Using Deep Learning Analytics in Proteomics Analysis of Acute Myeloid Leukemia Mutations

Andy Nguyen, M.D., M.S.

Medical Director, Hematopathology, Hematology and Coagulation Laboratory,
Memorial Hermann Laboratory

Professor of Pathology and Laboratory Medicine,
University of Texas-Houston, Medical School

Grand Round, Sept 12 2016

UTHealth | McGovern Medical School
The University of Texas Health Science Center at Houston

Outline of talk

- Introduction to acute myeloid leukemia (AML), AML without karyotype abnormality (CN-AML).
- Introduction to FLT3-ITD mutation in CN-AML
- Proteomics database (DREAM-9 Challenge) with newly diagnosed AML cases as data source for our project
- Deep Learning method for big-data analytics
- Our implementation of Deep Learning analytics to find correlation between FLT3-ITD status and proteomics (protein levels) in CN-AML cases [an update of previous prelim study]

Financial Disclosures:
No relevant financial relationships with commercial interests to disclose

AML

- A heterogeneous and complicated clonal disorder; characterized by dysregulation of multiple signal transduction pathways and differentiation defects, resulting in increased proliferation and survival of leukemic cells, at the expense of normal hematopoiesis.
- Clinical findings due to anemia, neutropenia, and thrombocytopenia.
- Typically presents with 20% or more blasts in bone marrow

2008 WHO Classification of AML:
AML with recurrent genetic abnormalities (a/w prognosis)

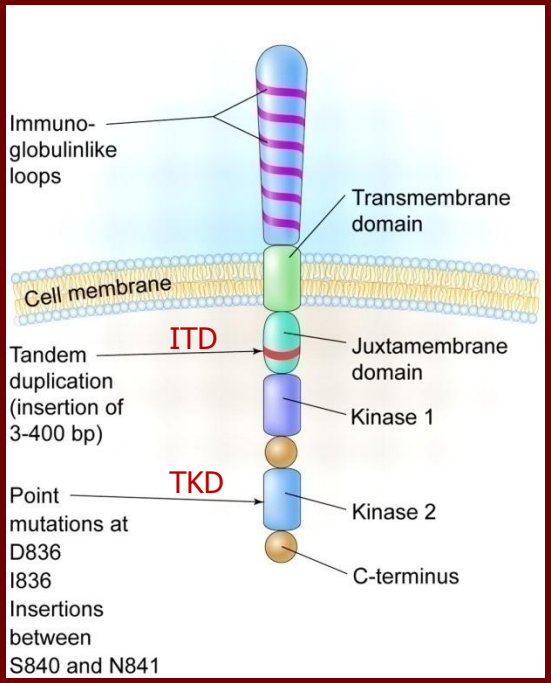
- AML with t(8;21)(q22;q22), RUNX1-RUNX1T1
- AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22), CBFβ-MYH11 **
- APL with t(15;17)(q22;q12), PML-RARA **
- AML with t(9;11)(p22;q23); MLLT3-MLL
- AML with t(6;9)(p23;q34); DEK-NUP214
- AML with inv(3)(q21q26.2) or t(3;3)(q21;q26.2): RPN1-EVI1
- AML (megakaryoblastic) with t(1;22)(p13;q13); RBM15-MKL1

** with specific morphology

- Approximately half of AML patients have no karyotype abnormality (CN-AML).
- Recently it has been demonstrated that the following mutations of are preferentially found in CN-AML with significant prognostic association.
  - FLT3 (Fms-like tyrosine kinase)
  - NPM1 (Nucleophosmin 1)
  - CEBPA (CCAAT/enhancer binding protein alpha)

FLT3 (Fms-like tyrosine kinase)

- The FLT3 protein: encoded by a gene located on 3q12, a member of the class III receptor-tyrosine kinase family (KIT, FMS and PDGFR receptors)
- FLT3 plays an important role in normal growth and differentiation of hematopoietic precursor cells.
- Mutant FLT3 is expressed at higher levels, ligand-independent, constitutive autophosphorylation and activation of downstream signaling
- FLT3-ITD (internal tandem duplication mutation): activates signal transduction pathways in the juxta-membranous region
- ITD mutation: approximately 23% of patients with de novo AML
  -Achieve complete remission: similar to FLT3 wild-type patients, but with a higher relapse rate and a poor response to salvage therapy
  -Inhibitors of FLT3 have so far not been effective



Triaging AML Patients for Treatment

- Approximately 75% of younger adults with AML and about 50% of patients older than 60 achieve a CR after treatment. However survival rate at 3 years is only 25% (with relapse and chemo-resistance)
- Good Cytogenetics and/or mutations: start induction chemo; if CR-> consolidation chemo and f/u; if relapse-> new chemo +/- SCT
- Poor Cytogenetics and/or mutations: start induction chemo; 1st CR-> SCT
- If Refractory (Resistant)-> clinical trial (new chemo +/- SCT)

- Issue to Explore in the Current Project

  -Is the protein profile in an AML case correlated to critical mutation (such as FLT3-ITD)?

  -If this is the case, the protein profile would help: determine key protein pathways in FLT3-ITD mutation, to explore pathogenesis involving the mutation, to monitor chemotherapy response, and to design personalized treatment.

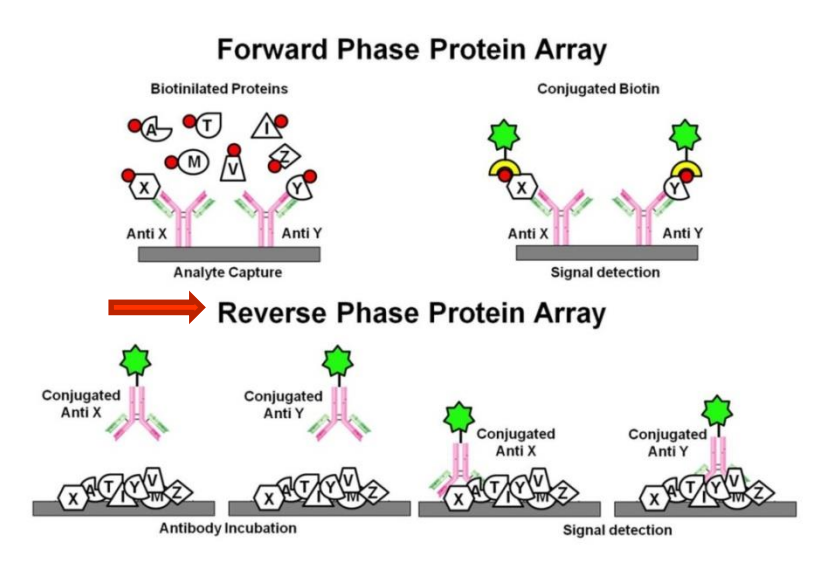
Data Source: DREAM-9 Challenge for AML

- The DREAM Challenges are crowd-sourcing challenges examining questions in biology and medicine.
- A non-profit, collaborative community effort consisting of contributors from across the research spectrum (universities, technology companies); organized by Sage Bionetworks; sponsored by companies and universities.
- Data from DREAM-9 Challenge: a comprehensive proteomic data base with 191 AML cases at MDACC (newly diagnosed, not previously treated), 231 protein levels (reverse-phase protein assay), clinical data, cytogenetics, and mutations.
- Select only CN-AML with isolated FLT3-ITD to avoid confounding factors -> 62 cases

DREAM 9 Challenge Organizer

-Hosted by Rice University

-Data were provided by Dr. S. Kornblau from The University of Texas MD Anderson Cancer Center and were obtained through Synapse syn2455683 as part of the AML DREAM Challenge

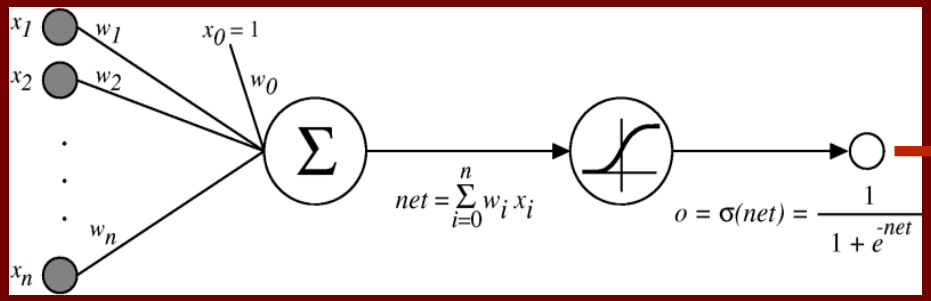


Big-Data Analytics and Deep Learning

- Big companies are analyzing large volumes of data for business analysis and decisions, using Deep Learning technology (Google's search engine, Google Photo, automobile companies: self-driving cars, IBM's Watson)

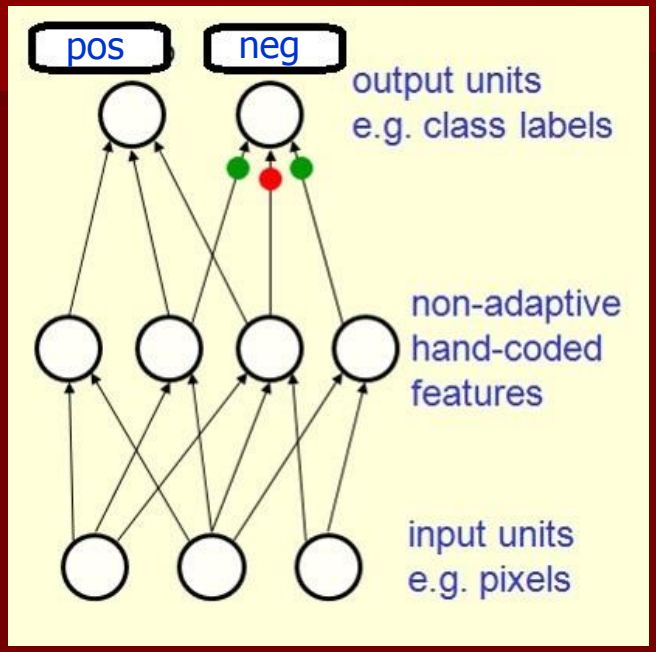
- Big data analytics in cancer proteomics and genomics can significantly be benefited from Deep Learning ***("We are drowning in information and starving for knowledge", Rutherford D. Roger)***

- Deep Learning is based on artificial neural networks (inspired by biological neural networks): artificial nodes ("neurons") are connected together to form a network for prediction/classification tasks

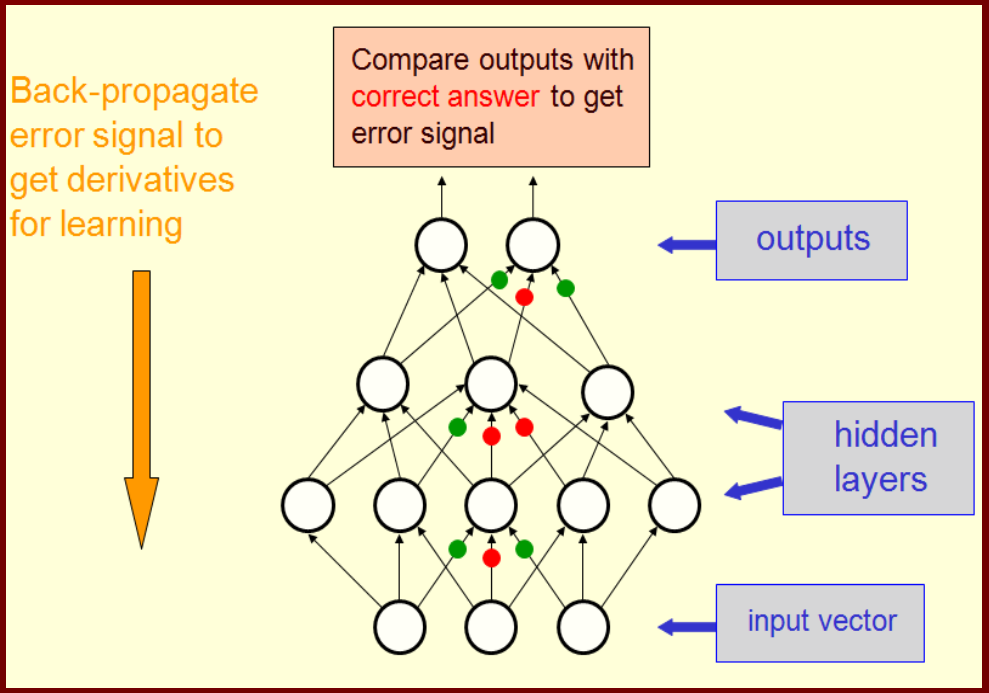


- In traditional programming, an engineer writes explicit, step-by-step instructions for computers to follow. In neural network, they do not encode software with instructions; instead they train the software

Early Generations of Neural Networks with Supervised Training (model is trained with known outcomes)



1st gen (1960's)



2nd gen (1980's)

- Early neural networks were based on supervised training often too difficult to train and they were found to be less effective than other methods.

Deep Learning (3rd Gen Neural Network)

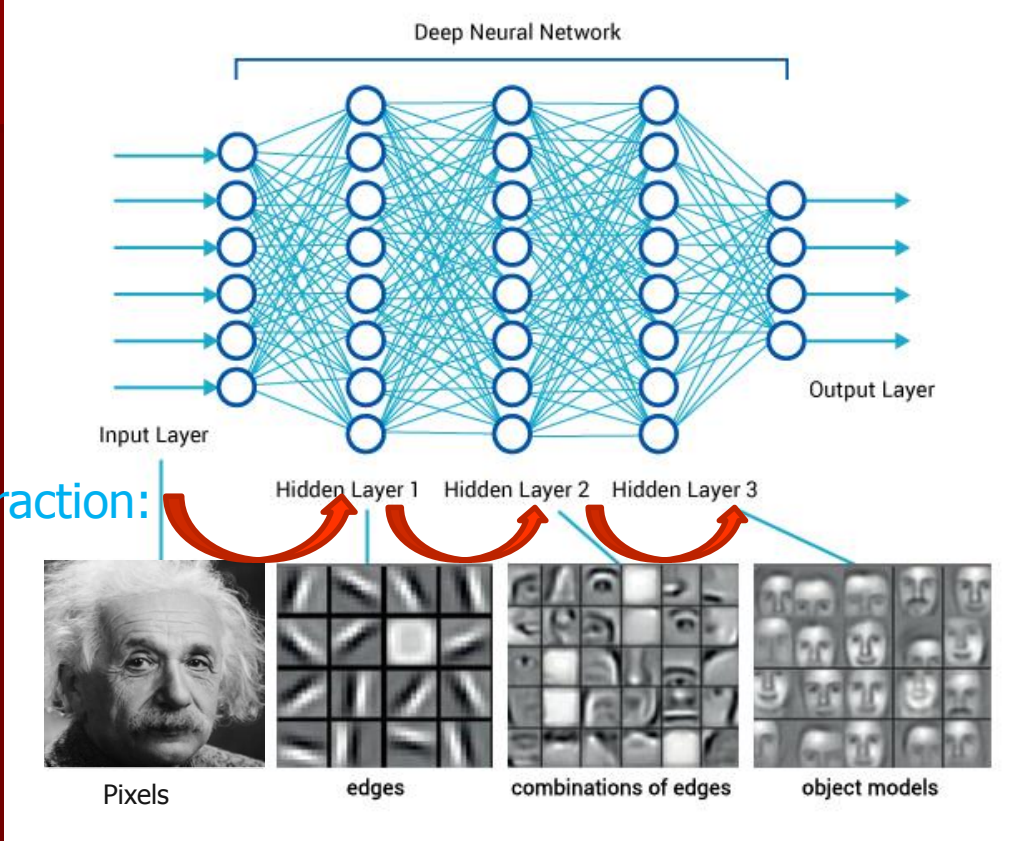
- A major breakthrough in 2006: Hinton (U of Toronto) won a contest held by Merck to identify molecules that could lead to new drugs. The group used deep learning to zero in on the molecules most likely to bind to their targets.

- Deep Learning algorithms:
  (1) Unsupervised learning ->allows a network to be fed with raw data (no known outcomes) and to automatically discover the representations needed for detection or classification

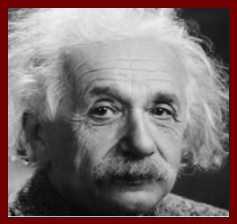
  (2) Extract high-level & complex data representations through multiple layers.
  -> allows for less interferences by background noise

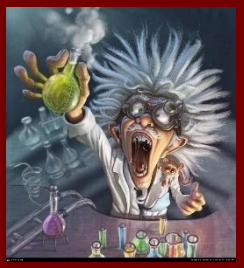
- Supporting hardware: multiple graphics processing units (GPU)

A Deep Learning Neural Network to Detect Image: Extracting higher-level Features With Unsupervised Learning



Feature extraction:





-Each hidden layer applies a nonlinear transformation on its input to transform the input to higher level of representation in its output.
-Multiple levels of abstraction of the image: from pixels to complex shapes and objects defining a human face

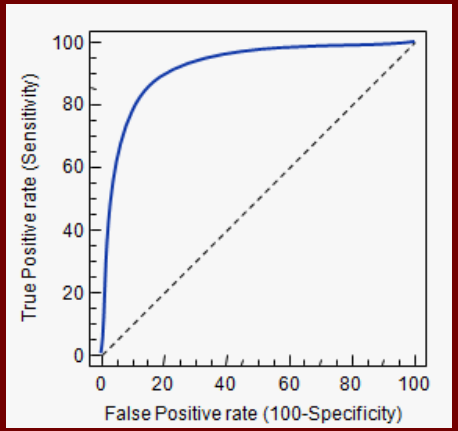
Deep Learning and Breast Cancer Detection, the Camelyon Grand Challenge 2016

- The International Symposium on Biomedical Imaging (ISBI) held a Grand Challenge to evaluate computational systems for the automated detection of metastatic breast cancer in whole slide images of sentinel lymph node biopsies.

- The Harvard & MIT team won the grand challenge:
  obtaining an area under the receiver operating curve (AUC) of 0.925 for the task of whole slide image classification (pos vs. neg)

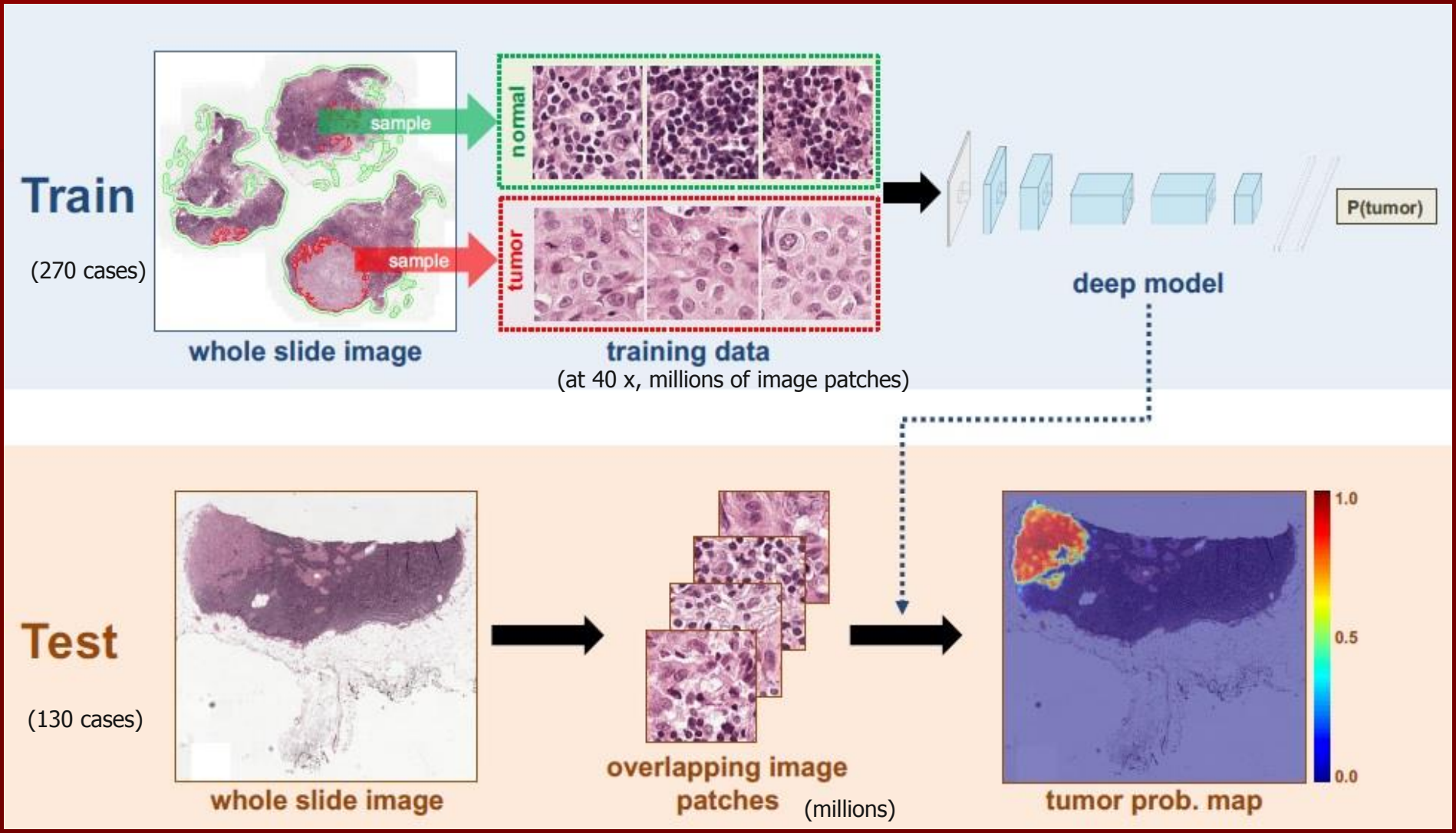
- A pathologist independently reviewed the same images, obtaining a whole slide image classification AUC of 0.966

- Combining this deep learning system's predictions with the human pathologist's diagnoses increased the pathologist's AUC to 0.995, representing an approximately 85 percent reduction in human error rate.



Receiver operating characteristic (ROC) curve

Deep Learning and Breast Cancer Detection (cont'd)



arXiv:1606.05718v1 [q-bio.QM] 18 Jun 2016

nature COMMUNICATIONS

ARTICLE

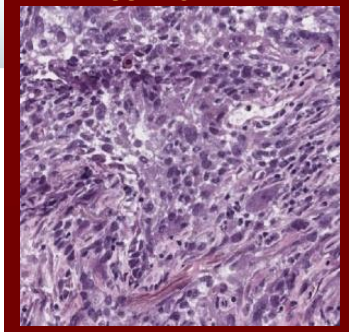
Received 24 Jan 2016 | Accepted 6 Jul 2016 | Published 16 Aug 2016

DOI: 10.1038/ncomms12474 OPEN

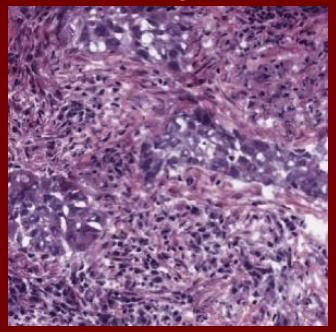
Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features

Kun-Hsing Yu[1,2], Ce Zhang[3], Gerald J. Berry[4], Russ B. Altman[1], Christopher Ré[3], Daniel L. Rubin[1,*] & Michael Snyder[2,*]

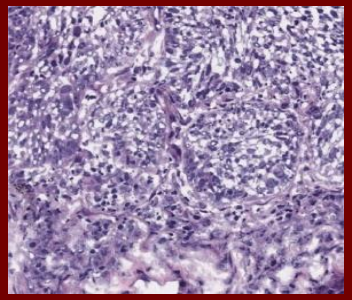
AdenoCA , 1B Gr 3 >99 mo



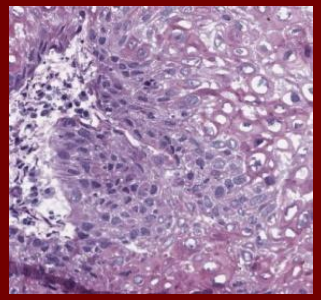
AdenoCA , 1B Gr 3 12 mo



SCC, 1 Gr 1 >70 mo



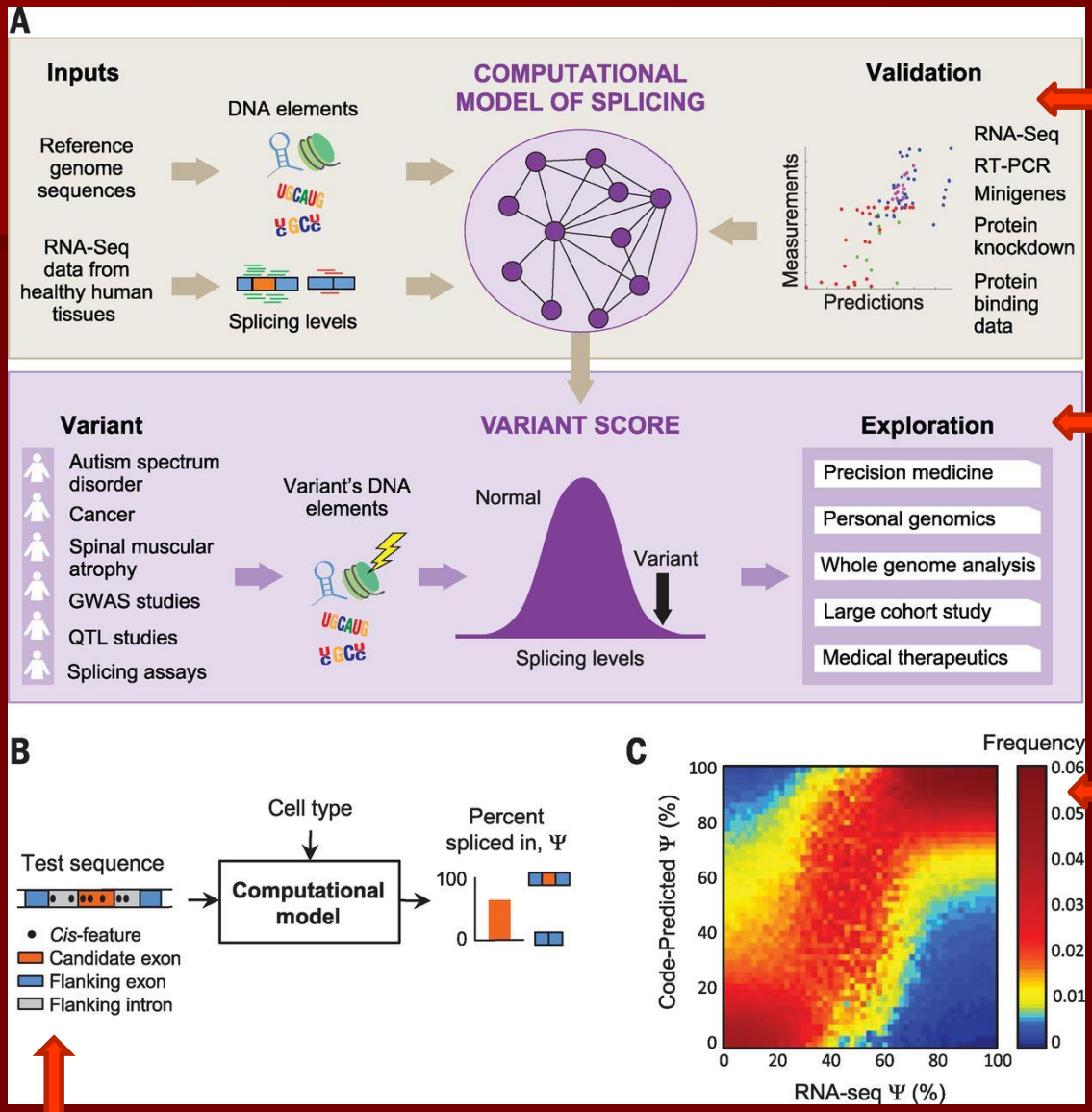
SCC, 1B Gr 3: 12 mo



- Tumour stages/grades are insufficient for predicting survival outcomes (diverse)-> room for improvement
- Study conducted by Department of Pathology, Stanford University
- 2,186 whole-slide images (H&E) of lung adenocarcinoma and squamous cell carcinoma patients from The Cancer Genome Atlas (TCGA)
- Extract 9,879 quantitative image features-> scale down to 240 key features -> distinguish shorter-term survivors from longer-term survivors with stage I adenocarcinoma ($p<0.003$) or squamous cell carcinoma ($p=0.023$)
- Methods are extensible to histopathology images of other organs.

Detecting pathologic genetic variants using a deep learning model of splicing:

Mutations in MLH1 and MSH2 arising in patients with colorectal cancer **(U. of Toronto)



A. Machine learning to infer a model of splicing, by correlating DNA elements with splicing levels in healthy tissues.

Model to learn variants a/w diseases

C. Predictions are made for 10,689 test exons profiled in 16 tissues -> AUC=94%

**Lynch syndrome, or hereditary nonpolyposis colorectal cancer

B. The model extracts the regulatory code from a test DNA sequence and predicts the percentage of transcripts with the central exon spliced in (Ψ)

[H. Y. Xiong et al. Science 2015;347:1254806]

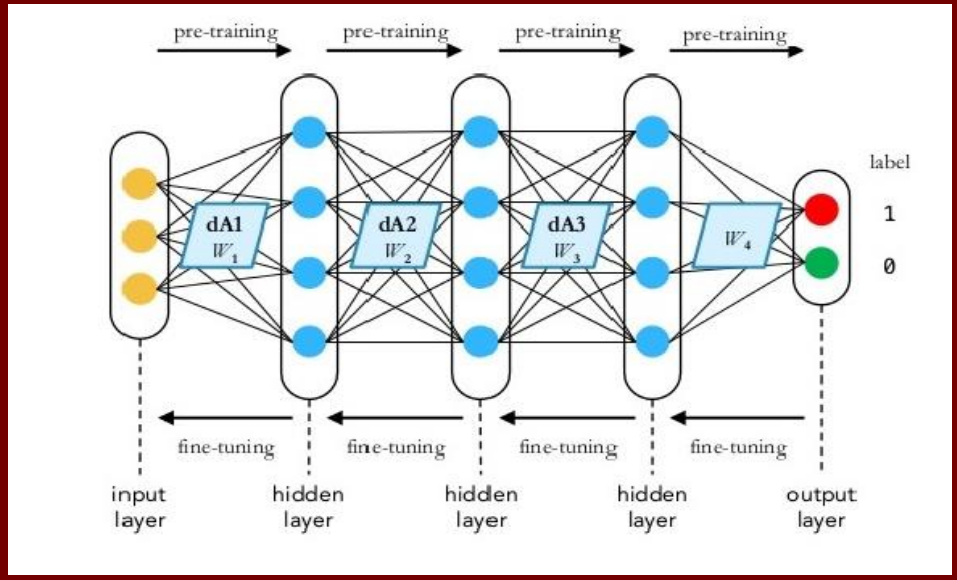
The Goal of our Study

In the present study, we explore how Deep Learning can be utilized for proteomics analysis in AML. Specifically we attempt to determine the correlation between FLT3-ITD mutation status with serum level of 231 proteins in newly-diagnosed AML patients.

Types of Deep Learning Architectures:

- 1 Convolutional neural networks
- 2 Recursive neural networks
- 3 Long short term memory (LSTM)
- 4 Deep belief networks
- 5 Convolutional deep belief networks
- 6 Deep Boltzmann machines
- 7 Stacked auto-encoders <<<<<<<<<<<
- 8 Tensor deep stacking networks
- 9 Spike-and-slab RBMs
- 10 Compound hierarchical-deep models
- 11 Deep coding networks
- 12 Deep q-networks
- 13 Encoder–decoder networks
- 14 Multilayer kernel machine
- etc.

Stacked Autoencoder Network



Our Programming Platform

- We design a Deep Learning neural network with stacked (multi-layered) auto-encoder in R language.
- R is a programming language for statistical computing and graphics supported by the R Foundation for Statistical Computing.
- R was derived from the S language which was developed at Bell Laboratories
- In this study, we use many Deep Learning functions obtained from an R package called “Deepnet” which is available from the Comprehensive R Archive Network, under the GNU General Public License
- We will also use a conventional neural network (EasyNN, Neural Planner Software, Cheshire, England) to compare performance [used in our previous study]

Stacked Autoencoder Algorithm

$$\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}\left[a_j^{(2)}(x^{(i)})\right]$$

$$\hat{\rho}_j = \rho,$$

$$\sum_{j=1}^{s_2} \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}.$$

$$\sum_{j=1}^{s_2} \mathrm{KL}(\rho||\hat{\rho}_j),$$

$$\mathrm{KL}(\rho||\hat{\rho}_j) = \rho \log \tfrac{\rho}{\hat{\rho}_j} + (1-\rho)\log\tfrac{1-\rho}{1-\hat{\rho}_j}$$

$$J_{\mathrm{sparse}}(W,b) = J(W,b) + \beta \sum_{j=1}^{s_2} \mathrm{KL}(\rho||\hat{\rho}_j),$$

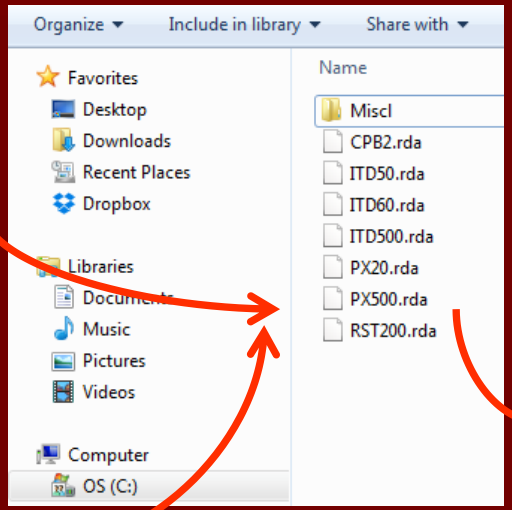
$$\delta_i^{(2)} = \left(\sum_{j=1}^{s_2} W_{ji}^{(2)}\delta_j^{(3)}\right) f'(z_i^{(2)}),$$

$$\delta_i^{(2)} = \left(\left(\sum_{j=1}^{s_2} W_{ji}^{(2)}\delta_j^{(3)}\right) + \beta\left(-\frac{\rho}{\hat{\rho}_i} + \frac{1-\rho}{1-\hat{\rho}_i}\right)\right) f'(z_i^{(2)}).$$

Deep Learning Analysis Workflow

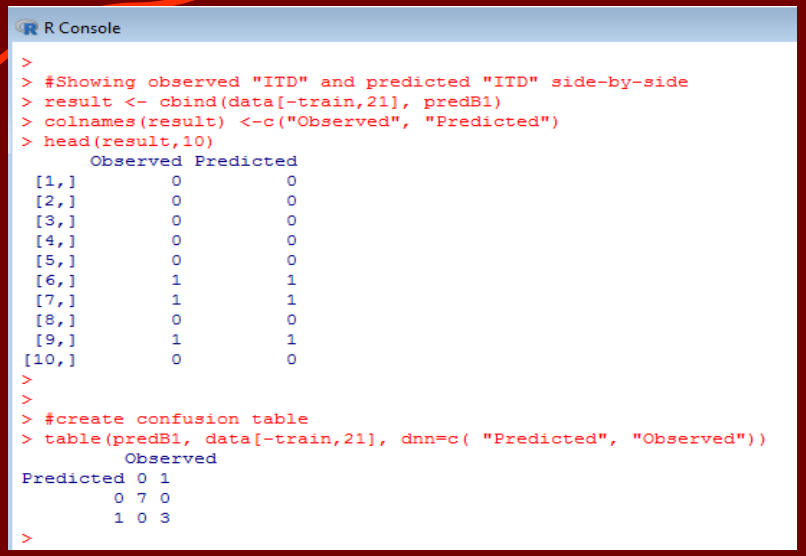
1. Convert data format: CSV / Excel -> RDA

	A	B	C	D	E	F	G	H
1	Pt_ID	INPPL1	CLPP	CDKN1B	BAD.pS155	TP53	DIABLO	PTPN11
2	1	0.656486	-1.25876	0.942978	-0.704074244	-0.3405	-0.5861	-0.18428
3	002	0.138922	-1.10418	1.518719	-0.866284883	0.137512	0.730412	1.33178
4	006	-0.66476	0.049663	0.484205	0.458060891	0.732413	1.354855	-1.54531
5	011	-1.11614	0.650572	-1.09003	0.555934117	0.158497	1.366448	-0.4657
6	012	-0.00163	1.396297	-0.63597	0.076881965	-0.22735	2.173629	0.260352
7	013	0.013124	1.455706	1.409363	1.788587769	0.298405	2.361544	-1.57627
8	015	-0.66865	0.23691	-0.25818	-0.417411308	0.739457	1.467231	-1.09529
9	017	-0.65688	1.616929	-0.97951	0.800944045	0.321133	2.124652	-0.73889
10	018	-0.64662	0.537541	0.571039	-0.851784736	-0.16169	0.70578	-0.64692
11	027	-0.40431	0.772609	0.416485	1.191255601	0.187356	2.224849	-1.02167
12	028	-0.41948	-0.08966	-0.41617	-0.081964146	-0.28816	0.545153	-1.47863



2. Wrote & Execute R program to retrieve data and run the Deep Learning network

```
library(deepnet)
set.seed(2016)
n=nrow(data)
train <-sample(1:n,52, FALSE)
X=data[train,1:20]
Y=data[train,21]
fitB<- sae.dnn.train(x=X,y=Y, hidden=c(20,15,10),
learningrate=0.5,momentum=0.5,
learningrate_scale=1,activationfun="sigm",output="sigm",
sae_output="linear",numepochs=970,batchsize=10,
hidden_dropout=0,visible_dropout=0)
Xtest <-data[-train,1:20]
Ytest <- data[-train,21]
predB <- nn.predict(fitB, Xtest)
predB1 <- ifelse(predB[,1]>=0.5,1,0)
result <- cbind(data[-train,21], predB1)
colnames(result) <-c("Observed", "Predicted")
head(result,10)
table(predB1, data[-train,21], dnn=c( "Predicted", "Observed"))
acc_rate=sum(predB1==sign(data[-train,21]))/10
round(acc_rate,2)
```



The Conventional Neural Network (EasyNN)

EasyNN-plus - [231 proteins FLT3_One Worksheet_set7.tvq]

File Edit Defaults View Zoom Insert Action Query Tools Macro Subset Window Help

	INPPL1	CLPP	CDKN1B	BAD#pS155	TP53	DIABLO	PTPN11	INPP5D
Q:0	0.6565	-1.2588	0.9430	-0.7041	-0.3405	-0.5861	-0.1843	0.2035
017	-0.6569	1.6169	-0.9795	0.8009	0.3211	2.1247	-0.7389	0.0049
018	-0.6466	0.5375	0.5710	-0.8518	-0.1617	0.7058	-0.6469	0.4357
039	-1.1188	0.5961	-0.6267	-0.8233	-0.1301	0.3634	-0.4419	0.8512
040	-0.4735	-0.7940	2.1253	-0.6282	-0.0377	-1.4185	0.3595	1.2376
047	-0.7940	1.0554	-0.4580	-0.2912	-0.4888	0.6168	-1.0629	-0.2325
050	-1.3011	1.7945	0.4580	-0.7403	-0.5329	-0.2363	-0.7729	0.5160
055	0.0674	0.2063	-0.6408	-0.1369	0.1439	1.3733	0.2856	0.4561
062	-0.6241	-0.0716	0.4145	-0.3515	-0.6492	0.7143	-0.6699	0.2033
080	1.5619	-1.0537	0.4188	-1.1927	-0.8135	-1.3643	-0.3588	-0.5821
086	0.9133	0.4326	-0.5745	-0.5775	0.1868	0.6111	1.0744	0.6587
142	0.8491	0.0696	-0.0672	-0.7066	-0.6306	-0.8644	-1.1175	-0.5032
146	0.6450	0.1643	-0.9670	1.2456	-0.0062	-1.5619	-0.4130	-0.6100
152	0.2236	0.3949	0.6140	0.6122	-0.6872	0.6877	0.2468	0.4575
168	-0.2590	0.6073	0.5436	1.3753	-0.1598	-0.1045	-0.3272	-0.9128
169	-0.0602	0.1403	0.8414	0.0824	-0.4433	-0.2014	-0.0404	0.3839

File: 231ProteinsClinData2-Set3.tvq User: Nghia Nguyen

General
231ProteinsClinData2-Set3.tvq
Learning cycles: 379 AutoSave cycles not set.
Training error: 0.009754 Validating error not known.
Validating results not known.

Grid
Input columns: 236
Output columns: 1
Serial columns: 0
Excluded columns: 1
Training example rows: 171
Validating example rows: 0
Querying example rows: 20
Excluded example rows: 0
Duplicated example rows: 0

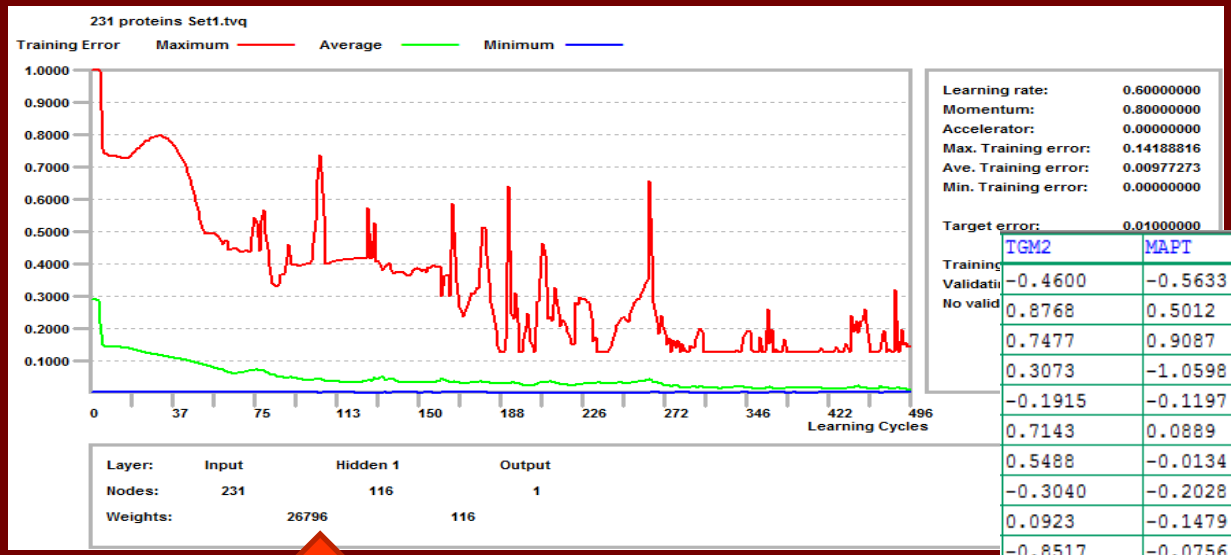
Network
Input nodes connected: 236
Hidden layer 1 nodes: 119
Hidden layer 2 nodes: 0
Hidden layer 3 nodes: 0
Output nodes: 1
Serial input nodes: 0
Serial output nodes: 0

Controls
Learning rate: 0.6000 Momentum: 0.8000
Target error: 0.0100 No extras enabled.

Validating rules
No columns have rules set.

Missing data action
The median value is used.

Show when a file is opened
History Save Refresh Continue



TGM2	MAPT	BIRC5	HSPB1	ITD	ITD-Origin+
-0.4600	-0.5633	-0.4048	-0.4527	~NEG	NEG
0.8768	0.5012	0.6318	1.4637	NEG	NEG
0.7477	0.9087	-1.0595	0.7779	NEG	NEG
0.3073	-1.0598	-1.1985	-0.6725	~~~NEG	POS
-0.1915	-0.1197	0.4362	0.0591	~~~POS	POS
0.7143	0.0889	0.0284	0.4699	NEG	NEG
0.5488	-0.0134	0.6210	0.0853	~~NEG	NEG
-0.3040	-0.2028	-0.0006	2.1193	~NEG	NEG
0.0923	-0.1479	0.3718	0.0229	~NEG	NEG
-0.8517	-0.0756	0.2088	0.0743	~~NEG	NEG

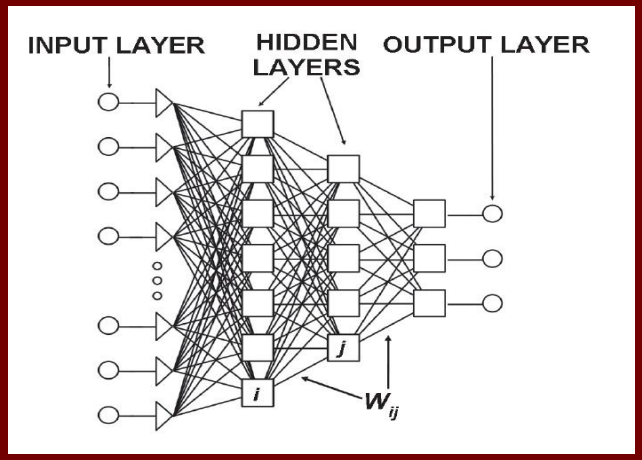
Analysis Steps

- First perform training for both the conventional and the Deep Learning neural networks with the original training set including all 231 proteins and compare the two networks in terms of accuracy in predicting FLT3-ITD mutation status in the cross-validation sets
- 10-fold cross-validation process: exclude 10 cases at a time to train the network and use the resultant network to test these 10 excluded cases)
- High dimensionality of protein expression data (231) is likely to introduce background noise in addition to relevant proteins in the training set. We try to reduce the dimensionality of the feature space to the most relevant number of proteins based on the ranking of the proteins in initial training
- The performance of the two neural networks in term of accuracy in predicting mutation status using this new scaled-down protein set will then be compared

Table 2. The List of the 20 Top-Ranking Proteins Used in Training

Column	Input Name	Importance
98	INPPL1	38.2049
46	CLPP	36.8005
165	CDKN1B	33.8749
13	BAD#pS155	33.2770
215	TP53	32.9059
54	DIABLO	29.3504
171	PTPN11	29.1863
97	INPP5D	28.4851
103	JMJD6	28.1224
182	SIRT1	28.0780
221	VHL	28.0422
8	ATF3	27.3000
66	ERBB2	27.0661
211	TAZ#pS89	26.7620
124	MET#pY1230_1234_1235	25.3951
5	ARC	24.3036
213	TGM2	23.8800
120	MAPT	23.6661
22	BIRC5	23.6211
94	HSPB1	23.4504

(The ranking of protein is based on the sum of the absolute weights of the connections from the input node to all the nodes in the first hidden layer)



Optimal Configurations

- Optimal network layers:
  -Conventional network: 1 hidden layer
  -Deep Learning network: 3 hidden layer
- Dimension of the Protein Set
  -Using a smaller or larger number of proteins than 20 do not yield better accuracy (data not shown) indicating that 20 is the optimal number of proteins for this study.

  -It appears that fewer than 20 proteins contain insufficient data for prediction. Conversely, more than 20 proteins would introduce much background noise compromising accuracy.

  -Reduction in data points for analysis:
  231 proteins x 62 cases = 14,322 data points
  down to: 20 proteins x 62 cases = 1,240 data points

Pre-Training in Deep Learning: Graphic display representing the original features (Left) and the more compact extracted features (Right)

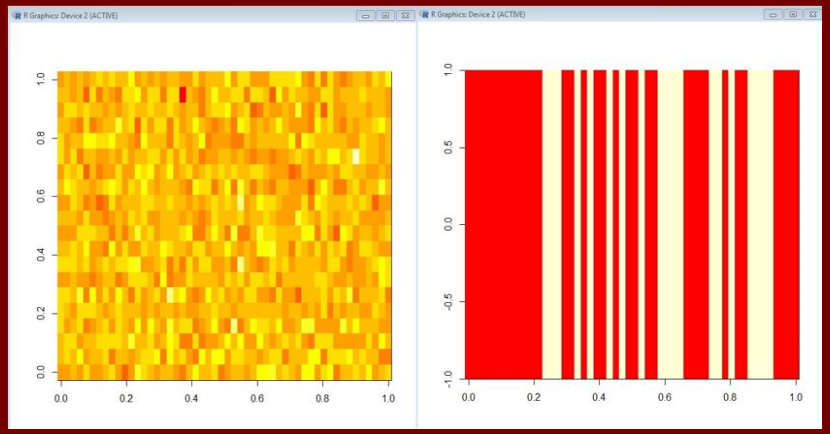


Table 1. Accuracy in Predicting FLT3-ITD Status with Different Protein Data Sets by Conventional Neural Networks vs. Deep Learning Networks

Neural Networks	231 Protein Data Set		20 Protein Data Set	
Conventional	Validation Set No.	Accuracy	Validation Set No.	Accuracy
	1	NC*	1	80%
	2	80%	2	80%
	3	60%	3	90%
	4	60%	4	90%
	5	NC*	5	90%
	6	90%	6	90%
	7	70%	7	90%
	Mean=	72%	Mean=	87%
Deep Learning	1	80%	1	100%
	2	90%	2	100%
	3	70%	3	80%
	4	80%	4	100%
	5	80%	5	100%
	6	90%	6	100%
	7	80%	7	100%
	Mean=	81%	Mean=	97%

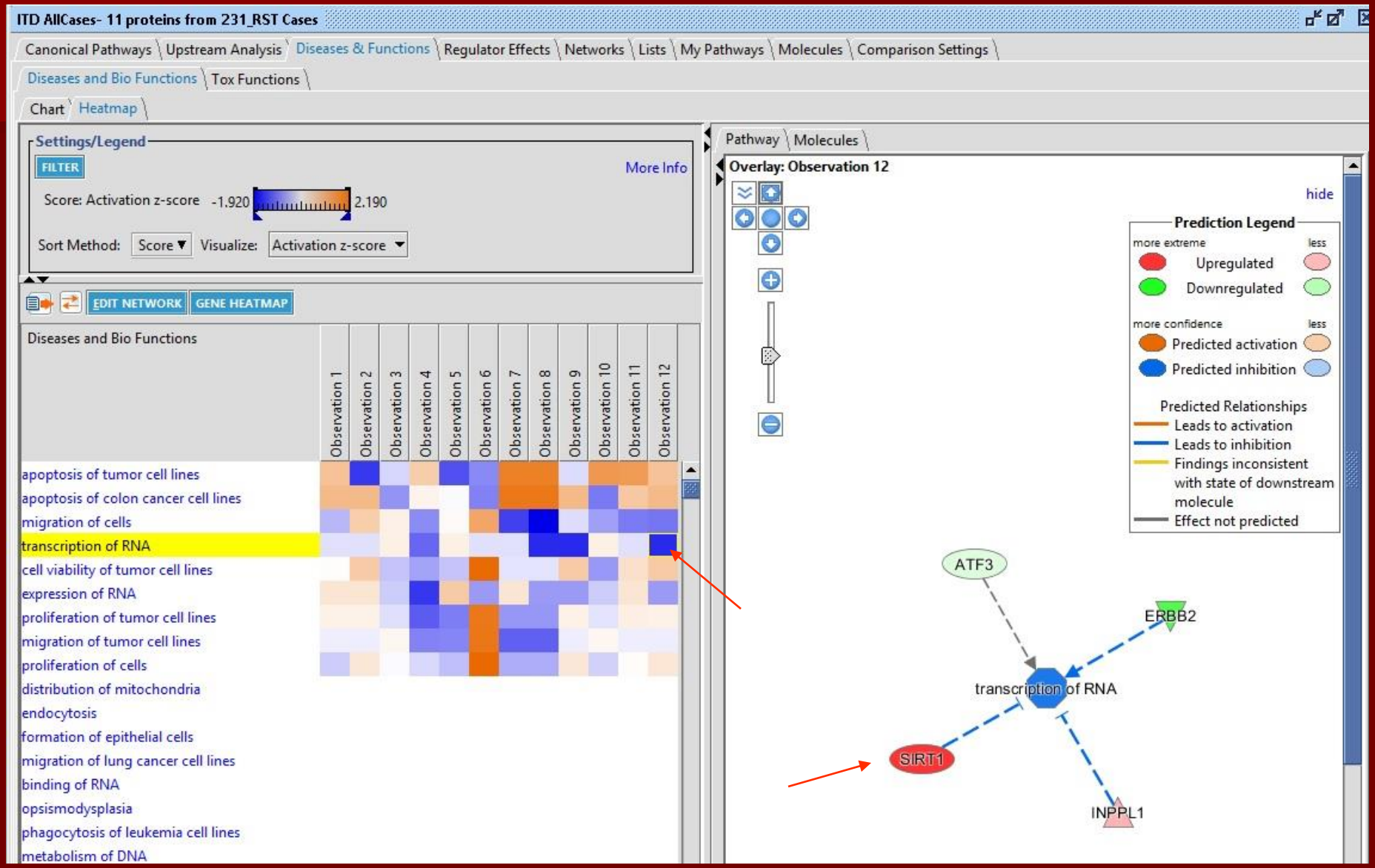
sensitivity of 75%, and specificity of 93%

sensitivity of 90%, and specificity of 100%

*: no convergence

Ingenuity Pathway Analysis: Heat map for Disease/Bio Functions

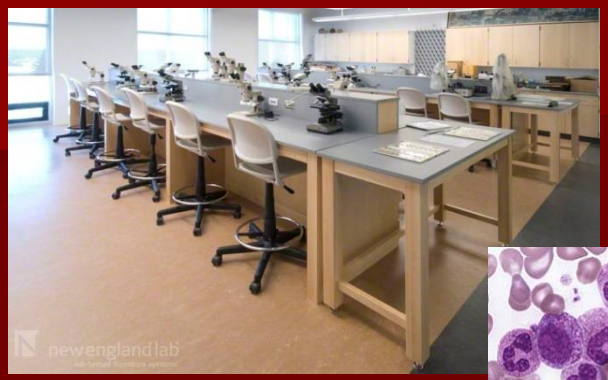
RST cases: 6 cases FLT3-neg, 6 cases FLT3-pos

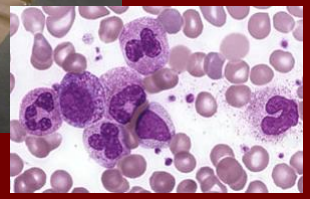


SUMMARY

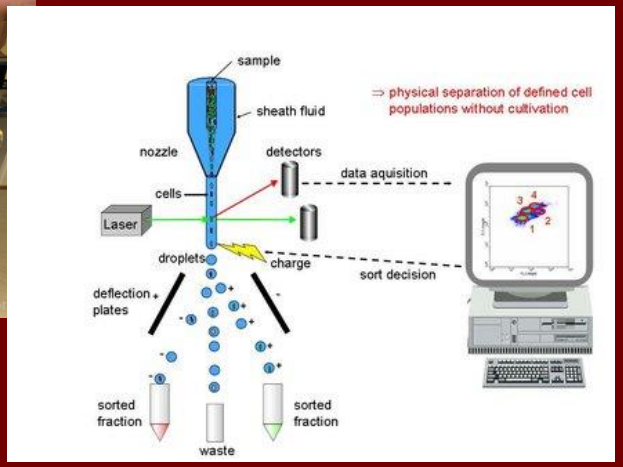
- In the present study, we explore how Deep Learning can be utilized for proteomics analysis in acute myeloid leukemia (AML). Specifically we attempt to determine a set of critical proteins that are associated with FLT3-ITD mutation out of 231 proteins available in 62 newly-diagnosed AML patients.
- Dimensional reduction was initially performed to reduce the number of critical proteins from 231 down to 20. We then show how Deep Learning which incorporates unsupervised feature training can be used to find excellent correlation between FLT3-ITD mutation with levels of these 20 proteins (**an accuracy of 97%, sensitivity of 90% and specificity of 100%**).
- Deep Learning against other algorithms (20 protein set):
  -Conventional neural network (86.7%)
  -K-Nearest Neighbor (88.6%)
  -Logistic regression (81.2%)
  -Support Vector Machines (89.4%)
  -Random Forest (85.7%)
- This study yield a critical dataset of 20 key proteins in FLT3-ITD mutation for potential further research to:
  -Determine important protein pathways for this mutation in AML, explore pathogenesis involving the mutation,
  -Monitor chemotherapy response, and design personalized treatment.
- This study provides a proof-of-concept for using Deep Learning neural network as a more accurate approach for modeling big data in cancer genomics and proteomics *[manuscript in preparation]*

Looking Forward: Deep Learning as a Disruptive Technology





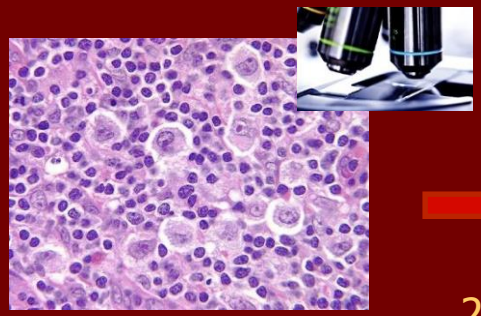
31 years



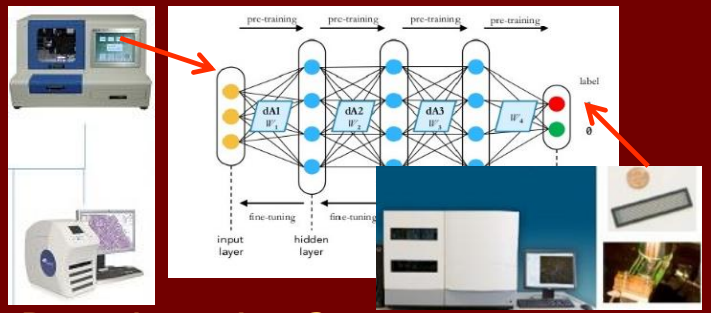
Hematology Lab
(600 bed hospital, circa 1985)
-Rudimentary CBC instruments
-10 microscope stations
for WBC differential counts

Hematology Lab
(900 bed hospital, circa 2016)
-Sophisticated CBC instruments
that release most WBC diff counts
-1 microscope station to check on WBC flags

What will pathology examination be like in another 31 years?



2016: microscope
->H&E, IHCs



2047: Deep Learning ?
-Digital whole slide imaging -> histol DX
-NGS-> genomic analysis -> mol DX