

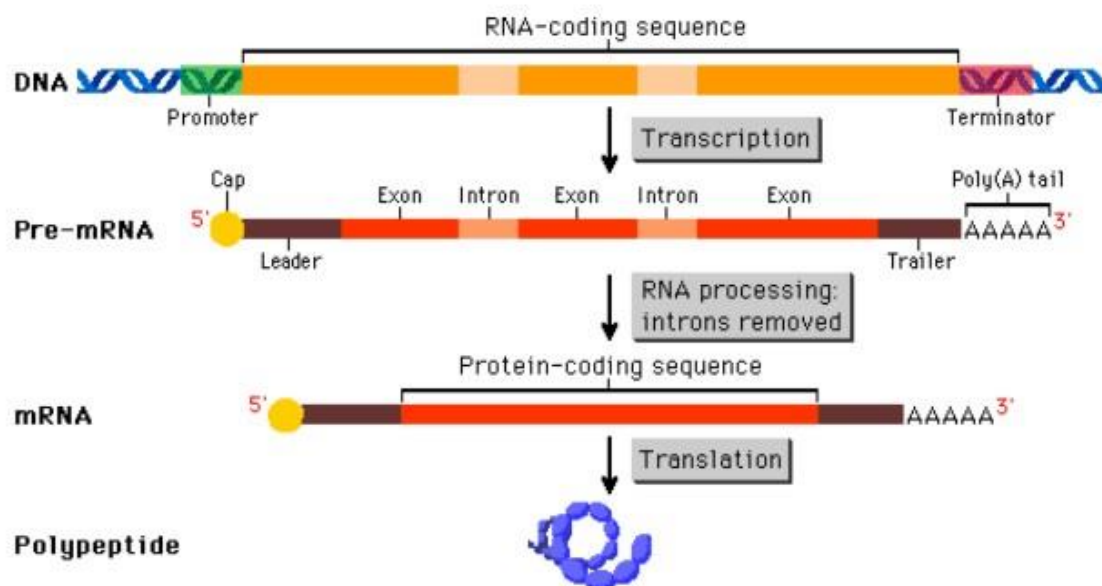
*Instytut Informatyki i Matematyki Komputerowej UJ,
opracowanie: mgr Ewa Matczyńska, dr Jacek Śmietański*

Przewidywanie genów

Pytanie na dzisiaj: **Jak odnaleźć gen w sekwencji DNA?**

```
CTTACTTCTGAAGGCTGTGCTCCGCTCACCATCCAGAGCGGAGGTGCGGACCTTAAACTCACTCCTGGAGA
AAGATCTGCAAGTGCGCAGGTAAAGTGCACGTGCTCCGCGGTGCGGAGGAAGGAGGCGAGGAGCCAGACT
AGCCTGGGAACAGGCAGGGAGGGTTTACACAGCCCCGGCTGAGTCGCGGCTTAGGAAGCAAGGCAAGTTC
CCCTAAAGGTTAGTGTGCACAGACGGGTGCGACGGAGCCGACCTAGCGCGGCTGAGTCCGCCTGGGCCTG
CAGCAGCTGCCCCCTGAGCACCCCCCTCCGGCTCTCTGCCAGGCGACCCAGGAAAAAGTCGCCCCCTGGTG
GGCCATGAGGTCATGGGTGGGGGGAGTTTGGAAAGGTTTCAGACAGCAAATGTTCCACTTGAATCCAGGG
CAGCATCTGGCACTGCGGGGCCCTCCTAGCCATGAGCCGTGGTCAGGCGTTTCCTTAGAATGGAATGCACT
GGAGTGAAACACTAAATCCCTCAAAGCTGCTTCTCTTTACTGTGGTCACACACAGTGAAATCAATGGGC
ATTAGTGCAGCTAGCTCTTTTCAAGGACACAATGTTAAGCACAGGAAGCCTGGTATGTGGACGCTCTGGG
TTTGAGAACCAGGCAGGGGCCAGGGGCTCAGGACAAGTGCCCGGTGCTCCCTTCCCATTTGGGCGAATCA
GAGCCTGGGGCCCGGCTGAAGCTCCCCAGGTGACTCTAATGTGCAGTGCACTTTGAGGAGCACTACTT
AGACCAATGTGACAGTCTACAATGTGTAGATTTAGGGTGAGTGATTCTGAGGAAAAGAAACCCGAGGCTG
TTAGCAGTTGTGGGCAGCTGCTGTCTACCTAAACCAGCTGCGGTTTGCTTGCTTGGTGAGCCTGAGCTTC
GCGGGGCGGGCATGGCATCTGCCATCCAGGACCCTGGGACGGGGCTGCCAGGGCAAGGAGCTGAGCATT
GAACGCATCTGGGGATAACTATCTCTCATAGAAGAACTAATGAGGATTGAACCTGAACACAGAGATGAAA
GAGCTGAGTAGACTGCAAAGAATTGCACAAAACTGCTTGTTCTGCAAATTTAGTTTACAACAATTTGAG
```

1. Budowa genu eukariotycznego



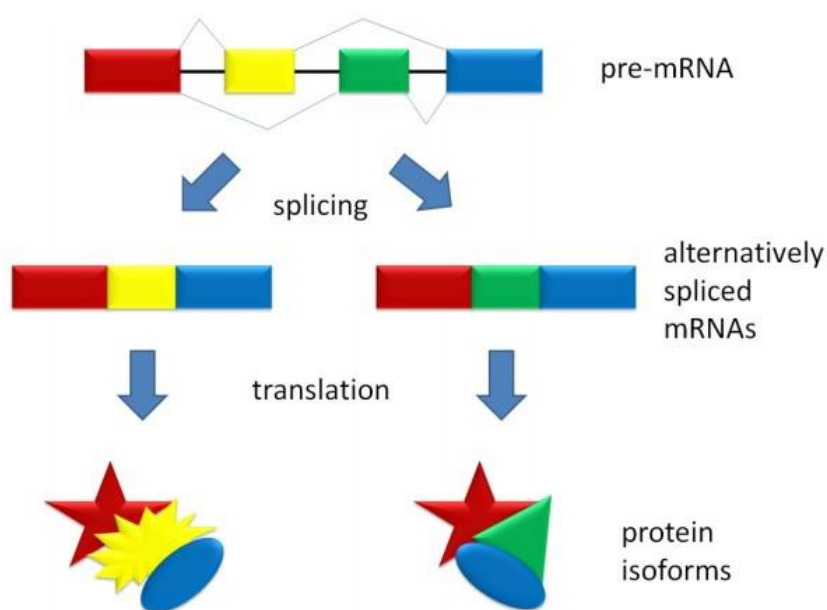
Rysunek 1: Budowa genu eukariotycznego i procesy, którym gen zostaje poddany aby został przetłumaczony na białko.

Gen to pewien ciągły fragment DNA, który koduje białko. Okazuje się, że u organizmów eukariotycznych (czyli takich, których komórki posiadają jądro w którym zawarte jest DNA)

budowa genu jest istotnie bardziej skomplikowana niż u prokariotów (organizmów bez jądra komórkowego, np. bakterii). Skupimy się teraz na budowie genu eukariotycznego.

Zwykle gen jest poprzedzony fragmentem DNA zwanym promotorem, który jest rozpoznawany przez maszynę komórkową. Przyłącza się ona do tego fragmentu i rozpoczyna transkrypcję czyli przepisywanie DNA na mRNA. Od charakterystyki promotora zależy regulacja takiego genu, czyli w uproszczeniu – jak dużo białka na podstawie tego genu powstaje w danej chwili w komórce.

Sama sekwencja genu jest podzielona na fragmenty dwóch typów: eksony i introny. Jak widać na powyższym rysunku zarówno eksony jak i introny są początkowo przepisywane na RNA, jest to tzw. pre-mRNA. Następnie z tego łańcucha usuwane są fragmenty intronów i w dojrzałym mRNA zostaje tylko sekwencja pochodząca z eksonów. Proces ten określa się jako splicing.



Rysunek 2: Przykład splicingu alternatywnego.

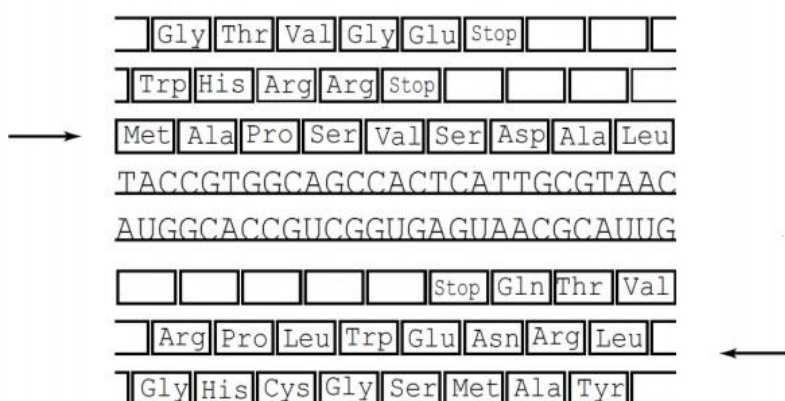
Możliwa jest sytuacja w której eksony składane są w innej kolejności niż było to określone w wyjściowej sekwencji DNA, bądź niektóre zostają pominięte. Taki proces nazywamy splicingiem alternatywnym. Dzięki niemu z jednego genu może powstać dużo wariantów białek. Kiedyś oceniano, że jest to proces stosunkowo rzadki, teraz wiadomo, że alternatywnemu splicingowi ulega ok 95% genów u człowieka.

U prokariotów czyli np. bakterii sytuacja jest łatwiejsza - introny nie występują.

Jeszcze jednym istotnym pojęciem związanym z sekwencją genu jest tzw. otwarta ramka odczytu (ORF - *open reading frame*). Jest to każda sekwencja zawarta pomiędzy kodonem start a kodonem stop w danej ramce odczytu. Może ona potencjalnie kodować białko.

2. Metody przewidywań oparte na charakterystycznych elementach sekwencji i statystyce

Pierwszym intuicyjnym podejściem dla wyszukania sekwencji genu, będzie odszukanie ORF we wszystkich ramkach odczytu (zwróć uwagę, że ORF mogą się nakładać). Możemy się spodziewać, że w losowej sekwencji raz na około 21 kodonów otrzymamy kodon STOP (dlaczego?). Geny mają zwykle dłuższą sekwencję niż losowo napotkane ORF. Naiwne podejście zakłada, że tam gdzie mamy odpowiednio długą ORF, możemy spodziewać się genu.



Rysunek 3: Sekwencja i kodowane aminokwasy w 6 ramkach odczytu. Z „Introduction to Bioinformatics Algorithms”, Pevzner, MIT Press 2004

W określeniu czy dana ORF jest sekwencją genu może pomóc tzw. *codon usage*. Jak pamiętamy, większość aminokwasów jest kodowana przez więcej niż jeden kodon, co wynika ze zdegenerowania kodu genetycznego. Okazuje się, że częstość pojawiania się danego kodonu dla danego aminokwasu jest znacząco różna w rejonach kodujących i niekodujących. Na podstawie zebranych danych można stworzyć tabele *codon usage*, określające częstości danych kodonów dla aminokwasów w sekwencjach kodujących i niekodujących.

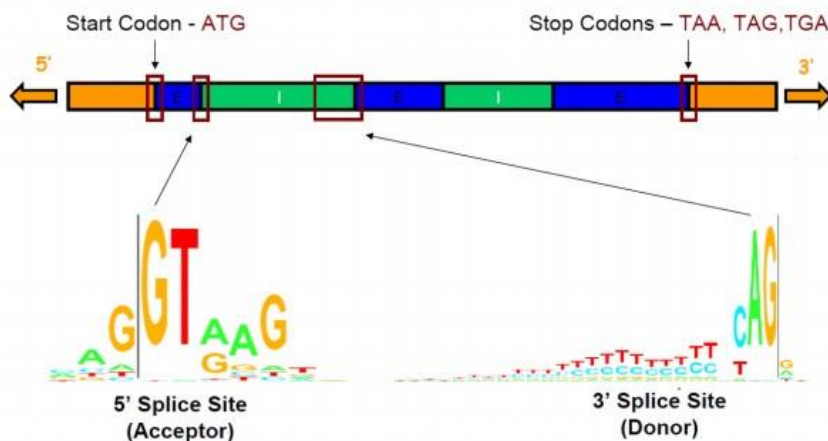
	U	C	A	G
U	UUU Phe 57 UUC Phe 43 UUA Leu 13 UUG Leu 13	UCU Ser 16 UCC Ser 15 UCA Ser 13 UCG Ser 15	UAU Tyr 58 UAC Tyr 42 UAA Stp 62 UAG Stp 8	UGU Cys 45 UGC Cys 55 UGA Stp 30 UGG Trp 100
C	CUU Leu 11 CUC Leu 10 CUA Leu 4 CUG Leu 49	CCU Pro 17 CCC Pro 17 CCA Pro 20 CCG Pro 51	CAU His 57 CAC His 43 CAA Gln 45 CAG Gln 66	CGU Arg 37 CGC Arg 38 CGA Arg 7 CGG Arg 10
A	AUU Ile 50 AUC Ile 41 AUA Ile 9 AUG Met 100	ACU Thr 18 ACC Thr 42 ACA Thr 15 ACG Thr 26	AAU Asn 46 AAC Asn 54 AAA Lys 75 AAG Lys 25	AGU Ser 15 AGC Ser 26 AGA Arg 5 AGG Arg 3
G	GUU Val 27 GUC Val 21 GUA Val 16 GUG Val 36	GCU Ala 17 GCC Ala 27 GCA Ala 22 GCG Ala 34	GAU Asp 63 GAC Asp 37 GAA Glu 68 GAG Glu 32	GGU Gly 34 GGC Gly 39 GGA Gly 12 GGG Gly 15

Rysunek 4: Tabela *codon usage* dla regionów kodujących u człowieka. Z „Introduction to Bioinformatics Algorithms”, Pevzner, MIT Press 2004

Następnie, używając prawdopodobieństw pojawienia się danego kodonu dla aminokwasu z tabeli *codon usage*, można obliczyć stosunek prawdopodobieństw warunkowych dla sekwencji s:

$$\frac{P(s \mid s \text{ jest kodująca})}{P(s \mid s \text{ jest niekodująca})}$$

Istnieje jeszcze wiele dodatkowych charakterystycznych elementów sekwencji genowej używanych do predykcji genów. Najważniejszymi z nich są tzw. wyspy CpG oraz miejsca początku i końca intronów. Wyspy CpG to dość długie fragmenty sekwencji (>200 bp) o znacząco wyższej częstości występowania nukleotydów C i G (>50%), zwykle wyspy te występują w promotorach genów. Miejsca początku i końca intronów mają najczęściej ściśle konserwatywne dinukleotydy odpowiednio GT i AG.



Rysunek 5: Sygnały w sekwencji genu. Z „Computational Biology: Genomes, Networks, Evolution” MIT OpenCourseWare

Oprócz tego istnieje oddzielna gałąź metod przewidywania genów oparta na podobieństwach sekwencji do już znanych genów (*comparative gene prediction*). Jest to wartościowa metoda, jeśli mamy blisko spokrewniony organizm z opisanymi sekwencjami genów, których możemy użyć do porównania. Metoda ta ma oczywiście pewne ograniczenia, nie znajdzie nam sekwencji genu, który nie wykazuje istotnego podobieństwa do tych już znanych.

3. *Hidden Markov Models (HMM)* dla przewidywania genów

Ukryte modele Markowa są modelami probabilistycznymi używanymi w wielu dziedzinach. Używa się ich np. do rozpoznawania mowy bądź pisma ręcznego. W bioinformatyce najbardziej znanym ich zastosowaniem jest właśnie przewidywanie genów. HMM zalicza się do modeli uczenia maszynowego (*machine learning*) tzn. że model początkowo uczy się charakterystycznych zależności na przedstawionych mu przykładach, a następnie sam powinien być w stanie podjąć decyzję co do nowo przedstawionego przykładu, bazując na wcześniej zdobytej wiedzy.

Jako wprowadzenie do HMM rozważmy pewien przykład nazywany „Fair Bet Casino”. Otóż mamy następującą sytuację: Wyobraźmy sobie, że znajdujemy się w kasynie i gramy w

następującą grę: krupier rzuca monetą, a my stawiamy pieniądze na to czy wypadnie orzeł czy reszka. Ponieważ w kasynie oszukują krupier może rzucać także monetą oszukaną, której prawdopodobieństwo rzucenia reszki wynosi $\frac{3}{4}$. Krupier podmienia monety, ale niezbyt często, bo boi się, że ktoś to zauważy, załóżmy, że podmienia monety z prawdopodobieństwem 0.1. Obserwując wyniki kolejnych rzutów chcemy wiedzieć czy krupier rzuca monetą dobrą czy oszukaną, co pozwoli nam zwiększyć szansę na wygranę pieniędzy.

Jeśli zaobserwujemy, że cały czas wypada reszka, możemy twierdzić, że rzucana jest moneta fałszywa, ale może również to być też oczywiście przypadek dla monety dobrej, jednak interesują nas rozwiązania najbardziej prawdopodobne.

Rozważmy najpierw sytuację w której krupier nigdy nie zmienia monety. Oznaczmy typy monet: F -fair, dobra moneta, B -biased, fałszywa moneta.

Wtedy dla sekwencji wyników kolejnych rzutów oznaczonej przez: $x = (x_1, x_2, x_3, \dots, x_n)$, możemy obliczyć prawdopodobieństwo pod warunkiem, że sekwencja została wyrzucona dobrą monetą:

$$P(x|F) = \prod_{i=1}^n p^F(x_i) = \frac{1}{2^n}$$

oraz prawdopodobieństwo pod warunkiem, że sekwencja została wyrzucona fałszywą monetą

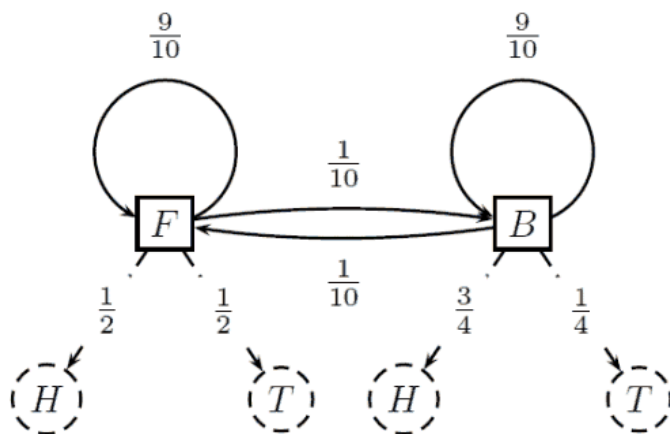
$$P(x|B) = \prod_{i=1}^n p^B(x_i) = \frac{1}{4^{n-k}} \frac{3^k}{4^k} = \frac{3^k}{4^n}$$

gdzie k - liczba reszek.

Teraz porównując prawdopodobieństwa, możemy określić czy bardziej prawdopodobne jest używanie monety dobrej czy fałszywej przy zadanej sekwencji. Okazuje się, że jeśli $k, \frac{n}{\log_2 3}$,

$\log_2 3$ wtedy bardziej prawdopodobna jest moneta dobra.

W przypadku kiedy krupier może zamieniać monety, dobrym modelem dla tej sytuacji będzie HMM. Ukryty model Markowa składa się z pewnych stanów między, którymi przemieszcza się z pewnymi prawdopodobieństwami. Oprócz tego w każdym stanie może wyemitować pewien symbol z danego alfabetu z pewnym prawdopodobieństwem. Dla naszego przykładu HMM będzie wyglądał następująco:



Rysunek 6: HMM dla problemu „Fair Bet Casino”. Oznaczenia: F-(fair) dobra moneta, B-(biased) fałszywa moneta, H - (head) reszka, T-(tails) orzeł.

W modelu mamy dwa stany oznaczone przez F i B oznaczające rzucanie danym typem monety. Rzucając monetą emitujemy pewne symbole - u nas H bądź T, czyli orzeł bądź reszka z pewnym prawdopodobieństwem w zależności od tego czy rzucamy dobrą czy fałszywą monetą. Monety czyli stany mogą się zmieniać z prawdopodobieństwem 0.1.

Zauważmy, że obserwujemy tylko emitowane symbole, czyli wyniki rzutu monetą, natomiast nie wiemy jaki aktualny stan modelu, czyli czy krupier rzuca monetą dobrą czy fałszywą. Stąd nazwa Ukryty model Markowa, ponieważ stany są niewidoczne dla obserwatora.

Celem obserwatora jest odnalezienie najbardziej prawdopodobnej ścieżki, czyli sekwencji stanów, która wyprodukowała daną sekwencję.

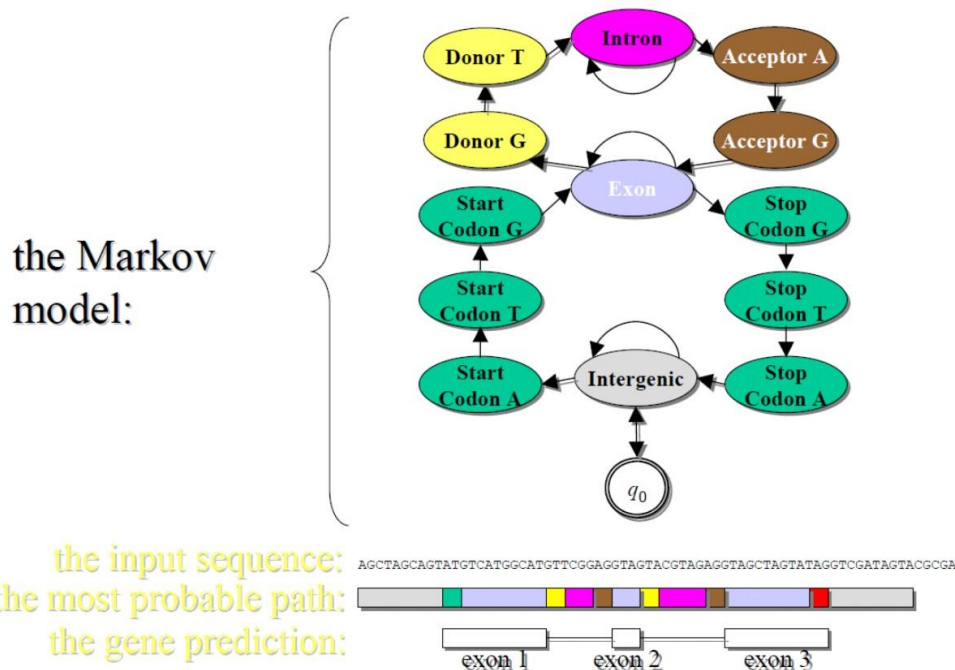
Bardziej formalnie HMM M składa się z :

- alfabetu emitowanych symboli
- Q : zbioru stanów, z których każdy emituje symbol z alfabetu
- $A = a_{kl}$ macierzy prawdopodobieństw przejść stanów rozmiaru $|Q_i| \times |Q_j|$
- $E = e_k(b)$ macierzy prawdopodobieństw emisji danego symbolu w danym stanie rozmiaru $|Q_i| \times |Q_j|$

Ścieżką π nazwiemy sekwencję stanów. Jak wyglądają poszczególne składniki dla wyżej opisanego problemu?

Ogólnie, szukamy takiej ścieżki π , która maksymalizuje prawdopodobieństwo uzyskania danej sekwencji emitowanych symboli $P(x|\pi)$.

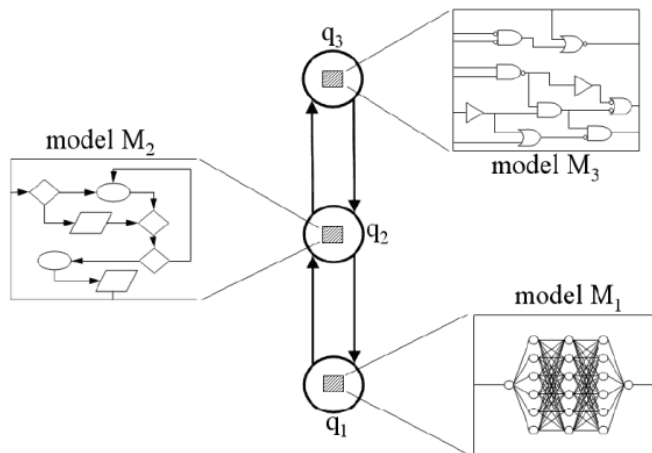
Bardzo łatwo znaleźć analogię gry z monetą do problemu predykcji genów. Emitowanym symbolem będzie kolejny nukleotyd sekwencji, natomiast ukrytym stanem, który chcemy odkryć, należenie danego nukleotydu do eksonu, intronu, sekwencji między genowej, itd.



Rysunek 7: Prosty HMM dla predykcji genów. Z „Computational Biology: Genomes, Networks, Evolution”, MIT OpenCourseWare

Wiele programów do predykcji genów bazuje na HMM-ach, wprowadzane są różne rozszerzenia tego modelu. W tych rozszerzonych modelach uwzględnia się również

charakterystykę częstości i sygnały, o których była mowa wcześniej. Jednym z najbardziej znanych programów przewidujących geny, nie opierającym się na podobieństwach do już znanych genów jest GENSCAN, opracowany na Uniwersytecie Stanford. GENSCAN jest narzędziem opierającym się na Ukrytym modelu Markowa, z tą różnicą, że w danym stanie możemy emitować więcej niż jeden nukleotyd. Emisja większej ilości nukleotydów jest kontrolowana przez z góry przyjęty model probabilistyczny dla każdego stanu. Jest to tzw. *generalized HMM* (GHMM).



W.H. Majoros (<http://geneprediction.org/book/classroom.html>)

Rysunek 8: Model GHMM.

Zadanie 1:

Wykorzystaj program GENSCAN dla poszukiwania genów w zadanych fragmentach DNA (pliki **gen1.txt** i **gen2.txt**).

1. Przeszukaj sekwencje DNA z obydwu plików. W poszukiwaniu genów wykorzystaj program GENSCAN (<http://genes.mit.edu/GENSCAN.html>):

W kolumnach tabeli mamy informacje o znalezionych eksonach:

Type - typ przewidzianego eksonu (*Sigl* - pojedynczy, *Init* - inicjujący, *Term* - ostatni, *Intr* - między innymi eksonami), może się zdarzyć, że GENSCAN wykryje również inne elementy charakterystyczne jak *Prom* - promotor, *Poly-A* - fragment złożony z samych adenin w przepisany na mRNA genie, charakterystyczny przy końcu genu, tzw. ogon polyA,

S - strand - nić: +wiodąca, -komplementarna

Begin End - początek i koniec eksonu

Length - długość eksonu

Fr - ramka odczytu

P - probability - prawdopodobieństwo eksonu

Tscore - punktacja eksonu określona przez GENSCAN

Obejrzyj wyniki. Znajdź eksony 3 różnych typów i zapisz współrzędne na których zostały odnalezione, nić i ramkę odczytu oraz ich prawdopodobieństwo.

2. Czy te sekwencje rzeczywiście zawierają geny? Odszukaj prawdziwie geny odpowiadające tym sekwencjom na NCBI, znajdź nazwy tych genów. Z czym związana jest ich funkcja?

3. Na stronie bazy GENE w sekcji *Genomic regions, transcripts and products* można obejrzeć strukturę eksonów w genie w oknie graficznym. Jeśli przejdziemy do formatu GenBank, znajdziemy dokładne współrzędne eksonów.

4. Sprawdź czy predykcja GENSCANA była prawidłowa, porównaj prawdziwe eksony z przewidywanymi dla obu genów. Warto przyglądać się tym przewidywanym eksonom, które mają wysokie prawdopodobieństwo (>0.8).

4. Klasyfikacja białek

Zadanie 2.

- wejdź do bazy struktur białkowych www.rcsb.org, w wyszukiwarce wpisz 1GZX, jest to symbol struktury hemoglobiny
- jaka struktura drugorzędowa dominuje w tym białku? jak myślisz, jak wyglądałaby mapa Ramachandrana dla hemoglobiny?
- wyświetl mapę Ramachandrana dla tej struktury, w tym celu wejdź na stronę <http://molprobity.biochem.duke.edu> i wpisz 1GZX w odpowiednie pole
- czy wynik jest zgodny z twoimi przewidywaniami?
- wejdź do bazy CATH, która klasyfikuje rodziny białkowe - www.cathdb.info, wybierz browse u dołu po lewej stronie
- obejrzyj jak baza CATH dzieli białka ze względu na ich strukturę drugorzędową, wyszukaj jakieś białko, które posiada dużo arkuszy β w swojej strukturze, obejrzyj je w PDB
- wyświetl dla niego mapę Ramachandrana jak powyżej, czy mapa wskazuje na posiadanie przez strukturę dużo arkuszy β ?

Rozwiązanie obu zadań prześlij mailem do wtorku, **4.12.2018** włącznie, na adres:
jacek.smietanski@ii.uj.edu.pl

Temat wiadomości proszę opatrzyć przedrostkiem **[Bio] Lab 09**. Rozwiązaniem ma być **tylko jeden plik** – dokument PDF. Proszę o nazwanie pliku wg schematu: **Imie.Nazwisko.09.pdf**.