functionality in order to target therapies
Population of Viral Genomes

```
ACTTCTTTTGTAAGGA
AGTACTTTGATAAGA
AGTTCTTTGTAAAGA
ACTTCTTTGATAGGA
ACTTCTTTGATAAGA
AGTTCTTTGTAAAGA
```

Reference

Variant

Genome
Obtaining the Data

NGS sequencer

Viral sample → Genome segments → Aligned genome

'Biinformatics' by Edward Boatman from the Noun Project [CC BY 3.0 US]
Aligned Genome Reads

100,000s of genomic fragments

100s of genome positions
Identify pairs of positions where mutations co-occur. Identify epistasis, indicating a functionality shift.

Analysis requires a maximum of sifting through $(\# \text{ positions})^2$ correlations.
Outline

Biological Background
bound our design space and exploration

Displaying occurrence relationships (in biology)
similar visual metaphors and related workflow techniques

MatrixViewer
exploring design decisions in the first iteration, learning from analyst confusion

CooccurViewer
analyst-guided exploration of ‘interesting’ co-occurrences

Case Study, Future Work
application to virology workflow, application to other data domains
Problem Abstraction

Event Axis

Observer

Observations
Excited

Ambivalent

Skeptical

Delight!

Alice

Bob

Carol

Dan

Erin

EuroVis 2016
Abstraction: grouping
Abstraction: interesting events

significant co-occurrence
Related Work: Genome Browsers

ensemble Genome Browser

UCSC Genome Browser
Visualization of Correlation


See Diaz, et al [2002] for more on graph construction of correlation. [doi]
Visualization of Correlation

Outline

**Biological Background**
bound our design space and exploration

**Displaying occurrence relationships (in biology)**
similar visual metaphors and related workflow techniques

**MatrixViewer**
exploring design decisions in the first iteration, learning from analyst confusion

**CooccurViewer**
analyst-guided exploration of ‘interesting’ co-occurrences

**Case Study, Future Work**
application to virology workflow, application to other data domains
Visualizing Pairwise Correlation

Collect counts of bases (A, C, T, G) for each pair of positions.
Visualizing Pairwise Correlation

Compute co-occurrence strength between every pair of genomic positions

\[ M_{i,j} = \Pr(j_- | i_-) - \Pr(j_- | i_+) \]
Show co-occurrences in full pairwise genomic space, in a web browser

Scale up to 20,000 x 20,000 about 280k correlations considered boost with WebGL and binary files

Color shows the co-occurrence strength

\[ M_{i,j} = Pr(j \mid i^-) - Pr(j \mid i^+) \]
Show co-occurrences in full pairwise genomic space, in a web browser.

Scale up to 20,000 x 20,000 about 280k correlations considered boost with WebGL and binary files.

Color shows the co-occurrence strength.

\[ M_{i,j} = \Pr(j \mid i) - \Pr(j \mid i^+) \]
Too much data to sift through
need to find the needle in the haystack by displaying the full space

Alignment errors produce false positives
only showing pairwise co-occurrences with 5% mutations makes matrix very sparse

Difficult to get an overview
Changing visualization parameters can have minute changes, easily missed by analyst
Learning from First Steps

Always present data in genomic sequence order
match mental model of the virologist, allows for rapid spatial identification

Display annotations alongside genome
annotations can provide valuable wayfinding with previous results and reading frames [nucleic bases to proteins]

Scaffold to navigate space of all pairwise correlation
support the analyst in discovering the most “interesting” co-occurrences

Support identifying synonymy
due to degeneracy in transcription, a change in the genome may not translate to a change in derived protein
Outline

Biological Background
bound our design space and exploration

Displaying occurrence relationships (in biology)
similar visual metaphors and related workflow techniques

MatrixViewer
exploring design decisions in the first iteration, learning from analyst confusion

CooccurrViewer
analyst-guided exploration of ‘interesting’ co-occurrences

Case Study, Future Work
application to virology workflow, application to other data domains
Interest Metrics

Coverage (read depth)
is there enough data to be meaningful?

Variation (mutations)
only showing pairwise co-occurrences with 5% mutations makes matrix very sparse

Co-occurrence strength
measure of interesting correlation, signed positive or negative correlation

\[ V_i = \Pr(i_\cdot) \]

\[ M_{i,j} = \Pr(j_\cdot | i_-) - \Pr(j_\cdot | i_+) \]
User-controlled metrics

http://graphics.cs.wisc.edu/Vis/CooccurViewer
An annotations positions with significant co-occurrences.

http://graphics.cs.wisc.edu/Vis/CooccurViewer
Pairwise co-occurrences with a particular position

http://graphics.cs.wisc.edu/Vis/CooccurViewer
Reads that do not overlap with the paired position
Position 9926 (found 1,808 reads)
Ref. nucleotide: T (synonym nucleotides: A, C, G)

282 (15.6% of total) reads are variant
33 (11.7%) do not overlap 9921
4 (1.4%) link to variants at 9921
245 (86.9%) link to reference at 9921

Nucleotides at this position:
- A: 35
- T: 1,526
- C: 163
- G: 84

Found 2 correspondences for 9926
Showing 1-2 of 2

<< Prev page   Next page >>
Outline

Biological Background
bound our design space and exploration

Displaying occurrence relationships (in biology)
similar visual metaphors and related workflow techniques

MatrixViewer
exploring design decisions in the first iteration, learning from analyst confusion

CooccurViewer
analyst-guided exploration of ‘interesting’ co-occurrences

Case Study, Future Work
application to virology workflow, application to other data domains
Sample of SIV: simian equivalent of HIV

Large cluster of correlated mutations in Nef protein to evade T cell recognition

Nearly no co-occurrences in structural proteins Gal & Pol

Scale: 238k reads [24-151 bp each], genome is 9,973 bp; 2.78M pairwise comparisons
Discussion

Use analyst-controlled metrics to focus exploration requires iteration to identify the key metrics interesting co-occurrences

Displaying the full space does not necessarily empower analysts design must enable the analyst to quickly target their analysis to the critical relationships

Providing usable context and scaffolding display of annotations and common genomic axis maintain virologists’ mental map of the viral genome
Future Work

Support comparison between multiple samples, and multi-step co-occurrence requires iteration to identify the key triggers, appropriate summarization.

Data aggregation and filtering techniques to support larger data sizes
filter and aggregate data using WebGL [e.g. imMens], compress data.

Application to other event-driven sequences
findings within this work can drive exploration of other domains, such as medical event data, event log data, etc.
Acknowledgments

Funding from the NIH and NSF
we graciously acknowledge the support of NSF IIS-1162037 and NIH award 5R01AI077376-07

Feedback from colleagues, virologists, and reviewers
for lifting this design study to benefit both the domain and vis communities

Code and working demo available online!
web:  http://graphics.cs.wisc.edu/Vis/CoocurViewer/
github:  http://github.com/uwgraphics/CooccurViewer/