CS 533C project update

Visualization
of the evolutions of a virus genetic sequence

Mahshid ZeinalyBaraghoush (mzeinaly@sfu.ca)

Overview

- The target domain + my goals
- Original dataset/tasks
- Translate dataset :

Data abstraction → Data visual encoding

- Translate task:

Task abstraction → Visual encoding/interaction

- Next steps

The target domain

- The IEEE Vast Challenge 2010
- Evaluate IMAS "Interactive Multi-genomic Analysis System"

Goal:

Enhance IMAS in an iterative process using infovis design guidelines

Domain vocabulary

- Gene sequence

  ATGCACCGCCCTGCGCAGTTCATAG

- Virus replication → Virus strains
- Substitutions → Mutations of virus strains

  ATGCACCGCCCTGCGCAGTTCATAG

  ATGCACCGCCGTGCGCAGTTCATAG

- Disease characteristics

Domain: dataset

Tabular

Virus strain sequences table:
100 virus strains
1000 nucleotides long

Sequence ID	1	2	3	4	5	6	7	8
118	A	T	C	G	A	C	T	C
200	A	T	C	G	C	G	T	A
19	A	T	G	G	C	C	T	C

Disease characteristics table:
100 virus strains
10 disease characteristics

Sequence ID	Symptoms	Drug Resistance
118	Mild	High
200	Severe	Mild
19	Moderate	High

Domain: task

- Identify substitutions that lead to an increase in different disease characteristic's severity.
- Output for a particular characteristics:

Severity-driven substitutions in the order of importance:

1. G → A, 513 + T → A, 907
2. A → G, 39

Data abstraction: Original dataset

Tabular

Row : strain sequence

Column: position in aligned sequences

Cell attribute: nominal data

Row: strain characteristics info

Column: a characteristics

Cell attribute: ordinal variable

Sequence ID	1	2	3	4	5	6	7	8
118	A	T	C	G	A	C	T	C
200	A	T	C	G	C	G	T	A
19	A	T	G	G	C	C	T	C

Sequence ID	Symptoms	Drug Resistance
118	Mild	High
200	Severe	Mild
19	Moderate	High

Data abstraction: derived dataset

Tabular

Row : strain sequence

Column: position in aligned sequences

Derived cell variable: whether there is a substitution from the first row

Sequence ID	1	2	3	4	5	6	7	8
118	A	T	C	G	A	C	T	C
200	0	0	0	0	1	1	0	1
19	0	0	1	0	1	0	0	0

Row: strain characteristics

Column: a Characteristics

Cell: ordinal variable

Derived characteristics from the originals

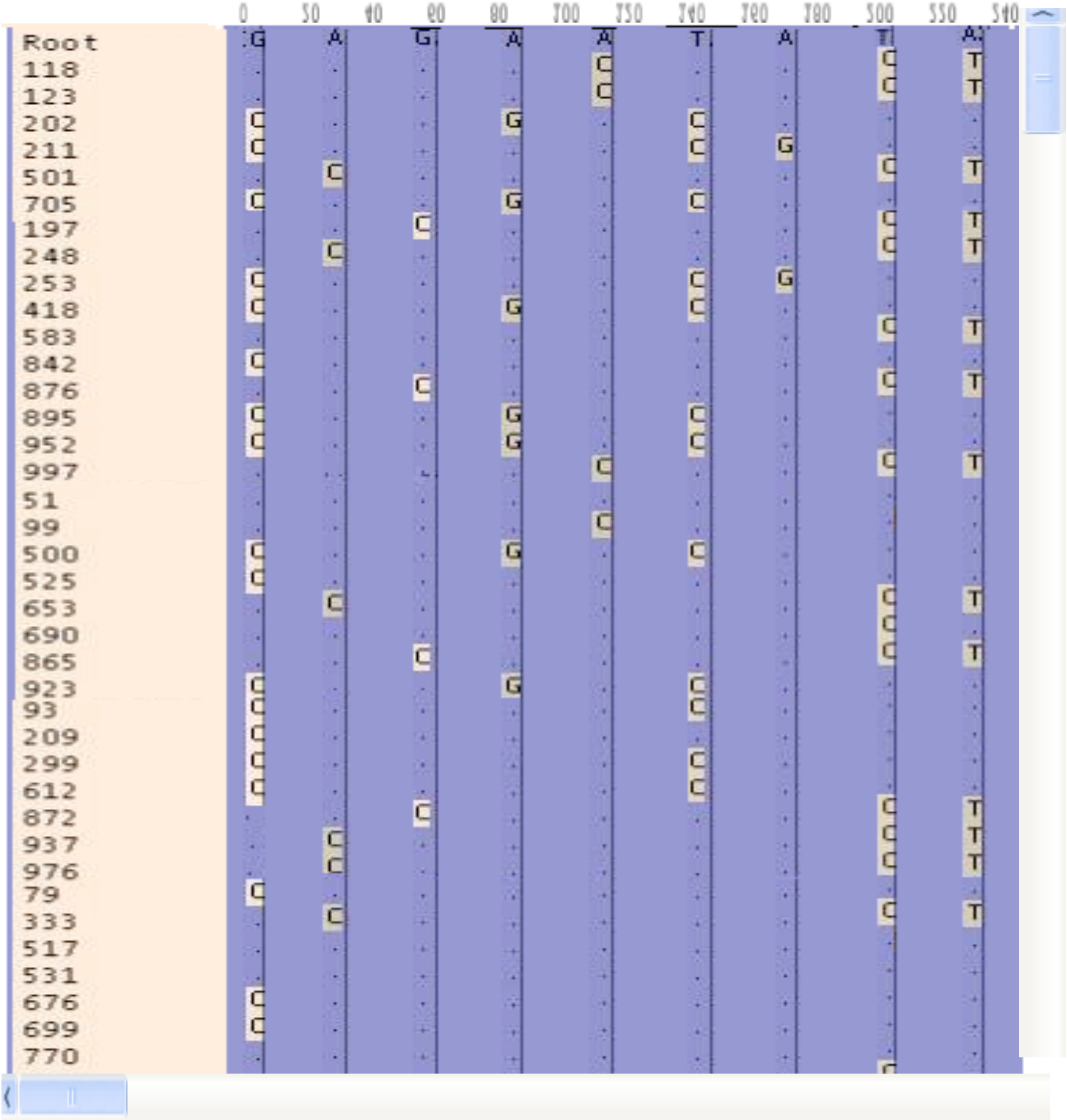
Sequence ID	Symptoms	Drug Resistance
118	Mild	High
200	Severe	Mild
19	Moderate	High

Data: visual encoding

	1	2	3	4	5	6
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	.	.	.	.
197	.	.	.	.	.	.
248	.	.	.	.	C	.
253	.	.	.	.	.	.
418	.	.	C	.	.	.
583	.	.	.	.	C	.

Note:
Color justifications

IMAS multi-alignment view



Root	G	A	G	A	A	T	A	T	A
118	.	.	.	.	C	.	.	C	T
123	.	.	.	.	C	.	.	C	T
202	C	.	.	G	.	C	.	.	.
211	C	.	.	.	.	C	G	.	.
501	.	C	.	.	.	.	.	C	T
705	C	.	.	G	.	C	.	.	.
197	.	.	C	.	.	.	.	C	T
248	.	C	.	.	.	.	.	C	T
253	C	.	.	.	.	C	G	.	.
418	C	.	.	G	.	C	.	.	.
583	.	.	.	.	.	.	.	C	T
842	C	.	.	.	.	.	.	.	.
876	.	.	C	.	.	.	.	C	T
895	C	.	.	G	.	C	.	.	.
952	C	.	.	G	.	C	.	.	.
997	.	.	.	.	C	.	.	C	T
51	.	.	.	.	.	.	.	.	.
99	.	.	.	.	C	.	.	.	.
500	C	.	.	G	.	C	.	.	.
525	C	.	.	.	.	.	.	.	.
653	.	C	.	.	.	.	.	C	T
690	.	.	.	.	.	.	.	C	.
865	.	.	C	.	.	.	.	C	T
923	C	.	.	G	.	C	.	.	.
93	C	.	.	.	.	C	.	.	.
209	C	.	.	.	.	.	.	.	.
299	C	.	.	.	.	C	.	.	.
612	C	.	.	.	.	C	.	.	.
872	.	.	C	.	.	.	.	C	T
937	.	C	.	.	.	.	.	C	T
976	.	C	.	.	.	.	.	C	T
79	C	.	.	.	.	.	.	.	.
333	.	C	.	.	.	.	.	C	T
517	.	.	.	.	.	.	.	.	.
531	.	.	.	.	.	.	.	.	.
676	C	.	.	.	.	.	.	.	.
699	C	.	.	.	.	.	.	.	.
770	.	.	.	.	.	.	.	.	.

Task abstraction

- Find columns that their substitution lead to the most severe rows. Report them individually or in a combination with other columns:

1. Sort the rows
2. Filter the columns until you find the answer:
   - 2.1 Find interesting columns
   - 2.2 Find related columns

     (and filter the rest)

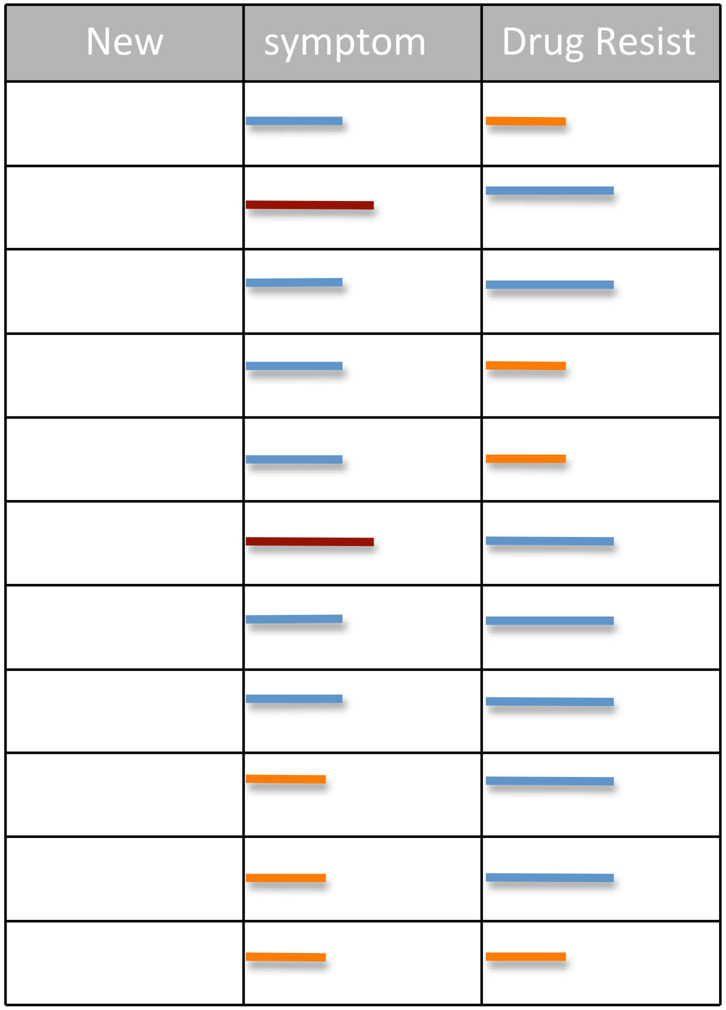
Task abstraction

- Find columns that their substitution lead to the most severe rows. Report them individually or in a combination with other columns:

1. Sort the rows
2. Filter the columns until you find the answer:
    - 2.1 Find interesting columns
    - 2.2 Find related columns

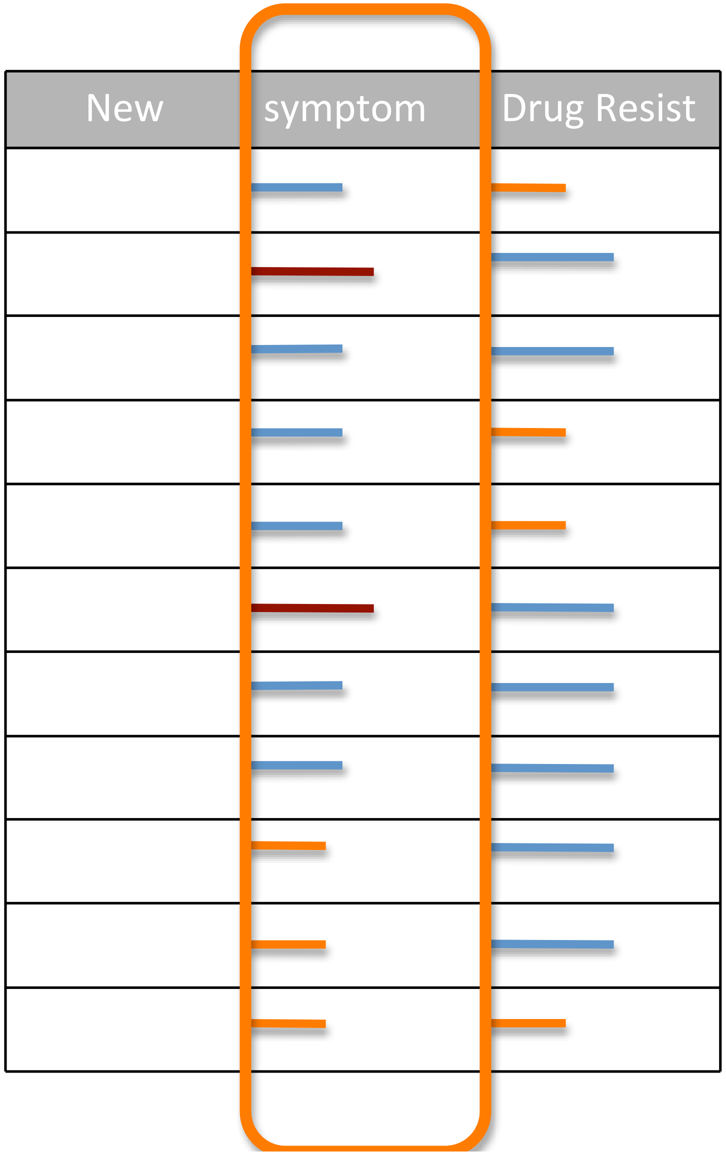
        (and filter the rest)

Dataset/Task visual encoding



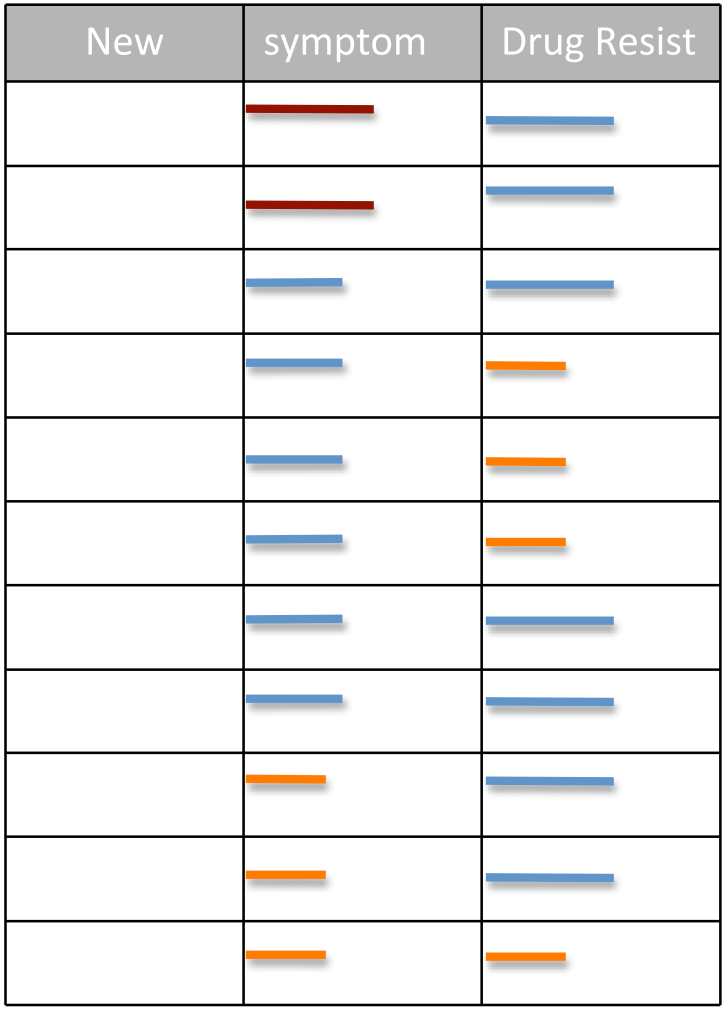
	0	1	2	3	4	5
Labels	A	A	T	A	T	A
705	G	.	C	.	.	.
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	G	.	C	.	.	.
501	.	.	.	.	C	T
118	.	C	.	.	C	T
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	.	.	C	G	.	.
583	.	.	.	.	C	T

“Table Lens” view



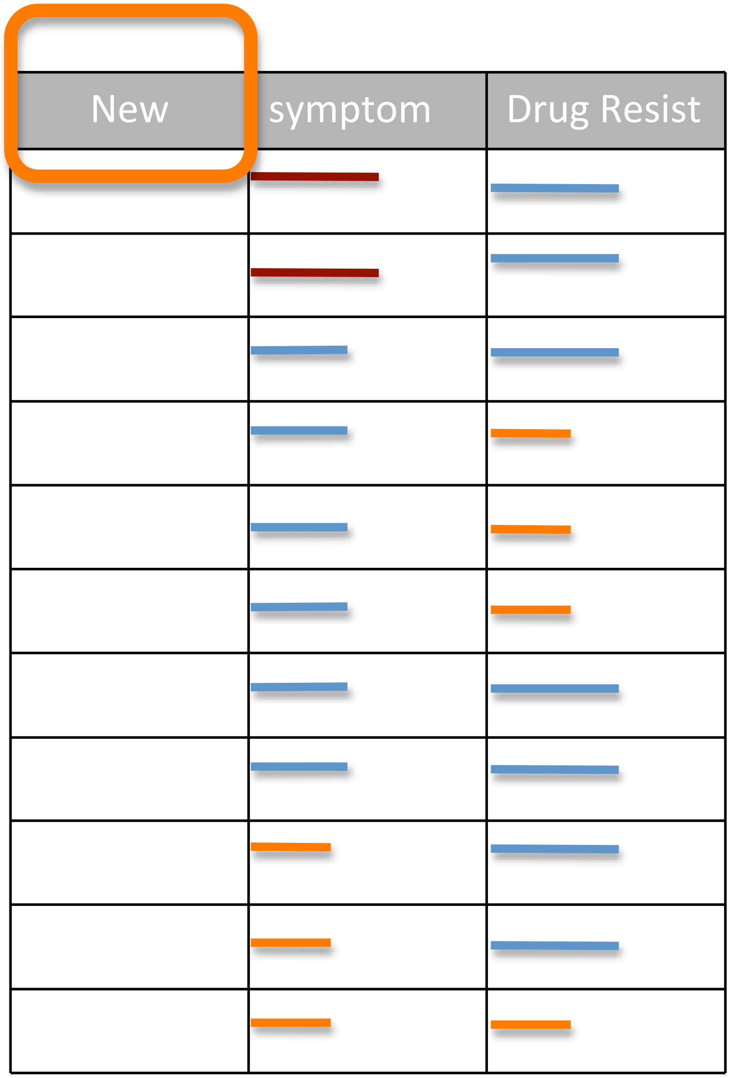
	0	1	2	3	4	5
Labels	A	A	T	A	T	A
705	G	.	C	.	.	.
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	G	.	C	.	.	.
501	.	.	.	.	C	T
118	.	C	.	.	C	T
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	.	.	C	G	.	.
583	.	.	.	.	C	T

Table Lens view



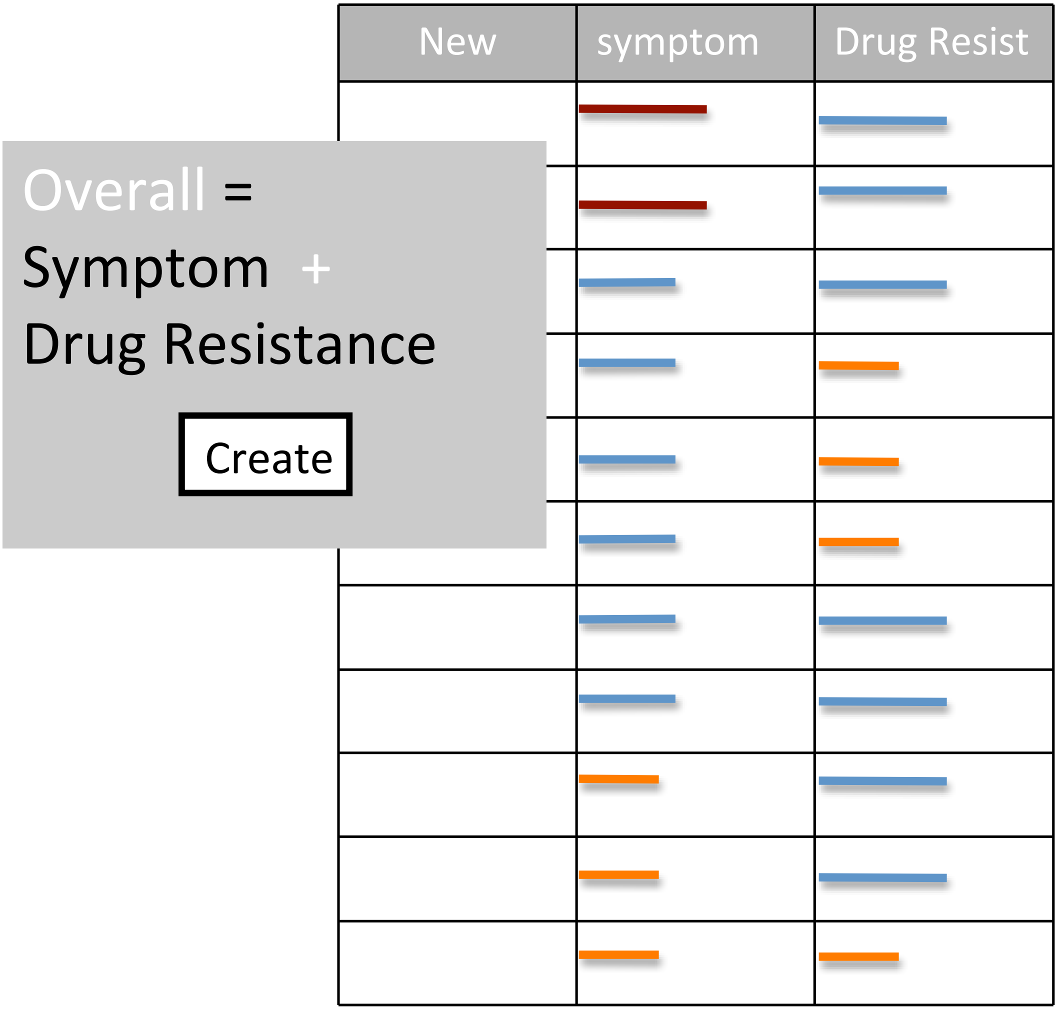
	0	1	2	3	4	5
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T

Table Lens view



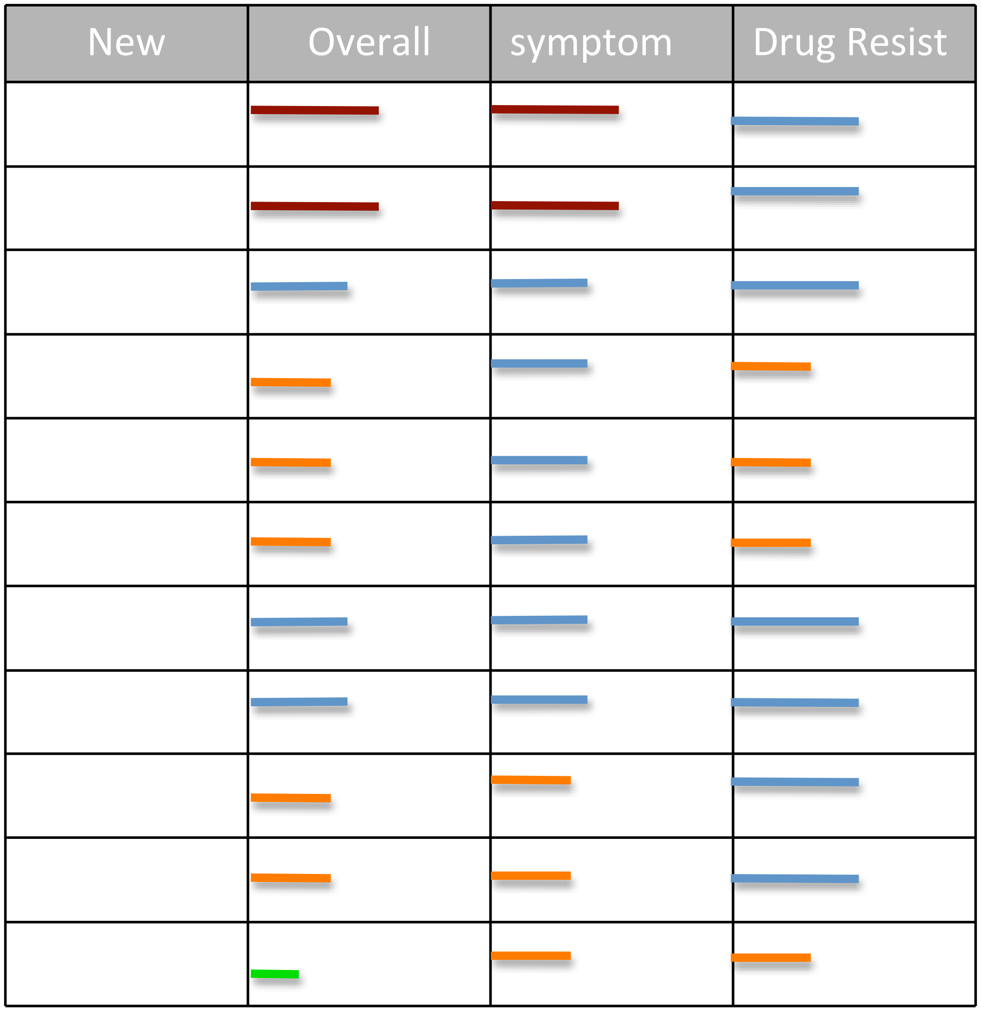
	0	1	2	3	4	5
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T

Table Lens view



	0	1	2	3	4	5
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T

Dataset/Task visual encoding



	0	1	2	3	4	5
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T

Task abstraction

- Find columns that their substitution lead to the most severe rows. Report them individually or in a combination with other columns:

1. Sort the rows ✔

1. Filter the columns until you find the answer:

    2.1 Find interesting columns and/or

    2.2 Find related columns

    (and filter the rest)

Task abstraction

- Find columns that their substitution lead to the most severe rows. Report them individually or in a combination with other columns:

1. Sort the rows ✓
2. Filter the columns until you find the answer:
   - 2.1 Find interesting columns and/or
   - 2.2 Find related columns

     (and filter the rest)

Comparison tasks challenges

- Navigation + Comparison: increases the memory load
- The order of columns should be remained

Interaction design: filter columns

	0	1	2	3	4	5	6	7	8	9	10	11	12
Labels	A	A	T	A	T	A	T	A	C	G	G	G	T
118	.	C	.	.	C	.	.	.	.	.	.	A	.
123	.	C	.	.	C	.	.	.	.	.	.	.	.
202	G	.	C	.	.	.	.	.	.	.	.	A	.
211	.	.	C	G	.	.	.	.	.	.	.	.	.
501	.	.	.	.	C	T	.	.	.	.	T	.	.
705	G	.	C	.	.	.	.	.	.	.	.	.	.
197	.	.	.	.	C	T	.	.	.	.	.	.	C
248	.	.	.	.	C	.	.	.	.	.	T	.	.
253	.	.	C	G	.	.	.		.	.	.	.	.
418	G	.	C	.	.	.	.	.	.	.	.	A	.
583													

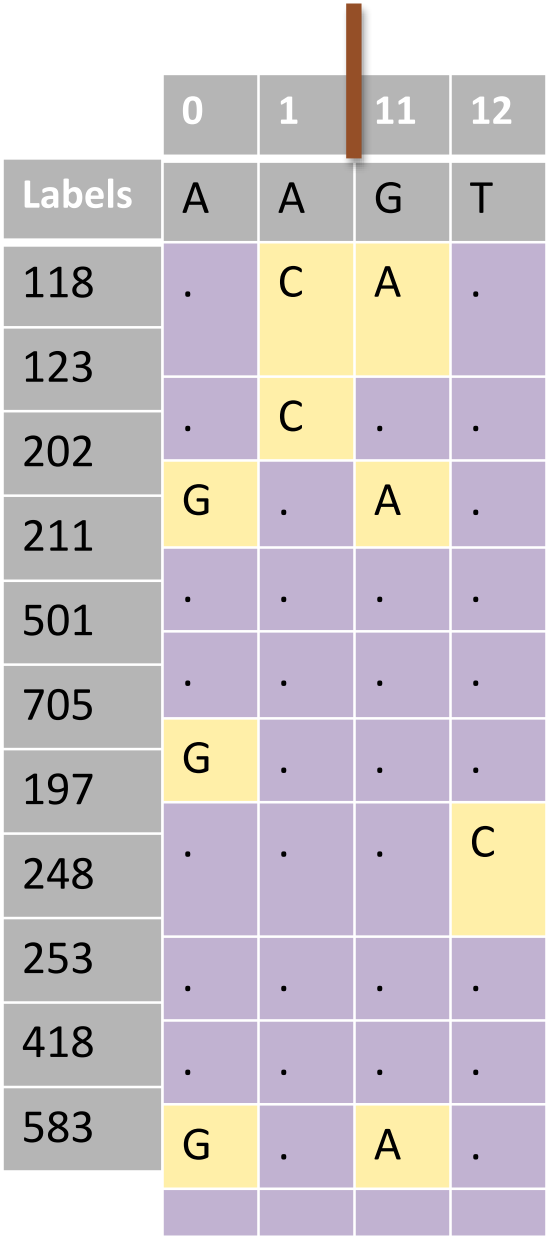
Filter columns: selecting the points

	0	1	2	3	4	5	6	7	8	9	10	11	12
Labels	A	A	T	A	T	A	T	A	C	G	G	G	T
118	.	C	.	.	C	.	.	.	.	.	.	A	.
123	.	C	.	.	C	.	.	.	.	.	.	.	.
202	G	.	C	.	.	.	.	.	.	.	.	A	.
211	.	.	C	G	.	.	.	.	.	.	.	.	.
501	.	.	.	.	C	T	.	.	.	.	T	.	.
705	G	.	C	.	.	.	.	.	.	.	.	.	.
197	.	.	.	.	C	T	.	.	.	.	.	.	C
248	.	.	.	.	C	.	.	.	.	.	T	.	.
253	.	.	C	G	.	.	.		.	.	.	.	.
418	G	.	C	.	.	.	.	.	.	.	.	A	.
583													

Filter columns: selecting the points

	0	1	2	3	4	5	6	7	8	9	10	11	12
Labels	A	A	T	A	T	A	T	A	C	G	G	G	T
118	.	C	.	.	C	.	.	.	.	.	.	A	.
123	.	C	.	.	C	.	.	.	.	.	.	.	.
202	G	.	C	.	.	.	.	.	.	.	.	A	.
211	.	.	C	G	.	.	.	.	.	.	.	.	.
501	.	.	.	.	C	T	.	.	.	.	T	.	.
705	G	.	C	.	.	.	.	.	.	.	.	.	.
197	.	.	.	.	C	T	.	.	.	.	.	.	C
248	.	.	.	.	C	.	.	.	.	.	T	.	.
253	.	.	C	G	.	.	.		.	.	.	.	.
418	G	.	C	.	.	.	.	.	.	.	.	A	.
583													

Hide/Unhide columns



	0	1	11	12
Labels	A	A	G	T
118	.	C	A	.
123	.	C	.	.
202	G	.	A	.
211	.	.	.	.
501	.	.	.	.
705	G	.	.	.
197	.	.	.	C
248				
253	.	.	.	.
418	.	.	.	.
583	G	.	A	.

Note:
The transition needs to be smooth

Redundant columns

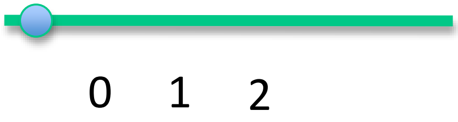
	0	1	2	3	4	5	6	7	8	9	10	11	12
Labels	A	A	T	A	T	A	T	A	C	G	G	G	T
118	.	C	.	.	C	.	.	.	.	.	.	A	.
123	.	C	.	.	C	.	.	.	.	.	.	.	.
202	G	.	C	.	.	.	.	.	.	.	.	A	.
211	.	.	C	G	.	.	.	.	.	.	.	.	.
501	.	.	.	.	C	.	.	.	.	.	T	.	.
705	G	.	C	.	.	.	.	.	.	.	.	.	.
197	.	.	.	.	C	.	.	.	.	.	.	.	C
248	.	.	.	.	C	.	.	.	.	.	T	.	.
253	.	.	C	G	.	.	.		.	.	.	.	.
418	G	.	C	.	.	.	.	.	.	.	.	A	.
583													

Redundant columns

	0	1	2	3	4	5	6	7	8	9	10	11	12
Labels	A	A	T	A	T	A	T	A	C	G	G	G	T
118	.	C	.	.	C	.	.	.	.	.	.	A	.
123	.	C	.	.	C	.	.	.	.	.	.	.	.
202	G	.	C	.	.	.	.	.	.	.	.	A	.
211	.	.	C	G	.	.	.	.	.	.	.	.	.
501	.	.	.	.	C	.	.	.	.	.	T	.	.
705	G	.	C	.	.	.	.	.	.	.	.	.	.
197	.	.	.	.	C	.	.	.	.	.	.	.	C
248	.	.	.	.	C	.	.	.	.	.	T	.	.
253	.	.	C	G	.	.	.		.	.	.	.	.
418	G	.	C	.	.	.	.	.		.	.	A	.
583													

Filtering columns

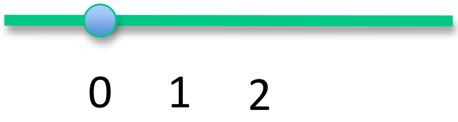
Filter columns



	0	1	2	3	4	5	6	7	8	9	10	11	12
Labels	A	A	T	A	T	A	T	A	C	G	G	G	T
118	.	C	.	.	C	.	.	.	.	.	.	A	.
123	.	C	.	.	C	.	.	.	.	.	.	.	.
202	G	.	C	.	.	.	.	.	.	.	.	A	.
211	.	.	C	G	.	.	.	.	.	.	.	.	.
501	.	.	.	.	C	.	.	.	.	.	T	.	.
705	G	.	C	.	.	.	.	.	.	.	.	.	.
197	.	.	.	.	C	.	.	.	.	.	.	.	C
248	.	.	.	.	C	.	.	.	.	.	T	.	.
253	.	.	C	G	.	.	.		.	.	.	.	.
418	G	.	C	.	.	.	.	.	.	.	.	A	.
583	.	.	.	.	C	T	.	.	.	.	.	.	.

Filtering columns

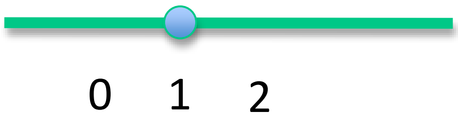
Filter columns



	0	1	2	3	4	5	10	11	12
Labels	A	A	T	A	T	A	G	G	T
118	.	C	.	.	C	.	.	A	.
123	.	C	.	.	C	.	.	.	.
202	G	.	C	.	.	.	.	A	.
211	.	.	C	G	.	.	.	.	.
501	.	.	.	.	C	.	T	.	.
705	G	.	C	.	.	.	.	.	.
197	.	.	.	.	C	.	.	.	C
248	.	.	.	.	C	.	T	.	.
253	.	.	C	G	.	.	.	.	.
418	G	.	C	.	.	.	.	A	.
583	.	.	.	.	C	T	.	.	.

Filtering columns

Filter columns



	0	1	2	3	4	10	11
Labels	A	A	T	A	T	G	G
118	.	C	.	.	C	.	A
123	.	C	.	.	C	.	.
202	G	.	C	.	.	.	A
211	.	.	C	G	.	.	.
501	.	.	.	.	C	T	.
705	G	.	C	.	.	.	.
197	.	.	.	.	C	.	.
248	.	.	.	.	C	T	.
253	.	.	C	G	.	.	.
418	G	.	C	.	.	.	A
583	.	.	.	.	C	.	.

Task abstraction

- Find columns that their substitution lead to the most severe rows. Report them individually or in a combination with other columns:

1. Sort the rows ✓
2. Filter the columns: Basic ✓

    2.1 Find interesting columns

    2.2 Find related columns

    (and filter the rest)

A pattern within a column

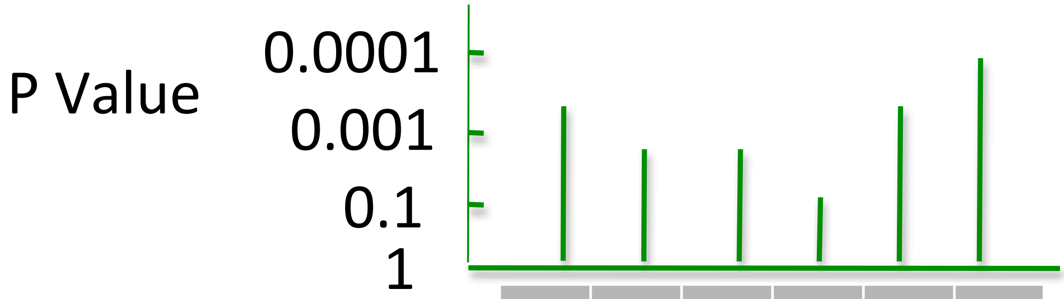
		12
	Labels	A
Severe	118	C
	123	C
	202	.
Moderate	211	.
	501	C
	705	.
	197	.
Mild	248	.
	253	.
	418	.
	583	.

Detecting the pattern:

- Challenging for humans
- Derived data for each column:

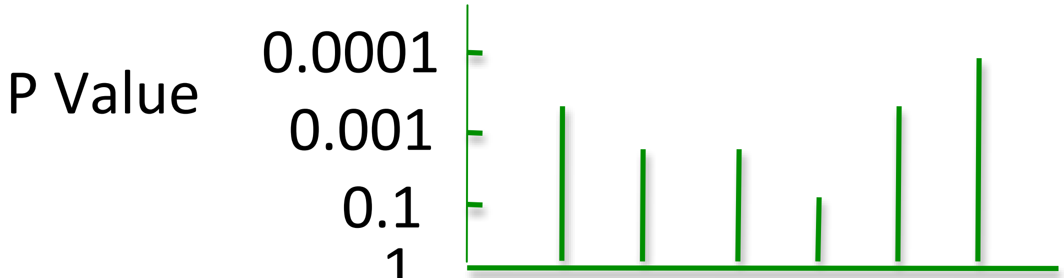
Statistical significance score in the Mann-Whitney U test (the P-value)

Visual encoding of the P-value

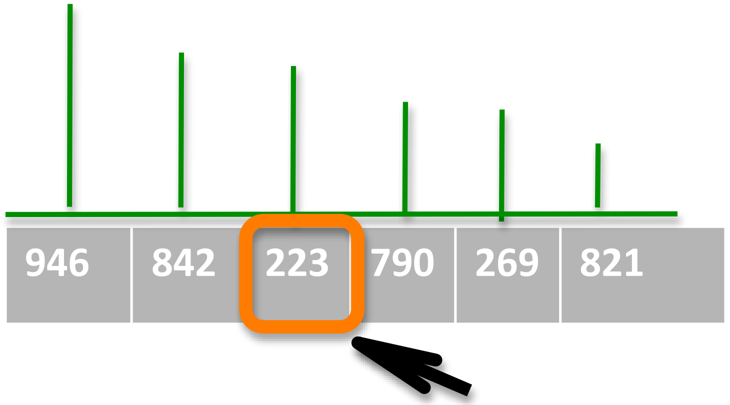


	223	269	790	821	842	946
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.

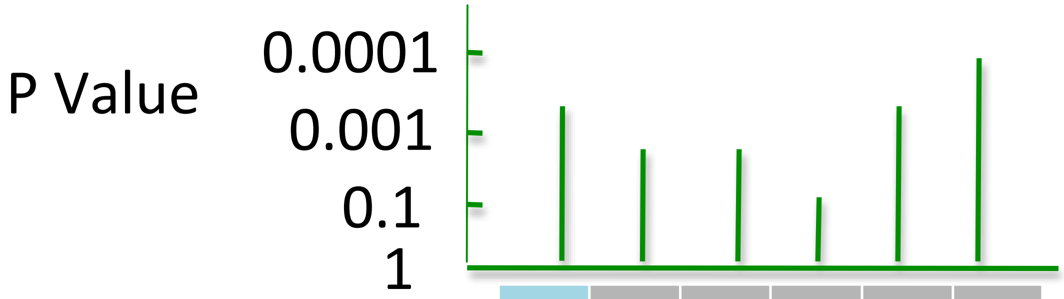
Sorting/Filtering columns by the P-value



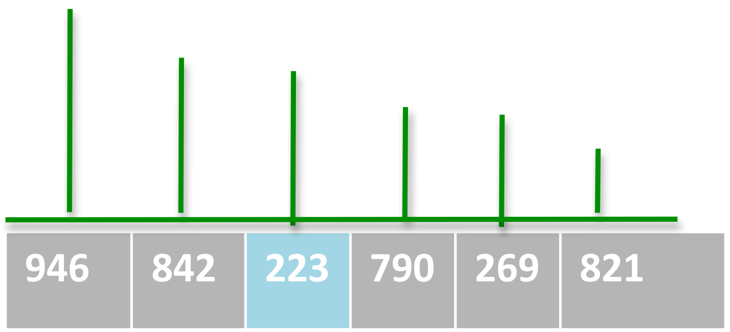
	223	269	790	821	842	946
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.



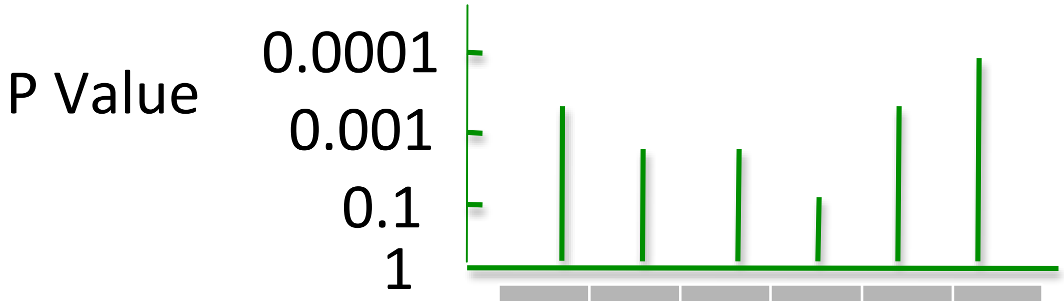
Sorting/Filtering columns by the P-value



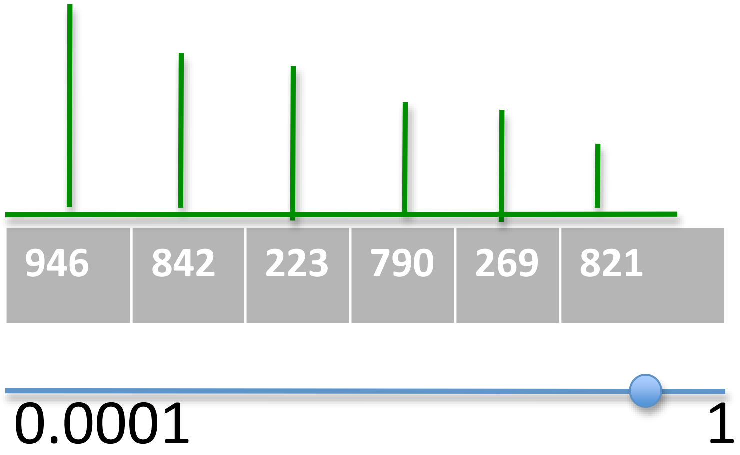
	223	269	790	821	842	946
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.



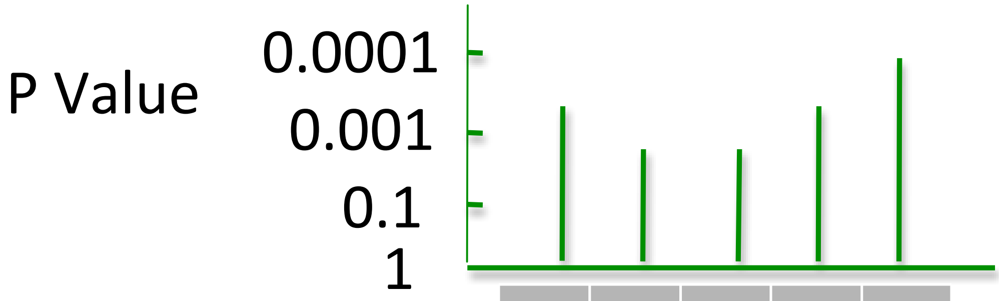
Sorting/Filtering columns by the P-value



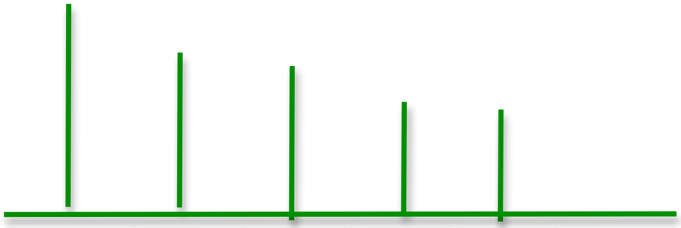
	223	269	790	821	842	946
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.



Sorting/Filtering columns by the P-value



	223	269	790	842	946
Labels	A	A	T	T	A
118	.	C	.	C	T
123	.	C	.	C	T
202	G	.	C	.	.
211	.	.	C	.	.
501	.	.	.	C	T
705	G	.	C	.	.
197	.	.	.	C	T
248	.	.	.	C	T
253	.	.	C	.	.



946	842	223	790	269

Task abstraction

- Find columns that their substitution lead to the most severe rows. Report them individually or in a combination with other columns:

1. Sort the rows ✓
2. Filter the columns Basic ✓

   2.1 Find interesting columns ✓

   2.2 Find related columns

   (and filter the rest)

Correlation pattern between the columns

	4	5
Labels	T	A
118	C	T
123	C	T
202	.	.
211	.	.
501	C	T
705	.	.
197	C	T
248	C	T
253	.	.
418	.	.
583	C	T

Complementary pattern between the columns

	2	5
Labels	T	A
118	.	T
123	.	T
202	C	.
211	C	.
501	.	T
705	C	.
197	.	T
248	.	T
253	C	.
418	C	.
583	.	T

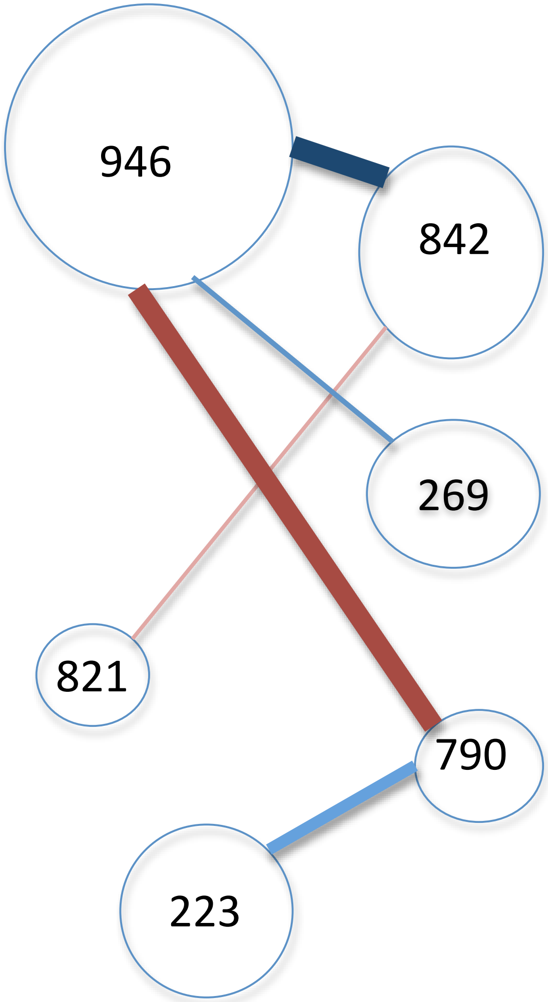
Derived attribute between any pair of columns

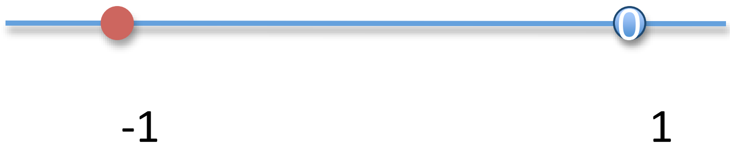
- Correlation: shows how much two columns are dependant

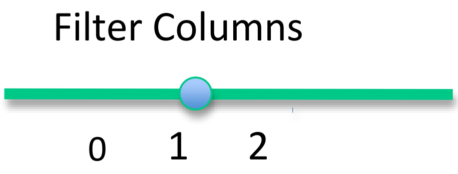
$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

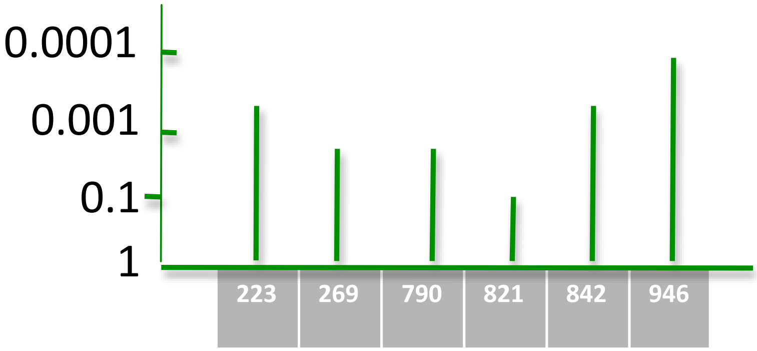
- A relationship between any two columns
- +1: highly correlated
- -1: highly complement

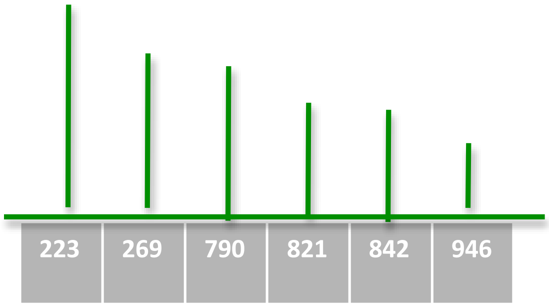
Visual encoding design of the correlation





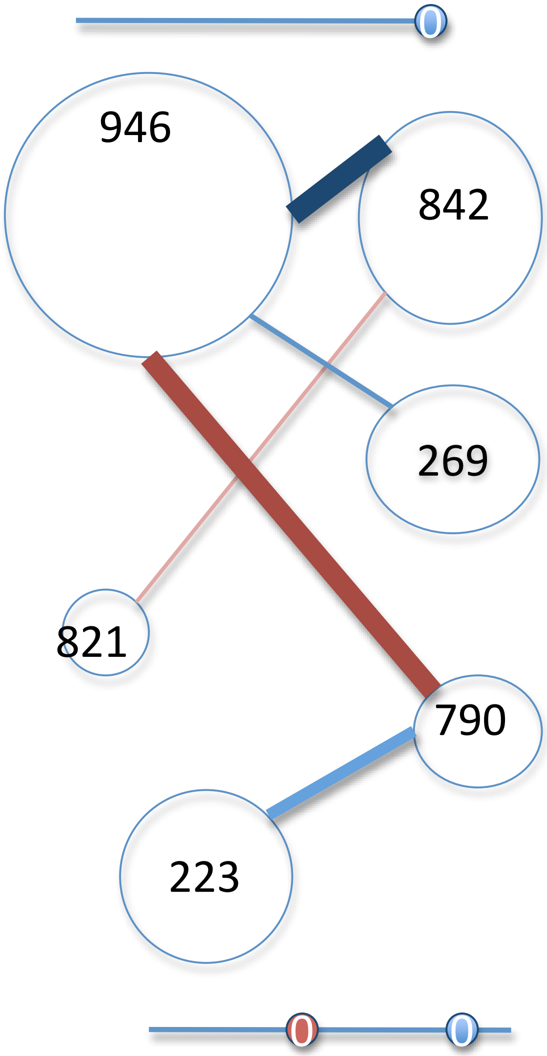


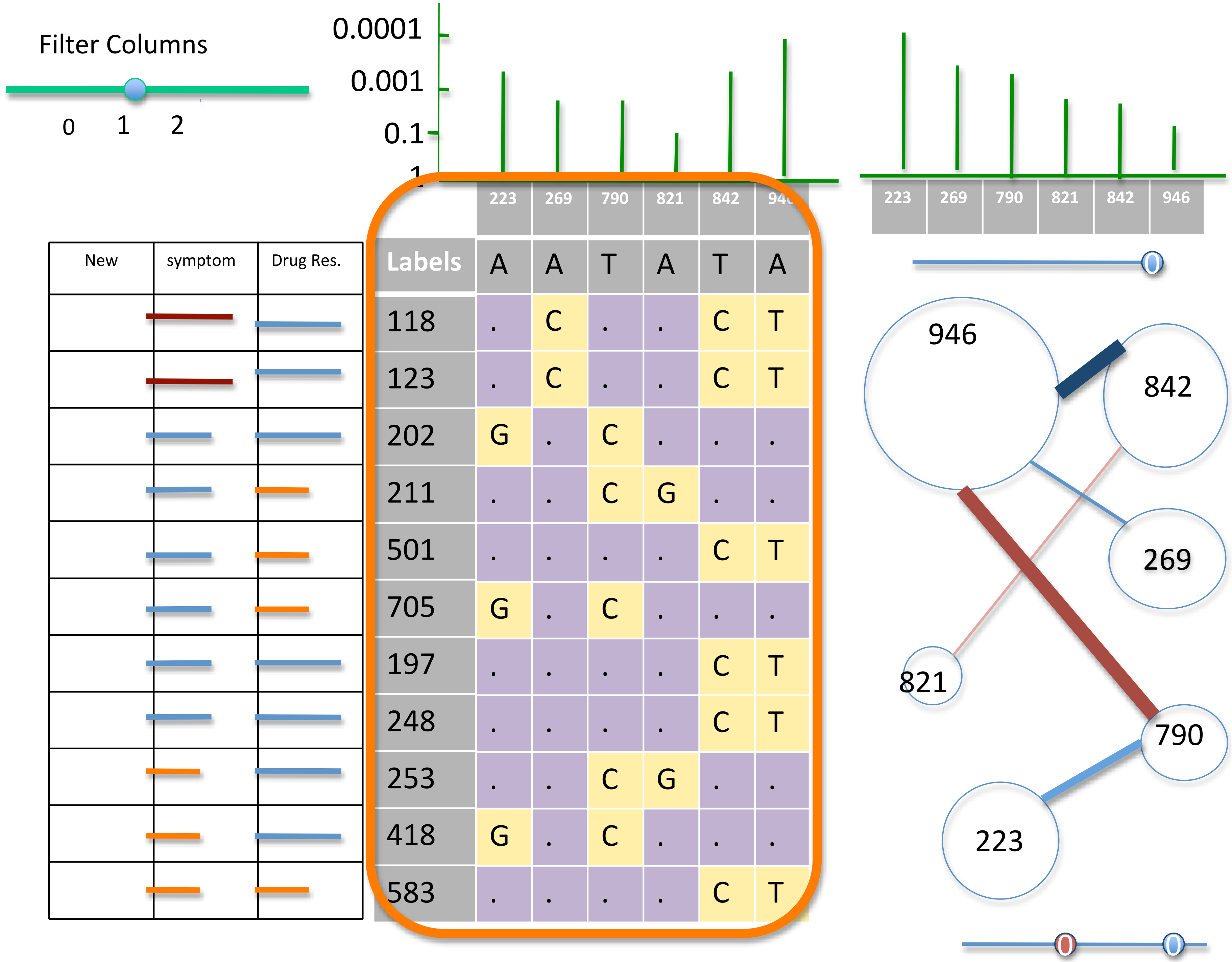




New	symptom	Drug Res.
	🟥	🟦
	🟥	🟦
	🟦	🟦
	🟦	🟧
	🟦	🟧
	🟦	🟧
	🟦	🟦
	🟦	🟦
	🟧	🟦
	🟧	🟦
	🟧	🟧

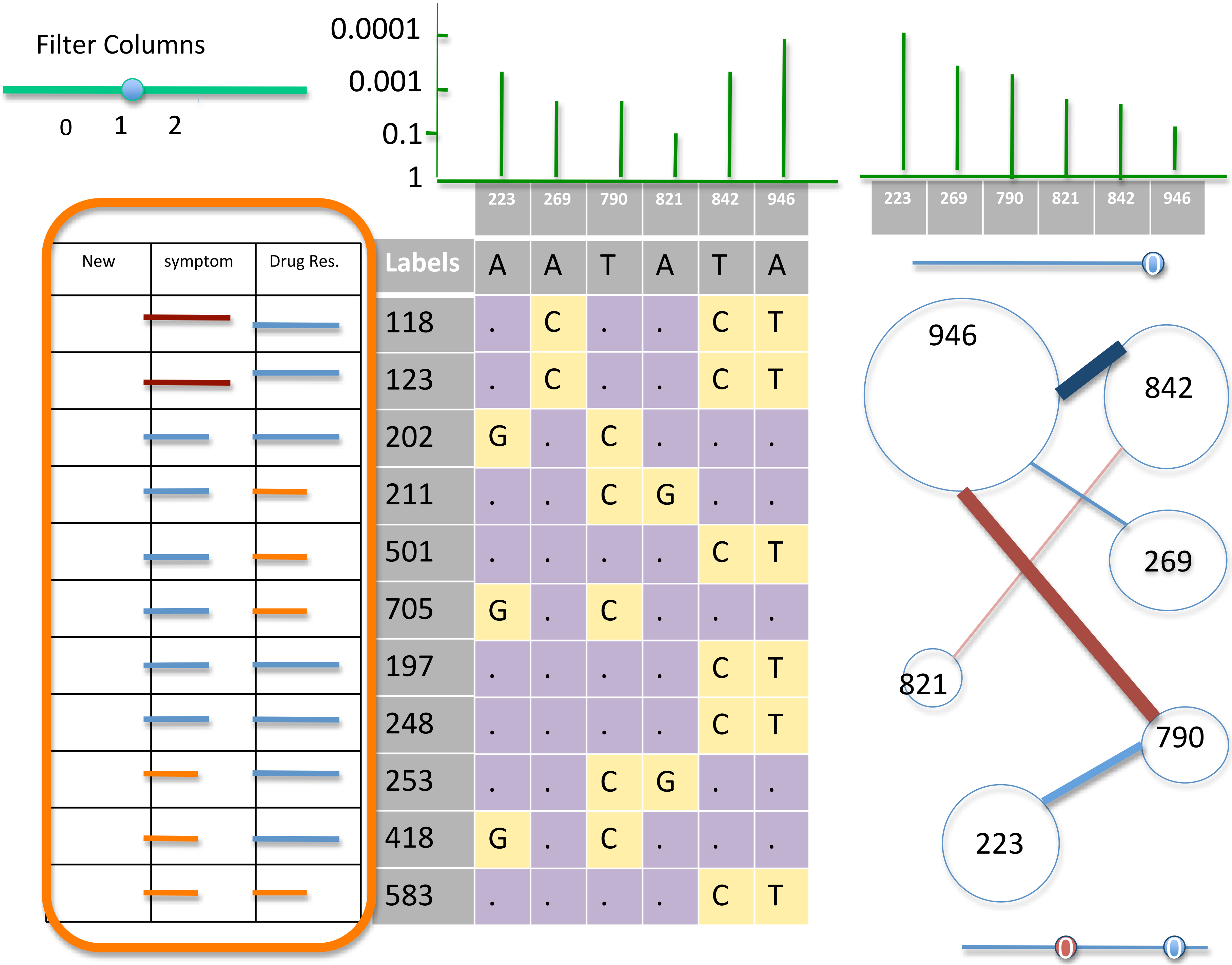
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T



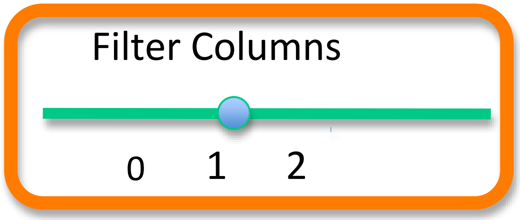


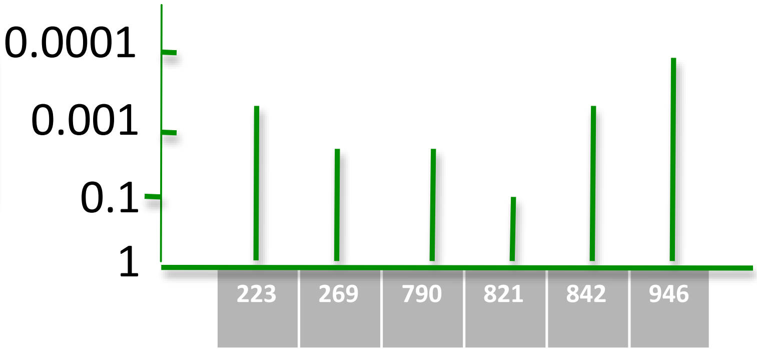
New	symptom	Drug Res.
	🟥	🟦
	🟥	🟦
	🟦	🟦
	🟦	🟧
	🟦	🟧
	🟦	🟧
	🟦	🟦
	🟦	🟦
	🟧	🟦
	🟧	🟦
	🟧	🟧

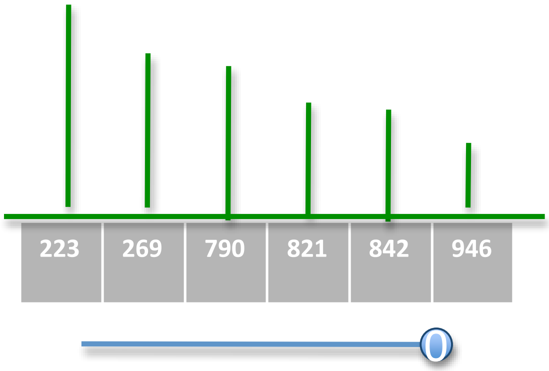
	223	269	790	821	842	94(
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T

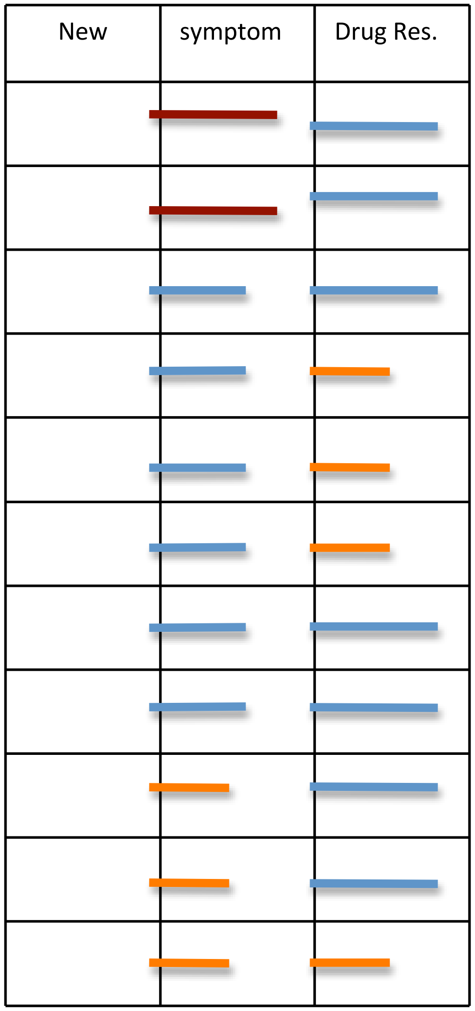


	223	269	790	821	842	946
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T

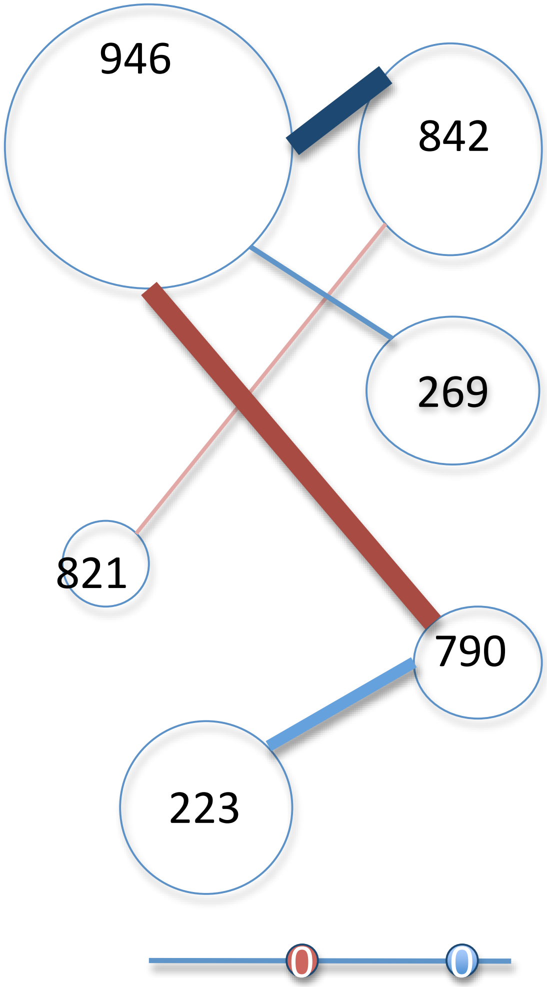


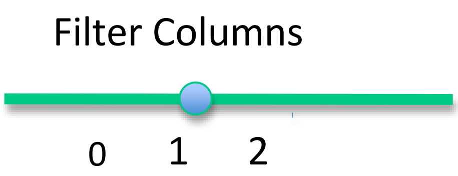


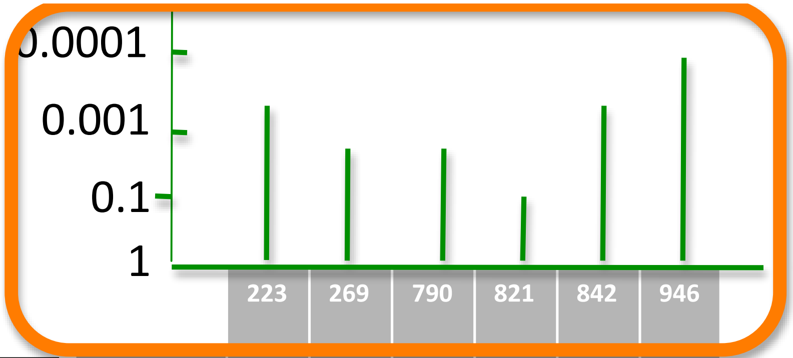


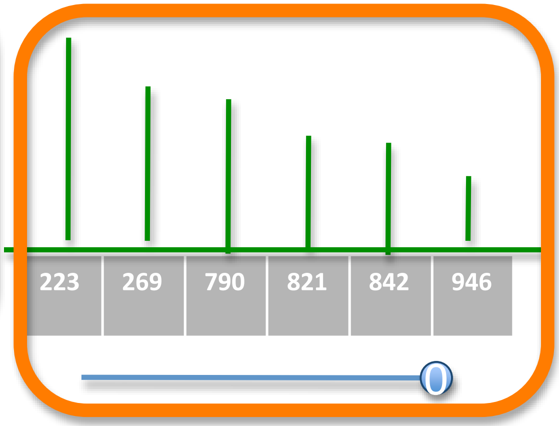


Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T

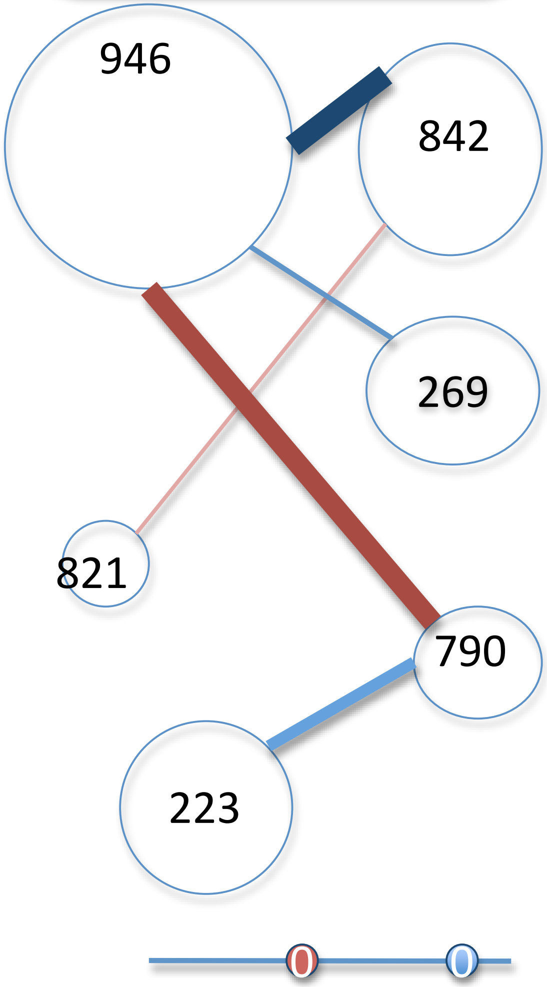


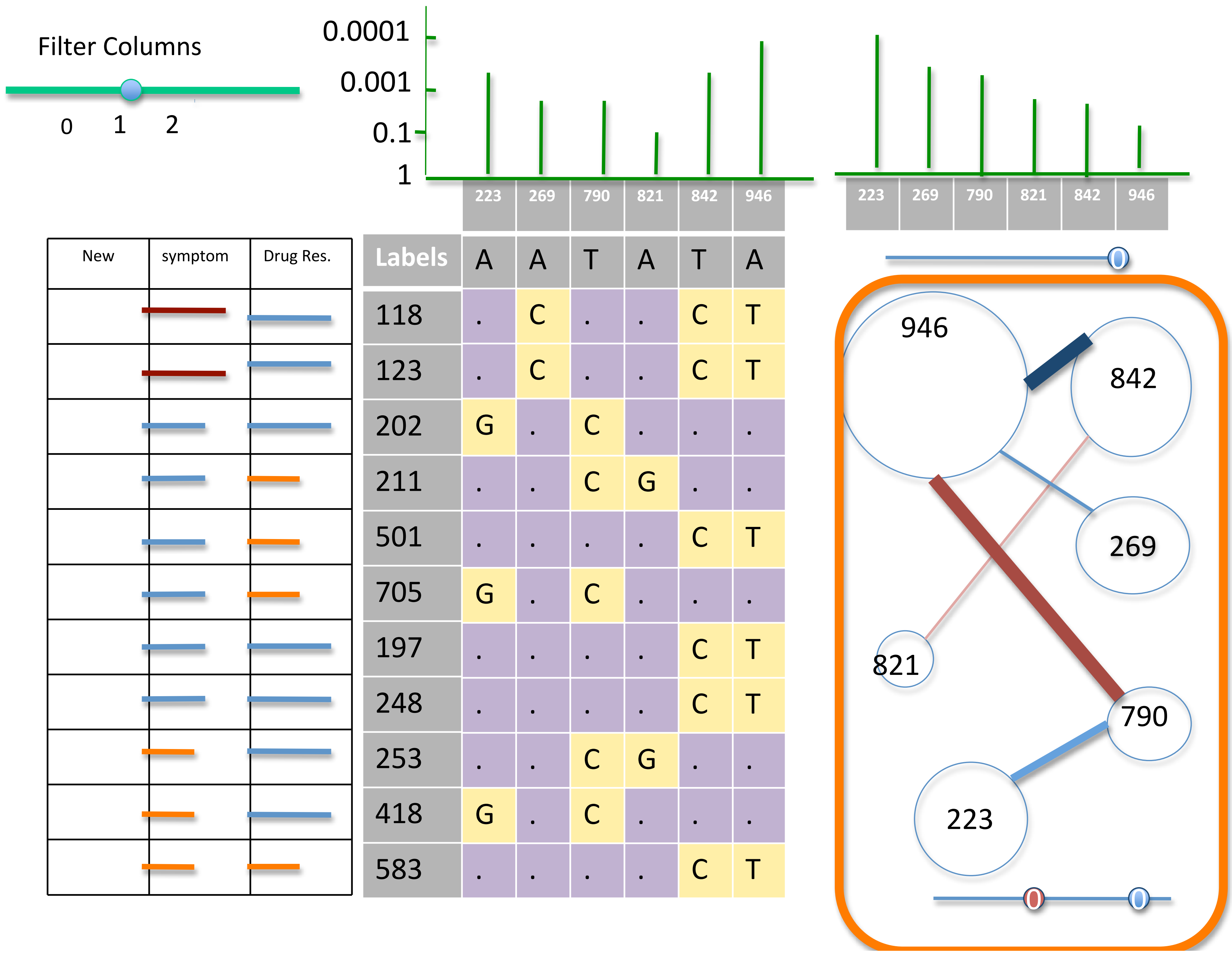






				223	269	790	821	842	946
New	symptom	Drug Res.	Labels	A	A	T	A	T	A
	🟥	🟦	118	.	C	.	.	C	T
	🟥	🟦	123	.	C	.	.	C	T
	🟦	🟦	202	G	.	C	.	.	.
	🟦	🟧	211	.	.	C	G	.	.
	🟦	🟧	501	.	.	.	.	C	T
	🟦	🟧	705	G	.	C	.	.	.
	🟦	🟦	197	.	.	.	.	C	T
	🟦	🟦	248	.	.	.	.	C	T
	🟧	🟦	253	.	.	C	G	.	.
	🟧	🟦	418	G	.	C	.	.	.
	🟧	🟧	583	.	.	.	.	C	T





	223	269	790	821	842	946
Labels	A	A	T	A	T	A
118	.	C	.	.	C	T
123	.	C	.	.	C	T
202	G	.	C	.	.	.
211	.	.	C	G	.	.
501	.	.	.	.	C	T
705	G	.	C	.	.	.
197	.	.	.	.	C	T
248	.	.	.	.	C	T
253	.	.	C	G	.	.
418	G	.	C	.	.	.
583	.	.	.	.	C	T

The next step

- Finalizing the ideas
- Implementation with Processing or Protovis
- Qualitatively testing on the real domain experts

Question and Feedbacks