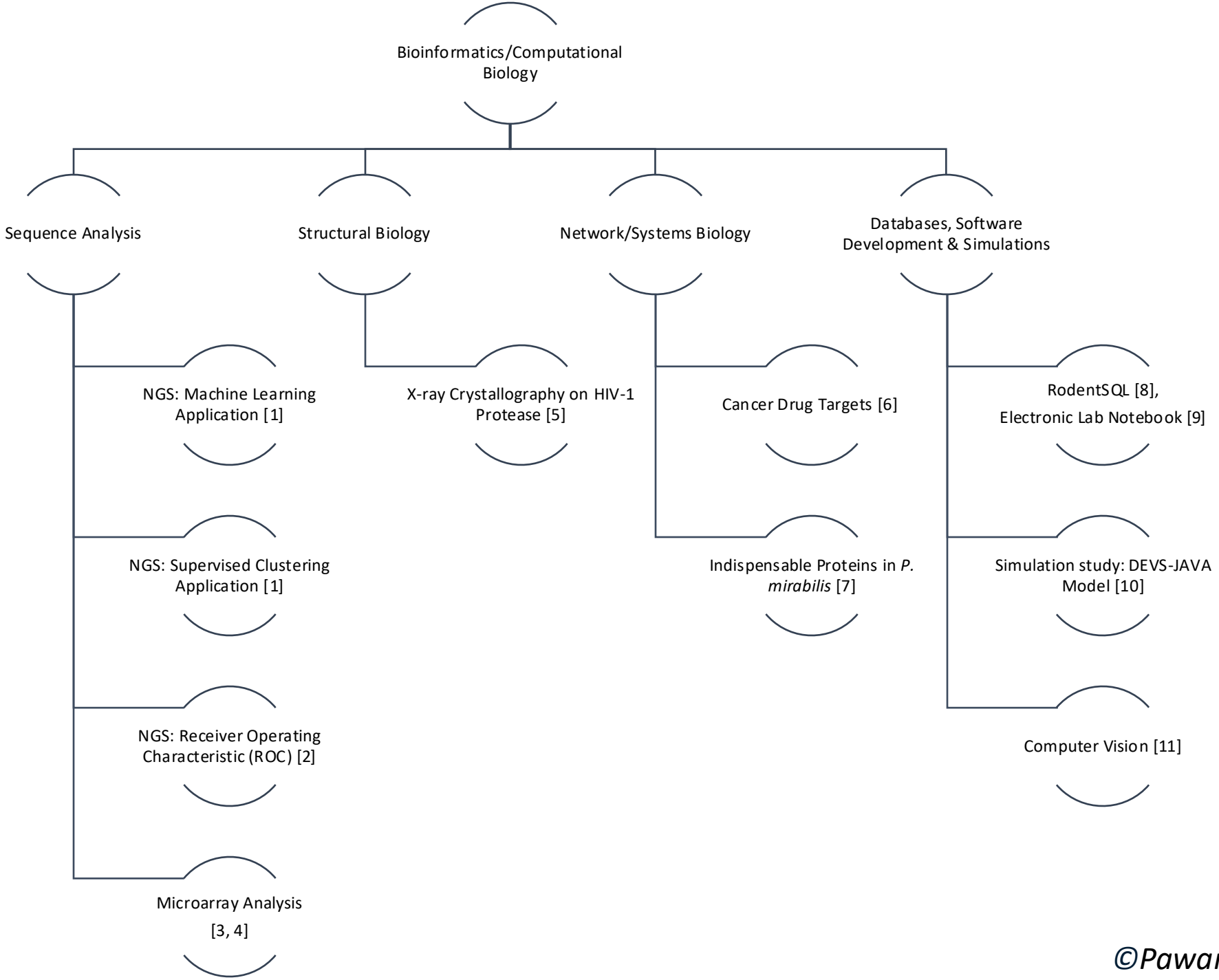
Machine learning techniques for analyzing and interpreting genomics and proteomics data

Pawar Shrikant, Ph.D.

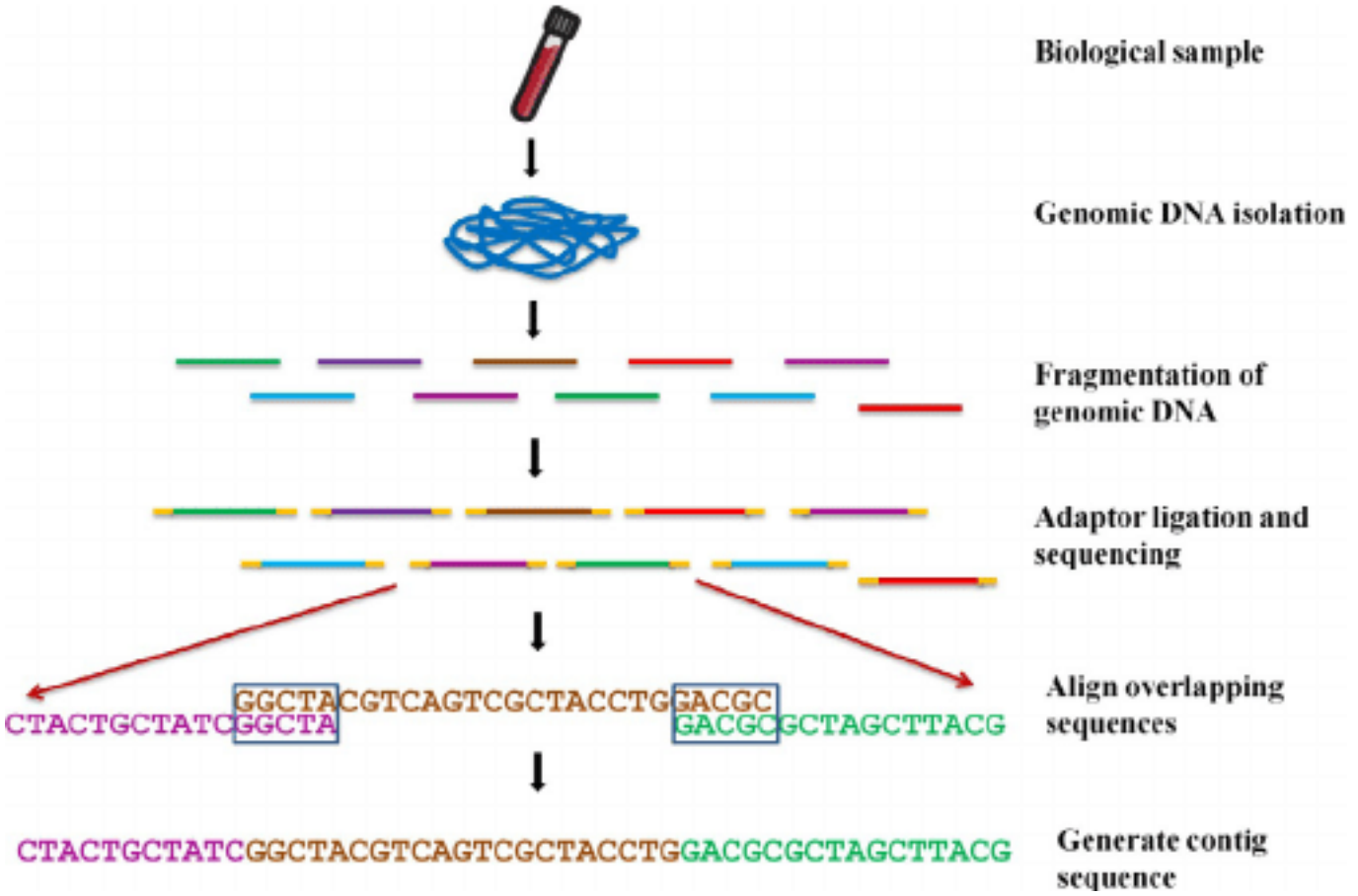


Sequence Analysis: NGS

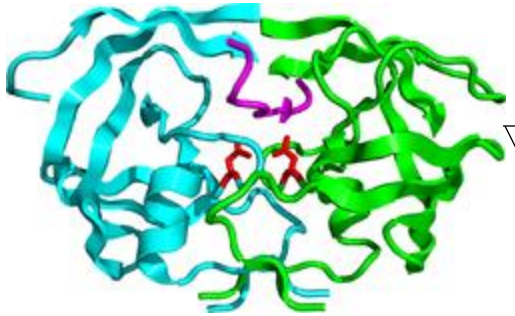
- Utilizing neural networks (Restricted Boltzmann machine's) and clustering algorithms to identify certain important, representative HIV-1 PR sequences from a pool of several hundred sequences.

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***

Next Generation Sequencing Technology (NGS): Align unknown sample contig sequence with known entire human genome to understand known and unknown genes



HIV-1 Protease Action

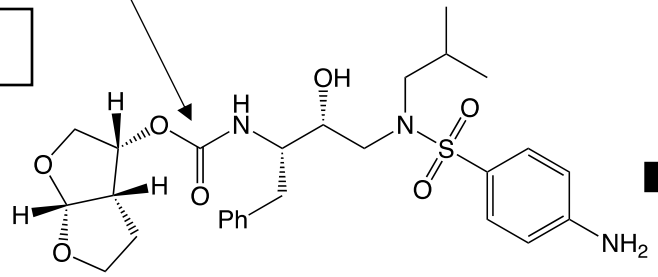


HIV-1 Protease

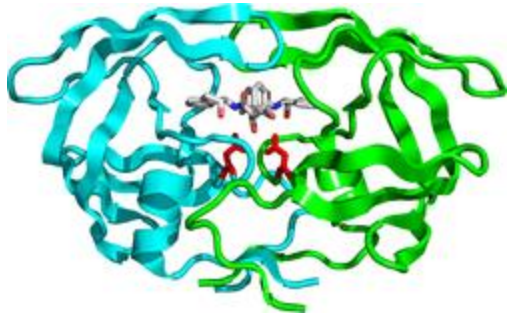
Hydrolyze peptide bonds on the Gag-Pol polyproteins

Mature, fully functional proteins

Infectious viral particle



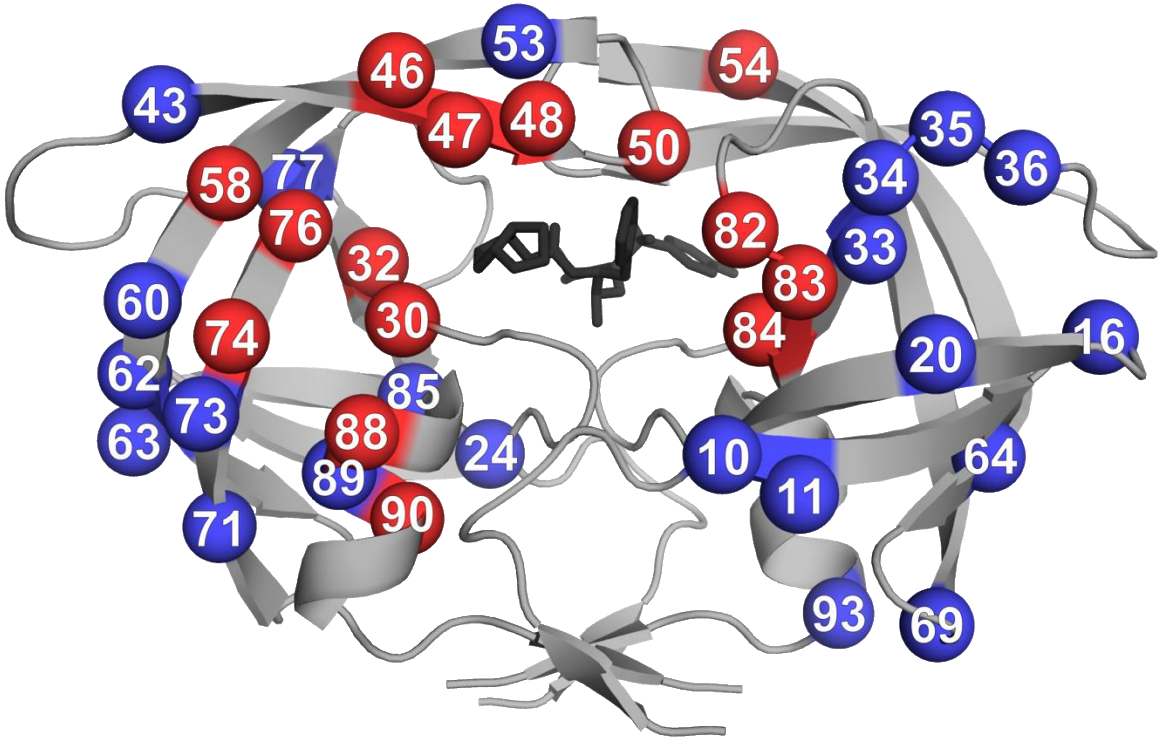
Protease inhibitor: DRV



Competitive inhibition of PR prevents virus maturation

No infectious viral particle

Drug resistance is a severe problem



PQVTLWQRPLVTIKIGGQLKEALLDTGADNTVLEEMSLPGKWKPKMIGGIGGFIKVRQYDQVSIEICGHKAIGTVLIGPTPVNIIGRNLLTQLGCTLNF

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 EALLDTGADN TVLEEMSLPG 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

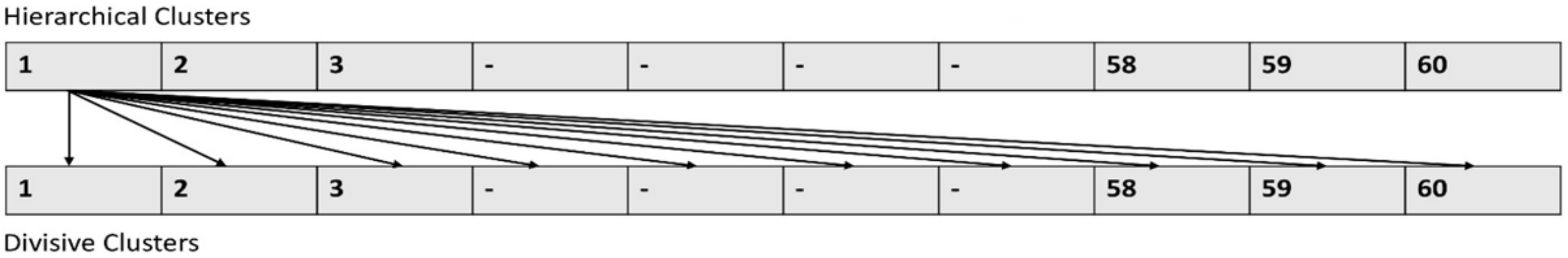
Encode WT protein sequence and structure with Delaunay triangulation

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

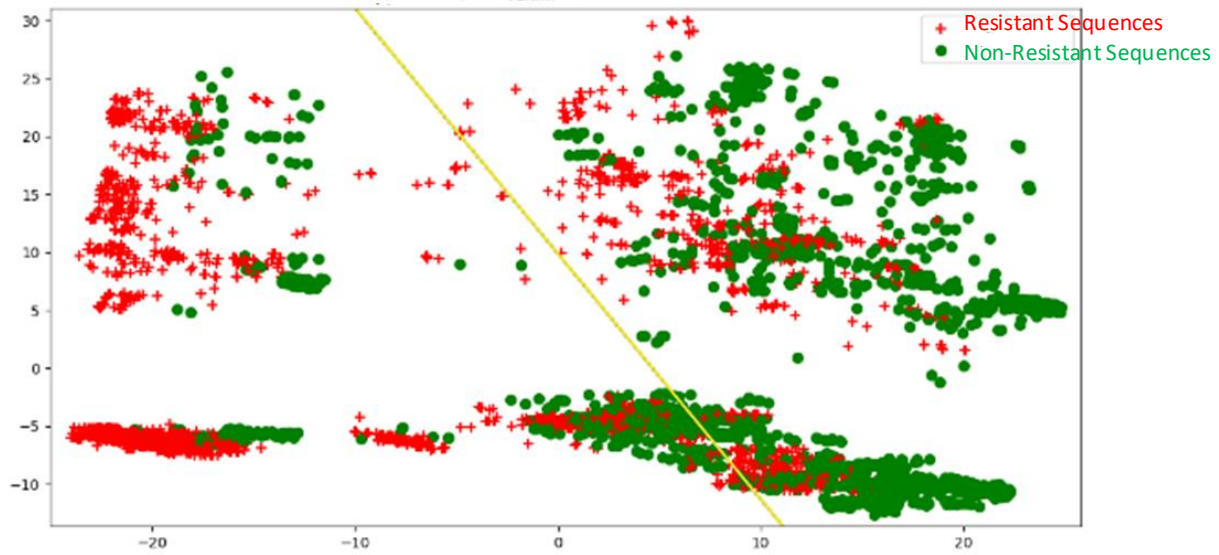
Apply RBM, SVM and RF for checking encoding accuracy

K means, Agglomerative and Divisive Clustering

Select sequences using intersection amongst techniques



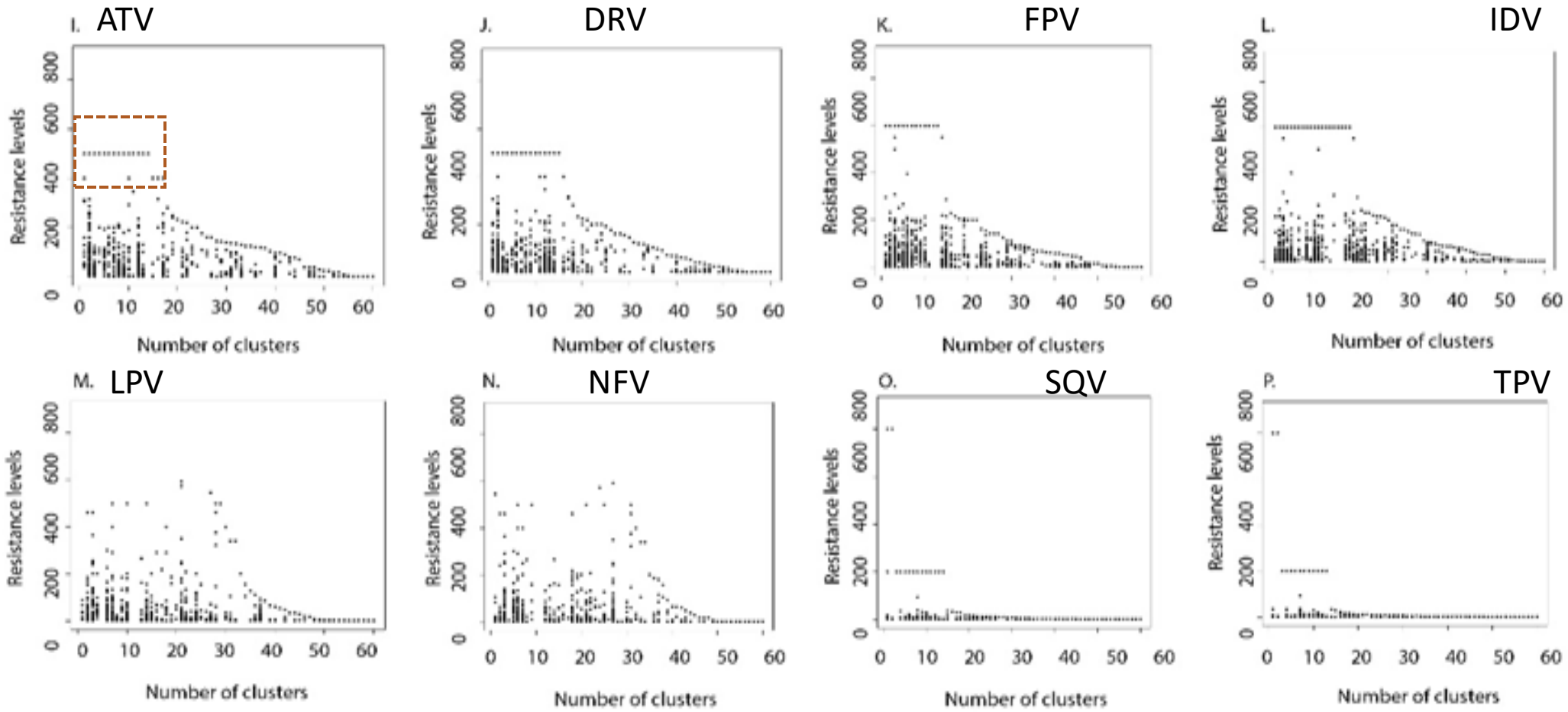
Representative plot with separation of sequences for inhibitor FPV with Support Vector Machine



X and Y axis are weight vectors

- Three, five- and ten-fold accuracies for SVM and RF ranged from 0.95-0.99. RBM also selected sequences in range 0.91-0.97.
- These accuracies are better than *Yu. Et. al. 2014* where she got SVM accuracies ranging from 0.93-0.96.
- These accuracies are comparable with *Shen. Et. al. 2016* where he got RF accuracies ranging from 0.98-0.99.

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

Category	ATV	DRV	FPV	IDV	LPV	NFV	SQV	TPV
H, D and K	0	0	20 (66)	0	35 (61) , 2 (31)	2 (12)	5 (58), 6 (63), 3 (73), 21 (85)	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.

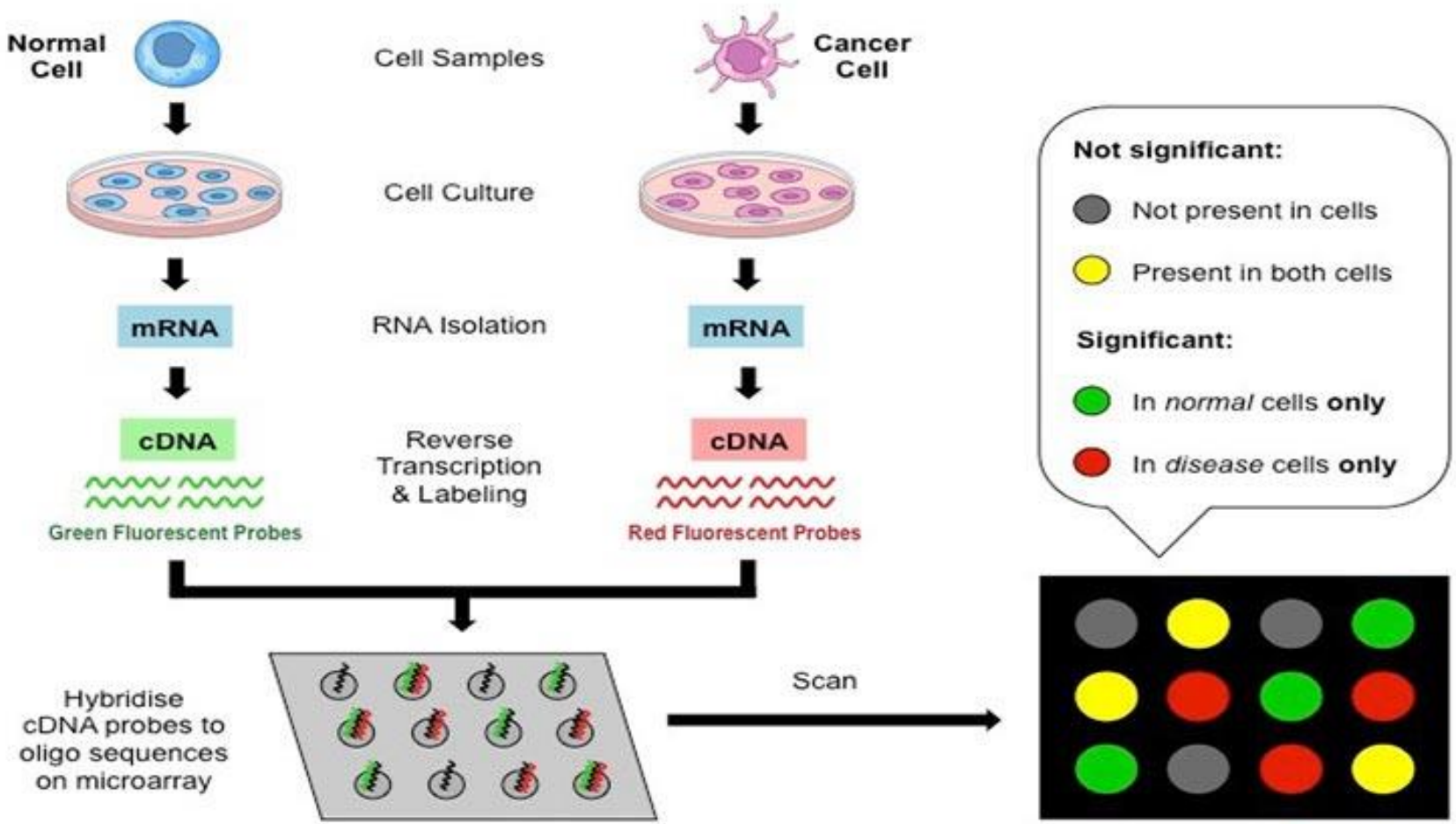
Sequence Analysis: Microarray

- 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***

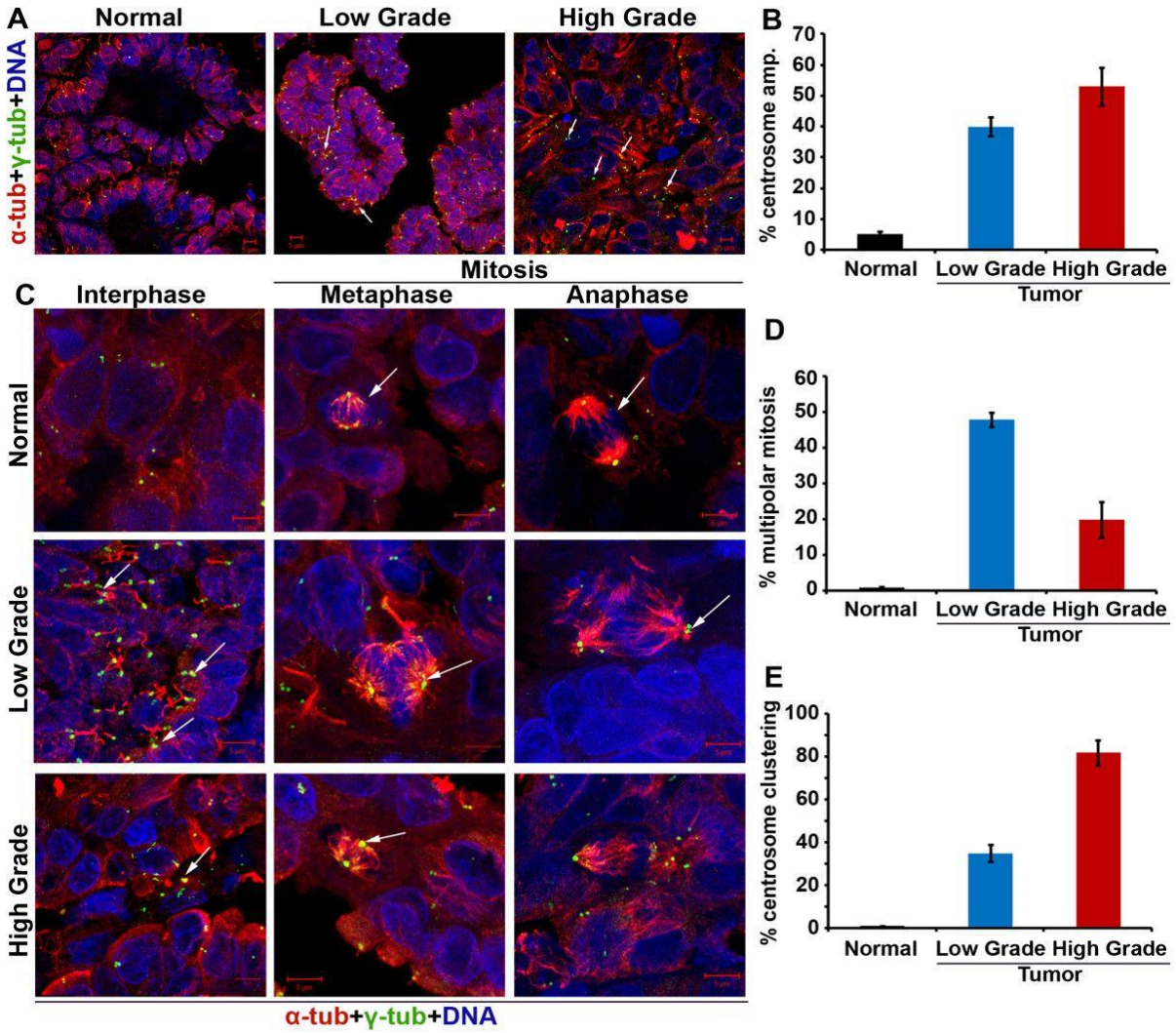
4. A centrosome clustering protein, KIFC1, predicts aggressive disease course in serous ovarian adenocarcinomas, Karuna Mittal, Da Hoon Choi, Sergey Klimov, ***Shrikant Pawar***, Ramneet Kaur, Anirban K. Mitra, Meenakshi V. Gupta, Ralph Sams, Guilherme Cantuaria, Padmashree C. G. Rida, Ritu Aneja, ***BMC: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

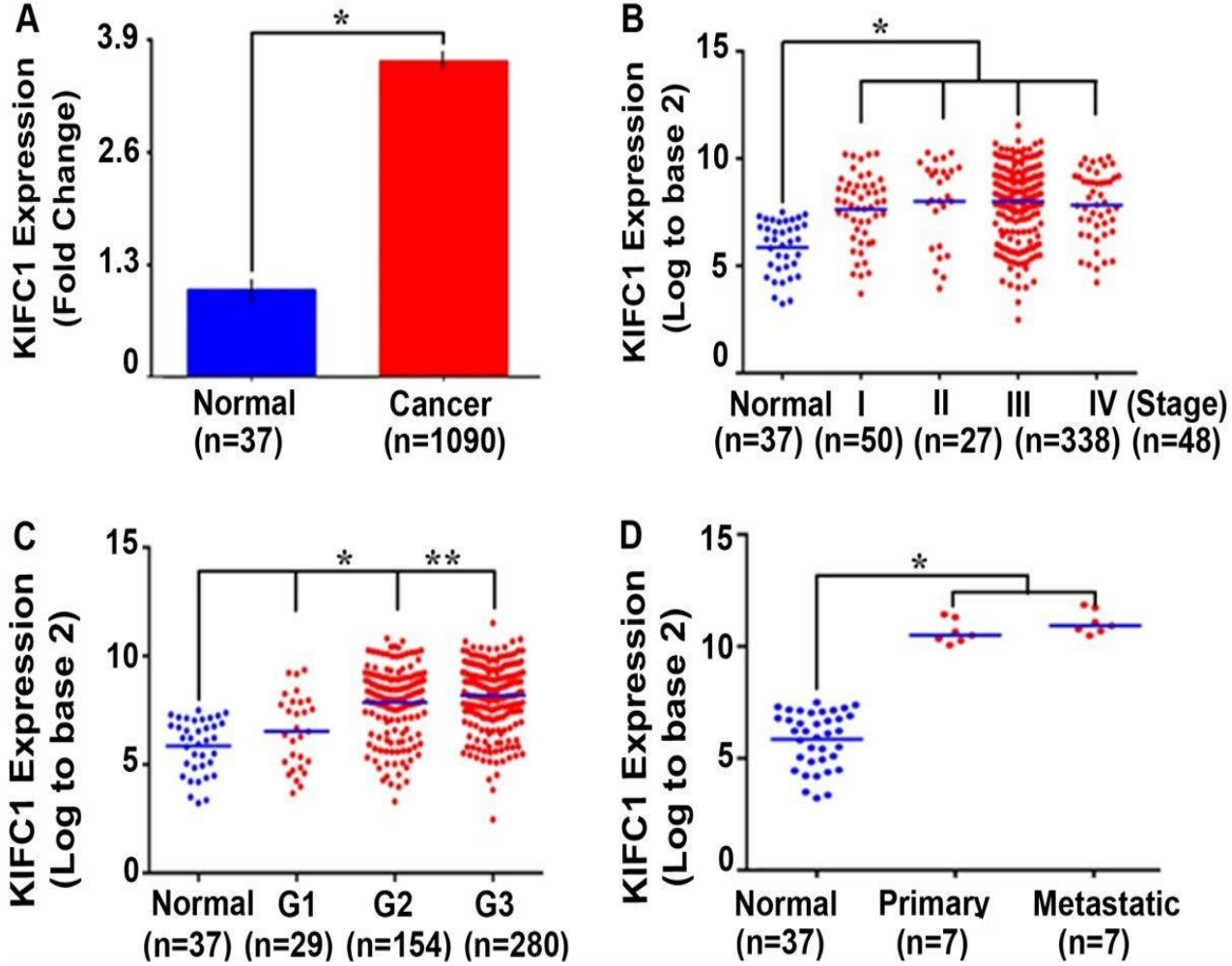
Microarray technology



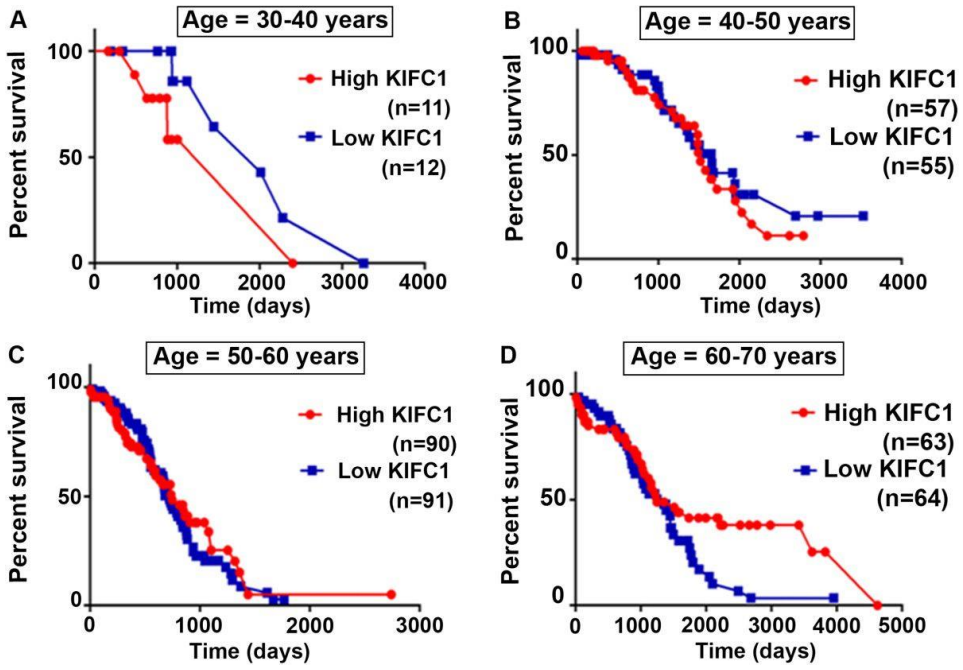
Centrosome amplification in ovarian cancer and high KIFC1 expression in ovarian cancer and normal tissue.

Grady

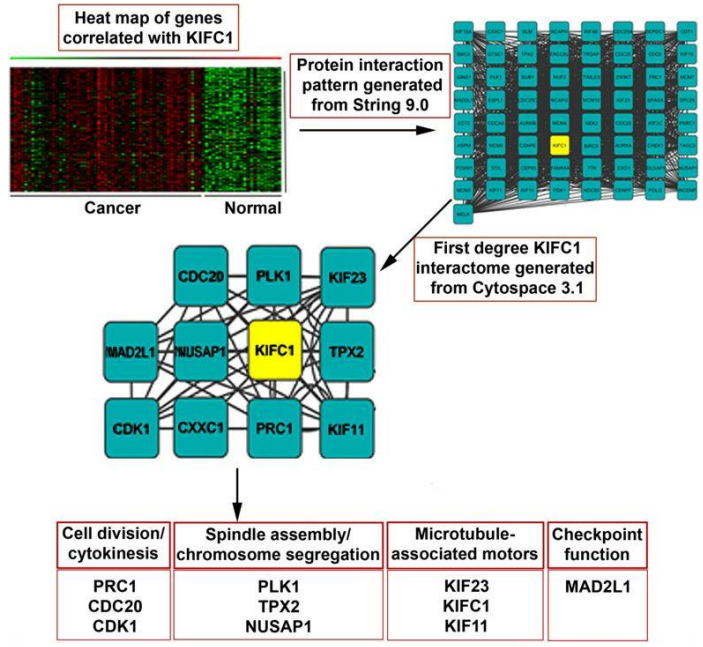




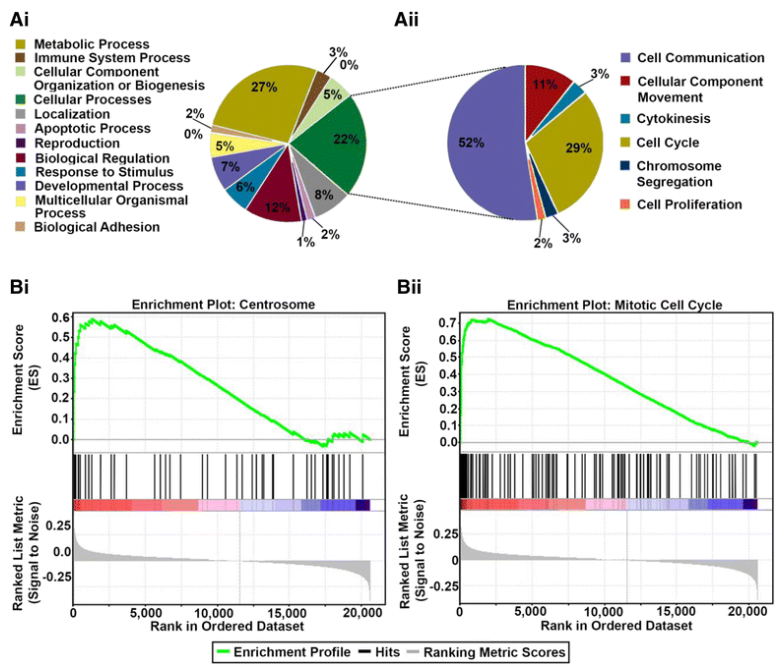
Increased KIFC1 expression is associated with poorer overall survival in age-specific ovarian cancer patients and pathways associated with first degree neighbors of KIFC1 protein



KM Survival Analysis



Protein Interactions



GSEA Analysis

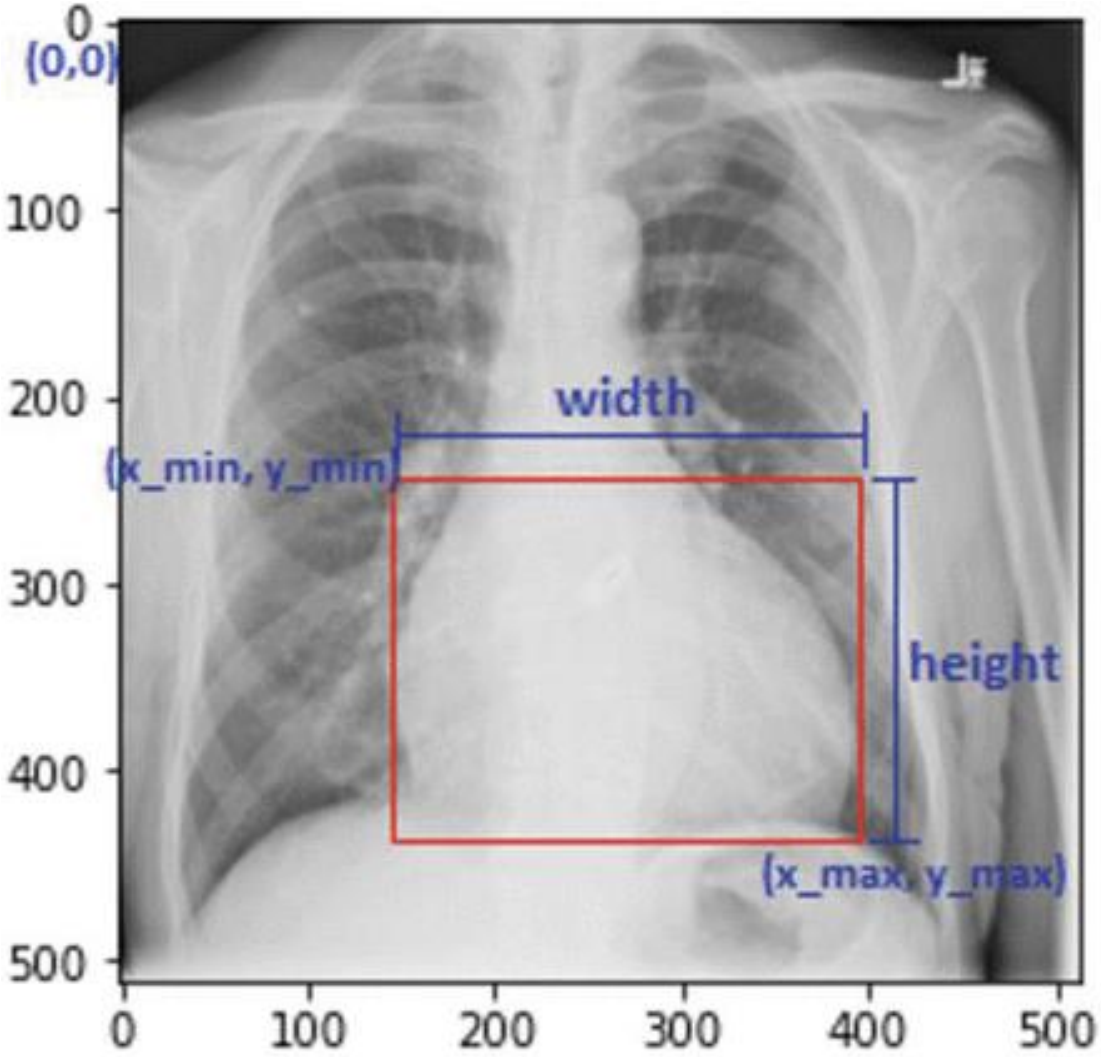
Computer Vision in Biomedical Imaging

Single shot detector application for image disease localization

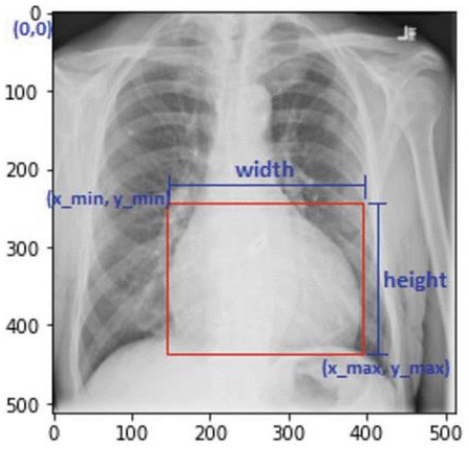
1. **Single shot detector application for image disease localization**, ***Shrikant Pawar***, Rushikesh Chopade, Aditya Stanam, ***bioRxiv***
2. Cyclical Learning Rates (CLR'S) for Improving Training Accuracies and Lowering Computational Cost, ***Shrikant Pawar***, Rushikesh Chopade, Aditya Stanam, ***Springer Lecture Notes in Computer Science***

Bounding boxes for disease localization

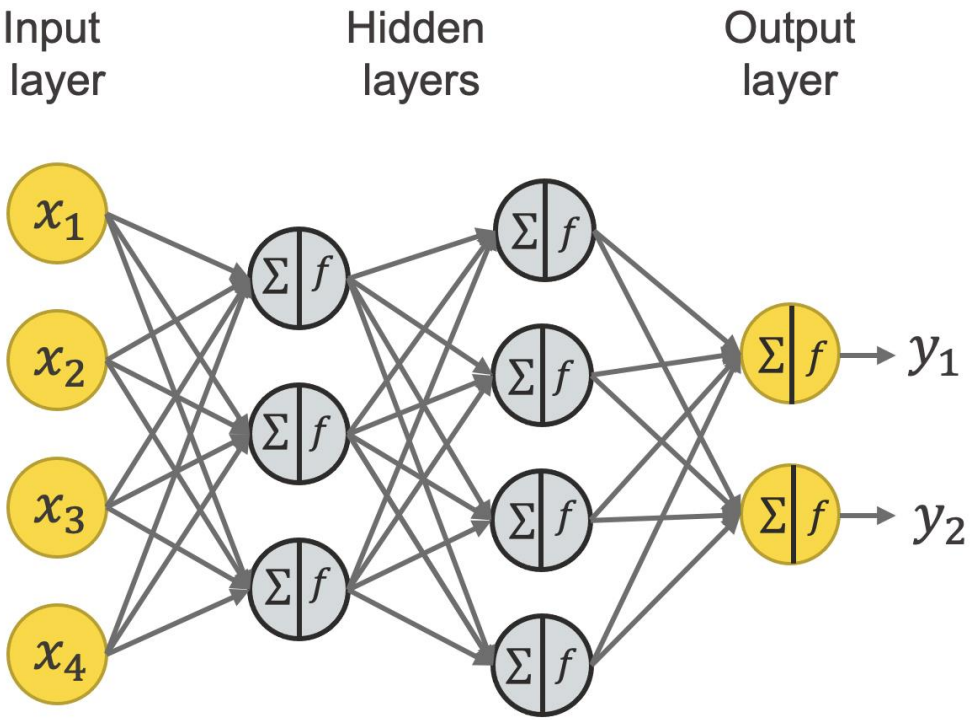
- Object localization is a subfield of computer vision that is used to detect the location of object in an image.
- Bounding box algorithms are useful in localization of image patterns. Recently, utilization of convolutional neural networks on X-ray images has proven a promising disease prediction technique.



Neural Network architecture



Unique pattern of cardiomegaly

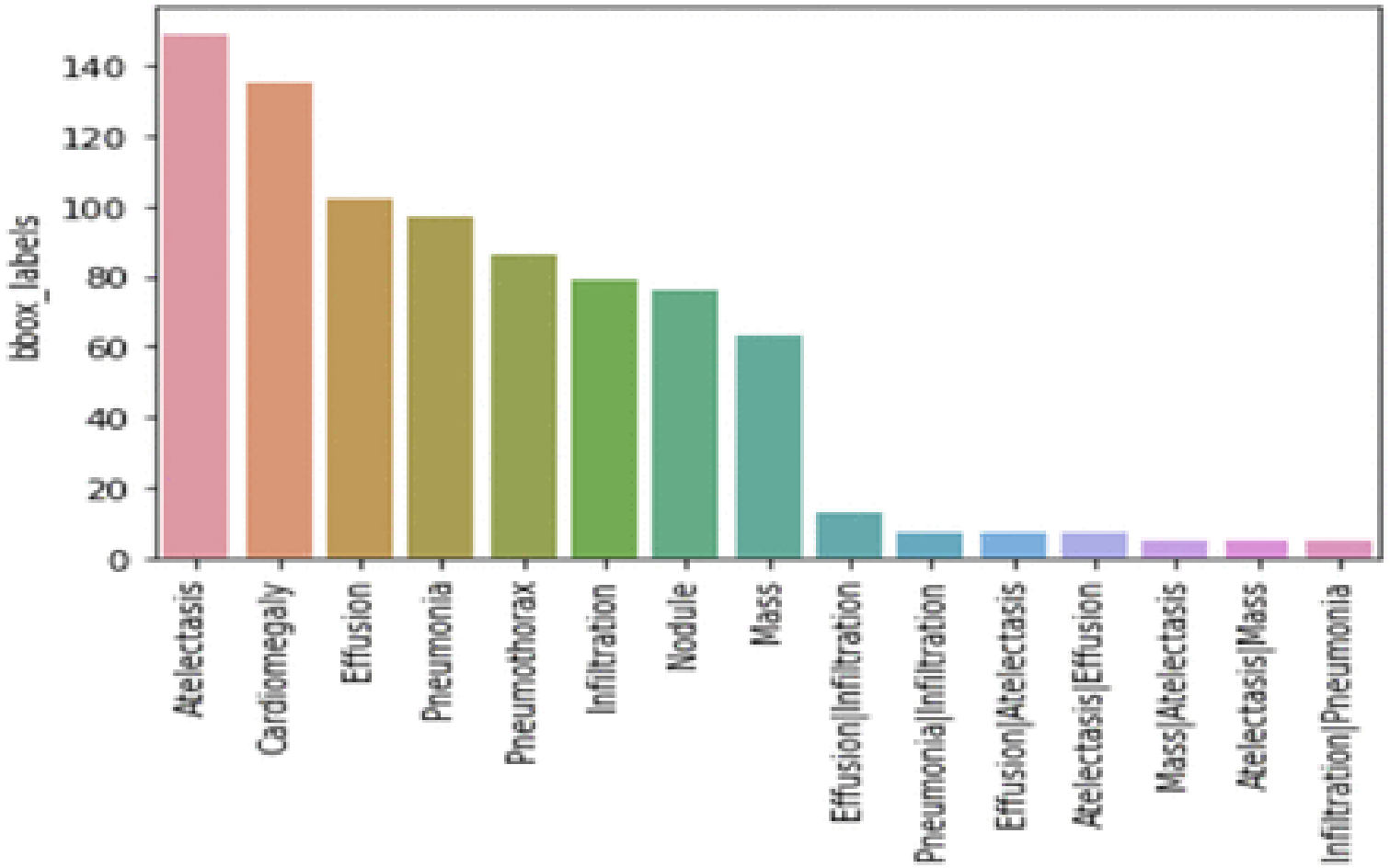


Passed into a network

Pattern captured in Output Layer. Output layer stored as a model

Model is used for predicting Patterns in new image

The dataset consists of 112,120 chest X-ray images, each image with a 1024*1024-pixel resolution. The images are divided into 15 classes ('No Finding', 'Atelectasis', 'Cardiomegaly', 'Consolidation', 'Effusion'; 'Emphysema', 'Edema', 'Fibrosis', 'Infiltration', 'Mass', 'Nodule', 'Pneumonia', 'Pneumothorax', 'Pleural Thickening' and 'Hernia')



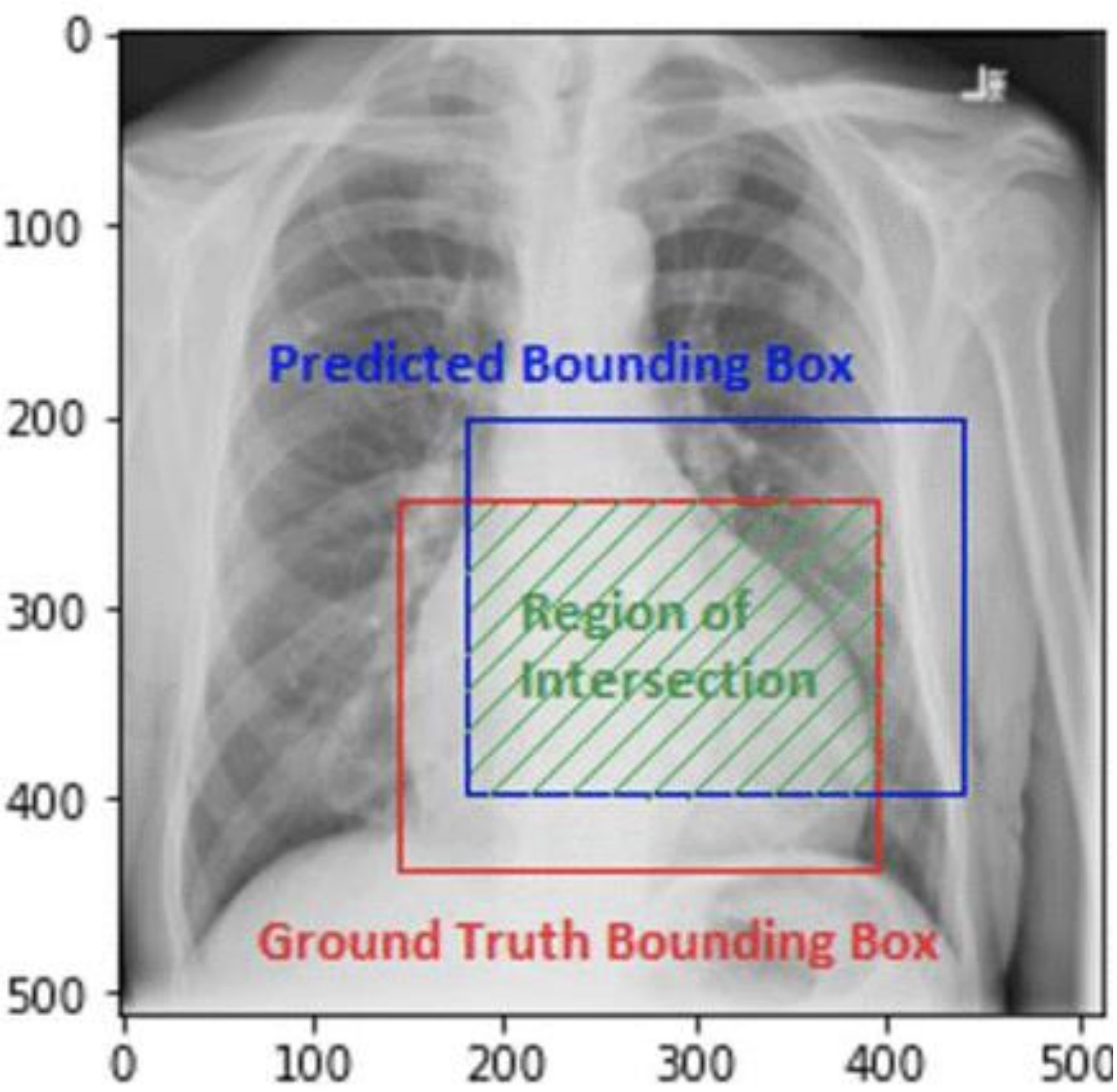


Image of a patient suffering from cardiomegaly showing intersection of original condition and prediction. The prediction accuracy is >95%

Interested in learning and working with above research projects?

- Students interested in working on biomedical imaging projects will receive a $6000/AY & summer stipend. Supported by 2023-2028 NSF SC EPSCoR ADAPT Track 1 Award.
- The research work can be a part of their thesis or project reports.
- Opportunities are also available to collaborate on microarray and sequencing projects but without stipend.

Acknowledgment's and Collaborators

Software Development

Sequence Analysis

Structural Biology

Sequence Analysis

Network Biology

Sequence Analysis

Network Biology

HPC Resources