COMPUTER SCIENCE

INDIANA UNIVERSITY

School of Informatics and Computing
Bloomington

LEARNING PROPORTIONS IN A SEMI-SUPERVISED SETTING: A CASE STUDY IN PRECISION MEDICINE

Predrag Radivojac

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATICS

INDIANA UNIVERSITY, BLOOMINGTON

November 28, 2016

From Hahn, Matthew William

Subject **Tweet by Eduardo Eyras on Twitter**

To Me <predrag@indiana.edu>

Reply Forward Archive

Eduardo Eyras (@EduEyras)

7/14/15, 9:10 PM

One Third of the Alternative Splicing Isoforms produce Functional Proteins (P. Kim's lab) cell.com/cell-reports/a…

Download the Twitter app

Cell Reports
Report

OPEN ACCESS
CellPress

Semi-supervised Learning Predicts Approximately One Third of the Alternative Splicing Isoforms as Functional Prot

Yanqi Hao,[1,2,10] Recep Colak,[1,2,10] Joa
Mathias Wilhelm,[4] Bernhard Kuster,[4,5]
[1]Terrence Donnelly Centre for Cellular and B
[2]Department of Computer Science, Universit
[3]Department of Medical Biophysics, Univers
[4]Chair for Proteomics and Bioanalytics

Predicted Functional Proteins and Characterization of Highly Spliced Genes

Given the relatively high accuracy of our predictions, we ran PULSE to label 15,639 unlabeled transcripts from the BodyMap data set (Cabili et al., 2011; Colak et al., 2013; see Figure S1B for the experimental setup). At a 90% true-positive rate, we predict that about 32% of isoforms are functional (Figure 2C), roughly consistent with one school of thought, which estimate around 20%–30% to be functional (Floris et al., 2011; Rodriguez et al., 2013). Thus, AS leads to a sizeable number of previously uncharacterized proteins (a total of 5,023 in this data set alone). To prevent any biases/errors that might be caused by using Uniprot canonical isoforms for anchoring (the mechanism by which we quantify how much an alternative isoform differs from its canonical pair—see Experimental Procedures), we repeated the analysis using APPRIS principal isoforms for anchoring. The score distribution remains qualitatively unchanged (Figure 2C).

From Hahn, Matthew William

Subject **Tweet by Eduardo Eyras on Twitter**

To Me <predrag@indiana.edu>

Reply Forward Archive

Eduardo Eyras (@EduEyras)

7/14/15, 9:10 PM

One Third of the Alternative Splicing Isoforms produce Functional Proteins (P. Kim's lab) cell.com/cell-reports/a…

Download the Twitter app

At some prediction threshold [one] Third of the Alternative Splicing Isoforms predicted to produce Functional Proteins….

Nucleic Acids Research, 1999, Vol. 27, No. 1 237–239

PhosphoBase, a database of phosphorylation sites: release 2.0

Andres Kreegipuu, Nikolaj Blom[1,*] and Søren Brunak[1]

INTRODUCTION

Protein phosphorylation is a key event in many signal transduction pathways of biological systems. Protein kinases recognize and phosphorylate specific amino acid residues (mainly serine, threonine or tyrosine) in the substrate proteins. The research of protein phosphorylation has been essential in understanding intracellular signaling pathways. The number of identified phosphorylation sites is steadily growing—several thousand are now known. However, this seems to be only a small fraction of all potential phosphorylation possibilities since the estimated fraction of phosphoproteins may be as high as 30–50% of the total protein repertoire (1).

According to recent estimates, at least 30% of proteins in a eukaryotic proteome are phosphorylated[2], and >100,000 potential phosphorylation sites exist[3]. This poses a major analytical challenge, especially given that phosphorylated proteins are often present at substoichiometric ratios relative to their nonphosphorylated counterparts, and there can be multiple phosphorylation isoforms of a given protein. Furthermore, some phosphorylation events are transient and their detection requires monitoring of very short timescales.

Phosphoproteomics takes it easy

Paola Picotti

The EasyPhos pipeline simplifies analysis of phosphorylation-dependent signaling networks at high temporal resolution.

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with

coordinate regulation of the genes in the clusters.

- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a

What is the Fraction of Enzymes in a Genome?

Proteins known to be enzymes:

The UniProt database in 2010: 28%

The UniProt database in 2014: 40%

Current state of the affairs:

E. coli: $\frac{1154}{4433} = 0.26$

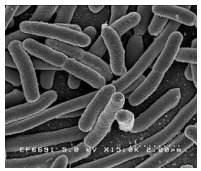
Yeast: $\frac{1477}{6621} = 0.22$

Arabidopsis: $\frac{1563}{13100} = 0.12$

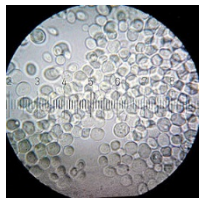
Mouse: $\frac{1250}{16676} = 0.07$

Human: $\frac{2647}{20193} = 0.13$

E. coli



Yeast



Arabidopsis

Mouse

Images from the Internet.

What is the Fraction of Enzymes in a Genome?

Expert opinion:

Charles Dann, Chemistry

E. coli:	0.35
Yeast:	0.45
Arabidopsis:	0.40
Mouse:	0.25
Human:	0.25

Tuli Mukhopadhyay, Biology

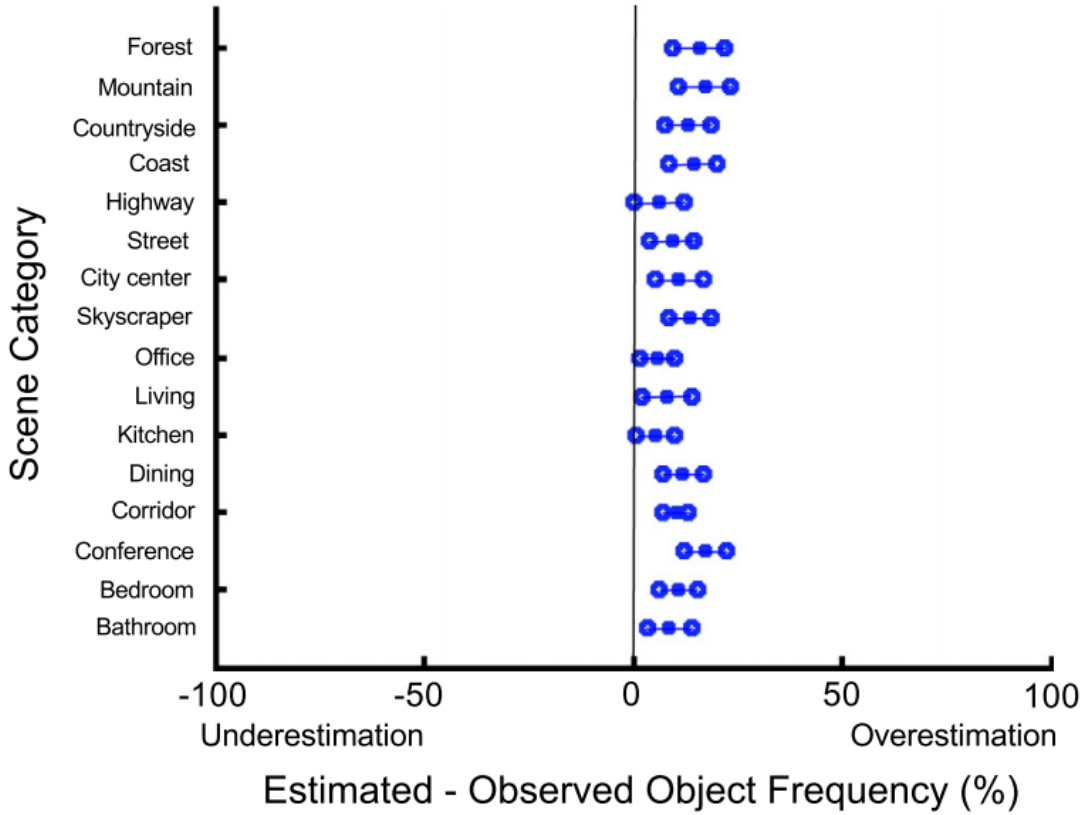
E. coli:	0.85
Yeast:	0.85
Arabidopsis:	0.85
Mouse:	0.85
Human:	0.85

Yuzhen Ye, Computer Science

E. coli:	0.40
Yeast:	0.40
Arabidopsis:	0.30
Mouse:	0.30
Human:	0.30

Example From Psychology

Estimations of object frequency are frequently overestimated!



Greene MR. Estimations of object frequency are frequently overestimated. *Cognition* (2016) 149: 6-10.

AND SO WE GO...

Estimation of proportions:

* Interesting problem

* Needs rigorous theoretical treatment

* Needs algorithms

* Needs to be more precisely communicated

Hi Pedja,
You pose interesting questions. I'd expect yeast to have the highest enzyme fraction as it does not need to have conserved genes for multicellular development, cognition, etc. (though many of these processes requires signaling pathways with enzymes). So here are my estimates for enzyme fraction, based entirely on intuition.

Yeast ~45 %; E. coli ~35 %; Mouse ~25 %; Human ~25 %; Arabidopsis ~40 % (no idea here)

I imagine I may hit low on all of these...

CD3

Supervised Learning Problem

Given:

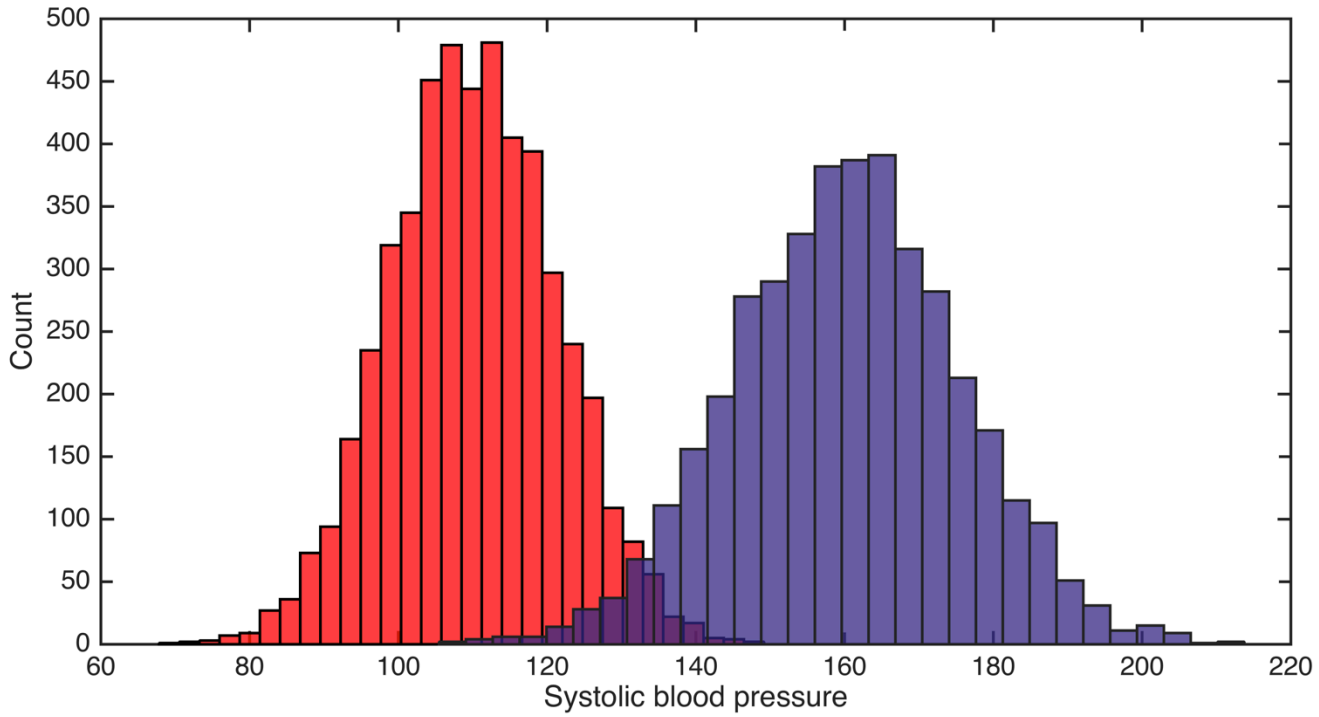
X_{red}: sample from people w/o heart disease

X_{blue}: sample from people w/ heart disease

Goal: predict heart disease

$x \in \mathbb{R}$

$y \in \{\text{disease}, \text{no disease}\}$



Supervised Learning Problem

Given:

$x \in \mathcal{X}, y \in \mathcal{Y} = \{0, 1\}$

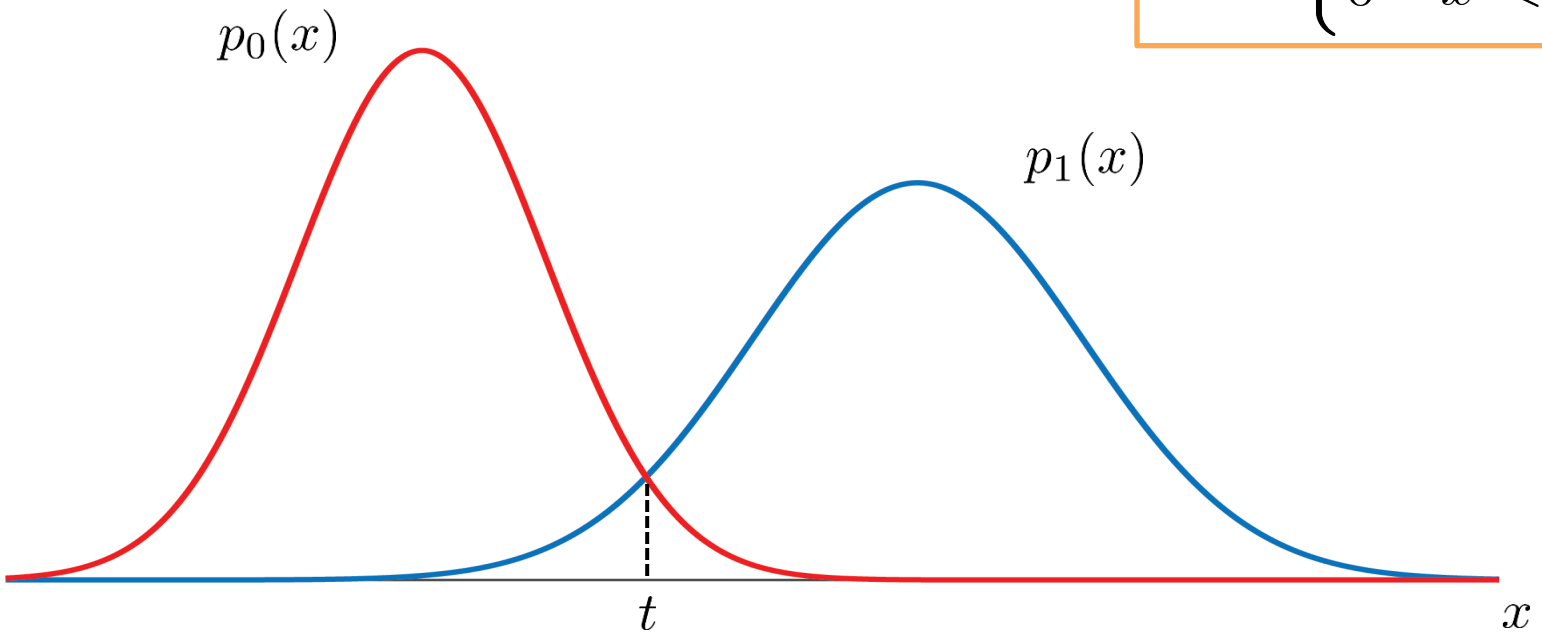
X_0: sample from $p_0(x) = p(x|y = 0)$

X_1: sample from $p_1(x) = p(x|y = 1)$

Goal: learn how x relates to y

Predictor:

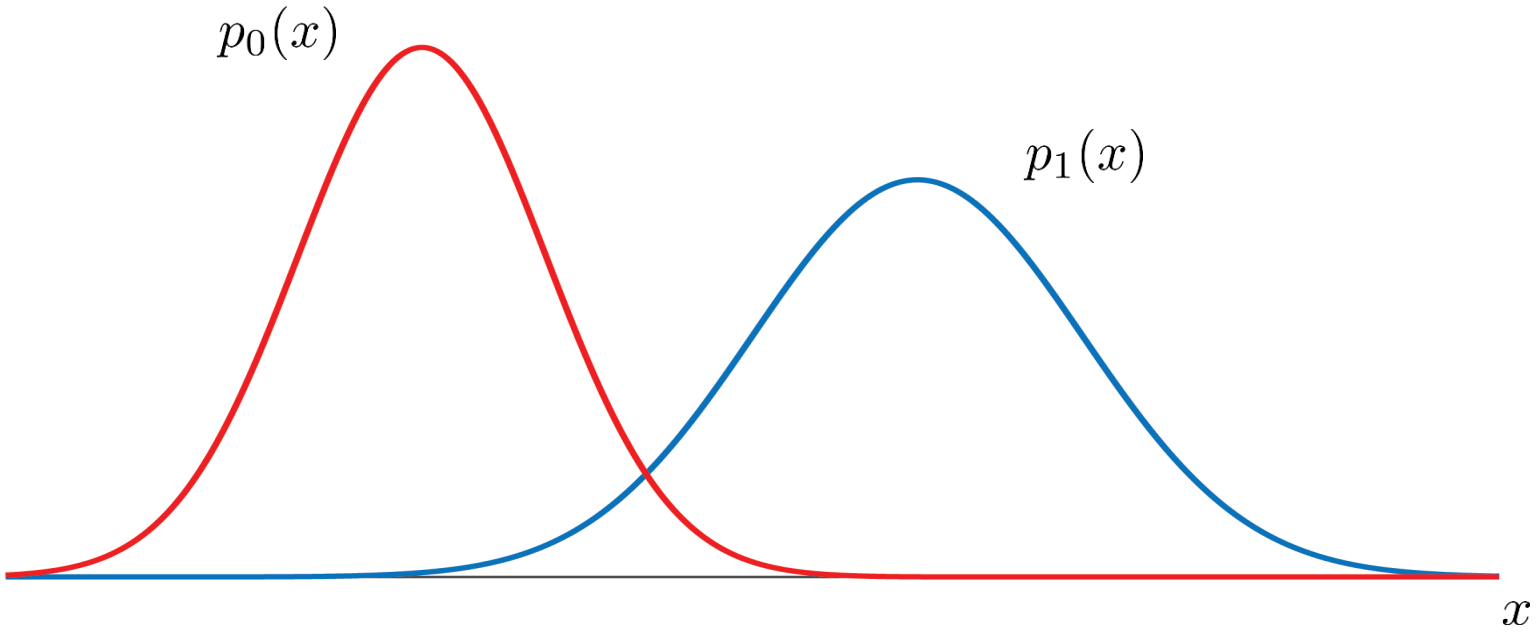
$$\hat{y} = \begin{cases} 1 & x \geq t \\ 0 & x < t \end{cases}$$



Supervised Learning Problem

Goal: learn how x relates to y $\quad x \in \mathcal{X}, y \in \mathcal{Y} = \{0, 1\}$

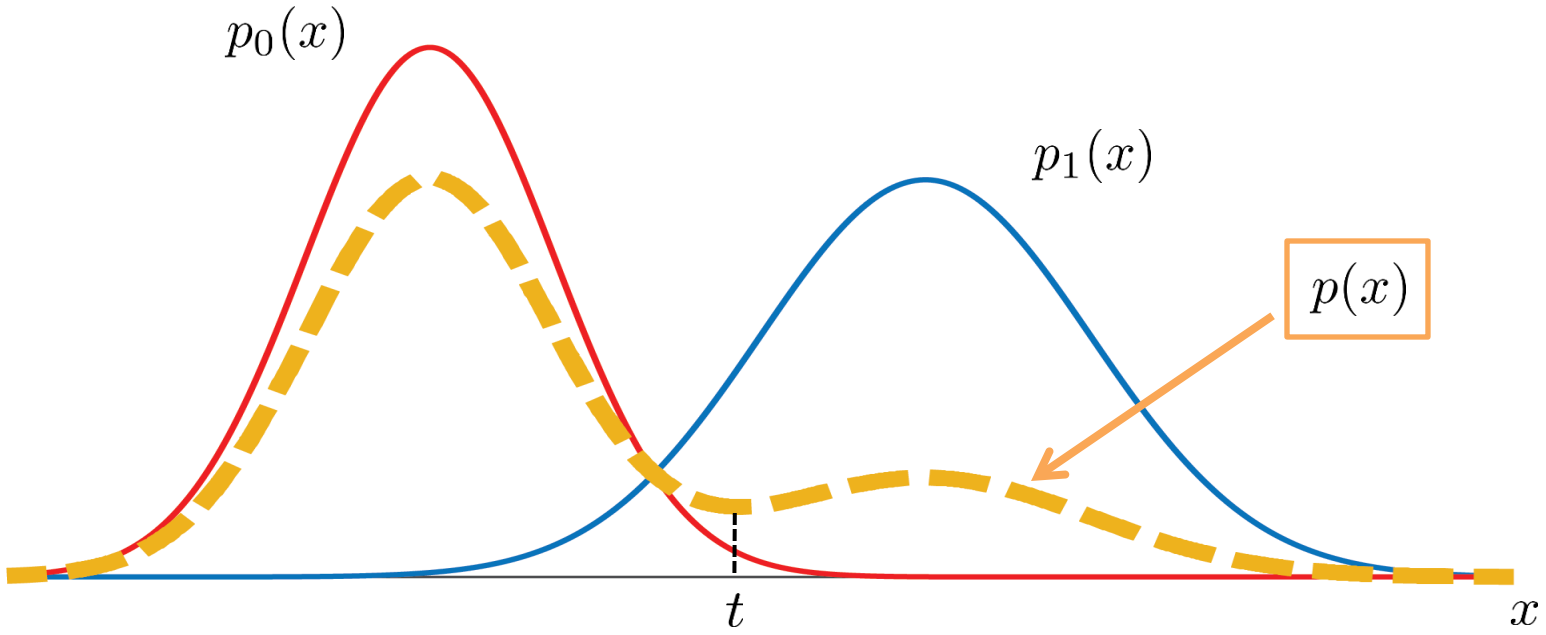
$$p(y|x) = \frac{p(x,y)}{p(x)} \propto p(x|y)p(y)$$



Supervised Learning Problem

Goal: learn how x relates to y $\quad x \in \mathcal{X}, y \in \mathcal{Y} = \{0, 1\}$

$$p(y|x) = \frac{p(x,y)}{p(x)} \propto p(x|y)p(y)$$



Semi-Supervised Learning Problem

Given:

$x \in \mathcal{X}, y \in \mathcal{Y} = \{0, 1\}$

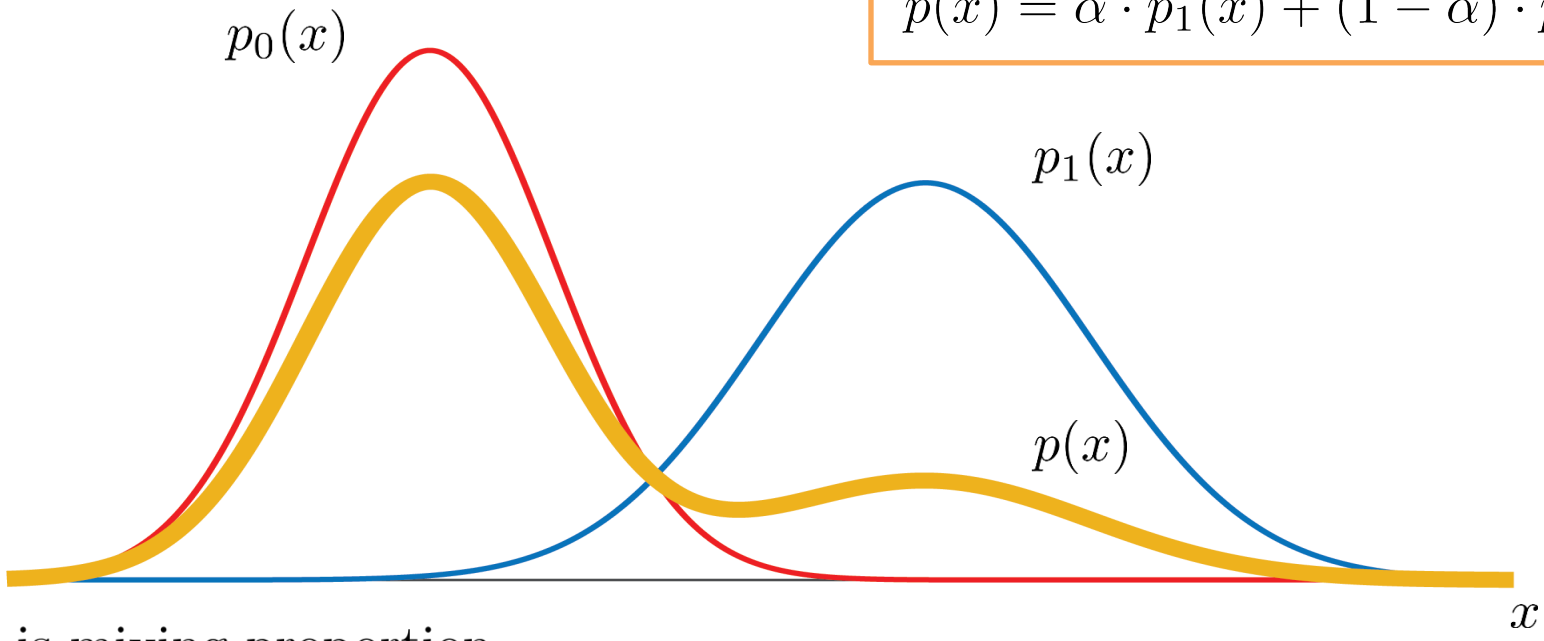
X_0: sample from $p_0(x) = p(x|y = 0)$

X_1: sample from $p_1(x) = p(x|y = 1)$

X: sample from $p(x)$

Goal: learn how x relates to y

$$p(x) = \alpha \cdot p_1(x) + (1 - \alpha) \cdot p_0(x)$$



$\alpha \in (0, 1)$ is mixing proportion

Unsupervised Learning Problem

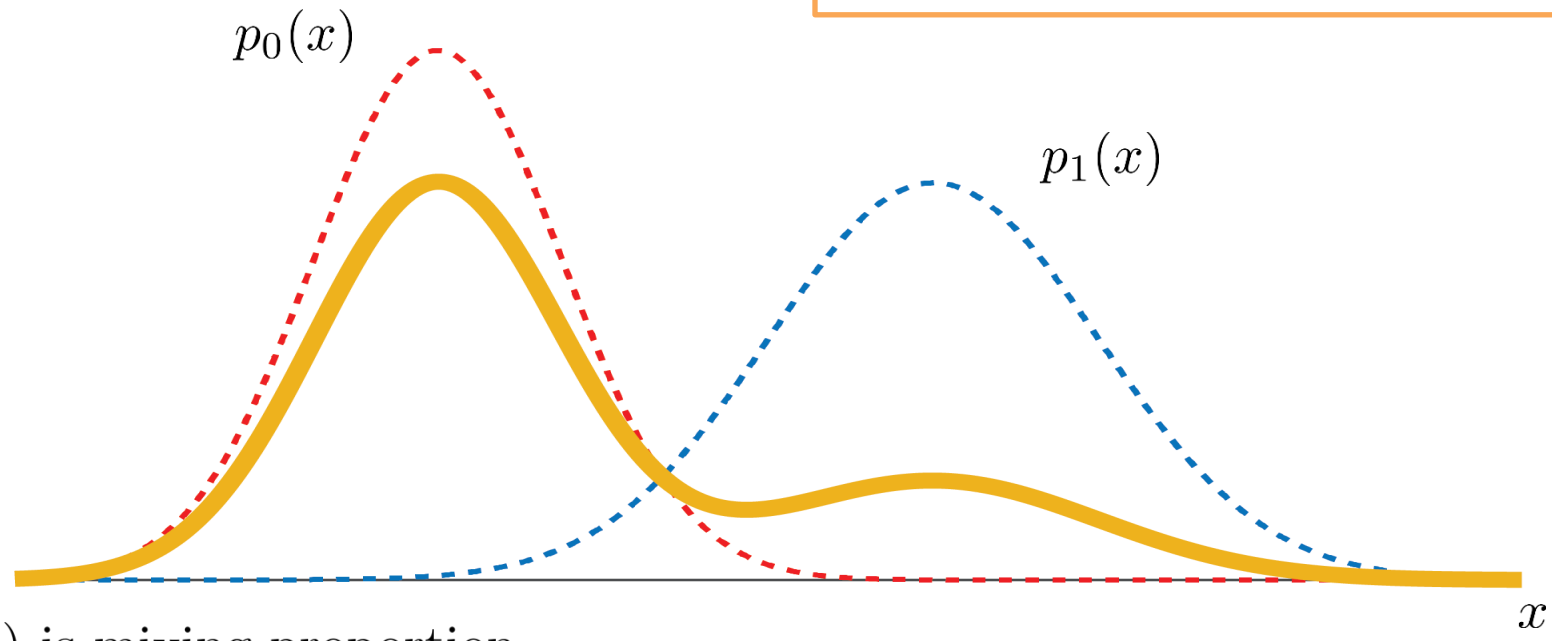
Given:

$x \in \mathcal{X}, y \in \mathcal{Y} = \{0, 1\}$

X = sample from $p(x)$

Goal: learn $p_0(x)$, $p_1(x)$ and α

$$p(x) = \alpha \cdot p_1(x) + (1 - \alpha) \cdot p_0(x)$$



$\alpha \in (0, 1)$ is mixing proportion

Positive-Unlabeled Learning Problem (PU)

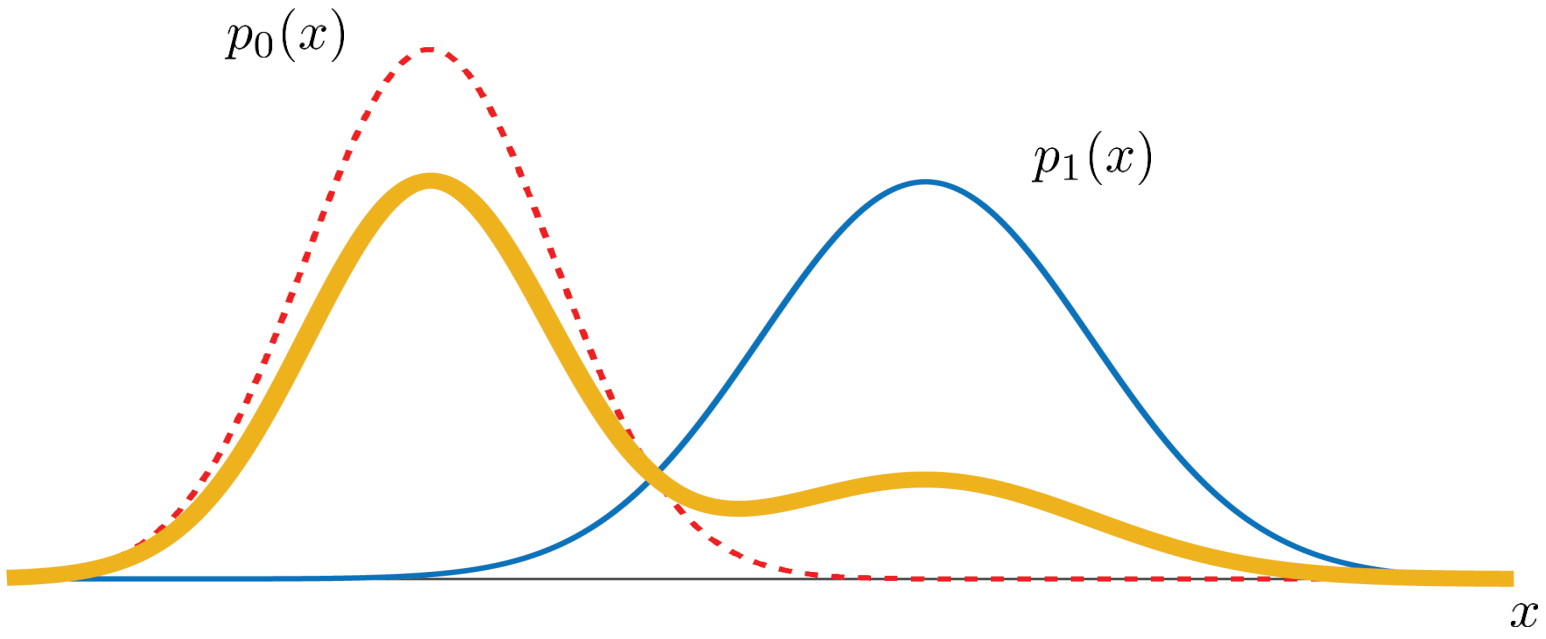
Given:

$x \in \mathcal{X}, y \in \mathcal{Y} = \{0, 1\}$

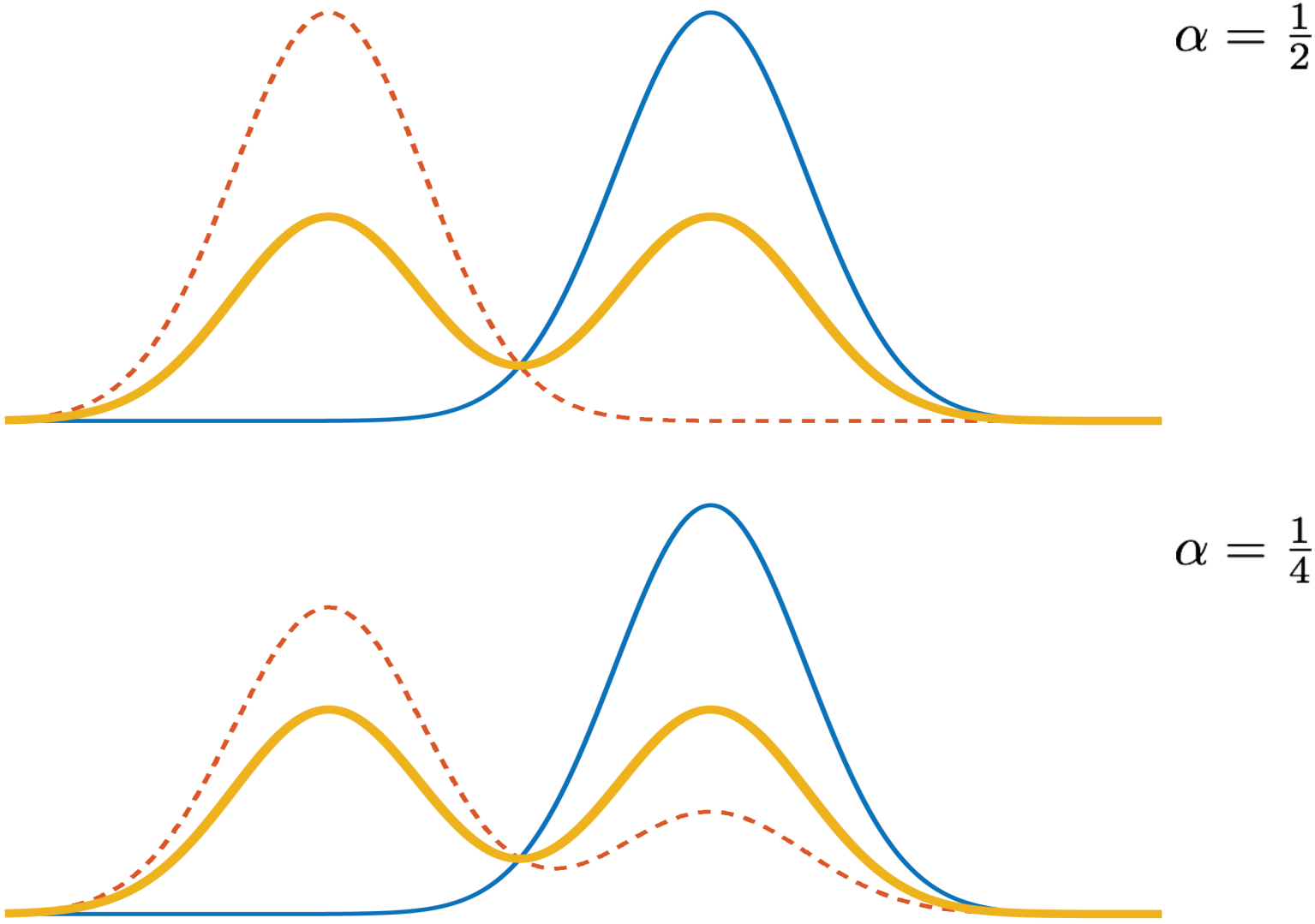
X_1: sample from $p_1(x) = p(x|y = 1)$

X: sample from $p(x)$

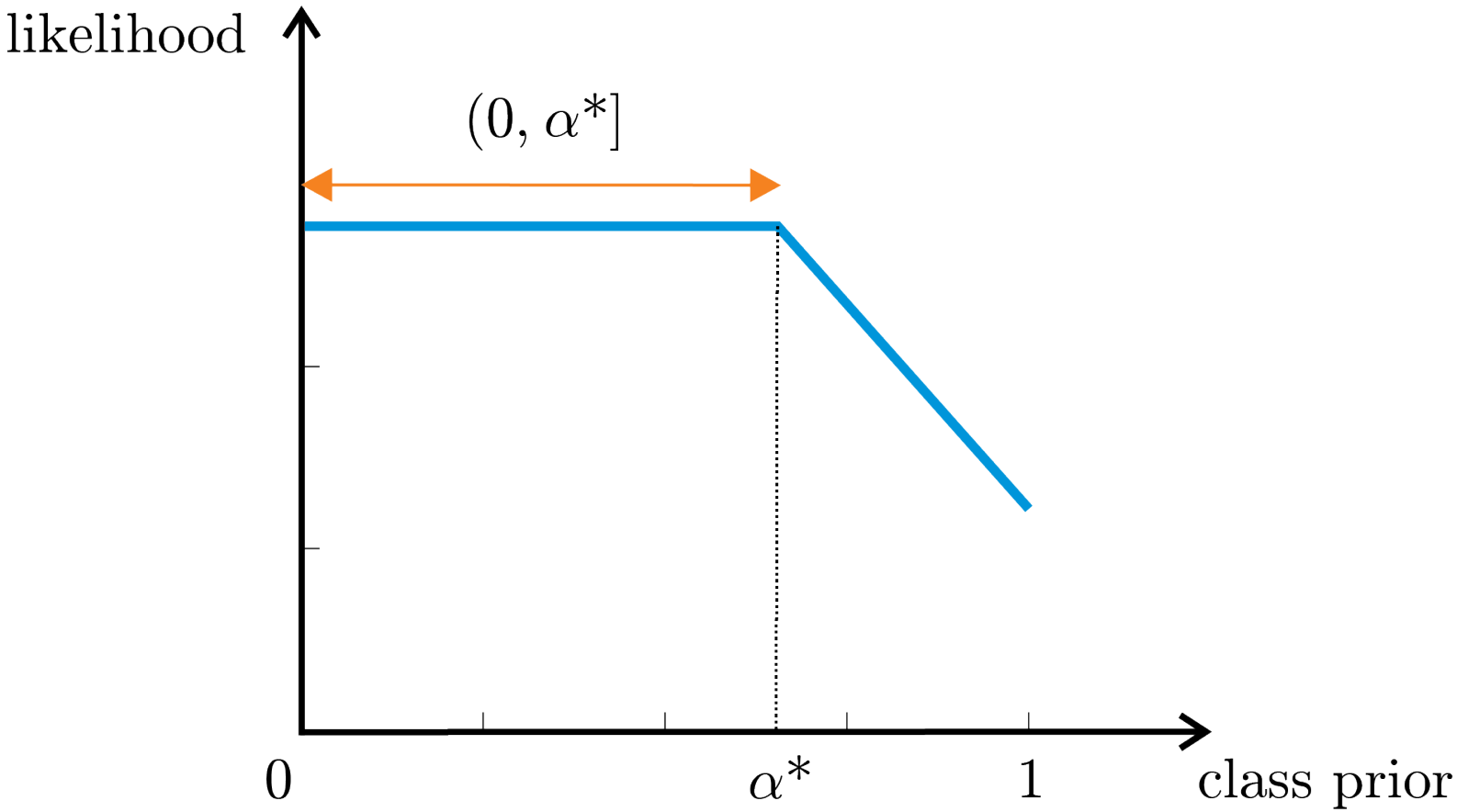
Goal: learn how x relates to y



IDENTIFIABILITY



Anticipated Likelihood Function



ENZYMES: EXPERIMENTAL PROTOCOL

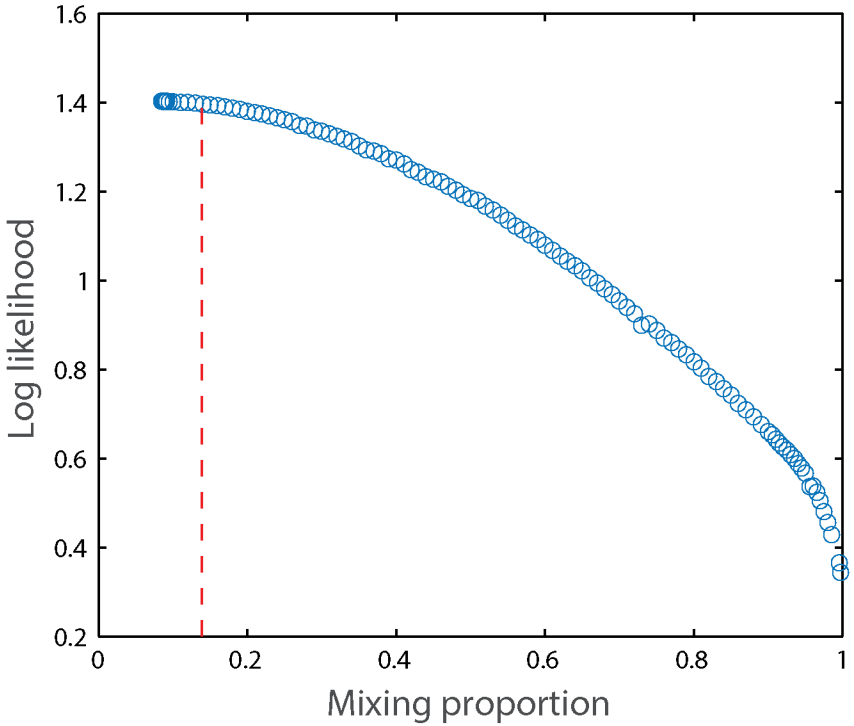
>sp|P04637|P53_HUMAN Cellular tumor antigen p53
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHS**VVVP**YEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

Develop an SVM predictor

- Linear kernel
- AUC ≈ 75%
- Platt's correction

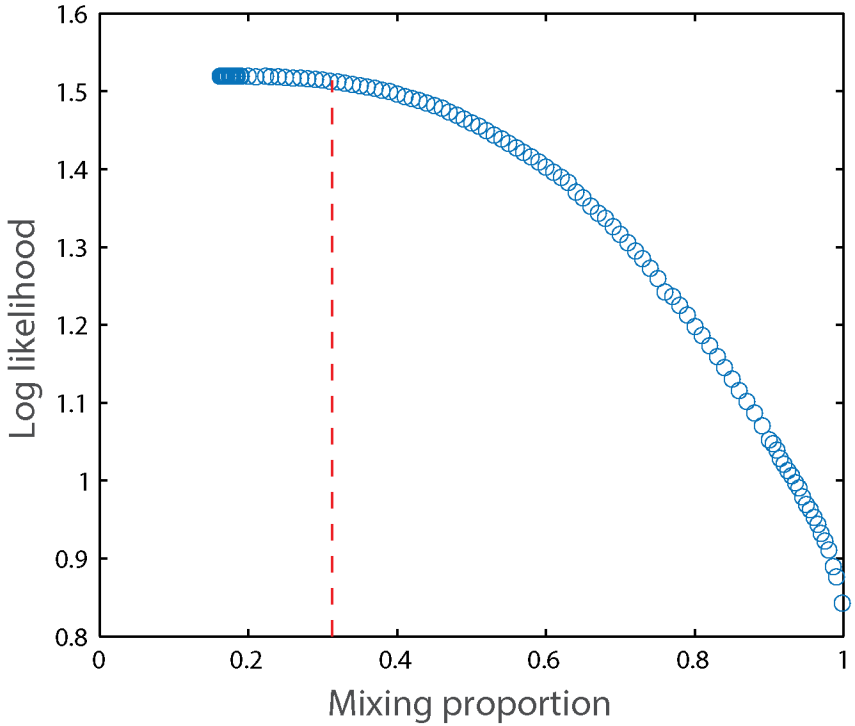
Results: Enzymes

Yeast



We: 0.33
Elkan-Noto: 0.62
Charles: 0.45
Yuzhen: 0.40
Tuli: 0.85

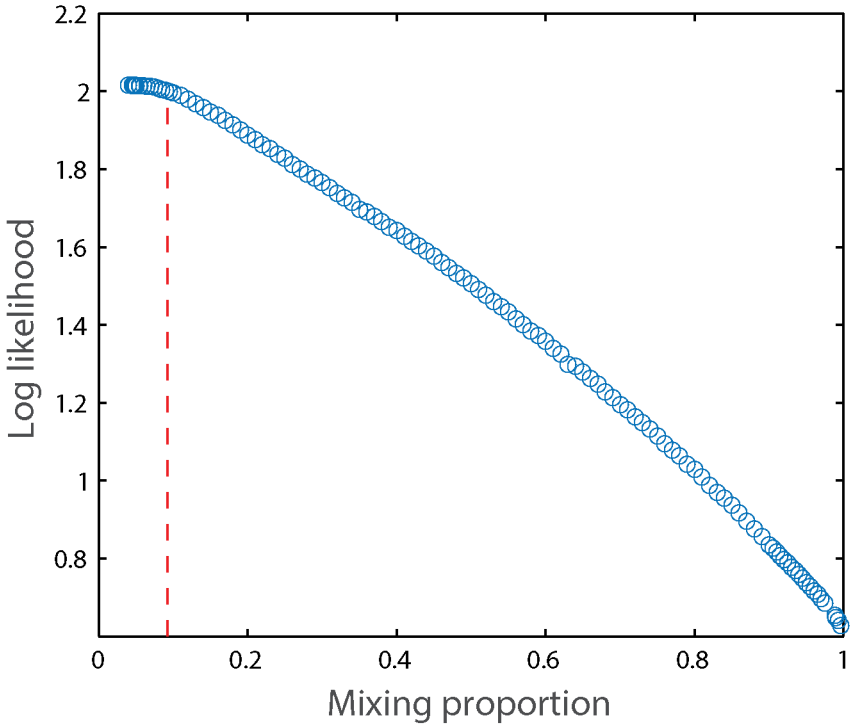
E. coli



We: 0.49
Elkan-Noto: 0.76
Charles: 0.35
Yuzhen: 0.40
Tuli: 0.85

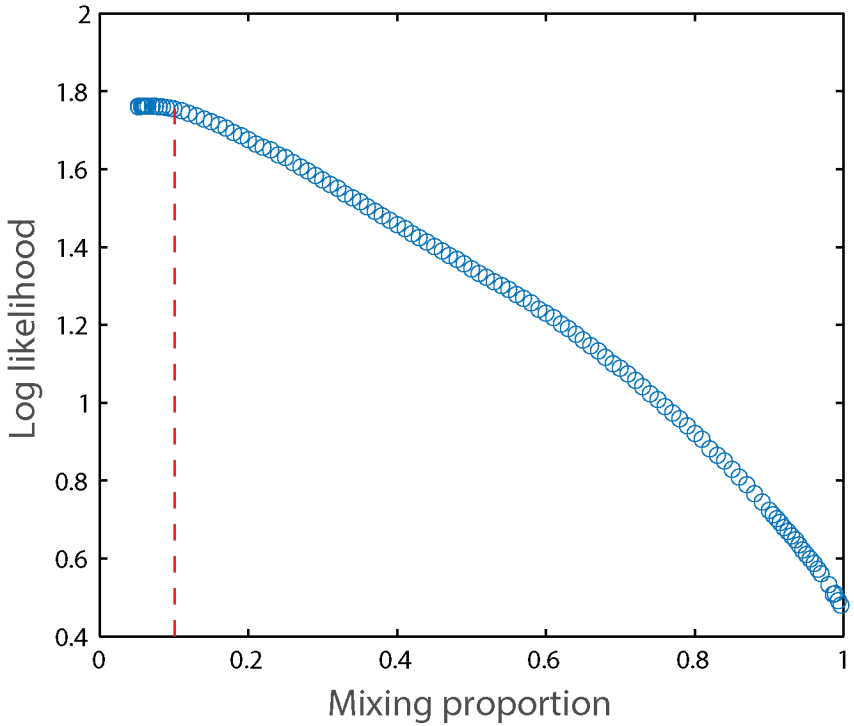
Results: Enzymes

Arabidopsis



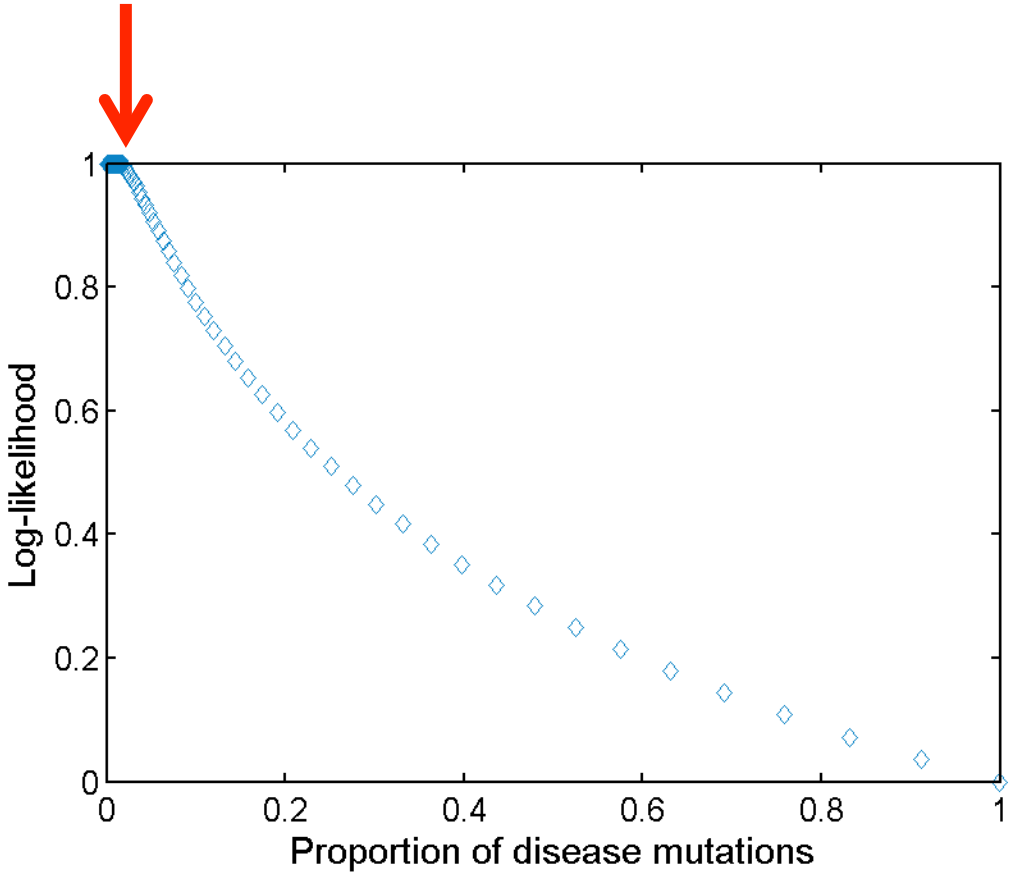
We: 0.20
Elkan-Noto: 0.41
Charles: 0.40
Yuzhen: 0.30
Tuli: 0.85

Human

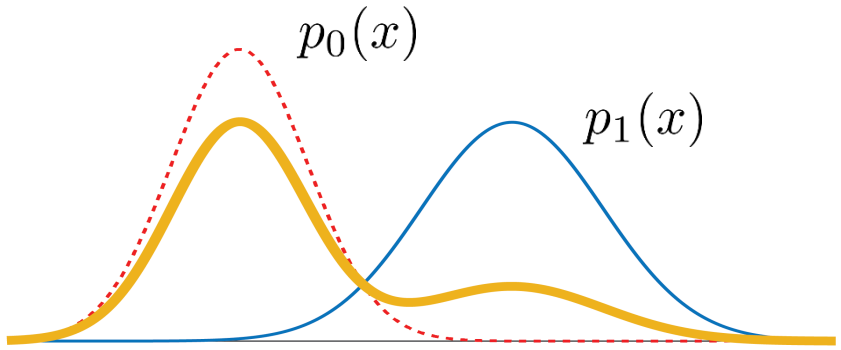


We: 0.21
Elkan-Noto: 0.45
Charles: 0.25
Yuzhen: 0.30
Tuli: 0.85

Disease Mutations in Human Exome



About 1.5% of human nsSNPs are disease-associated!



PRECISION MEDICINE

01/20/2015

Precision Medicine

the science and practice of matching the best diagnostic, therapeutic and prevention strategies to promote health that are tailored to an individual's genetic, biological, behavioral and psychosocial characteristics

"So tonight, I'm launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes, and to give all of us access to the personalized information we need to keep ourselves and our families healthier. We can do this." – President Barack Obama.

PRECISION MEDICINE

All of Us[SM] Research Program

WHAT IS IT?

Precision medicine is a groundbreaking approach to disease prevention and treatment based on people's individual difference in environment, genes and lifestyle.

The *All of Us* Research Program will lay the foundation for using this approach in **clinical practice.**

WHAT ARE THE GOALS?

Engage a group of **1 million or more U.S. research participants** who will share biological samples, genetic data and diet/lifestyle information, all linked to their electronic health records. This data will allow researchers to develop more precise treatments for **many diseases and conditions**.

Pioneer a new model of research that emphasizes **engaged research participants, responsible data sharing and privacy protection.**

Research based on the cohort data will:

- Lay **scientific foundation** for precision medicine
- Help identify new ways to **treat and prevent disease**
- Test whether **mobile devices**, such as phones and tablets, ca encourage healthy behaviors
- Help develop the **right drug** for the **right person** at the **right dose**

WHY NOW?

The **time is right** because:

We have a greater understanding of human genes

People are more engaged in healthcare and research

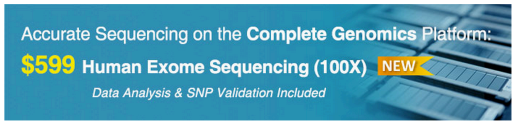
We have the tools to track health information and use large databases

Research technologies have improved

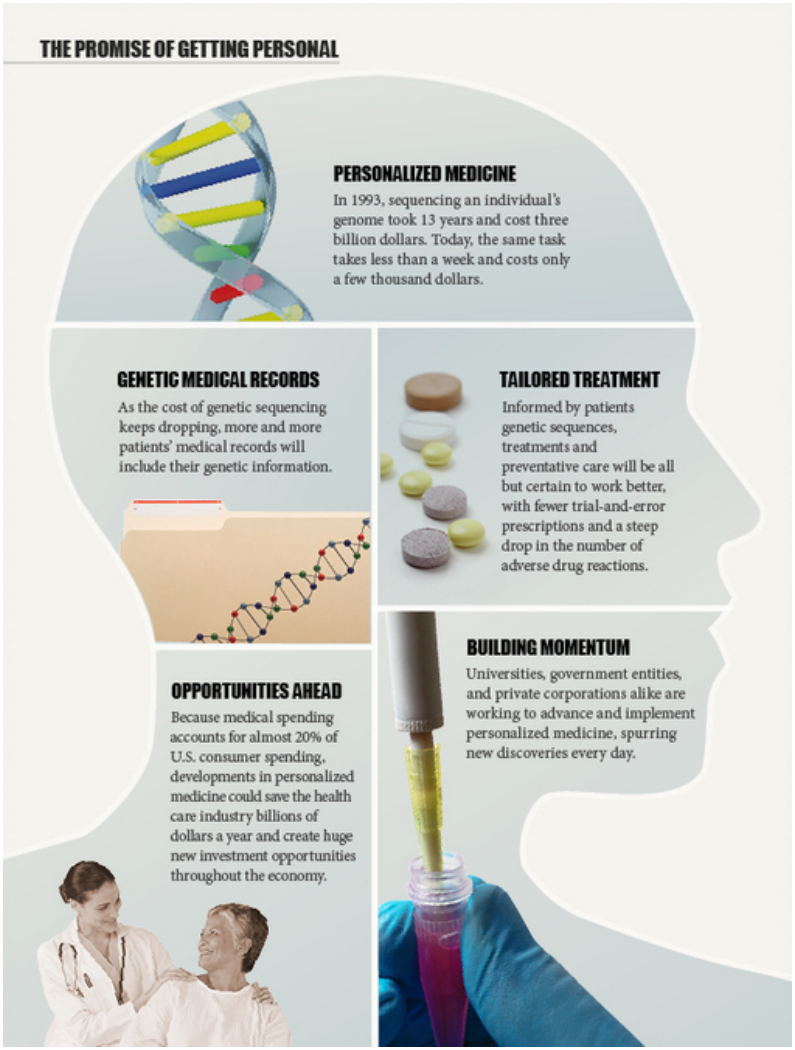
Follow the Program's progress and be one of the first to join this landmark effort.

www.nih.gov/AllofUs-Research-Program

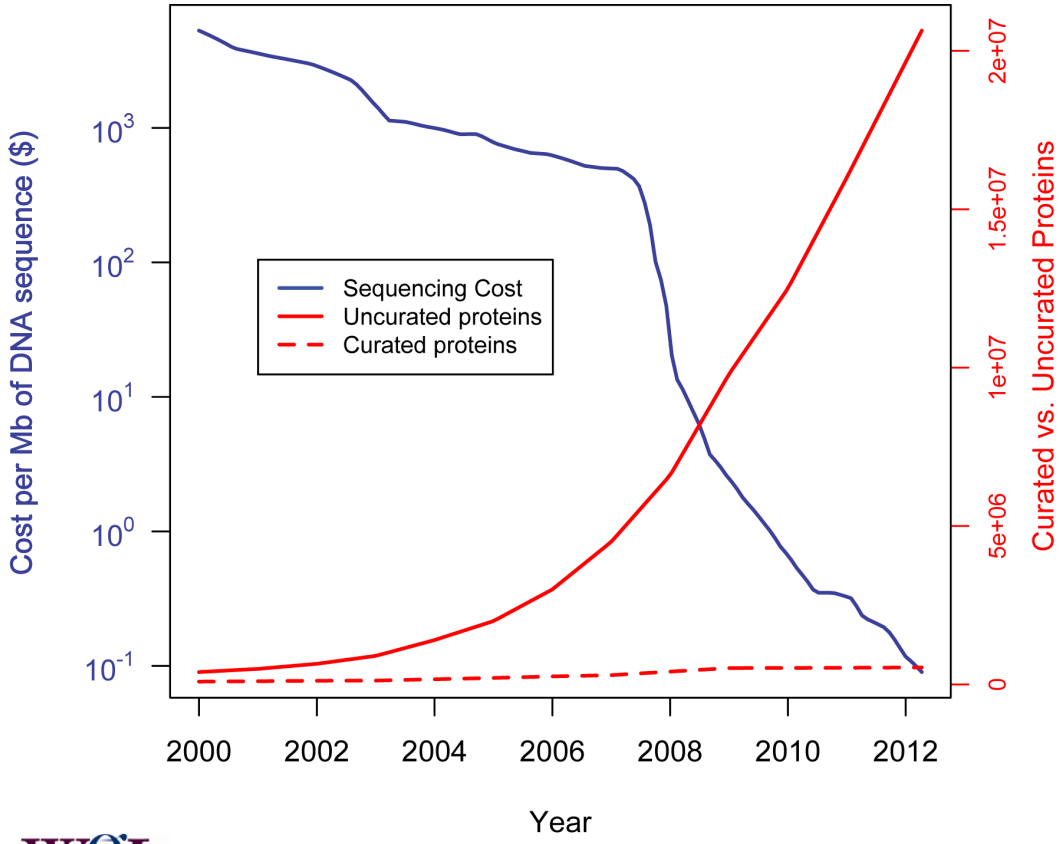
GENOME SEQUENCING



http://bgiamericas.com



The Atlantic



WGL
WHOLE GENOME LABORATORY

Subject	Platform	SNPs
USA individual (CV)	Sanger	3,213,400
German individual	Illumina	3,258,774
Estonian individual	SOLiD	3,482,975
USA individual (JW)	454	3,322,090
Irish individual	Illumina	3,125,825

HUMAN GENOME AND ITS IMPACT ON PHENOTYPE
WHAT IS THE MOLECULAR BASIS OF IT?

… TTTACCGAGC …

… AGCATACCGA …

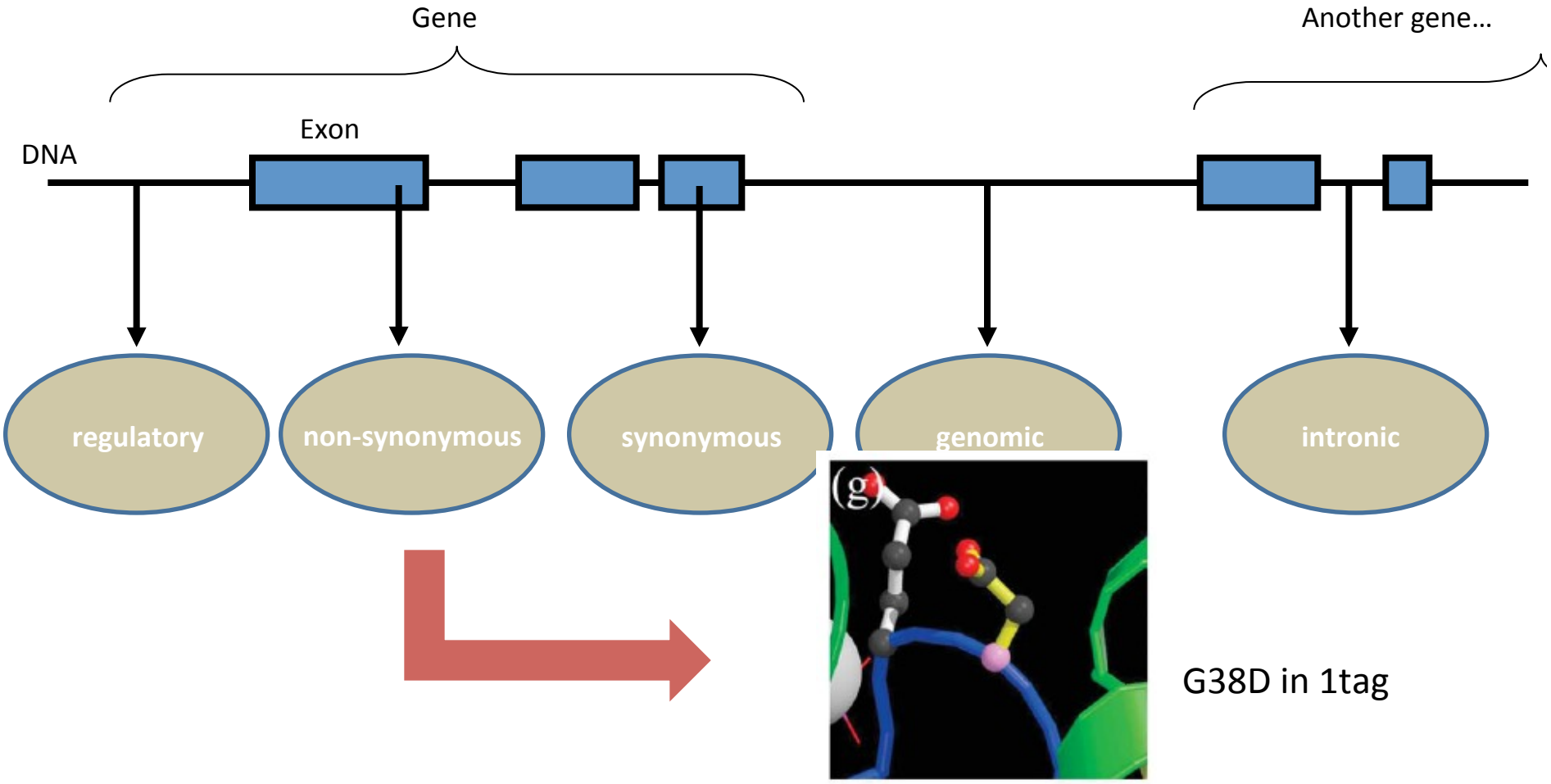
… AGCATAGCGA …

PHENOME

GENOME

Base Changes Resulting in Different Proteins

>40 million known unique sites of variation!



Adapted from: http://snp.ims.u-tokyo.ac.jp/samplesMethods.html#SNP

Yue *et al*. *J Mol Biol*, 353: 459 (2005).

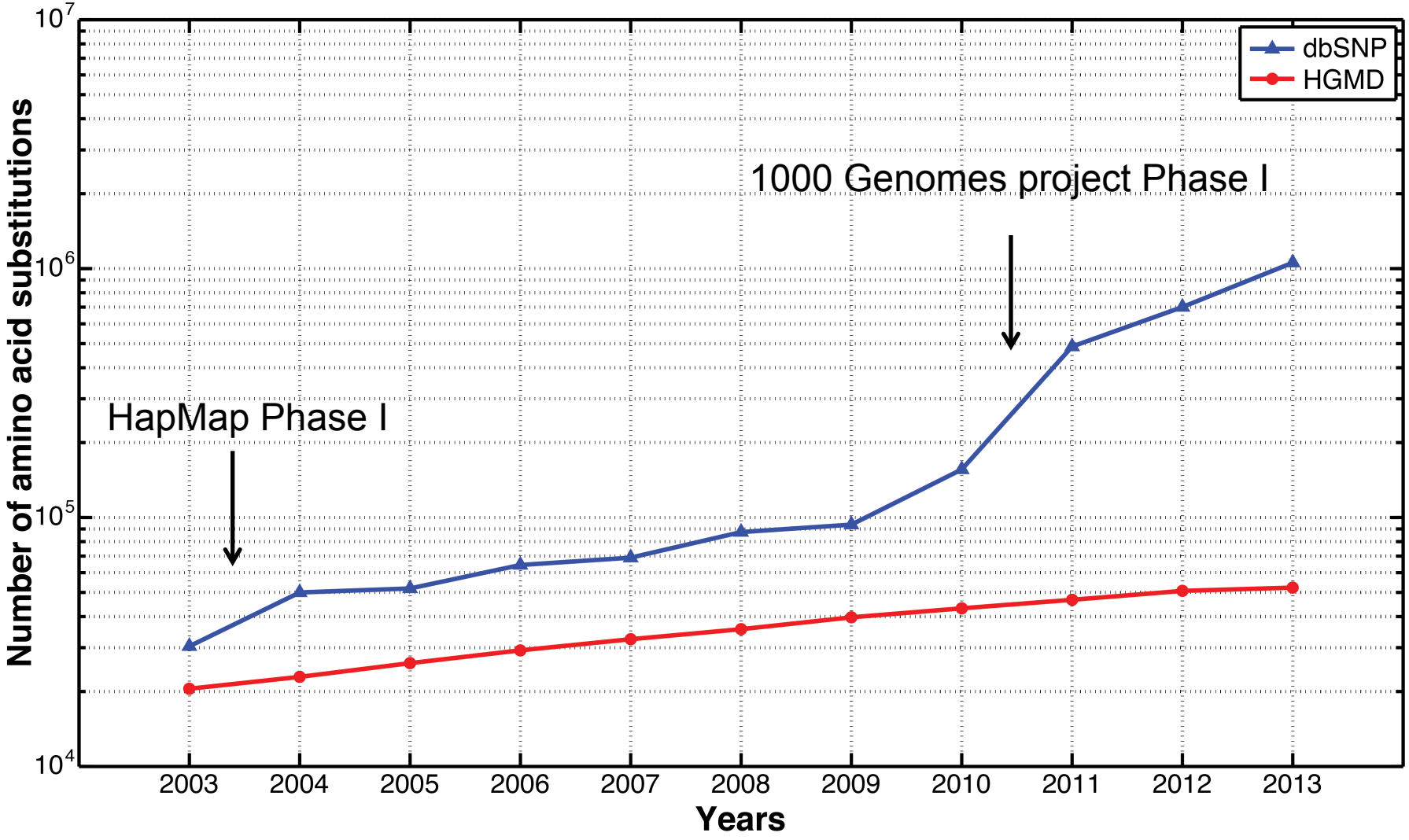
CURRENT TOOLS PREDICT EFFECTS OF VARIANTS

no.	SNP_ID	RefSNP_ID	Allele	Polymorph	MAF	SIFT	Gene	BlinkR	SeqR	LocalR	PolyPhenR	SNP3DR	LS-SNPR	Panther	Pmut_pred
2	th1808	rs1126673	C/T	SNP	0.033	V374I	ADH4	neutral	neutral	neutral	neutral	neutral	neutral		neutral
3	th1819	rs1126671	T/C	SNP	0.000	I309V	ADH4	neutral	neutral	neutral	neutral	neutral	neutral		neutral
4	th1823	NEW	C/A	SNP	0.033	V158L	ADH4	neutral	neutral	neutral	neutral	neutral		neutral	damage
5	th509	rs4646422	G/A	SNP	0.063	G45D	CYP1A1	neutral	neutral	neutral	neutral	damage	neutral	damage	neutral
6	th512	rs2839942	C/G	SNP	0.016	T173R	CYP1A1	neutral	neutral	neutral	neutral	neutral	damage	neutral	damage
7	th514	rs1048943	A/G	SNP	0.234	I462V	CYP1A1	neutral	neutral	neutral	neutral	damage	neutral	neutral	neutral
8	th515	NEW	G/T	SNP	0.016	R511L	CYP1A1	damage	damage	damage	damage	damage			damage
9	th1845	NEW	C/G	SNP	0.031	F21L	CYP1A2	neutral	neutral	neutral	neutral			neutral	neutral
10	th1846	NEW	T/A	SNP	0.016	F125I	CYP1A2	damage	damage	damage	damage			damage	neutral
11	th1847	NEW	A/G	SNP	0.016	M180V	CYP1A2	neutral	damage	neutral	damage				neutral
12	th1850	NEW	G/A	SNP	0.016	G299S	CYP1A2	neutral	neutral	neutral	neutral			neutral	neutral
13	th1861	rs10012	C/G	SNP	0.281	R48G	CYP1B1	neutral	neutral	neutral	neutral	neutral	damage	neutral	neutral
14	th1862	NEW	C/T	SNP	0.016	S112L	CYP1B1	neutral	damage	neutral	neutral	neutral		neutral	neutral
15	th1863	rs1056827	G/T	SNP	0.281	A119S	CYP1B1	neutral	damage	neutral	neutral	neutral	neutral	neutral	neutral
16	th1865	rs1056836	G/C	SNP	0.083	V432L	CYP1B1	neutral	neutral	damage	neutral	damage	neutral	neutral	neutral
17	th1872	rs5031017	C/A	SNP	0.234	G479V	CYP2A6	damage	damage	damage	neutral	damage	damage		damage
18	th1876	rs2839944	A/G	SNP	0.031	S224P	CYP2A6	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
19	th1887	rs3758581	G/A	SNP	0.047	**I331V**	CYP2C19	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
20	th1921	rs2837175	T/C	SNP	0.031	L293P	CYP3A4	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
21	th1926	NEW	C/T	SNP	0.017	R35W	CYP4B1	neutral	neutral	neutral	damage	neutral		damage	damage
22	th1933	rs3215983	AT/__	Deletion	0.234	D294G	CYP4B1	neutral	neutral	neutral	neutral	neutral		neutral	neutral
23	th1934	rs2297810	G/A	SNP	0.234	M331I	CYP4B1	neutral	damage	damage	neutral	damage	neutral	neutral	neutral
24	th1935	rs4646491	C/T	SNP	0.250	R340C	CYP4B1	damage	damage	damage	damage	damage	damage	damage	damage
25	th1937	rs2297809	C/T	SNP	0.229	R375C	CYP4B1	damage	damage	damage	damage	damage	damage	damage	damage
26	th1943	rs3738046	G/C	SNP	0.047	R43T	EPHX1	neutral	neutral	neutral	neutral	neutral	neutral	neutral	damage
27	th1949	rs1051740	T/C	SNP	0.438	Y113H	EPHX1	damage	damage	damage	damage	damage	damage	damage	neutral
28	**th1952**	**NEW**	**C/T**	**SNP**	**0.016**	**H139Y**	**EPHX1**	neutral	neutral	neutral	damage	damage		neutral	neutral
29	**th1953**	**rs2234922**	**A/G**	**SNP**	**0.141**	**H139R**	**EPHX1**	neutral	neutral	neutral	neutral	neutral		neutral	neutral
30	th1961	rs947894	A/G	SNP	0.266	I105V	GSTP1	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
31	th1026	rs1805158	C/T	SNP	0.016	R64W	NAT2	damage	damage	damage	neutral	damage		damage	damage
32	th1028	rs1801280	T/C	SNP	0.094	I114T	NAT2	neutral	damage	neutral	neutral	damage	damage	neutral	neutral
33	th1030	rs1799930	G/A	SNP	0.344	R197Q	NAT2	neutral	damage	neutral	damage	neutral	damage	neutral	damage
34	**th1031**	**rs1208**	**G/A**	**SNP**	**0.156**	**R268K**	**NAT2**	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
35	th1032	rs1799931	G/A	SNP	0.141	G286E	NAT2	neutral	neutral	neutral	neutral	neutral	neutral	neutral	damage
36	th2245	NEW	C/A	SNP	0.071	L71I	NNMT	neutral	neutral	neutral	neutral	neutral		neutral	neutral
37	th2296	rs2837445	A/G	SNP	0.479	F247L	SULT1A1	neutral	neutral	neutral	neutral	damage	neutral	neutral	neutral
38	th2297	rs1801030	C/T	SNP	0.000	V223M	SULT1A1	damage	damage	neutral	neutral	neutral	neutral	neutral	neutral
39	th2313	rs27742	T/C	SNP	0.000	**E282K**	SULT1A2	neutral	neutral	neutral	neutral	neutral	neutral	neutral	damage
40	th2314	NEW	T/A	SNP	0.016	K258N	SULT1A2	damage	damage	damage	neutral	damage		damage	neutral
41	th2316	rs1059491	T/G	SNP	0.125	N235T	SULT1A2	damage	damage	damage	damage	damage	damage	neutral	neutral
42	th2331	rs1703610	T/G	SNP	0.056	S255A	SULT1C1	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
43	**th2239**	**rs1109826**	**A/G**	**SNP**	**0.000**	**M368I**	**UGT8**	neutral	neutral	neutral	neutral	neutral	neutral	damage	neutral

When applied to 43 nsSNPs of 18 drug related genes from the Thai SNP resequencing project there were strong correlations

Slide from Sean Mooney's group.

Growth of Data



ACMG STANDARDS AND GUIDELINES | Genetics inMedicine

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

Disclaimer: These ACMG Standards and Guidelines were developed primarily as an educational resource for clinical laboratory geneticists to help them provide quality clinical laboratory services. Adherence to these standards and guidelines is voluntary and does not necessarily assure a successful medical outcome. These Standards and Guidelines should not be considered inclusive of all proper procedures and tests or exclusive of other procedures and tests that are reasonably directed to obtaining the same results. In determining the propriety of any specific procedure or test, the clinical laboratory geneticist should apply his or

Table 2 In silico predictive algorithms

Category	Name	Website	Basis
Missense prediction	ConSurf	http://consurftest.tau.ac.il	Evolutionary conservation
	FATHMM	http://fathmm.biocompute.org.uk	Evolutionary conservation
	MutationAssessor	http://mutationassessor.org	Evolutionary conservation
	PANTHER	http://www.pantherdb.org/tools/csnpScoreForm.jsp	Evolutionary conservation
	PhD-SNP	http://snps.biofold.org/phd-snp/phd-snp.html	Evolutionary conservation
	SIFT	http://sift.jcvi.org	Evolutionary conservation
	SNPs&GO	http://snps-and-go.biocomp.unibo.it/snps-and-go	Protein structure/function
	Align GVGD	http://agvgd.iarc.fr/agvgd_input.php	Protein structure/function and evolutionary conservation
	MAPP	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	Protein structure/function and evolutionary conservation
	MutationTaster	http://www.mutationtaster.org	Protein structure/function and evolutionary conservation
	MutPred	http://mutpred.mutdb.org	Protein structure/function and evolutionary conservation
	PolyPhen-2	http://genetics.bwh.harvard.edu/pph2	Protein structure/function and evolutionary conservation
	PROVEAN	http://provean.jcvi.org/index.php	Alignment and measurement of similarity between variant sequence and protein sequence homolog
	nsSNPAnalyzer	http://snpanalyzer.uthsc.edu	Multiple sequence alignment and protein structure analysis
	Condel	http://bg.upf.edu/fannsdb/	Combines SIFT, PolyPhen-2, and MutationAssessor
	CADD	http://cadd.gs.washington.edu	Contrasts annotations of fixed/nearly fixed derived alleles in humans with simulated variants

TUMOR BOARD

Person: 68 year old woman
Cancer type: colon cancer, metastatic
Mutations:

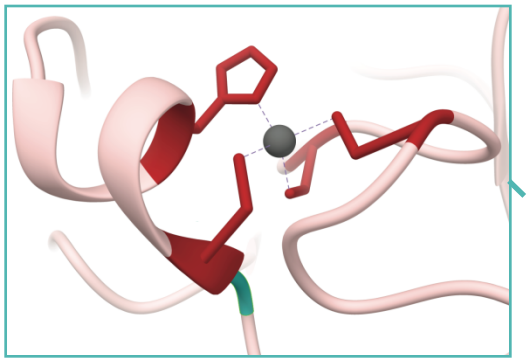
KRAS, C27F
BRCA1, H57R
TP53, T98*

Treatment options:

- clinical trial at MD Anderson
- continue with chemotherapy
- treat with new drug for breast cancer

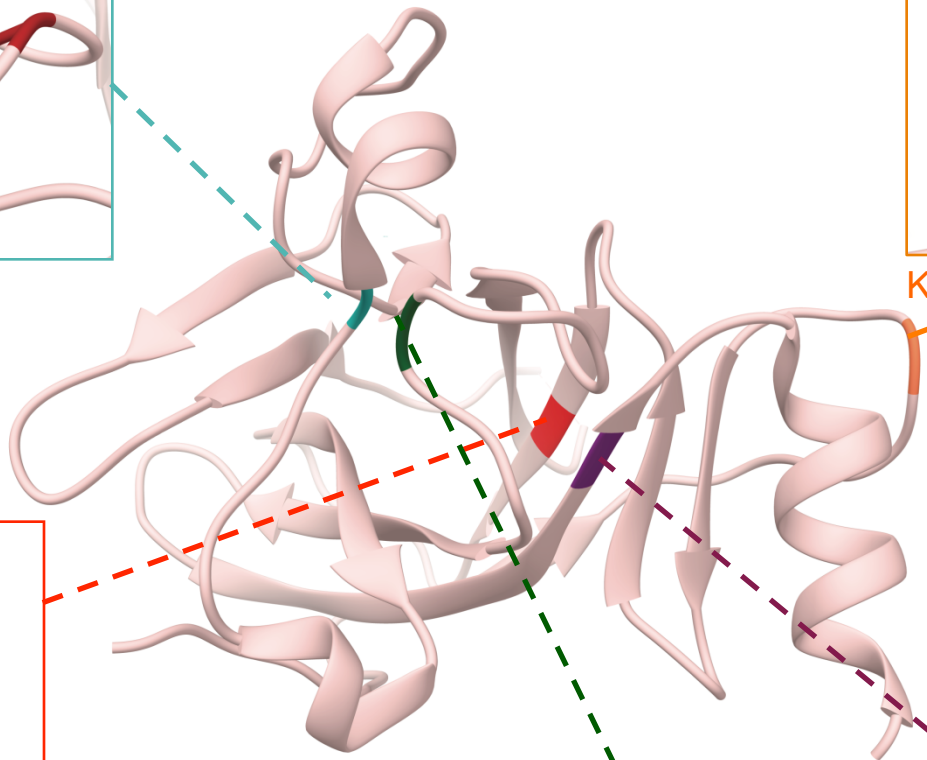
http://www.med.umich.edu/cancer/images/urologic-oncology-tumor-board.jpg

Molecular Consequences on P53



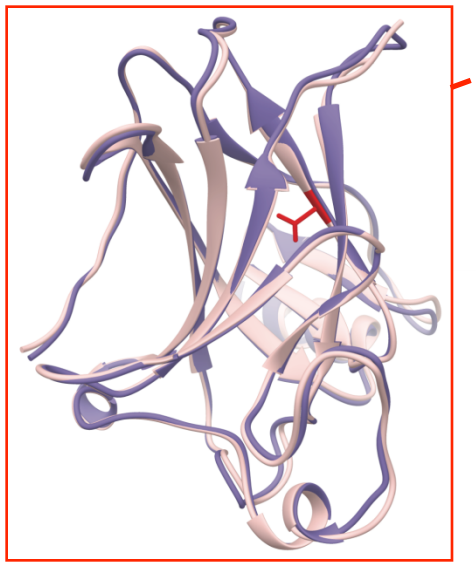
R175H: Metal-binding

K120R: Acetylation

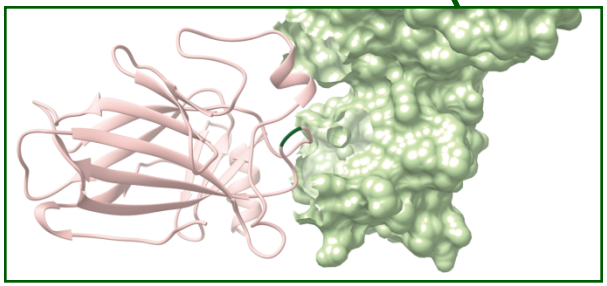


p53 – tumor suppressor protein

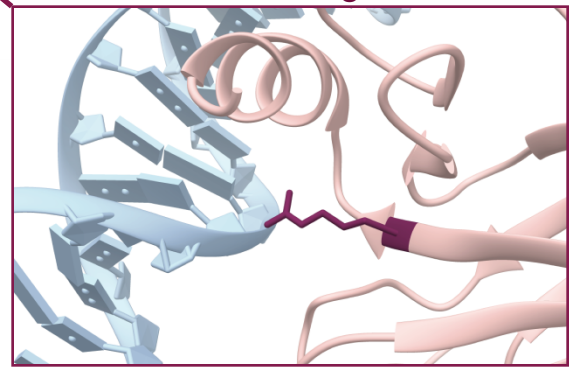
V143A: Stability



G245S: Protein-binding

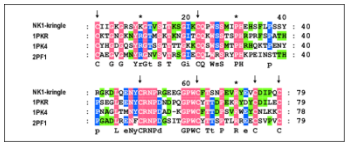


R273H: DNA-binding



PDB structures: 2ybg, 2j1w, 1ycs and 1tup

MUTPRED 2.0



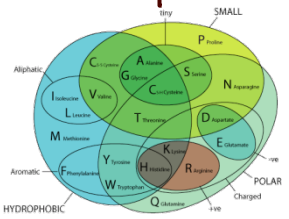
```
>sp|P04637|P53_HUMAN
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPRVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
```

$$P(loss\ of\ p\ at\ s_i|xjy) = P(s_i = s_i^p|s) \cdot \left(1 - P(s_i = s_i^p|s_{xjy})\right)$$

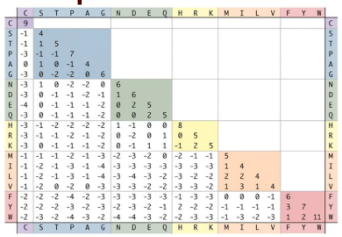
$$P(loss\ of\ p\ at\ s_i|xjy) = P(s_i = s_i^p|s) \cdot \left(1 - P(s_i = s_i^p|s_{xjy})\right)$$

Conservation | **Sequence** | **Struct. Prop.** | **Funct. Prop.**

Neural networks



Physicochemical



Substitution matrices

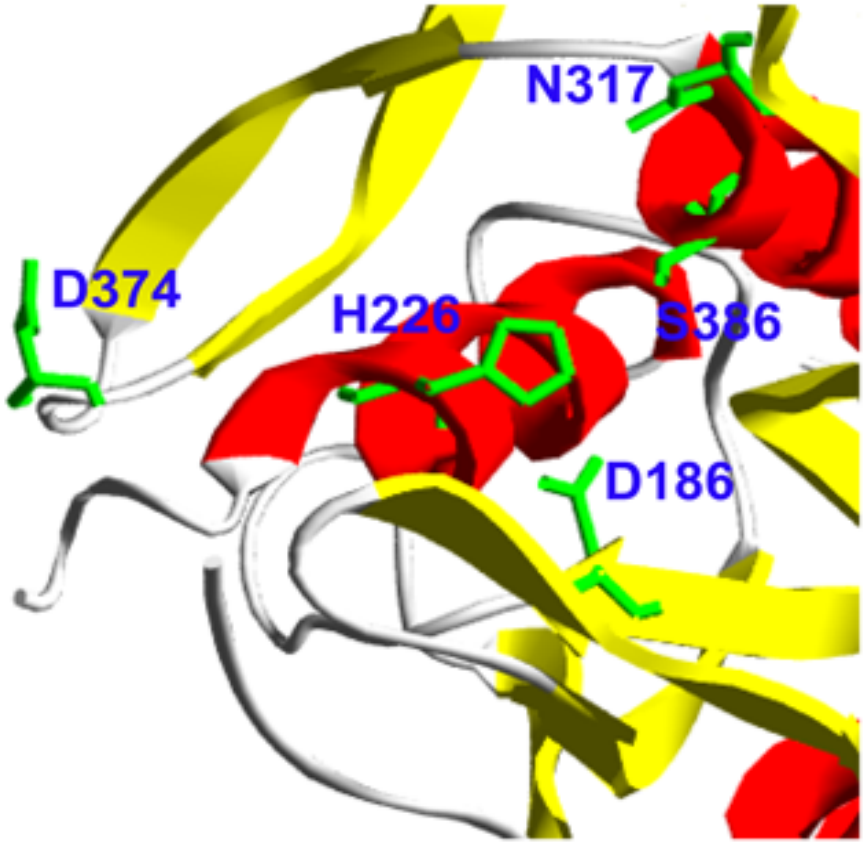
Neural network ensemble:

- Z-score normalized and PCA'd
- 30 feed-forward networks
- bootstrap aggregating, balanced training
- trained using resilient propagation

Gain of Catalytic Activity Causes Disease

- a member of the proteinase K sub-family of subtilases that reduces the number of LDL receptors in liver through a posttranscriptional mechanism.
- D374Y leads to a 10-fold increase in catalytic activity that causes hypercholesterolemia.

Catalytic residues of PCSK9 (2qtw)



Lagace *et al. J Clin Invest*, 116: 2995 (2006).

Xin *et al. Bioinformatics* 26: 1975-1982 (2010).