Metody Statystyczne w Biologii

Tomasz Suchocki

Łukasz Pawelec

Wstęp

1. Metody statystyczne w biologii ???
2. Pracownia Biostatystyki – aktualne badania
3. Charakterystyka przedmiotu
4. Kontakt
5. Literatura

Metody statystyczne w biologii ???

"...science is not data. Data are the raw material of science. It is what you do with the data that is science – the interpretation you make, the story you tell."

Metody statystyczne w biologii - SNP

[Header]	
BSGT Version	3.2.32
Processing Date	11/24/2008 10:14 AM
Content	BovineSNP50_A.bpm
Num SNPs	54001
Total SNPs	54001
Num Samples	32
Total Samples	2636
[Data]	

SNP Name	Sample ID	GC Score	SNP Index	Allele1 - AB	Allele2 - AB	Chr	Position	GT Score
ARS-BFGL-BAC-10172	4408169492_K	0.883	1	B	B	14	4736993	0.849
ARS-BFGL-BAC-1020	4408169492_K	0.899	2	B	B	14	6339014	0.8626
ARS-BFGL-BAC-10245	4408169492_K	0.6582	3	B	B	14	30073020	0.71
ARS-BFGL-BAC-10345	4408169492_K	0.9092	4	A	B	14	4497877	0.8721
ARS-BFGL-BAC-10365	4408169492_K	0.8021	5	B	B	14	25140301	0.833
ARS-BFGL-BAC-10375	4408169492_K	0.8858	6	A	B	14	4983527	0.8513
ARS-BFGL-BAC-10591	4408169492_K	0.867	7	A	B	14	15446975	0.8363
ARS-BFGL-BAC-10793	4408169492_K	0.8722	8	B	B	14	27452258	0.8403
ARS-BFGL-BAC-10867	4408169492_K	0.9316	9	A	B	14	32700054	0.8949
ARS-BFGL-BAC-10919	4408169492_K	0.7805	10	A	B	14	29520816	0.778
ARS-BFGL-BAC-10952	4408169492_K	0.9314	11	B	B	10	19315327	0.8947
ARS-BFGL-BAC-10960	4408169492_K	0.6543	12	B	B	10	21056606	0.7079
ARS-BFGL-BAC-10975	4408169492_K	0.8622	13	A	B	10	21682679	0.8358
ARS-BFGL-BAC-10986	4408169492_K	0.8687	14	A	B	10	25897020	0.8376
ARS-BFGL-BAC-10993	4408169492_K	0.8146	15	A	B	10	80403647	0.7993
ARS-BFGL-BAC-11000	4408169492_K	0.9135	16	A	A	10	81191638	0.8762

N = 19 778 743 834

Metody statystyczne w biologii - SNP

```
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  BSWCHEM120014887571
Chr1 182   C       T       30   DP=2;VDB=7.200000e-02;AF1=1;AC1=2;DP4=0,0,2,0;MQ=34;FQ=-33   GT:PL:GQ 1/1:62,6,0:10
Chr1 300   A       G       87   DP=6;VDB=8.330040e-02;RPB=0.000000e+00;AF1=0.5;AC1=1;DP4      GT:PL:GQ 0/1:117,0,52:5
Chr1 324   A       G       34   DP=9;VDB=6.733101e-02;RPB=-1.711553e+00;AF1=0.5;AC1=1;DP4=    GT:PL:GQ 0/1:64,0,160:6
Chr1 340   G       A       90   DP=14;VDB=8.462522e-02;RPB=-1.333333e-01;AF1=0.5;AC1=1;DP4=   GT:PL:GQ 0/1:120,0,209:9
Chr1 353   T       A       136  DP=14;VDB=1.121465e-01;RPB=1.219070e+00;AF1=0.5;AC1=1;DP      GT:PL:GQ 0/1:166,0,49:5
Chr1 355   T       A       141  DP=14;VDB=9.310645e-02;RPB=1.219070e+00;AF1=0.5;AC1=1;DP4=   GT:PL:GQ 0/1:171,0,50:53
Chr1 380   G       T       103  DP=18;VDB=1.049857e-01;RPB=8.897565e-01;AF1=0.5;AC1=1;DP      GT:PL:GQ 0/1:133,0,241:9
Chr1 420   T       A       211  DP=19;VDB=1.566941e-01;RPB=-7.964914e-01;AF1=0.5;AC1=1;DP     GT:PL:GQ 0/1:241,0,81:84
```

7 198 552 wariantów polimorficznych dla 1 osobnika

Metody statystyczne w biologii - CNV

Duplication	chr1:4001-16300	12300	2.10151	0	1.76985e-38	0	2.2605e-49	1
deletion	chr1:16301-20400	4100	0.535091	3.73056e-06		15163.9	3.33459	1
duplication	chr1:20401-24500	4100	1.81889	0.000438811		9.48454	10.3995	1
duplication	chr1:43501-62600	19100	2.19581	0		2.21431	0	1
duplication	chr1:64901-68800	3900	2.29307	0		1.84995	0.000141891	1
deletion	chr1:215901-217800	1900	0	8.38803e-11		9.73635	1	1
deletion	chr1:319601-320500	900	0.026701	0.000351021		9.17077	1	1
deletion	chr1:518401-519200	800	0.171095	0.188201		1.02726	1	1
deletion	chr1:531101-537700	6600	0.553887	1.11078e-09		3577.68	5.86046e-05	1901.93 1
deletion	chr1:541901-542900	1000	0.056984	0.00162457		4.12208	1	1
deletion	chr1:626501-627200	700	0.020635	0.00758167		2.79548	1	1
deletion	chr1:665101-671700	6600	0.707933	1.16982e-06		399149	0.000237804	1
deletion	chr1:761501-762300	800	0.037549	0.0396925		3.13366	1	1
deletion	chr1:1044501-1045500	1000	0.0142217	1.22398e-07		3.16665	1	1

13 149 – 22 496 delecji & 1 694 – 5 187 duplikacji dla 1go osobnika

Metody statystyczne w biologii – ekspresja genów

Name	TPM	TPM	TPM	TPM	TPM	TPM	TPM	TPM	TPM
ENSRNOT00000047550.4	4884.697493	4738.277530	2604.213519	3985.752318	5732.663362	5023.493819	3846.591490	4279.177573	
ENSRNOT00000040993.4	2482.340658	2689.680829	1272.801658	2311.170740	2822.156368	3353.882591	2641.689557	3297.695371	
ENSRNOT00000050156.3	18523.723689	20188.548961	12589.329509	16281.978897	31121.720782	19718.666924	15821.412541	17712.661162	
ENSRNOT00000043693.3	13479.177781	16415.673632	8908.516205	13599.408288	15707.617971	8904.178691	8911.699095	9638.886602	
ENSRNOT00000046201.3	2879.664120	4460.470710	1839.333709	6918.925769	7377.920279	4084.025521	3817.763433	4287.996188	
ENSRNOT00000046108.3	18639.827814	20342.476378	9689.638476	15075.380942	23089.394108	16351.850792	13938.721661	16260.457165	
ENSRNOT00000049683.3	19454.307846	21761.261482	10506.336626	15560.190788	24647.725602	15938.942147	14507.621638	17490.580003	
ENSRNOT00000041241.3	11171.724511	15702.243427	3103.880485	6857.837582	4621.824177	4210.072799	3550.848633	3658.665273	
ENSRNOT00000044582.3	1190.921111	1397.940532	407.491550	694.439063	834.559687	1701.483113	1679.903406	1856.376706	
ENSRNOT00000042928.3	2605.634407	2628.967598	1081.744293	2533.522841	3279.023333	3215.011148	2400.983250	2679.926014	
ENSRNOT00000048767.3	446.009903	687.564853	451.596293	420.324947	537.425809	484.384952	396.101451	451.476732	
ENSRNOT00000051268.3	2022.698641	2570.953131	2389.831386	2304.667829	3815.315728	1304.296149	930.693816	1261.863284	
ENSRNOT00000042098.3	5833.192678	6818.090134	3627.982640	4174.378585	9000.851702	5441.383826	5087.312394	5584.573047	
ENSRNOT00000067916.3	0.214827	0.199681	0.486458	0.382470	0.260718	0.036366	0.000000	0.156468	
ENSRNOT00000000471.6	1.476165	2.108333	0.871890	1.427177	0.853302	1.041466	1.081356	0.940819	
ENSRNOT00000084420.1	5.157895	5.356282	8.540623	4.538862	5.614132	3.840417	3.619653	3.680542	
ENSRNOT00000017829.4	1.240067	0.685103	1.516322	0.955789	1.346555	1.033335	0.768898	1.599833	
ENSRNOT00000072054.2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
ENSRNOT00000061930.4	0.189653	0.087897	0.473867	0.119897	0.098815	0.126786	0.064671	0.054892	
ENSRNOT00000009763.5	1.465691	1.645721	2.010367	1.559923	1.051099	1.896500	2.822518	1.941354	
ENSRNOT00000020339.7	2.330289	2.292182	4.488911	3.495487	4.478677	3.063810	3.325385	4.819330	
ENSRNOT00000005273.7	11.001476	10.727183	8.638224	10.590056	14.116366	14.380876	16.135767	14.224954	
ENSRNOT00000018987.3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
ENSRNOT00000071707.3	0.000000	0.000000	0.000000	0.000000	0.000000	0.444121	0.000000	0.000000	
ENSRNOT00000073435.2	0.554749	0.150595	0.000000	0.152206	0.327402	0.247616	0.281225	0.355818	
ENSRNOT00000048033.3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
ENSRNOT00000059467.4	235.816117	227.821360	97.863179	140.032364	101.698302	106.670608	82.425375	98.970228	
ENSRNOT00000043686.2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
ENSRNOT00000001973.3	0.000000	0.019016	0.000000	0.026666	0.000000	0.000000	0.112665	0.000000	
ENSRNOT00000023584.6	0.024841	0.045247	0.000000	0.037102	0.098838	0.000000	0.053498	0.000000	
ENSRNOT00000078650.1	0.940966	1.267720	1.035848	1.567569	0.552175	0.299765	0.767257	0.569210	
ENSRNOT00000007309.6	0.000000	0.014477	0.000000	0.040847	0.000000	0.000000	0.021474	0.000000	
ENSRNOT00000026875.5	26.278341	22.653593	34.107159	26.502578	35.471675	33.311650	36.176177	36.026062	
ENSRNOT00000068623.2	0.034889	0.101411	0.142327	0.000000	0.047457	0.029624	0.090096	0.152714	
ENSRNOT00000004104.6	0.000000	0.000000	0.000000	0.029258	0.000000	0.000000	0.000000	0.026235	
ENSRNOT00000034448.5	3.164855	3.822186	3.516888	3.984256	3.453593	3.713983	3.348950	4.486245	
ENSRNOT00000006953.6	0.234321	0.169933	0.232924	0.140613	0.660402	0.197127	0.250716	0.042534	

Ekspresja 40 104 transkryptów dla 10 osobników *Rattus norvegicus*

projekt 1

Pracownia biostatystyki – projekt

scientific reports

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › scientific reports › articles › article

Article | Open Access | Published: 10 May 2022

Identification of functional features underlying heat stress response in Sprague–Dawley rats using mixed linear models

Krzysztof Kotlarz, Magda Mielczarek, Yachun Wang, Jinhuan Dou, Tomasz Suchocki & Joanna Szyda ✉

Scientific Reports **12**, Article number: 7671 (2022) | Cite this article

Motywacja

- Identyfikacja elementów funkcjonalnych genomu odpowiedzialnych za stres cieplny
- Organizm modelowy – Szczur wędrowny
- Elementy funkcjonalne:
  - transkrypty
  - ontologie genów (GO)
  - ścieżki metaboliczne (Reactome)

Pracownia biostatystyki – projekt

Metodyka → liniowy model mieszany

Statistical modelling of expression data

The $\log_2$ fold changes ($\log_2$FC) calculated based on the transcript expression levels pooled over the control and heat-stressed animals respectively, were analysed in four mixed linear models.

The *transcript-based model* (M1) is given by:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z}_{M1}\boldsymbol{t} + \boldsymbol{e}_{M1}, \quad (1$$

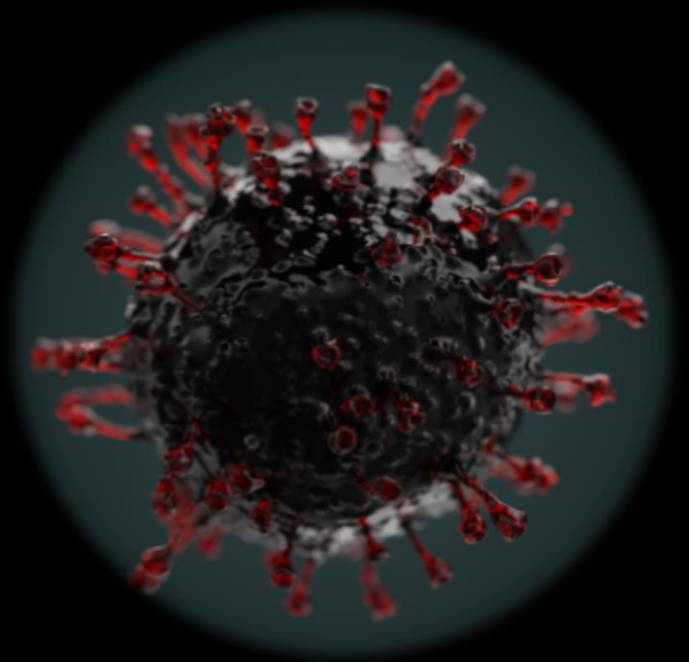
where $\boldsymbol{y}$ is the vector of $\log_2$FC of transcript expression, $\boldsymbol{\mu}$ represents the general mean, $\boldsymbol{t}$ is the random transcript effect with a predisposed normal distribution defined by $N\left(0, \boldsymbol{V}_{M1}\sigma_t^2\right)$, $\boldsymbol{e}_{M1}$ is a vector of residuals distributed as $N\left(0, \boldsymbol{I}\sigma_{e_{M1}}^2\right)$, $\boldsymbol{Z}_{M1}$ is an incidence matrix for $\boldsymbol{t}$. In this model, the similarity between transcripts *i* and *j*, was introduced into the model by incorporating a nondiagonal transcript covariance matrix $\boldsymbol{V}_{M1}$. The covariance between

Pracownia biostatystyki – projekt

Wyniki

Tissue	Model	ID name	Effect	P
Liver	Transcript (M1)	ENSRNOT00000074131 *PNKD*	2.76	4.7×10^{-10}
	Transcript (M1)	ENSRNOT00000093245 *TRIP12*	2.15	1.2×10^{-6}
	Transcript (M1)	ENSRNOT00000093735 *TRIP12*	1.67	0.00016
	Transcript (M1)	ENSRNOT00000022822 *TRIP12*	1.65	0.00021
	Transcript (M1)	ENSRNOT00000079452 *TRIP12*	1.63	0.00023
	Gene (M2)	ENSRNOG00000016963 *TRIP12*	1.66	3.5×10^{-5}
	Gene (M2)	ENSRNOG00000014806 *PNKD*	1.12	0.00520
	Gene ontology (M3)	GO:1901315 Negative regulation of histone H2A K63-linked ubiquitination	0.52	0.00055
	Gene ontology (M3)	GO:2000780 Negative regulation of double-strand break repair	0.48	0.00130
	Gene ontology (M3)	GO:2000779 Regulation of double-strand break repair	0.37	0.01300
	Gene ontology (M3)	GO:0045995 Regulation of embryonic development	0.30	0.04200
	Gene ontology (M3)	GO:0002181 Cytoplasmic translation	0.29	0.05200
	Reactome pathway (M4)	R-RNO-6791226 Major pathway of rRNA processing in the nucleolus and cytosol	0.29	0.05500
Adrenal	Transcript (M1)	ENSRNOT00000075998 *SUCO*	3.48	9.2×10^{-11}
	Transcript (M1)	ENSRNOT00000084058 *SUCO*	2.18	4.9×10^{-5}
	Transcript (M1)	ENSRNOT00000082271 *PLEC*	2.10	8.9×10^{-10}
	Transcript (M1)	ENSRNOT00000075936 *SUCO*	1.95	0.00029
	Transcript (M1)	ENSRNOT00000088945 *PLEC*	1.70	0.00160
	Gene (M2)	ENSRNOG00000026542 *SUCO*	2.26	8.4×10^{-6}
	Gene (M2)	ENSRNOG00000018413 *PER3*	1.63	0.00130
	Reactome pathway (M4)	R-RNO-212436 Generic Transcription Pathway	0.49	0.00500

projekt 2



Pracownia biostatystyki – projekt

Modelowanie przebiegu pandemii COVID-19

Dane

- Public SARS-CoV-2 Data Repository → Johns Hopkins University → GitHub
- Liczba zakażonych, zmarłych, ozdrowieńców
- od 21.01.2020
- 191 krajów

Pracownia biostatystyki – projekt

Metodyka → model SIRD

SIRD (model epidemiologiczny, obserwacje do dnia [illegible])

- Liczba podatnych → $S(t) = S(t-1) - \frac{\alpha}{N} S(t-1) I(t-1)$
- Liczba zakażonych → $I(t) = I(t-1) + \frac{\alpha}{N} S(t-1) I(t-1) - \beta I(t-1) - \gamma I(t-1)$
- Liczba ozdrowieńców → $R(t) = R(t-1) + \beta I(t-1)$
- Liczba zmarłych → $D(t) = D(t-1) + \gamma I(t-1)$

time [days]

infection rate

population

recovery rate

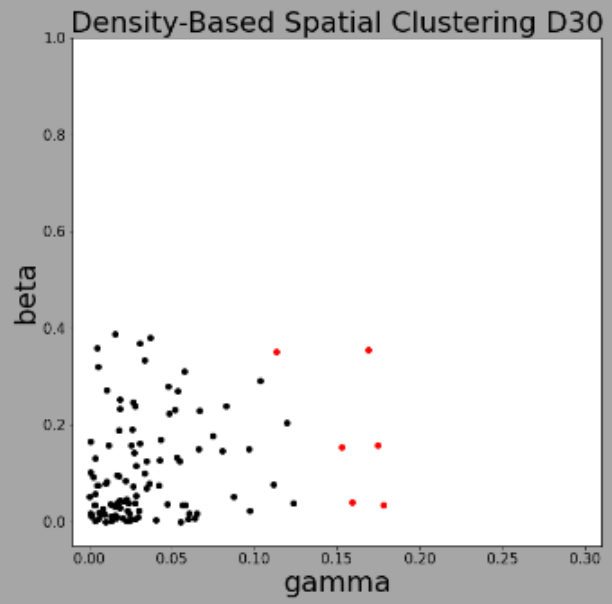
mortality rate

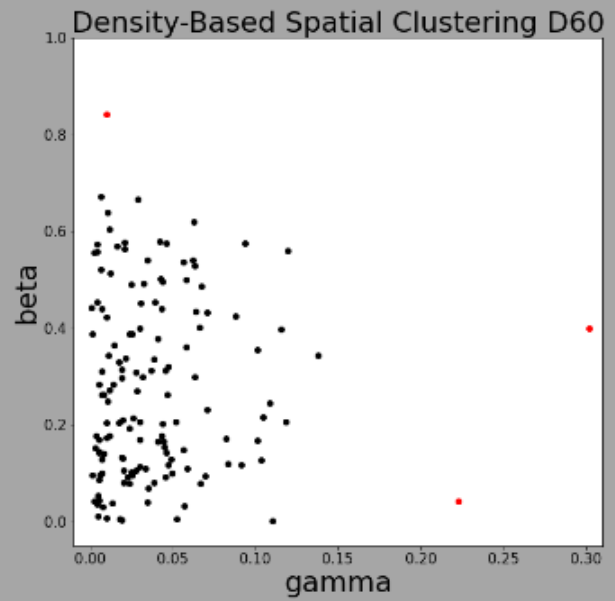
Liniowy model mieszany (modelowanie statystyczne, obserwacje do dnia 300)

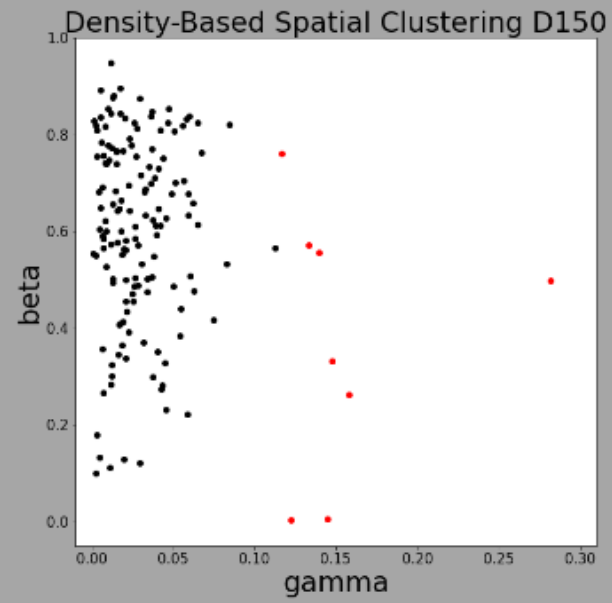
- Liczba zakażonych

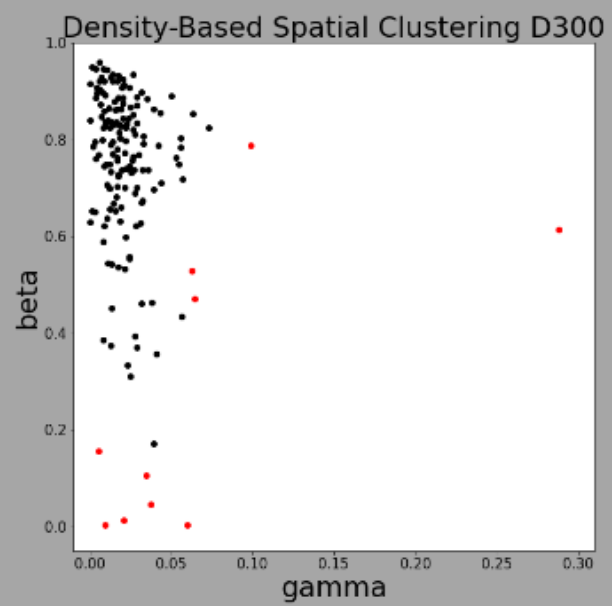
Pracownia biostatystyki – projekt

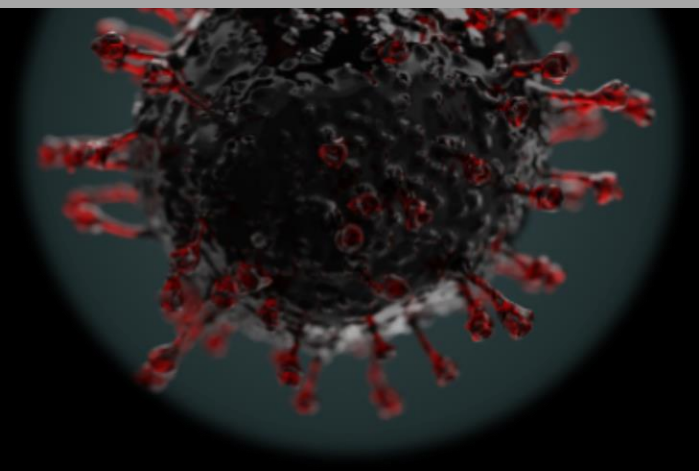
Wyniki → model SIRD → klastry





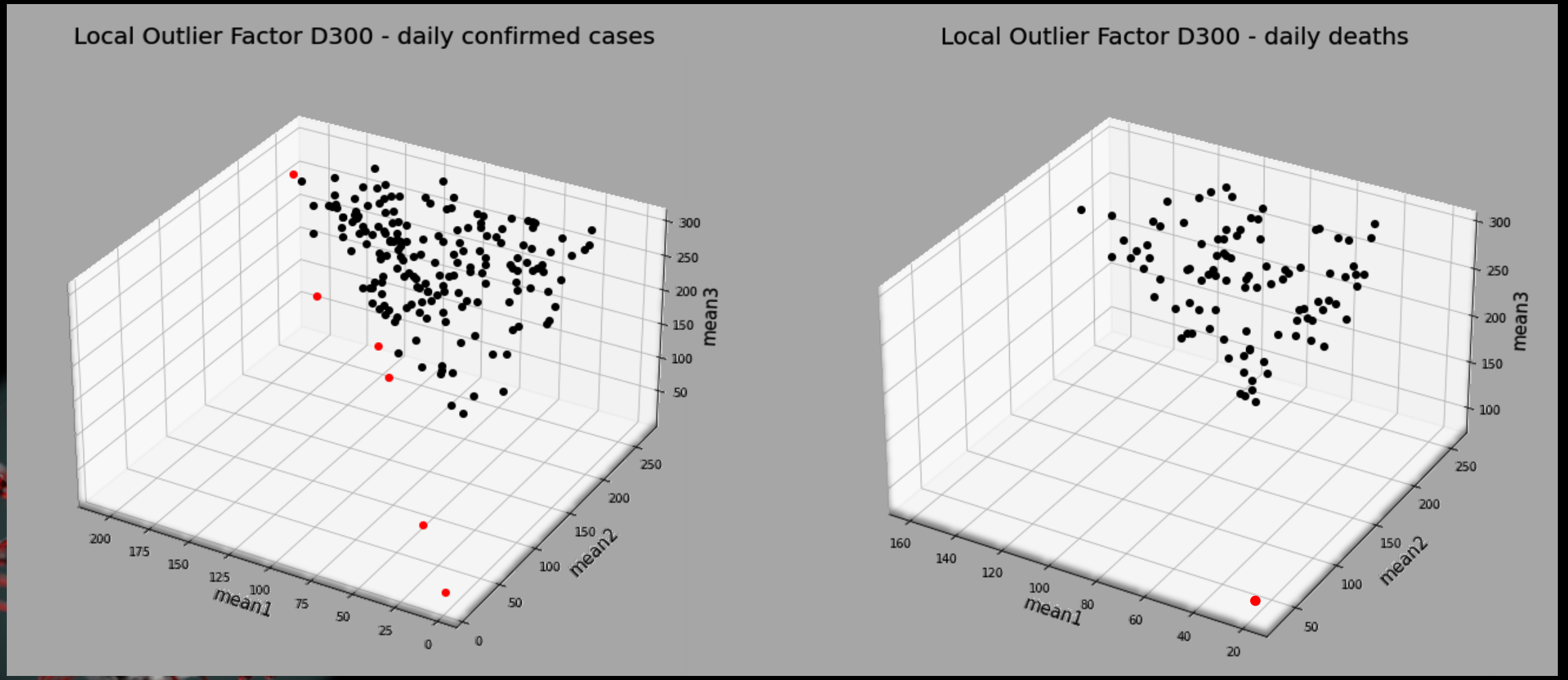






Pracownia biostatystyki – projekt

Wyniki → model mieszany

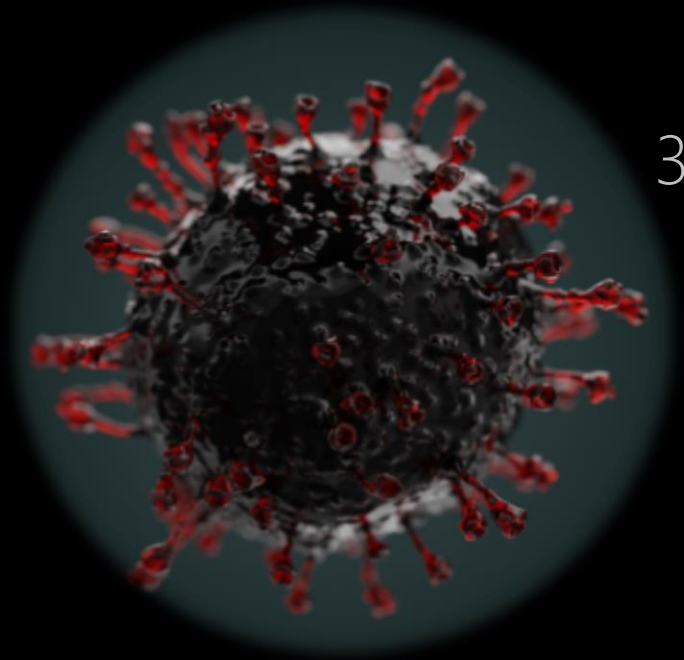


Pracownia biostatystyki – projekt

Modelowanie przebiegu pandemii COVID-19

1. Zróżnicowanie między krajami – wyróżniono kraje odstające

2. Lichtenstein → "pozytywnie" odstający
   - niska śmiertelność i wysoka liczba wyzdrowień



3. Yemen → "negatywnie" odstający
   - wysoka śmiertelność i niska liczba wyzdrowień

Wykłady

Charakterystyka wykładów

1. Umiejętność analizy danych biologicznych o zróżnicowanej strukturze
2. Statystyczne podstawy analizy danych
3. Umiejętność interpretacji wyników
4. Aktywne uczestnictwo → pytania

Tematyka wykładów

Podstawy statystycznej analizy danych

1. Wykład wstępny
2. Populacje i próby danych
3. Testowanie hipotez i estymacja parametrów
4. Planowanie eksperymentów biologicznych
5. Najczęściej wykorzystywane testy statystyczne I
6. Najczęściej wykorzystywane testy statystyczne II

Tematyka wykładów

Elementy statystycznego modelowania danych

7. Regresja liniowa
8. Regresja nieliniowa
9. Określenie jakości dopasowania równania regresji liniowej i nieliniowej
10. Korelacja
11. Elementy statystycznego modelowania danych
12. Porównywanie modeli
13. Analiza wariancji
14. Analiza kowariancji
15. Podsumowanie dotychczasowego materiału, wspólna analiza przykładów, dyskusja

Ćwiczenia

Tematyka ćwiczeń

1. Zaliczenie (bez poprawek !!!) – średnia ocen

2. Oceny:

- 2 kolokwia
- wykłady + ćwiczenia
- aktywność na ćwiczeniach

Charakterystyka ćwiczeń

Podstawowe zagadnienia biostatystyki

1. Ćwiczenia wstępne
2. Populacje i próby danych
3. Estymacja parametrów
4. Testowanie hipotez statystycznych I
5. Testowanie hipotez statystycznych II
6. Kolokwium 1

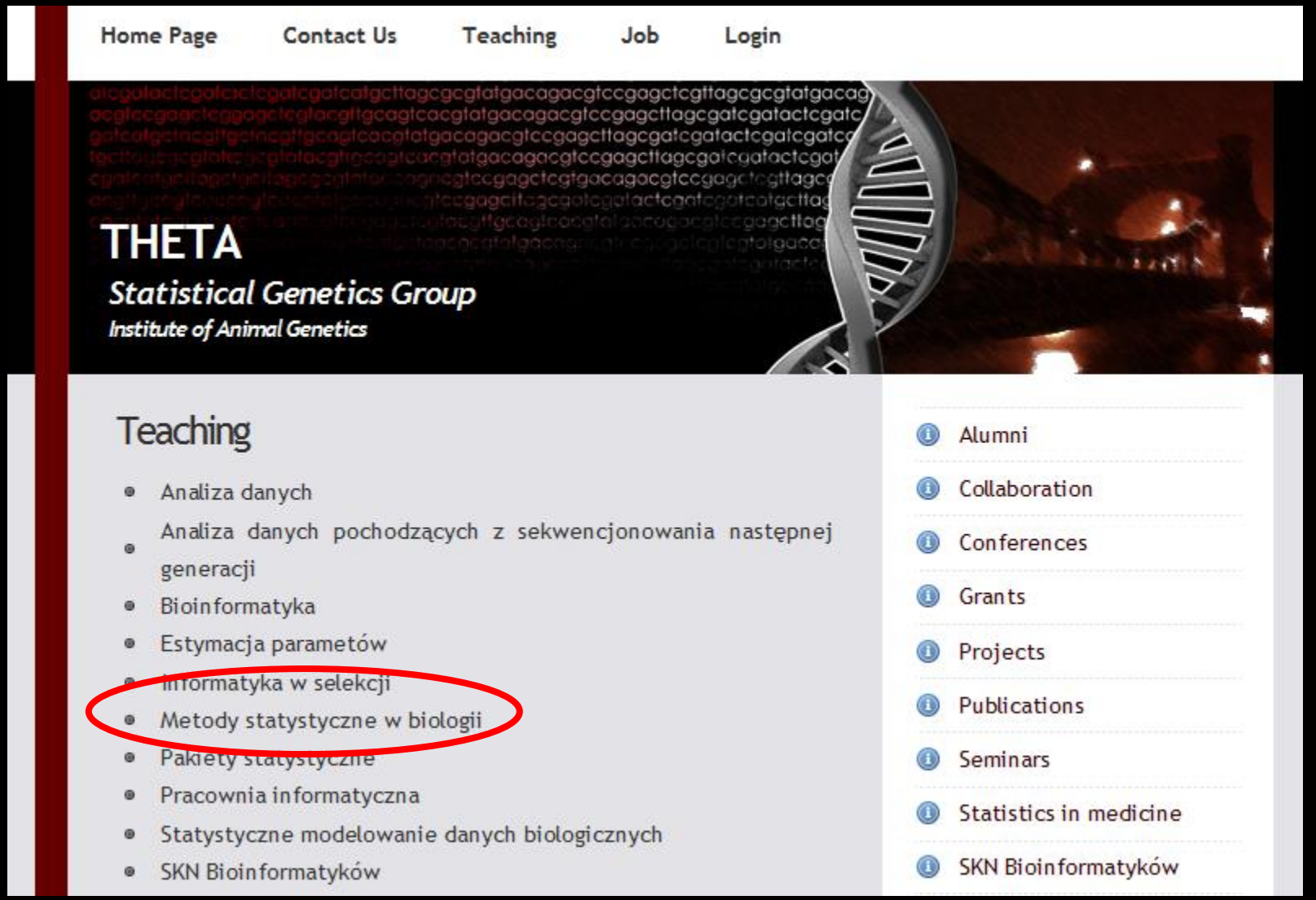
Tematyka ćwiczeń

Elementy statystycznego modelowania danych

7. Korelacja
8. Regresja liniowa
9. Regresja nieliniowa
10. Interpretacja wyników różnych modeli regresji
11. Kolokwium 2
12. Porównywanie modeli
13. Analiza wariancji
14. Prezentowanie przez grupy robocze wyników analizy danych
15. Zaliczenie ćwiczeń

Kontakt

http://theta.edu.pl/teaching/ → Metody statystyczne



Kontakt

Katedra Genetyki, Kożuchowska 7

Konsultacje

- termin ustalony indywidualnie
- online
- stacjonarne
- Email

Literatura

1. Wykłady
2. Książki statystyczne → statystyka dla biologów

1. Metody statystyczne w biologii ???
2. Pracownia Biostatystyki – aktualne badania
3. Charakterystyka przedmiotu
4. Kontakt
5. Literatura