

Przewidywanie przynależności krótkich sekwencji (50nt - 150nt) CDS *A. thaliana* do genomów: jądrowego, mitochondrialnego lub plastydowego.

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A. thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

Metody

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A. thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

Statystyki adnotacji

CDS		gene	

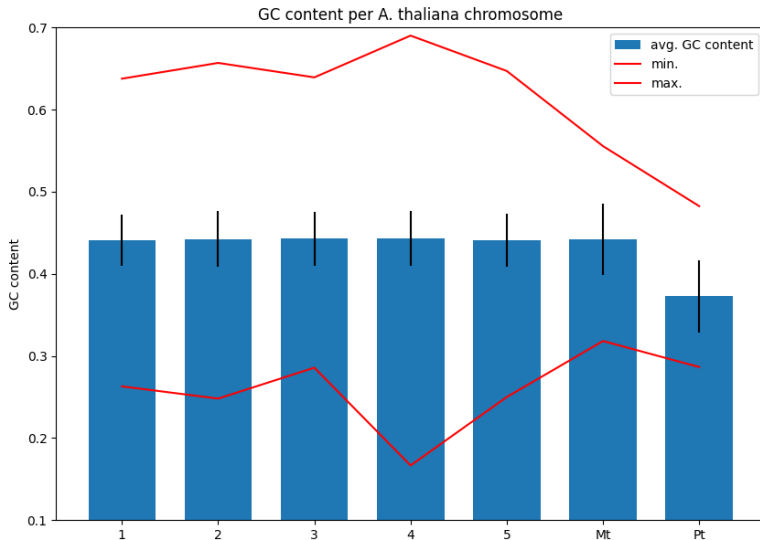
77519	1	7156	1
43804	2	4317	2
54853	3	5460	3
44061	4	4180	4
65591	5	6332	5
135	Mt	122	Mt
104	Pt	88	Pt

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A.
thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

per molecule GC content



Genetic code usage, nuclear

***Arabidopsis thaliana* [gbpln]: 80395 CDS's (31098475 codons)**

fields: [triplet] [frequency: per thousand] ([number])

UUU 21.8(678320)	UCU 25.2(782818)	UAU 14.6(455089)	UGU 10.5(327640)
UUC 20.7(642407)	UCC 11.2(348173)	UAC 13.7(427132)	UGC 7.2(222769)
UUA 12.7(394867)	UCA 18.3(568570)	UAA 0.9(29405)	UGA 1.2(36260)
UUG 20.9(649150)	UCG 9.3(290158)	UAG 0.5(16417)	UGG 12.5(388049)
CUU 24.1(750114)	CCU 18.7(580962)	CAU 13.8(428694)	CGU 9.0(280392)
CUC 16.1(500524)	CCC 5.3(165252)	CAC 8.7(271155)	CGC 3.8(117543)
CUA 9.9(307000)	CCA 16.1(502101)	CAA 19.4(604800)	CGA 6.3(195736)
CUG 9.8(305822)	CCG 8.6(268115)	CAG 15.2(473809)	CGG 4.9(151572)
AUU 21.5(668227)	ACU 17.5(544807)	AAU 22.3(693344)	AGU 14.0(435738)
AUC 18.5(576287)	ACC 10.3(321640)	AAC 20.9(650826)	AGC 11.3(352568)
AUA 12.6(391867)	ACA 15.7(487161)	AAA 30.8(957374)	AGA 19.0(589788)
AUG 24.5(762852)	ACG 7.7(240652)	AAG 32.7(1016176)	AGG 11.0(340922)
GUU 27.2(847061)	GCU 28.3(880808)	GAU 36.6(1139637)	GGU 22.2(689891)
GUC 12.8(397008)	GCC 10.3(321500)	GAC 17.2(535668)	GGC 9.2(284681)
GUA 9.9(308605)	GCA 17.5(543180)	GAA 34.3(1068012)	GGA 24.2(751489)
GUG 17.4(539873)	GCG 9.0(280804)	GAG 32.2(1002594)	GGG 10.2(316620)

Genetic code usage, mitochondrial

mitochondrion Arabidopsis thaliana [gbpln]: 129 CDS's (26228 cc

fields: [triplet] [frequency: per thousand] ([number])

UUU	36.1(	946)	UCU	22.6(	592)	UAU	20.7(	542)	UGU	10.9(	286)
UUC	27.2(	714)	UCC	15.4(	405)	UAC	10.0(	262)	UGC	7.6(	199)
UUA	22.3(	584)	UCA	15.6(	410)	UAA	2.4(	62)	UGA	2.0(	53)
UUG	20.6(	541)	UCG	10.9(	287)	UAG	1.0(	25)	UGG	14.6(	382)
CUU	24.8(	650)	CCU	19.9(	521)	CAU	17.2(	452)	CGU	10.9(	285)
CUC	17.1(	448)	CCC	10.8(	282)	CAC	7.3(	192)	CGC	6.8(	178)
CUA	15.5(	407)	CCA	14.3(	376)	CAA	20.9(	549)	CGA	11.6(	303)
CUG	13.2(	347)	CCG	9.2(	241)	CAG	12.2(	321)	CGG	8.0(	210)
AUU	26.3(	689)	ACU	17.9(	469)	AAU	21.2(	557)	AGU	14.5(	380)
AUC	19.8(	519)	ACC	12.1(	318)	AAC	13.6(	357)	AGC	11.8(	310)
AUA	18.7(	490)	ACA	13.9(	365)	AAA	27.6(	723)	AGA	16.4(	429)
AUG	26.6(	697)	ACG	8.0(	209)	AAG	21.0(	552)	AGG	12.0(	315)
GUU	17.4(	456)	GCU	23.8(	625)	GAU	22.6(	594)	GGU	18.6(	489)
GUC	11.5(	302)	GCC	13.5(	353)	GAC	12.0(	315)	GGC	9.1(	238)
GUA	14.8(	389)	GCA	15.1(	397)	GAA	26.4(	693)	GGA	22.0(	576)
GUG	12.8(	335)	GCG	8.3(	218)	GAG	18.2(	477)	GGG	13.0(	340)

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS A.
thaliana do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

Genetic code usage, plastid

***chloroplast Arabidopsis thaliana* [gbpln]: 162 CDS's (41837 codons)**

fields: [triplet] [frequency: per thousand] ([number])

UUU	35.2(	1473)	UCU	22.7(	950)	UAU	24.3(	1017)	UGU	8.2(	344)
UUC	19.3(	808)	UCC	12.9(	540)	UAC	8.6(	361)	UGC	3.7(	156)
UUA	29.0(	1215)	UCA	16.2(	677)	UAA	2.0(	84)	UGA	1.2(	52)
UUG	19.5(	817)	UCG	6.9(	289)	UAG	0.7(	29)	UGG	15.8(	662)
CUU	22.8(	953)	CCU	17.1(	717)	CAU	15.3(	642)	CGU	12.3(	516)
CUC	9.3(	391)	CCC	6.6(	275)	CAC	5.7(	239)	CGC	4.0(	169)
CUA	13.9(	582)	CCA	11.7(	488)	CAA	23.8(	997)	CGA	11.0(	461)
CUG	7.8(	327)	CCG	5.2(	216)	CAG	9.1(	379)	CGG	4.6(	192)
AUU	37.6(	1571)	ACU	20.9(	876)	AAU	32.1(	1345)	AGU	14.7(	616)
AUC	17.1(	714)	ACC	9.5(	398)	AAC	15.1(	632)	AGC	6.1(	255)
AUA	24.9(	1040)	ACA	17.3(	725)	AAA	41.0(	1714)	AGA	16.9(	706)
AUG	21.9(	917)	ACG	4.8(	200)	AAG	19.8(	828)	AGG	6.7(	279)
GUU	22.4(	936)	GCU	26.4(	1105)	GAU	30.9(	1291)	GGU	22.3(	932)
GUC	8.6(	360)	GCC	9.3(	388)	GAC	10.3(	433)	GGC	7.1(	299)
GUA	16.9(	706)	GCA	16.2(	676)	GAA	39.3(	1645)	GGA	28.2(	1179)
GUG	11.9(	499)	GCG	7.8(	326)	GAG	18.3(	767)	GGG	11.0(	461)

Tworzenie zestawu danych

Sekwencje CDS genomu ref. TAIR10 pobrano z Ensembl Plants.

Sekwencje pocięto na nakładające lub nienakładające się fragmenty równej długości (zależnie od podejścia, między 15 a 150 nt), i obliczono wartości:

Cechy (features) do ML

- 1 GC%
- 2 entropia Shannona
- 3 skośność GC
- 4 skośność AT
- 5 Stosunek obserwowanej do przewidywanej liczby dinukleotydów w oknie (m.in. CpG) [Gardiner-Garden]
- 6 Zliczenia k-merów (testowano 3-, 4- i 5-mery)

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A. thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

Etykiety (labels)

Testowano układ z trzema klasami (Nuc, Mt, Pt), siedmioma klasami (1-5, Mt, Pt).

Modele

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A. thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

Drzewa decyzyjne, RandomForest, LightGBM

cont.

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A. thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

Problem nierównowagę klas (znacznie więcej CDSów z 1-5 niż Mt i Pt razem wziętych) próbowano rozwiązać przez undersampling. Wszystkie dane Mt i Pt + losowa próba z zestawu Nuc lub 1-5, żeby było po równo.

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A. thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

Wyniki

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A. thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

3 klasy, nakładające się sekwencje wejściowe, 15 nt

Metody

Wyniki

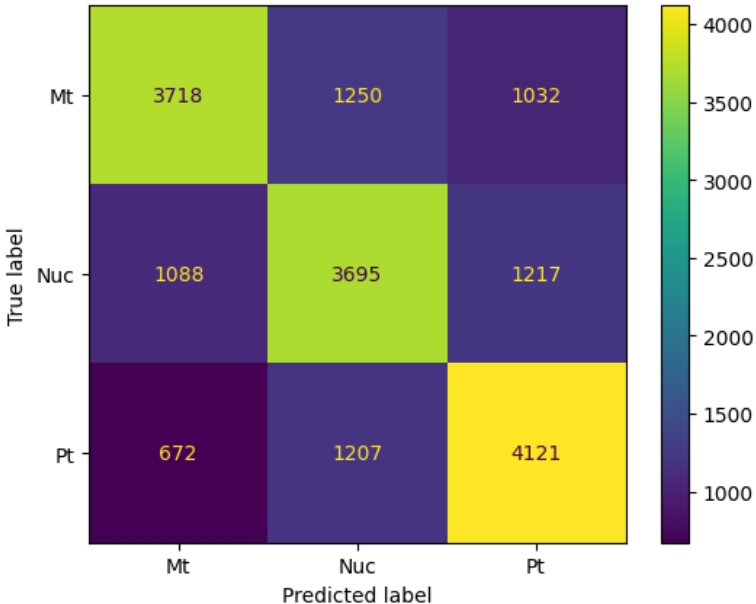
	precision	recall	f1-score	support
Mt	0.68	0.62	0.65	6000
Nuc	0.60	0.62	0.61	6000
Pt	0.65	0.69	0.67	6000
accuracy			0.64	18000
macro avg	0.64	0.64	0.64	18000
weighted avg	0.64	0.64	0.64	18000

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A.
thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

cont.



3 klasy, nakładające się sekwencje, 150 nt

Drzewo decyzyjne

	precision	recall	f1-score	support
Mt	0.87	0.94	0.90	3118
Nuc	0.99	0.98	0.99	52730
Pt	0.90	0.96	0.93	4152
accuracy			0.98	60000
macro avg	0.92	0.96	0.94	60000
weighted avg	0.98	0.98	0.98	60000

cont.

RandomForest

	precision	recall	f1-score	support
Mt	1.00	0.97	0.98	3118
Nuc	1.00	1.00	1.00	52730
Pt	1.00	0.99	0.99	4152
accuracy			1.00	60000
macro avg	1.00	0.99	0.99	60000
weighted avg	1.00	1.00	1.00	60000

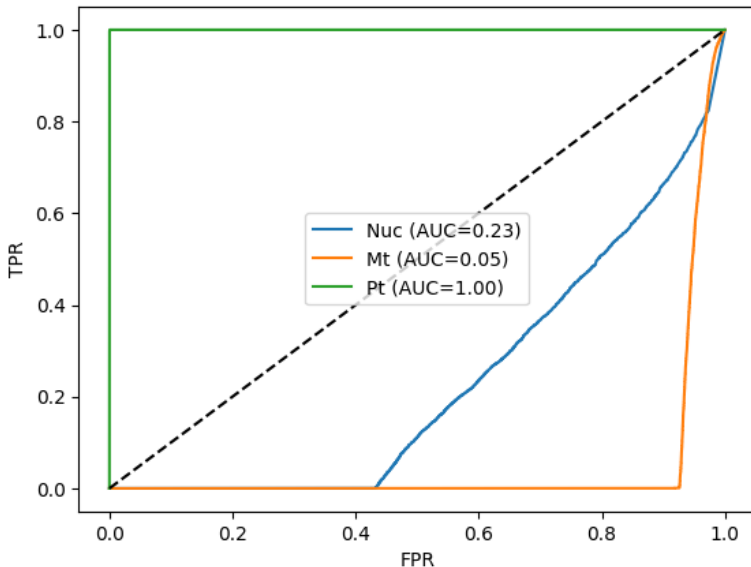
Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A.
thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

cont. 2

One-vs-rest ROC



Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A. thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

3 klasy, nienakładające się sekwencje, 30 nt, 3, 4 i 5 mery

Metody

Wyniki

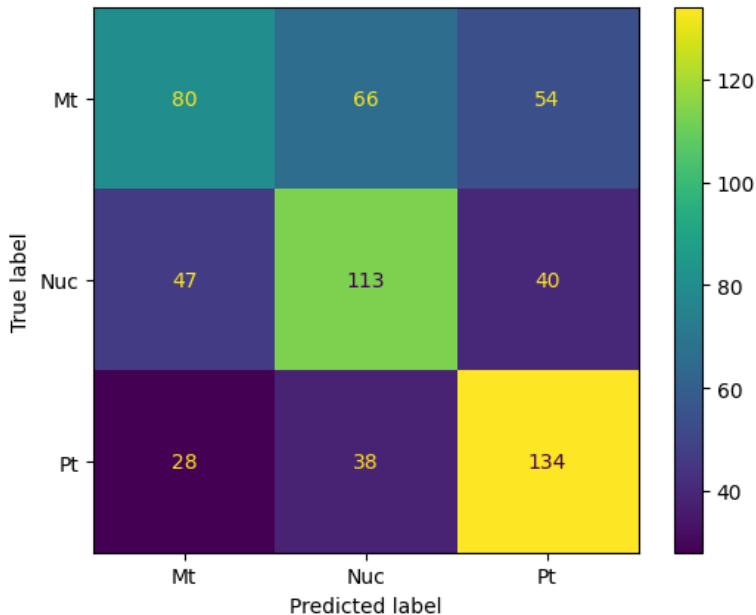
	precision	recall	f1-score	support
Mt	0.52	0.40	0.45	200
Nuc	0.52	0.56	0.54	200
Pt	0.59	0.67	0.63	200
accuracy			0.55	600
macro avg	0.54	0.55	0.54	600
weighted avg	0.54	0.55	0.54	600

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A.
thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

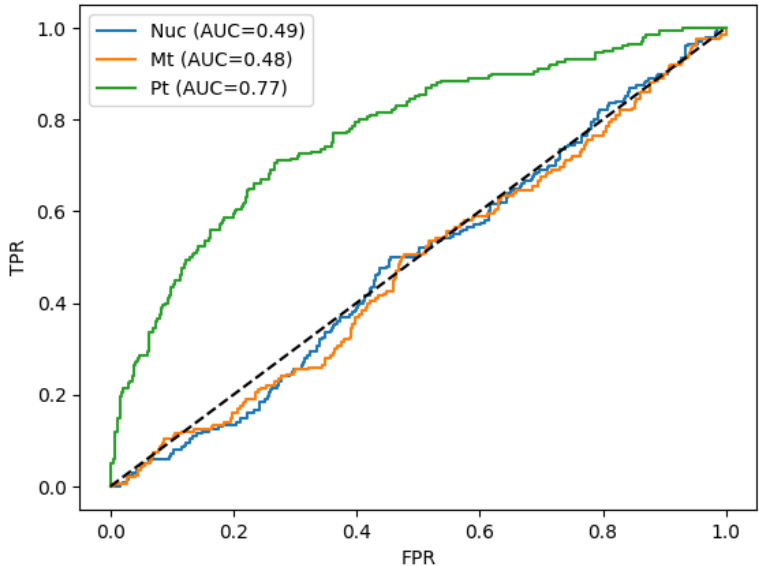
Metody

Wyniki

cont.



One-vs-rest ROC



Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A. thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

3 klasy, nienakładające się sekwencje, 30 nt, Gardiner-Gardner zamiast k-merów

Metody

Wyniki

	precision	recall	f1-score	support
Mt	0.39	0.30	0.34	438
Nuc	0.44	0.46	0.45	500
Pt	0.51	0.59	0.55	518
accuracy			0.46	1456
macro avg	0.45	0.45	0.45	1456
weighted avg	0.45	0.46	0.45	1456

Przewidywanie
przynależności
krótkich
sekwencji
(50nt - 150nt)
CDS *A.
thaliana* do
genomów:
jądrowego,
mitochondrial-
nego lub
plastydowego.

Metody

Wyniki

cont.

