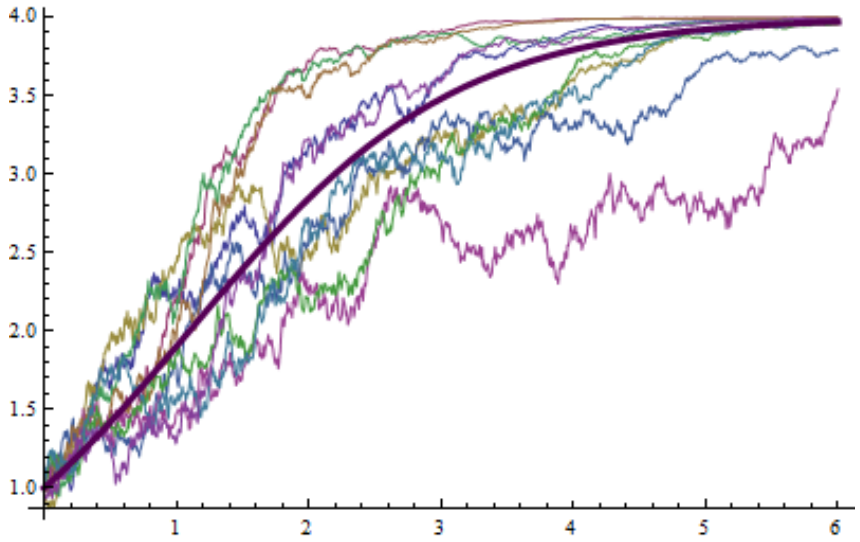


MODELOWANIE STOCHASTYCZNE CZĘŚĆ II - ŁAŃCUCHY MARKOWA

Biomatematyka
Dr Wioleta Drobik-Czwarno

METODA BOOTSTRAP

Sprawdź: https://seeing-theory.brown.edu/frequentist-inference/index.html#section1

Stosowana gdy nie znamy rozkładu z którego pochodzi próba, a jej wielkość jest zbyt mała by stosować metody oparte na prawach wielkich liczb

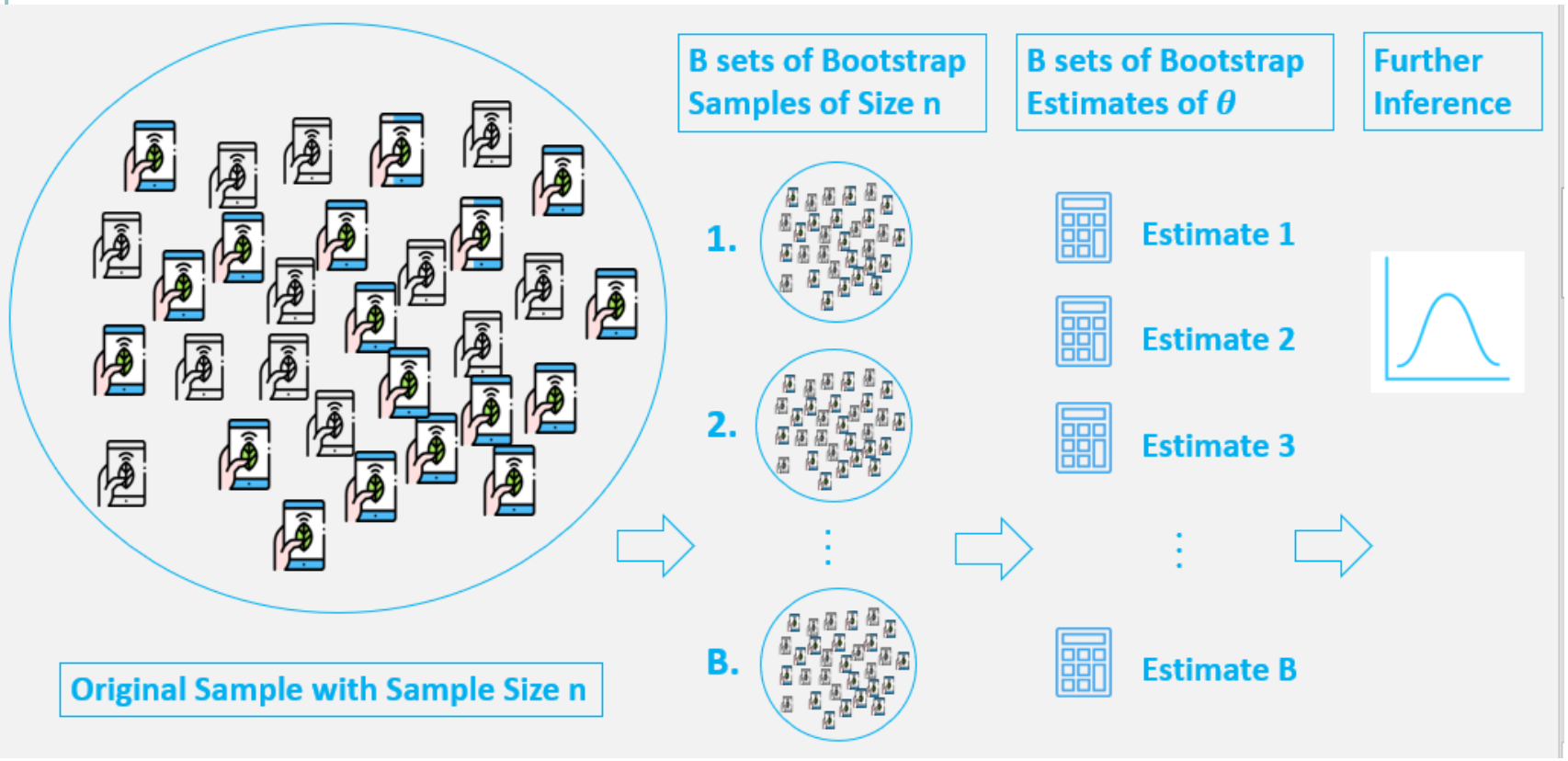
Metoda: Losowanie wielu prób z wyjściowej próbki ze zwracaniem

Algorytm:

1. Wybrać liczbę próbek do wylosowania (N)
2. Wybrać wielkość wylosowanych próbek – zazwyczaj losujemy tyle samo ile liczy bazowa próbka (n)
3. Dla każdej próbki:
   - Wylosuj z próby wyjściowej próbkę o określonej liczebności (n)
   - Oblicz szukaną statystykę dla próbki
4. Oblicz średnią ze statystyki dla wszystkich próbek (N)

Sprawdź: https://seeing-theory.brown.edu/frequentist-inference/index.html#section1

BOOTSTRAP



https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60

METODA BOOTSTRAP W FILOGENETYCE

„… keep all of original species while sampling characters with replacement, under the assumption that the characters have been independently drawn by the systematist and have evolved independently” (Felsenstein J. 1985)

Metoda:

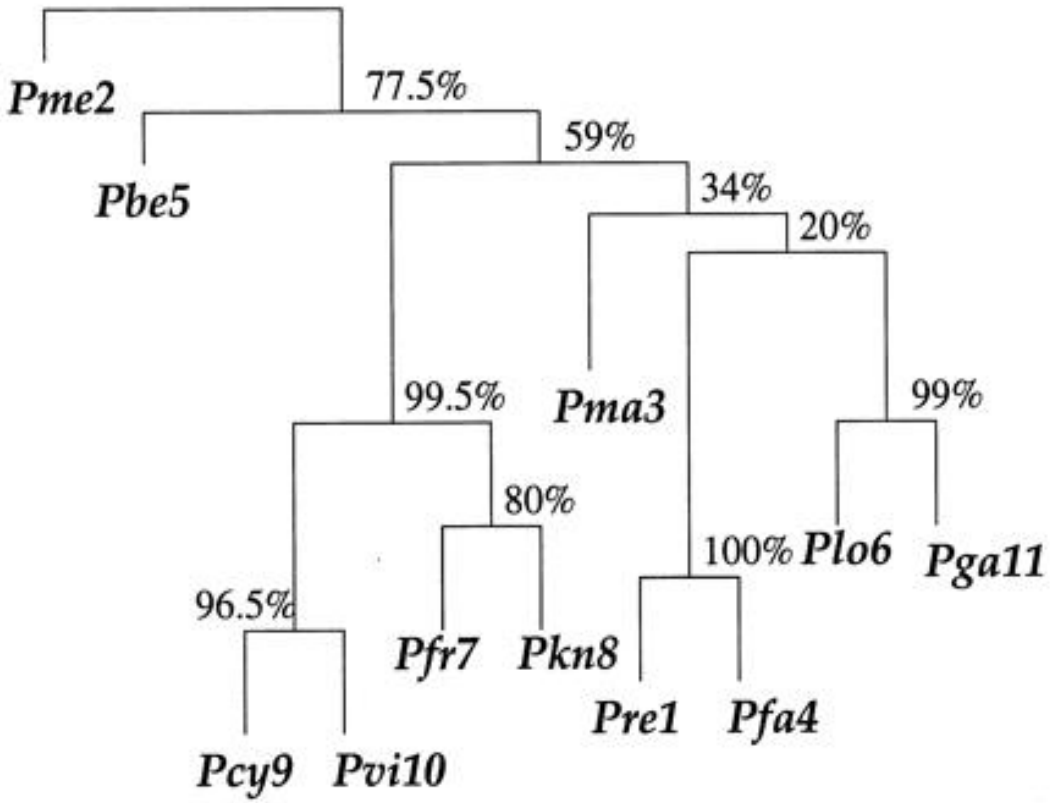
- Losowy wybór kolumn z oryginalnej ramki danych ze zwracaniem – pojedyncza próbka
- Algorytm budowy drzewa jest stosowany do danych
- Proces wyboru kolumn i budowy drzewa jest powtarzany N razy
- Proporcja drzew które są zgodne (topologicznie) z drzewem zbudowanym na danych oryginalnych jest podawana w % jako wynik

METODA BOOTSTRAP

Jak dokładne są drzewa filogenetyczne?

- Metoda opisana w Felsenstein J. 1985. „Confidence Limits on Phylogenies: An Approach Using the Bootstrap". Evolution 39, 783-791

	Site:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Species	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	Pre (Chimp)	C	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
2	Pme (Lizard)	T	C	T	A	A	A	A	G	A	T	T	A	T	A	T	A	G	A	T	A
3	Pma (Human)	T	T	T	A	A	G	G	A	A	A	T	T	C	T	T	A	A	A	T	T
4	Pfa (Human)	T	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
5	Pbe (Rodent)	T	T	T	A	A	G	A	A	A	A	T	T	T	A	T	A	A	A	T	A
6	Plo (Bird)	T	T	T	A	A	G	A	A	A	A	C	T	C	A	C	A	A	A	T	C
7	Pfr (Monkey)	C	T	T	A	A	G	A	A	G	A	T	T	C	T	T	A	G	G	A	A
8	Pkn (Monkey)	C	T	T	A	A	G	A	A	A	G	T	T	C	T	T	A	G	A	T	A
9	Pcy (Monkey)	C	T	C	A	T	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
10	Pv (Human)	C	T	T	A	T	G	A	A	A	A	T	T	C	T	C	G	G	A	T	A
11	Pga(Bird)	T	T	T	A	A	G	A	A	A	A	T	T	T	T	C	A	A	A	T	C



CO OZNACZAJĄ WARTOŚCI PROCENTOWE?

Metoda bootstrap

PROCESY STOCHASTYCZNE

Procesem stochastycznym nazywamy zbiór zmiennych losowych $\{X_t\}_{t\in T}$, które przyjmują wartości w przestrzeni i są indeksowane przez zbiór T

- Np. X_1, X_2, X_3

Prawdopodobieństwa dla różnych wartości zmiennych losowych mogą zależeć od:

- niczego - są niezależne np. proces Bernoulliego, Poissona
- wyników wszystkich poprzednich doświadczeń
- tylko od wyniku poprzedniego doświadczenia - **Procesy Markowa**

PROCES POISSONA

Proces $\{N(t), t \geq 0\}$ nazywamy **procesem zliczającym** jeśli $N(t)$ oznacza całkowitą liczbę badanych zdarzeń zaobserwowanych do chwili t

Proces Poissona jest **procesem zliczającym:**

- **o przyrostach niezależnych** – rozkłady liczby zdarzeń obserwowane w niezależnych przedziałach czasu są niezależne
- **jednorodnym w czasie** (stacjonarnym) – rozkład liczby zdarzeń zaobserwowanych w przedziale czasu zależy wyłącznie od długości tego przedziału

Generuje zmienną losową o rozkładzie Poissona

PROCES POISSONA

Zmienna losowa X ma rozkład Poissona $X \sim Poiss(\lambda)$

Gęstość rozkładu prawdopodobieństwa wyraża się wzorem:

$$P_\mu(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- $\lambda > 0$ – wartość oczekiwana musi być liczbą dodatnią
- $k = 0,1,2,\ldots$ - liczba zdarzeń nie może być ujemna i do tego musi być liczbą całkowitą
- Wartość oczekiwana $EX = \lambda$
- Wariancja $VarX = \lambda$

ROZKŁAD POISSONA

Rzadkie typy komórek

- Np. komórki macierzyste szpiku (HSC ang. Hemapoietic stem cells) stanowią niewielką część populacji wszystkich komórek, załóżmy 1/100,000 czyli 0,00001

- Dlaczego rozkład Poissona?
  - Liczba komórek to liczby całkowite
  - Rzadkie typy komórek są losowo rozmieszczone w próbkach
  - Przy małej próbie prawdopodobieństwo znalezienia rzadkiej komórki spada niemal do zera

ROZKŁAD POISSONA

- Jakie jest prawdopodobieństwo znalezienia dokładnie jednej komórki HSC w próbce 50 000 komórek?
  - Oczekiwana liczba rzadkich komórek w próbie 50 tysięcy.

$$EX = \lambda = 50000 * \frac{1}{100000} = 0.5$$

  - Jakie jest prawdopodobieństwo znalezienia dokładnie jednej komórki w tej próbie?

$$P_{\mu}(k) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{0.5^1 e^{-0.5}}{1!} = 0.303$$

ROZKŁAD POISSONA

Ile komórek powinna liczyć próba, aby prawdopodobieństwo znalezienia w niej co najmniej 20 komórek było równe 0.95?

$$P_\mu(k \geq 20) = 0.95 = 1 - \sum_{k=0}^{19} \left[\left(\frac{N}{100000} \right)^k \frac{e^{-\left(\frac{N}{100000}\right)}}{k!} \right]$$

Rozwiązujemy np. podstawiając kolejno różne wartości pod N.

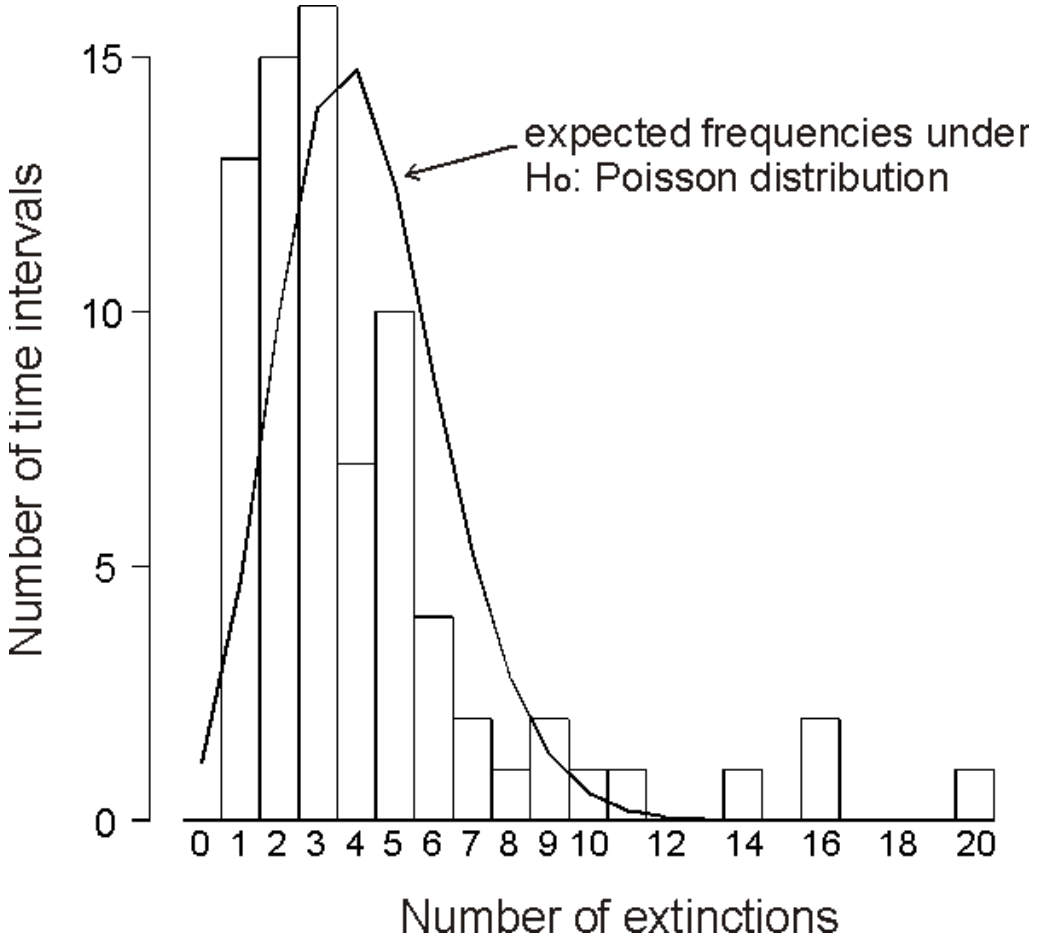
Rozwiązanie: potrzebujemy próby złożonej z ~2.8 mln komórek.

ROZKŁAD POISSONA

Czy liczba wymarłych gatunków podlega rozkładowi Poissona?

Test chi-kwadrat wykazał, że nie jest to rozkład Poissona

Wniosek: Wymieranie nie następuje losowo, równomiernie w czasie, ale w okresach tzw. wielkiego wymierania



http://www.zoology.ubc.ca/~bio300b/poissonnotes.html

PROCESY MARKOWA

Ciąg doświadczeń w których wynik kolejnego eksperymentu zależy tylko od wyniku poprzedniego

- Oznacza to, że dochodząc do każdego stanu, łańcuch „zapomina", skąd przyszedł, a prawdopodobieństwa przejścia w następnym ruchu zależą tylko od położenia bieżącego

Wyniki poszczególnych prób są stanami łańcucha

Ciąg zmiennych losowych o wartościach całkowitych $\{X_t\}_{t=0}^{\infty}$

nazwiemy **łańcuchem Markowa** jeśli:

$$P(X_t = j | X_0 = k_0, X_1 = k_1, \ldots, X_{t-1} = i) = P(X_t = j | X_{t-1} = i)$$

jeśli $P(X_t = j | X_{t-1} = i)$ nie zależy od numeru próby (t) to taki łańcuch nazwiemy **jednorodnym**

MODELE Z CZASEM DYSKRETNYM

Czym różnią się łańcuchy Markowa od równań rekurencyjnych i modeli z czasem dyskretnym?

Łańcuchy Markowa:

- zmiana stanów ma charakter losowy
- można przewidzieć kolejny stan tylko z pewnym prawdopodobieństwem
- to w jakim stanie aktualnie znajduje się model w dalszych etapach łańcucha również wyznaczamy z pewnym prawdopodobieństwem

ŁAŃCUCHY MARKOWA

Jednorodny łańcuch Markowa - przykłady:

- rzut monetą – dwa możliwe wyniki, w każdej próbie, orzeł lub reszka, a prawdopodobieństwo otrzymania wynosi zawsze 0,5 niezależnie od tego jakie wyniki otrzymamy w przeszłości
- rosyjska ruletka, w której używamy sześciostrzałowego rewolweru z jedną kulą. Za każdym razem kręcimy bębenkiem, więc prawdopodobieństwo, że pistolet wystrzeli jest zawsze jednakowe.

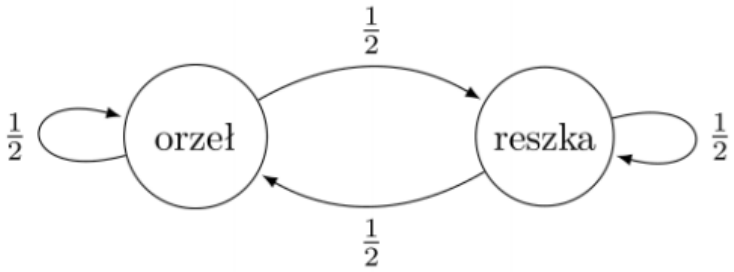
Niejednorodny łańcuch Markowa – przykład:

- Rosyjska ruletka, w której używamy sześciostrzałowego rewolweru z jedną kulą, ale bębenkiem kręcimy tylko raz

RZUT MONETĄ

Jest jednorodnym łańcuchem markowa, dwa możliwe wyniki O lub R

Graf



Macierz przejścia – prawdopodobieństwo przejścia z jednego stanu do drugiego

Przyszły stan → orzeł reszka

$$P = \begin{pmatrix} 0{,}5 & 0{,}5 \\ 0{,}5 & 0{,}5 \end{pmatrix}$$

orzeł

reszka

↑ Obecny stan

Suma wierszy jest zawsze równa 1

S_0 = stan początkowy

orzeł reszka

$$S_0 = (0\ 1)$$

JAK ZNALEŹĆ PRAWDOPODOBIEŃSTWA PRZEJŚCIA?

Prawdopodobieństwa przejścia łańcucha Markowa można znaleźć doświadczalnie, zliczając przypadki przejść między stanami

Znaleziony łańcuch może być użyty do symulacji badanego zjawiska

CO SIĘ STANIE W PRZYSZŁOŚCI?

Rozkład stacjonarny

- Rozkład prawdopodobieństw do którego dąży łańcuch Markowa

$$\lim_{n \to +\infty} p_{ij}^{(n)} = \pi_j$$

- Odpowiednio długo symulowany łańcuch Markowa zbiega do swojego stanu stacjonarnego (jeżeli taki istnieje)

- To, czy taki rozkład istnieje i jest jednoznacznie wyznaczony, zależy m.in. od rodzaju stanów występujących w danym łańcuchu

JAK ZNALEŹĆ ROZKŁAD STACJONARNY?

Sposób 1: Oblicz P^t, gdzie t jest wysoką potęgą i zwróć jeden z wierszy otrzymanej macierzy

Sposób 2: MCMC (*Marcov Chain Monte Carlo*)

- Symulujemy wstępnie dużą liczbę kroków łańcucha tak by zbiegł do rozkładu stacjonarnego
- Dla określonego punktu przez N iteracji zliczamy ilości stanów jakie przyjął łańcuch
- Za prawdopodobieństwo przyjęcia stanu i przyjmujemy:

$$\pi_i = \frac{ilość\ kroków\ w\ których\ łańcuch\ był\ w\ stanie\ i-tym}{N}$$

POLECANE

Łańcuchy Markowa wizualnie:

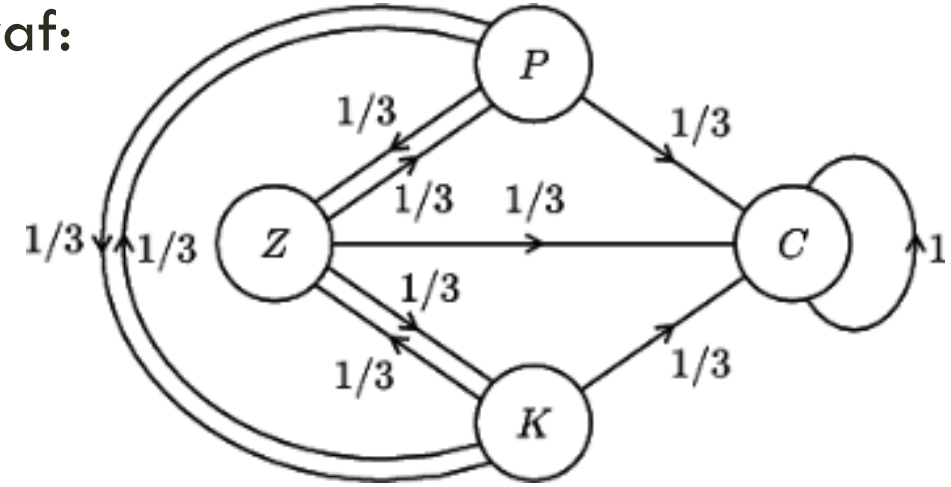
http://setosa.io/ev/markov-chains/

PCHŁA

Pchła skacze między podłogą, kotem, psem a człowiekiem

- Za każdym razem wybiera miejsce docelowe z takim samym prawdopodobieństwem (1/3)
- Pchła ginie po wskoczeniu na człowieka

Graf:



Źródło:

http://www.deltami.edu.pl/temat/matematyka/rachunek_prawdopodobienstwa/2013/08/29/O_polowaniu_na_pchle_i_czekaniu/

MACIERZ PRZEJŚCIA

Z definicji łańcucha Markowa wynika, że prawdopodobieństwo przejścia ze stanu i do j jest zawsze jednakowe. Oznaczmy je p_{ij}

Reprezentacja macierzowa łańcucha Markowa

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix} = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

1 – podłoga 2 – pies 3 – kot 4 – człowiek

ZADANIA

Wyznacz macierz przejścia i graf dla sytuacji, w której, pchła skacze po kocie, podłodze i człowieku. Skacząc na człowieka ginie.

	Kot	Podłoga	Człowiek
Kot	0	1/2	1/2
Podłoga	1/2	0	1/2
Człowiek	0	0	1

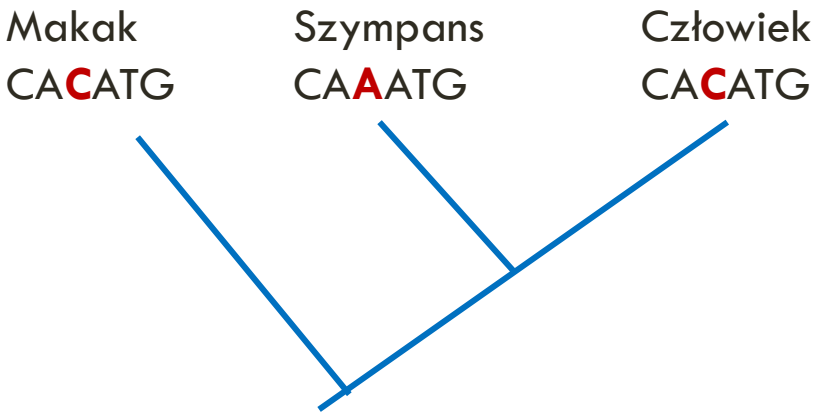
MODELE MARKOWA W BIOLOGII

Przykładowe wykorzystanie w biologii molekularnej:

- Ewolucja molekularna i konstrukcja drzew filogenetycznych
- Poszukiwanie charakterystycznych wzorów sekwencji np. sekwencje regulatorowe
- Analiza struktury białek

MODELE EWOLUCJI SEKWENCJI

Po co tworzymy modele ewolucji sekwencji?

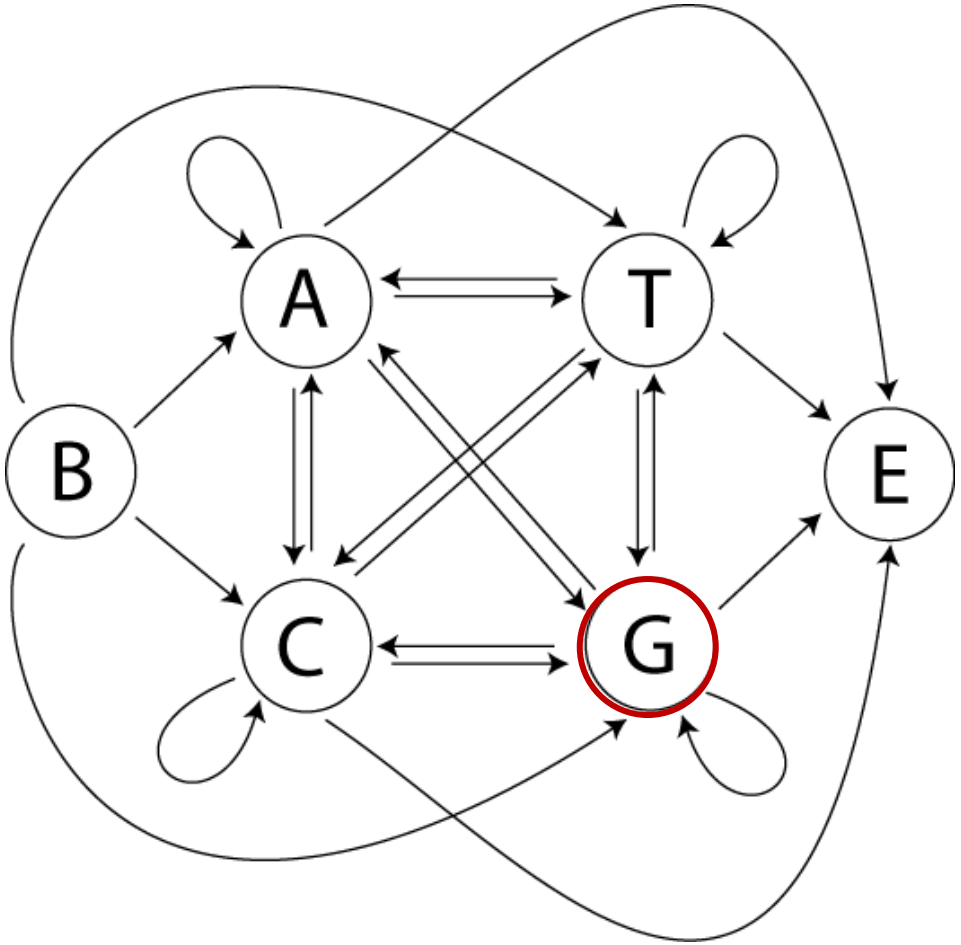


Jak bardzo analizowane sekwencje różnią się od siebie?

- Często brak jest sekwencji przodka
- Oszacowanie liczby mutacji musi opierać się na różnicach pomiędzy współczesnymi sekwencjami
- Konieczne jest uwzględnienie wielokrotnych mutacji tej samej pozycji

Najprostszy sposób: dopasowanie sekwencji oraz wyznaczenie odsetka pozycji na których obserwujemy różnice

MODELOWANIE SEKWENCJI DNA



Przykład:
Jakie będzie prawdopodobieństwo, że kolejnym nukleotydem (x_i) będzie A,C,G lub T jeżeli poprzednim (x_{i-1}) było G?

Przykładowe prawdopodobieństwa przejścia:

- $P(x_i = A \mid x_{i-1} = G) = 0,16$
- $P(x_i = C \mid x_{i-1} = G) = 0,34$
- $P(x_i = G \mid x_{i-1} = G) = 0,38$
- $P(x_i = T \mid x_{i-1} = G) = 0,12$

Zakończenie sekwencji (E) pozwala na uwzględnienie długości sekwencji w modelu

MODEL JUKESA-CANTORA (JC)

Model ewolucji sekwencji opracowany przez Jukesa i Cantora w 1969.

Opisuje pojedynczą pozycje w dopasowaniu pary sekwencji DNA, na której może znajdować się jeden z nukleotydów A,C,G lub T

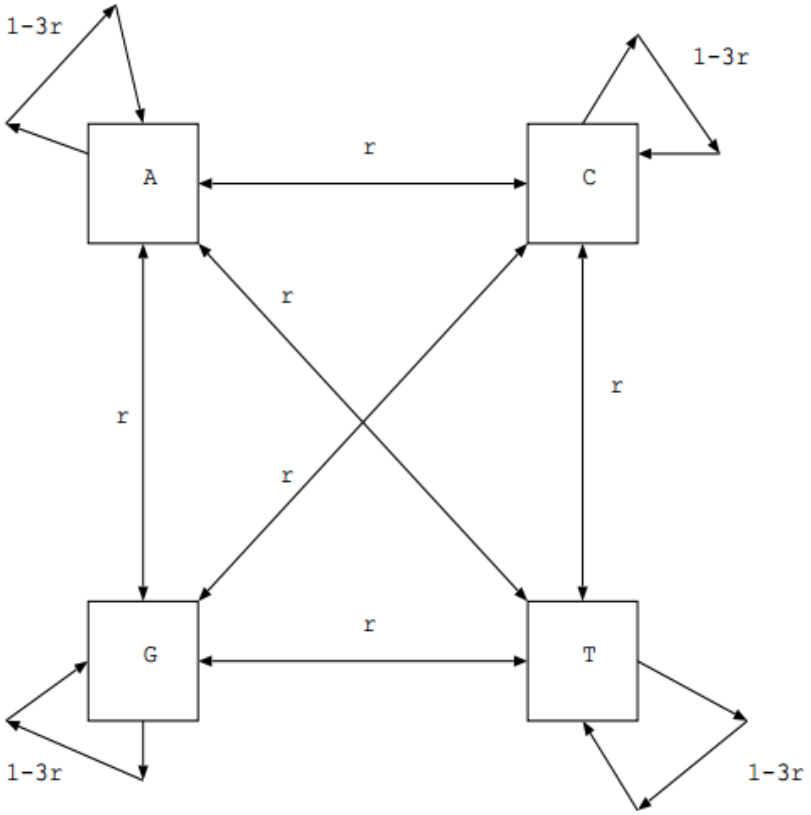
Stosowany dla niekodujących fragmentów DNA

Założenia:

- Zajście mutacji w jednym miejscu nie wpływa na pozostałe miejsca
- Jednakowe częstości nukleotydów (f=0,25)
- Jednakowe prawdopodobieństwo każdej substytucji

MODEL JUKESA-CANTORA (JC)

Graf:



http://www.uccs.edu/~rcarlson/cs484/notes3.pdf

Prawdopodobieństwo, że na pojedynczej pozycji obserwujemy nukleotyd A, C, G lub T w czasie t

$\Pi(t)=(P_A, P_C, P_G, P_T)$

Macierz przejścia:

$$\Pi(1) = \Pi(0) \begin{pmatrix} 1-3r & r & r & r \\ r & 1-3r & r & r \\ r & r & 1-3r & r \\ r & r & r & 1-3r \end{pmatrix}.$$

r – tempo mutacji (liczba mutacji po jednym kroku podzielona przez długość fragmentu DNA)

UKRYTE MODELE MARKOWA (HMM)

Skrót HMM pochodzi od angielskiego Hidden Markov Models

Są one rozszerzeniem definicji łańcucha Markowa. Modelują proces stochastyczny, którego pewne właściwości nie są znane

Przyjmujemy, że udaje się obserwować symbole (liczby, znaki) emitowane przez układ, a nie jego stany wewnętrzne

Definiuje się stany ukryte oraz stany początkowe i końcowe

Prawdopodobieństwo wystąpienia znaku będzie zależeć od:

- Znaków występujących na poprzedzających pozycjach
- Ukrytego stanu w jakim znajduje się model

NIEUCZCIWE KASYNO

Kasyno ma dwa rodzaje kości:

- Uczciwa – symetryczna, prawdopodobieństwo wyrzucenia określonej liczby oczek jest równe zawsze 1/6

- Nieuczciwa – Prawdopodobieństwo wyrzucenia 6 jest równe 1/2, natomiast pozostałych oczek 1/10

Możliwe są dwa stany: kostka uczciwa oraz nieuczciwa

Układ może zmieniać swój stan z pewnym prawdopodobieństwem, ale samego stanu nie jesteśmy w stanie zaobserwować. Obserwujemy jedynie wyniki rzutu kostką.

DEKODOWANIE

Chcąc poznać najbardziej prawdopodobne rozmieszczenie poszczególnych stanów w sekwencji musimy przeprowadzić **dekodowanie** modelu

Algorytm Viterbiego

- Identyfikacja najbardziej prawdopodobnej ścieżki przejścia przez model, czyli sekwencji przejścia przez model stanów ukrytych, której prawdopodobieństwo jest największe
- Dla każdej pozycji wyznaczamy prawdopodobieństwo jej przynależności do jednego ze stanów

DEKODOWANIE

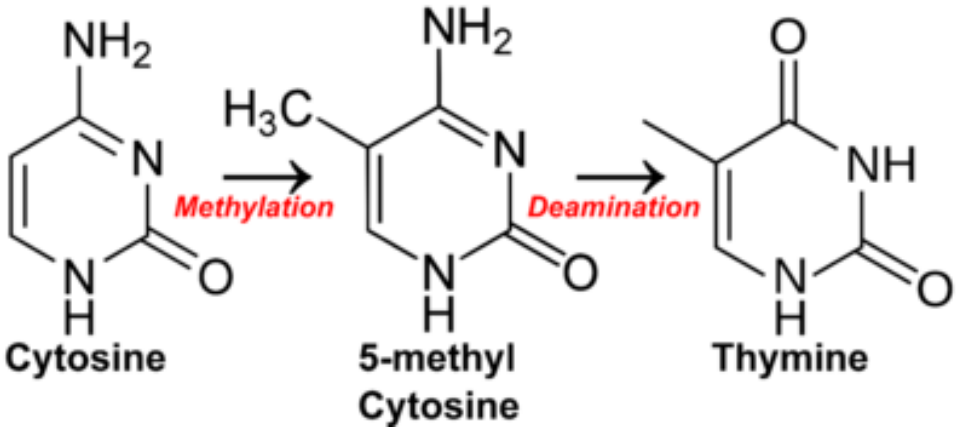
Algorytmy sufiksowy i prefiksowy (ang. *forward and backward algorithms*)

- Rozważamy wszystkie możliwe ścieżki w modelu w celu wyznaczenia dla każdej pozycji przynależności do określonego stanu
- Dla pewnych pozycji w sekwencji prawdopodobieństwo jest bliskie 1, natomiast dla pozostałych zbliża się do 0
- Dla pozostałych pozycji prawdopodobieństwo będzie mniej jednoznaczne – mogą one znaleźć się na krańcach jednoznacznie przypisanych fragmentów lub w ich środku

WYSPY CPG

Wyspa CpG to odcinek DNA o długości co najmniej 200 pz, charakteryzujący się zawartością par CG powyżej 50%

W dinukleotydach CpG, poza wyspami CpG, C często ulega metylacji co z dużym prawdopodobieństwem prowadzi do deaminacji i mutacji w T

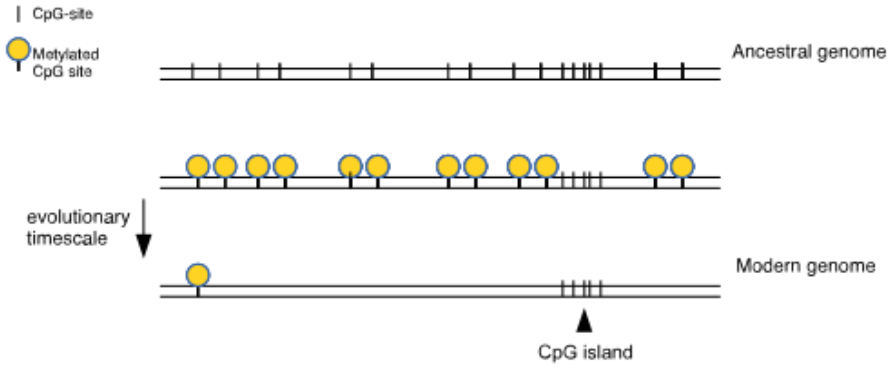


WYSPY CPG

Występują w 60% sekwencji promotorowych, gdzie metylacja cytozyny mogłaby wyciszać ekspresję genów na poziomie transkrypcji

Ponieważ wyspy CpG występują w pobliżu genów często są **chronione przed metylacją** dzięki specyficznym białkom wiążącym się z DNA lub pozycjonowaniu nukleosomów, co blokuje dostęp metylotransferaz

Rozpoznanie wysp CpG może być wykorzystane jako wskazówka dla znalezienia ciekawych biologicznie fragmentów genomu



WYSPY CPG A HMM

Czy krótka sekwencja DNA pochodzi z wyspy CpG?

Zaproponowany model oparty na HMM:

Model ma osiem stanów:

- A_+, C_+, G_+, T_+, A_-, C_-, G_-, T_-.
- Stan z plusem np. A_+ oznacza , że znajdujemy się w rejonie wyspy CpG i czytamy A
- Stan z minusem, np. A_- ,oznacza, że znajdujemy się poza rejonem wyspy CpG i czytamy A

Emitowane symbole: będąc w stanie emitujemy A_ξ (gdzie $\xi \in \{-, +\}$ z prawdopodobieństwem 1, każdy inny symbol jest emitowany w z prawdopodobieństwem 0

WYSPY CPG A HMM

Wymagane jest posiadane danych w których stany są znane, na ich podstawie wyznaczamy prawdopodobieństwa przejścia oddzielnie dla wysp CpG oraz pozostałych fragmentów sekwencji

$$a_{ij}^{+} = \frac{c_{ij}^{+}}{\sum_k c_{ij}^{+}}$$

Ile razy nukleotyd i pojawia się za nukleotydem j w sekwencji oznaczonej + (wyspa CpG)

Analogiczne obliczenia prowadzimy dla sekwencji oznaczonej -

Wzory: http://www.cs.hunter.cuny.edu/~saad/courses/compbio/lectures/lecture9.pdf

WYSPY CPG A HMM

W obszarze wysp CpG zaobserwujemy większe prawdopodobieństwo przejścia w C i G niż poza nią

$$P^+ = \begin{bmatrix} & A & C & G & T \\ A & 0.18 & 0.27 & 0.43 & 0.12 \\ C & 0.17 & 0.37 & 0.27 & 0.19 \\ G & 0.16 & 0.34 & 0.37 & 0.13 \\ T & 0.08 & 0.36 & 0.38 & 0.18 \end{bmatrix} \quad P^- = \begin{bmatrix} & A & C & G & T \\ A & 0.30 & 0.20 & 0.29 & 0.21 \\ C & 0.32 & 0.30 & 0.08 & 0.30 \\ G & 0.25 & 0.25 & 0.29 & 0.21 \\ T & 0.18 & 0.24 & 0.29 & 0.29 \end{bmatrix}$$

http://www.cs.hunter.cuny.edu/~saad/courses/compbio/lectures/lecture9.pdf

Posiadając krótką sekwencje x obliczymy p(x) dla każdego łańcucha Markowa szansę p(x|+) i p(x|-)

WYSPY CPG

$$P^+ = \begin{bmatrix} & A & C & G & T \\ A & 0.18 & 0.27 & 0.43 & 0.12 \\ C & 0.17 & 0.37 & 0.27 & 0.19 \\ G & 0.16 & 0.34 & 0.37 & 0.13 \\ T & 0.08 & 0.36 & 0.38 & 0.18 \end{bmatrix} \quad P^- = \begin{bmatrix} & A & C & G & T \\ A & 0.30 & 0.20 & 0.29 & 0.21 \\ C & 0.32 & 0.30 & 0.08 & 0.30 \\ G & 0.25 & 0.25 & 0.29 & 0.21 \\ T & 0.18 & 0.24 & 0.29 & 0.29 \end{bmatrix}$$

Iloraz szans:

$$\log \frac{p(x|+)}{p(x|-)} = \log \frac{\prod_{i=0}^{n} a^+_{x_i x_{i+1}}}{\prod_{i=0}^{n} a^-_{x_i x_{i+1}}} = \sum_{i=1}^{n-1} \log \frac{a^+_{x_i x_{i+1}}}{a^-_{x_i x_{i+1}}}$$

Jeżeli suma ilorazów szans będzie większa od 0 wnioskujemy, że x pochodzi z wyspy CpG

CGCG

$$\log \frac{0.27}{0.08} + \log \frac{0.34}{0.25} + \log \frac{0.27}{0.08} > 0$$

Wzory: http://www.cs.hunter.cuny.edu/~saad/courses/compbio/lectures/lecture9.pdf

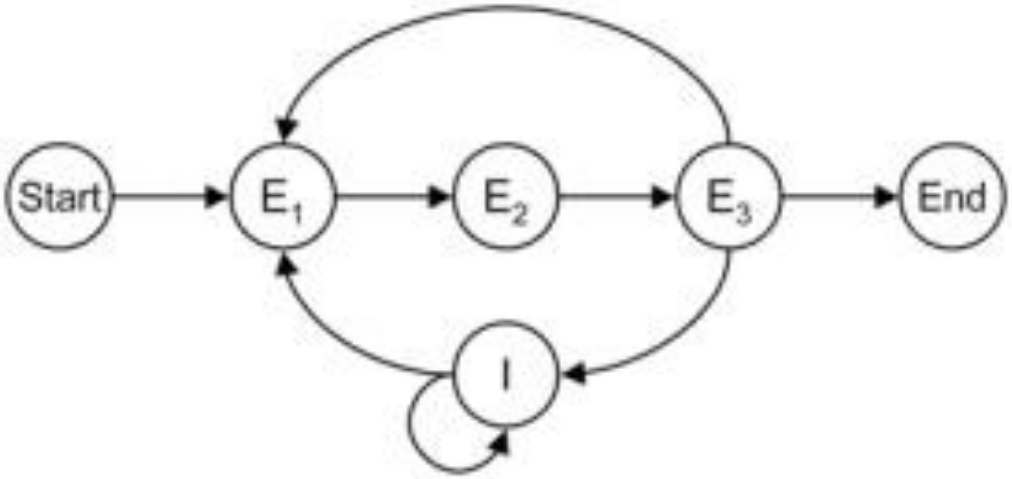
ZADANIE:

Czy sekwencja CCACG pochodzi z wyspy CpG?

$$P^{+} = \begin{bmatrix} & A & C & G & T \\ A & 0.18 & 0.27 & 0.43 & 0.12 \\ C & 0.17 & 0.37 & 0.27 & 0.19 \\ G & 0.16 & 0.34 & 0.37 & 0.13 \\ T & 0.08 & 0.36 & 0.38 & 0.18 \end{bmatrix} \quad P^{-} = \begin{bmatrix} & A & C & G & T \\ A & 0.30 & 0.20 & 0.29 & 0.21 \\ C & 0.32 & 0.30 & 0.08 & 0.30 \\ G & 0.25 & 0.25 & 0.29 & 0.21 \\ T & 0.18 & 0.24 & 0.29 & 0.29 \end{bmatrix}$$

GENY EUKARIOTYCZNE

JAK POPRAWIE ZIDENTYFIKOWAĆ EKSONY I INTRONY?



E_1 : 1st position of codon
E_2 : 2nd position of codon
E_3 : 3rd position of codon
} exon
I : intron

x	A	T	G	C	G	A	C	T	G	C	A	T	A	G	C	A	C	T	T	observed symbols	
y	E_1	E_2	E_3	E_1	E_2	E_3	E_1	E_2	E_3	I	I	I	I	E_1	E_2	E_3	E_1	E_2	E_3	hidden states	

exon | intron | exon

DLA ZAINTERESOWANYCH

Current Genomics

BENTHAM SCIENCE PUBLISHERS LTD.

Curr Genomics. 2009 Sep; 10(6): 402–415. PMCID: PMC2766791
doi: 10.2174/138920209789177575

Hidden Markov Models and their Applications in Biological Sequence Analysis

Byung-Jun Yoon*

Abstract Go to:

Hidden Markov models (HMMs) have been extensively used in biological sequence analysis. In this paper, we give a tutorial review of HMMs and their applications in a variety of problems in molecular biology. We especially focus on three types of HMMs: the profile-HMMs, pair-HMMs, and context-sensitive HMMs. We show how these HMMs can be used to solve various sequence analysis problems, such as pairwise and multiple sequence alignments, gene annotation, classification, similarity search, and many others.

LITERATURA

Foryś U. 2011. Modelowanie matematyczne w biologii i medycynie.

Higgs P., Atwood T. 2011. Bioinformatyka i ewolucja molekularna. PWN.

http://www.deltami.edu.pl/temat/matematyka/rachunek_prawdopodobienstwa/2013/08/29/O_polowaniu_na_pchle_i_czekaniu/

http://www.cs.hunter.cuny.edu/~saad/courses/compbio/lectures/lecture9.pdf

Tiuryn J. 2006. Wstęp do obliczeniowej biologii molekularnej. Wykłady