Monte Carlo Feature Selection and Rough Sets - A New Approach to Combinatorial Modeling in Systems Biology

Jan Komorowski

ICM, Uppsala University
and
ICM*, University of Warsaw

Monte Carlo Feature Selection and Rough Sets - A New Approach to Combinatorial Modeling in Systems Biology

Jan Komorowski

ICM, Uppsala University
and
ICM*, University of Warsaw

*Interdisciplinary Centre for Mathematical and Computational Modelling

Machine Learning in Bioinformatics and Systems Biology - the CSc perspective

- A well-known paradigm
- Traditionally:
  - usually a small number of cases
  - ability to discern between decision (outcome) classes
  - quality of classification
- Current and forthcoming:
  - very large number of attributes (variables) and ill-defined systems (attributes >> cases)
  - structure of the classifier: which variables and possibly in which order of significance (ranking)
  - local classifiers (no high quality global classifiers)

Give me...

- ... the most significantly expressed X
  - gene
  - protein
  - binding
  - etc
- but biology is not one parameter science!

What the life scientist needs and expects

- Changing the focus in biological research:
  - from single to interacting variables (features, attributes)
  - from analytical models (lines, hyperplanes) to descriptive rule models
- Methodology:
  - Monte Carlo feature selection
  - Rule-based learning - the rough set approach
  - Selection of interacting variables
  - Visualization and interpretation
- Examples of questions
  - which histone modifications and in what combinations associate with exon expression
  - which sequences are cleavable by a protease
  - which mutations of RT play a significant role in drug resistance
- Implications and applications

The setting

- Classification systems (decision tables) where the number of features is >> than the number of objects:
  - gene expression profiles of 50 cancer samples - benign and malicious; 1000 genes will be changing expression levels; 1000 >> 50
  - 500 sequences of RT and the clinical outcome on drug resistance; each sequence has 590 aa’s, each with 7 physico-chemical properties; 4130 >> 500
- Such systems are often ill-defined and most approaches will not work
  - one may discover artifacts in the data, not valid relationships

Rough sets -
an approach to approximate modeling

Z. Pawlak

Rough Set in Gene Expression Analysis

Gene	Tissue1 (T1)	Tissue 2 (T2)	Tissue 3 (T3)	Process
g_1	+	+	+	**A**
g_2	+	0	-	**B**
g_3	-	+	+	**B**
g_4	0	+	-	**A**
g_5	0	+	-	**B**
g_6	+	+	+	**A**
g_7	+	-	0	**A**
g_8	-	-	+	**B**

Rough Set in Gene Expression Analysis

Gene	Tissue1 (T1)	Tissue 2 (T2)	Tissue 3 (T3)	Process
g_1	+	+	+	**A**
g_2	+	0	-	**B**
g_3	-	+	+	**B**
g_4	0	+	-	**A**
g_5	0	+	-	**B**
g_6	+	+	+	**A**
g_7	+	-	0	**A**
g_8	-	-	+	**B**

Equivalence classes:

$\{\boldsymbol{g}_{1,}\boldsymbol{g}_6\}$, $\{\boldsymbol{g}_2\}$, $\{\boldsymbol{g}_3\}$, $\{\boldsymbol{g}_{4,}\boldsymbol{g}_5\}$, $\{\boldsymbol{g}_7\}$, $\{\boldsymbol{g}_8\}$

Decision classes:

$\{\boldsymbol{g}_{1,}\boldsymbol{g}_4, \boldsymbol{g}_6, \boldsymbol{g}_7\}_{\mathrm{A}}$, $\{\boldsymbol{g}_2, \boldsymbol{g}_3, \boldsymbol{g}_5, \boldsymbol{g}_8\}_{\mathrm{B}}$

Rough Set in Gene Expression Analysis

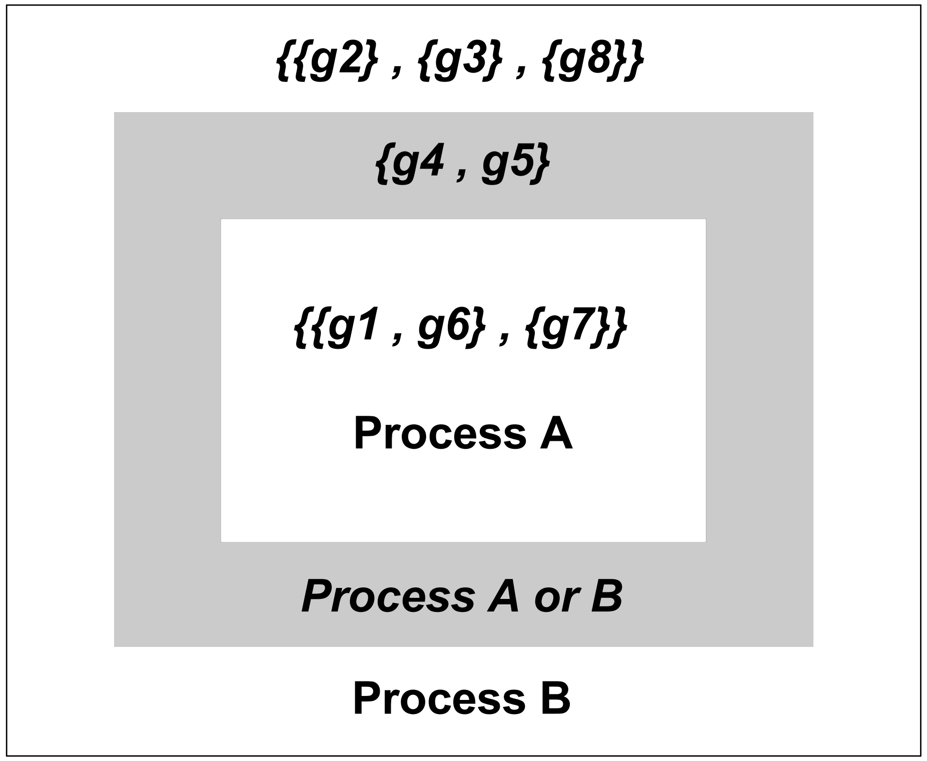
Gene	Tissue1 (T1)	Tissue 2 (T2)	Tissue 3 (T3)	Process
g_1	+	+	+	**A**
g_2	+	0	-	**B**
g_3	-	+	+	**B**
g_4	0	+	-	**A**
g_5	0	+	-	**B**
g_6	+	+	+	**A**
g_7	+	-	0	**A**
g_8	-	-	+	**B**

Equivalence classes:

$\{g_1, g_6\}$, $\{g_2\}$, $\{g_3\}$, $\{g_4, g_5\}$, $\{g_7\}$, $\{g_8\}$

Decision classes:

$\{g_1, g_4, g_6, g_7\}_A$, $\{g_2, g_3, g_5, g_8\}_B$



{{g2} , {g3} , {g8}}

{g4 , g5}

{{g1 , g6} , {g7}}

Process A

Process A or B

Process B

Discernibility matrix modulo decision:

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
g_1	∅							
g_2	T2,T3	∅						
g_3	T1,T3	∅	∅					
g_4	∅	T1,T2,T3	T1,T3	∅				
g_5	T1,T3	∅	∅	∅	∅			
g_6	∅	T2,T3	T1	∅	T1,T3	∅		
g_7	∅	T2,T3	T1,T2	∅	T1,T2,T3	∅	∅	
g_8	T1,T2	∅	∅	T1,T2,T3	∅	T1,T2	T1,T3	∅

Discernibility matrix modulo decision:

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
g_1	∅							
g_2	T2,T3	∅						
g_3	T1,T3	∅	∅					
g_4	∅	T1,T2,T3	T1,T3	∅				
g_5	T1,T3	∅	∅	∅	∅			
g_6	∅	T2,T3	T1	∅	T1,T3	∅		
g_7	∅	T2,T3	T1,T2	∅	T1,T2,T3	∅	∅	
g_8	T1,T2	∅	∅	T1,T2,T3	∅	T1,T2	T1,T3	∅

Discernibility function modulo decision:

f(T1,T2,T3) = (T2,T3)(T1,T3)(T1,T3)(T1,T2)

(T1,T2,T3)(T2,T3)(T2,T3)

(T1,T3)(T1)(T1,T2)

(T1,T2,T3)

(T1,T3)(T1,T2,T3)

(T1,T3)

(T1,T3)

This is a Boolean CD formula that may be simplified

Reduct or prime implicant: {Tissue 1, Tissue 2} , {Tissue1, Tissue 3}

Gene	Tissue1 (T1)	Tissue 2 (T2)	Tissue 3 (T3)	Process
g_1	+	+	+	**A**
g_2	+	0	-	**B**
g_3	-	+	+	**B**
g_4	0	+	-	**A**
g_5	0	+	-	**B**
g_6	+	+	+	**A**
g_7	+	-	0	**A**
g_8	-	-	+	**B**

Gene	**Tissue1 (T1)**	**Tissue 2 (T2)**	**Tissue 3 (T3)**	**Process**
g_1	+	+	+	**A**
g_2	+	0	-	**B**
g_3	-	+	+	**B**
g_4	0	+	-	**A**
g_5	0	+	-	**B**
g_6	+	+	+	**A**
g_7	+	-	0	**A**
g_8	-	-	+	**B**

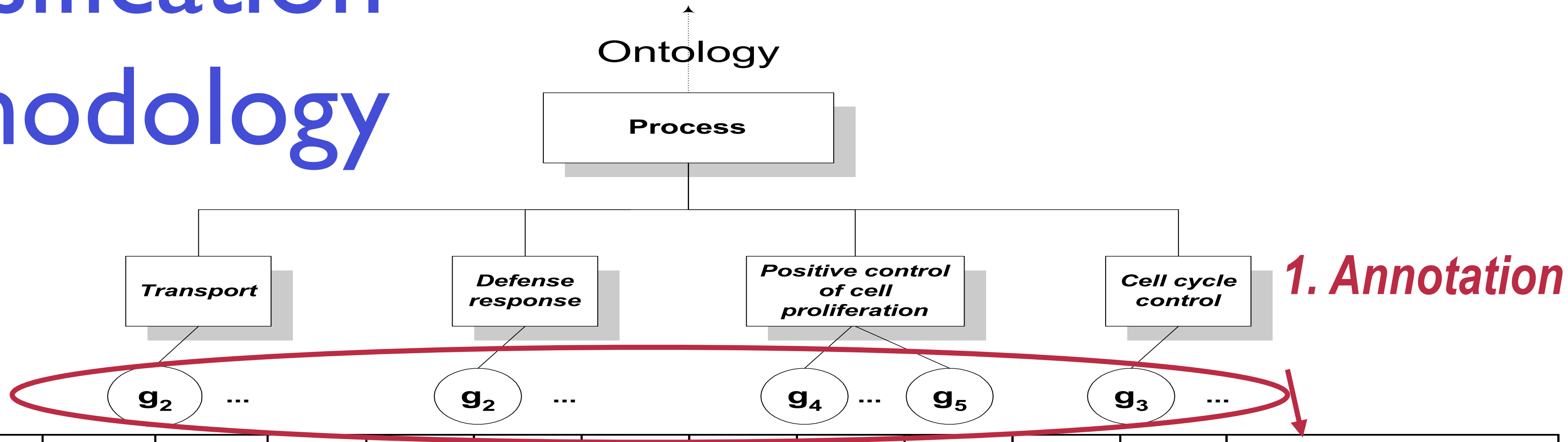
Reduct: {Tissue 1, Tissue 2} , {Tissue1, Tissue 3}

Gene	Tissue1 (T1)	Tissue 2 (T2)	Tissue 3 (T3)	Process
g_1	+	+	+	**A**
g_2	+	0	-	**B**
g_3	-	+	+	**B**
g_4	0	+	-	**A**
g_5	0	+	-	**B**
g_6	+	+	+	**A**
g_7	+	-	0	**A**
g_8	-	-	+	**B**

Reduct: {Tissue 1, Tissue 2} , {Tissue1, Tissue 3}

Rules: Tissue1(+) AND Tissue 2(0) => Process(B)
Tissue1(-) AND Tissue 2(+) => Process(B)
Tissue1(0) AND Tissue 2(+) => Process(A) OR Process(B)
Tissue1(+) AND Tissue 2(+) => Process(A)
Tissue1(+) AND Tissue 2(-) => Process(A)
Tissue1(-) AND Tissue 2(-) => Process(B)
Tissue1(+) AND Tissue 3(-) => Process(B)
Tissue1(-) AND Tissue 3(+) => Process(B)
Tissue1(0) AND Tissue 3(-) => Process(A) OR Process(B)
Tissue1(+) AND Tissue 3(+) => Process(A)
Tissue1(+) AND Tissue 3(0) => Process(A)

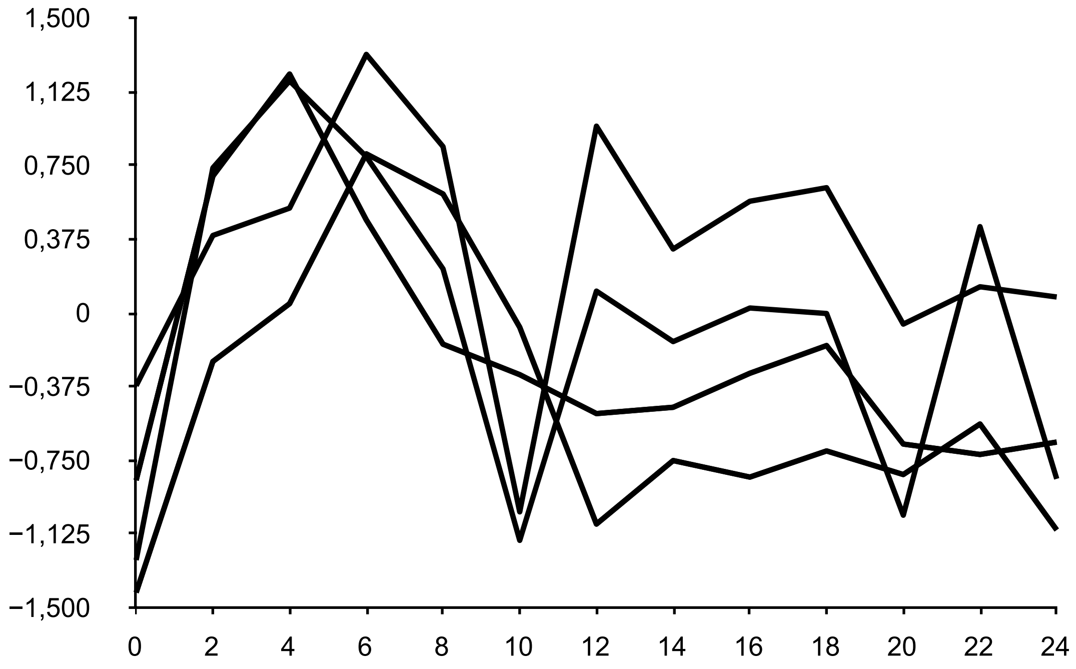
Classification Methodology



1. Annotation

Gene	0HR	15MIN	30MIN	1HR	2HR	4HR	6HR	8HR	12HR	16HR	20HR	24HR	Process
g_1	0.00	-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84	-1.00	-0.60	-0.94	Unknown
g_2	0.00	0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29	-0.15	-0.45	-0.42	Transport and defense response
g_3	0.00	0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38	-0.49	-0.81	-1.12	Cell cycle control
g_4	0.00	-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09	-0.71	-0.76	-0.62	Positive control of cell proliferation
g_5	0.00	0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58	-0.79	-0.29	-0.74	Positive control of cell proliferation
...	...	...	...	...	...	...	...	...	...	...	...	...	...

2. Extracting features for learning



3. Inducing minimal decision rules using rough sets

0 - 4(Increasing) AND 6 - 10(Decreasing) AND 14 - 18(Constant) => GO(cell proliferation)

4. The function of uncharacterized genes is predicted using the rules

!

A. Lægreid, et al Genome Res. 2003 May;13(5):965-79

Rules are generative

- Keep the original coordinates and have no projections:

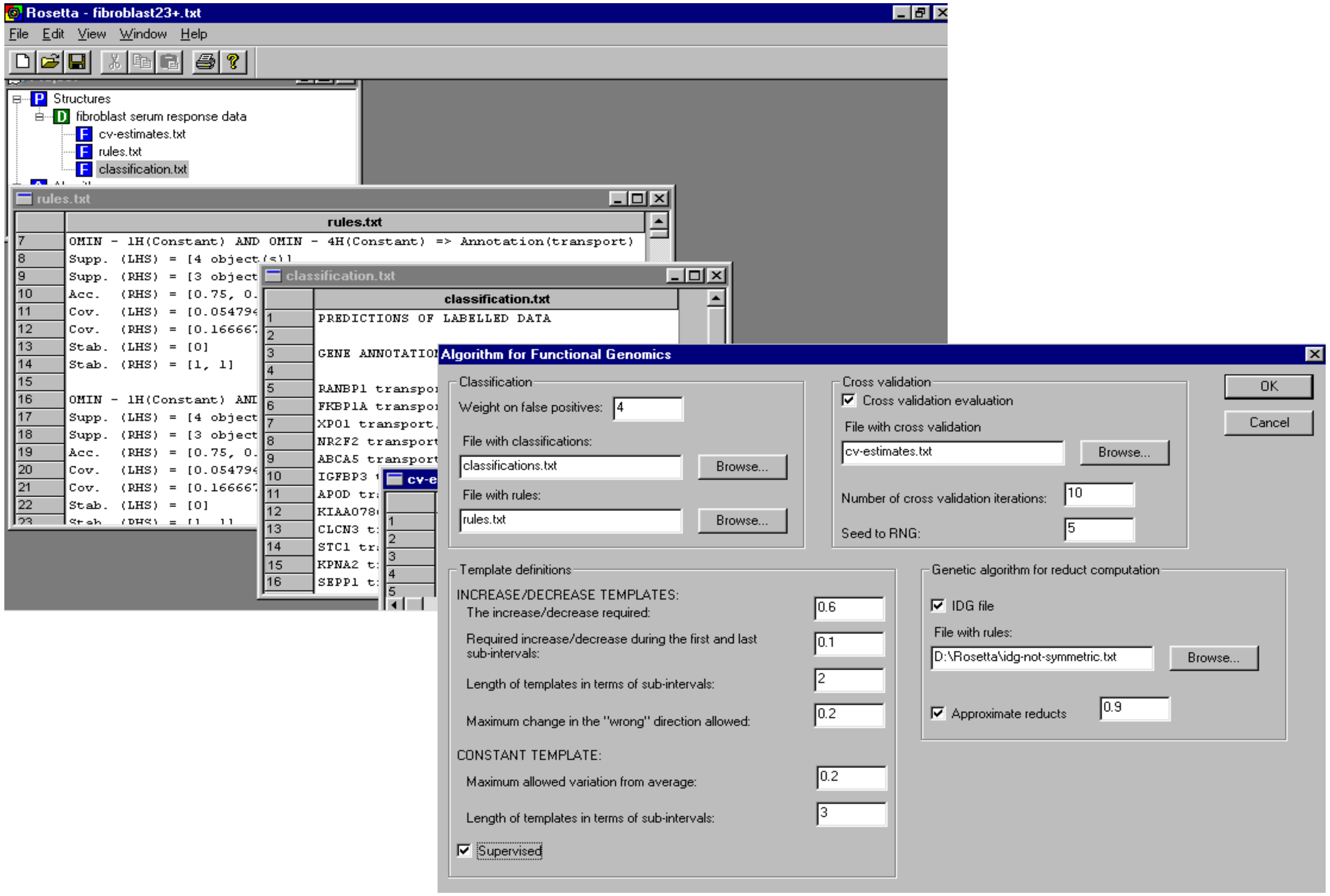
IF (P101 polarity(-inf, 2.1)) AND ... THEN resistant

IF (P101 (D or E or H or K or N or Q or R)) AND ... THEN resistant

And other operations using the Rosetta rough set system

- Rule training
- Cross validation
- Randomization tests, all this in the Rosetta system

http://www.lcb.uu.se/tools/rosetta/



rules.txt

```
7   OMIN - 1H(Constant) AND OMIN - 4H(Constant) => Annotation(transport)
8   Supp. (LHS) = [4 object(s)]
9   Supp. (RHS) = [3 object
10  Acc.  (RHS) = [0.75, 0.
11  Cov.  (LHS) = [0.054794
12  Cov.  (RHS) = [0.166667
13  Stab. (LHS) = [0]
14  Stab. (RHS) = [1, 1]
15
16  OMIN - 1H(Constant) ANI
17  Supp. (LHS) = [4 object
18  Supp. (RHS) = [3 object
19  Acc.  (RHS) = [0.75, 0.
20  Cov.  (LHS) = [0.054794
21  Cov.  (RHS) = [0.166667
22  Stab. (LHS) = [0]
```

classification.txt

```
1   PREDICTIONS OF LABELLED DATA
2
3   GENE ANNOTATIOI
4
5   RANBP1 transpo
6   FKBP1A transpo
7   XPO1 transport.
8   NR2F2 transport
9   ABCA5 transport
10  IGFBP3 t
11  APOD tr
12  KIAA078
13  CLCN3 t
14  STC1 tr
15  KPNA2 t
16  SEPP1 t
```

Algorithm for Functional Genomics

Classification

Weight on false positives: 4

File with classifications: classifications.txt [Browse...]

File with rules: rules.txt [Browse...]

Cross validation

☑ Cross validation evaluation

File with cross validation: cv-estimates.txt [Browse...]

Number of cross validation iterations: 10

Seed to RNG: 5

[OK] [Cancel]

Template definitions

INCREASE/DECREASE TEMPLATES:

The increase/decrease required: 0.6

Required increase/decrease during the first and last sub-intervals: 0.1

Length of templates in terms of sub-intervals: 2

Maximum change in the "wrong" direction allowed: 0.2

CONSTANT TEMPLATE:

Maximum allowed variation from average: 0.2

Length of templates in terms of sub-intervals: 3

☑ Supervised

Genetic algorithm for reduct computation

☑ IDG file

File with rules: D:\Rosetta\idg-not-symmetric.txt [Browse...]

☑ Approximate reducts 0.9

Classification: quality versus interpretability

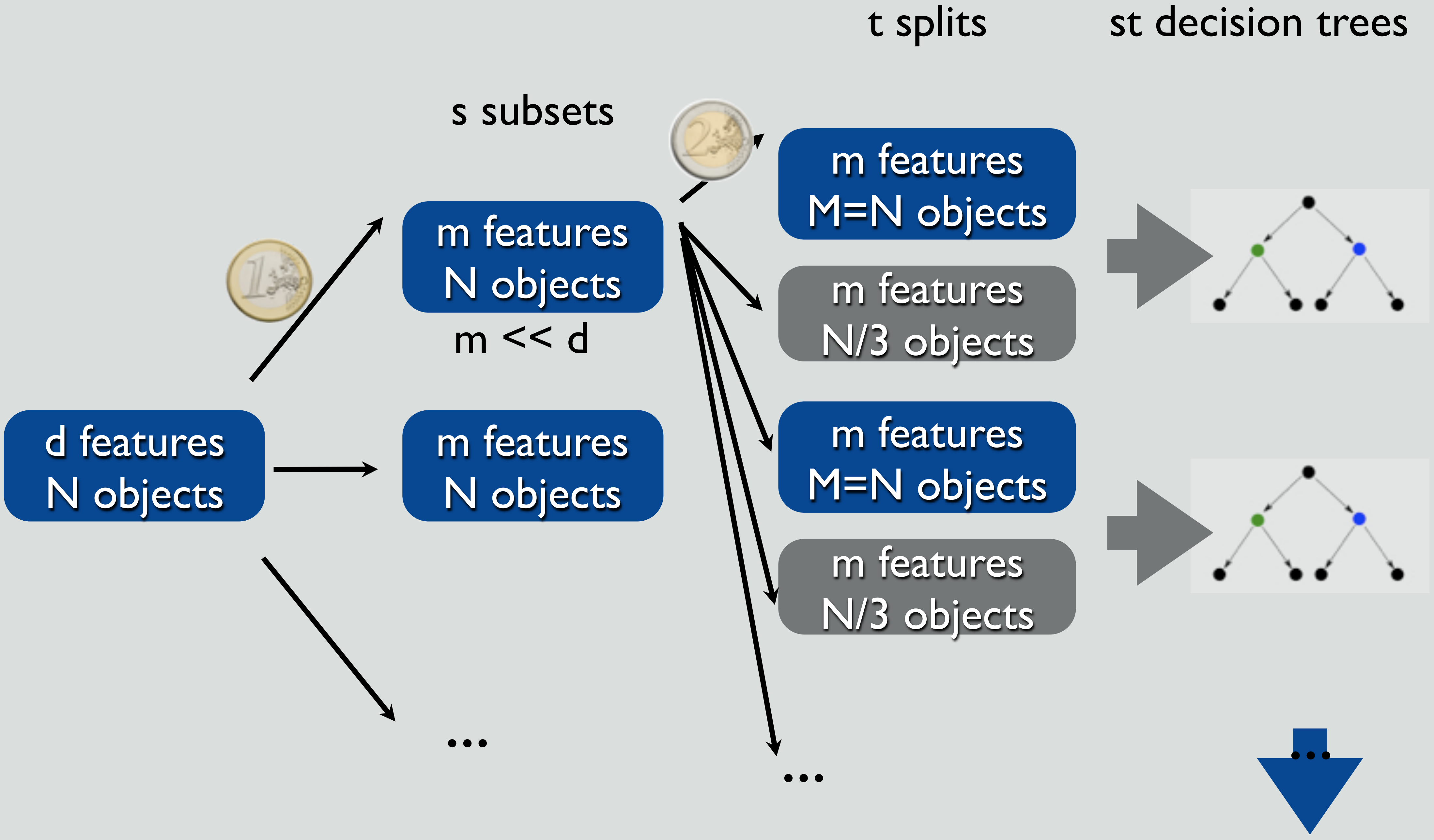
- Traditionally: quality of classification
- Our view: the structure and interpretation of the model in the original (untransformed) language of the experiment:
  - significant features and their ranking
  - easily interpretable results -> networks of interacting features

Transformed	Preserved
neural networks	decision trees
Support Vector Machines	rough sets
linear regression	fuzzy sets
...	...

Model structure rather than classification ability

- The combined approach of Monte Carlo Feature Selection and Rosetta:
  - features are significance-ranked,
  - models are based on the original coordinates, human legible and have a structure that is amenable to interpretation
  - an novel approach to systems biology
- How:
  - Feature selection
  - rule-based modeling
  - networks of interdependent features
- But: there may be many interacting pairs, triplets, etc, features

Monte Carlo Feature Selection



t splits

st decision trees

s subsets

d features
N objects

m features
N objects

m << d

m features
N objects

...

m features
M=N objects

m features
N/3 objects

m features
M=N objects

m features
N/3 objects

...

...

feature
importance
ranking

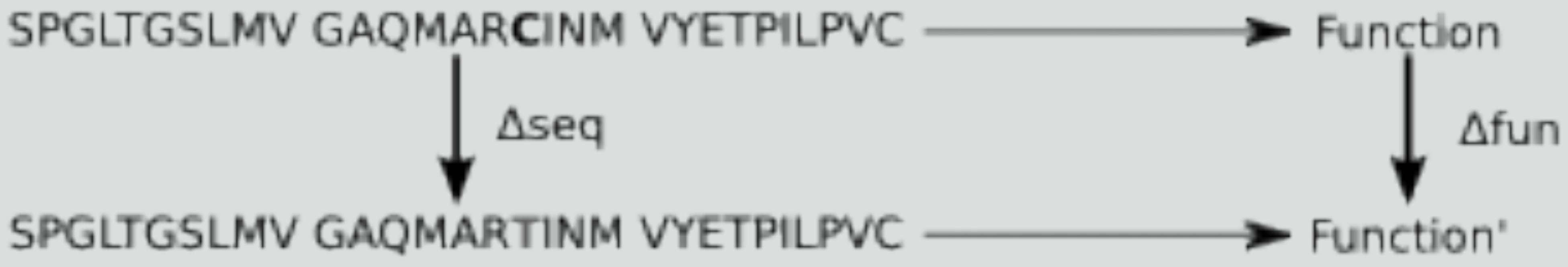
Informal analogy: MCFS = signal amplifier

s increases until the subsequent
rankings converge enough

randomization test => p-value

Dramińnski et al. Bioinformatics, 24:110–117, 2008.

First case: modeling protein function



SPGLTGSLMV GAQMARCINM VYETPILPVC → Function

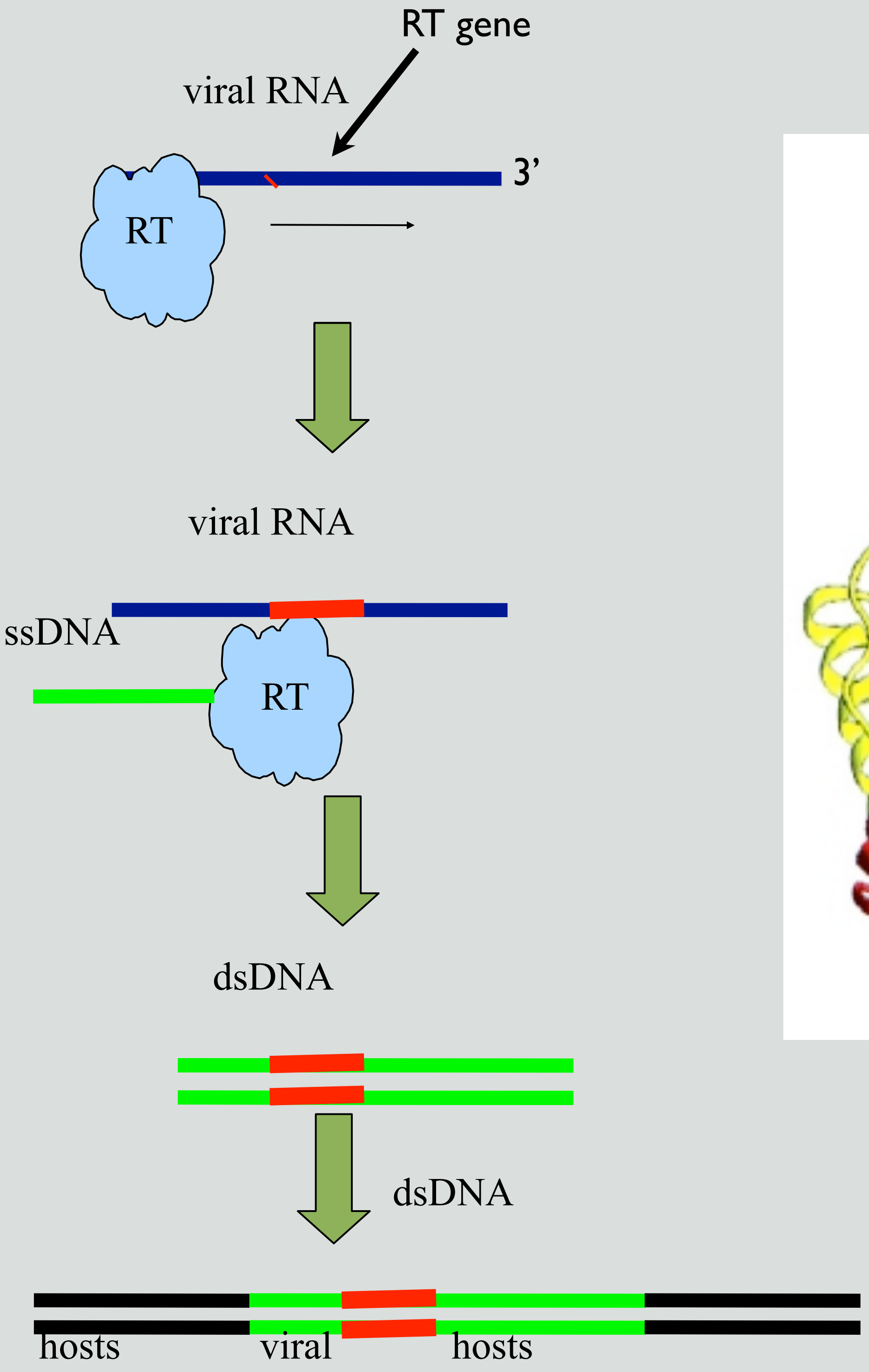
Δseq

Δfun

SPGLTGSLMV GAQMARTINM VYETPILPVC → Function'

O-sialoglycoprotein metal-dependent endopeptidase [Leptospira borgpetersenii serovar Hardjo-bovis L550]

Reverse transcription



after: Huifang H. et al., 1998

- One of the major targets in anti-HIV
therapies
- High error rate (no proof-reading activity)

When and how HIV-1 RT is susceptible to drugs?

- Which aa (positions) in the Reverse Transcriptase contribute to (changing) drug susceptibility?
- And, in what combinations?

The learning data set

Data pre-processing: alignment and removal of highly incomplete sequences.

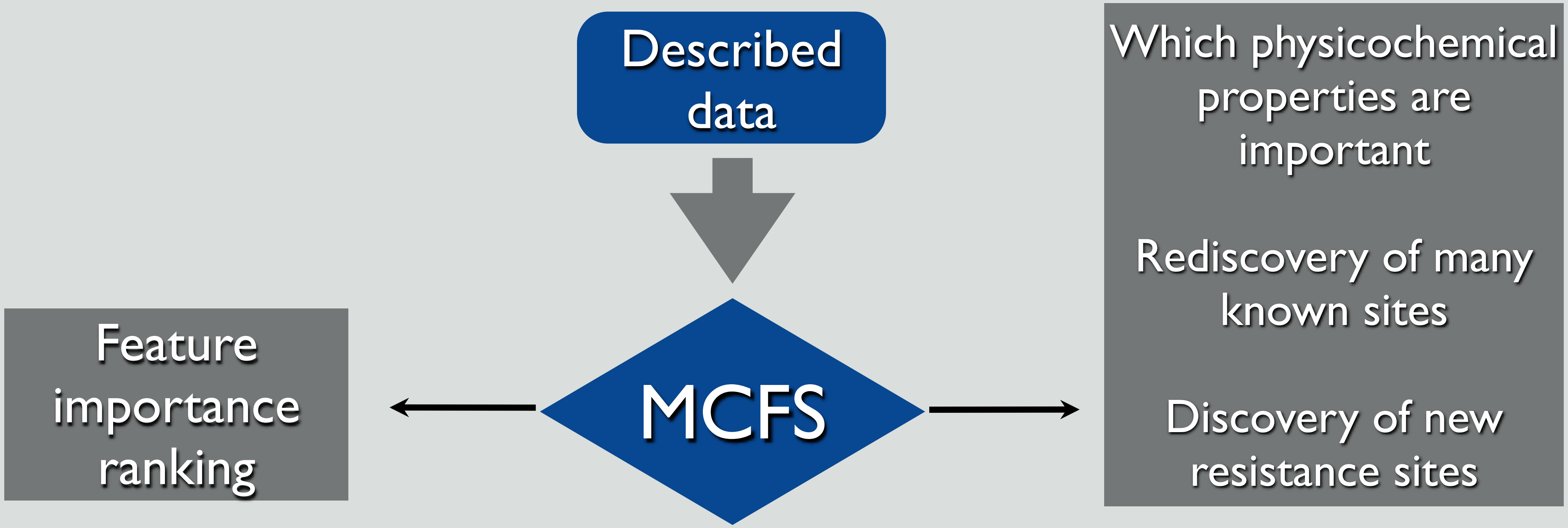
For each of the analyzed drugs label it with lab results:
susceptible, moderate, resistant

```
pispiapvpv klkpgmdgpk ... meqegkisri gpenpyntpi wild-type

pispiapvpv klkpgmdgpk ... meqegkisri gpenryntpi susceptible
...........................................................
pispiapvpv klkpgmdgpk ... meqegkisri gpqnpyntpi susceptible
pispiapvpv klkprmdgpk ... meqegkisri gpenpyntpi moderate
...........................................................
pispiapvpv klkpgmdgpk ... meqegkisri gpenpyntpi moderate
pispiapvpv klkpgvdgpk ... meqegkisri gpenpyntpi resistant
...........................................................
pispiapvpv klkvgmdgpk ... meqegkisri gpenpyntpi resistant
```

source: http://hivdb.stanford.edu

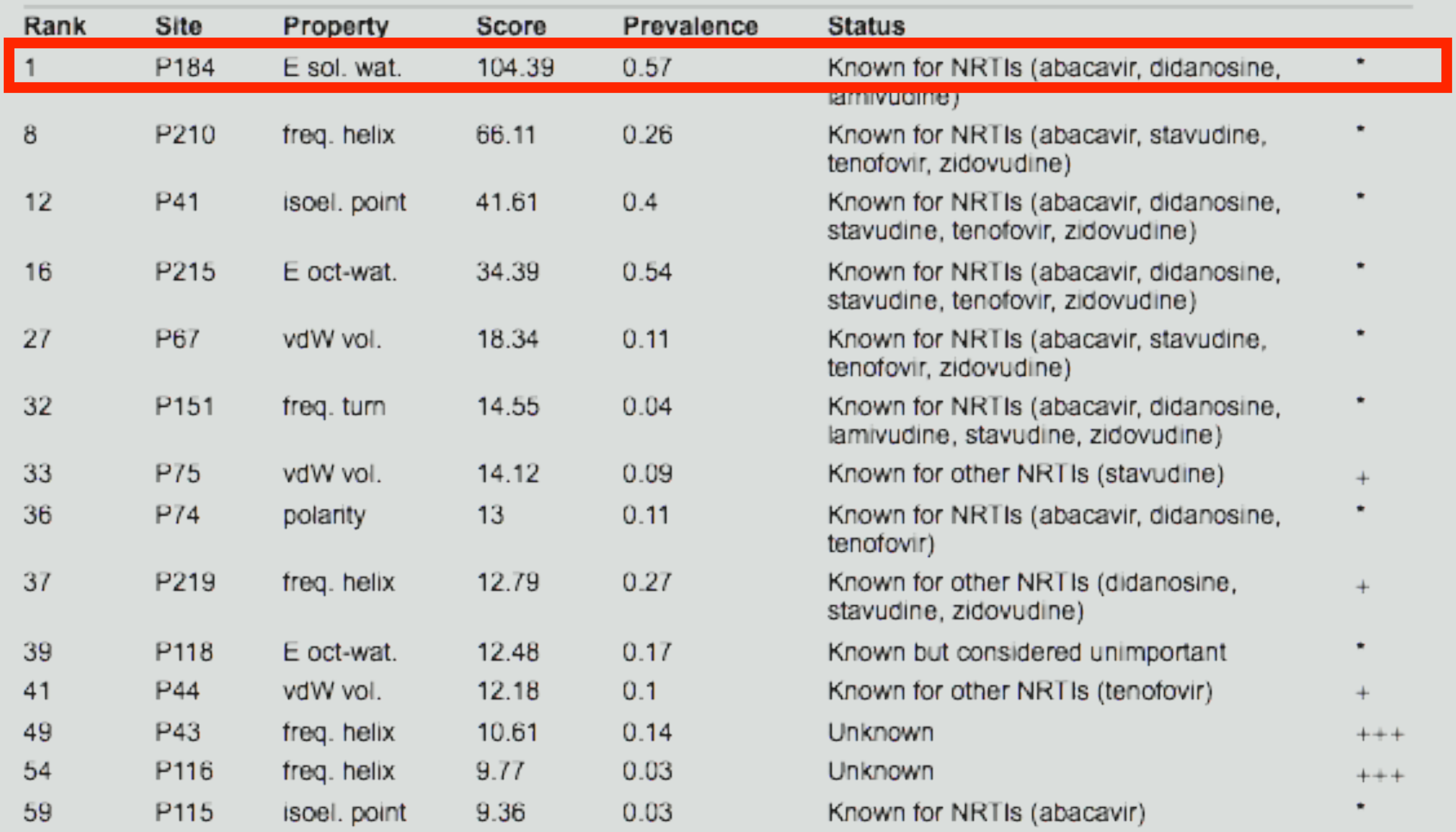
Methodology


Described data
MCFS
Feature importance ranking
Which physicochemical properties are important
Rediscovery of many known sites
Discovery of new resistance sites

- *Monte Carlo feature selection and interdependency discovery*, Draminski M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J.Bioinformatics, 2008 Jan 1;24(1):110-7; Advances in Machine Learning 2010, II:371-385.

Results - importance rankings

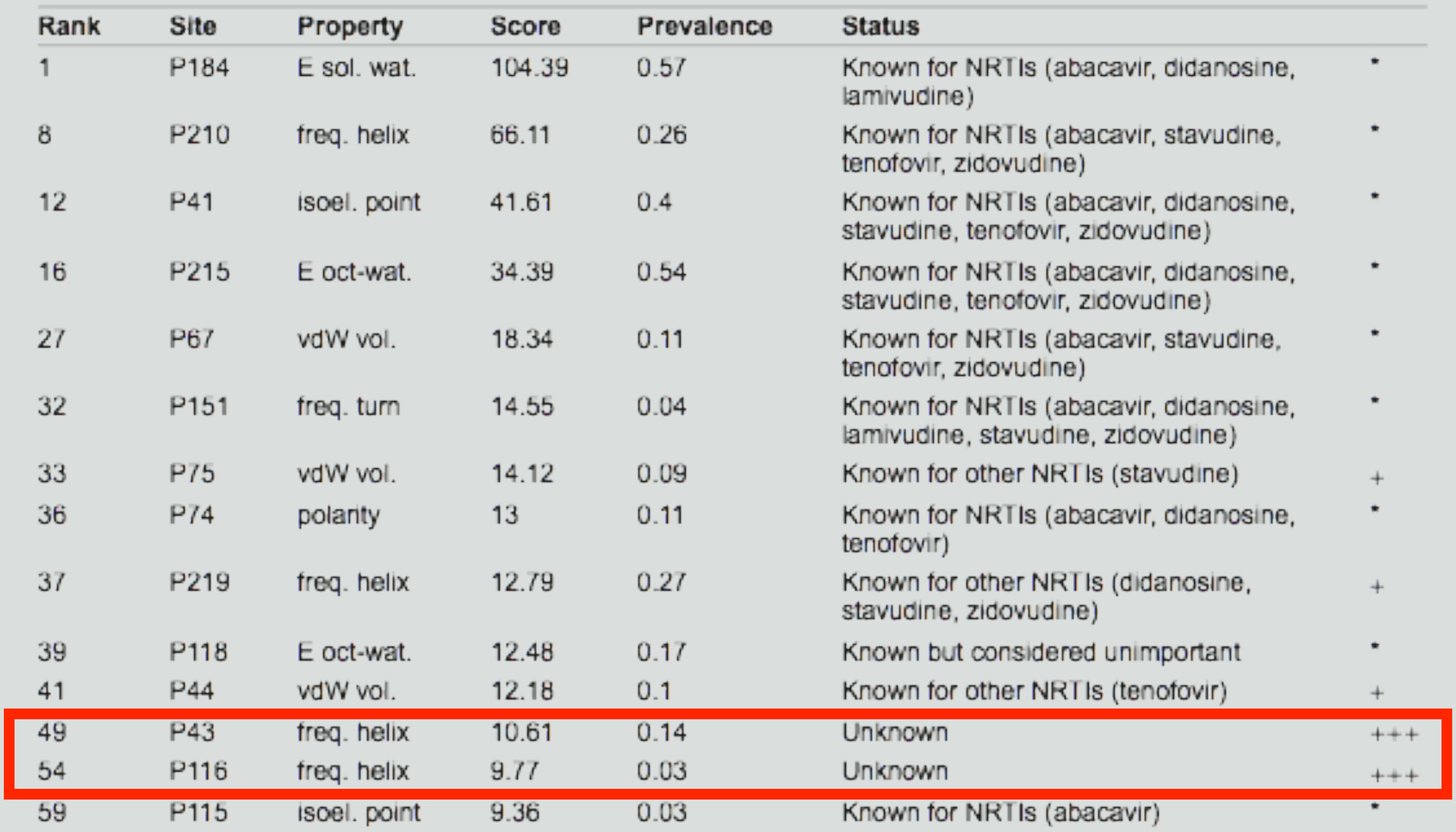
Resistance to Abacavir



Rank	Site	Property	Score	Prevalence	Status	
1	P184	E sol. wat.	104.39	0.57	Known for NRTIs (abacavir, didanosine, lamivudine)	*
8	P210	freq. helix	66.11	0.26	Known for NRTIs (abacavir, stavudine, tenofovir, zidovudine)	*
12	P41	isoel. point	41.61	0.4	Known for NRTIs (abacavir, didanosine, stavudine, tenofovir, zidovudine)	*
16	P215	E oct-wat.	34.39	0.54	Known for NRTIs (abacavir, didanosine, stavudine, tenofovir, zidovudine)	*
27	P67	vdW vol.	18.34	0.11	Known for NRTIs (abacavir, stavudine, tenofovir, zidovudine)	*
32	P151	freq. turn	14.55	0.04	Known for NRTIs (abacavir, didanosine, lamivudine, stavudine, zidovudine)	*
33	P75	vdW vol.	14.12	0.09	Known for other NRTIs (stavudine)	+
36	P74	polarity	13	0.11	Known for NRTIs (abacavir, didanosine, tenofovir)	*
37	P219	freq. helix	12.79	0.27	Known for other NRTIs (didanosine, stavudine, zidovudine)	+
39	P118	E oct-wat.	12.48	0.17	Known but considered unimportant	*
41	P44	vdW vol.	12.18	0.1	Known for other NRTIs (tenofovir)	+
49	P43	freq. helix	10.61	0.14	Unknown	+++
54	P116	freq. helix	9.77	0.03	Unknown	+++
59	P115	isoel. point	9.36	0.03	Known for NRTIs (abacavir)	*

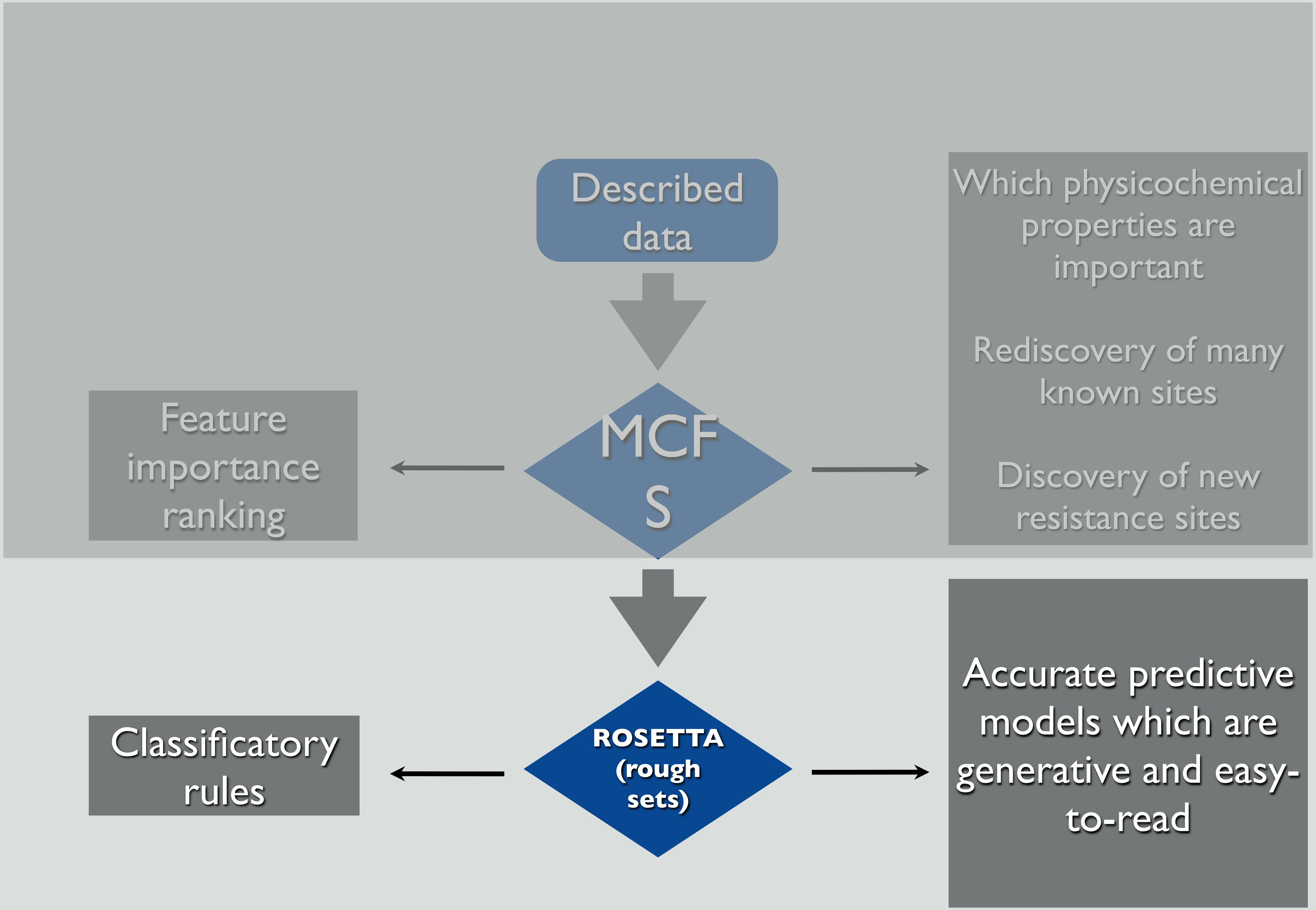
Results - importance rankings

Resistance to Abacavir



Rank	Site	Property	Score	Prevalence	Status	
1	P184	E sol. wat.	104.39	0.57	Known for NRTIs (abacavir, didanosine, lamivudine)	*
8	P210	freq. helix	66.11	0.26	Known for NRTIs (abacavir, stavudine, tenofovir, zidovudine)	*
12	P41	isoel. point	41.61	0.4	Known for NRTIs (abacavir, didanosine, stavudine, tenofovir, zidovudine)	*
16	P215	E oct-wat.	34.39	0.54	Known for NRTIs (abacavir, didanosine, stavudine, tenofovir, zidovudine)	*
27	P67	vdW vol.	18.34	0.11	Known for NRTIs (abacavir, stavudine, tenofovir, zidovudine)	*
32	P151	freq. turn	14.55	0.04	Known for NRTIs (abacavir, didanosine, lamivudine, stavudine, zidovudine)	*
33	P75	vdW vol.	14.12	0.09	Known for other NRTIs (stavudine)	+
36	P74	polarity	13	0.11	Known for NRTIs (abacavir, didanosine, tenofovir)	*
37	P219	freq. helix	12.79	0.27	Known for other NRTIs (didanosine, stavudine, zidovudine)	+
39	P118	E oct-wat.	12.48	0.17	Known but considered unimportant	*
41	P44	vdW vol.	12.18	0.1	Known for other NRTIs (tenofovir)	+
49	P43	freq. helix	10.61	0.14	Unknown	+++
54	P116	freq. helix	9.77	0.03	Unknown	+++
59	P115	isoel. point	9.36	0.03	Known for NRTIs (abacavir)	*

From features to models



Described data

MCFS

Feature importance ranking

Which physicochemical properties are important

Rediscovery of many known sites

Discovery of new resistance sites

ROSETTA (rough sets)

Classificatory rules

Accurate predictive models which are generative and easy-to-read

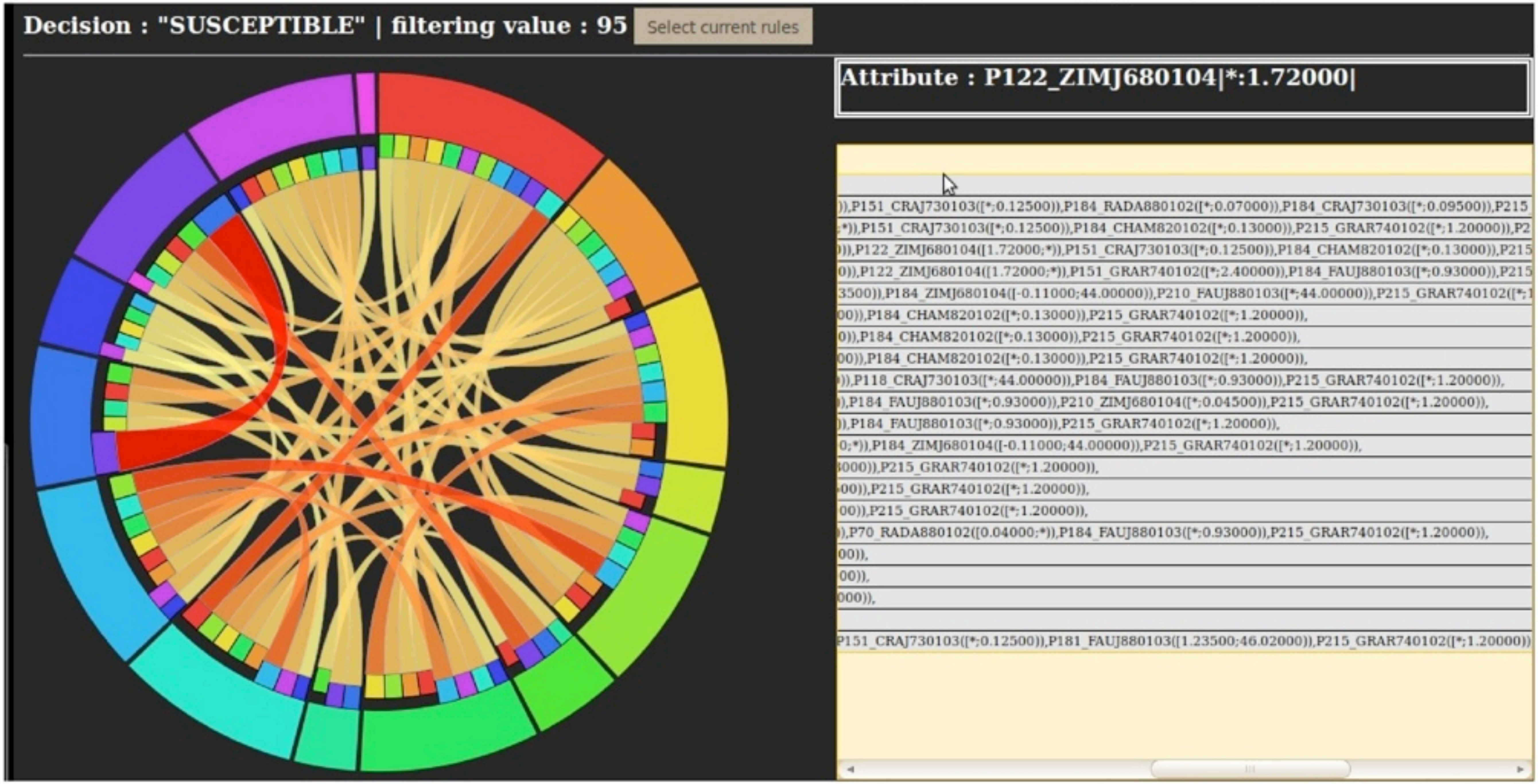
Handbook of Data Mining and Knowledge Discovery, W. Klösgen and J. Zytkow (eds.), ch. D.2.3, Oxford University Press. ISBN 0-19-511831-6

HIV RT

Critical positions : 65, 215, 151, 185, 210, 122

All are among with the selected sites mentioned in the article.

Three of them also show the mentioned top-scoring property.



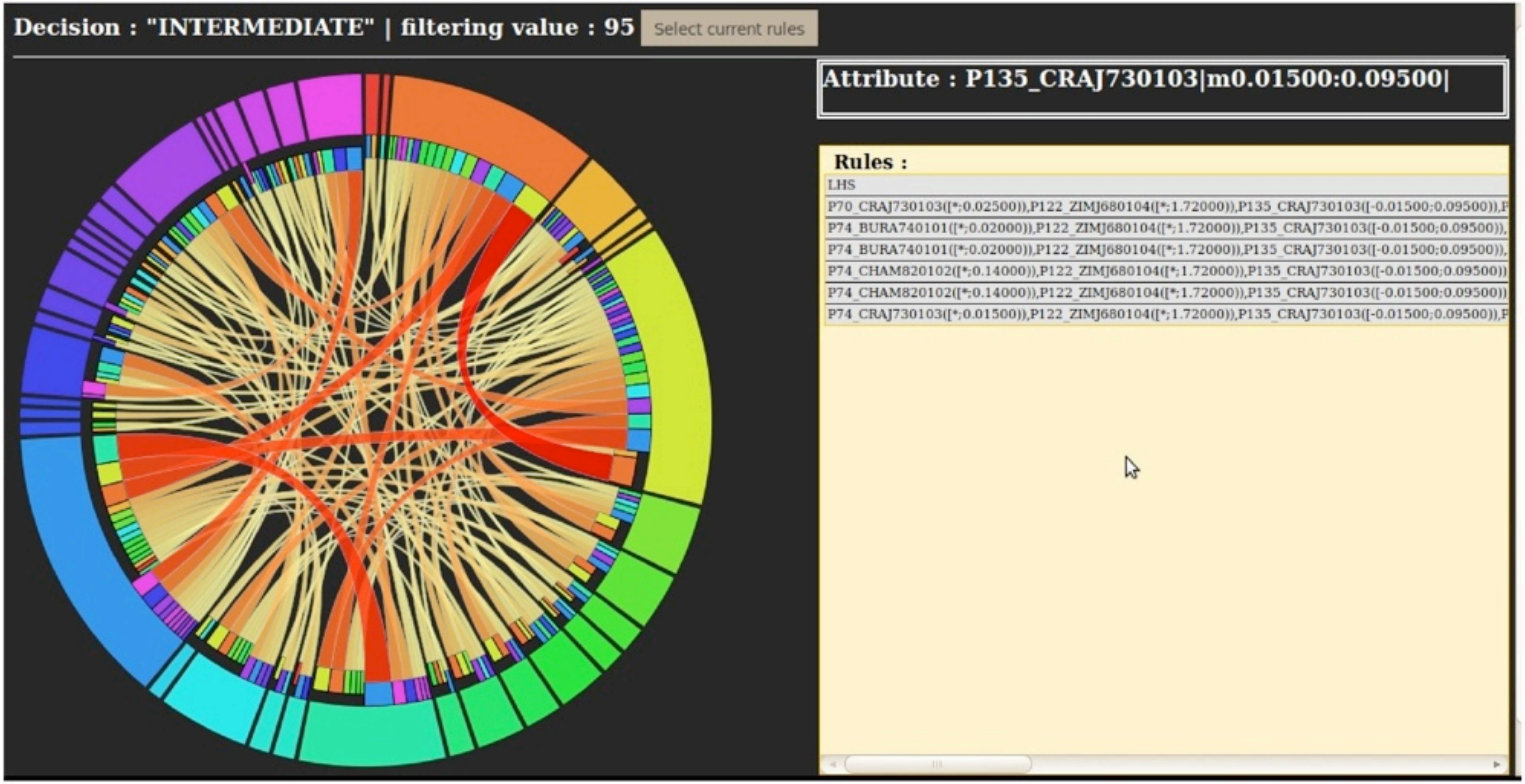
Attribute : P122_ZIMJ680104|*:1.72000|

```
)),P151_CRAJ730103([*;0.12500)),P184_RADA880102([*;0.07000)),P184_CRAJ730103([*;0.09500)),P215
;*)),P151_CRAJ730103([*;0.12500)),P184_CHAM820102([*;0.13000)),P215_GRAR740102([*;1.20000)),P2
)),P122_ZIMJ680104([1.72000;*)),P151_CRAJ730103([*;0.12500)),P184_CHAM820102([*;0.13000)),P215
0)),P122_ZIMJ680104([1.72000;*)),P151_GRAR740102([*;2.40000)),P184_FAUJ880103([*;0.93000)),P215
3500)),P184_ZIMJ680104([-0.11000;44.00000)),P210_FAUJ880103([*;44.00000)),P215_GRAR740102([*;1
00)),P184_CHAM820102([*;0.13000)),P215_GRAR740102([*;1.20000)),
0)),P184_CHAM820102([*;0.13000)),P215_GRAR740102([*;1.20000)),
00)),P184_CHAM820102([*;0.13000)),P215_GRAR740102([*;1.20000)),
)),P118_CRAJ730103([*;44.00000)),P184_FAUJ880103([*;0.93000)),P215_GRAR740102([*;1.20000)),
),P184_FAUJ880103([*;0.93000)),P210_ZIMJ680104([*;0.04500)),P215_GRAR740102([*;1.20000)),
)),P184_FAUJ880103([*;0.93000)),P215_GRAR740102([*;1.20000)),
0;*)),P184_ZIMJ680104([-0.11000;44.00000)),P215_GRAR740102([*;1.20000)),
3000)),P215_GRAR740102([*;1.20000)),
00)),P215_GRAR740102([*;1.20000)),
00)),P215_GRAR740102([*;1.20000)),
),P70_RADA880102([0.04000;*)),P184_FAUJ880103([*;0.93000)),P215_GRAR740102([*;1.20000)),
00)),
00)),
000)),

P151_CRAJ730103([*;0.12500)),P181_FAUJ880103([1.23500;46.02000)),P215_GRAR740102([*;1.20000))
```

HIV-RT

Critical positions : 122, 184, 215, 135, 74, 70, 228
All except 70 are among the selected positions in the article
Three of them also show the mentioned top-scoring property.



Decision : "INTERMEDIATE" | filtering value : 95

Attribute : P135_CRAJ730103|m0.01500:0.09500|

Rules :

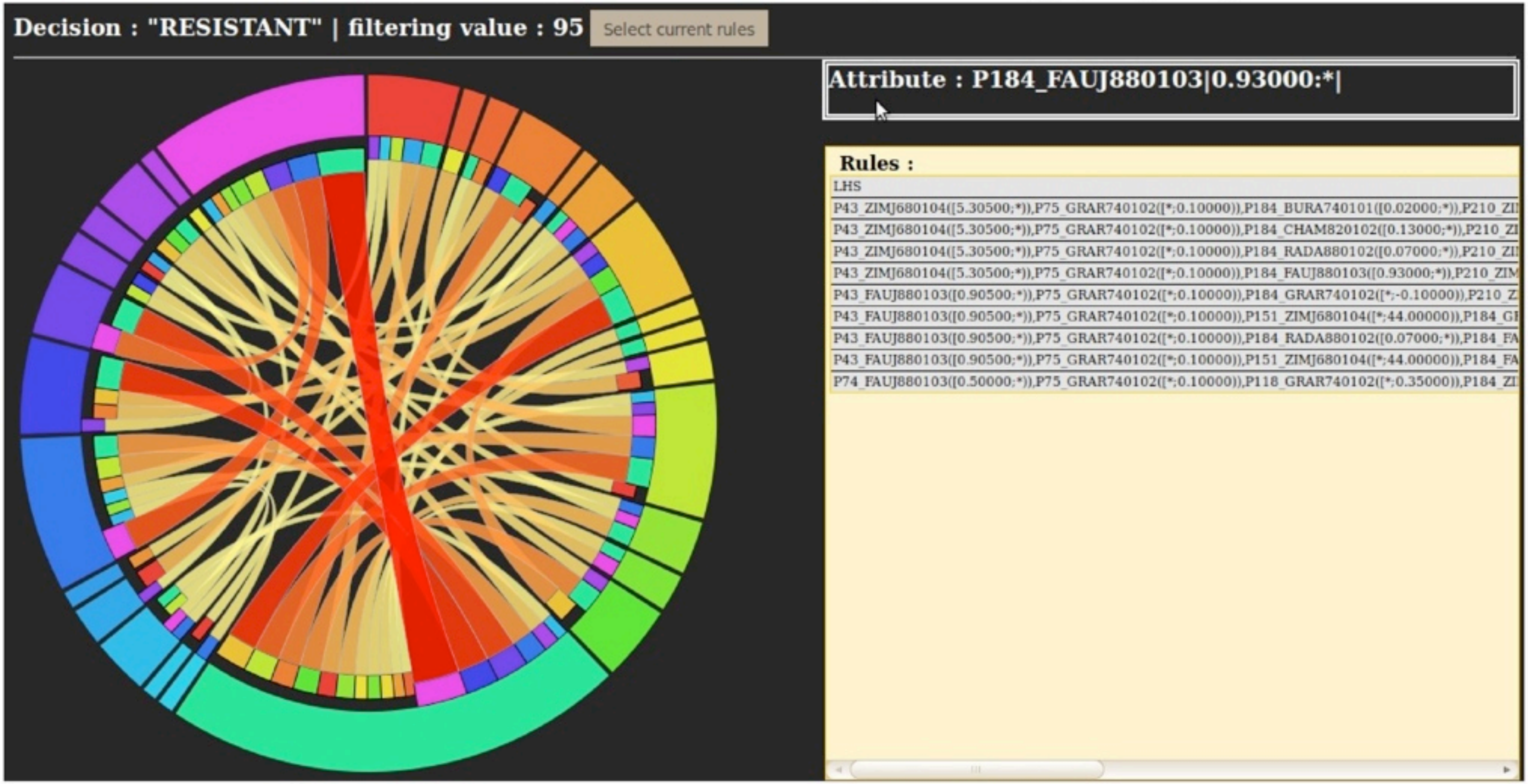
LHS
P70_CRAJ730103([*;0.02500)),P122_ZIMJ680104([*;1.72000)),P135_CRAJ730103([-0.01500;0.09500)),P
P74_BURA740101([*;0.02000)),P122_ZIMJ680104([*;1.72000)),P135_CRAJ730103([-0.01500;0.09500)),
P74_BURA740101([*;0.02000)),P122_ZIMJ680104([*;1.72000)),P135_CRAJ730103([-0.01500;0.09500)),
P74_CHAM820102([*;0.14000)),P122_ZIMJ680104([*;1.72000)),P135_CRAJ730103([-0.01500;0.09500))
P74_CHAM820102([*;0.14000)),P122_ZIMJ680104([*;1.72000)),P135_CRAJ730103([-0.01500;0.09500))
P74_CRAJ730103([*;0.01500)),P122_ZIMJ680104([*;1.72000)),P135_CRAJ730103([-0.01500;0.09500)),P

HIV-RT

Critical positions : 75, 210, 181, 43, 184, 135, 118

All except are among the selected positions in the article

One of them also show the mentioned top-scoring property.



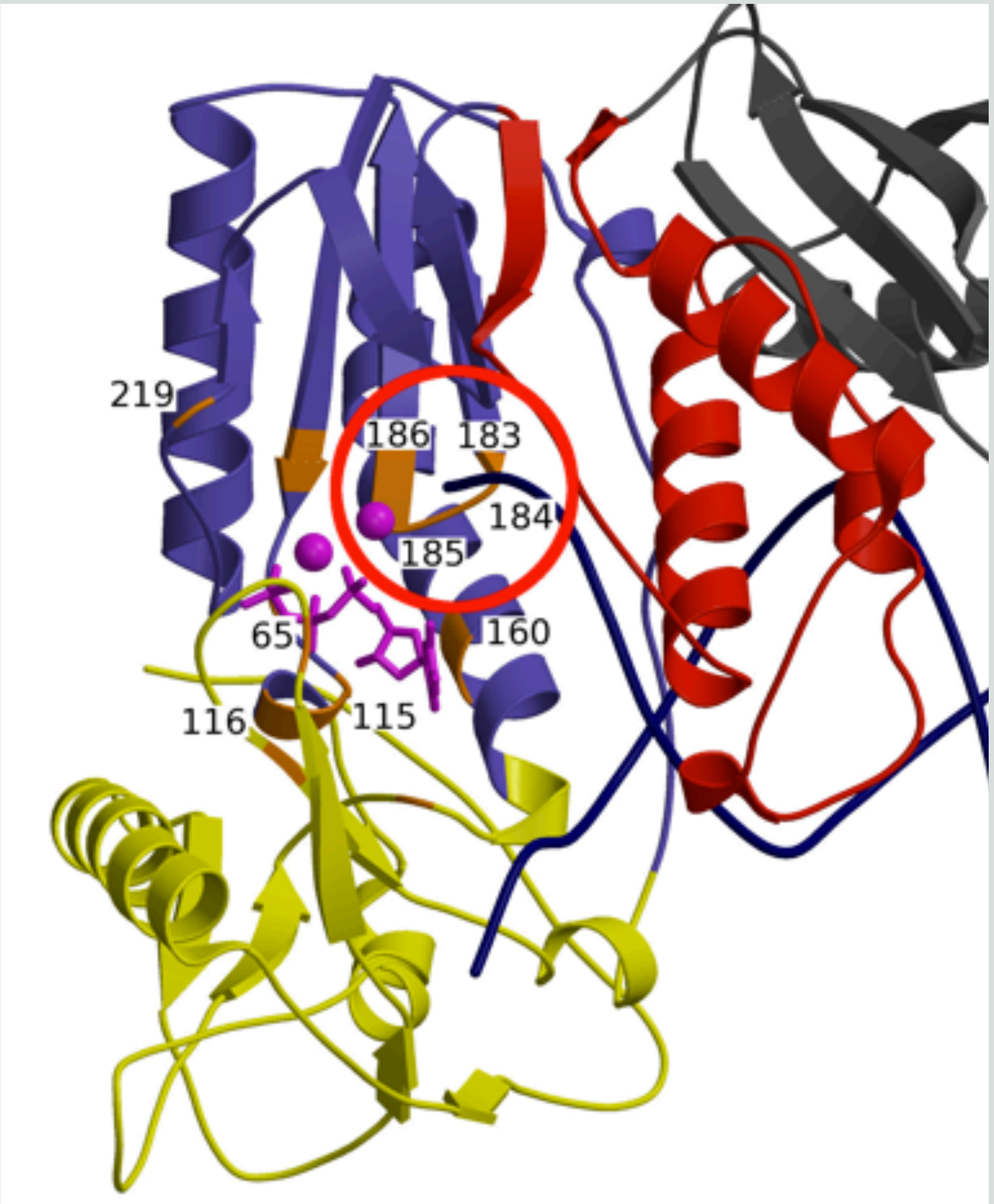
Decision : "RESISTANT" | filtering value : 95

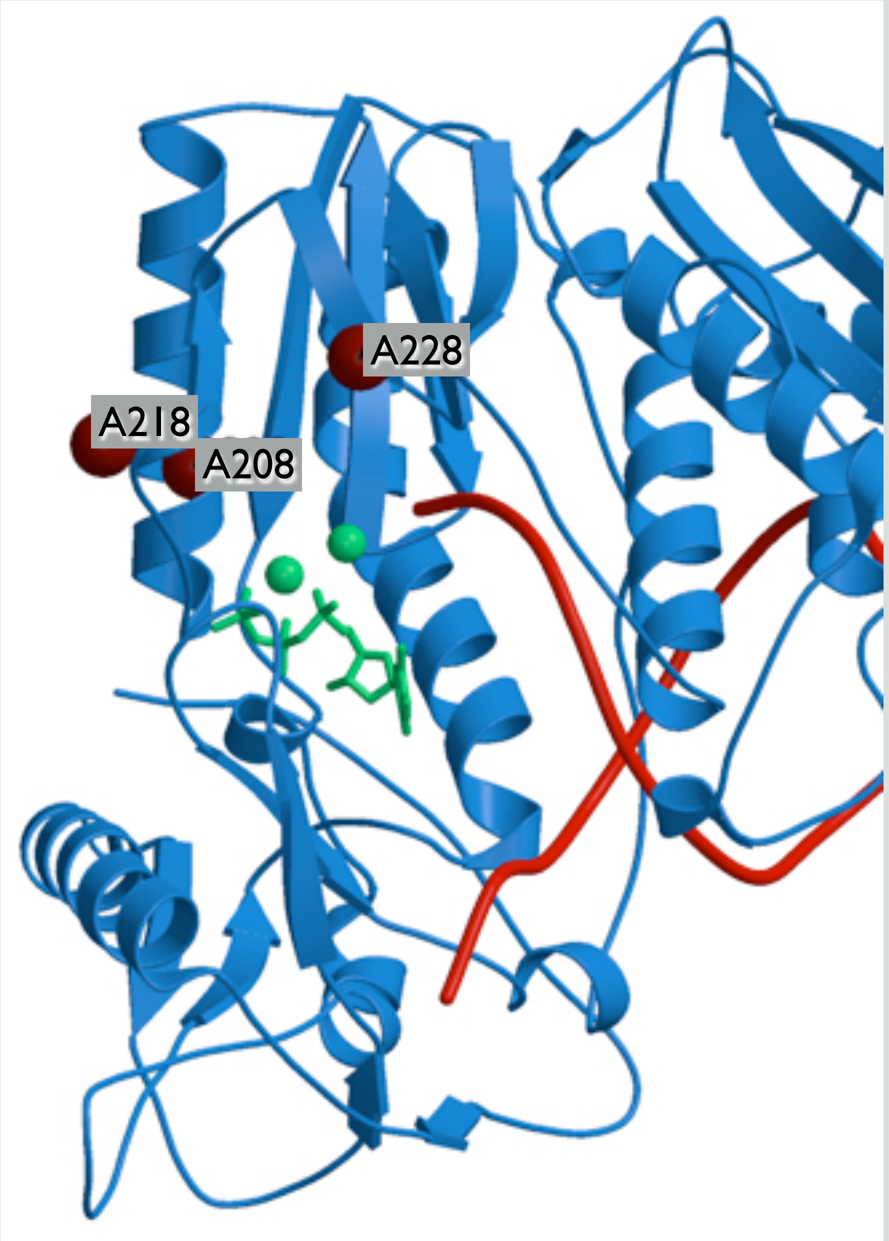
Attribute : P184_FAUJ880103|0.93000:*|

Rules :

LHS
P43_ZIMJ680104([5.30500;*)),P75_GRAR740102([*;0.10000)),P184_BURA740101([0.02000;*)),P210_ZI
P43_ZIMJ680104([5.30500;*)),P75_GRAR740102([*;0.10000)),P184_CHAM820102([0.13000;*)),P210_ZI
P43_ZIMJ680104([5.30500;*)),P75_GRAR740102([*;0.10000)),P184_RADA880102([0.07000;*)),P210_ZI
P43_ZIMJ680104([5.30500;*)),P75_GRAR740102([*;0.10000)),P184_FAUJ880103([0.93000;*)),P210_ZIM
P43_FAUJ880103([0.90500;*)),P75_GRAR740102([*;0.10000)),P184_GRAR740102([*;-0.10000)),P210_Z
P43_FAUJ880103([0.90500;*)),P75_GRAR740102([*;0.10000)),P151_ZIMJ680104([*;44.00000)),P184_G
P43_FAUJ880103([0.90500;*)),P75_GRAR740102([*;0.10000)),P184_RADA880102([0.07000;*)),P184_FA
P43_FAUJ880103([0.90500;*)),P75_GRAR740102([*;0.10000)),P151_ZIMJ680104([*;44.00000)),P184_FA
P74_FAUJ880103([0.50000;*)),P75_GRAR740102([*;0.10000)),P118_GRAR740102([*;0.35000)),P184_ZI

Conclusions after 3D mapping





For the indicated sites - in vitro testing recommended.

Results

1. Physicochemical property/site rankings.

2. Re-discovery and discovery of several sites.

3. Networks of interdependencies between physicochemical properties.

4. Validation: literature and 3D structure.

5. Ultimate validation: wet-lab experiments.

Shadowing

- **Shadowing** occurs when two or more variables (features) in the system carry at least partially the same information correlated with the decision attribute.

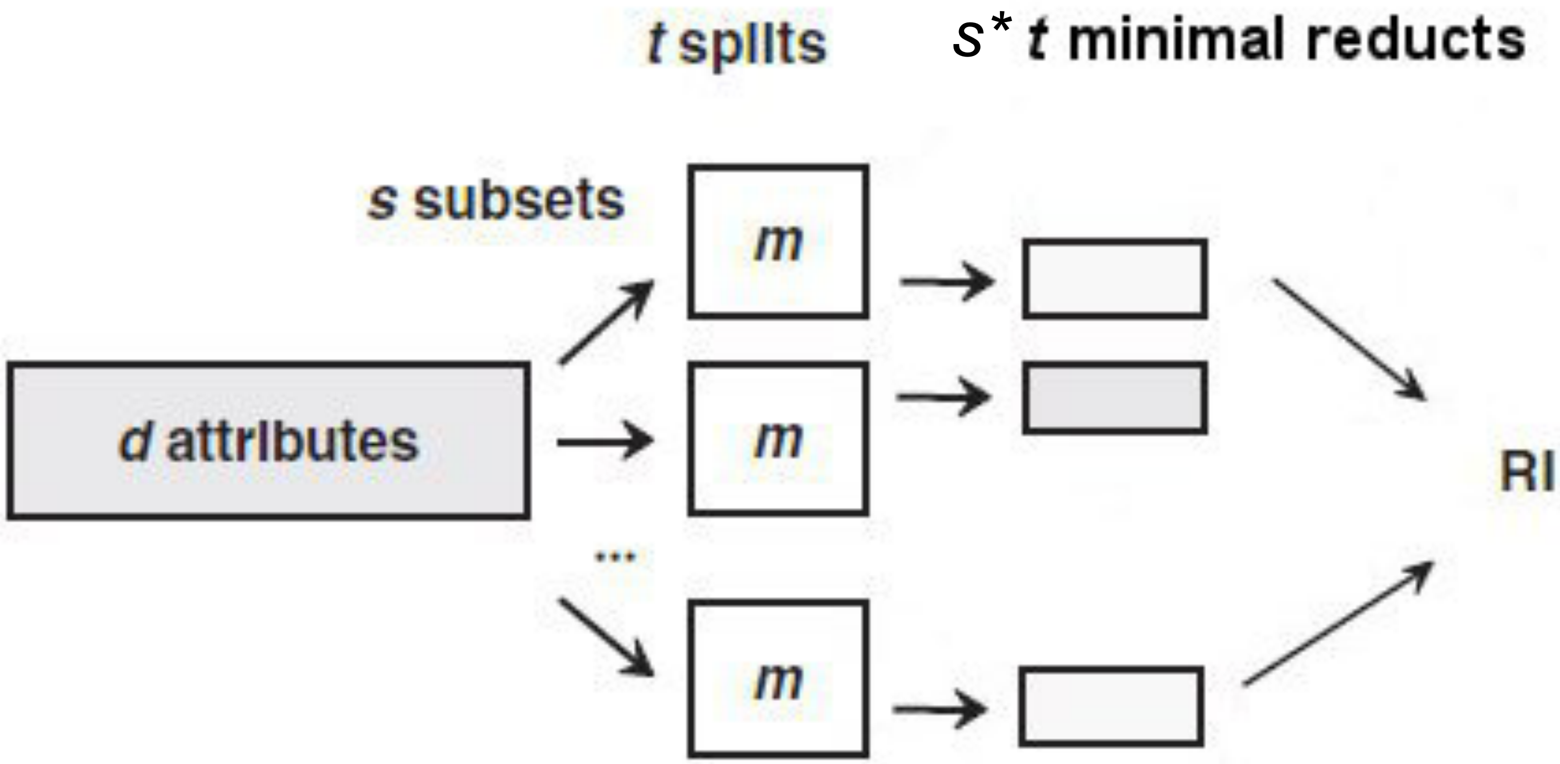
Patient	Fat ratio	Weight	Height	Risk of heart disease
Patient no 1	low	low	medium	low
Patient no 2	low	high	tall	low
Patient no 3	high	high	medium	high

- To determine the risk, measure **Fat ratio** and **Weight**, or **Fat ratio** and **Height**. Therefore, Weight shadows Height and vice versa.

How to deal with shadowing

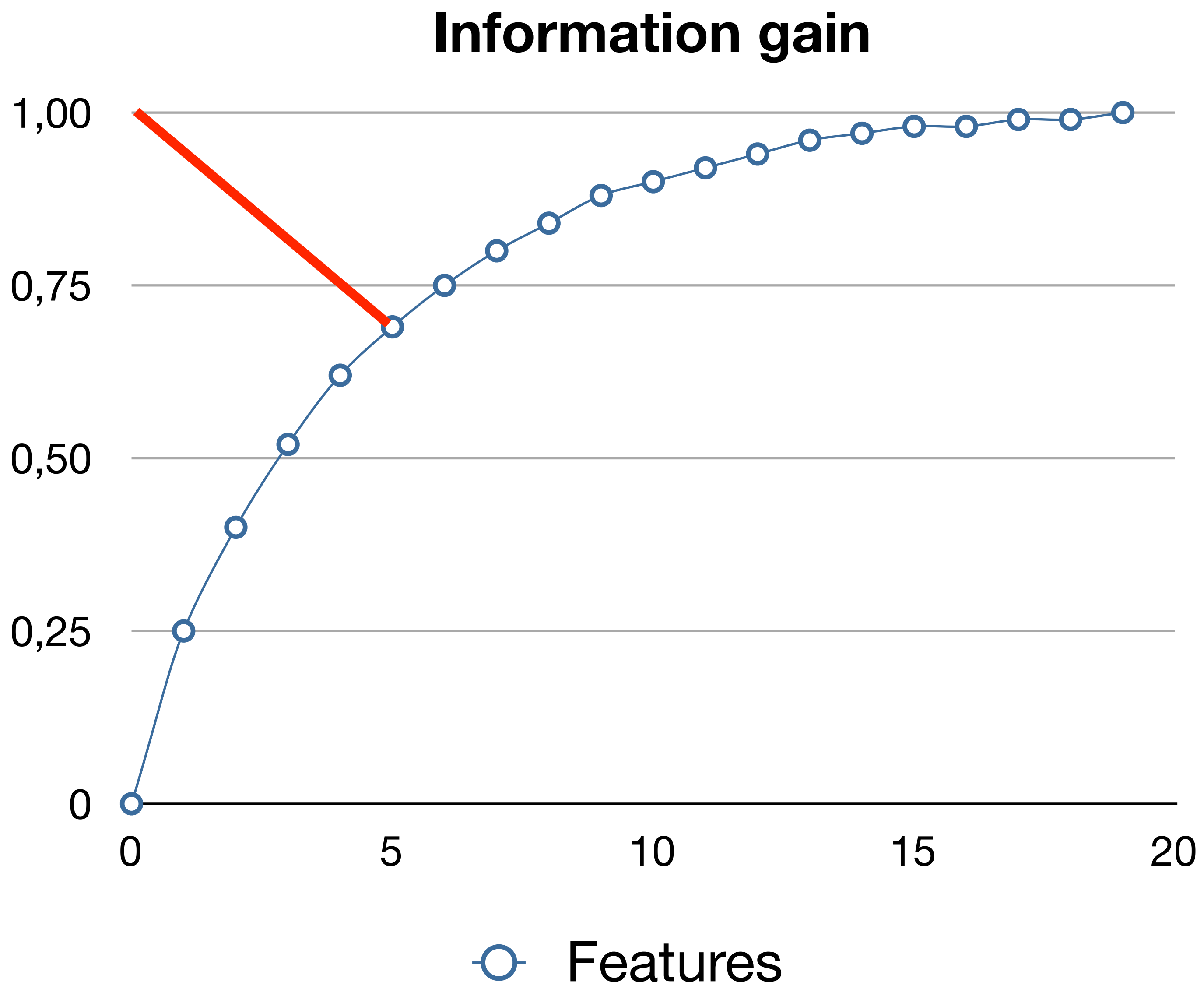
- Proposal:
  - instead of computing decision trees compute <u>random</u> reducts
  - if efficiency is the issue, replace the randomization step with ROC-based selection

The Random Reduct algorithm (RR)



Cut-off point calculation

- MCFS and Random Reducts use simple Permutation Tests for cut-off point calculation (decision is made basing on p-values)
- Random Reducts has so called *fast option*



Simulation Data Preparation

- Generate data with features correlated to the decision on different levels
- Add copies of some of the features to the dataset (fully shadowing features)

MCFS and RR comparison in terms of shadowing

MCFS results		RR results	
Feature ID	Correlation to decision	Feature ID	Correlation to decision
F1	72,07%	F1	72,07%
F2	69,41%	F2	69,41%
F5	67,20%	F3	69,41%
F6	65,74%	F4	67,20%
F7	61,13%	F5	67,20%
F8	59,68%	F6	65,74%
F10	57,93%	F8	59,68%
F12	53,61%	F7	61,13%
F11	53,66%	F9	57,93%
F14	51,94%	F10	57,93%
F15	49,33%	F11	53,66%
F16	48,89%	F12	53,61%
F17	46,01%	F13	51,94%
F19	45,47%	F14	51,94%
F20	42,93%	F16	48,89%
F22	40,35%	F15	49,33%
F23	36,35%	F17	46,01%
F24	37,21%	F18	45,47%
F25	35,28%	F19	45,47%

Biological data tests

The tests have been performed on a virus protein sequences dataset containing over 4000 features and 752 objects

	MCFS (10k projections, 30 permutations)	RR (10k projections, 30 permutations)	RR (10k projections, fast option - 1 iteration	RR (10k projections, fast option - 6 iterations)
Time of evaluation	60,345 s	21,642 s	497 s	1207 s
Number of features selected	305	1847	1163	22
Accuracy of the model	95,1%	97,1%	95,3%	91,8%
Speed Difference		2.787x	121.418x	49.99x

Summary of the methodology

- MCFS/RR to select ranked and significant features
- Rule-based model to provide a legible model
- If model not accurate for all outcomes, choose locally acurate ones
- RR to remove shadowing and run very efficiently

Conclusions

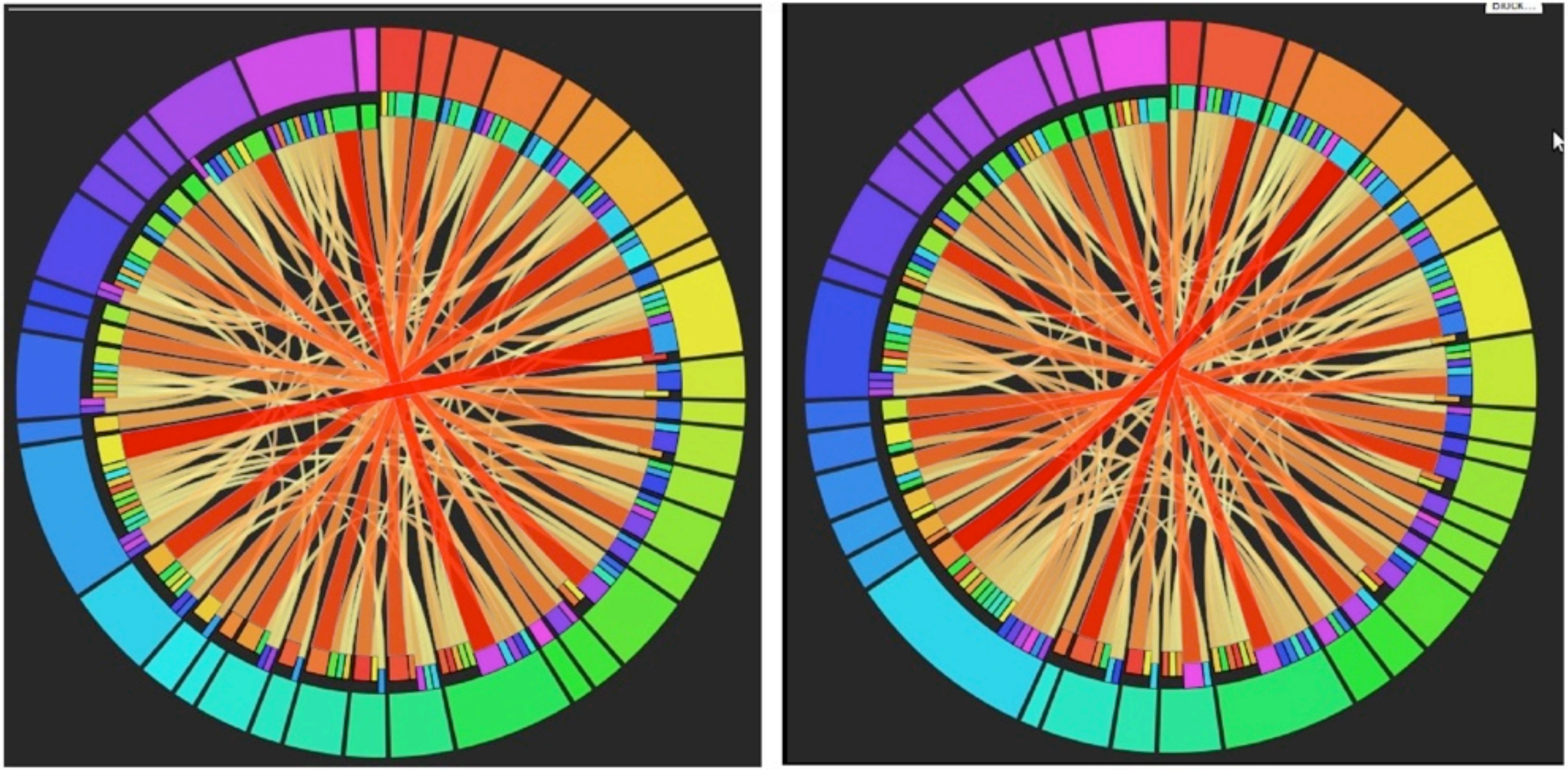
- The structure (features and rules) of a classifier is more important to explaining the modeled phenomenon (outcome, decision) than its numerical qualities (accuracy, AUC)
- Generative property allows constructing chimeric cases and biological validation
- A new approach to network (system) biology
  - Pairs of interacting features extracted from the rules give a well-defined systems (differential) view of the features that determine the outcome

Acknowledgements

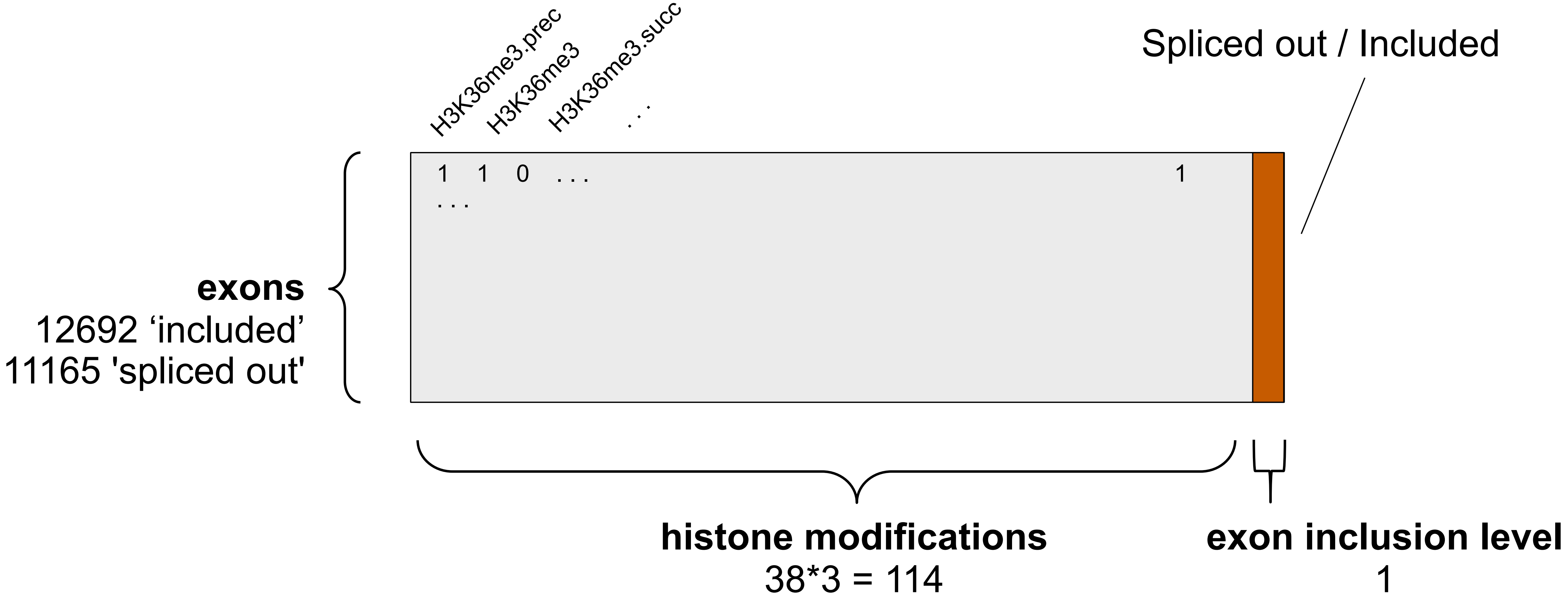
- Uppsala
  - PhD students: Susanne Bornelöv, Marcin Kruczyk
  - Masters students: Zeeshan Khaliq, Simon Marillet
  - Rudbeck Collaborators: Ola Wallerman, Helene Nord, and Claes Wadelius
- Warszawa
  - Krzysztof Ginalski, Michal Draminski and Jacek Koronacki
- Former members, collaborators and visitors at the LCB
  - Former PhD students: Marcin Kierczak, Stefan Enroth, Adam Ameur, Robin Andersson, Aleksejs Kontijevskis, Torgeir Hvidsten
  - Mats Gustafsson, Jarl Wikberg
  - Witold Rudnicki, ICM UW

Test on artificial data

The related attributes are on the opposite side of the circle



Another classification problem



A sample rule:
IF H2BK5me1.prec=1 **AND** H2BK5me1.succ=1 **AND** H3K4me1.succ=0 **AND** H3K36me3.prec=0 **AND** H3K36me3.succ=0 **AND** H4K20me1.prec=1 **AND** H4K91ac.prec=1 **THEN** Inclusion_level='Spliced out'

Oberdoerffer S et al. (2008) Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. Science 321: 686-691.