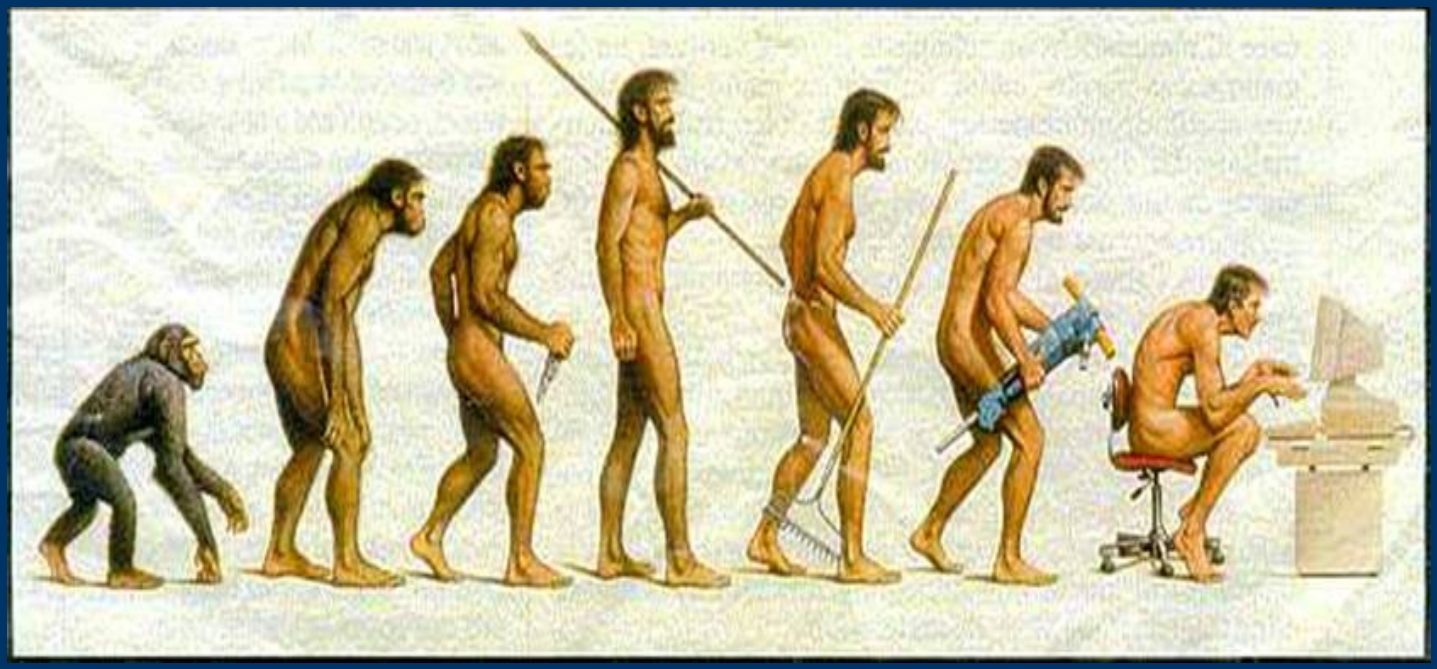
MACIERZE MUTACYJNE W ANALIZIE GENOMÓW – czy możliwa jest rekonstrukcja filogenetyczna?



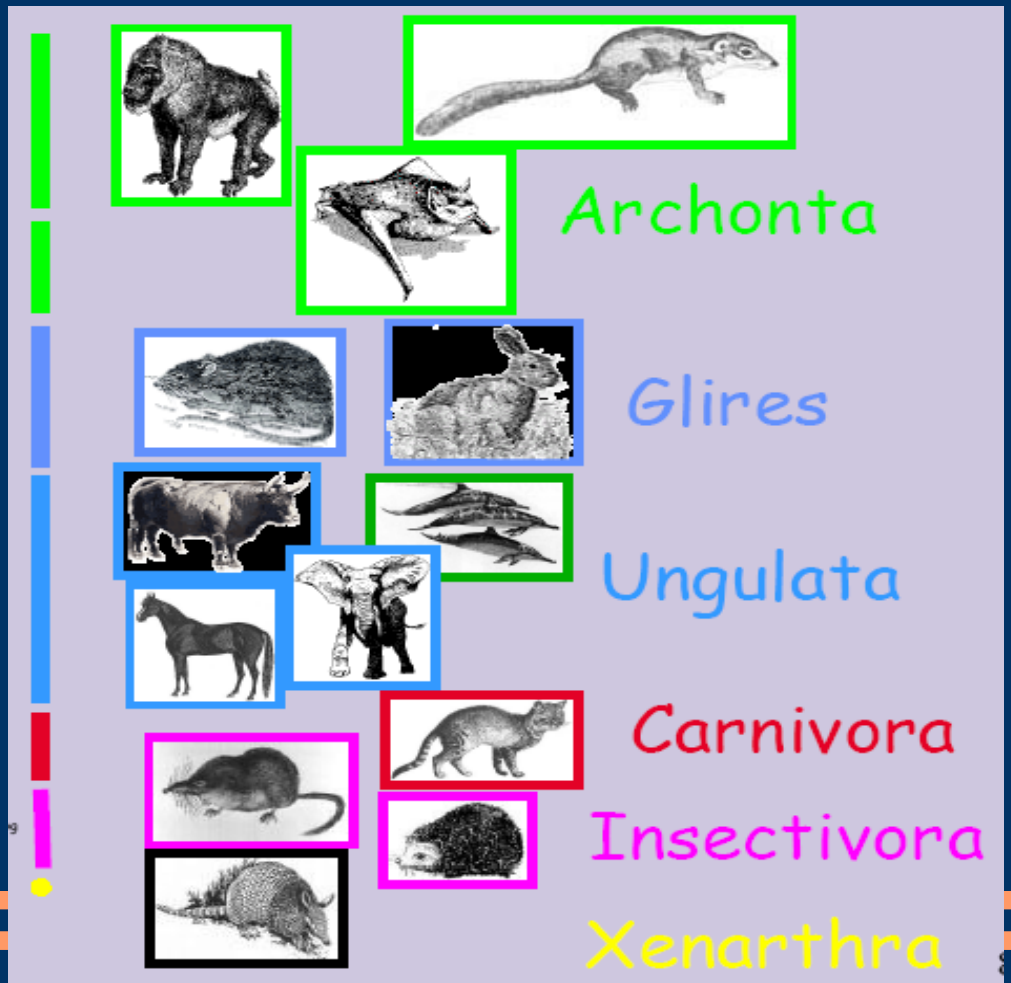
Aleksandra Nowicka

Zadaniem FILOGENETYKI jest :

- zrekonstruowanie ewolucyjnej historii wszystkich organizmów
- odkrycie przodka wszystkich organizmów żyjących na ziemi
- segregacja i klasyfikacja organizmów
- poznanie mechanizmów ewolucji

- **FILOGENETYKA W PODEJŚCIU KLASYCZNYM**

– rekonstrukcja historii ewolucji głównie w oparciu o cechy morfologiczne np. długość dzioba u ptaków, nóg itd.

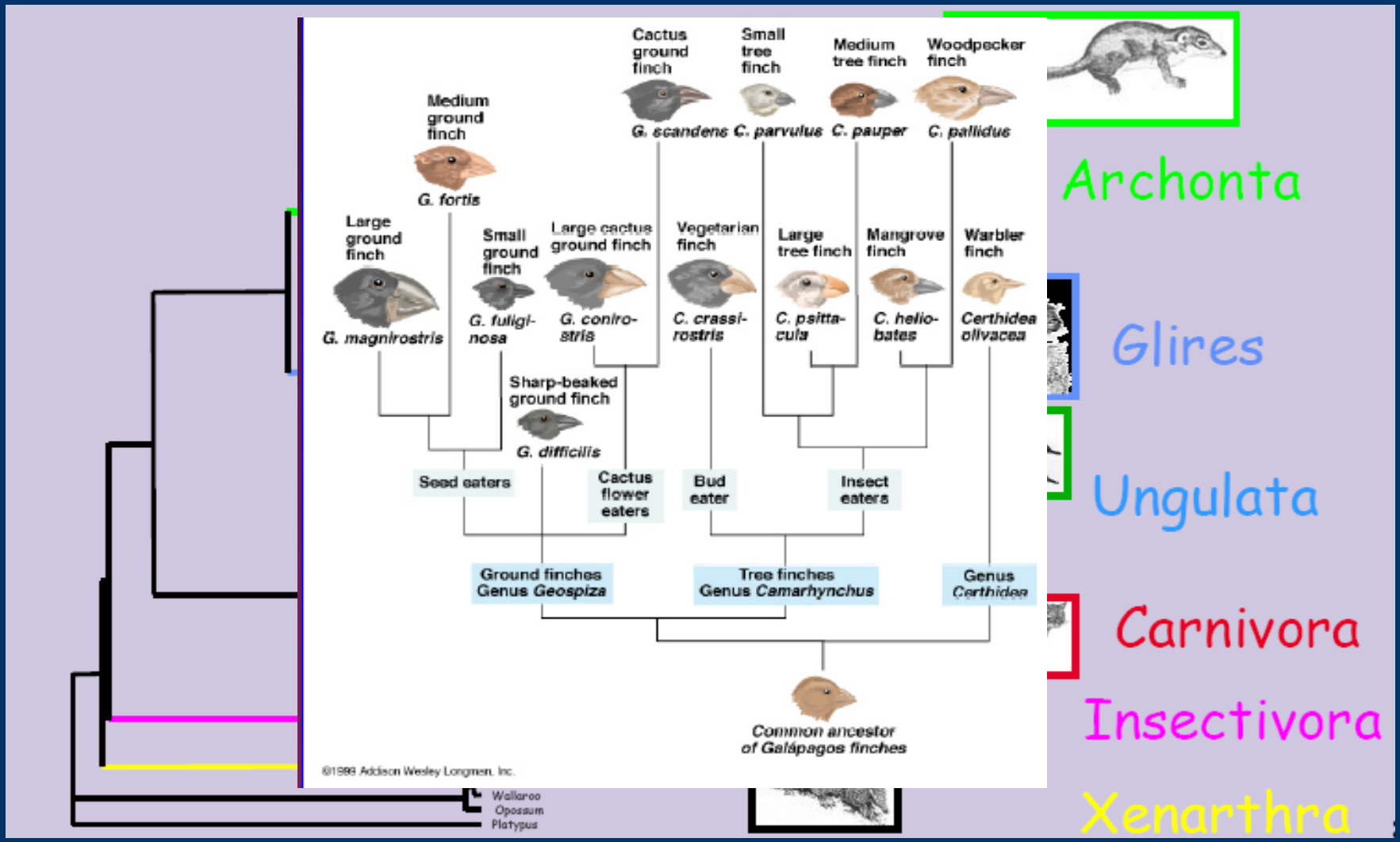


FILOGENETYKA MOLEKULARNA - rekonstrukcja historii ewolucji poprzez porównanie sekwencji nukleotydowych lub aminokwasowych pochodzących z różnych organizmów

Podstawowe założenia w **filogenetyce molekularnej**:

- każdy organizm posiada informację genetyczną (sekwencję nukleotydową) zwaną genomem
- sekwencje przodka mutują w sekwencje potomków
- podobne gatunki są genetycznie blisko spokrewnione
- różnice w sekwencjach nukleotydowych i aminokwasowych są wprost skorelowane z czasem ewolucji
- istnieje jeden wspólny przodek dla wszystkich form życia

- wyrazem analiz filogenetycznych są *drzewa filogenetyczne*



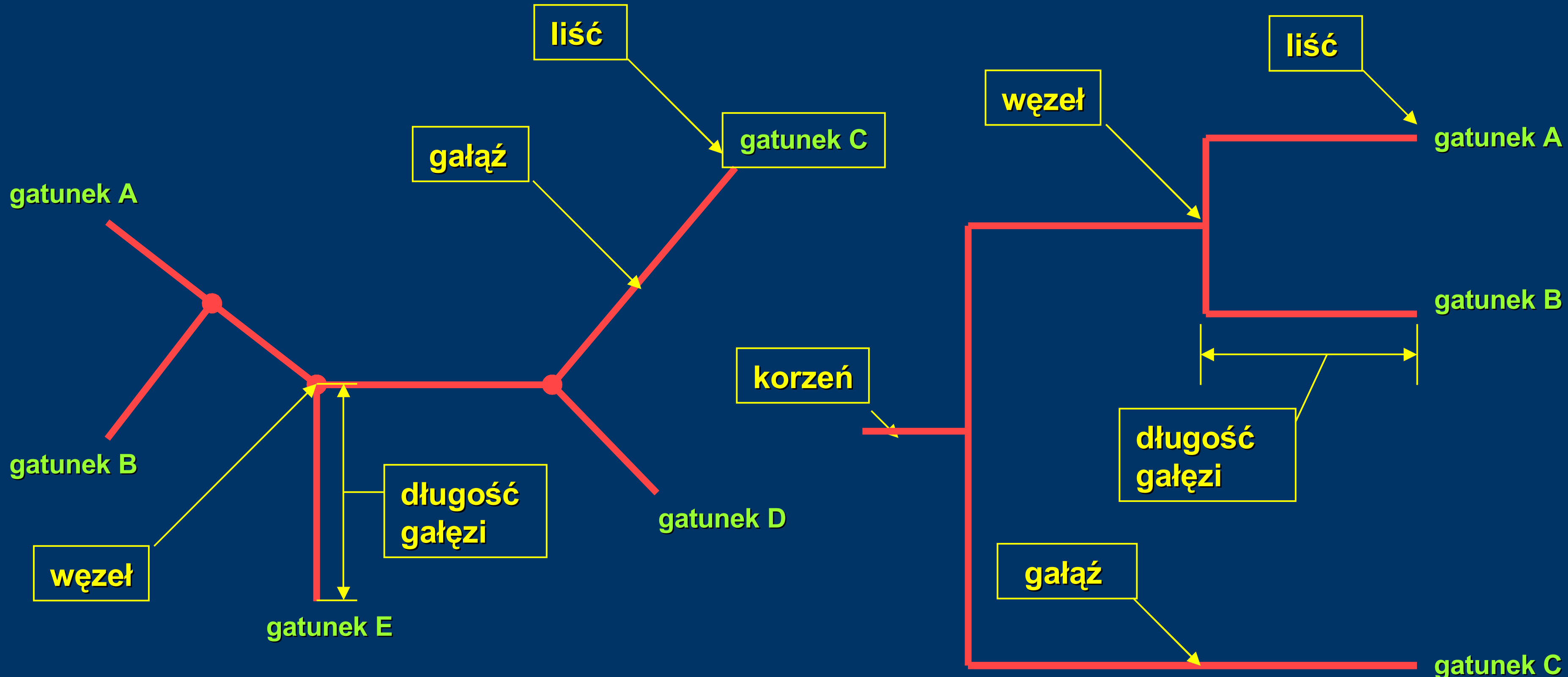
Liść - reprezentuje aktualnie analizowaną jednostkę taksonomiczną.

Gałąź - obrazuje związki ewolucyjne między jednostkami taksonomicznymi.

Długość gałęzi - reprezentuje liczbę zmian, które zaszły w danej linii ewolucyjnej.

Korzeń - wspólny przodek dla wszystkich taksonów.

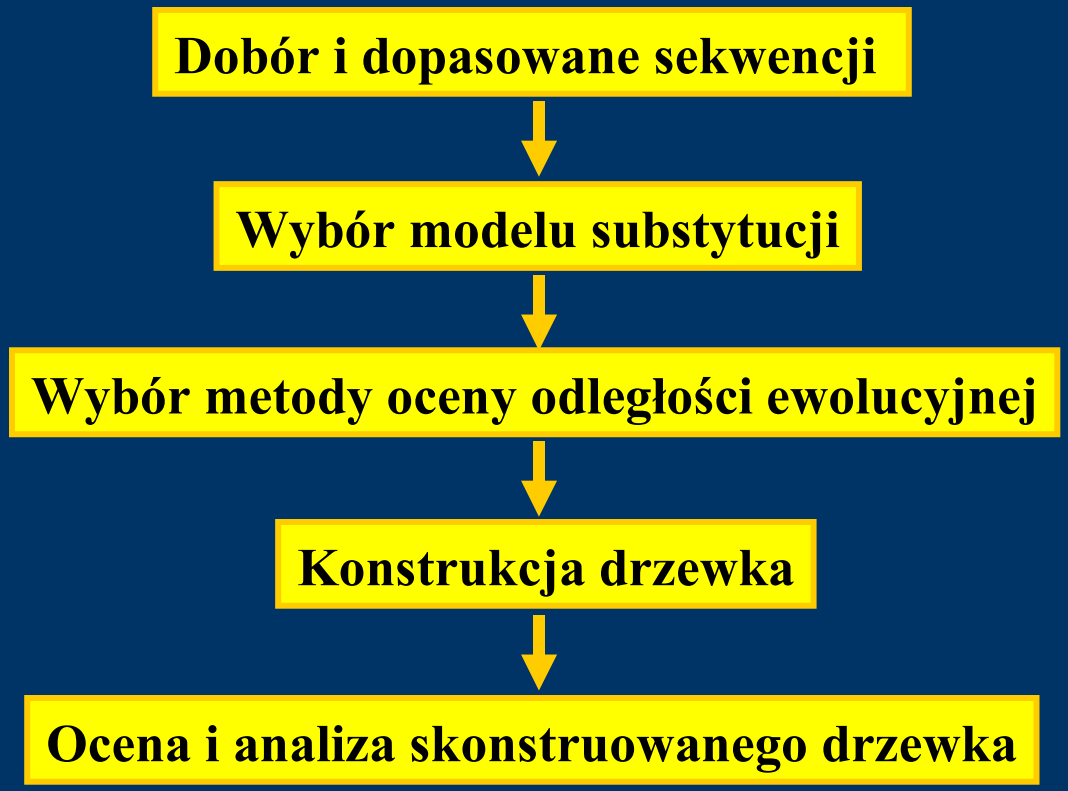
Węzeł - reprezentuje jednostkę taksonomiczną (populację, organizm, gen).



przykładowe *nieukorzenione* drzewo filogenetyczne

przykładowe *ukorzenione* drzewo filogenetyczne

Etapy analizy filogenetycznej



typy mutacji punktowych

substytucja

ACC TAT TTG CTG

↓

ACC TCT TTG CTG

insercja

ACC TAT TTG CTG

↓

ACC TAC TTT GCT G

delecja

ACC TAT TTG CTG

↓

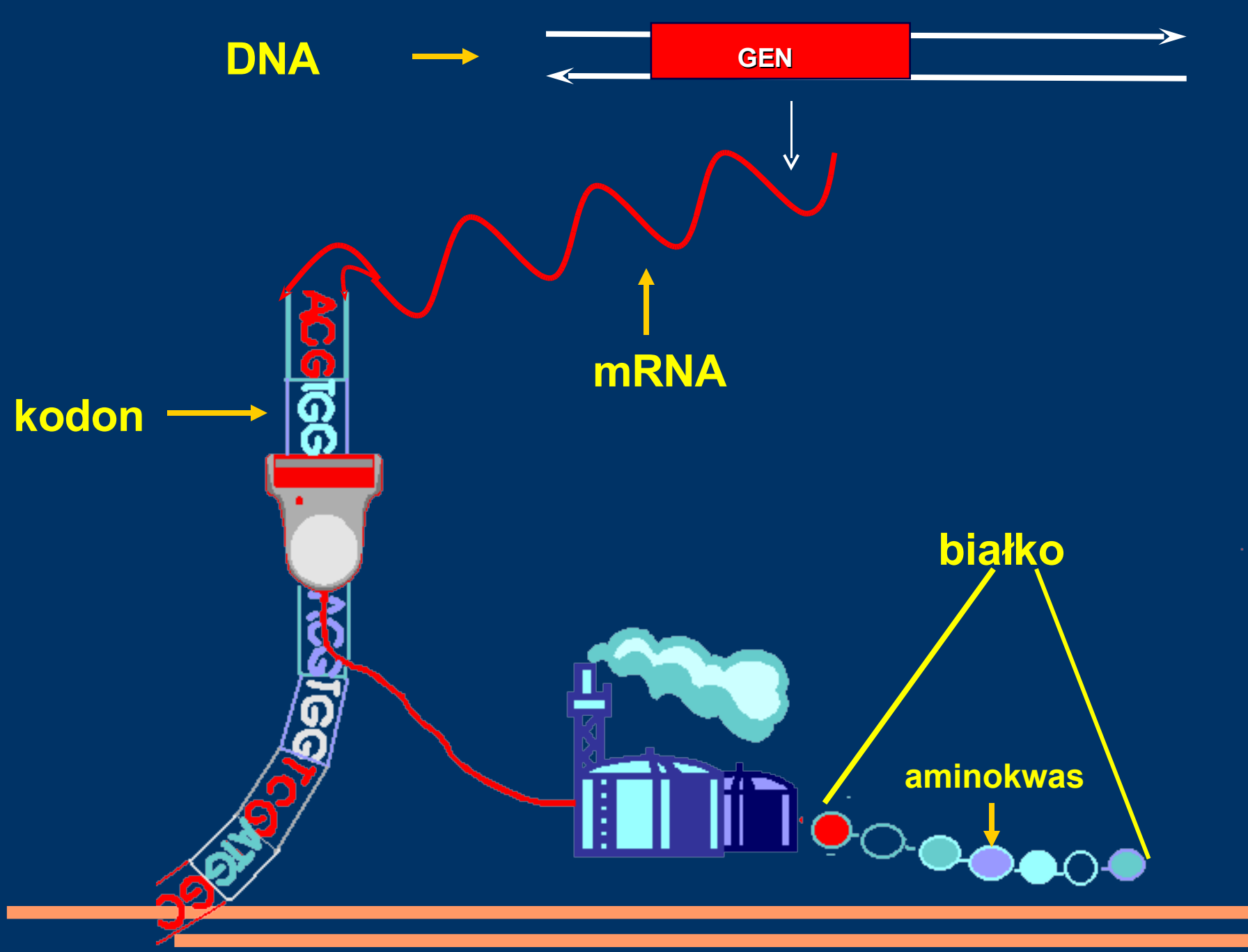
ACC TAT TGC TG-

inwersja

ACC TAT TTG CTG

↓

ACC TTT ATG CTG


DNA
GEN
mRNA
kodon
ACGTGG
ACGTGGTCGAGGC
białko
aminokwas

substytucja

Thr Tyr Leu Leu

ACC TAT TTG CTG

ACC TCT TTG CTG

Thr Tyr Leu Leu

insercja

Thr Tyr Leu Leu

ACC TAT TTG CTG

ACC TAC TTT GCT G

Thr Tyr Phe Ala

delecja

Thr Tyr Leu Leu

ACC TAT TTG CTG

ACC TAT TGC TG-

Thr Tyr Cys

inwersja

Thr Tyr Leu Leu

ACC TAT TTG CTG

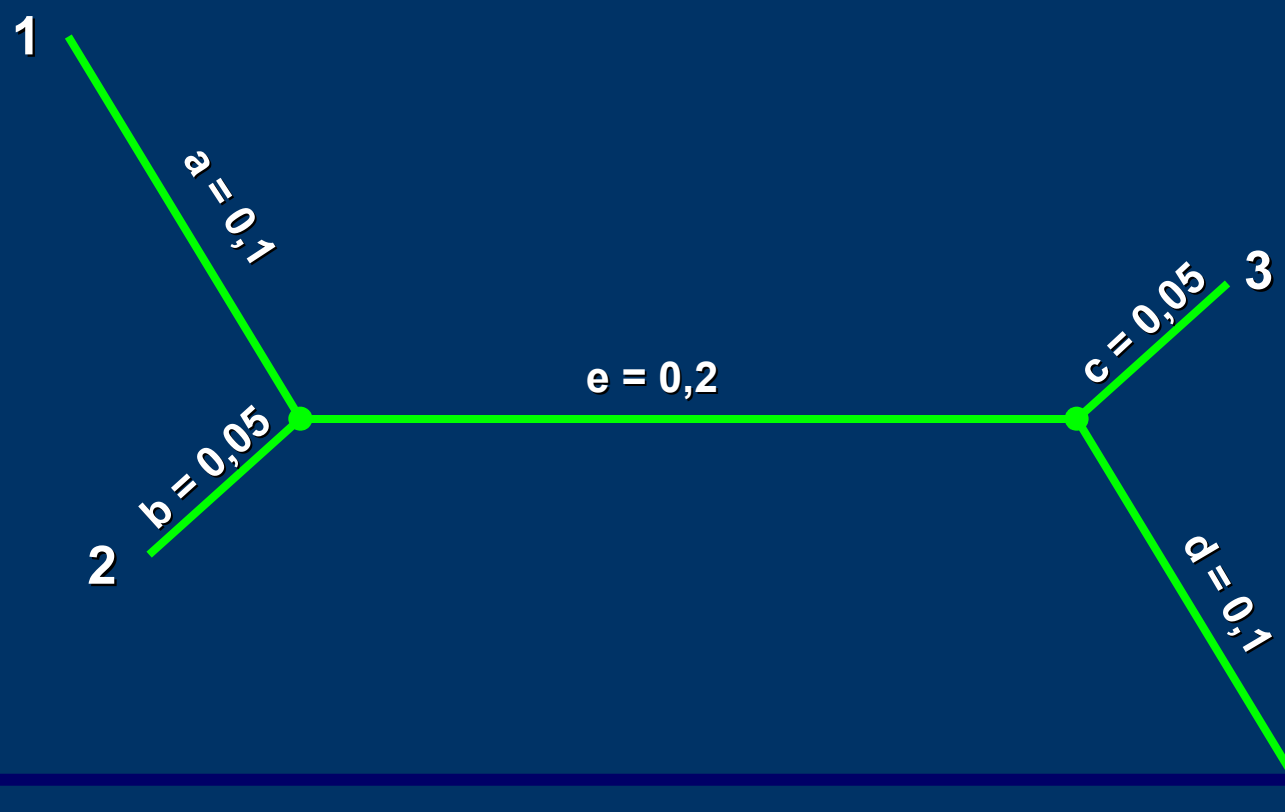
ACC TTT ATG CTG

Thr Phe Met Leu

sek.1	A	G	D	A	E	R	G	K	K	L	F	E	S	R	A	A	Q	C	S	A
sek.2	A	G	D	A	E	R	G	K	K	L	F	E	S	S	A	A	R	C	S	C
sek.3	A	G	D	A	N	R	G	K	I	I	M	E	S	R	A	N	R	C	S	C
sek.4	A	G	N	A	N	R	G	K	I	L	M	E	S	R	S	N	R	C	S	C

20

	1	2	3	4
1	-	3/20 = 0,15	7/20 = 0,35	8/20 = 0,4
2		-	6/20 = 0,3	7/20 = 0,35
3			-	3/20 = 0,15
4				-



przodek

MELSKLTGDPAREKELKMLMELSKLTGDPAPFVYRVLKRL

2 zmiany

2 zmiany w stosunku do przodka MELSKTTGDPARRKELKMLMELSKLTGDPAPFVYRVLKRL

3 zmiany

5 zmian w stosunku do przodka MELSKTTGDPARRKELSMLMKLSKLTGDPAPFVYRVGKRL

2 zmiany

6 zmian w stosunku do przodka MELSKTTGDPARQKELSMLMKLSKLTGDPAPFYYRVGKRL

2 zmiana

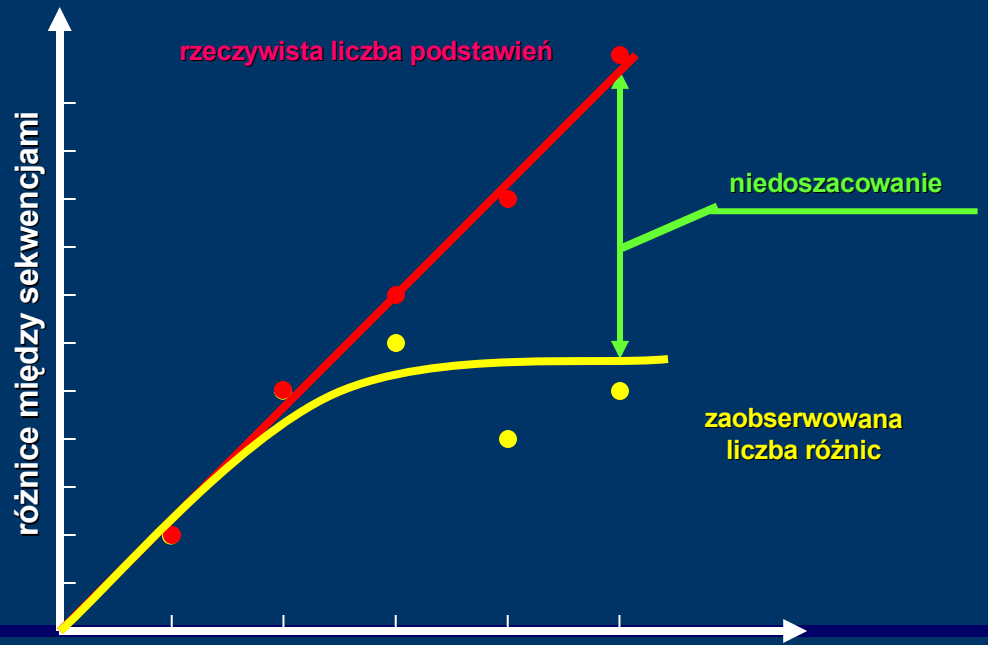
4 zmian w stosunku do przodka MELSKLTGDPARQKELSMLMKLSKLTGDPAPFVYRVGKRL

3 zmiany

5 zmian w stosunku do przodka MELSKLTGDPARQKELSMLWKLSKLTGDRAPFVYRVLKRL

= 12 zmian

potomek



- możliwość wystąpienia wielokrotnych podstawień
- rewersja
- częściej obserwuje się podstawienia między aminokwasami podobnymi do siebie, ze względu na swoje właściwości biochemiczne, biofizyczne np.:
  - izoleucyna *(I)* ↔ leucyna *(L)*,
  - walina *(V)* ↔ izoleucyna *(I)*,
  - kwas asparaginowy *(D)* ↔ kwas glutaminowy (*E)*,
- rzadko obserwuje się podstawienia między aminokwasami bardzo różniącymi się swoimi własnościami
  - tryptofan *(W)* ↔ izoleucyna *(I)*
- rzadko obserwuje się podstawienia między aminokwasami pełniącymi ważne role w białkach, jak: cysteina *(C)* czy tryptofan *(W)*
- niektóre aminokwasy, takie jak: asparagina *(N)*, kwas asparaginowy *(D)*, seryna *(S)* mutują częściej niż inne

Jak więc możemy obliczyć rzeczywistą liczbę podstawień?

Macierze PAMs (Percent Accepted Mutations)

- skonstruowane zostały po raz pierwszy przez M. Dayhoff w 1978 roku
- wykorzystano 71 grup blisko spokrewnionych białek, zaobserwowano 1572 zmiany
- sekwencje w grupach były przynajmniej w 85% identyczne
- uwzględniono mutabilność i częstość wystepowania danego aminokwasu
- określają prawdopodobieństwo przejścia jednego aminokwasu w drugi

Percent Accepted Mutation PAM1 - M. Dayhoff 1978r.

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0																
Asn N	4	1	9822	36																
Asp D	6	0	42	9859																
Cys C	1	1	0	0	99															
Gln Q	3	9	4																	
Glu E	10	0	7	56																
Gly G	21	1	12	11																
His H	1	8	18	3																
Ile I	2	2	3	1																
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13													1	9926	12	4	0	0	2
Ser S	28	1												3	17	9840	38	5	2	2
Thr T	22													1	5	32	9871	0	2	9
Trp W	0													1	0	1	0	9976	1	0
Tyr T	1													21	0	1	1	2	9945	1
Val V	13													1	3	2	10	0	2	9901

element M_{IJ} tej macierzy reprezentuje prawdopodobieństwo z jakim aminokwas w *kolumnie j* zostanie podstawiony przez aminokwas z *wiersza i* w czasie ewolucyjnym 1 PAM

element diagonalny M_{ii} określa prawdopodobieństwo, że dany aminokwas nie ulegnie substytucji w tym czasie

Elementy pomnożone zostały przez 10 000

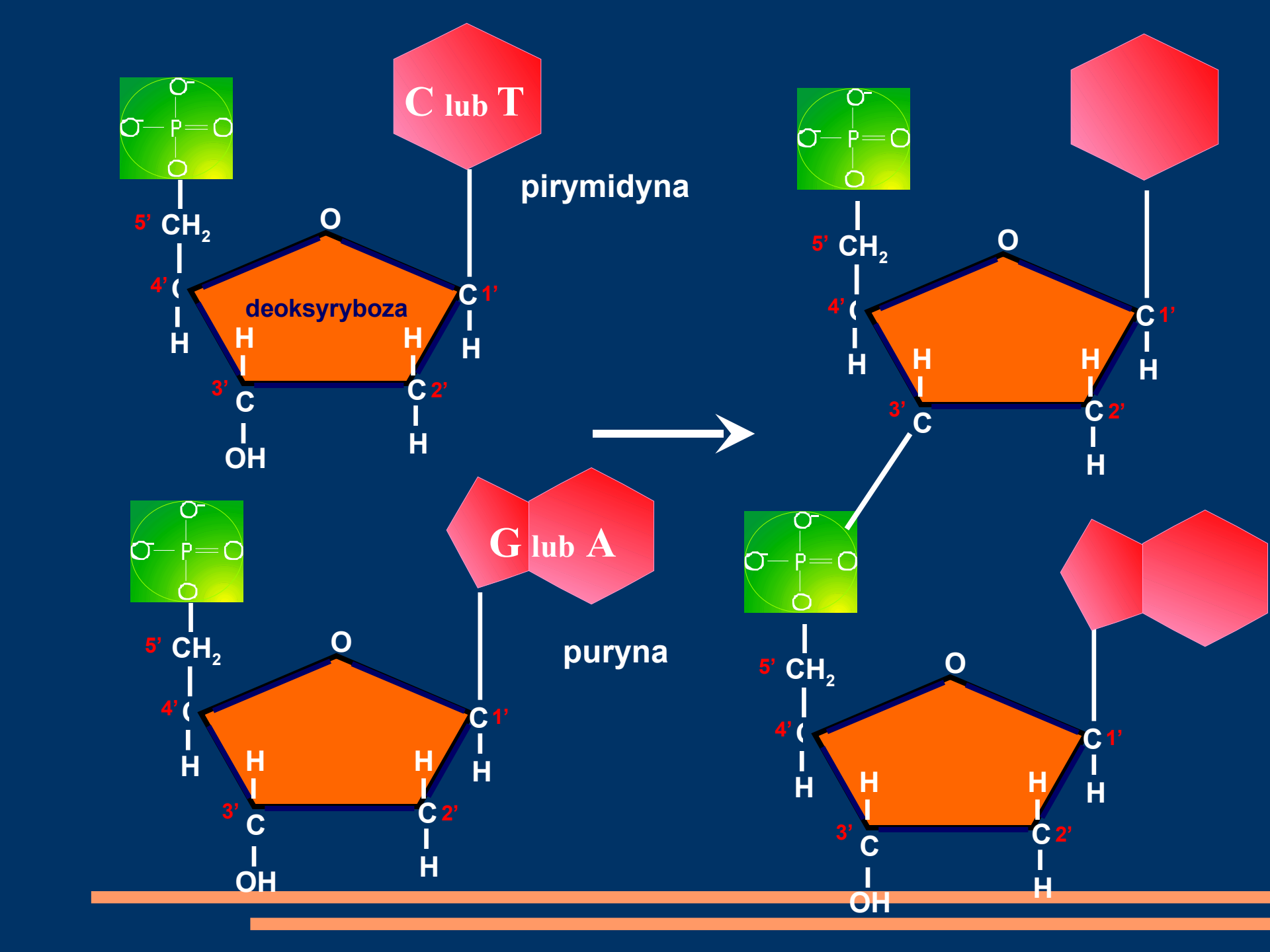
M. Dayhoff i współpracownicy – 1978r.

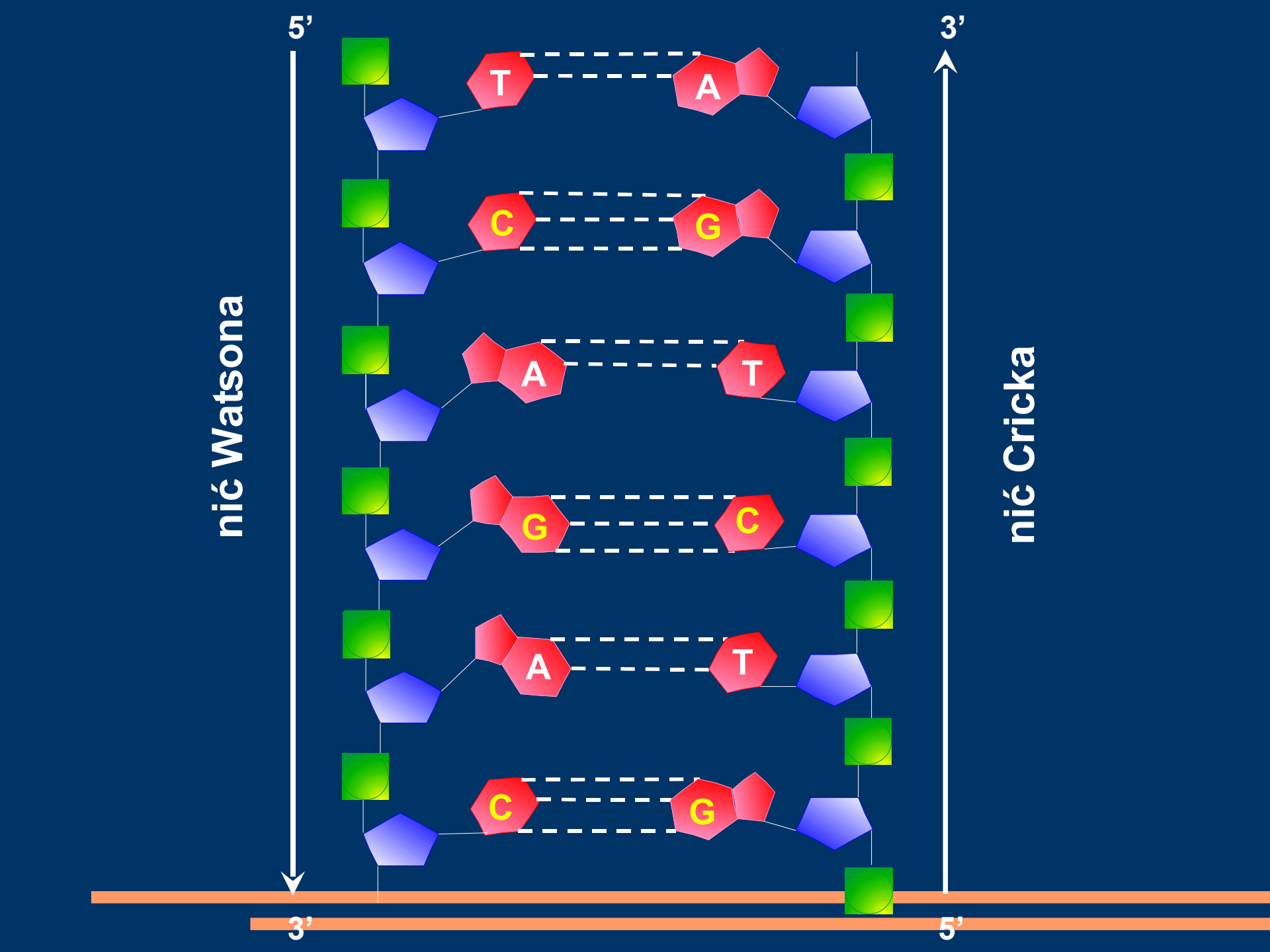
JEDNOSTKA PAM (*Percent Accepted Mutation*)

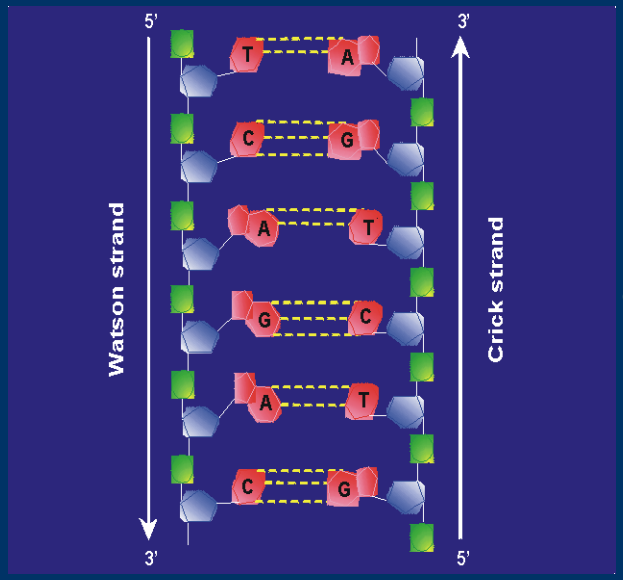
– miara odległości ewolucyjnej między sekwencjami.

1 PAM – odpowiada takiemu czasowi ewolucyjnemu, podczas którego, w porównywanych sekwencjach, zmianie ulegnie 1 aminokwas na 100.

Zmianie uległo 10/1000 = 1/100 aminokwasów, czyli 1%


C lub T
pirymidyna
deoksyryboza
G lub A
puryna


5’
3’
nić Watsona
nić Cricka
T
A
C
G
A
T
G
C
A
T
C
G
3’
5’

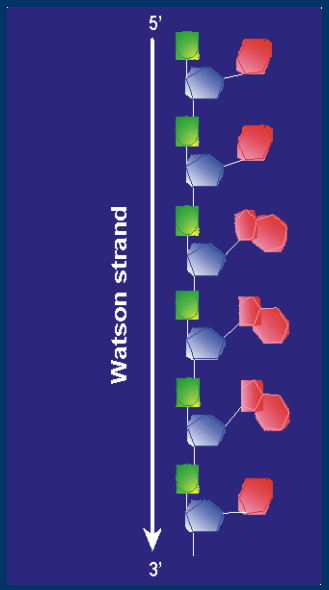


Asymetria w genomach prokariotycznych

dwuniciowa cząsteczka DNA – zasada komplementarności

[G] = [C] i [A] = [T]

PR1 rule – pierwsza reguła parowania



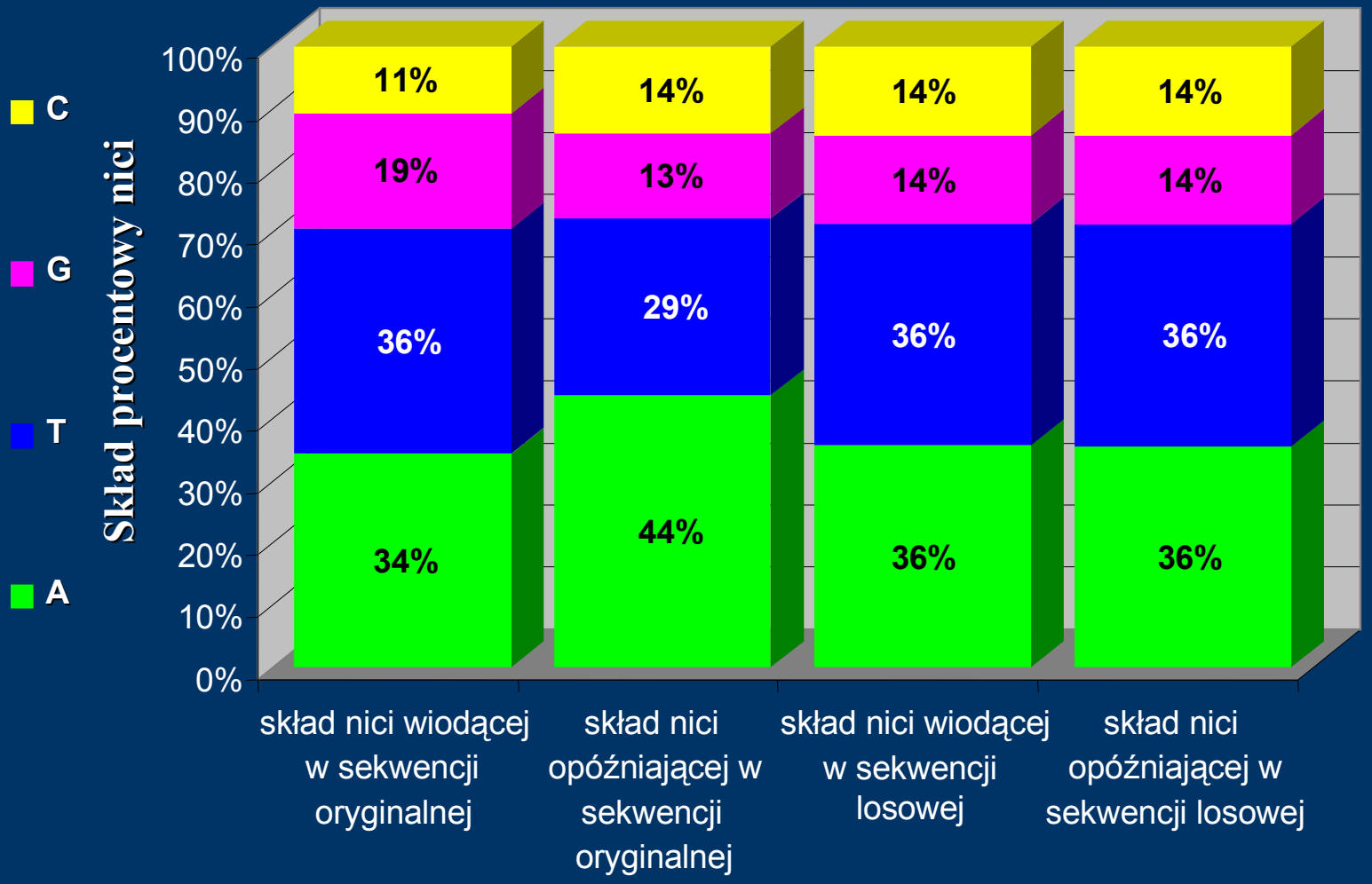
losowa pojedyncza nić DNA

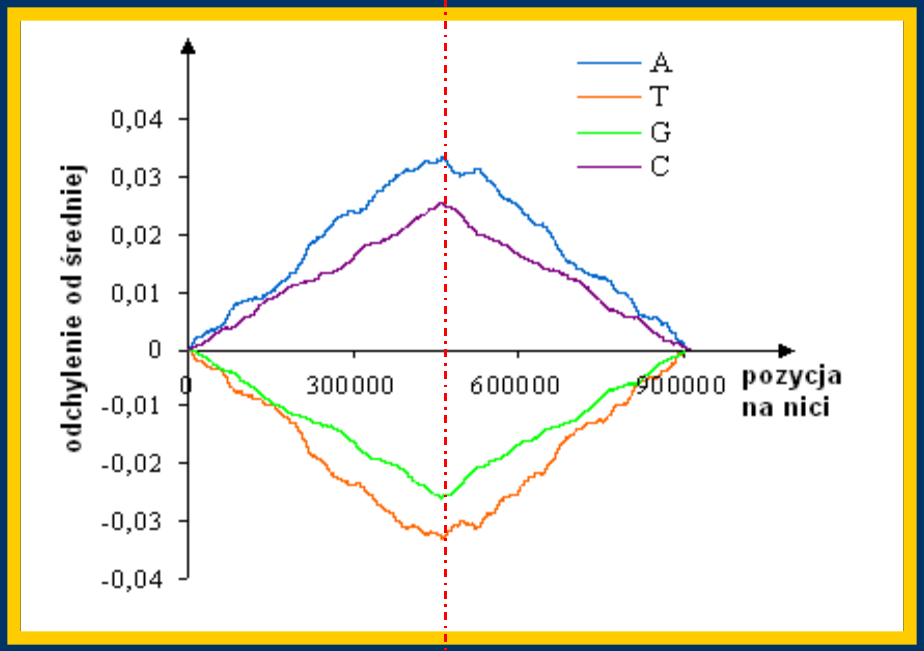
[G] = [C] i [A] = [T]

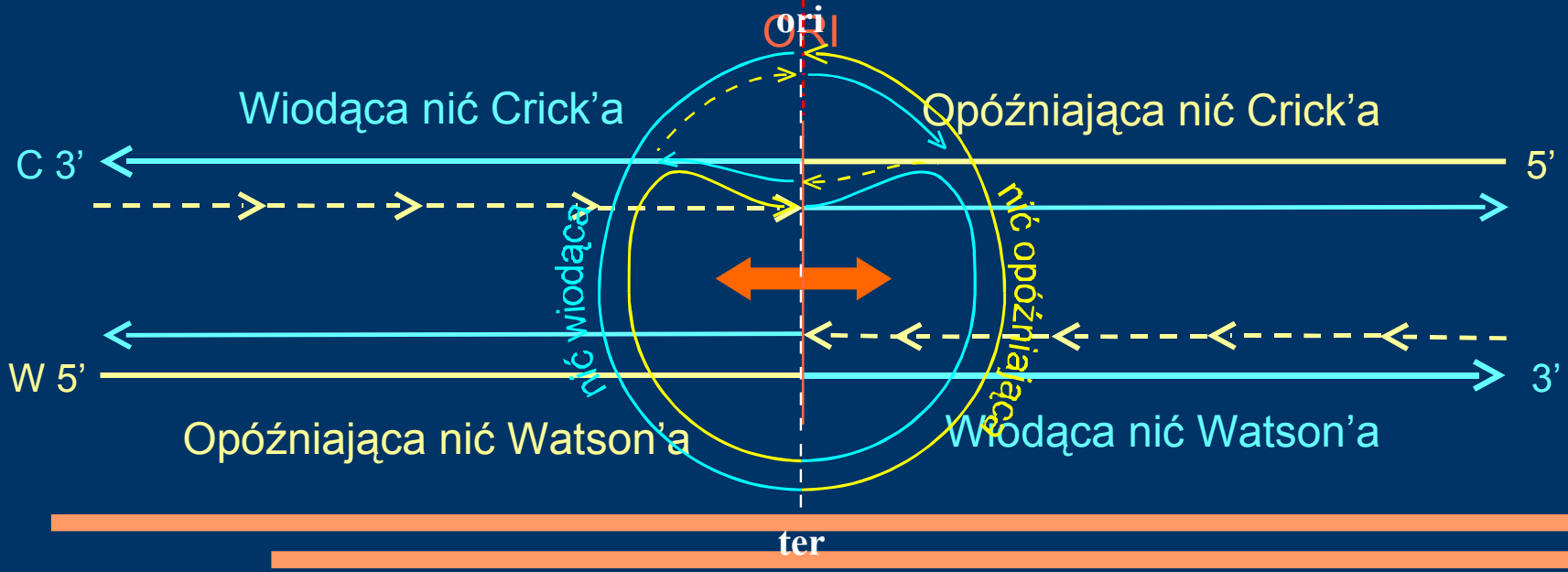
PR2 rule – druga reguła parowania

Asymetrią nazywamy odchylenie od drugiej reguły parowania

Asymetria DNA




A
T
G
C
0,04
0,03
0,02
0,01
0
-0,01
-0,02
-0,03
-0,04
odchylenie od średniej
0
300000
600000
900000
pozycja na nici


ori
Wiodąca nić Crick'a
Opóźniająca nić Crick'a
C 3'
5'
nić wiodąca
nić opóźniająca
W 5'
3'
Opóźniająca nić Watson'a
Wiodąca nić Watson'a
ter

Przykładowa sekwencja

MELSKLTGDPAPFVY

15 aminokwasów

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr T	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

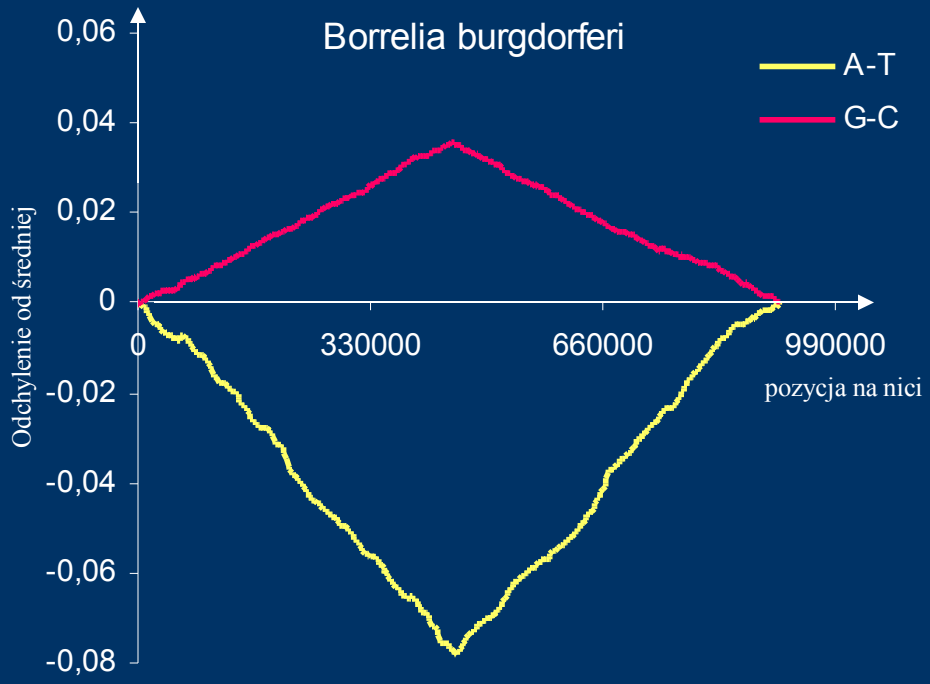
Macierz PAM1 pomnożona przez 10 000

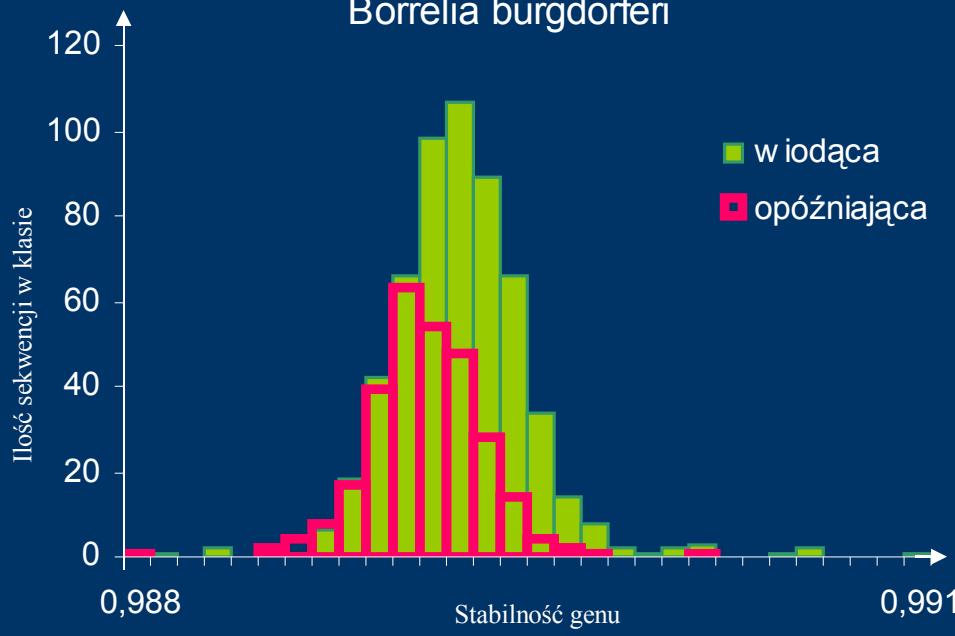
Suma przeżywalności aminokwasów (pomnożona przez 10 000):

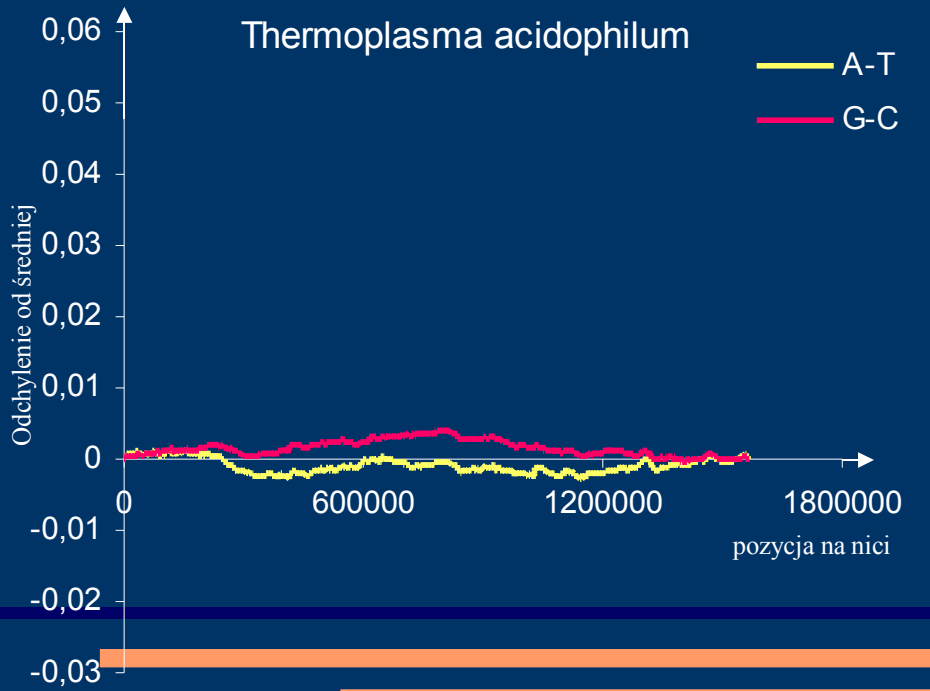
$$9874 + 9865 + 9947 + 9840 + 9926 + 9947 + 9871 + 9935 + 9859 + 9926 + 9867 + 9926 + 9946 + 9901 + 9945 = 148575$$

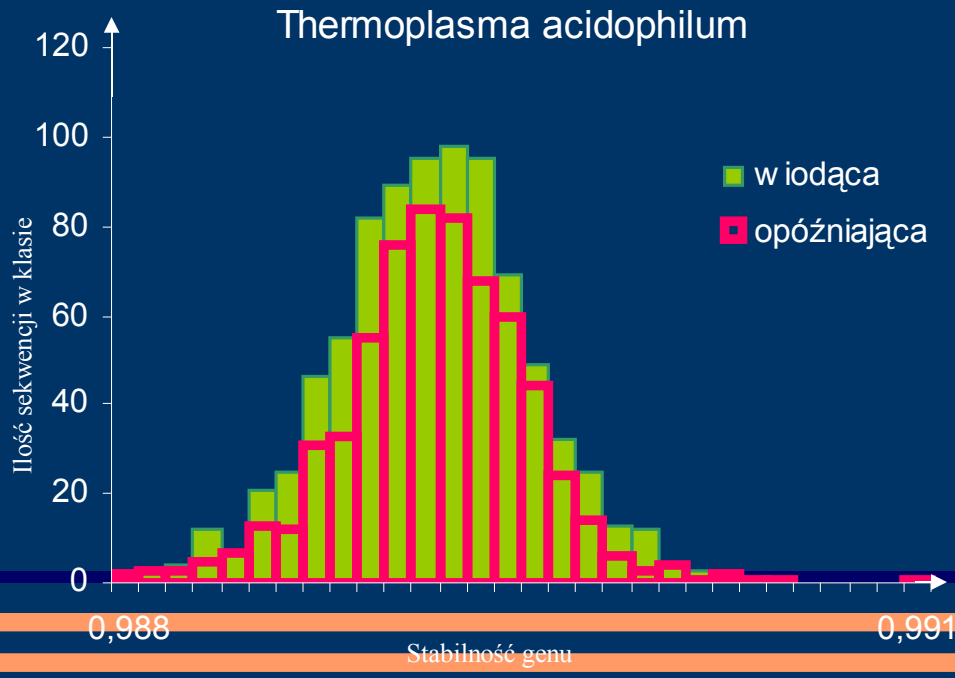
Otrzymany wynik dzielimy przez liczbę aminokwasów, w celu uzyskania średniego prawdopodobieństwa (tu dodatkowo dzielimy przez 10000, aby uzyskać rzeczywiste prawdopodobieństwo):

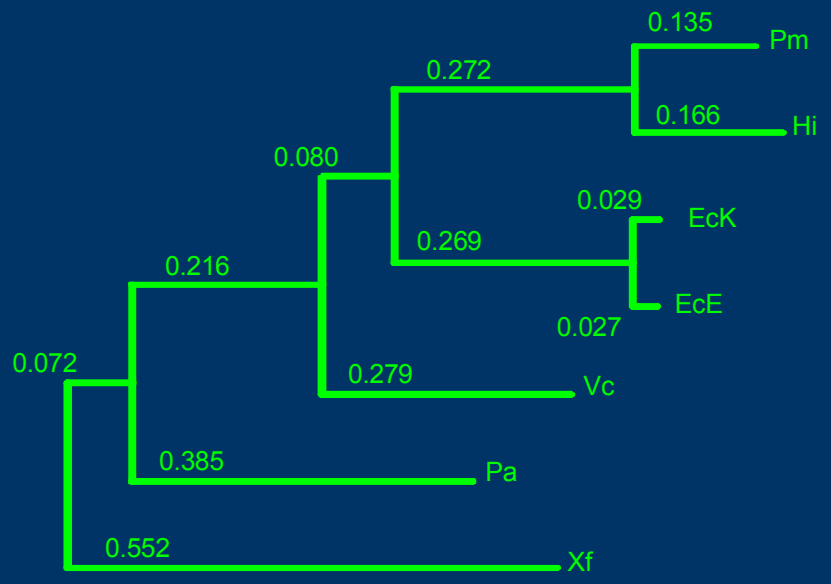
$$\frac{148\,575}{15 \times 10000} = 0{,}9905 = 99{,}05\ \%$$


Borrelia burgdorferi
A-T
G-C
0,06
0,04
0,02
0
-0,02
-0,04
-0,06
-0,08
Odchylenie od średniej
0
330000
660000
990000
pozycja na nici

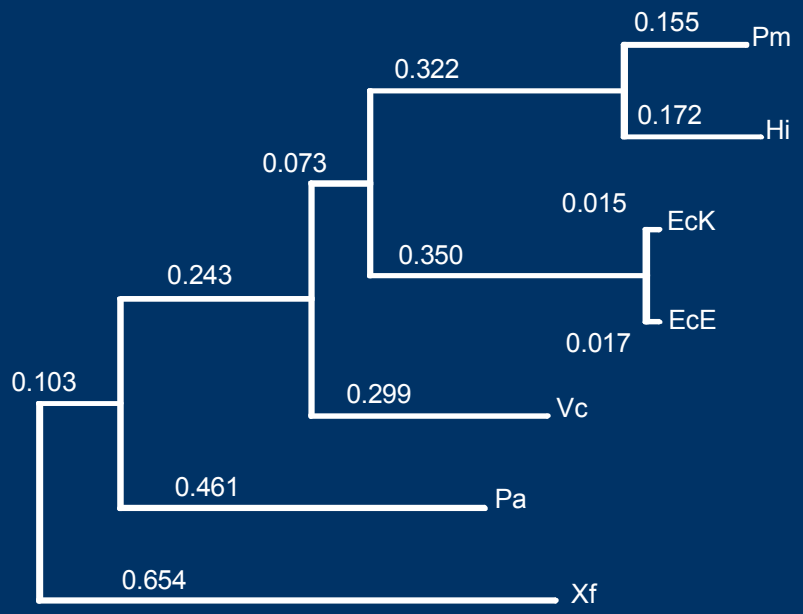

Borrelia burgdorferi
w iodąca
opóźniająca
120
100
80
60
40
20
0
Ilość sekwencji w klasie
0,988
Stabilność genu
0,991


Thermoplasma acidophilum
A-T
G-C
0,06
0,05
0,04
0,03
0,02
0,01
0
-0,01
-0,02
-0,03
Odchylenie od średniej
0
600000
1200000
1800000
pozycja na nici

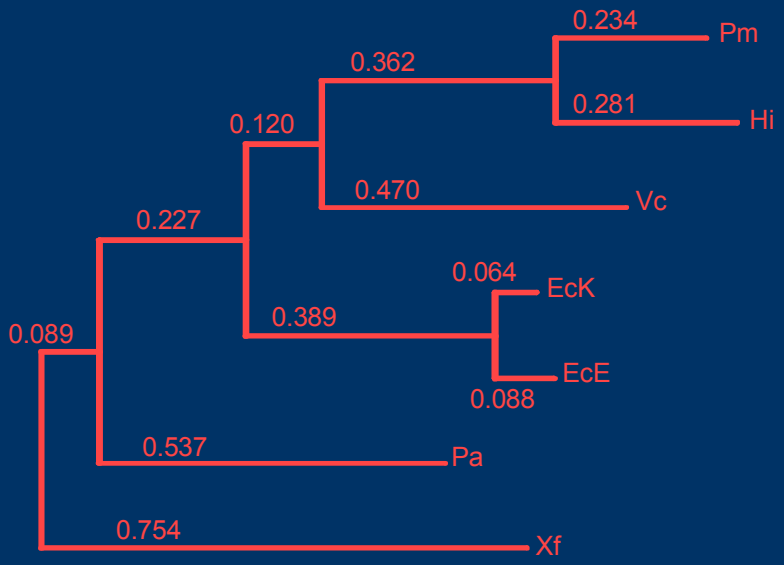

Thermoplasma acidophilum
w iodąca
opóźniająca
120
100
80
60
40
20
0
Ilość sekwencji w klasie
0,988
Stabilność genu
0,991



drzewo dla sekwencji z nici wiodącej

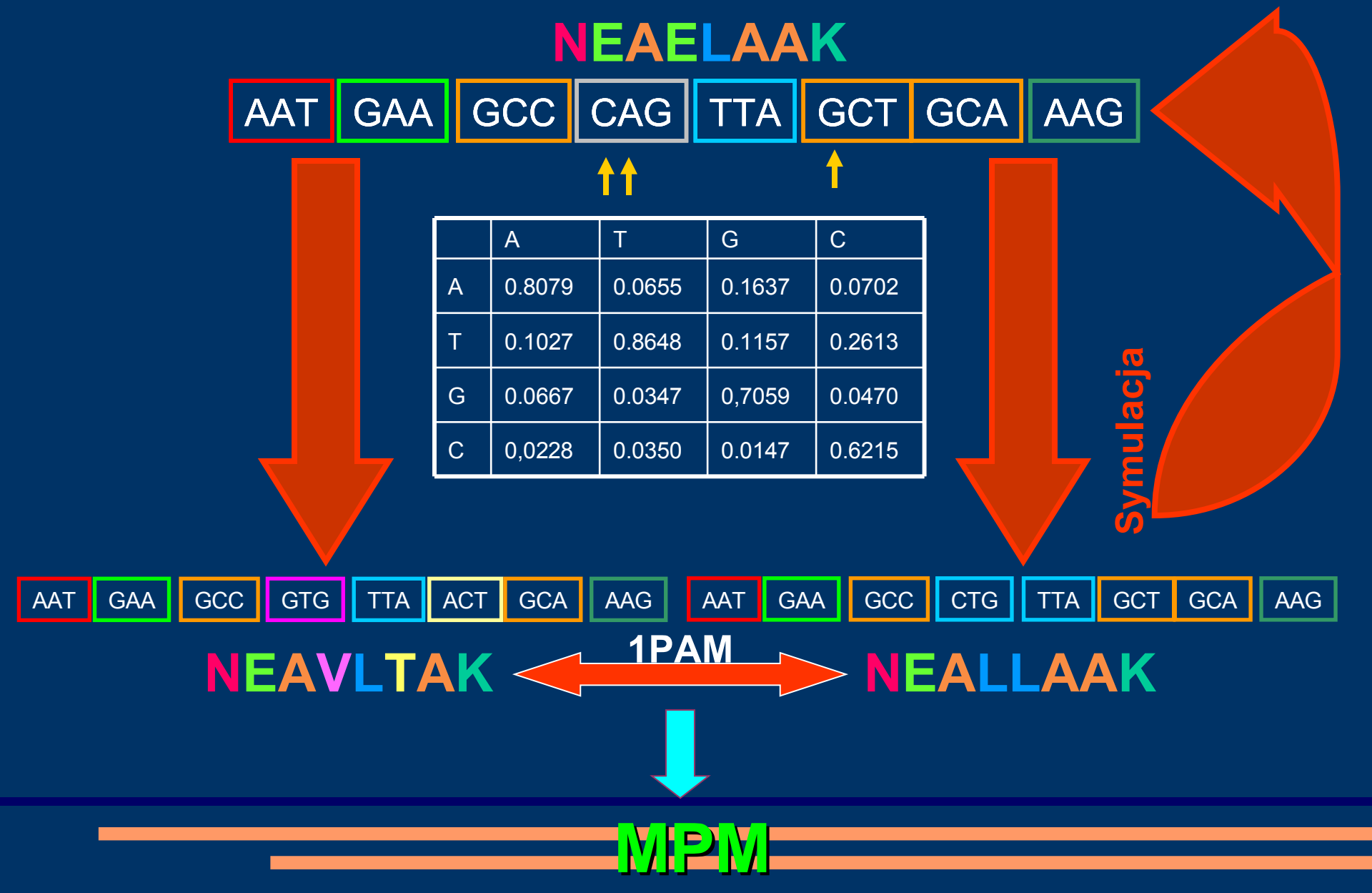


drzewo dla sekwencji z nici opóźniającej



drzewo dla sekwencji z obu nici

TWORZENIE Mutation Pressure Matrix



	A	T	G	C
A	0.8079	0.0655	0.1637	0.0702
T	0.1027	0.8648	0.1157	0.2613
G	0.0667	0.0347	0,7059	0.0470
C	0,0228	0.0350	0.0147	0.6215

MPM - *M*utation *P*ressure *M*atrix

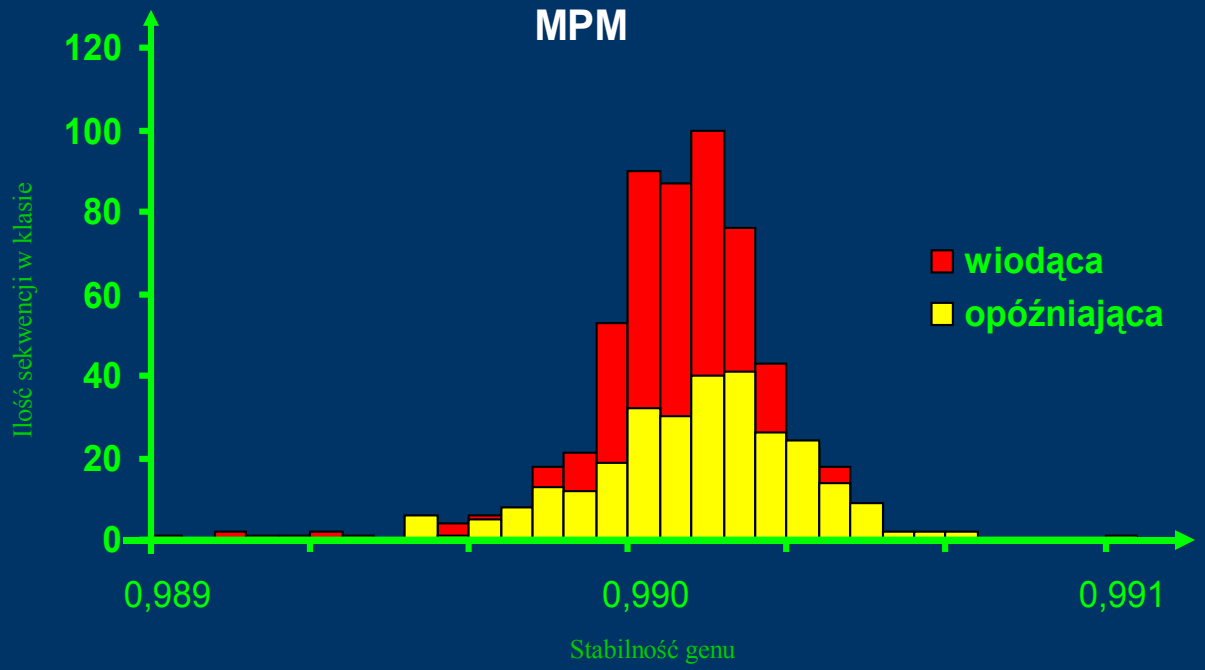
		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala	A	99027	0,16	0,18	59,94	0,56	0,07	44,17	55,93	0,15	0,48	0,22	0,11	0,2	0,33	79,41	93,79	303,9	0,14	0,19	229,7
Arg	R	0,13	98784	0,62	0,06	1															
Asn	N	0,26	1,14	98925	255,2																
Asp	D	76,9	0,09	220	98935																
Cys	C	0,092	33,48	0,28	0,3	97															
Gln	Q	0,03	50,97	0,17																	
Glu	E	63,05	1,06	0,84	258,1																
Gly	G	68,86	227,2	0,47																	
His	H	0,04	37,94	36,72	25,37																
Ile	I	1,03	103,2	180,5	0,62																
Leu	L	0,51	60,76	0,12	0,08							99267									
Lys	K	0,23	325,1	294,6	0,99	0,01	184,4	276,2	0,65	0,98	94,85	0,31	99245	230,7	0,002	0,1	0,56	97,39	1,32	0,83	0,53
Met	M	0,08	46,66												0,2	0,04	0,13	38	0,87	0,001	37,92
Phe	F	0,5	0,23													1,21	153,1	0,74	1,01	221,3	122,7
Pro	P	41,9	34,7													98948	107,4	77,27	0,09	0,18	0,14
Ser	S	165,6	218,7													359,4	98951	261	46,97	85,4	0,84
Thr	T	224,7	31,3	46,37	0,21	0,56	0,12	0,16	0,19	0,22	126,7	0,27	34,91	68,87	0,36	108,3	109,2	98724	0,14	0,2	0,78
Trp	W	0,014	42,64	3E-04	3E-04	127,5	0,11	0,09	23,91	0,001	2E-04	13,48	0,07	0,22	0,07	0,02	2,77	0,019	98827	0,09	0,05
Tyr	T	0,18	0,21	153,2	149,9	580,5	0,91	0,45	0,32	413,8	0,29	0,41	0,38	0,003	136,9	0,32	45,75	0,25	0,84	99034	0,33
Val	V	329	0,55	0,53	161,8	1,43	0,18	131,7	166	0,42	174,8	75,52	0,37	133,2	114,9	0,38	0,68	1,52	0,7	0,5	98802

element M_{IJ} tej macierzy reprezentuje prawdopodobieństwo z jakim aminokwas w *kolumnie j* zostanie podstawiony przez aminokwas z *wiersza i* w czasie ewolucyjnym 1 PAM, gdy sekwencje będą poddane działaniu tylko presji mutacyjnej

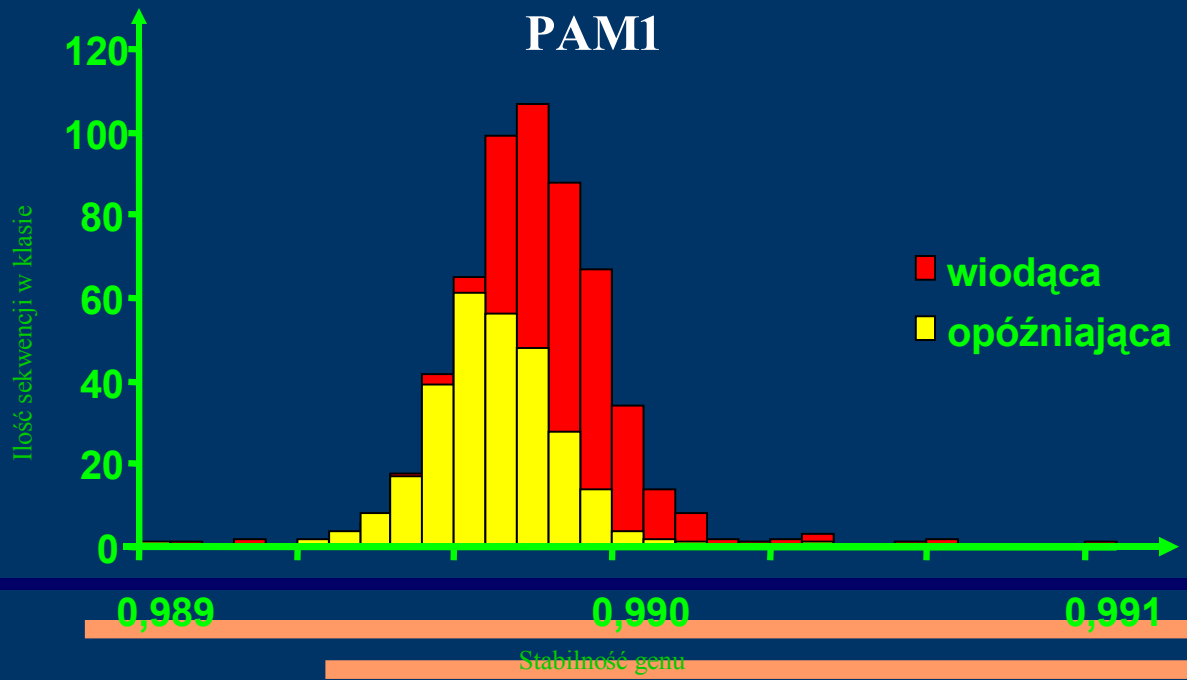
element diagonalny M_{ii} określa prawdopodobieństwo, że dany aminokwas nie ulegnie substytucji

Elementy pomnożone zostały przez 10 000

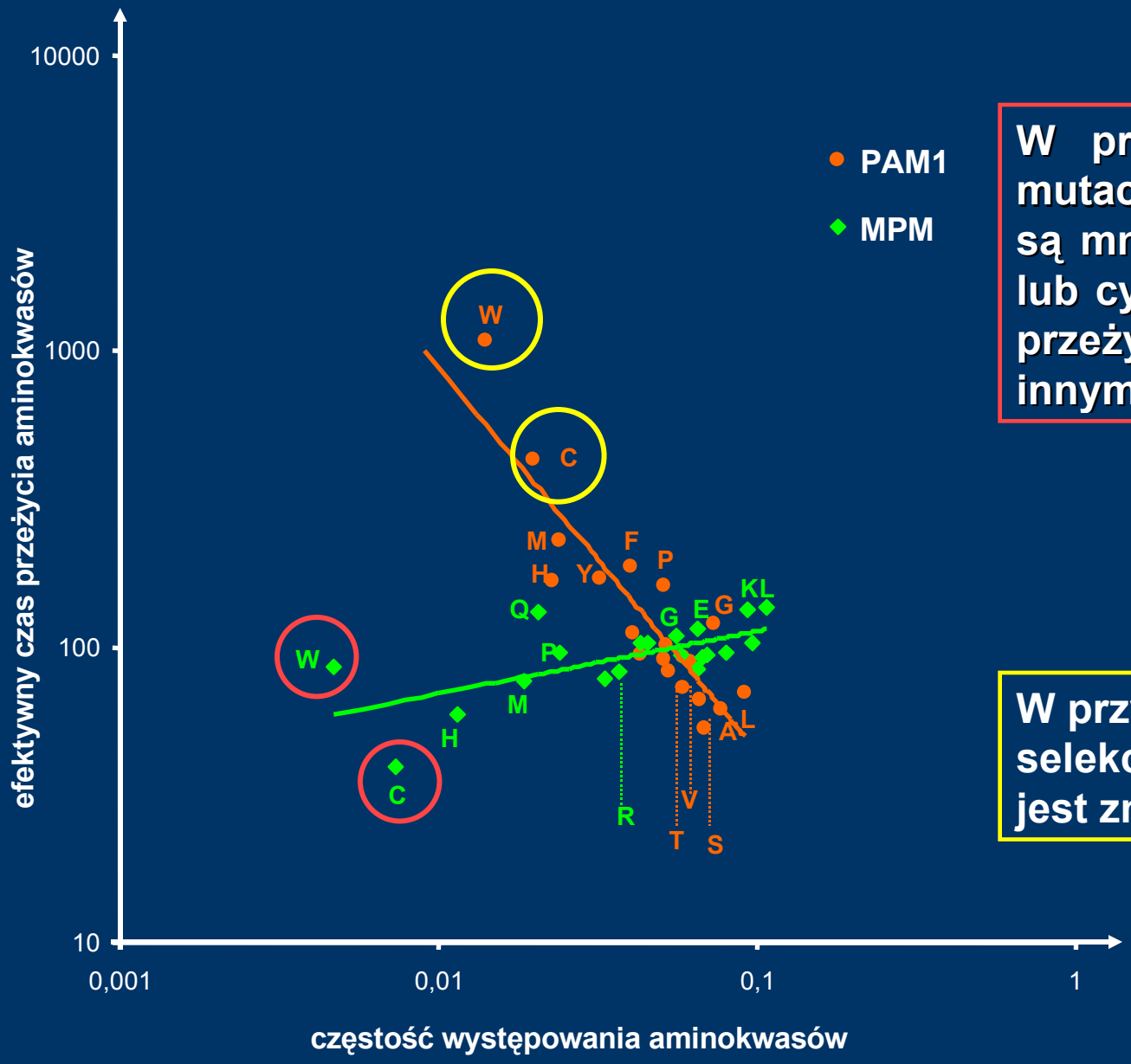
Uzyskana tablica reprezentuje czystą presję mutacyjną.



Geny z nici opóźniającej i wiodącej mają średnio zbliżone prawdopo-dobieństwo przeżycia.

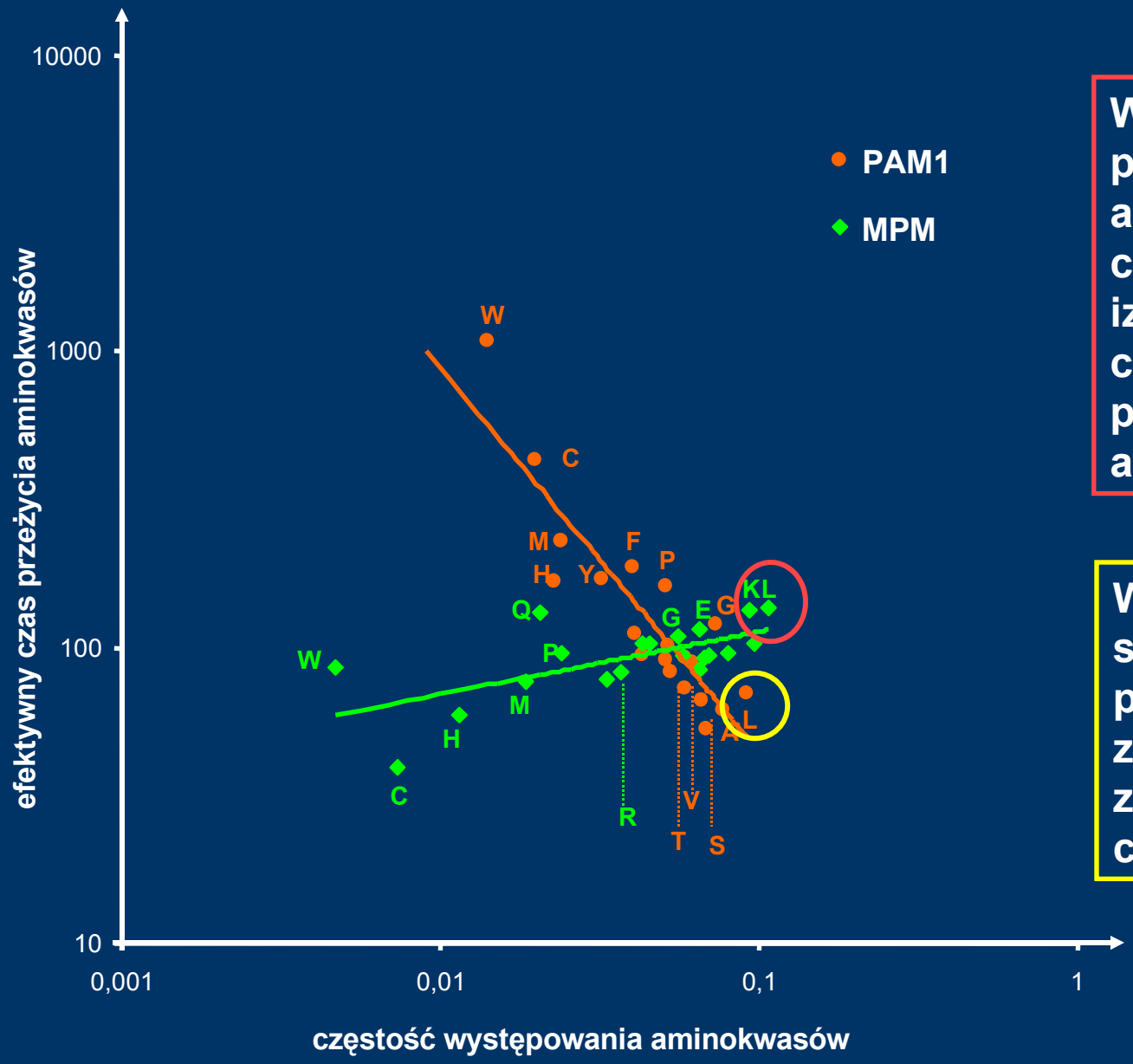


Geny z nici wiodącej mają średnio większe prawdopodobieństwo przeżycia niż geny z nici przeciwnej.



W przypadku czystej presji mutacyjnej aminokwasy, które są mniej częste, jak tryptofan lub cysteina, mają krótki czas przeżycia, w porównaniu z innymi aminokwasami.

W przypadku presji selekcyjnej czas ich przeżycia jest znacznie dłuższy.



W przypadku czystej presji mutacyjnej aminokwasy, które są częste, jak leucyna lub izoleucyna, mają długi czas przeżycia w porównaniu z innymi aminokwasami.

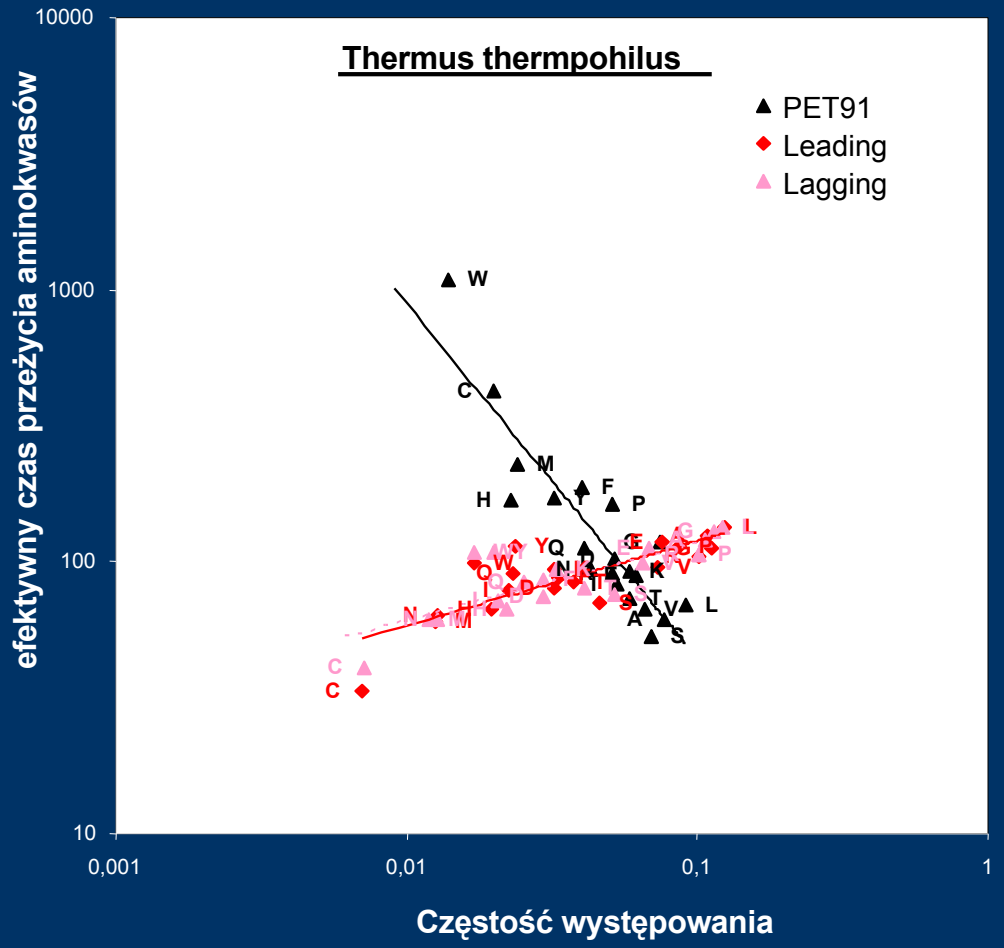
W przypadku presji selekcyjnej czas ich przeżycia jest znacznie krótszy – są znacznie słabiej chronione.

Bakterie termofilne

	A:4	R:6	N:2	D:2	C:2	Q:2	
A:4	99163	0,73	1,66	183,8	0,76	0,56	…
R:6	0,69	99060	3,42	1,03	723,9	349,2	…
N:2	0,2	0,44	98352	183,2	1,14	0,44	…
D:2	50,39	0,3	420,7	98647	1,25	1,05	…
C:2	0,06	51,16	0,64	0,31	97539	0,01	…
Q:2	0,14	87,63	0,88	0,91	0,03	98808	…
E:2	51,98	1,29	3,45	353	0,01	296,6	…
G:4	147,2	252,6	3,25	366,3	191,1	1,9	…
H:2	0,16	74,95	139,8	127,9	5,39	191,2	…
I:3	0,62	1,63	106,1	0,37	0,38	0,04	…
L:6	0,92	62,63	0,48	0,34	2,87	105,7	…
K:2	0,23	87,1	378,9	1,48	0,01	124,9	…
M:1	0,37	9,59	0,48	0,01	0	0,2	…
F:2	0,14	0,16	0,27	0,11	195,3	0,01	…
P:4	138,6	117,4	0,57	0,57	3,55	117,7	…
S:6	35,01	71,08	314,6	1,55	415,1	0,43	…
T:4	174,8	30,82	163	0,7	0,69	0,26	…
W:1	0,04	89,53	0,01	0,01	314,6	1,07	…
Y:2	0,05	0,38	108,8	49,26	604,5	0,88	…
V:4	235,9	0,29	0,88	82,5	0,34	0,54	…

Thermus thermpohilus

Fig. 5b



WNIOSKI

- Macierze odległościowe PAM mogą wykazać inne odległości filogenetyczne, gdy badana jest niereprezentatywna pula genów. Ma to szczególne znaczenie w przypadku analiz filogenetycznych asymetrycznych genomów bakteryjnych. Ważnym jest aby w badanej puli sekwencji były odpowiednio reprezentowane geny leżące na nici wiodącej i opóźniającej.

- Opracowana metoda analizy stabilności aminokwasów i całych sekwencji kodujących, pozwala na badanie pewnych aspektów ewolucji poszczególnych grup bakterii

- Porównanie wpływu presji mutacyjnej i presji selekcyjnej na poszczególne aminokwasy wykazało, że aminokwasy najmniej stabilne pod presją mutacyjną są najrzadziej wbudowywane w białko ale najbardziej pilnowane przez selekcję.