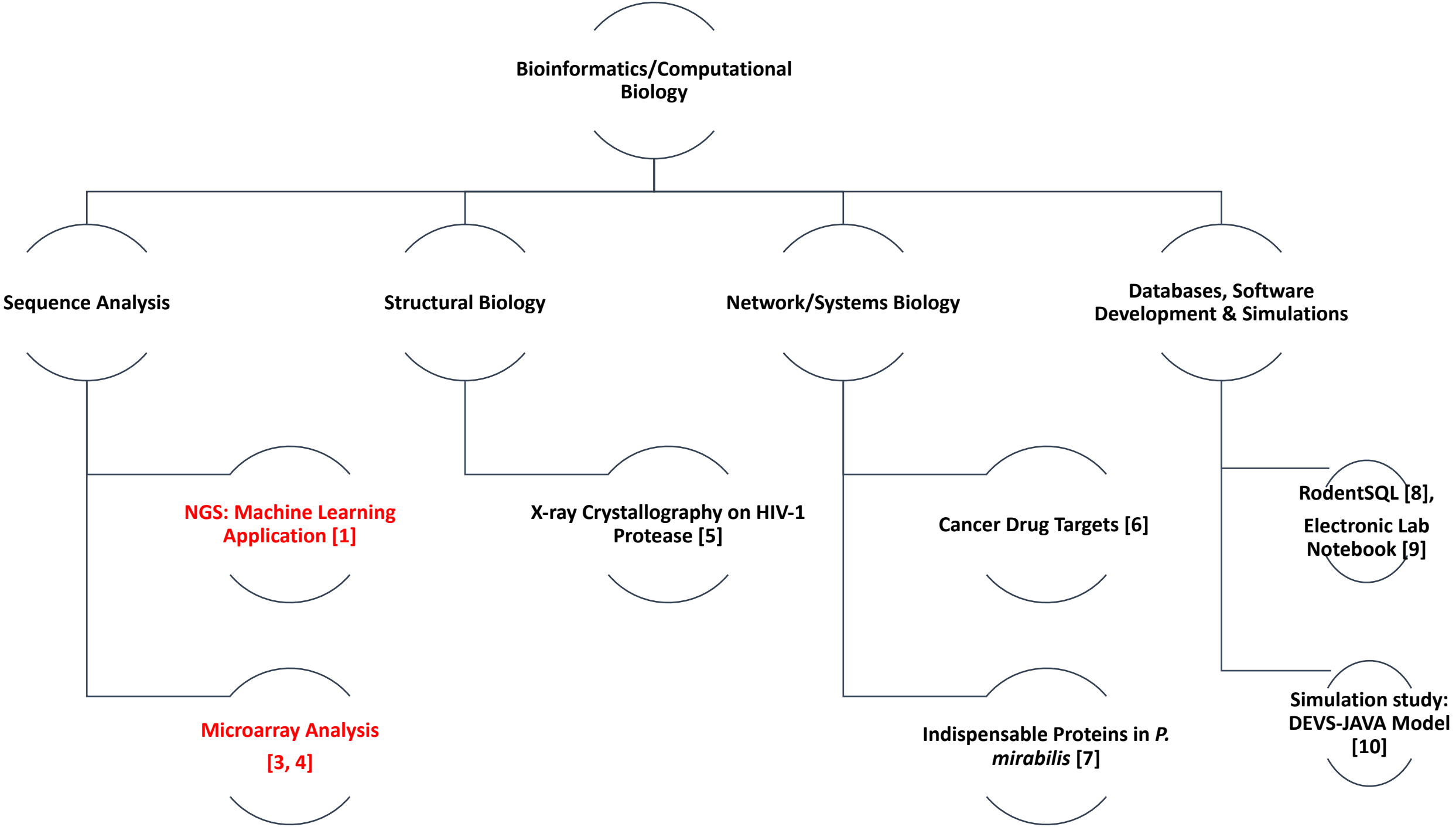
Machine learning techniques for analyzing and interpreting genomics and proteomics data

Shrikant Pawar, Ph.D.

Claflin University

https://www.claflin-computation.com/

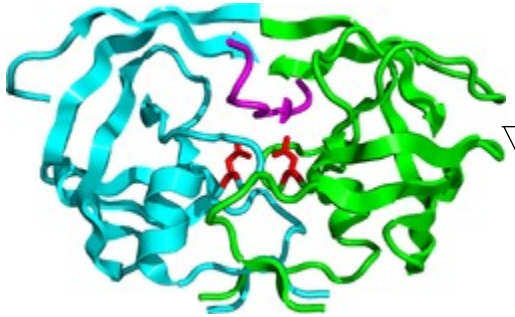

Bioinformatics/Computational Biology
Sequence Analysis
Structural Biology
Network/Systems Biology
Databases, Software Development & Simulations
NGS: Machine Learning Application [1]
Microarray Analysis [3, 4]
X-ray Crystallography on HIV-1 Protease [5]
Cancer Drug Targets [6]
Indispensable Proteins in P. mirabilis [7]
RodentSQL [8], Electronic Lab Notebook [9]
Simulation study: DEVS-JAVA Model [10]

Sequence Analysis: NGS

- Utilizing neural networks (Restricted Boltzmann machine's) and clustering algorithms to identify certain important, representative HIV-1 PR sequences from a pool of several hundred sequences.

1. Analysis of drug resistance in HIV protease, ***Shrikant Pawar***, Chris Freas, Robert W. Harrison, and Irene T. Weber, ***BMC: Bioinformatics***
2. Structural studies of antiviral inhibitor with HIV-1 protease bearing drug resistant substitutions of V32I, I47V and V82I, ***Shrikant Pawar****, Yuan-FangWang, Andres Wong-Sam, Johnson Agniswamy, Arun K. Ghosh, Robert W. Harrison, and Irene T. Weber,* ***Elsevier: Biochemical and Biophysical Research Communications***

HIV-1 Protease Action

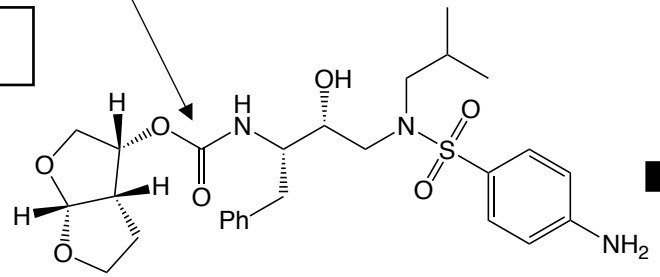


HIV-1 Protease

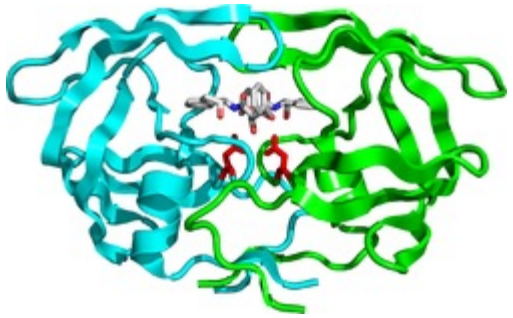
Hydrolyze peptide bonds on the Gag-Pol polyproteins

Mature, fully functional proteins

Infectious viral particle



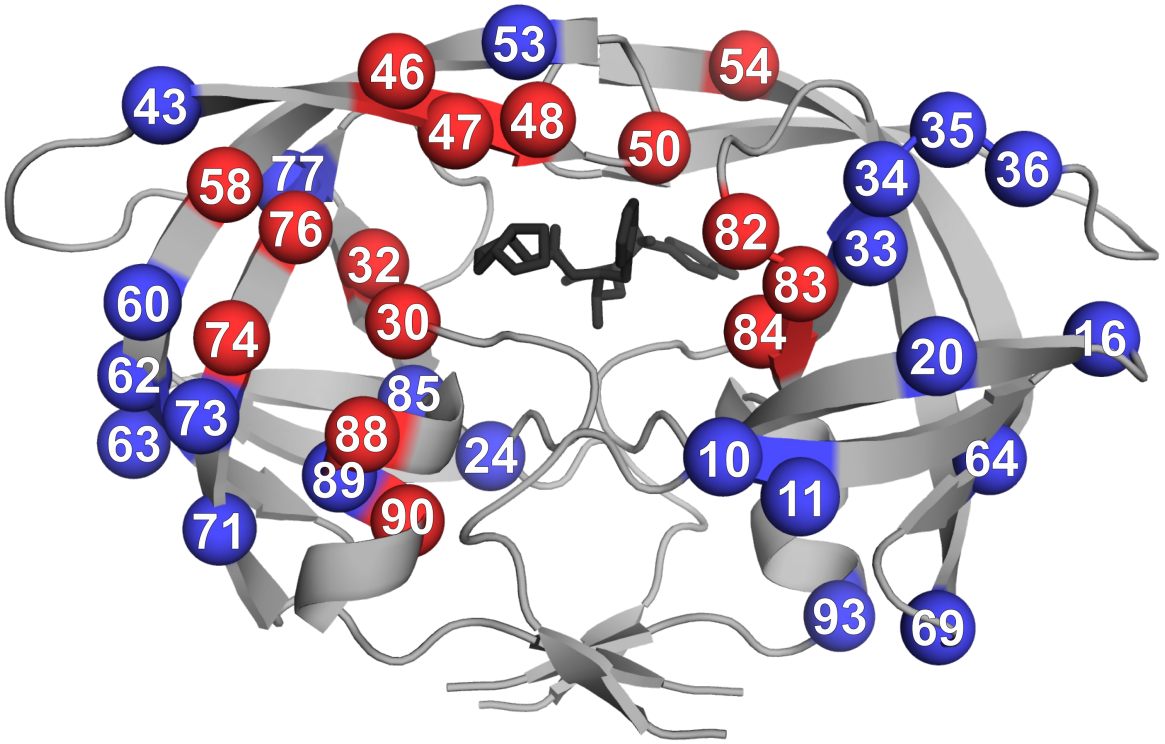
Protease inhibitor: DRV



Competitive inhibition of PR prevents virus maturation

No infectious viral particle

Drug resistance is a severe problem



PQVTLWQRPI VTIKIGGQLK EALLDTGADN TVLEEMSLPG KWKPKMIGGI GGFIKVRQYD QVSIEICGHI AIGTVLIGPT PVNIIGRNLL TQLGCTLNF

Patient 1

Patient 2

Patient 3

Patient 4

Patient 5

Patient 6

Patient 7

Patient n

~100,000 sequences

Major and **minor** mutations
associated with resistance to all clinical protease inhibitors

Adapted from Weber, Kneller, Wong-Sam. Future Med Chem 2015

Stanford University

HIV DRUG RESISTANCE DATABASE

A curated public database to represent, store and analyze HIV drug resistance data.

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

Phenotype Data

Resistance Fold

Genotype Data

Next-generation deep sequencing

Overall 5000 isolates for PI, NRTI

Seq ID	FPV	IDV	NFV	SQV
12861	0.4	0.5	7.1	0.5

~100,000 sequences can be classified as resistant or non-resistant

PQVTLWQRPI VTIKIGGQLK EALLDTGADNTVLEEMSLPG KWKPKMIGGI GGFIKVRQYD QVSIEICGHI AIGTVLIGPT PVNIIGRNLL TQLGCTLNF

Can machine learning help in selecting few drug resistant PR sequences for structure guided drug design?

Analysis Pipeline

Encoding

Selection

Coordinates for WT HIV-1 PR with its protein sequence information

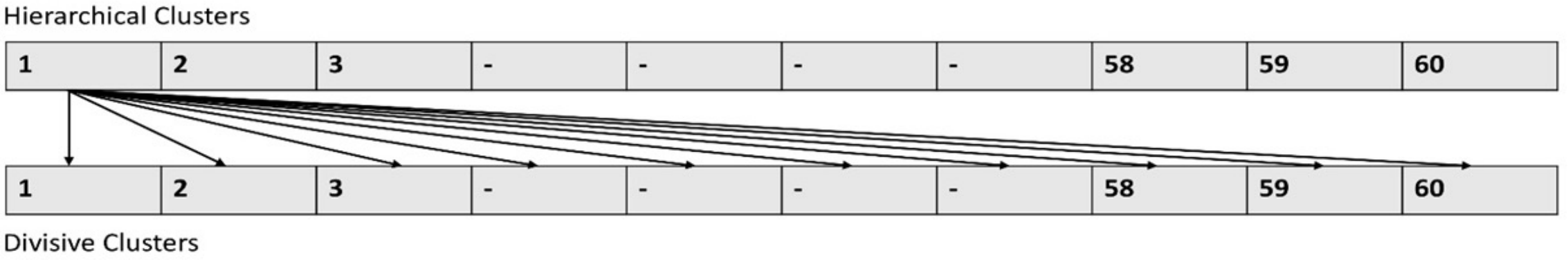
Encode WT protein sequence and structure with Delaunay triangulation

Represent all the isolate sequences from Stanford database in form of 210 vectors

Apply RBM, SVM and RF for checking encoding accuracy

K means, Agglomerative and Divisive Clustering

Select sequences using intersection amongst techniques

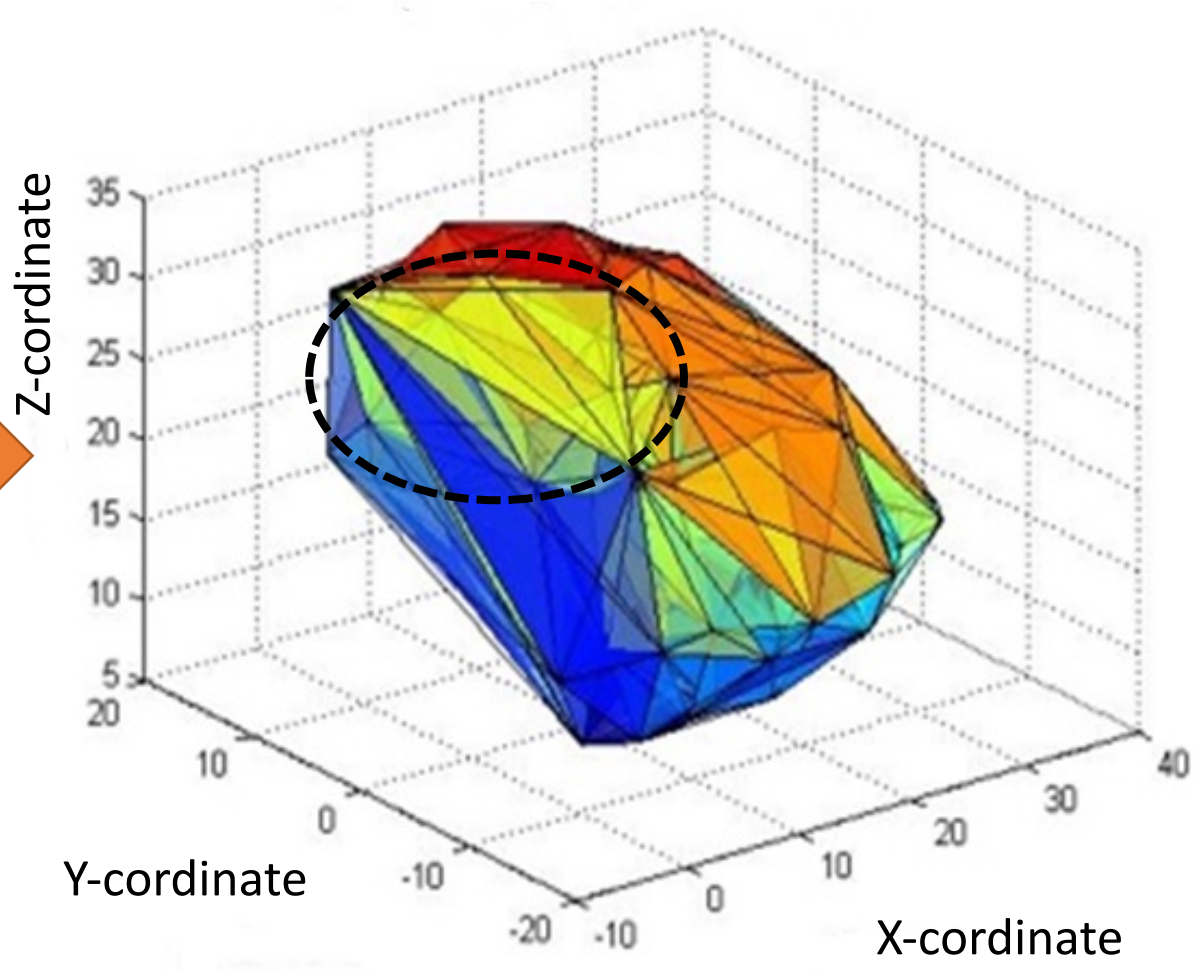


Encoding sequence-structure information

HIV-1 WT Protease

```
ATOM      1  N   PRO A   1      -12.889  38.692  31.300  1.00 24.90
ATOM      2  CA  PRO A   1      -12.932  39.182  29.909  1.00 22.91
ATOM      3  C   PRO A   1      -13.637  38.169  29.029  1.00 22.17
ATOM      4  O   PRO A   1      -14.041  37.131  29.511  1.00 21.69
ATOM      5  CB  PRO A   1      -11.480  39.420  29.521  1.00 22.50
ATOM      6  CG  PRO A   1      -10.712  38.712  30.620  1.00 23.24
ATOM      7  CD  PRO A   1      -11.562  38.972  31.858  1.00 22.41
ATOM      8  H2  PRO A   1      -13.120  37.673  31.261  1.00 -1.00
ATOM      9  H3  PRO A   1      -13.680  39.011  31.916  1.00 -1.00
```

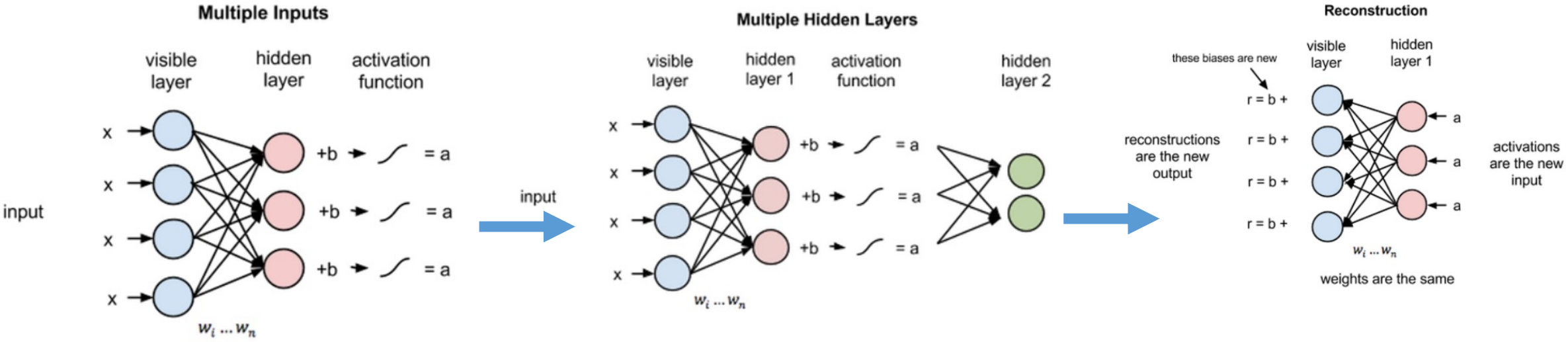
Delaunay triangulation on Wild Type HIV-1 Protease



A Delaunay triangulation for a set P of points in a plane is a triangulation DT(P) such that no point in P is inside the circumcircle of any triangle in DT(P).

Yu, X., Weber, I. T., & Harrison, R. W. Prediction of HIV drug resistance from genotype with encoded three-dimensional protein structure. *BMC genomics*, 2014

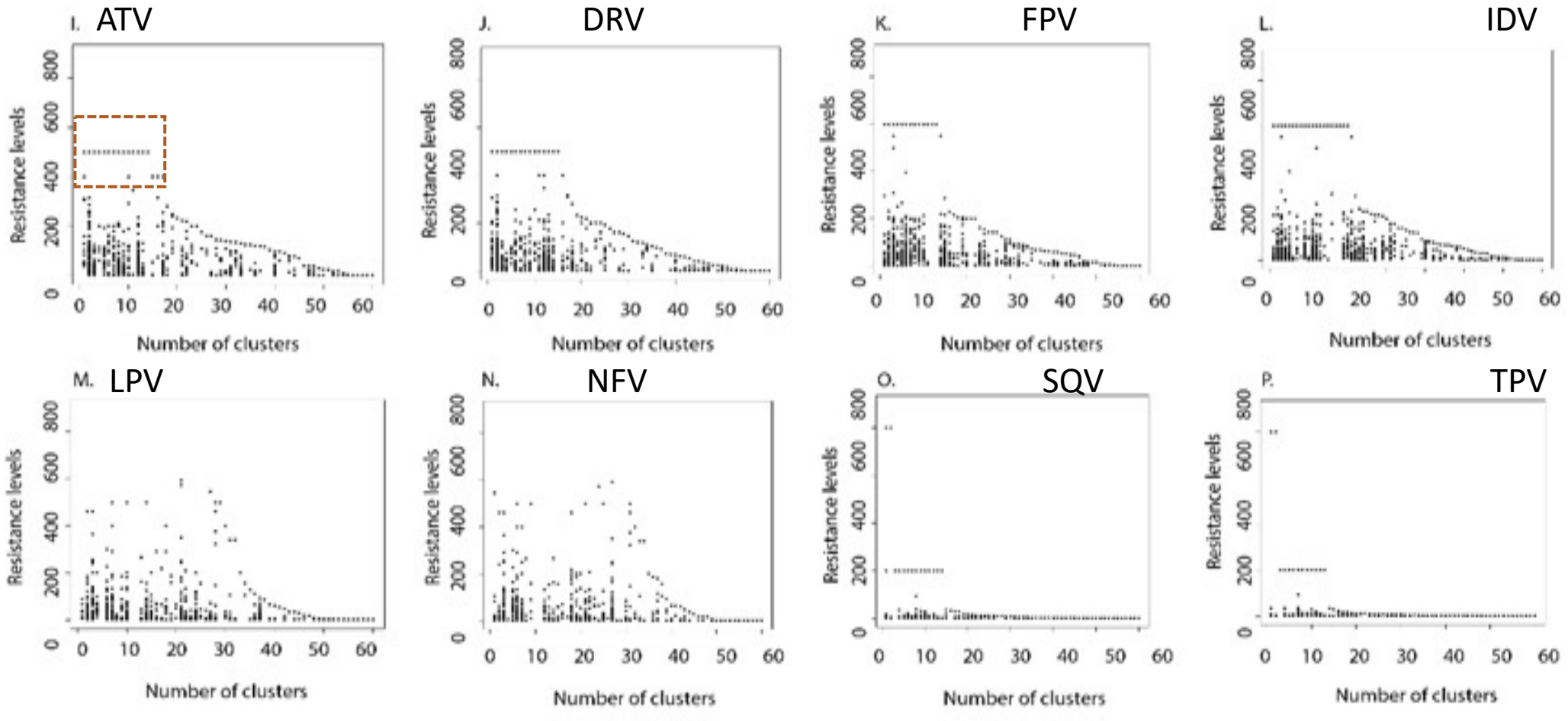
Restricted Boltzmann machine



$$\frac{dU}{dW_{i,j}} = H_j V_i - < \frac{dU}{dW_{i,j}} >$$

$$R = \frac{H_i \sum_j W_{i,j} V_j}{|H_i| \sum_j |W_{i,j}| \, |C_j|}$$ where C is the perfect reconstruction.

Most of the high resistance fold sequences with class 2 were clustered in first 10 clusters for most of the selected inhibitors through both hierarchical and divisive clustering delineating a clean separation between non-resistant and resistant sequences.



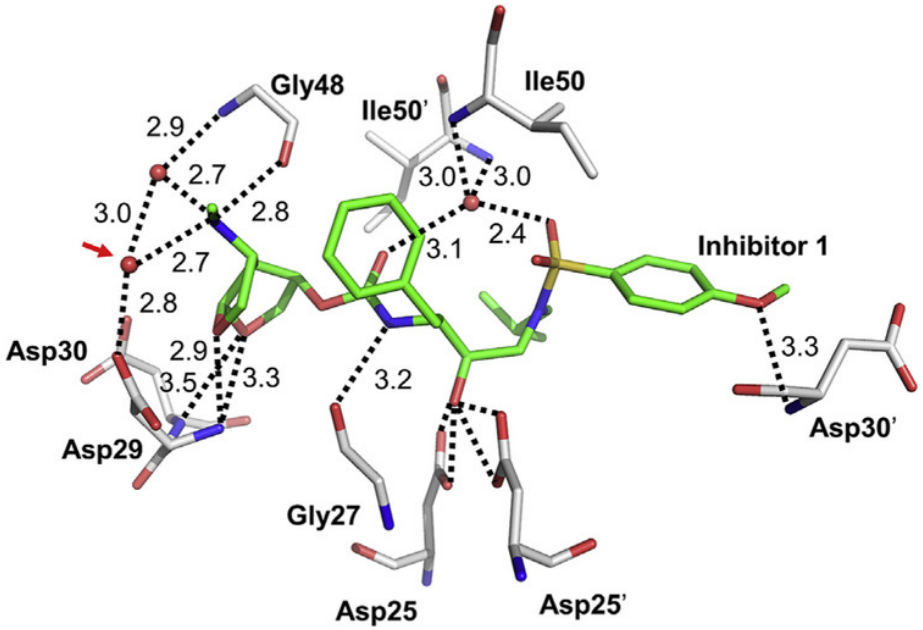
Divisive Clustering

From a pool of 100,000 only 2-35 sequences were selected common through all the 3 approaches, further utilized for structure guided drug design

Category	ATV	DRV	FPV	IDV	LPV	NFV	SQV	TPV
H, D and K	0	0	20 (66)	0	35 (61)	2 (12)	5 (58)	0

Numbers in parenthesis are the cluster from which they were selected.

1. The resistance status of the selected sequences should be identified.
2. Minimum number of sequences selected for inhibitors, NFV, SQV or LPV would be some of the ideal candidates for testing in laboratory.



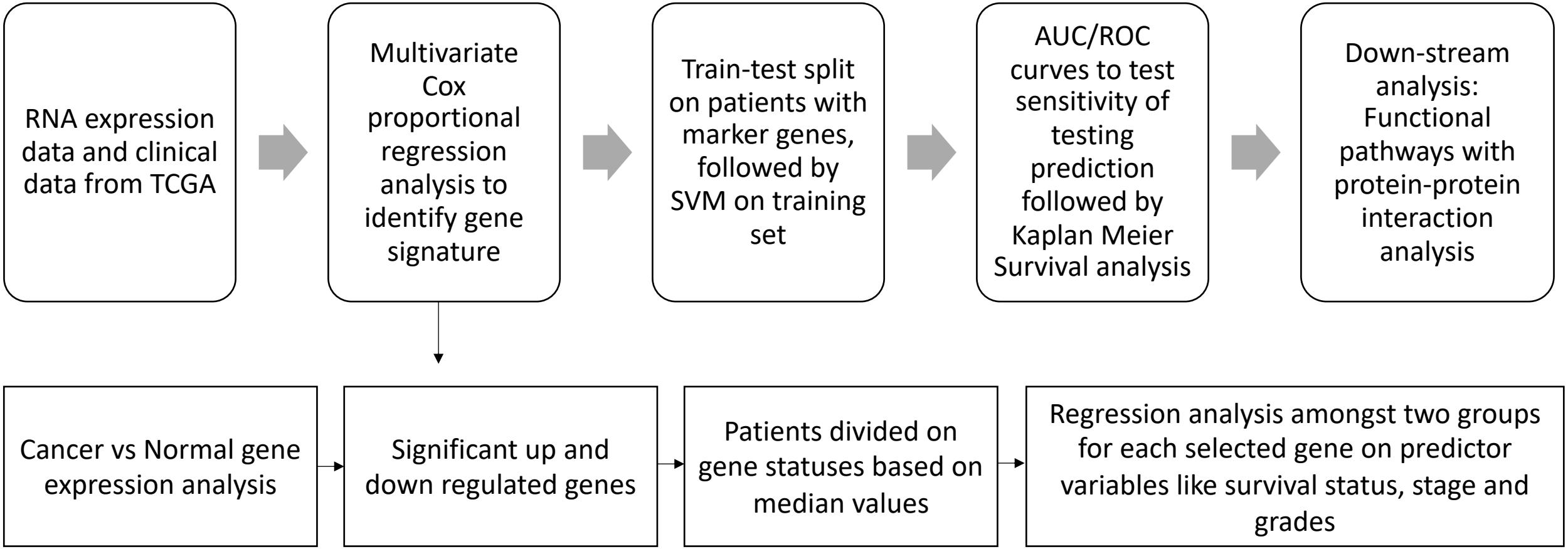
Sequence Analysis: NGS

- A Six-Gene-Based Prognostic Model Predicts Survival in Head and Neck Squamous Cell Carcinoma Patients.

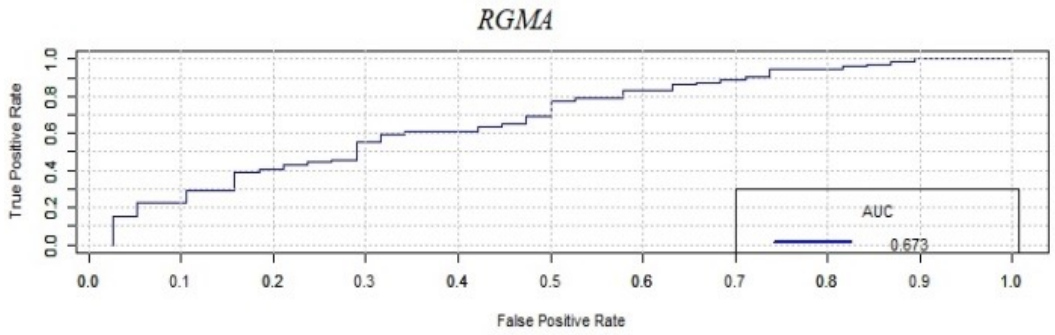
2. A Six-Gene-Based Prognostic Model Predicts Survival in Head and Neck Squamous Cell Carcinoma Patients, ***Shrikant Pawar*** and Aditya Stanam, ***Springer: Journal of Maxillofacial and Oral Surgery***

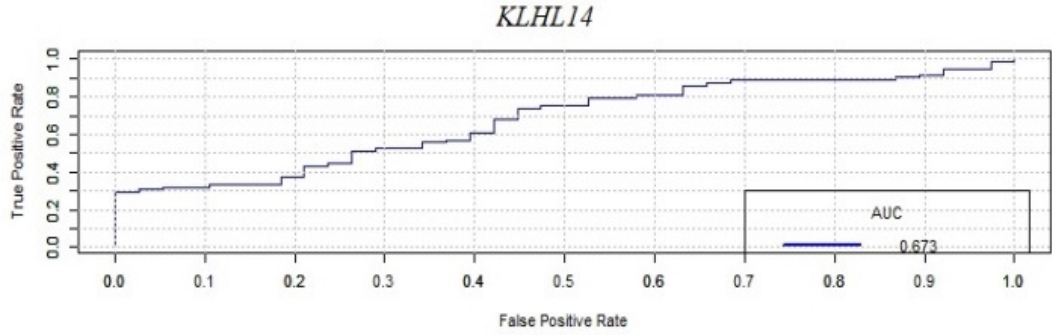
3. Common cancer biomarkers of breast and ovarian types identified through artificial intelligence, ***Shrikant Pawar****, Tuck Onn Liew, Aditya Stanam, Chandrajit Lahiri,* ***Wiley: Chemical Biology & Drug Design***

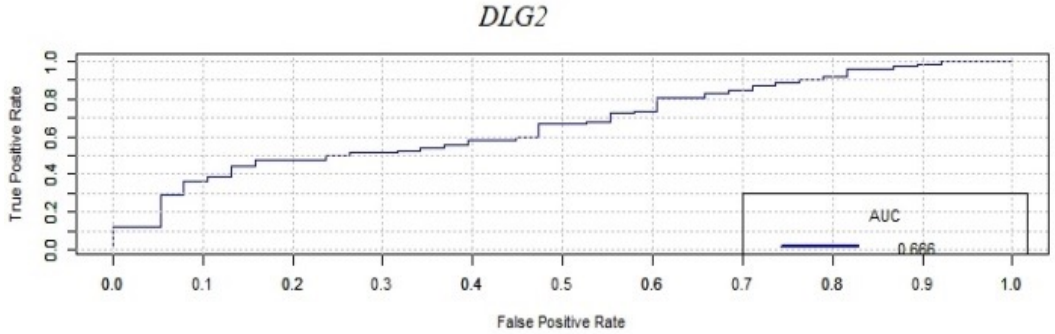
Analysis Pipeline

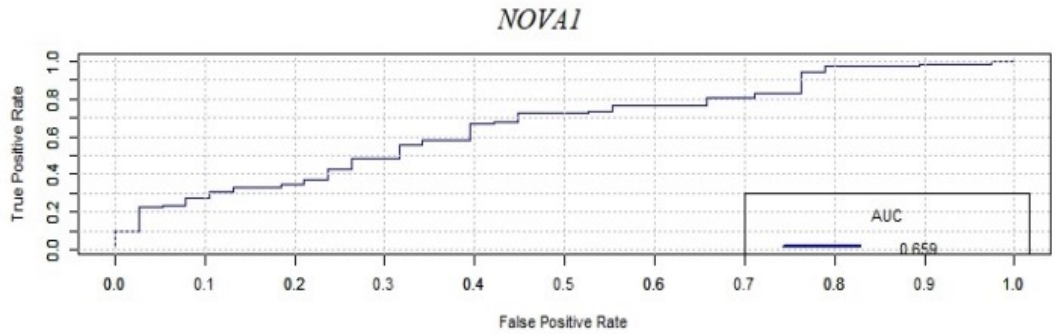


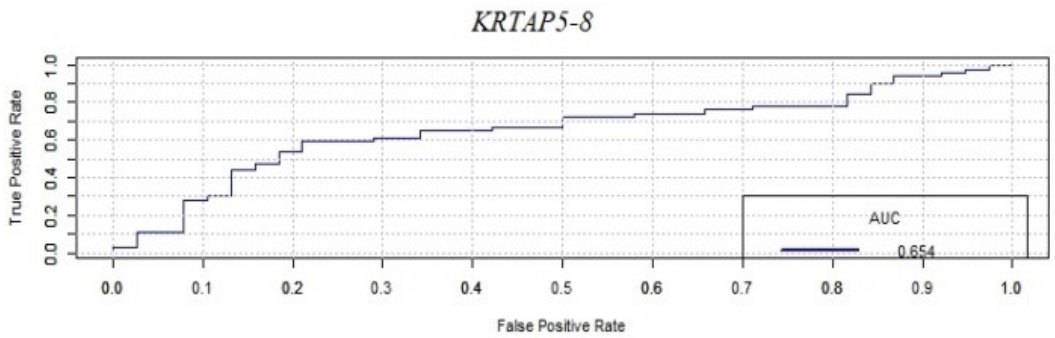
ROC curves, AUC values for the selected biomarker six genes, A reasonable prediction accuracy of 85.38, 85.89 and 86.20 % were found on test dataset with SVM

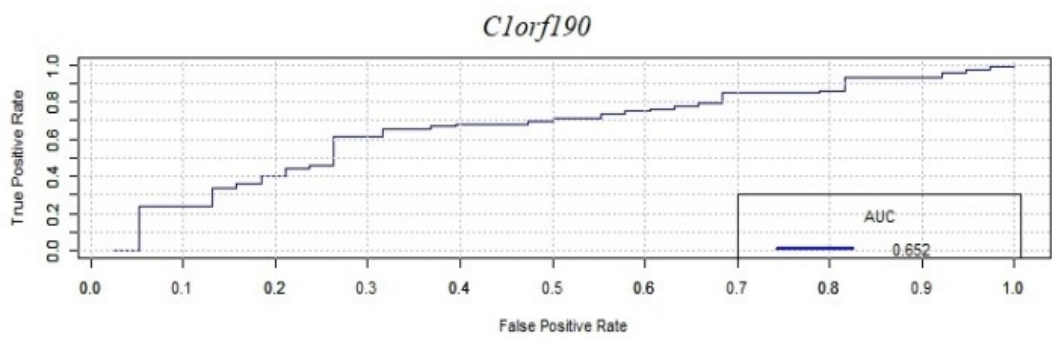




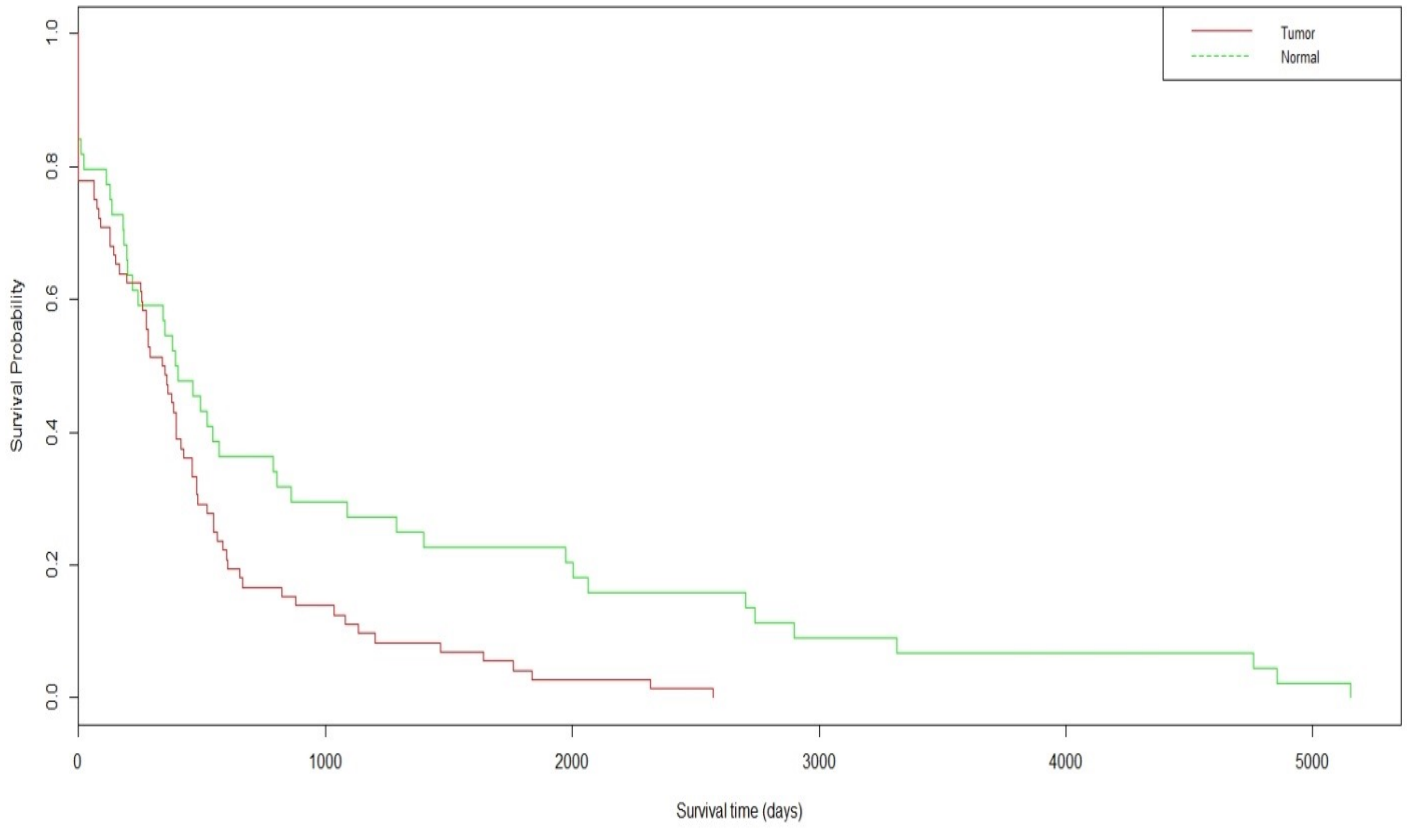








Kaplan-Meier survival (KM) curve comparing survival probability of patients with high six gene expression index in tumor and tumor free patients (P-value < 0.001).



Sequence Analysis: Microarray

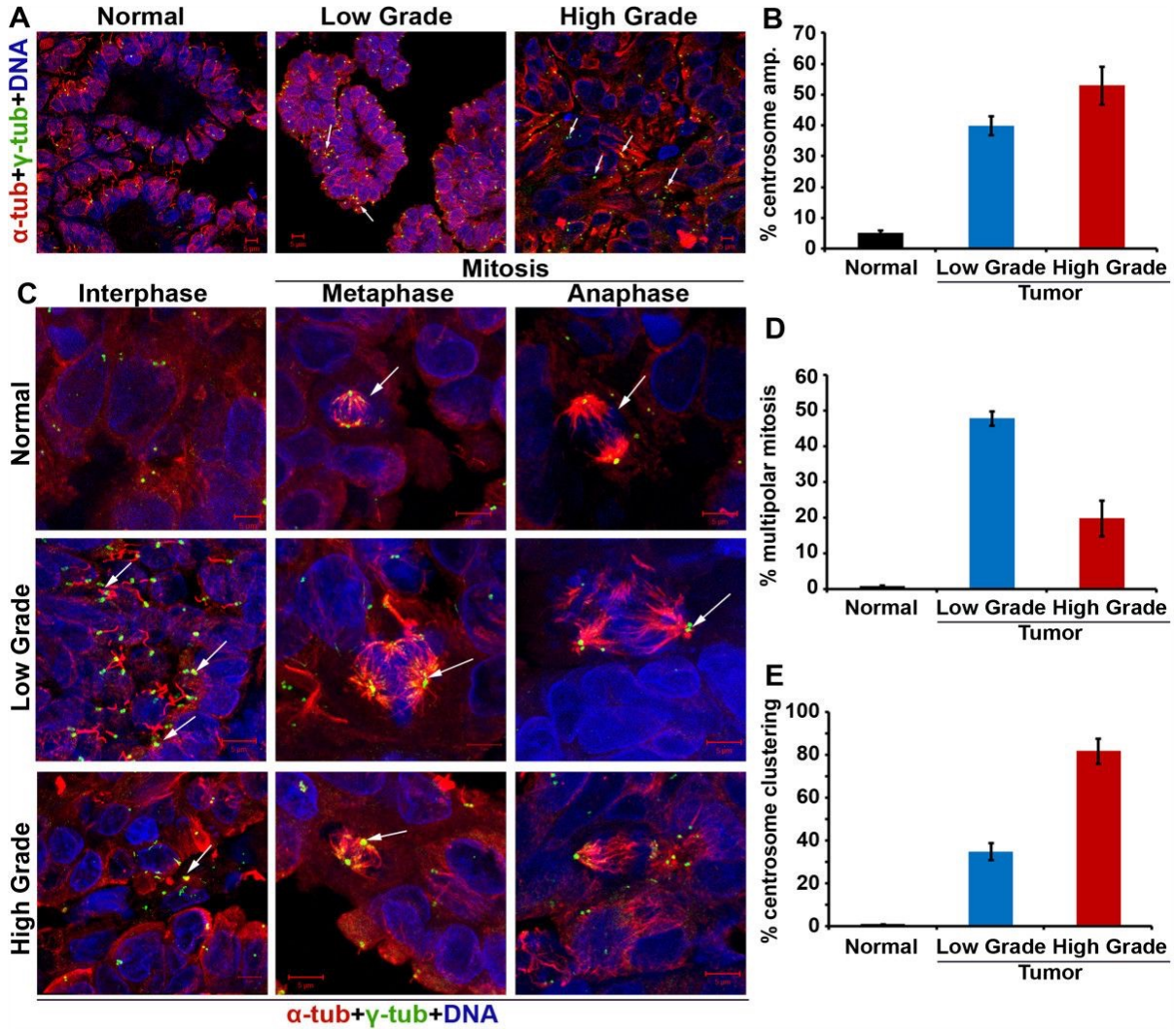
- KIFCI, a novel putative prognostic biomarker for ovarian adenocarcinomas.

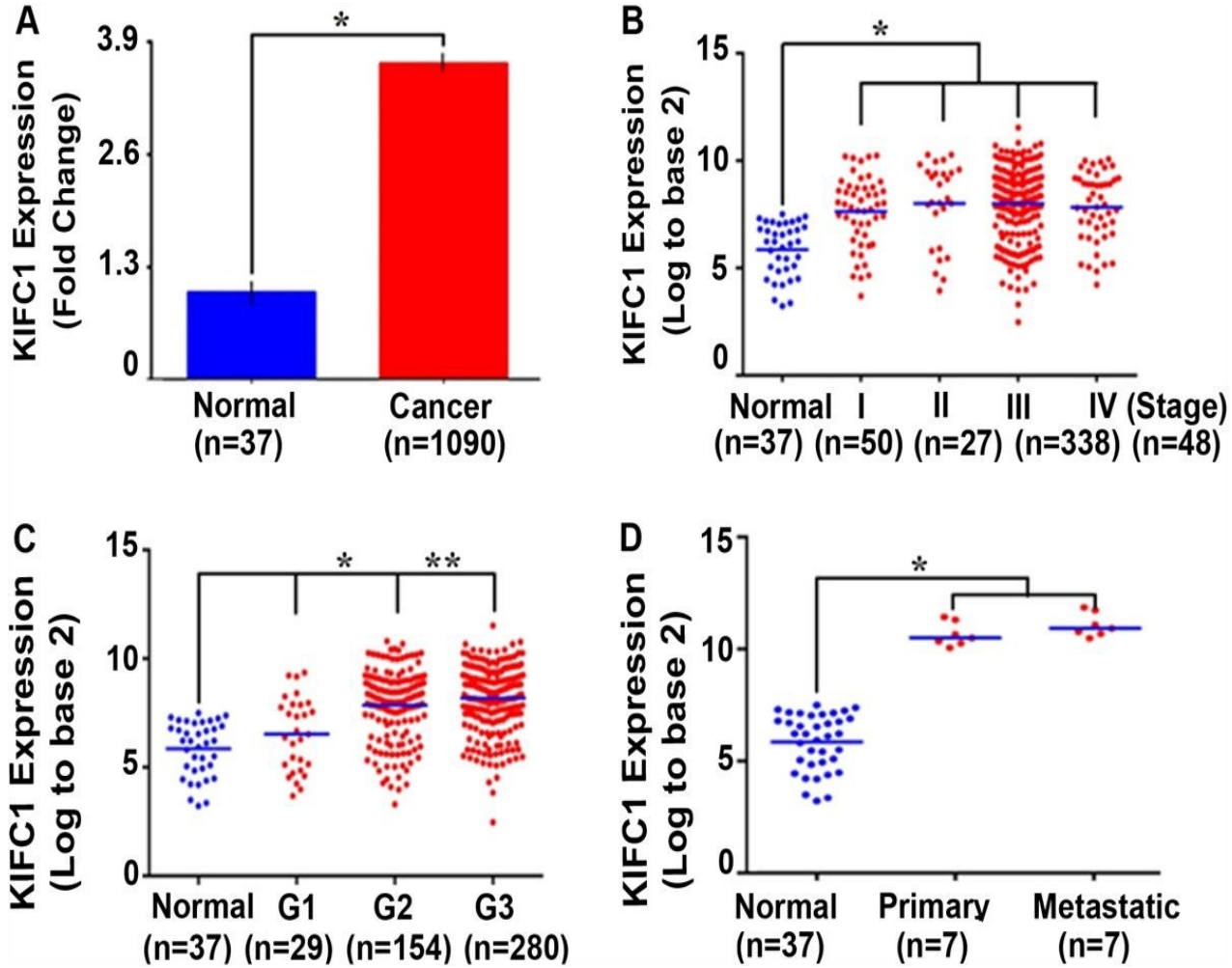
3. KIFCI, a novel putative prognostic biomarker for ovarian adenocarcinomas: delineating protein interaction networks and signaling circuitries, ***Shrikant Pawar***, Shashikiran Donthamsetty, Vaishali Pannu, Padmashree Rida, Angela Ogden, Nathan Bowen, Remus Osan, Guilherme Cantuaria, and Ritu Aneja, ***BMC: Journal of Ovarian Research***

4. A centrosome clustering protein, KIFC1, predicts aggressive disease course in serous ovarian adenocarcinomas, Karuna Mittal, Da Hoon Choi, Sergey Klimov, ***Shrikant Pawar***, Ramneet Kaur, Anirban K. Mitra, Meenakshi V. Gupta, Ralph Sams, Guilherme Cantuaria, Padmashree C. G. Rida, Ritu Aneja, ***BMC: Journal of Ovarian Research***

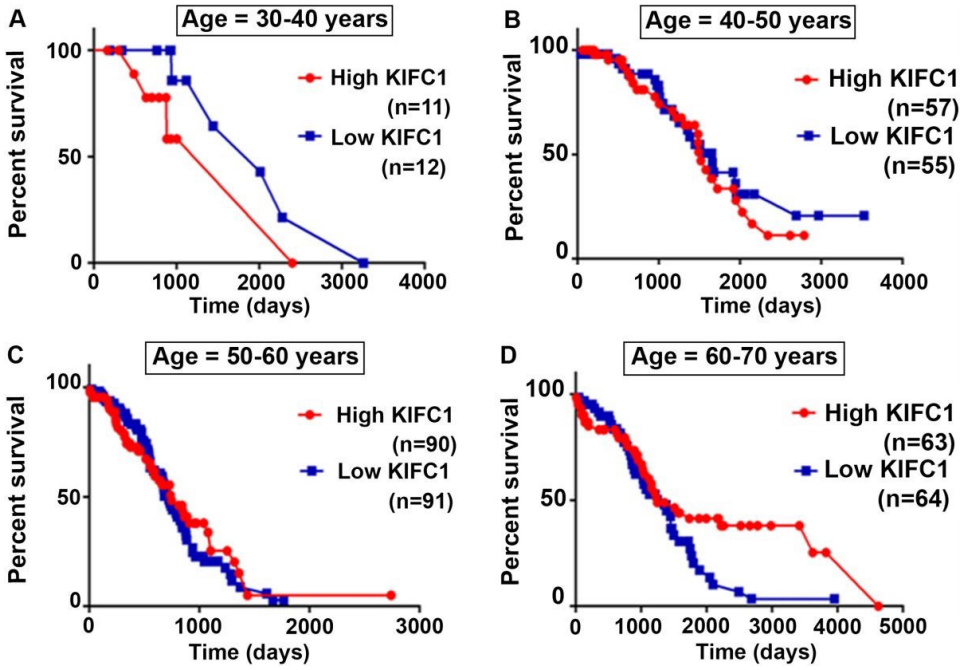
Centrosome amplification in ovarian cancer and high KIFC1 expression in ovarian cancer and normal tissue.

Grady

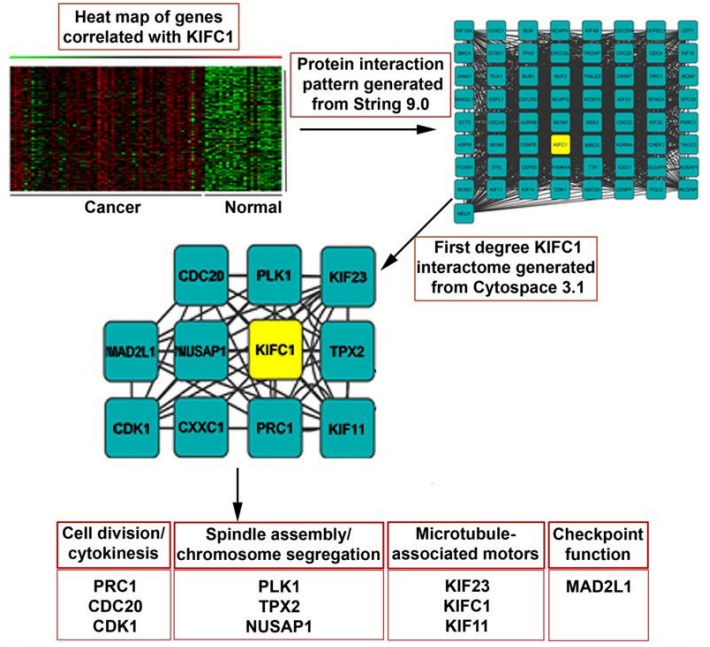




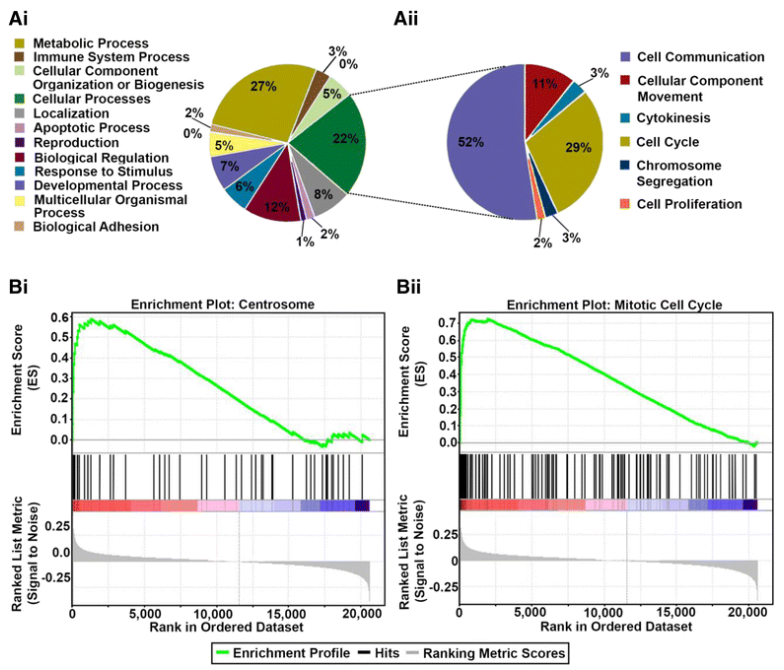
Increased KIFC1 expression is associated with poorer overall survival in age-specific ovarian cancer patients and pathways associated with first degree neighbors of KIFC1 protein



KM Survival Analysis



Protein Interactions



GSEA Analysis

Acknowledgment's and Collaborators

Software Development

Sequence Analysis

Structural Biology

Sequence Analysis

Network Biology

Sequence Analysis

Network Biology

HPC Resources