**Variant View**

*Visualizing Sequence Variants in their Gene Context*

**Tamar Munzner**
Department of Computer Science
University of British Columbia

*BioIT World, Data Visualizations and Exploratory Tools - Talk*
20April 2014 Boston MA

http://www.cs.ubc.ca/~tmm/talks.html#bioit14

**Variant View: Visualization Design Study**

*• a specific real-world problem*
  *real users and real data*
  *collaboration (often) fundamental*
*design a visualization system*
  *implications: requirements, multiple ideas*
*validate the design*
  *at appropriate levels*

• reflect about lessons learned
  *translatable research improve design guidelines for vis in general*
  *confirm, refine, repeat, propose*

- Exons
- Gene
- Transcript
- Protein
- Genome
- Translation
- Protein Regions

**How To Do Design Studies**

*• definitions*

*9-stage framework*

*• 32 pitfalls and how to avoid them*

**When To Do Design Studies**

**INFORMATION LOCATION**

*computerhead*

**TASK CLARITY**

*fuzzy crisp*

*NOT ENOUGH DATA*

**DESIGN STUDY METHODOLOGY**

*SUITEABLE*

**ALGORITHM**

*POSSIBLE*

**Variant View**

*Visualizing Sequence Variants in their Gene Context*

**Cancer Research**

• collaboration with analysts at BC Genome Sciences Center
  *analyzing genetic basis of leukemia*
  *driving task*
  *discover new candidate genes with harmful variants*
  *two big questions*
  *what to show*
  *how to show it*
  *visual encoding idiom*

**Selectivity for representations cross-cuts domains!**

**Reference Genome DNA:** ATG TCA ACA CTT

**Sample 1 Genome DNA:** ATG TGC ACA CTT

**Sample 2 Genome DNA:** ATG TCA ACA CTT

**Abstractions and Idioms**

• abstractions
  *translate from specifics of domain to vocabulary of vis*
  *task abstraction: why they are looking at it*
  *data abstraction: what to draw*
  *transform data into forms useful for task at hand*
  *don’t just draw what you’re given; decide what is the right thing!*

• idioms
  *visual encoding idiom: how to draw*
  *interaction idiom: how to manipulate*

• focus today: two mappings
  *from domain to abstraction*
  *from abstraction to idiom*

**Option 1: DNA visualization**

**Option 2: Protein visualization**

**Option 3: RNA visualization**

**Sequence Variant Definition**

• difference between reference and given genome

**Reference Genome DNA:** ATG TCA ACA CTT

**Sample 1 Genome DNA:** ATG TGC ACA CTT

**Sample 2 Genome DNA:** ATG TCA ACA CTT

**Data: Filtering to relevant biological levels and scales**

**Abstractions**

**Data abstraction**

• from specifics of domain to vocabulary of vis

**Validation**

• personal
• outward-facing

**Design Study Methodology**

*Reflections from the Trenches and from the Stacks*

**Design Study Methodology**

*• a specific real-world problem*
  *real users and real data*
  *collaboration (often) fundamental*
*design a visualization system*
  *implications: requirements, multiple ideas*
*validate the design*
  *at appropriate levels*

• reflect about lessons learned
  *translatable research improve design guidelines for vis in general*
  *confirm, refine, repeat, propose*

- Exons
- Gene
- Transcript
- Protein
- Genome
- Translation
- Protein Regions

**Data abstraction:**

• horizontal tracks: user data
• shared coordinate system: genome coordinate (bp)

**Problems**

- Idiosyncrasies of interest spread out across largeextent
- must zoom in to see exons, known features
- must zoom out to see human scale
- high cognitive load for interactions
- must already know where to look

**Variant View**

*Visualizing Sequence Variants in their Gene Context*

**Cancer Research**

• collaboration with analysts at BC Genome Sciences Center
  *analyzing genetic basis of leukemia*
  *driving task*
  *discover new candidate genes with harmful variants*
  *two big questions*
  *what to show*
  *how to show it*
  *visual encoding idiom*

**Selectivity for representations cross-cuts domains!**

**Reference Genome DNA:** ATG TCA ACA CTT

**Sample 1 Genome DNA:** ATG TGC ACA CTT

**Sample 2 Genome DNA:** ATG TCA ACA CTT

**Abstractions and Idioms**

• abstractions
  *translate from specifics of domain to vocabulary of vis*
  *task abstraction: why they are looking at it*
  *data abstraction: what to draw*
  *transform data into forms useful for task at hand*
  *don’t just draw what you’re given; decide what is the right thing!*

• idioms
  *visual encoding idiom: how to draw*
  *interaction idiom: how to manipulate*

• focus today: two mappings
  *from domain to abstraction*
  *from abstraction to idiom*

**Option 1: DNA visualization**

**Option 2: Protein visualization**

**Option 3: RNA visualization**

**Sequence Variant Definition**

• difference between reference and given genome

**Reference Genome DNA:** ATG TCA ACA CTT

**Sample 1 Genome DNA:** ATG TGC ACA CTT

**Sample 2 Genome DNA:** ATG TCA ACA CTT

**Data: Filtering to relevant biological levels and scales**

**Abstractions**

**Data abstraction**

• from specifics of domain to vocabulary of vis

**Validation**

• personal
• outward-facing

**Design Study Methodology**

*Reflections from the Trenches and from the Stacks*

**Design Study Methodology**

*• a specific real-world problem*
  *real users and real data*
  *collaboration (often) fundamental*
*design a visualization system*
  *implications: requirements, multiple ideas*
*validate the design*
  *at appropriate levels*

• reflect about lessons learned
  *translatable research improve design guidelines for vis in general*
  *confirm, refine, repeat, propose*

- Exons
- Gene
- Transcript
- Protein
- Genome
- Translation
- Protein Regions

**Data abstraction:**

• horizontal tracks: user data
• shared coordinate system: genome coordinate (bp)

**Problems**

- Idiosyncrasies of interest spread out across largeextent
- must zoom in to see exons, known features
- must zoom out to see human scale
- high cognitive load for interactions
- must already know where to look
Side-by-side comparison: MuSiC vs Variant View

Protein regions can overlap
Regions get separate lanes

Large bloom of repeated elements: more salient

Results

Verify known leukemia gene: Highly scored by sorting metric

Visual inspection reveals collocation of variants
Several functional protein regions affected

Highly scored by metric: not previously known, good candidate

In contrast, low scoring gene
No collocation of variants
Mostly unaffected protein regions

Additional tasks
• task 2: compare patients
  – clinical setting application
  – compare patient data to known harmful variants
  – challenge
  • similarity is loosely understood rather than fully characterized
  • visual inspection for what constitutes a match

Adapted Variant View with minimal changes
Navigate through patient data with list
Patient data emphasized with arrows
Patient has same harmful L to P mutation
Nonmatching variants

Additional tasks
• task 3: debug pipeline
  – data cleaning before analysis
  – analysts originally thought pipeline fully debugged
  – no perceived need for vis support

Tool revealed errors in the data
• The tool exposed artifacts in the data that slid past at least two rounds of quality metric filtering … this type of problem would not have been caught by our previous, automated methods.
  - Analyst 3

Reflections: vis design guidelines
• transferrable to other domains
  – specialize first, generalize later
• good for domains where with complex, multi-scale data
  – difficult to judge a priori which design elements will generalize
• high-level considerations
  – identifying scales of interest
  – what to visually encode directly vs what to support through interaction
  – when (and how) to abstract/visualize

Conclusions
• visual variant impact assessment
  – designed, implemented, and deployed tool for
  – originally designed for Discover Genes task
  – adapted to two others with minimal changes
• features
  – navigation-free main overview at gene level
  – reveal genes of interest through sorting by new derived metrics
• major considerations
  – what to show
    • filtering data scope
  – how to show it
    • carefully selected visual encodings

Further Information
• paper page
• open source software download
• further info
  – http://www.cs.ubc.ca/~tmm/talks.html#bioit14 (this talk, and many others)
  – http://www.cs.ubc.ca/group/infovis (papers, software, videos)
• acknowledgements
  – funding: Vancouver Institute for Visual Analytics (VIVA), Aeroinfo/Boeing, Mitacs
  – Dr. Aly Karsan
  – Rod Docking
  – Dr. Linda Chang
  – Dr. Gerben Duns
  – Simon Chang

•

Took revealed errors in the data
• The tool exposed artifacts in the data that slid past at least two rounds of quality metric filtering … this type of problem would not have been caught by our previous, automated methods.
  - Analyst 3

Reflections: vis design guidelines
• transferrable to other domains
  – specialize first, generalize later
• good for domains where with complex, multi-scale data
  – difficult to judge a priori which design elements will generalize
• high-level considerations
  – identifying scales of interest
  – what to visually encode directly vs what to support through interaction
  – when (and how) to abstract/visualize

Conclusions
• visual variant impact assessment
  – designed, implemented, and deployed tool for
  – originally designed for Discover Genes task
  – adapted to two others with minimal changes
• features
  – navigation-free main overview at gene level
  – reveal genes of interest through sorting by new derived metrics
• major considerations
  – what to show
    • filtering data scope
  – how to show it
    • carefully selected visual encodings

Further Information
• paper page
• open source software download
• further info
  – http://www.cs.ubc.ca/~tmm/talks.html#bioit14 (this talk, and many others)
  – http://www.cs.ubc.ca/group/infovis (papers, software, videos)
• acknowledgements
  – funding: Vancouver Institute for Visual Analytics (VIVA), Aeroinfo/Boeing, Mitacs
  – Dr. Aly Karsan
  – Rod Docking
  – Dr. Linda Chang
  – Dr. Gerben Duns
  – Simon Chang

•