SOC Summer School 2017

A logical introduction to computational biology

Limsoon Wong

National University of Singapore

About Part III

Let us now work on a couple of simple projects for you to practice using logical analysis to solve problems in computational biology…

Part III

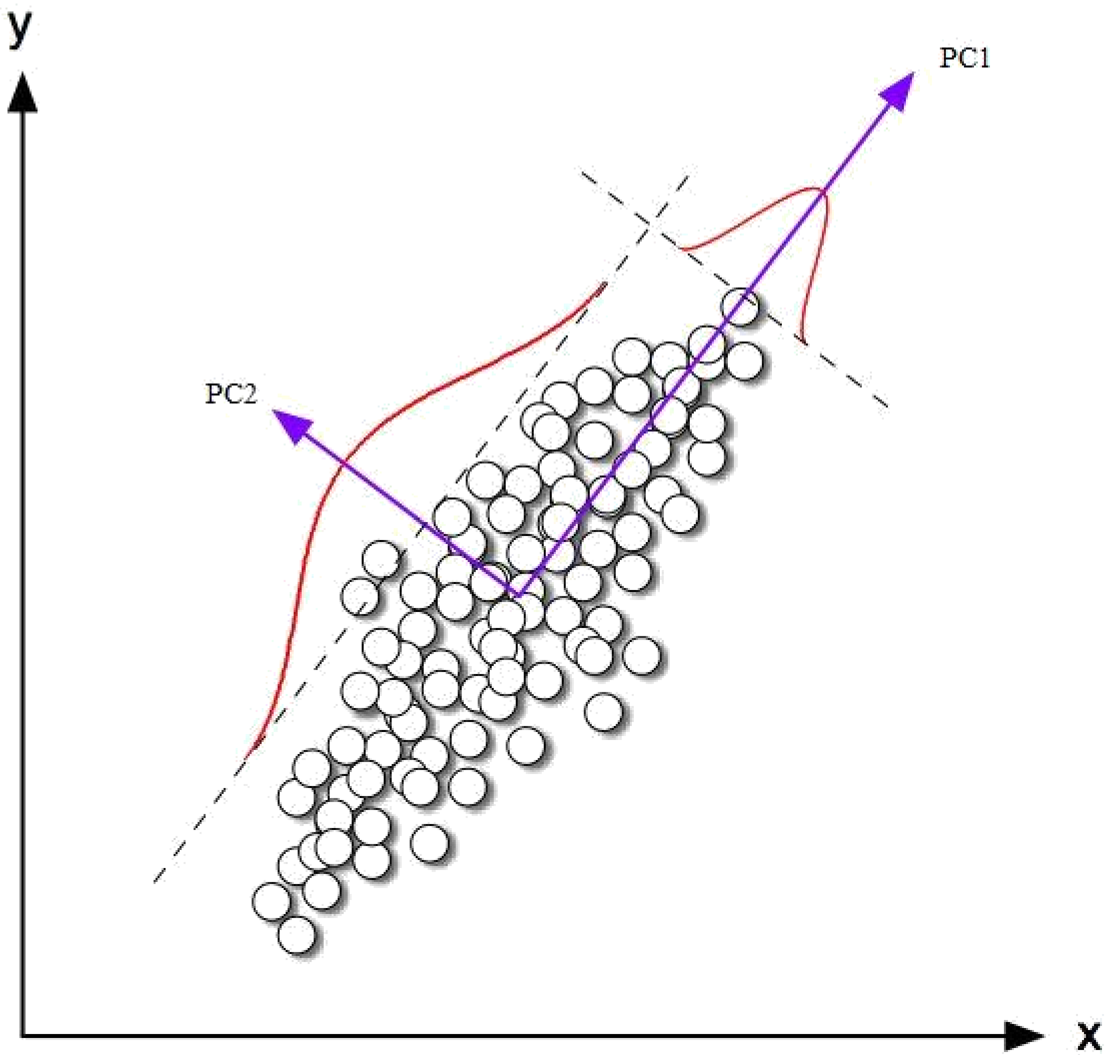
A couple of projects to round things up

NUS National University of Singapore

DISCOVERING INVARIANTS

We have talked quite a bit about logical analysis and exploitation of invariants

Let us now learn one simple way for discovering invariants



Principal component analysis (PCA)

Credit: Alessandro Giuliani

PCA, a la Pearson (1901)

)(98)(

SULLE FUNZIONI BILINEARI

DI

E. BELTRAMI

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London**.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y=a_0+a_1x, \quad \text{or} \quad z=a_0+a_1x+b_1y,$$

$$\text{or} \quad z=a_0+a_1x_1+a_2x_2+a_3x_3+\ldots+a_nx_n,$$

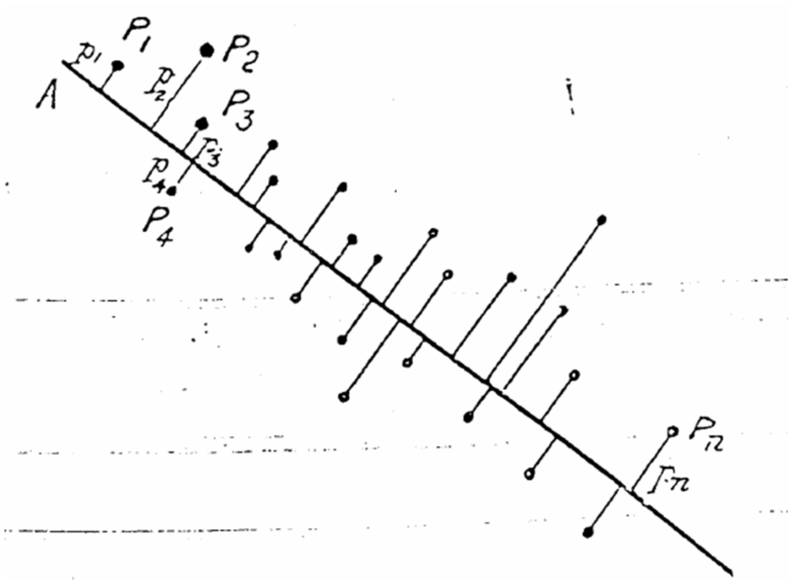
where $y, x, z, x_1, x_2, \ldots x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \ldots a_n$

For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

$$U=S(p^2)=\text{a minimum.}$$

If y were the dependent variable, we should have made

$$S(y'-y)^2=\text{a minimum}$$



Credit: Alessandro Giuliani

Growth, 1960, **24**, 339-354.

SIZE AND SHAPE VARIATION IN THE PAINTED TURTLE.[1] A PRINCIPAL COMPONENT ANALYSIS

PIERRE JOLICOEUR AND JAMES E. MOSIMANN[2]

Walker Museum, University of Chicago
and
Institut de Biologie, Université de Montréal

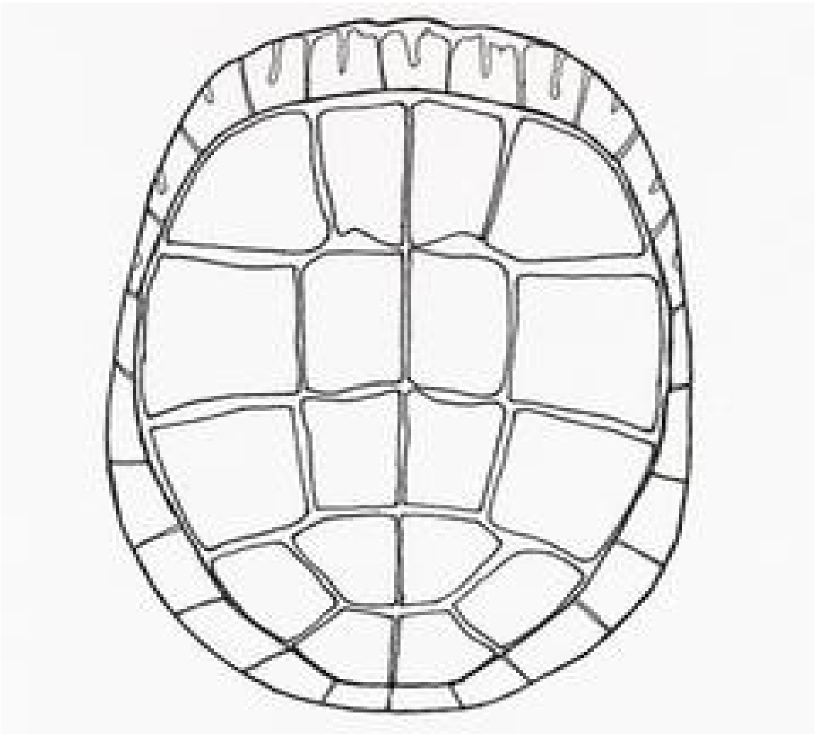
(Received for publication July 11, 1960)

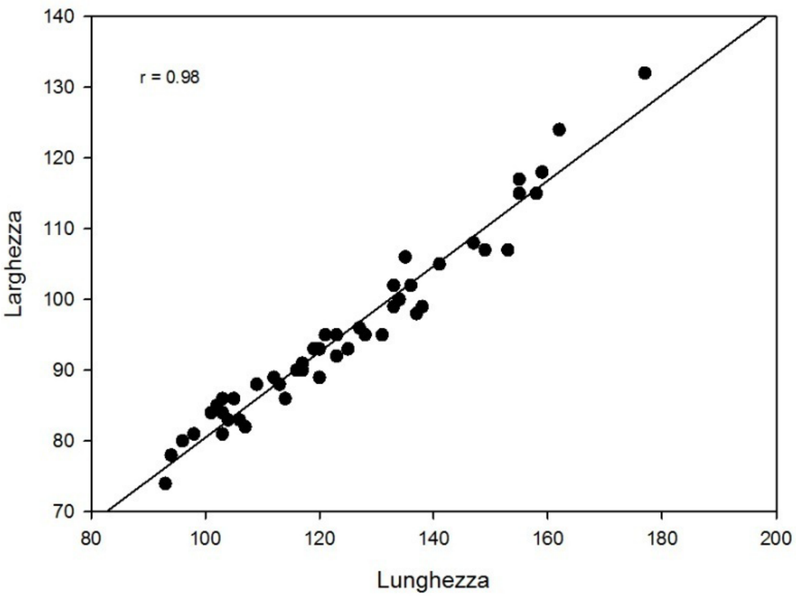
Credit: Alessandro Giuliani

TABLE 1
CARAPACE DIMENSIONS OF PAINTED TURTLES (*Chrysemys picta marginata*) IN MM.

24 Males			24 Females		
length	width	height	length	width	height
93	74	37	98	81	38
94	78	35	103	84	38
96	80	35	103	86	42
101	84	39	105	86	40
102	85	38	109	88	44
103	81	37	123	92	50
104	83	39	123	95	46
106	83	39	133	99	51
107	82	38	133	102	51
112	89	40	133	102	51
113	88	40	134	100	48
114	86	40	136	102	49
116	90	43	137	98	51
117	90	41	138	99	51
117	91	41	141	105	53
119	93	41	147	108	57
120	89	40	149	107	55
120	93	44	153	107	56
121	95	42	155	115	63
125	93	45	155	117	60
127	96	45	158	115	62
128	95	45	159	118	63
131	95	46	162	124	61
135	106	47	177	132	67

Credit: Alessandro Giuliani





Pearson Correlation Coefficients,

	length	width	height
length	1.00000	0.97831	0.96469
width	0.97831	1.00000	0.96057
height	0.96469	0.96057	1.00000

Width = 19,94 + 0,605*Length

Credit: Alessandro Giuliani

Credit: Alessandro Giuliani

Interesting info are often in the 2nd principal component

	PC1 (98%)	PC2 (1.4%)
Length	0,992	-0,067
Width	0,990	-0,100
Height	0,986	0,168

PC1= 33.78*Length +33.73*Width + 33.57*Height

PC2 = -1.57*Length – 2.33*Width + 3.93*Height

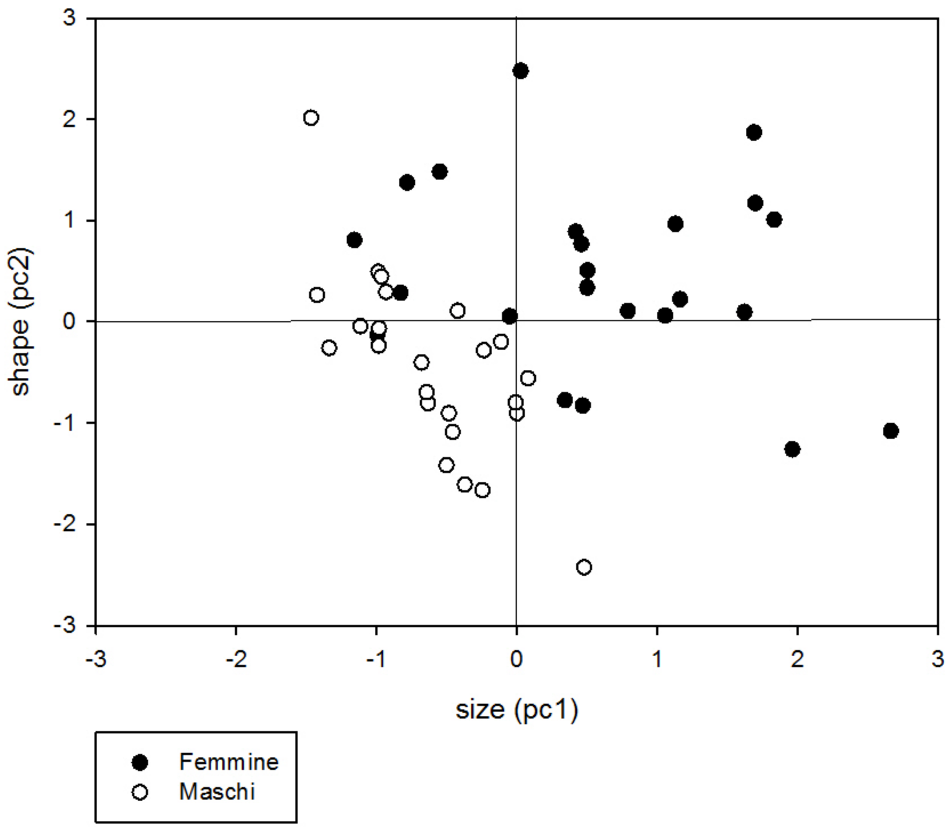
- Presence of an overwhelming size component explaining system variance comes from the presence of a ‘typical’ common shape
- Displacement along pc1 = size variation (all positive terms)
- Displacement along pc2 = shape deformation (both positive and negative terms)

unit	sex	Length	Width	Height	PC1(size)	PC2(shape)
T25	F	98	81	38	-1,15774	0,80754832
T26	F	103	84	38	-0,99544	-0,1285916
T27	F	103	86	42	-0,7822	1,37433475
T28	F	105	86	40	-0,82922	0,28526912
T29	F	109	88	44	-0,55001	1,4815252
T30	F	123	92	50	0,027368	2,47830153
T31	F	123	95	46	-0,05281	0,05403839
T32	F	133	99	51	0,418589	0,88961967
T33	F	133	102	51	0,498425	0,33681756
T34	F	133	102	51	0,498425	0,33681756
T35	F	134	100	48	0,341684	-0,774911
T36	F	136	102	49	0,467898	-0,8289156
T37	F	137	98	51	0,457949	0,76721682
T38	F	138	99	51	0,501055	0,50628189
T39	F	141	105	53	0,790215	0,10640554
T40	F	147	108	57	1,129025	0,96505915
T41	F	149	107	55	1,055392	0,06026089
T42	F	153	107	56	1,161368	0,22145593
T43	F	155	115	63	1,687277	1,86903869
T44	F	158	115	62	1,696753	1,17117077
T45	F	159	118	63	1,833086	1,00956637
T46	F	162	124	61	1,962232	-1,261771
T47	F	177	132	67	2,662548	-1,0787317
T48	F	155	117	60	1,620491	0,09690818
T1	M	93	74	37	-1,46649	2,01289241
T2	M	94	78	35	-1,42356	0,26342486
T3	M	96	80	35	-1,33735	-0,258445
T4	M	101	84	39	-0,98842	0,49260881
T5	M	102	85	38	-0,98532	-0,2361914
T6	M	103	81	37	-1,11528	-0,0436547
T7	M	104	83	39	-0,96555	0,44687352
T8	M	106	83	39	-0,93257	0,29353841
T9	M	107	82	38	-0,98269	-0,066727
T10	M	112	89	40	-0,63393	-0,8042059
T11	M	113	88	40	-0,64405	-0,6966061
T12	M	114	86	40	-0,68078	-0,4047389
T13	M	116	90	43	-0,42133	0,10845233
T14	M	117	90	41	-0,48485	-0,9039457
T15	M	117	91	41	-0,45824	-1,0882131
T16	M	119	93	41	-0,37202	-1,610083
T17	M	120	89	40	-0,50198	-1,4175463
T18	M	120	93	44	-0,23552	-0,2831547
T19	M	121	95	42	-0,24581	-1,6640875
T20	M	125	93	45	-0,11305	-0,1986272
T21	M	127	96	45	-0,00023	-0,9047645
T22	M	128	95	45	-0,01035	-0,7971646
T23	M	131	95	46	0,079136	-0,559302
T24	M	135	106	47	0,477846	-2,4250481

Credit: Alessandro Giuliani

National University of Singapore

Female turtles are larger and have more exaggerated height ☺



Exercise

- **Madrid and Warsaw are at almost the same distance to Latium cities**

  Are Madrid and Warsaw near each other?

Distances of European cities (km) from the main cities of Latium

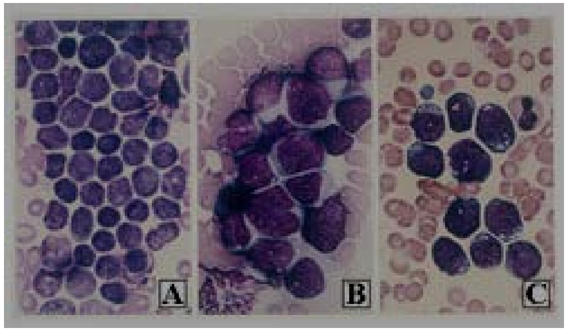
	Rome	Latina	Frosinone	Viterbo	Rieti
Amsterdam	430	447	449	415	409
Athens	347	321	331	346	364
Barcelona	283	305	293	292	271
Beograd	227	222	236	220	238
Berlin	393	400	409	374	373
Bern	227	249	247	220	205
Bonn	353	370	372	339	330
Bruselles	388	406	406	371	365
Bucharest	364	355	368	359	378
Budapest	268	261	274	246	259
Calais	418	448	446	418	405
Copenhagen	510	522	527	492	491
Dublin	622	645	641	615	600
Edinburgh	637	655	655	625	615
Frankfurt	318	333	336	302	295
Hamburg	435	448	453	417	414
Helsinki	727	729	739	706	713
Istanbul	452	430	443	443	464
Lisbon	615	637	622	624	604
London	474	494	493	464	456
Luxembourg	325	346	346	315	307
Madrid	449	470	458	460	440
Marseille	200	223	213	202	183
Moscow	782	773	785	759	774
Munich	230	245	250	216	213
Oslo	664	675	682	646	645
Paris	365	386	383	357	343
Prague	305	313	320	286	290
Sofia	294	273	286	280	301
Stockholm	653	658	668	632	636
Warsaw	435	433	444	413	421
Vienna	255	254	265	233	240
Zurich	227	246	246	214	205

Giuliani et al., Physics Letters A, 247:47-52, 1998

Exercise

- **Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse

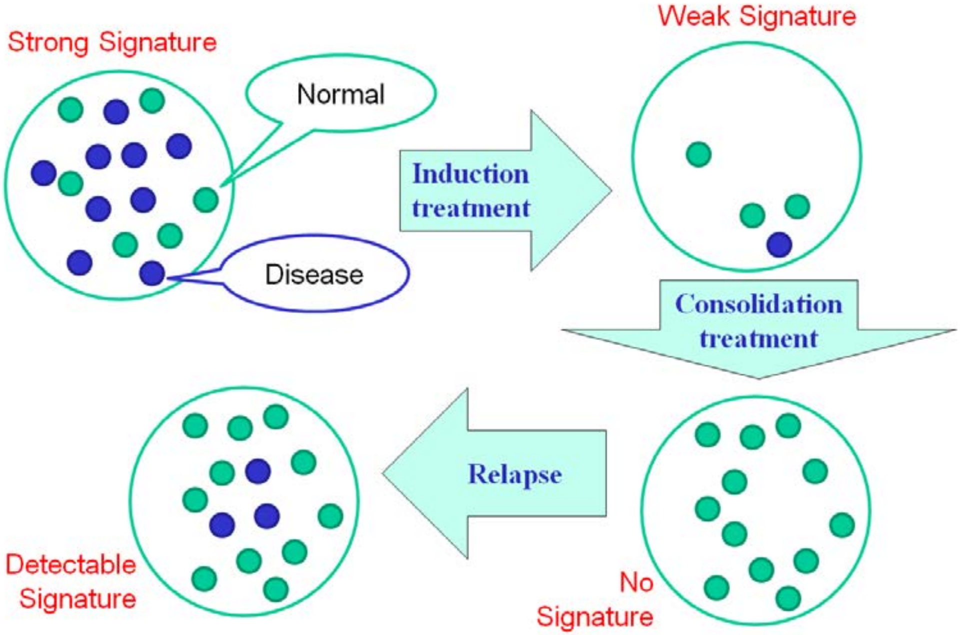
- **The subtypes look similar**



- **Can we diagnosis the subtypes based on gene expression profiling?**

Exercise

- **Treatment gradually removes leukemic cells in patient**



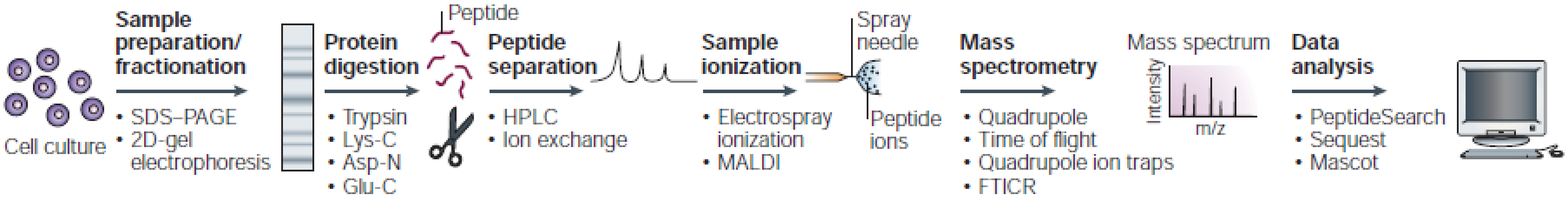
- **Diagnostic GEP captures leukemic subtype signature**

- **Hypothesis: Poor genetic response suggests high risk of relapse**

- **Suggest a leukemia relapse prediction model based on gene expression profiling**

MISSING PROTEINS IN PROTEOMIC PROFILES

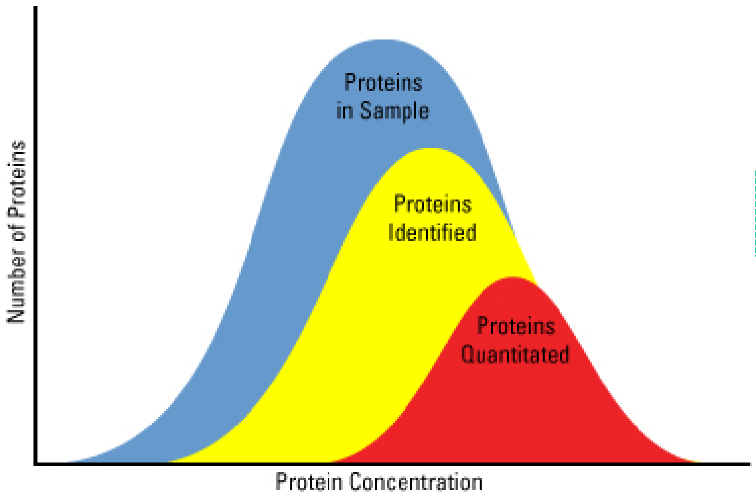
Proteomics profiling



- **Proteomic profile**
  - Which protein is found in the sample
  - How abundant it is
- **Difficulties**
  - Complexity: 20k genes vs 500k proteins
  - Dynamic range: > 10 orders of magnitude in plasma. Proteins cannot be amplified

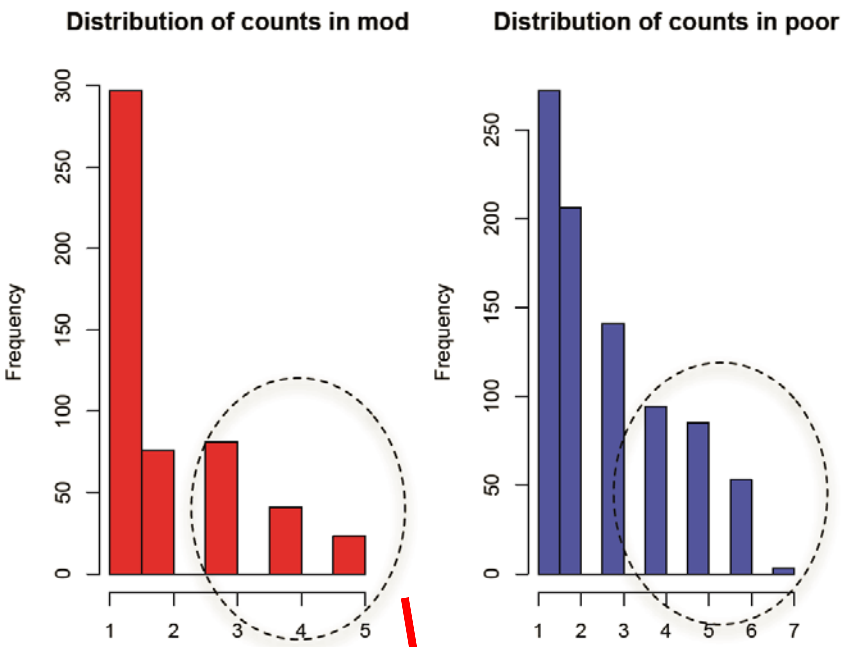
Issues in proteomics: Coverage and consistency

Technical incompleteness



Per sample

How it affects real data



Only 25 out of 800+ proteins are common to all 5 mod-stage HCC patients!

Guo et al. *Nature Medicine*, *21*, 407, 2015

Lots of missing values in real proteomics datasets

nm.3807-S4.xls [Read-Only] [Compatibility Mode] - Microsoft Excel

X30 | NA

	protein	GeneSymbol	kidneyTisue1	kidneyTissue2	kidneyTissue3	kidneyTissue4	kidneyTissue5	kidneyTissue6	kidneyTissue7	kidneyTissue8	kidneyTissue9	kidneyTissue10	kidneyTissue11	kidneyTissue12	kidneyTissue13	kidneyTissue14	kidneyTissue15	kidneyTissue16	kidneyTissue17	kidneyTissue18	kidneyTissue19	kidneyTissue20	kidneyTissue21	kidneyTissue22	kidneyTissue23	kidneyTissue24	kidneyTissue25	kidneyTissue26	kidneyTissue27
2	P09110	ACAA1	288001.7778	46353.28	237958.5	30102.47	297711.2	37098.09	67454.84	92200.62	231528.4	12617.18	263299.1	NA	222387.2	NA	177211	27857.94	84689.84	43497.89	280540.3	77962.17	235242.5	23827.06	302761.4	41190.07	2064.747	97756.44	122386.3
3	P05166	PCCB	246687.75	70504.27	253890.9	NA	314250.1	33680.65	108554.7	321442.7	260389.5	183399.7	258247.1	139288.5	284934.5	115138	245595.9	30488.41	221565	280540.3	240054.8	65477.99	250479.3	NA	327799	41974.24	125103	321442.7	175808.5
4	Q96RP9	GFM1	37872.59722	NA	40359.89	NA	73975.35	NA	64601.65	56815.28	34506.99	35176.2	98642.34	23060.3	91995.3	NA	37735.48	33491.8	48208.46	47858.24	39584.44	NA	67976.03	23631.74	46763.48	NA	2064.747	53619.99	67555.47
5	Q15417	CNN3	28364.89722	NA	NA	NA	NA	44156.47	52272.02	27128.03	10577.49	32524.27	14171.12	33388.93	27593.38	49821.32	23144.21	24964.95	32403	NA	24907.94	46053.92	NA	NA	25129.86	42948.4	2064.747	26438.35	23207.51
6	Q96FQ6	S100A16	NA	35176.2	NA	66058.39	NA	30674.6	1804.538	21706.65	NA	NA	11359.64	NA	18677.58	41493.97	12617.18	22496.77	NA	NA	NA	36422.79	NA	75858.83	20589.93	31161.06	2064.747	20398.13	NA
7	P62820	RAB1A	NA	NA	NA	NA	NA	NA	54417.16	3130.811	NA	68503.39	NA	NA	NA	NA	NA	NA	NA	NA	32596.28	NA	NA	54839	NA	48748.28	2064.747	NA	NA
8	P27169	PON1	NA	47101.83	58436.31	18128.35	NA	33573.36	112930.6	NA	NA	NA	NA	59432.1	NA	39084.55	36282.92	16953.34	NA	NA	NA	45107.13	NA	19506.67	NA	38130.55	109838.9	NA	NA
9	Q9UL46	PSME2	33680.65278	99968.93	59047.33	145114.2	33256.26	141575.7	77962.17	75727.38	64365.04	121022.2	40286.83	114480.8	40567.01	104458.4	42876.78	83666.14	55954.92	62742.03	33768.27	111940.8	59915.42	151558.9	38443.16	113145.5	79024.33	73747.38	40140.37
10	P08237	PFKM	39644.09722	NA	54240.61	NA	136064	NA	1804.538	62845.97	141296.3	100616.3	137596.7	NA	140860.9	NA	96590.73	NA	92823.65	51085.24	155550.8	NA	47697.29	NA	136064	NA	2064.747	58618.05	143381.1
11	P04040	CAT	292456.0528	149632.6	239229.2	24964.95	258247.1	220764.4	540115.8	133921.9	284934.5	367784.7	293727.3	179981.9	259314.6	124294.3	204722.1	77070.33	109006.7	136875.9	290924.4	163095.2	237958.5	31389.75	271920.4	227900.3	499422.8	150524.5	294964.3
12	Q8WYA6	CTNNBL1	NA	NA	NA	NA	NA	NA	1804.538	NA	NA	NA	NA	NA	NA	NA	NA	27646.1	37621.73	26686.24	NA	NA	NA	NA	NA	NA	2064.747	NA	NA
13	Q9H0W9	C11orf54	454591.5833	77225.75	393512.7	55431.72	365975.5	180535.1	188742.5	77348.17	352898.9	119242.7	417999.9	263299.1	474797	229655.9	427428	143697	124568	146454.4	441856.5	74156.41	370040.5	44605.86	363784.6	187566.8	129074.8	104101.6	375463.4
14	P31948	STIP1	76018.00556	83236.9	83516.5	137596.7	75613.89	110367.2	98642.34	195146	77709.53	282315.9	65948.94	122386.3	81635.42	129969.2	67749.81	124568	108554.7	135737.2	69039.96	92656.4	85600.47	147792.9	65262.99	109273.7	91127.04	218888	122047.2
15	O94901	SUN1	57623.33889	NA	NA	NA	72273.86	NA	1804.538	NA	NA	NA	58063.49	NA	NA	NA	NA	NA	NA	NA	60013.66	NA	NA	NA	71252.19	NA	2064.747	NA	NA
16	Q99714	HSD17B10	175372.7444	114480.8	181096.8	75400.28	222387.2	91466.47	218888	269679.7	179177.4	165285.9	202618.2	117389.5	191537	41135.21	196208.5	151044.7	210269.6	294964.3	183893	82644.38	179981.9	102286.8	233372.9	91325.89	196996.8	293727.3	174540.8
17	Q15833	STXBP2	14224.84722	24264.99	14303.05	19690.86	16316.33	NA	1804.538	NA	14303.05	17309.98	11459.84	14224.85	12617.18	NA	14224.85	9837.458	21131.38	5634.228	13283.71	28846.59	20057.06	12924.71	17380.49	NA	2064.747	11880.63	13166.66
18	P08195	SLC3A2	50797.625	42825.82	63302.14	26628.24	85345.18	NA	1804.538	NA	77850.57	NA	100616.3	NA	76579.02	NA	44010.16	17146.31	NA	NA	80199.58	41362.6	72273.86	32198.97	75858.83	NA	2064.747	NA	76292.57
19	P26038	MSN	333342.6833	438752.3	421056.2	381249.5	241992.3	404349.8	164343.5	172028.6	446678.9	167923.7	367784.7	310472.5	404349.8	393512.7	292456.1	427428	390317.5	244865.7	273261.7	446678.9	404349.8	306071.8	222387.2	423963.5	191537	182241.6	441856.5
20	P09104	ENO2	NA	144058.2	NA	184650.5	NA	137596.7	126146.3	21831.56	NA	NA	NA	119650.8	NA	404349.8	NA	48438.29	57080.76	NA	NA	151558.9	NA	181096.8	NA	123793.9	2064.747	NA	NA
21	P07148	FABP1	1219163.714	34579.48	861796.3	NA	940142	NA	1804.538	NA	1130692	NA	1057986	NA	789446.1	NA	221565	NA	NA	NA	1162786	32336.43	805128.4	NA	970053.3	NA	2064.747	NA	1300718
22	Q96Q11	TRNT1	NA	NA	NA	NA	NA	NA	1804.538	NA	NA	NA	NA	NA	NA	NA	NA	NA	37098.09	35565.03	NA	NA	NA	NA	NA	NA	2064.747	NA	NA
23	O15083	ERC2	NA	NA	NA	85740.42	NA	NA	1804.538	NA	83390.33	NA	NA	NA	NA	NA	NA	142306.8	NA	NA	NA	NA	NA	72396.48	NA	NA	2064.747	NA	70213.43
24	Q15911	ZFHX3	NA	NA	178745.3	393512.7	205865.1	682653.9	1804.538	NA	243050.1	NA	189860.5	NA	NA	NA	NA	457756.2	NA	NA	NA	NA	NA	NA	NA	NA	2064.747	NA	252846.2
25	Q9BUR5	APOO	35479.70278	NA	27260.11	15459.06	40140.37	NA	1804.538	46154.89	30730.15	54737.36	47185.33	13642.38	28517.17	NA	40140.37	NA	NA	10649.17	34436.2	NA	36956.08	16653.18	47858.24	NA	2064.747	33003.64	20057.06
26	Q9UJ83	HACL1	417999.9306	NA	435248.4	NA	336790.8	227161.7	1804.538	174111.8	276628.6	NA	274264.6	NA	317227.1	271920.4	336790.8	NA	NA	372485.6	446678.9	NA	390317.5	NA	307205	211073.8	2064.747	169817.6	333342.7
27	Q8WUM4	PDCD6IP	50008.50556	34991.44	70504.27	50108.55	59047.33	41611.18	84319.78	97140.59	56715.96	134561.7	52110.31	61553.77	67555.47	65262.99	68597.03	59827.38	73200.35	75049.44	64108.37	40359.89	70903.29	49636.31	49821.32	37258.59	76579.02	76685.11	37386.23
28	P53597	SUCLG1	387432.1583	99433.59	228946.3	94932.09	310472.5	150524.5	187002.3	299487.5	275420.7	308775.7	299487.5	101732.7	245595.9	108554.7	270810.9	89524.72	192915.6	276628.6	357417.6	96737.9	205171.6	95793.82	288001.8	162300.5	193664.8	299487.5	245595.9
29	O00186	STXBP3	NA	28468.21	NA	NA	NA	19019.68	1804.538	NA	NA	NA	NA	21949.83	NA	NA	NA	NA	NA	NA	15575.29	29005.53	NA	NA	NA	NA	2064.747	NA	NA
30	Q8N335	GPD1L	52415.71111	NA	59328.51	NA	54240.61	21949.83	109838.9	91466.47	45427.61	109273.7	50443.03	NA	52700.48	22321.01	45502.32	NA	57623.34	41362.6	54737.36	NA	62380.69	NA	54839	23827.06	152627.3	71658.52	49636.31
31	P08621	SNRNP70	48594.65	51791.05	47269.07	86082.28	44306.32	53026.19	1804.538	NA	59432.1	54839	49636.31	60605.33	52477.21	NA	NA	72977.35	74546.25	82242.07	33003.64	60605.33	49636.31	93224.91	NA	56917.54	2064.747	NA	50797.63
32	Q969V6	MKL1	NA	91325.89	55954.92	NA	74269.09	80102.57	1804.538	NA	71906.43	NA	NA	152627.3	72497.5	72497.5	89662.88	51690.71	68707.95	41576.85	72021.55	92973.8	NA	NA	NA	88904.66	2064.747	NA	NA
33	P08311	CTSG	NA	NA	46154.89	NA	NA	67879.78	1804.538	NA	53026.19	NA	NA	68927.99	NA	NA	NA	NA	218057.1	78414.15	NA	NA	46895.88	NA	NA	56514.53	66379.24	NA	NA
34	Q9UKU7	ACAD8	46053.91944	31797.32	50179.16	NA	46601.65	NA	75160.02	49228.15	44010.16	28070.84	41974.24	NA	41840.21	NA	42678.39	NA	24335.52	32270.84	46053.92	NA	49467.07	NA	61900.08	NA	2064.747	46053.92	44605.86
35	Q86X76	NIT1	75613.88611	NA	61068.98	63988.55	80199.58	69590.71	1804.538	55745.15	70389.43	NA	84009.8	75506.47	78547.77	84980.21	76153.19	NA	57523.94	40935.27	70713.02	NA	59540.84	70713.02	78753.85	73278.36	55745.15	58932	52415.71
36	P05162	LGALS2	33491.8	NA	35565.03	NA	52415.71	36825.06	1804.538	23560.07	18592.77	NA	36763.92	72761.18	35479.7	50008.51	24907.94	NA	16653.18	22730.31	34916.06	NA	30730.15	NA	32815.68	71139.86	2064.747	NA	25737.06
37	P23946	CMA1	NA	NA	NA	NA	NA	NA	1804.538	NA	NA	NA	NA	NA	NA	NA	NA	NA	61155.07	14049.16	NA	NA	NA	NA	NA	NA	53240.82	NA	NA
38	P01834	IGKC	462133.8694	885197.1	692332.5	484624	296507.9	462133.9	1219164	319228.4	659554.4	351190.2	312295.6	524995.4	566103.9	692332.5	325019.6	494067.2	286640.3	263299.1	499422.8	1130692	706520.3	469971.2	322906.2	438752.3	913960	310472.5	643593
39	P14868	DARS	12567.36389	110112	54554.37	136875.9	30209.1	121022.2	1804.538	114195.5	43350.86	95493.71	29430.84	182241.6	61667.11	201171.9	81193.99	247871.5	161420	94484.9	76929.26	114678.3	54839	177772	50108.55	141996.6	2064.747	95951.08	53026.19
40	Q9H773	DCTPP1	NA	NA	NA	NA	NA	NA	1804.538	46303.49	NA	11589.48	NA	27509.79	NA	NA	NA	26314.17	87070.11	74656.39	NA	NA	NA	NA	NA	NA	2064.747	22251.11	NA

sTable3 | legend

Webb-Robertson, *JPR*, 14(5):1993-2001, 2015

Missing values are not due mostly to low-abundance proteins

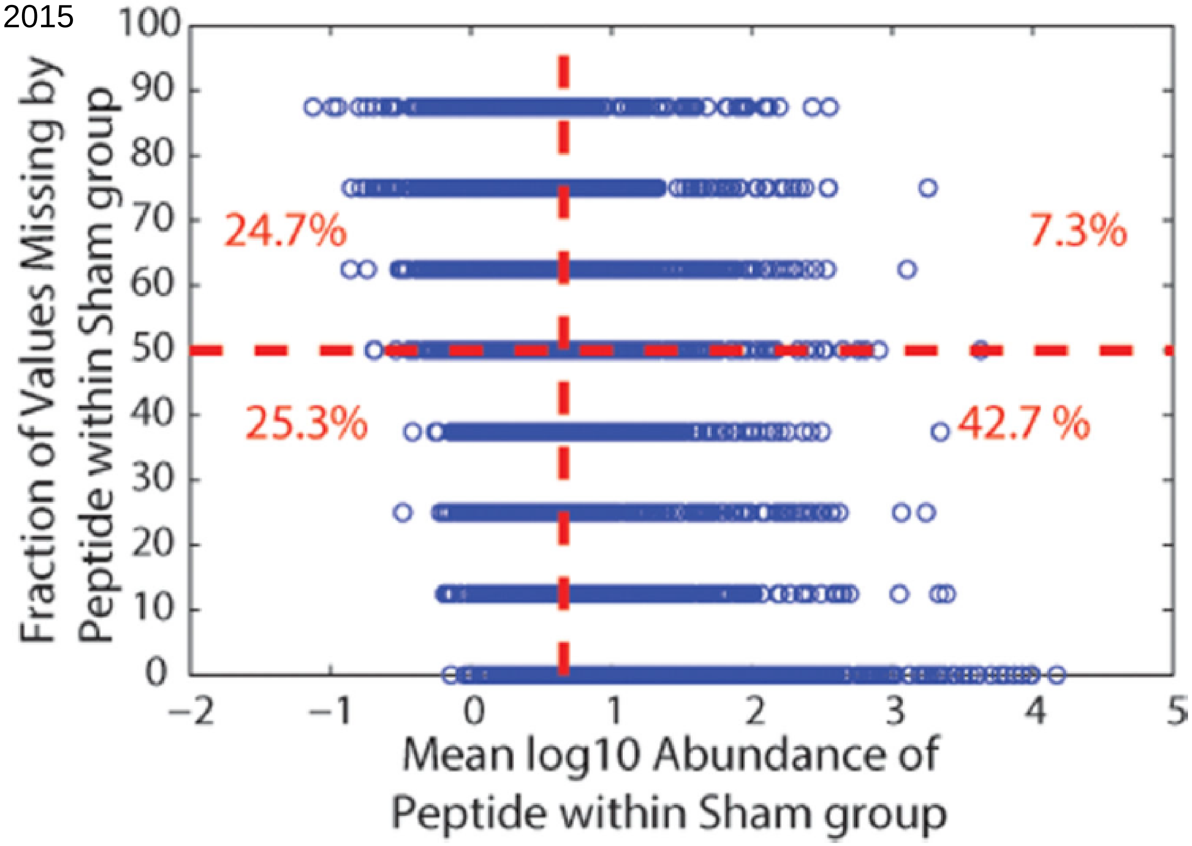


Figure 1.

Average $\log_{10}$ intensity as measured by peptide peak area in the control group versus fraction of missing values and peptide counts associated with bins corresponding to the fraction of missing data comparing phenotypes and exposures for datasets from (A) human plasma and (B) mouse lung. The control group for the human plasma is the normal glucose tolerant (NGT) samples, and the sham group for the mouse lung is the regular weight mice with no lipopolysaccharide (LPS) exposure. The vertical red line represents median average intensity, and the horizontal red line represents the point that 50% of the values are missing.

Webb-Robertson, *JPR*, 14(5):1993-2001, 2015

CV(RMSE) ~ 20% at 75th percentile

Current imputation methods don't work very well

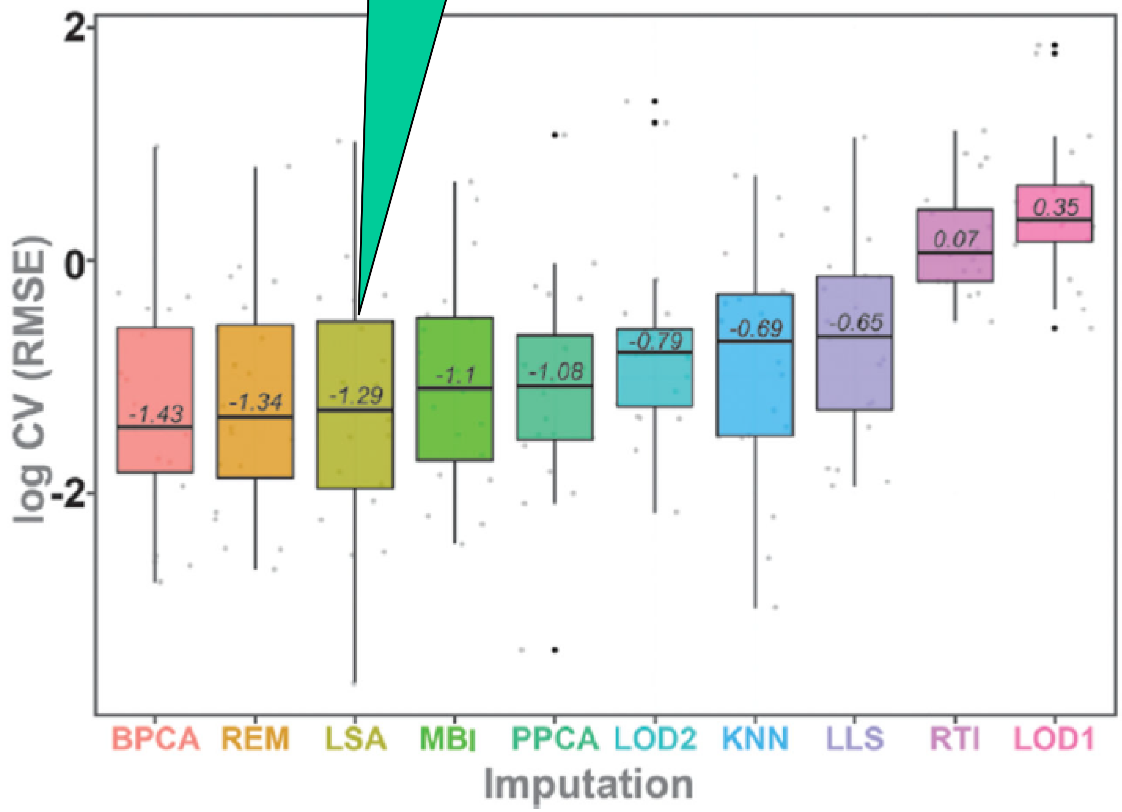


Figure 2.

Boxplot of the average $\log_{10}$ CV(RMSE) for the imputed dilution series datasets (Table 1) at the (A) peptide and (B) protein levels. The lower line represents the 25th percentile, the upper line of the box represents the 75th percentile, and the inner line corresponds to the median $\log_{10}$ CV(RMSE).

Exercise

- **Postulate: The chance of a protein complex being present in a sample is proportional to the fraction of its constituent proteins being correctly reported in the sample**

- **Derive from the postulate above an approach/index for predicting which proteins are likely to be present in a sample even though the proteomic screen does not report it**
- **You may assume a list of protein complexes (and their constituent proteins) is available**

Exercise

- **What experiments can you do to demonstrate that the proposed solution is effective in predicting missing proteins?**

I HOPE YOU HAVE ENJOYED THESE EXERCISES ☺