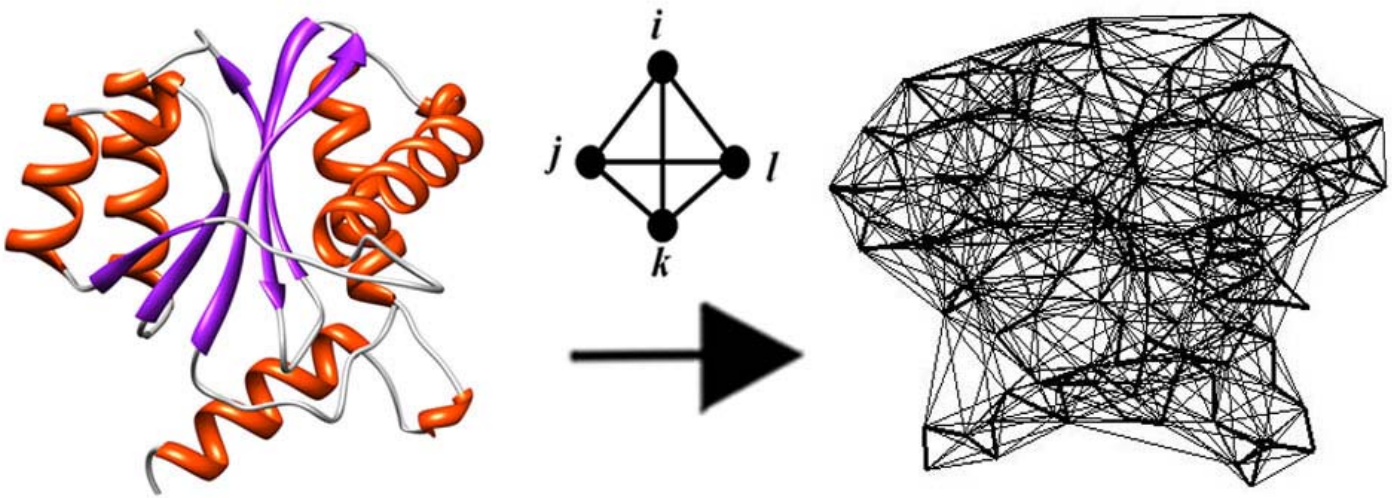
Predicting Drug Resistance: Probability and Statistics Meet the Building Blocks of Proteins



Majid Masso

School of Systems Biology

George Mason University

2014 Joint Mathematics Meeting

Amino Acids – The Protein Building Blocks

- 20 distinct amino acid (aa) types, each assigned a letter: {A, C, D, E, F ,G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}

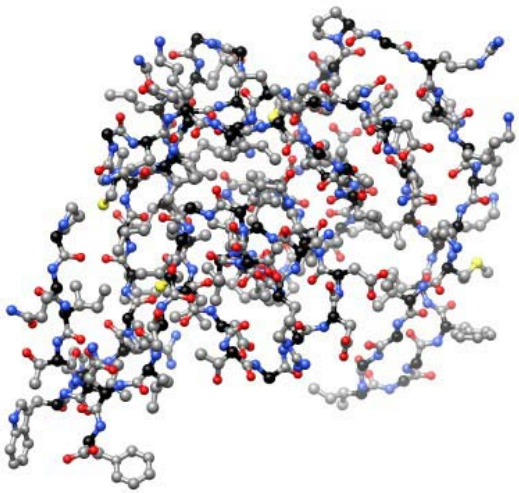
- What is a protein? A linear sequence (chain) of consecutively linked aa's, selected w/replacement from above set (avg. size ~ 300 aa's), which folds into a precise 3D structure

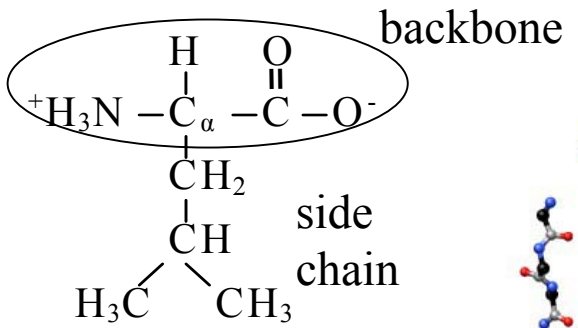
- Protein structure maintained by atomic interactions between structurally neighboring aa's (may be far apart in linear sequence)

- Genes (DNA) are blueprints or codes for making proteins (the workhorses: enzymes, hormones, receptors, antibodies, etc.)

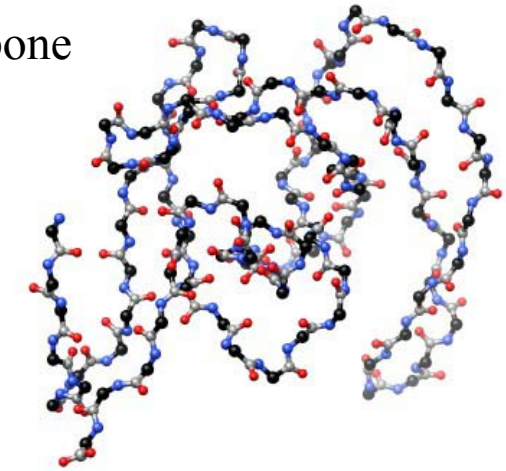
Protein Example: HIV-1 Protease

Each aa comprised of several atoms: identical backbones, unique side chains



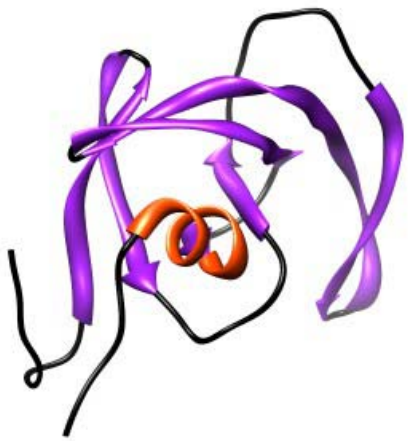


Leucine (Leu or L)



Backbone atoms reveal path

Each CA point has 2 labels:
1. Amino acid letter
2. Sequence position number



Coarse-grained model: one CA atom per aa

Protein Data Bank (PDB, http://www.pdb.org)

RCSB PDB : Structure Explorer - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.pdb.org/pdb/explore/explore.do?structureId=3PHV

Most Visited | Getting Started | Latest Headlines | HIV DBs | Binf | Res | Libs | GMU | Accts | Services | Search | Schools | Tutorials | News | Misc | Confs

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB MyPDB: Login | Register

An Information Portal to Biological Macromolecular Structures

As of Tuesday Mar 10, 2009 there are 56366 Structures | PDB Statistics

CONTACT US | FEEDBACK | HELP | PRINT PDB ID or keyword Author Site Search | Advanced Search

Home | Search | Structure | Queries

- 3PHV
- Download Files
- FASTA Sequence
- Download Original Files
- Display Files
- Display Molecule
- Structural Reports
- External Links
- Structure Analysis
- Help

Quick Tips:

Click the PDB file icon above to view the PDB file.

Some chains and/or residues have been updated. Click here for details, or here for details about the remediation process

Help | Structure Summary | Sequence Details | Biology & Chemistry | Materials & Methods | Geometry | Remediation | External Links

3phv

Learn more: [M]
DOI 10.2210/pdb3phv/pdb

Red - Derived Information

Title	X-RAY ANALYSIS OF HIV-1 PROTEINASE AT 2.7 ANGSTROMS RESOLUTION CONFIRMS STRUCTURAL HOMOLOGY AMONG RETROVIRAL ENZYMES
Authors	Lapatto, R., Blundell, T.L., Hemmings, A., Wilderspin, A., Wood, S.P., Danley, D.E., Geoghegan, K.F., Hawrylik, S.J., Hobart, P.M.
Primary Citation	Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J.R., Whittle, P.J., Danley, D.E., Geoghegan, K.F., *et al.* (1989) X-ray analysis of HIV-1 proteinase at 2.7 A resolution confirms structural homology among retroviral enzymes. *Nature* **342:** *299-302* [Abstract] PubMed
History	Deposition 1991-11-04 Release 1992-01-15 Last Modified (REVDAT) 2003-04-01

Images and Visualization

<< Asymmetric Unit >>

Display Options

KiNG
Jmol
WebMol
MBT SimpleViewer
MBT Protein Workshop

Done

PDB File Format

```
HEADER    HYDROLASE(ASPARTIC PROTEINASE)          04-NOV-91   3PHV
TITLE     X-RAY ANALYSIS OF HIV-1 PROTEINASE AT 2.7 ANGSTROMS
TITLE    2 RESOLUTION CONFIRMS STRUCTURAL HOMOLOGY AMONG RETROVIRAL
TITLE    3 ENZYMES
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: UNLIGANDED HIV-1 PROTEASE;
  .
  .
  .
SEQRES   1 A   99  PRO GLN ILE THR LEU TRP GLN ARG PRO LEU VAL THR ILE
SEQRES   2 A   99  LYS ILE GLY GLY GLN LEU LYS GLU ALA LEU LEU ASP THR
SEQRES   3 A   99  GLY ALA ASP ASP THR VAL LEU GLU GLU MET SER LEU PRO
SEQRES   4 A   99  GLY ARG TRP LYS PRO LYS MET ILE GLY GLY ILE GLY GLY
SEQRES   5 A   99  PHE ILE LYS VAL ARG GLN TYR ASP GLN ILE LEU ILE GLU
SEQRES   6 A   99  ILE CYS GLY HIS LYS ALA ILE GLY THR VAL LEU VAL GLY
SEQRES   7 A   99  PRO THR PRO VAL ASN ILE ILE GLY ARG ASN LEU LEU THR
SEQRES   8 A   99  GLN ILE GLY CYS THR LEU ASN PHE
  .
  .                                   X       Y       Z
  .
ATOM      1  N   PRO A   1      22.644  34.004  35.541  1.00  0.00          N
ATOM      2  CA  PRO A   1      23.698  34.424  34.629  1.00  0.00          C
ATOM      3  C   PRO A   1      23.670  33.634  33.311  1.00  0.00          C
ATOM      4  O   PRO A   1      23.732  32.407  33.378  1.00  0.00          O
ATOM      5  CB  PRO A   1      24.942  33.969  35.398  1.00  0.00          C
ATOM      6  CG  PRO A   1      24.473  32.997  36.472  1.00  0.00          C
ATOM      7  CD  PRO A   1      23.105  33.581  36.872  1.00  0.00          C
ATOM      8  N   GLN A   2      23.620  34.346  32.222  1.00  0.00          N
ATOM      9  CA  GLN A   2      23.686  33.843  30.844  1.00  0.00          C
ATOM     10  C   GLN A   2      25.109  34.080  30.312  1.00  0.00          C
ATOM     11  O   GLN A   2      25.656  35.175  30.522  1.00  0.00          O
ATOM     12  CB  GLN A   2      22.644  34.435  29.949  1.00  0.00          C
ATOM     13  CG  GLN A   2      23.093  34.632  28.515  1.00  0.00          C
ATOM     14  CD  GLN A   2      24.214  35.667  28.411  1.00  0.00          C
ATOM     15  OE1 GLN A   2      25.432  35.285  28.025  1.00  0.00          O
ATOM     16  NE2 GLN A   2      23.974  36.937  28.720  1.00  0.00          N
ATOM     17  N   ILE A   3      25.696  33.055  29.732  1.00  0.00          N
ATOM     18  CA  ILE A   3      27.062  33.029  29.263  1.00  0.00          C
ATOM     19  C   ILE A   3      27.209  32.567  27.802  1.00  0.00          C
ATOM     20  O   ILE A   3      26.648  31.543  27.438  1.00  0.00          O
ATOM     21  CB  ILE A   3      27.898  32.019  30.081  1.00  0.00          C
ATOM     22  CG1 ILE A   3      27.202  30.675  30.070  1.00  0.00          C
ATOM     23  CG2 ILE A   3      28.195  32.529  31.457  1.00  0.00          C
ATOM     24  CD1 ILE A   3      26.556  30.287  31.392  1.00  0.00          C
  .
  .
  .
```

HIV-1 Protease CA Coordinate Data

```
#!/usr/bin/perl
open(PDB,"3PHV.pdb");
open(OUTPUT, ">3PHV_CA_coords.txt");
while(<PDB>){
        chomp($_);
        @linevector=split(/\s+/,$_);
        if($linevector[0] eq 'ATOM' && $linevector[2] eq 'CA'){
                print OUTPUT "@linevector\n";
        }
}
close(OUTPUT);
close(PDB);
```

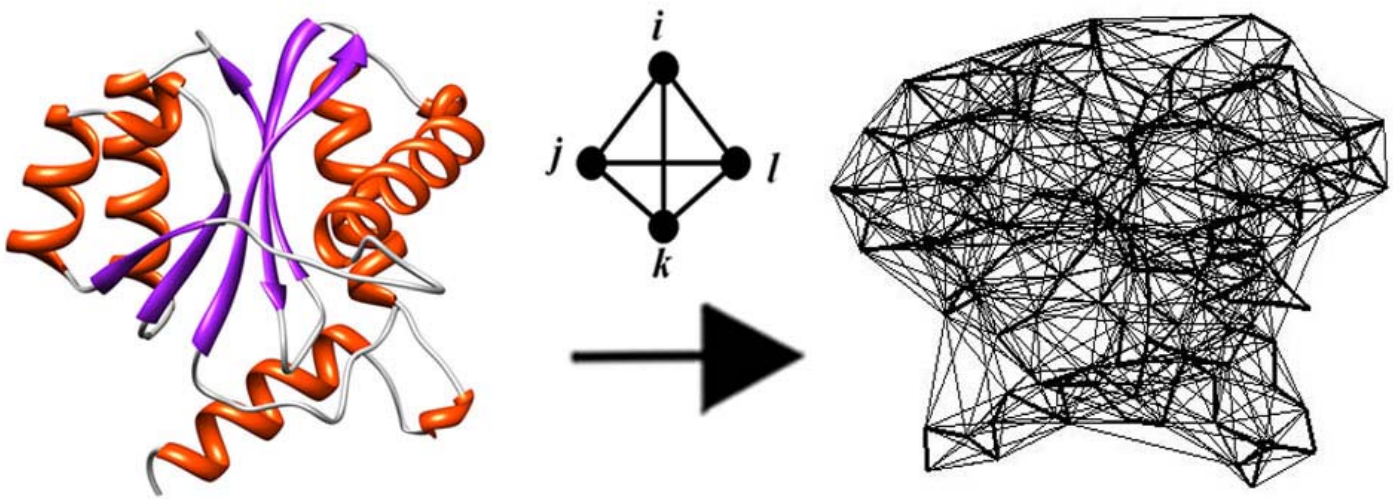
	A	B	C	D	E	F	G	H
1						X	Y	Z
2	ATOM	CA	PRO	A	1	23.698	34.424	34.629
3	ATOM	CA	GLN	A	2	23.686	33.843	30.844
4	ATOM	CA	ILE	A	3	27.062	33.029	29.262
5	ATOM	CA	THR	A	4	28.426	33.077	25.718
6	ATOM	CA	LEU	A	5	30.738	30.518	24.158
7	ATOM	CA	TRP	A	6	33.436	32.724	22.604
8	ATOM	CA	GLN	A	7	35.862	31.228	25.107
9	ATOM	CA	ARG	A	8	35.677	28.307	27.53
10	ATOM	CA	PRO	A	9	32.728	28.303	29.863
11	ATOM	CA	LEU	A	10	34.326	28.493	33.308
12	ATOM	CA	VAL	A	11	32.406	29.637	36.403
13	ATOM	CA	THR	A	12	33.031	29.494	40.159
14	ATOM	CA	ILE	A	13	31.807	26.736	42.446
15	ATOM	CA	LYS	A	14	31.406	25.988	46.122
16	ATOM	CA	ILE	A	15	31.756	22.457	47.446
17	ATOM	CA	GLY	A	16	31.721	22.691	51.261
18	ATOM	CA	GLY	A	17	33.076	26.171	51.947
19	ATOM	CA	GLN	A	18	35.737	25.835	49.251
20	ATOM	CA	LEU	A	19	35.495	28.32	46.372
21	ATOM	CA	LYS	A	20	36.239	26.546	43.058
22	ATOM	CA	GLU	A	21	36.094	26.838	39.258
23	ATOM	CA	ALA	A	22	34.676	24.579	36.537
24	ATOM	CA	LEU	A	23	33.434	24.022	33.005

Counting Interacting Amino Acids: One Approach

- Consider pairs of neighbor amino acids whose points are within a given distance of each other in the structure
- 20 x 20 = 400 possible ordered pairs (i.e., AC and CA different)
  No permutation (more approp.): 20 + (20 choose 2) = 210 pairs
- Obtain all pairs from a large diverse set of proteins, and calculate observed relative frequency of interaction for each pair f_{ij}
- Calculate a rate expected by chance for each pair by using a multinomial distribution (n = 2 trials, k = 20 outcomes) p_{ij}
- Inverted Boltzmann principle: propensity for pairwise interaction, also known as a "knowledge-based" empirical potential energy of pairwise interaction, is proportional to $s_{ij} = \log (f_{ij} / p_{ij})$

Our Approach: Four-Body Interactions

- Identify a diverse set of over 1400 structures of protein chains
- Apply Delaunay tessellation (3D) to amino acid points of each protein: convex hull of space-filling tetrahedra, each objectively identifies a quadruplet of nearest neighbor amino acids
- Qhull (free) at http://www.qhull.org, or Matlab (delaunay3)
- Tessellation edges longer than 12 Angstroms removed



Counting Amino Acid Quadruplets

n **= size of amino acid alphabet = 20;** ***r*** **= size of the subsets = 4**

Repetitions Allowed?	Permutations Allowed?	Number of Quadruplets
yes	yes	$n^r = 20^4 = 160{,}000$
yes	no	$\binom{n+r-1}{r} = \binom{23}{4} = 8855$
no	yes	$\frac{n!}{(n-r)!} = \frac{20!}{16!} = 116{,}280$
no	no	$\frac{n!}{r!(n-r)!} = \binom{n}{r} = \binom{20}{4} = 4845$

only realistic choice for proteins

only realistic choice when identifying quadruplets of interacting amino acids based on the four unordered vertices of tetrahedra in a protein tessellation

Counting Amino Acid Quadruplets

Repetitions – yes, permutations – no:

a more "hands-on" counting approach

$\underbrace{C}\ \underbrace{D}\ \underbrace{E}\ \underbrace{F}$	$\binom{20}{4}$
$C\ \ C\ \ \underbrace{D}\ \underbrace{E}$	$20 \cdot \binom{19}{2}$
$\underbrace{C\ \ C}\ \underbrace{D\ \ D}$	$\binom{20}{2}$
$C\ \ C\ \ C\ \ D$	$20 \cdot 19$
$C\ \ C\ \ C\ \ C$	20

Total: 8,855 distinct quadruplets

Four-Body Statistical Potential

- Knowledge-based, modeled after the inverted Boltzmann principle from statistical mechanics
- f_{ijkl} = observed proportion of all tetrahedra in the 1400+ tessellations whose four vertex amino acid residues are *i,j,k,l*
- p_{ijkl} = rate expected by chance (multinomial distribution, based on proportions of amino acids of types *i,j,k,l* in the 1400+ proteins)
- For amino acid quadruplet (*i,j,k,l*), a log-likelihood score (energy of interaction) is given by $s(i,j,k,l) = \log(f_{ijkl} / p_{ijkl})$
- Four-body statistical potential: the collection of 8855 amino acid quadruplet types with their respective scores

Multinomial Reference Distribution

$n =$ number of independent trials of an experiment

$k =$ number of mutually exclusive and exhaustive outcomes for the experiment, say $A_1, A_2, \ldots, A_k$

$P(A_i) = p_i, i = 1, 2, \ldots, k$ on each trial with $\sum_{i=1}^{k} p_i = 1$

Let random variable X_i be the number of times A_i occurs in the n trials, $i = 1, 2, \ldots, k$.

If $x_1, x_2, \ldots, x_k$ are nonnegative integers such that $\sum_{i=1}^{k} x_i = n$, then the probability that A_i occurs x_i times, $i = 1, 2, \ldots, k$ is given by

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

In our case, each experiment consists of selecting an amino acid ($k = 20$), and there are $n = 4$ trials..

Each A_i represents a different amino acid type, where p_i is the proportion of all amino acids in the $1400+$ proteins that are of type i, and x_i is the number of times that amino acid A_i occurs in the quadruplet. So,

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_{20} = x_{20}) = \frac{4!}{\prod_{i=1}^{20} x_i!} \prod_{i=1}^{20} p_i^{x_i}$$

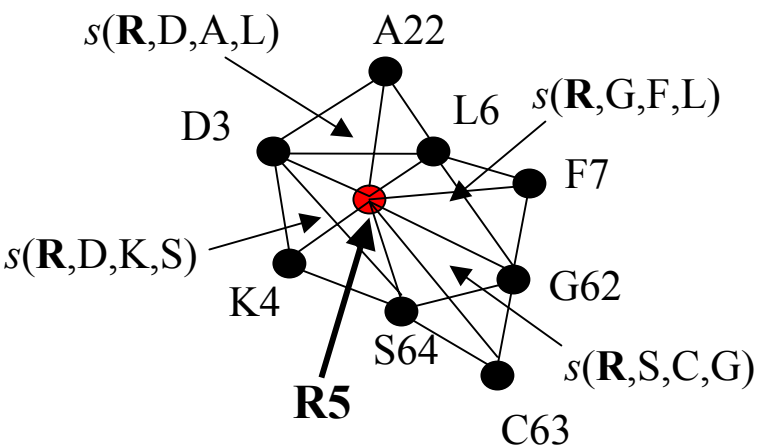
is the random chance of occurence of any given quadruplet, where $\sum_{i=1}^{20} x_i = 4$.

Four-Body Statistical Potential

Amino Acid Quadruplet	"Pseudo-Energy" Log-likelihood s(i,j,k,l)
CCCC	3.29042538
CCCH	2.09542785
CCCS	1.96177162
CCCG	1.84022021
CCCI	1.79961166
CCCF	1.77139046
CCCT	1.76378293
CCCP	1.74840641
ACCC	1.74777711
CCCW	1.74711265
CCHH	1.70747111
CCCN	1.69741431
HHHH	1.61473339
.	.
.	.
.	.
HMNP	0.000221495
DGGY	0.000178988
DRSV	9.45855E-05
EHHV	4.979E-06
LRYY	-6.29797E-05
DGKP	-9.73563E-05
NPSS	-0.000100914
IPRW	-0.000136526
MMRT	-0.000168007
GLLP	-0.000294376
EKNT	-0.000312593
EKQR	-0.000343148
.	.
.	.
.	.
HKKW	-0.66398714
KKKP	-0.66875323
CDEQ	-0.67215257
CKKW	-0.75315166
CDDM	-0.76390474
HHKK	-0.85974
CKKR	-0.88002907
CIKR	-0.90372634
CHKW	-0.94458122
CEEE	-1.02439761
HKKM	-1.14234339

Amino Acid (Residue) Environment Scores

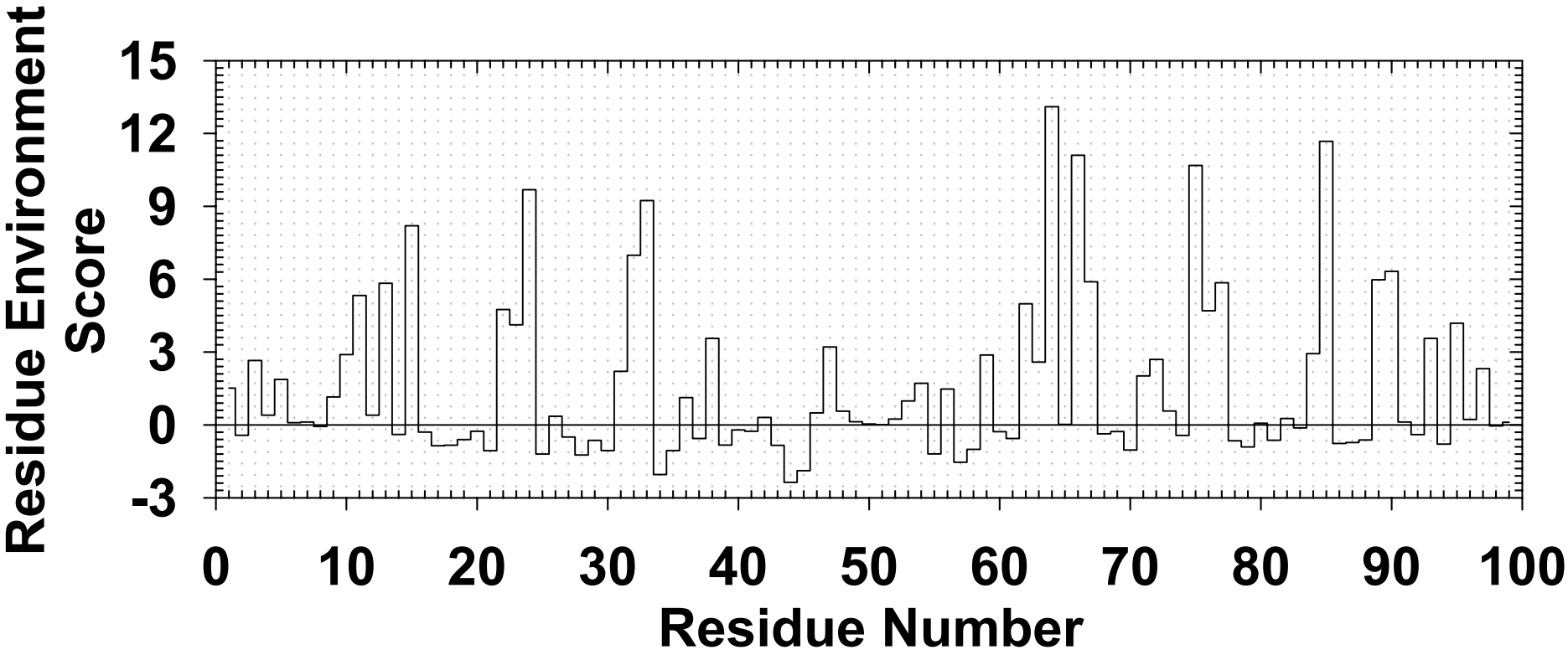
- For each amino acid position, locally sum scores $s(i,j,k,l)$ of the tetrahedral quadruplets that use the position as a vertex



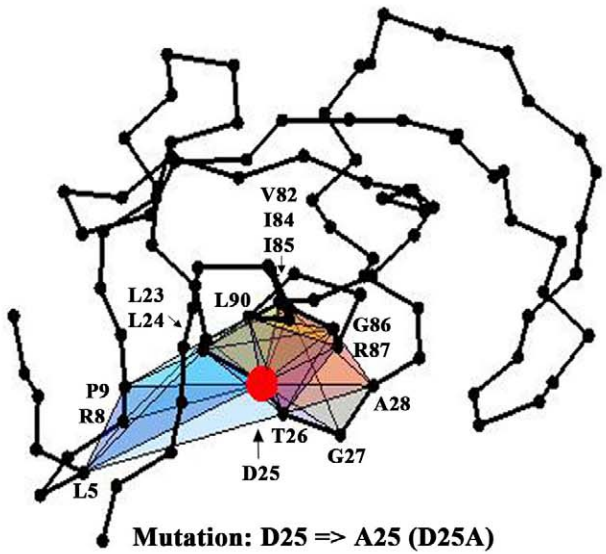
Example: $q_5 = q(\mathrm{R5}) = \sum_{(i,j,k,l)} s(i,j,k,l)$, sum is taken **only** over all tetrahedral quadruplets (i,j,k,l) that include R5

- The scores q_i of all amino acid residue positions i in a protein form a **3D-1D Potential Profile** vector $\mathbf{Q} = < q_1, q_2, q_3, \ldots, q_N >$ (N = length of the protein sequence in the structure)

3D-1D Potential Profile: HIV-1 Protease



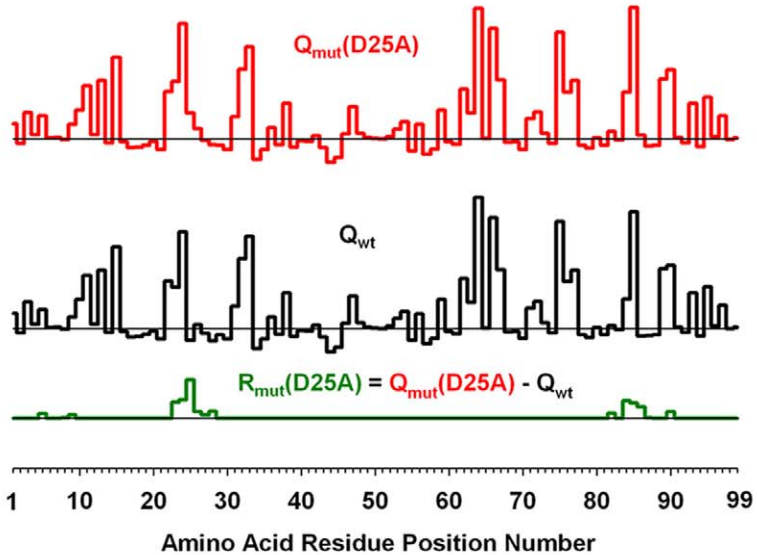
Computational Mutagenesis: HIV-1 Protease



(A) Mutant Profile

(B) Native Profile

(A) – (B) = R
Environmental
Perturbation Scores



Profile R used for representing mutant

Experimental Data

- Ritonavir is one of many HIV-1 protease inhibitor drugs, amino acid mutations in protease alter its susceptibility to the drugs
- Susceptibility given by a fold change (FC) value, which can be obtained for each distinct mutant protein by a phenotypic test (expensive and has a long turnaround time)
- Sequencing patient virus (genotypic test) is much faster, cheaper; hence, high interest in predicting phenotype from genotype!
- Dataset: 473 mutant HIV-1 protease proteins, each with an already known phenotype (FC value) WRT drug ritonavir; can be categorized as Sensitive/Resistant (FC cutoff known)
- Question: Can we predict mutant FC values or classes (output) based on R vectors of environmental perturbation scores (inputs)

Experimental Data

Genotype-Phenotype Datasets - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Genotype-Phenotype Datasets

hivdb.stanford.edu/cgi-bin/GenoPhenoDS.cgi

Latest Headlines | HIV DBs | Binf | Res | Libs | GMU | Accts | Services | Search | Schools | Tutorials | News | Misc | Confs | Journals

STANFORD UNIVERSITY
HIV DRUG RESISTANCE DATABASE
A curated public database designed to represent, store, and analyze the divergent forms of data underlying HIV drug resistance.

HOME GENOTYPE-RX GENOTYPE-PHENO GENOTYPE-CLINICAL HIVdb PROGRAM

Genotype-Phenotype Datasets

Version 5.0, March, 2012

- To access high quality filtered datasets from HIVDB and analyses using methods described in the paper by Rhee SY, Taylor J, Wadhera G, Ben-Hur A, Brutlag DL, Shafer RW., **Genotypic predictors of human immunodeficiency virus type 1 drug resistance.**,*Proceedings of National Academy of Sciences of the United States of America*, Oct 25, 2006, click here.
- To access complete unfiltered datasets from HIVDB by gene and phenotype assay, click the links below:

Gene	Method	Data
PR	PhenoSense (ViroLogic™)	11731 phenotype results from 1727 isolates: Tab delimited | Comma separated (Excel)
	Antivirogram (Virco™)	10026 phenotype results from 1434 isolates: Tab delimited | Comma separated (Excel)
	All Others	1040 phenotype results from 319 isolates: Tab delimited | Comma separated (Excel)
RT	PhenoSense (ViroLogic™)	8884 phenotype results from 1033 isolates: Tab delimited | Comma separated (Excel)
	Antivirogram (Virco™)	12357 phenotype results from 1748 isolates: Tab delimited | Comma separated (Excel)
	All Others	1286 phenotype results from 356 isolates: Tab delimited | Comma separated (Excel)

Description of fields in the datasets

Field Name	Description

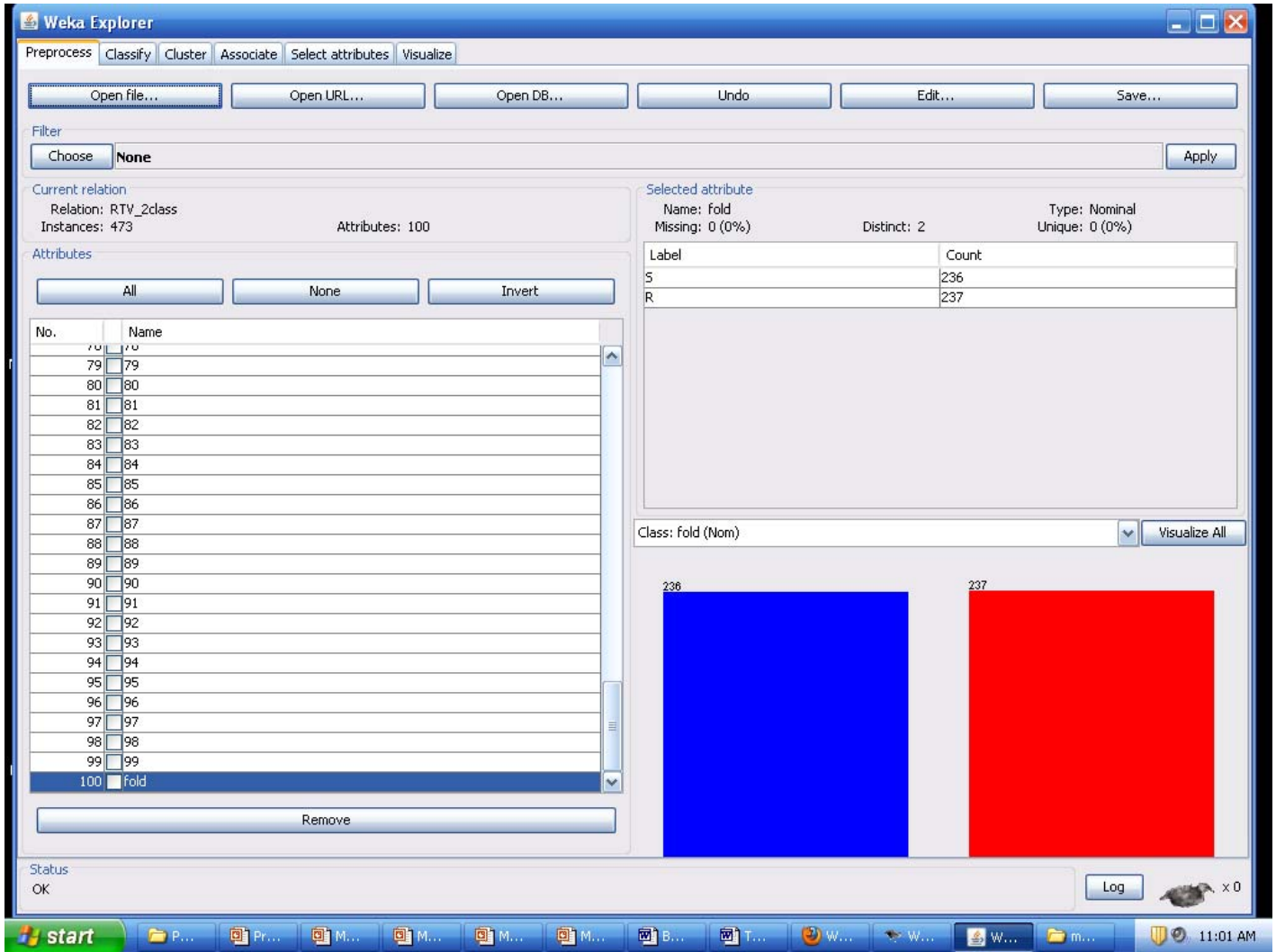
start | Paper | Predic... | MAAta... | MAA_... | MAA2... | MAA2... | BMCR... | Table1... | Genot... | 10:02 AM

Statistical Machine Learning Algorithms

- Classification or regression tree, neural network, support vector machine or regression, random forest, Bayesian network, etc
- Predictive models are trained using the available data, learned models are complex nonlinear functions of the inputs
- Free software: Weka (http://www.cs.waikato.ac.nz/ml/weka/)
- User friendly GUI, opens the door to discussing concepts such as:
  - model training, validation, and testing
  - evaluating model performance using resubstitution (training set itself), independent test set, cross-validation, % split
  - defining measures (accuracy, sensitivity, specificity, precision, kappa stat, Matthew's / Pearson's correlation, ROC curves, etc)

Comma Separated Data File for Weka

For each protease mutant, R vector components (inputs) separated by commas, and FC value (or class label) as last component (output)



Classification

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose: J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) fold

Start | Stop

Result list (right-click for options)

- 10:21:15 - trees.REPTree
- 10:36:10 - trees.J48

Classifier output

```
Number of Leaves  :     12

Size of the tree :      23


Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         445               94.0803 %
Incorrectly Classified Instances        28                5.9197 %
Kappa statistic                          0.8816
Mean absolute error                      0.0687
Root mean squared error                  0.2338
Relative absolute error                 13.7376 %
Root relative squared error             46.7667 %
Total Number of Instances              473

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.928     0.046      0.952     0.928     0.94     S
  0.954     0.072      0.93      0.954     0.942    R

=== Confusion Matrix ===

   a   b   <-- classified as
 219  17 |   a = S
  11 226 |   b = R
```

Weka GUI Choo...

Waikato Environment for Knowledge Analysis

Version 3.4.12

(c) 1999 - 2007
University of Waikato
New Zealand

GUI

Simple CLI | Explorer | Experimenter | KnowledgeFlow

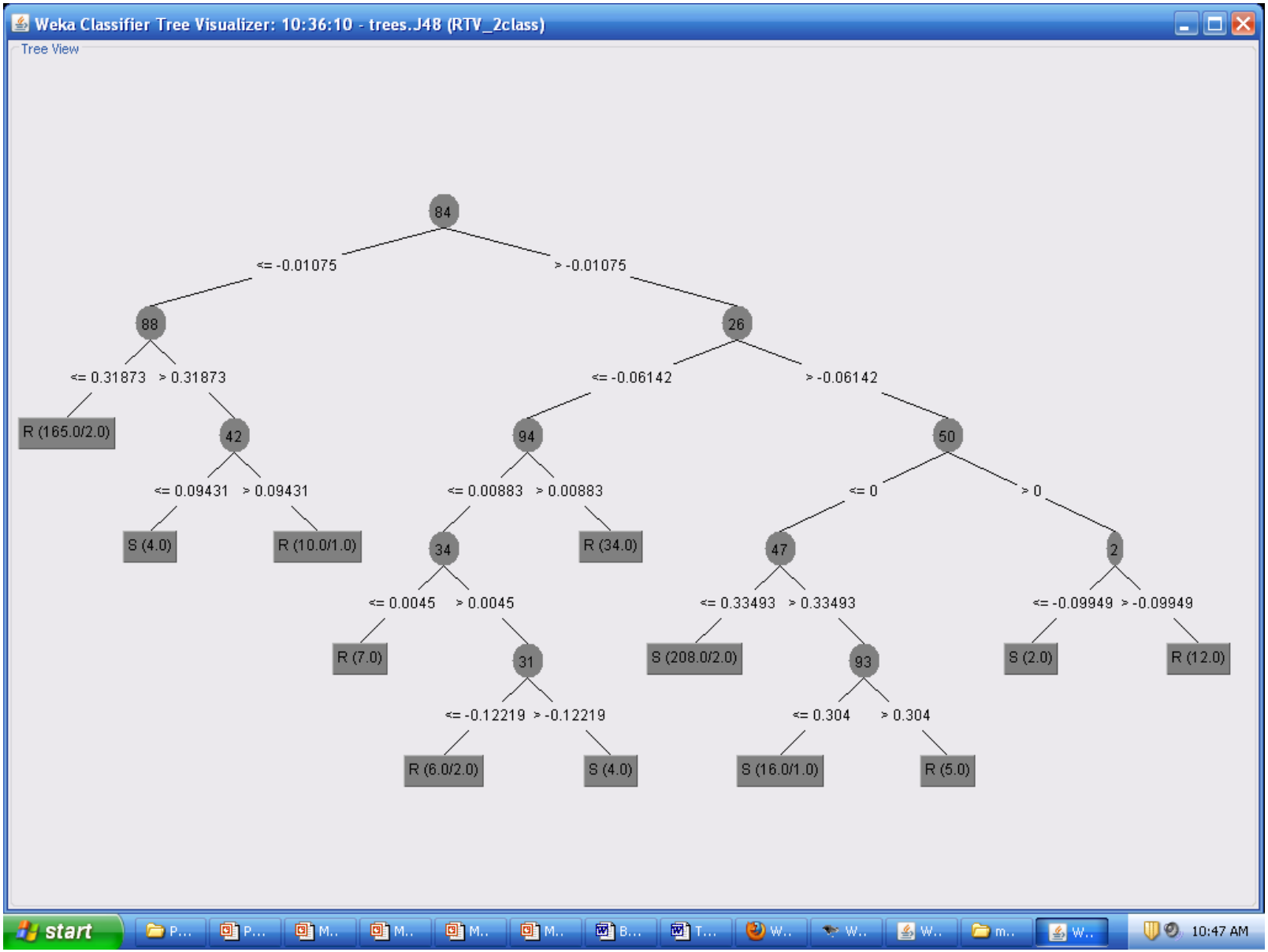
Status: OK

Log

x 0

start | 10:43 AM

Trained Classification Tree Model



Regression

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose: J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) fold

Start | Stop

Result list (right-click for options)

- 10:21:15 - trees.REPTree
- 10:36:10 - trees.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1
Relation:     RTV_train
Instances:    473
Attributes:   100
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8595
Mean absolute error                      0.3886
Root mean squared error                  0.5499
Relative absolute error                 41.2769 %
Root relative squared error             51.2453 %
Total Number of Instances              473
```

Weka GUI Choo...

Waikato Environment for Knowledge Analysis

Version 3.4.12

(c) 1999 - 2007 University of Waikato New Zealand

GUI

Simple CLI | Explorer | Experimenter | KnowledgeFlow

Status: OK

Log

x 0

start | P... | Pr... | M... | M... | M... | M... | B... | T... | W... | W... | W... | m... | 10:51 AM

Relevant Links

- These slides: http://binf.gmu.edu/mmasso/JMM2014.pdf
- Delaunay tessellation software:
  Qhull (free) at http://www.qhull.org, or Matlab (delaunay3)
- Weka machine learning software (free):
  http://www.cs.waikato.ac.nz/ml/weka/)
- PDB codes for dataset of 1417 diverse protein structures:
  http://proteins.gmu.edu/automute/tessellatable1417.txt
- Four-body statistical potential derived from above dataset:
  http://proteins.gmu.edu/automute/potential_1417_cut12.txt
- Weka-formatted datasets of 473 HIV-1 protease mutants with known phenotypes, represented using our *in silico* method:
  - (regression) http://proteins.gmu.edu/automute/RTV_train.csv
  - (class.) http://proteins.gmu.edu/automute/RTV_2class.csv